Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Assessing incomplete neutralization of final devoicing in German

T.B. Roettger[a,*], B. Winter[b], S. Grawunder[c], J. Kirby[d], M. Grice[a]

[a] IfL Phonetik, University of Cologne, Herbert-Levin-Str. 6, D-50931 Köln, Germany
[b] Department of Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA
[c] Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany
[d] School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh EH8 9AD, Scotland, UK

ARTICLE INFO

Article history:
Received 12 June 2013
Received in revised form
3 January 2014
Accepted 7 January 2014

ABSTRACT

It has been claimed that the long established neutralization of the voicing distinction in domain final position in German is phonetically incomplete. However, many studies that have advanced this claim have subsequently been criticized on methodological grounds, calling incomplete neutralization into question. In three production experiments and one perception experiment we address these methodological criticisms.

In the first production study, we address the role of orthography. In a large scale auditory task using pseudowords, we confirm that neutralization is indeed incomplete and suggest that previous null results may simply be due to lack of statistical power. In two follow-up production studies (Experiments 2 and 3), we rule out a potential confound of Experiment 1, namely that the effect might be due to accommodation to the presented auditory stimuli, by manipulating the duration of the preceding vowel. While the between-items design (Experiment 2) replicated the findings of Experiment 1, the between-subjects version (Experiment 3) failed to find a statistically significant incomplete neutralization effect, although we found numerical tendencies in the expected direction. Finally, in a perception study (Experiment 4), we demonstrate that the subphonemic differences between final voiceless and "devoiced" stops are audible, but only barely so. Even though the present findings provide evidence for the robustness of incomplete neutralization in German, the small effect sizes highlight the challenges of investigating this phenomenon. We argue that without necessarily postulating functional relevance, incomplete neutralization can be accounted for by recent models of lexical organization.

1. Introduction

Many languages such as Catalan, Dutch, German, Polish, Russian and Turkish contrast voiced obstruents intervocalically but neutralize the contrast syllable or word finally in favor of voiceless obstruents. An example from German is given in (1) and (2): in syllable final position, the voicing of the alveolar stop is neutralized, leading to apparent homophony between e.g. *Rad* [ʁaːt] 'wheel' and *Rat* [ʁaːt] 'council'.

(i) Rad [ʁaːt] 'wheel'; Räder [ʁæːdɐ] 'wheels'

(ii) Rat [ʁaːt] 'council'; Räte [ʁæːtə] 'councils' (1)

(i) Radschlag [ʁaːtʃlaːk] 'cartwheel'

(ii) Ratschlag [ʁaːtʃlaːk] 'advice' (2)

This asymmetrical distribution is commonly described in terms of final devoicing, a process that is often described in purely phonological terms. In fact, final devoicing in German[1] has been called the "universally recognized archetype of phonological neutralization" (Fourakis & Iverson, 1984: 141) and described as a "classic example of a phonological rule" (Wiese, 1996: 204).

In traditional formal theories of phonology, *Rad* and *Rat* are thought to differ only in their "underlying" lexical representations, while the surface form of the voiced stop is thought to be phonetically indistinguishable from that of the corresponding voiceless stop. In other words, neutralization of the final voicing distinction is assumed to be phonetically complete, resulting in homophony between the two lexical items. However, numerous experimental studies have

* Corresponding author. Tel.: +49 221 4707047; fax: +49 221 470 5938.
E-mail address: timo.roettger@uni-koeln.de (T.B. Roettger).

[1] Kohler (1984) argues that German voiced and voiceless stops are better characterized as fortis and lenis. To remain consistent with the terminology adopted in the incomplete neutralization debate, we retain the terms "voiced", "voiceless" and "final devoicing".

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2014.01.002

argued that there are small acoustic and articulatory differences between words such as *Rad* and *Rat*, suggesting that in German this neutralization is in fact *incomplete* (Charles-Luce, 1985; Dinnsen, 1985; Dinnsen & Garcia-Zamor, 1971; Fuchs, 2005; Greisbach, 2001; Mitleb, 1981; O'Dell & Port, 1983; Port & Crawford, 1989; Port, Mitleb, & O'Dell, 1981; Port & O'Dell, 1985; Piroth & Janker, 2004). Further studies suggest that listeners can distinguish “devoiced”[2] stops from voiceless ones with above-chance accuracy (Kleber, John, & Harrington, 2010; Port & Crawford, 1989; Port & O'Dell, 1985).

The results obtained in the above mentioned experiments are difficult to reconcile with traditional linguistic descriptions of German (Jespersen, 1913; Trubetzkoy, 1939; Wiese, 1996; Zifonun et al., 1997) that assume abstract phonological categories devoid of gradient phonetic information. Accounts based on this view have problems incorporating intermediate categories as the purported “semi-voiced” final obstruents. Most early formal attempts to incorporate incomplete neutralization (e.g., Charles-Luce, 1985; Port & O'Dell, 1985) involved a proliferation of post-hoc repairs (such as the “phonetic implementation rules” of e.g., Dinnsen & Charles-Luce, 1984) which led Port & Crawford (1989: 257) to claim that incomplete neutralization poses “a threat to phonological theory” (see also Port & Leary, 2005).

More recent attempts to account for incomplete neutralization are rooted in psycholinguistic models of lexical organization. There is mounting evidence suggesting that, far from being impoverished, lexical representations are rich in information, and may contain both detailed phonetic information of individual word forms (e.g., Brown & McNeill, 1966; Bybee, 1994; Goldinger, 1996, 1997; Palmeri, Goldinger, & Pisoni, 1993; Pisoni, 1997) as well as completely inflected forms (e.g., Alegre & Gordon, 1999; Baayen, Dijkstra, & Schreuder, 1997; Butterworth, 1983; Bybee, 1995; Manelis & Tharp, 1977; Sereno & Jongman, 1997). Such models of lexical organization and access assume that German speakers have inflected forms such as *Räder* in their mental lexicon. Due to its phonological and semantic relations with the singular form *Rad*, these two forms will be closely connected to each other. Ernestus and Baayen (2006) consider the possibility of incomplete neutralization effects being due to the co-activation of these related forms, i.e., when speakers pronounce *Rad*, they also activate the non-neutralized *Räder*. If some or most of the co-activated forms contain a non-neutralized segment that is fully voiced, these voiced forms could influence the motor commands used in speech production in subtle ways, leading to the observed incomplete neutralization effects.

A similar account has been advanced to explain the finding that speakers are able to distinguish forms like *Rat* and *Rad* with above-chance accuracy. Kleber et al. (2010) found that there is a greater probability of identifying a stop as voiceless after lax than after tense vowels. They further found that, following tense vowels, the (putatively neutralized) stop voicing contrast in syllable final position was recoverable more often when the stop was alveolar than when it was velar. Since in German phonologically short/lax vowels tend to occur more often before bilabial and velar voiceless stops, this suggests that sensitivity to statistical patterns of the German lexicon may affect the perception of incomplete neutralization, and thus it seems plausible that knowledge of phonotactic probabilities might play a role in production as well.

It seems safe to say that the predominant response to incomplete neutralization studies has been one of skepticism. Given that several early studies found no evidence for incomplete neutralization (Fourakis & Iverson, 1984; Jassem & Richter, 1989), some researchers have considered the debate to be settled (e.g., Kohler, 2007, 2012). However, other researchers have continued to investigate the phenomenon (e.g., Kleber et al., 2010; Piroth & Janker, 2004), and studies have since been carried out on both incomplete neutralization of final devoicing in other languages (e.g., in Dutch (e.g., Warner, Jongman, Sereno, & Kemps, 2004), Catalan (e.g., Charles-Luce & Dinnsen, 1987), Polish (e.g., Slowiaczek & Dinnsen, 1985) and Russian (e.g., Dmitrieva, Jongman, & Sereno, 2010; Kharlamov, 2012)) as well as incomplete neutralization of other processes (Bishop, 2007; Braver & Kawahara, 2012; de Jong, 2011; Dinnsen, 1985; Gerfen, 2002; Gerfen & Hall, 2001; Simonet, Rohena-Madrazo & Paz, 2008).

Thus, the debate surrounding incomplete neutralization is still very much ongoing. However, the numerically small effect sizes common across incomplete neutralization studies have attracted serious criticism on methodological grounds (Kohler, 2007; Manaster-Ramer, 1996). Fuchs (2005: 25) points out that the debate surrounding incomplete neutralization has become increasingly a debate about methodology rather than the phenomenon per se. As such, our first and foremost aim in the present work is to address the methodological and conceptual concerns raised against previous studies, thereby placing the debate surrounding incomplete neutralization on firmer empirical footing. Our second aim is to interpret our findings in light of recent psycholinguistic models of lexical organization.

In Section 2, we summarize previous empirical findings as well as their critiques, with a particular focus on Fourakis and Iverson (1984) and Jassem and Richter (1989). In Sections 3–5 we discuss the results of three production experiments that were inspired by Fourakis and Iverson's study. Section 6 presents the results of a perception experiment. In Section 7, we discuss the implications of our work for an assessment of the status of incomplete neutralization in German in light of co-activation accounts.

2. Methodological debate and the problem of “proving the null”

Across different studies, numerous phonetic properties have been found to distinguish voiceless from devoiced stops in final position. These include the duration of the preceding vowel, the closure duration, the duration of the “voicing-into-the-closure”, as well as the burst and aspiration durations. Across different studies and languages, the duration of the preceding vowel has been shown to be the most reliable correlate of obstruent “voicing” in final position. Thus in the present study we shall focus on this acoustic parameter. This has the added advantage that we avoid statistical issues surrounding multiple comparisons: with each additional measure taken into account we have an added probability of rejecting the global null hypothesis that there is no acoustic correlate of incomplete neutralization at all. Standard ways of correcting for multiple comparisons, such as Bonferroni correction, increase the probability of missing a true effect and according to Bender and Lange (2001: 347) the “easiest and best interpretable approach is to avoid multiplicity as far as possible”. We do this by focusing on vowel duration.

The direction of the vowel duration difference mirrors the durational difference in the intervocalic context, i.e., vowels tend to be longer before final devoiced stops than before final voiceless stops. Numerically, incomplete neutralization effects of vowel duration are minute. For example, Port and Crawford (1989) report a difference of 1.2–6.2 ms between devoiced and voiceless stops in German, while Warner et al. (2004) report a difference of 3.5 ms in Dutch. The magnitude of the incomplete neutralization effect appears to be dialect- and speaker-dependent (Piroth & Janker, 2004), as well as highly sensitive to the phonetic, semantic and pragmatic context (Charles-Luce, 1985, 1993; Ernestus & Baayen, 2006; Port & Crawford, 1989; Slowiaczek & Dinnsen, 1985).

As German maintains an orthographic contrast between voiced/devoiced and voiceless stops in all positions, the biggest issue surrounding previous results was the influence of this orthographic representation.[3] Most of the above-mentioned experiments used stimuli that had to be read aloud by the participants,

[2] We refer to the segment in words such as *Rad* as “devoiced”. This term is theoretically loaded because it assumes the presence of an underlying voiced segment. However, for this paper, we merely use the term as shorthand to refer to a segment corresponding to an intervocalic voiced segment within the same morphological paradigm, e.g., *Räder* [d] vs. *Rad* [t], without necessarily invoking a phonological process of devoicing.

[3] There are other concerns with incomplete neutralization studies. These include minimal pair awareness, second language proficiency of experimenter and participants and stimuli selection. These concerns have been dealt with at length in Fourakis and Iverson (1984), Manaster-Ramer (1996), Kohler (2007) and Winter and Roettger (2011).

inviting the criticism that participants used a form of hypercorrection or spelling pronunciation: as laboratory settings tend to elicit more formal and clear speech, participants might have produced words based on the written language in a way that they would not do in everyday speech.

In Fourakis and Iverson (1984) (henceforth FI), four native speakers were asked to conjugate neutralized verb forms such as *mied* (‘avoid.PST.1+3SG’) when presented auditorily with non-neutralized forms such as *meiden* (‘to avoid’). Both the duration of the preceding vowel and the closure duration were measured. No statistically significant incomplete neutralization effect was obtained. Jassem and Richter (1989) (henceforth JR) conducted a very similar study in Polish in which participants answered questions constructed by the experimenter such that the answer could be expected to consist of a single word utterance. They measured the duration of the preceding vowel, voicing-into-the-closure/frication, closure/frication duration, and, where relevant, release duration. Again, four speakers were recorded and no incomplete neutralization effect was found.

In both cases, it was concluded that the lack of a statistically significant effect supports an orthography-based explanation of incomplete neutralization. Since then, many have cited FI and JR as evidence against incomplete neutralization (e.g., Kohler, 2007, 2012; Wiese, 1996). However, these studies have methodological shortcomings of their own. For example, FI did not use minimal pairs, but instead compared words such as *mied* and *riet* (‘avoid.PST.1+3SG’ and ‘advice.PST.1+3SG’). As pointed out by Dinnsen and Charles-Luce (1984) and Port and Crawford (1989), this leaves the potential influence of the syllable onset uncontrolled for. In other words, the durational differences due to final voicing are confounded with durational differences due to properties of the initial consonant.

Both FI and JR interpret their null results as evidence for the absence of incomplete neutralization. There is a logical problem with “accepting the null”, and most researchers would argue that it is not logically sound to accept null hypotheses (e.g., Cohen, 1990; Weitzman, 1984), in line with the saying that “absence of evidence is not evidence for absence”. If anything, one can only demonstrate “sufficiently good effort” to disprove the null hypothesis (Frick, 1995). FI and JR only tested four speakers – less than most of the previous and following investigations of incomplete neutralization that *did* find an effect. Their null results may thus well be due to a lack of statistical power.

Another concern related to statistical power is that FI conducted statistical tests within speakers. Thus, for each individual test there were only a few data points. Indeed, an across-speaker re-analysis of the published FI data conducted by Port and Crawford (1989) did find significant differences consistent with incomplete neutralization. Given the low statistical power (due to lack of minimal pairs, a small number of speakers and the fact that subset analyses were conducted), it is possible that both studies committed a Type II error (i.e., failing to reject a false null hypothesis). This would not be the first time this has happened with respect to incomplete neutralization. For Dutch final devoicing, Baumann (1995) and Jongman, Sereno, Raaijmakers, and Lahiri (1992) failed to find significant incomplete neutralization effects, but Warner et al. (2004), with more speakers, did find effects.

At a bare minimum, any study that wants to demonstrate “sufficiently good effort” to disprove the null needs to have at least as many subjects and items as previous investigations *in support of* the purported phenomenon. Thus while the studies by FI and JR certainly suggest that effect sizes for incomplete neutralization are small, their results cannot be taken as counter-evidence against the phenomenon.

With regard to the perceptibility of incomplete neutralization, previous studies investigated accuracy in forced choice identification tasks. The identification accuracies were reported to be generally lower than in experiments with non-neutralized contrasts (see Brockhaus, 1995: 244, for an overview) and, in some studies, even barely above chance performance (Port & O'Dell, 1985). This leads to the question as to whether incomplete neutralization has any function in speech communication.

Previous studies used auditory stimuli for the perception experiments which come from a small set of speakers (e.g., Port & Crawford, 1989), or in some cases from just a single speaker (e.g., Kleber et al., 2010). This, together with many repetitions, gives participants ample opportunity to familiarize themselves with speaker characteristics. This in turn might make it easier for participants to detect subtle cues to voicing in a neutralizing context, enhancing the likelihood that they might be attending to cues that they would not use in listening situations outside of the laboratory.

Thus, although there is some evidence that listeners are able to exploit subtle cues to distinguish devoiced from voiceless stops in final position, the results must be interpreted with caution. While some see this as genuine evidence for incomplete neutralization as a perceptual phenomenon with potential real-world relevance, others are more inclined to view it as the result of task demands (e.g., Slowiaczek & Szymanska, 1989; Warner et al., 2004). Brockhaus (1995: 244), among many others, points out that it is not clear whether the perceptual difference between syllable-final devoiced and voiceless obstruents is actually “salient enough to be relied upon in normal communication”. Although it is not known how accurate a contrast needs to be perceived in order to play a role outside the laboratory (Xu, 2010: 334), the low accuracy scores and high variability suggest that incomplete neutralization would have little if any functional relevance in everyday communicative situations.

In summary, a number of methodological shortcomings have been identified in previous studies arguing for the existence of incomplete neutralization. However, studies that failed to find incomplete neutralization effects are, themselves, subject to methodological criticism, especially since *failure* to find an effect cannot be taken as evidence for the *absence* of that effect. The present study aims to circumvent these concerns.

Our production studies are inspired by Fourakis and Iverson's (1984) study of German final devoicing, but employ a design that has increased statistical power (more speakers, more items). We also address the concern that incomplete neutralization is potentially a result of an orthographically induced contrast. It is known that speakers automatically activate orthographic representations even in completely auditory tasks (Dehaene et al., 2010; Perre, Midgley, & Ziegler, 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998). Given that all previous studies on incomplete neutralization used real word stimuli, literate speakers inevitably know their written forms. Thus in our first experiment, we employed pseudowords, such as *Gobe* or *Gope,* in order to reduce the effect of orthography. Subjects were presented with a plural form auditorily in which the target consonant is intervocalic ([goːbə]), and were instructed to produce the singular form ([goːp]) in which the target consonant is word final. Pseudowords, which effectively have a frequency of 0, presumably lack existing orthographic representations. While it is still possible that participants think of our auditorily presented pseudowords in terms of orthography (for example, they might think of how they would spell a given pseudoword in order to produce its related singular form), the design minimizes the role of orthography *relative to other studies on incomplete neutralization*, in particular relative to FI. To the extent that orthography impacts the realization of incomplete neutralization, this should make the effect less likely to emerge.

This design, however, potentially introduces another confound: accommodation to the auditory stimuli. Phonetic accommodation, also known as phonetic convergence or phonetic imitation, involves the adaptation of a talker's speech to that of his or her interlocutor (e.g., Goldinger, 1998; Gregory & Hoyt, 1982; Natale, 1975a,b). This process happens even in situations with minimal social interaction: a number of laboratory studies have found that participants shift their pronunciation of single words towards productions of auditorily presented voices they have just heard (e.g., Babel, 2012; Goldinger 1996, 1997, 1998; Nielsen, 2011). Thus in a task in which participants are exposed to the intervocalic contrast auditorily (they hear e.g. [goːbə] or [goːpʰə]) and respond with the corresponding singular form right away, they may merely imitate the acoustics of the stimulus they have just heard. To address this issue we conducted two additional experiments, eliminating this potential confound by manipulating the acoustic cues of the intervocalic voicing distinction.

Finally, given the small effect sizes reported in the literature, we sought to evaluate the functional relevance of incomplete neutralization for speech communication. To assess the perceptibility of incomplete neutralization in a more ecologically valid design, we replicated earlier perception studies utilizing a number of different voices. If speakers consistently fail to fully neutralize the voicing contrast in final position, and/or if listeners are able to distinguish between voiceless and devoiced stops with greater than chance accuracy, this suggests that neutralization is indeed incomplete. Even if it arguably has no functional utility for communication, an explanation of this effect is nonetheless warranted, given its implications for foundational theories of phonological processing and lexical organization.

3. Production Experiment 1

3.1. Methodology

3.1.1. Participants and experimental procedure

Sixteen native speakers of German participated in the experiment (mean age: 25 years; nine women). All were undergraduates or PhD students in the humanities living in Cologne or in the area surrounding Cologne. Most of them grew up in this area and all participants claimed to speak non-dialectal Standard German. None of the participants were familiar with the concept of incomplete neutralization prior to the post-experiment debriefing.

The recording session was managed by a native speaker (the first author) and conducted entirely in German. Participants were seated in a well-illuminated sound-treated booth in front of a computer screen. They were given written instructions that stated that the experiment investigates German plural formation. None of the participants reported noticing the presence of minimal pairs in the post-experimental interview. This addresses previous concerns surrounding the idea that incomplete neutralization effects might be artificially enhanced because of hyperarticulation due to participants noticing the final voicing alternation (see discussion in Winter & Roettger, 2011). After the written instructions, the remaining procedure was conducted auditorily. In each trial, participants first heard a stimulus sentence such as (3) and then produced a corresponding sentence such as (4).

Plural stimulus :

Aus Dortmund kamen die ***Drude***. (3)

From Dortmund come.3PL.PST DEFDET.PL.NOM NONCE-PL
From Dortmund came the NONCE-PL.

Singular response :

Ein ***Drud*** *wollte nicht mehr*. (4)

INDFDET.SG.M.NOM. NONCE-SG want.3SG.PST NEG longer
One NONCE-SG did not want to continue.

The experiment was run using Superlab 2.04 (Abboud, 1991). At the beginning of each trial, a cross appeared in the center of the screen (+) and participants heard the plural sentence through headphones. After presenting a blank screen for 500 ms three question marks appeared on the screen. Participants were now asked to produce the corresponding singular sentence. The experiment was self-paced and there were no time constraints.

Prior to the actual experiment, participants listened to eight demonstration stimuli, each of which was a plural sentence followed by a singular response. None of these demonstration items were potential critical items, and none included a voiced/voiceless obstruent distinction. This was done so as not to bias our participants' responses with respect to incomplete neutralization. After the demonstration, participants performed eight practice trials where they had to produce the response sentences themselves. The actual experiment was divided into four blocks. After each block, there was an obligatory break of at least ten seconds. On average, the entire experiment (including instruction and debriefing) took about 30 min.

3.1.2. Speech material

The experimental items consisted of 24 pseudoword pairs such as (5)–(7) (see Appendix A):

Gobe [goːbə] vs. Gope [goːpʰə] (5)

Frade [fraːdə] vs. Frate [fraːtʰə] (6)

Schuge [ʃuːgə] vs. Schuke [ʃuːkʰə] (7)

All pseudowords were trochaic and complied with German phonotactic rules. There were eight bilabial, seven alveolar and nine velar stimulus pairs, each containing one of the vowels /aː, oː, uː, iː, aʊ̯/. Each experimental item was introduced as a masculine noun inflected for the plural. Plural inflection was indicated through the regular plural marker for masculine nouns (/-ə/), the plural determiner /diː/ and number agreement on the verb. The German plural system exhibits many irregularities, and we chose the particular plural form used in this study because it is the most likely plural of monosyllabic masculine nouns (e.g., *Arm/Arme* ‘arm/arms’, *Stift/Stifte* ‘pen/pens’, etc.). We did not choose the commonly occurring plural ending *–en* because speakers are more insecure as to which singular form corresponds to pseudowords ending in *–en* (as a pilot study demonstrated), and because this marker often involves schwa deletion and a nasal release, which might in turn lead to an additional lengthening of the preceding vowel.

As German plural formation is very complex, we needed to norm our stimuli with respect to their morphology. A list of the intended singular forms was given to a group of five participants who were asked to provide the respective plural forms. Indeed, the schwa-plural (/-ə/) was the most frequent response pattern (84% of all responses). However, as expected, some nonsense words were more consistently formed with this morpheme than others. The extent to which a stimulus was identified as schwa-plural was included in the statistical analyses presented below.

To further alleviate the concern of hyperarticulation due to minimal pair awareness, we included 96 fillers (2/3 of the total stimulus set), 70% of which contained an umlaut vowel. As plural forms with umlaut vowels sometimes do and sometimes do not require a vowel change (e.g., *Turm*>*Türme* ‘tower/towers’ but *Bär*>*Bären* ‘bear/bears’), we hoped this would increase the salience of the fillers, simultaneously detracting attention from the critical stimuli. Forty different city names (randomized over stimulus pairs) were embedded in the carrier phrase to introduce an additional distracting element, but in all other respects the carrier phrase (‘Aus (CITY NAME) kamen die (NONCE PLURAL)’) remained constant. We avoided repetition of items to further decrease the salience of the relevant minimal pairs. The 144 stimuli and the 16 demonstration and practice items were spoken by a native speaker of German (male, trained phonetician) and recorded in a sound-treated booth with an AKG C420 III microphone. All stimuli were randomized and divided into four blocks. Members of a stimulus pair were always within different blocks. At the beginning of the experiment, each participant was randomly assigned to one of eight block orders.

3.1.3. Acoustic analysis of stimuli

We performed acoustic analyses of the plural stimuli that were presented to participants to ensure that the stimuli have the typical acoustic characteristics of German voiced and voiceless stops (Keating, 1984; Kohler, 1984). Using Praat (Boersma & Weenink, 2011), we measured the duration of the vowel preceding the critical stop, the closure duration, the duration of the following vowel, the burst duration, the voice onset time and the median intensity of the burst. In addition, we analyzed the mean fundamental frequency (f_0), as well as the f_0 in the first quintile of the vowel following the stop release.

The vowel preceding the critical stop was on average 28 ms (SE = 3.7) longer before intervocalic voiced stops than before intervocalic voiceless stops ($\chi^2(1)=30.27$, $p<0.001$)[4]; there was no significant difference in the following vowel ($\chi^2(1)=0.04$, $p=0.99$). Voice onset times were on average 42 ms (SE = 2.3) longer for voiceless stops ($\chi^2(1)=65.57$, $p<0.0001$), and closures were on average 21 ms (SE = 1.56) longer for voiceless stops ($\chi^2(1)=52.8$, $p<0.0001$). There were no significant differences for burst duration ($\chi^2(1)=1.57$, $p=0.85$) or burst intensity ($\chi^2(1)=0.37$, $p=0.99$), nor were there differences of mean f_0 ($\chi^2(1)=2.17$, $p=0.7$) or of the first quintile f_0 in the second vowel ($\chi^2(1)=2.75$, $p=0.56$). Furthermore, all but one of the voiced stimuli had consistent voicing during the closure, meaning that vocal fold vibration was a reliable and consistent cue. Thus, the stimuli that were given to participants are relatively typical German voiced and voiceless stops. We found large differences in vowel durations before voiced and voiceless stops, the closure duration and the voice onset time in addition to voicing during the closure. This means that there are at least four robust cues for participants to distinguish between the voiced and the voiceless stimuli intervocalically. We presented these stimuli to five male and five female German participants who were able to retrieve the voicing status with 98% accuracy. Having described the phonetic details of the stimuli, we now turn to the measurements of the responses.

3.1.4. Acoustic analysis of responses

Participant responses were digitized at a sampling rate of 44.1 kHz (16 bit). The durations of the vowels preceding the final stops were measured by the first author. If the sound preceding the vowel/diphthong was a stop, the onset of the vowel was defined as the onset of voicing in cases of voiceless stop or as the end of the burst in cases of voiced stops. A sudden discontinuity in the spectrogram was taken as the onset of vowels following fricatives ([ʃ]), nasals ([m] and [n]), laterals ([l]) and palatal approximants ([j]) (e.g., [ʃuːk], [muːp], [bloːk] or [jiːt]). The end of the vowel was defined as the end of the second formant of the vowel, which usually coincided with a sudden drop in amplitude of voicing. To assess the interaction between incomplete neutralization and prosodic factors, we also coded certain aspects of the prosodic realization including the accent position and the presence of a potential prosodic boundary following the critical item.

3.1.5. Statistics

All data were analyzed with generalized linear mixed models, using R (R Core Team, 2012) and the package *lme4* (Bates, Maechler, & Bolker, 2012). For the production experiments (Experiments 1, 2 and 3), we used a Gaussian error distribution (assuming normality). We adhere to the random effect specification principles outlined in Barr, Levy, Scheepers, and Tily (2013). We included a term for random intercepts for participants and items, which quantifies by-participant and by-item variability in overall vowel duration (i.e., the fact that some speakers tend to produce longer or shorter vowels). The critical fixed effect in question was VOICING (i.e., voiced vs. voiceless in the plural form, where voiced = /b,d,g/ and voiceless = /p,t,k/), and for this fixed effect, we included correlated random slopes for participants and items (this quantifies by-participant and by-item variability in the effect of VOICING).

In our model selection process, we conceptually separated the fixed effects into control variables and the test variable (VOICING). ACCENT TYPE and PROSODIC BOUNDARY were two prosodic control variables. If either one of these had led to a significant interaction with VOICING, this would have indicated that the amount of incomplete neutralization depends on prosodic conditions, and could have been the result of hyperarticulation due to e.g. prosodic strengthening (e.g. Cho & Keating, 2009). VOWEL QUALITY and PLACE OF ARTICULATION (bilabial vs. alveolar vs. velar) were also included to explain residual variance. Since processing differences could lead participants to perceive some singular forms as better matches to their corresponding plural forms than others, we included the results of our stimulus norming as an additional control variable, PLURAL ASSOCIATION.

We first tested whether VOICING interacted with the control variables by performing a likelihood ratio test between a model containing interactions and a model containing main effects only. We then excluded the interaction between VOICING and all control variables. *P*-values were generated using likelihood ratio tests.

3.2. Results

VOICING had a significant effect on vowel duration in the singular form ($\chi^2(1)=13.76$, $p<0.0002$), with vowels estimated to be 8.6 ms longer before devoiced stops rather than to voiceless stops (SE = 2.03 ms). The effect of VOICING on vowel duration was fairly consistent across participants and items, as can be seen in Fig. 1. Overall, 14 out of 16 participants and 20 out of 22 items exhibited longer vowels preceding devoiced stops than preceding voiceless stops. Descriptive inspection of the data did not indicate that the effect was dependent on any item specific phonotactic characteristics (cf. Appendix A for a detailed listing of vowel differences for each stimulus pair separately). This was statistically validated: there were no interactions between VOICING and any of the control variables ($\chi^2(10)=9.45$, $p=0.49$). This means that the effect of VOICING on vowel duration did not depend on either of the prosodic variables (ACCENT TYPE or PROSODIC BOUNDARY),[5] nor on the variables VOWEL QUALITY, PLACE OF ARTICULATION, or PLURAL ASSOCIATION.

[4] Here and subsequently we report likelihood ratio tests between hierarchical linear regression models ("mixed models") with the fixed effect VOICING and random intercepts Item (no random effect for Speaker is needed as there is only one Speaker), as well as random slopes for Voicing dependent on Item. *P*-values were corrected for multiple testing by means of Dunn–Šidák correction.

[5] The target word was deaccented (EIN *Gop wollte nicht mehr*), or accented in prenuclear (*Ein Gop* WOLLTE *nicht mehr; Ein Gop wollte* NICHT *mehr*) or nuclear position (*Ein* GOP *wollte nicht mehr*).

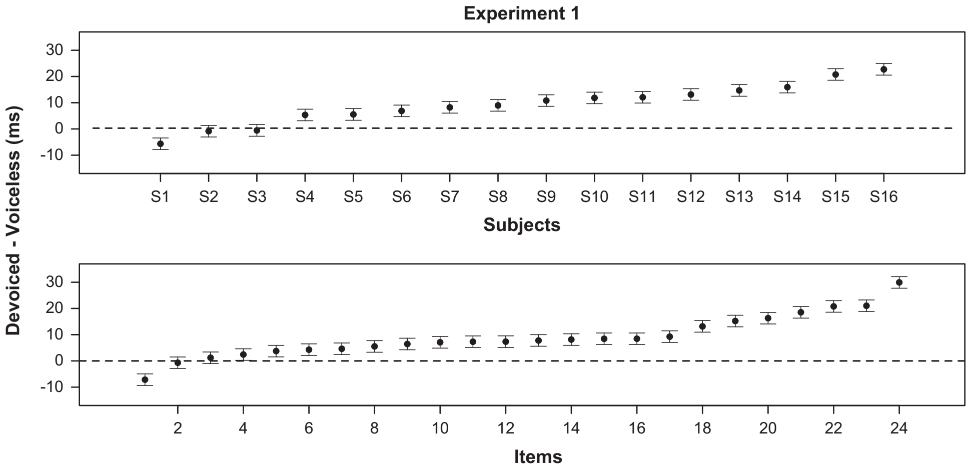

Fig. 1. Results of Experiment 1. Difference in vowel duration between stops in final position corresponding to voiced and voiceless stops in intervocalic position (‘devoiced’ and ‘voiceless’, respectively). Means are arranged according to size for all subjects (upper plot) and items (lower plot) separately. Error bars indicate standard errors taken from the model described in Section 3.1.5. Dashed lines indicate no difference between devoiced and voiceless stops.

Pseudowords are by definition unusual for our participants, and so we were concerned about problems with the task. To see whether our results might be disproportionately affected by a few extremely unusual responses, we performed subset analyses in which we excluded all responses where the pseudoword was either incorrectly remembered (e.g., substituting the vowel /iː/ for /eː/), or produced with considerable hesitation. This led to a removal of 10.02% of our data points, a considerable reduction of the size of the dataset. Nevertheless, even with these responses excluded, the main effect still obtains ($\chi^2(1)=13.34$, $p<0.0003$), with vowels being 8.6 ms longer (SE = 2.04 ms) before devoiced stops.

Our results indicate a successful extension of the FI study of incomplete neutralization in German, i.e., speakers produce longer vowels before stops corresponding to voiced stops in the plural form (‘devoiced’) than to the voiceless stops in the plural form (‘voiceless’). However, there is a potential confound inherent to our design: namely, the pronounced differences in vowel duration between intervocalic voiceless and voiced stops in the acoustic stimuli (Section 3.1.3). This raises the possibility that the observed effect is simply an artifact of phonetic accommodation: since all stimuli were produced by a single speaker, participants may simply have been imitating the vowel duration differences when producing the singular forms with the stop in final position. This would be in line with previous research demonstrating that speakers shift their pronunciation towards productions of auditorily presented voices they have just heard (e.g., Babel, 2012; Goldinger, 1998; Nielsen, 2011). Experiments 2 and 3 were conducted to address this issue. By systematically manipulating the durational cue of the intervocalic stops in the plural forms, we can evaluate the potential impact of phonetic accommodation on the observed incomplete neutralization effect.

4. Production Experiment 2

4.1. Methodology

4.1.1. Participants and experimental procedure

Sixteen speakers participated in Experiment 2 (mean age: 27 years; 9 women). Background details of participants are as stated for E1. None of the participants in E2 had participated in E1. All details of the procedure were the same as in E1 if not stated otherwise.

In Experiment 2, we used a different carrier sentence. For each trial, participants first heard a sentence such as (8) and then produced a corresponding sentence such as (9).

Plural stimulus :

Peter weiβ nun, wie die Bauge aussehen. (8)

Peter know 3SG.PRS now how DET.PL.NOM NONCE-PL look
Peter knows now what the NONCE-PL look like.

Singular response :

Denn nur der Baug sieht so aus. (9)

Because only DET.SG.M.NOM NONCE-SG look.3SG.PRS like PART
As only the NONCE-SG looks like this.

4.1.2. Speech material, stimulus manipulation and norming

The experimental items consisted of 48 pseudoword pairs (see Appendix B). There were 24 stimulus pairs with labial and 24 with velar stops, each of which followed one of the vowels /iː, eː, aː, aʊ̯, oː, uː/.[6] To minimize measurement difficulties, alveolar stops (which may show coarticulatory effects of the following word) were excluded.

[6] Acoustical analyses of the stimuli show that the vowel preceding the critical stop was on average 23.78 ms (SE = 2.62) shorter before voiceless stops ($\chi^2(1)=48.547$, $p<0.0001$). The mean closure duration was 13.7 ms (SE = 1.74) longer for voiceless stops ($\chi^2(1)=40.32$, $p<0.0001$) and VOTs were on average 47.34 ms longer for voiceless stops (SE = 1.75, $\chi^2(1)=208.28$, $p<0.0001$). All of the voiced stimuli had voicing during the closure. Thus, as in E1, robust cues to the voicing status of the critical stop were present in intervocalic position.

Stimuli were balanced for vowel quality. As in E1, each experimental item was introduced as a masculine noun inflected for plural, and a norming study found that the schwa-plural (/-e/) was the most frequent response pattern (82% of all responses). Unlike E1, there were no fillers, making the contrast between the corresponding members of a minimal pair more obvious to participants and potentially leading to an enhancement of the effect under investigation (cf., Jassem & Richter, 1989), which in turn might make a potential confound effect of accommodation easier to detect. The 48 stimulus pairs were spoken by a native speaker of German (male, trained phonetician) along with the demonstration and practice items in a sound-treated booth recorded with an AKG C420 III microphone.

To evaluate the potential impact of phonetic accommodation, we manipulated the duration of the vowel preceding the intervocalic stops in the plural forms (e.g. /oː/ in /goːpə/). We took the mean difference in vowel duration preceding voiced and voiceless stops produced by the speaker as a baseline: Vowels preceding voiced stops were 16% longer than vowels preceding voiceless stops. We then manipulated vowel durations of both members of a minimal pair using TD-PSOLA (Time-Domain Pitch Synchronous OverLap-Add, Moulines & Charpentier, 1990) resynthesis as implemented in Praat (Boersma & Weenink, 2011) selecting a 10 ms Hanning window for analysis. Fundamental frequency was not manipulated. As this was a between-items design, each stimulus pair was manipulated only once and assigned to one of four sets. Stimuli in set A were edited to have a difference in vowel duration of 32% (henceforth *enhanced*); that is, vowels preceding voiced stops were 32% longer than vowels preceding voiceless stops (twice as long as the baseline condition). Stimuli in set B were edited to have a difference in vowel duration of 16% (henceforth *original*), similar to the baseline. Vowel durations of stimuli in set C did not differ at all (henceforth *neutralized*), meaning that vowel duration as a cue to intervocalic voicing was neutralized. Stimuli in set D were manipulated so that vowels preceding *voiceless* stops were 16% longer than those preceding voiced stops (henceforth *reversed*). In other words, set D contained stimuli where the effect of voicing on vowel duration was the mirror image of the baseline. The stimuli were judged to sound natural by a native speaker of German.

Additionally, we examined the perceptual robustness of the voicing distinction in the manipulated forms by conducting a norming study. Five native speakers of German (mean age: 25) were asked to decide whether the presented stimuli were voiced or voiceless in a forced-choice identification task. The norming study confirmed that the voicing contrast is very easy to perceive for all manipulation conditions: participants did not make any errors in identifying the voicing category. Even though we manipulated one perceptual cue to the voicing distinction, participants were able to rely on other cues like voicing during the closure, VOT and closure duration.

4.1.3. *Stimulus presentation, acoustic analyses of responses and statistics*

All stimulus presentations were randomized for each participant. The actual experiment was divided into four blocks. The first two blocks contained all 48 critical pairs, balanced for place of articulation of the stop, vowel quality and condition (voiced or voiceless). A subset of these items was repeated twice in blocks three and four. Corresponding members of a minimal pair in the first two blocks were separated by one block (so by at least 24 items). The acoustic analysis and statistical analysis was performed as specified for E1. In our model selection process, we separated the fixed effects into control variables (PLACE OF ARTICULATION, VOWEL QUALITY, PLURAL ASSOCIATION and REPETITION) and test variables (VOICING and MANIPULATION CONDITION). Statistical analyses were performed as specified for E1.

4.2. *Results*

We found an interaction between MANIPULATION CONDITION and VOICING ($\chi^2(1)=7.01$, $p=0.008$). For each manipulation step (*enhanced*⩾*original*⩾*neutralized*⩾*reversed*), the estimated difference between devoiced and voiceless stops became 1.49 ms smaller (SE = 0.57 ms). Interestingly, an incomplete neutralization effect was observed even in the *neutralized* and the *reversed* conditions, where the duration cue was at best uninformative. There was also a main effect of VOICING ($\chi^2(1)=12.76$, $p=0.00035$), with vowels being overall 4.3 ms shorter (SE = 1.02 ms) before voiceless stops than before devoiced stops (pooled across different manipulation conditions; see Fig. 2).

As in Experiment 1, we did not find any interactions between VOICING and any of the control variables ($\chi^2(8)=3.54$, $p=0.89$), suggesting that PLURAL ASSOCIATION, PLACE OF ARTICULATION, VOWEL QUALITY and REPETITION, did not have an effect. Subsequent inspection did not reveal any interaction of item specific phonotactic characteristics (see Appendix B for a detailed listing of vowel differences for each stimulus pair).

The results of Experiment 2 show that manipulation of the vowel duration in the plural stimulus affected the degree to which neutralization was incomplete. Nonetheless, there was still a significant overall effect of incomplete neutralization in the expected direction, even in the *reversed* condition. In other words, even though in a quarter of cases the input stimuli exhibited shorter vowel durations preceding *voiced* stops, participants produced shorter vowel durations preceding *voiceless* stops, suggesting that the effect of accommodation, if present, was at best small.

However, since this experiment employed a between-item design, all participants were prompted with items from all four manipulation conditions. Thus, we cannot rule out the possibility that stimuli of one condition might have influenced those of other conditions ("carry-over effects"). In addition, the manipulation conditions were not perfectly balanced; there was an overall duration advantage for vowels preceding voiced stops of +16% (adding up all conditions, 32%, 16%, 0%, –16%). This advantage is actually biased towards incomplete neutralization, as participants might have adapted to the overall 16% vowel duration difference, which might explain the persistence of the effect even in the *neutralized* and *reversed* conditions. To rule out this possibility, we conducted a third experiment with a between-*subjects* design and balanced manipulation conditions.

5. Production Experiment 3

5.1. *Methodology*

5.1.1. *Participants and experimental procedure*

Sixteen speakers participated in Experiment 3 (mean age: 24 years; 10 women). Background details of participants are as stated for E1 and E2. None had participated in the previous experiments. All details of the procedure were the same as in E2 if not stated otherwise.

(*footnote continued*)
As there was no interaction between manipulation condition and voicing for the parameters ($\chi^2(1)\leq 3.91$, $p\geq 0.27$), we may conclude that there were no differences of intervocalic voicing cues between conditions.

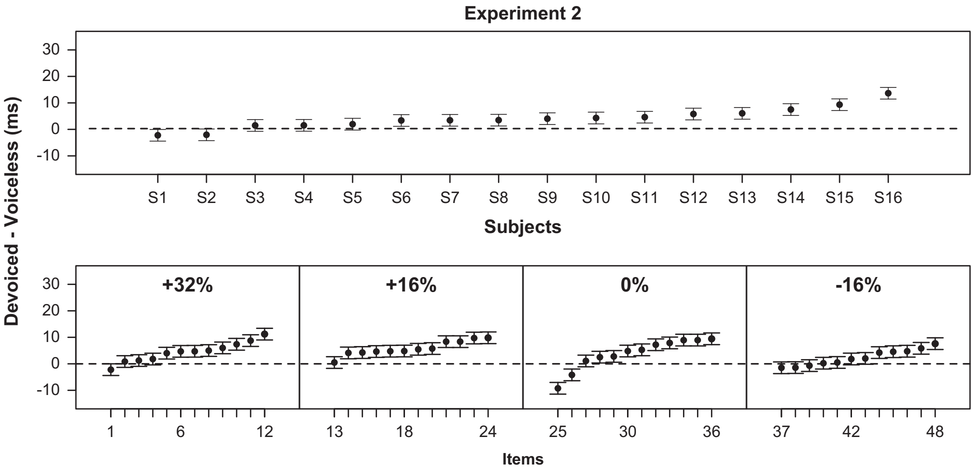

Fig. 2. Results of Experiment 2. Difference in vowel duration between devoiced and voiceless stops. Means are arranged according to size for all subjects (upper plot) and items (lower plot) separately. The lower plot shows vertical solid lines that separate the between-item manipulations of vowel duration. Dashed lines indicate no difference between voiced and voiceless stops.

5.1.2. Speech material, stimulus manipulation and norming

The 24 stimulus pairs consisted of a subset of the items used in E2 (cf. Appendix B).[7,8] Again, we manipulated the vowel durations as stated for E2. Each minimal pair was manipulated twice resulting in two stimulus sets: In set A, the difference in vowel duration was 32% (henceforth *enhanced*), that is vowels preceding underlying voiced stops were 32% longer than vowels preceding voiceless stops. In set B, the difference in vowel duration was 32% in the opposite direction (henceforth *reversed*).

As was done for E2, we examined the perceptual robustness of the voicing distinction in the manipulated forms by conducting a norming study. Five native speakers of German (mean age: 24) were asked to decide whether the presented stimuli were voiced or voiceless in a forced-choice identification task. All stimuli in both manipulation conditions were presented to all participants. The norming study confirmed that the voicing contrast is very easy to perceive: for the enhanced condition, participants were 100% correct in identifying the voicing of a stop, and for the reversed condition they were 99% (=3 incorrect tokens) correct.

5.1.3. Stimulus presentation, acoustic analyses of responses and statistics

Subjects were randomly assigned to one of two groups. Group A was presented with stimuli from set A only (*enhanced* stimuli), and group B was presented with stimuli from set B only (*reversed* stimuli). All stimuli were randomized for each participant. The actual experiment was divided into three blocks. In each block each stimulus was presented once resulting in three productions of each stimulus. The acoustic and statistical analyses were performed as specified for E2.

5.2. Results

As opposed to Experiment 2, there was no interaction between MANIPULATION CONDITION and VOICING ($\chi^2(1)=1.25$, $p=0.26$). However, numerically there was a small impact of manipulation condition, with the incomplete neutralization effect being 2.69 ms (SE=2.42) smaller in the *reversed* condition. For the *enhanced* condition, the predicted difference between devoiced and voiceless stops was 4.1 ms (SE=3.12). The difference between the two manipulation conditions, while not significant, resembles the effect seen in E2; however, the main effect of VOICING did not reach significance ($\chi^2(1)=1.62$, $p=0.2$), with vowels being only 1.75 ms shorter (SE=1.32 ms) before voiceless stops (cf. Fig. 3). Experiment 3 thus marks a failure to replicate the incomplete neutralization effect. Subsequent inspection of the results did not indicate any interaction of underlying voicing with item-specific phonotactic characteristics (cf. Appendix B for a detailed listing of vowel differences for each stimulus pair).

Before we proceed to the perception experiment, we summarize the results of the three production experiments and discuss their implications for the incomplete neutralization debate.

5.3. Discussion of production results

In Experiment 1 we found a difference in vowel duration depending on the voicing status of the intervocalic stop in the (plural) stimulus form. Thus, we were able to demonstrate that neutralization of the voicing contrast in final position is incomplete in terms of the duration of the preceding vowel, even when the influence of orthography was minimized by using auditory presentation of pseudowords (which are presumed to lack pre-existing orthographic representations). The pattern was found to be consistent across different individuals and stimuli even when controlling for variation between different participants and items. Furthermore, we found no interactions between the incomplete neutralization effect and any of the other variables that we controlled for. This is noteworthy, as it suggests that the incomplete neutralization effects were not altered by prosodic characteristics or place of articulation, suggesting relative independence from these factors.

[7] Acoustical analyses of the stimuli show that the vowel preceding the critical stop was 16.56 ms (SE=3.86) shorter before voiceless stops ($\chi^2(1)=14.12$, $p=0.00017$). The closure duration was 13.71 ms (SE=1.95) longer for voiceless stops ($\chi^2(1)=27.62$, $p<0.0001$). VOTs were on average 47.54 ms long for voiceless stops (SE=2.84, $\chi^2(1)=94.04$, $p<0.0001$). All of the voiced stimuli had voicing during the closure. So as stated for E1 and E2, there were robust cues for the voicing status of the critical stop in intervocalic position.

[8] Due to a coding error, one stimulus pair had to be excluded from the analysis.

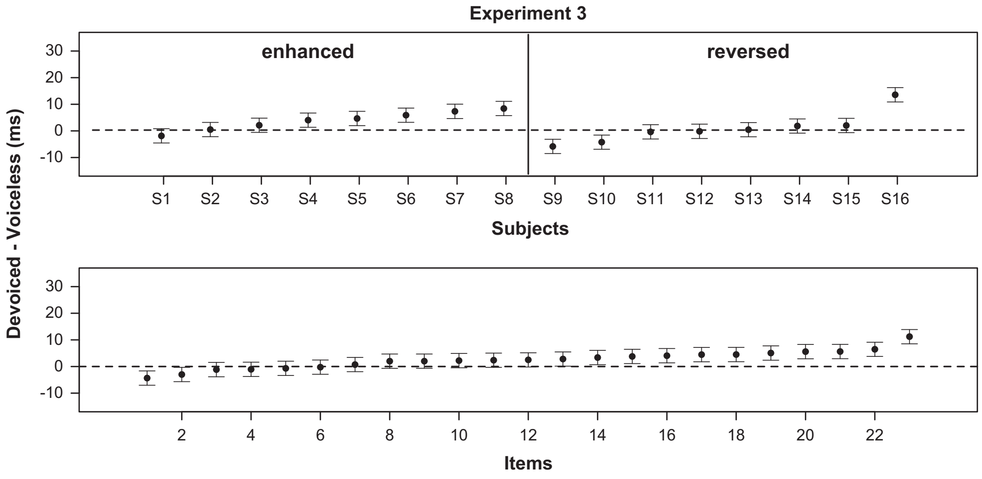

Fig. 3. Results of Experiment 3. Difference in vowel duration between devoiced and voiceless stops. Means are arranged according to size for all subjects (upper plot) and items (lower plot) separately. The upper plot shows a vertical solid line that separates the between-participants manipulation. Dashed lines indicate no difference between devoiced and voiceless stops. The lower plot shows 23/24 items because one item had to be excluded due to a coding error.

Finally, our debriefing indicated that participants perceived the task to be a morphological one – none were aware of the fact that we were looking specifically at minimal pairs such as [goːbə] and [goːpʰə]. This suggests that our distraction devices (instructions, difficult fillers, different city names) were successful, and that task demands and strategic responses were unlikely to play a significant role. We can then safely conclude that we have demonstrated the existence of an incomplete neutralization effect while avoiding the methodological shortcomings that may have impacted previous findings.

Experiment 2 replicated the findings of Experiment 1 and ruled out a potential confound, namely accommodation to the input stimuli. We demonstrated an incomplete neutralization effect of vowel duration in four manipulation conditions. Participants produced incomplete neutralization effects in the expected direction even when they were prompted with intervocalic cues providing evidence in the opposite direction. Although there was a statistically significant difference between the manipulation conditions, and thus potentially a residual effect of accommodation, this effect was numerically very small.

The incomplete neutralization effect was even smaller in Experiment 3, where we manipulated vowel durations in a between-subjects design. Whereas in the *enhanced* condition, there still was a numerical difference between vowels before devoiced and voiceless stops that was of similar magnitude as in the other experiments, this difference was even further diminished in the *reversed* condition. The latter condition is strongly biased against an incomplete neutralization effect, as all of the stimuli are manipulated so as to make accommodation counteract the vowel duration differences predicted by incomplete neutralization. It should also be pointed out that a between-subjects design inherently reduces statistical power. It is therefore unsurprising that we failed to replicate an incomplete neutralization effect in Experiment 3. As has been observed repeatedly in the literature on final devoicing in Dutch and German, incomplete neutralization effects are brittle and can be difficult to detect with inferential statistics (Baumann, 1995; Fourakis & Iverson, 1984; Jongman et al., 1992; Warner et al., 2004).

We now turn to the role of orthography. Given that literate adult speakers constantly and habitually associate phonological with orthographic forms (Perre et al., 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998), participants might have mentally constructed orthographic representations "on the fly". Thus, a given participant that has just heard a pseudoword such as [goːbə] might have activated an orthographic mental representation of that word, despite our solely auditory task design. It should be emphasized, however, that the magnitude of the effect that we obtained for vowel duration is comparable to previous studies that *did* have orthographic representations as the input.

Although we minimized the role of orthography at least to the same extent as Fourakis and Iverson (1984), our use of pseudowords comes with its own set of problems. For one thing, as pseudowords are necessarily unknown and unfamiliar (and thus have a frequency of 0), they may be more likely to be hyperarticulated (see Whalen, 1991, 1992). This, however, does not seem to be the case in our data. The overall vowel durations in Experiment 1, for example, are lower than previously reported ones: our mean was 156 ms (SD = 44 ms), whereas Port and O'Dell (1985: 459) reported 202–305 ms and Charles-Luce (1985: 315) reported 184–211 ms, suggesting that relative to these other experiments, our participants were if anything hyperarticulating *less*.

Furthermore, pseudowords always introduce the possibility of analogy to real words. The difference in frequency of different V-C sequences in the lexicon is the main source of potential analogical asymmetries. For example, tense/long vowels tend to precede voiced bilabial/velar stops (e.g., /liːbə/ 'love', /fliːgə/ 'fly'), while lax/short vowels before voiced bilabial/velar stops are very rare (e.g., /ɛbə/ 'tide'). We addressed this issue by manual inspection of the data and adding place of articulation as an effect to our statistical models. We found no noteworthy pattern, suggesting that any item-specific effects of individual pseudowords are marginal.

Finally, we checked for the possibility that singular-plural formation was more difficult for some stimuli than for others, which could introduce a potential confound. Data collected in our norming studies were used to predict production results, but no effect of plural formation preference was observed. We conclude that any processing difficulties due to idiosyncratic properties of the stimuli are of minor importance for our results.

We now turn to the perception experiment.

6. Perception experiment

The production experiments confirmed that neutralization of the voicing contrast in final position is indeed incomplete. We have ruled out a number of potential methodological reasons for this incompleteness. However, as mentioned in Section 1, it is not clear what, if any, functional role incomplete neutralization plays in speech communication. To further our understanding of the perceptibility of incomplete neutralization, our fourth experiment sets out to replicate and extend earlier studies of incomplete neutralization in perception. Previous studies used auditory stimuli from a small set of speakers, or even just a single speaker (e.g., Kleber et al., 2010; Port & Crawford, 1989). But are listeners able to discriminate between final stops corresponding to intervocalic voiced and voiceless counterparts when they are confronted with a multitude of speakers? For a more ecologically valid assessment of incomplete neutralization in perception, our experiment confronted listeners with productions from all of the participants of Experiment 1.

6.1. *Methodology*

6.1.1. *Participants and experimental procedure*

Sixteen listeners participated in the experiment, none of whom had participated in any of the preceding experiments. All participants were native speakers of German with no reported hearing deficits (mean age: 30 years; five women). Two of the participants were authors of this study (the first and the second author, both from the Cologne/Rhine region), neither of whom performed remarkably better or worse than naïve participants, thus showing that even extensive familiarity with the training stimuli does not affect the results of this experiment. The remaining participants were either living in Cologne or in Leipzig. Regardless of their origins, all participants claimed to speak non-dialectal Standard German.

Participants heard the response sentences spoken by the speakers of Experiment 1. They were asked to decide whether the presented stimulus corresponded to an intervocalic voiced or voiceless stop by choosing the appropriate written presentation of a word (e.g., *Gob* vs. *Gop*). These were presented on the left and the right side of the screen (counterbalanced), and participants had to press a left or right button on the computer keyboard. Because we expected ceiling effects in the direction of the voiceless response, the instructions emphasized that exactly half of the stimuli were from the set <b,d,g> and half were from the set <p,t,k>. In order to control for the possibility of a speed-accuracy trade-off, we also measured reaction times. The procedure was run using E-Prime 2.0 (Schneider, Eschman, & Zuccolotto, 2002).

6.1.2. *Speech material*

The experiment was designed to capture immediate success in perceiving the distinction between stops corresponding to voiced and voiceless stops in intervocalic position as well as long-term success over many trials and repetitions. Recall that Experiment 1 contained 748 critical stimuli (16 speakers*24 items*2 voicing conditions). In order to mitigate the potential effects of participant fatigue, we sampled a subset of these data (192 items) to use as stimuli in the perceptual study.

In order not to handpick particular items, semi-random subsets of 192 items (12 items per speaker) were sampled from the set of critical stimuli (the sampling was semi-random in order to insure that each speaker and each item was represented). In order to make sure that the stimuli included acoustic evidence of incomplete neutralization, we chose the first subset with a significant incomplete neutralization effect. Out of this subset we constructed four lists that constituted the four blocks of the experimental procedure. In each block, each stimulus pair (e.g., *Gob* vs. *Gop*) and each speaker appeared at least once. Each devoiced/voiceless combination came from the same speaker (e.g., *Gob* and *Gop* in list 1 were both from speaker 4). Given that there were 24 item pairs but only 16 speakers, 8 speakers had to be re-used and their voices appeared twice per block.

6.1.3. *Statistics*

There are several ways of analyzing this type of perception data. The traditional way works with d', a sensitivity index derived from Signal Detection Theory (Green & Swets, 1966). d' takes into account a subject's response bias (here, their inclination to respond with “voiced” or “voiceless”) in order to compute a measure of perceptual sensitivity. We calculated d' per subject, per item and per speaker voice and performed one-sample *t*-tests against $d'=0$ for each of these measures. While this traditional analysis already takes bias into account, we cannot use it to analyze the effects of other measures on accuracy, such as response times and trial order. For this, we used a mixed logistic regression model with “accuracy” (0 or 1) as the dependent measure. As fixed effects we included mean-centered RESPONSE TIMES, TRIAL ORDER and REPETITION. We included correlated random slopes for subject, item and speaker voice. If the *intercept* of this model is significantly above zero, we can conclude that participants are able to perceive the voicing contrast in the neutralization context with above chance accuracy.

6.2. *Results*

Fig. 4 displays d' per subject, item and speaker voice. Overall, d' was fairly low, indicating that sensitivity to the voicing of final stops, when response bias was controlled for, was poor. *T*-tests indicate that d' is significantly above zero by subjects ($t(15)=7.1$, $p<0.0001$), items ($t(23)=4.419$, $p=0.00019$) and speaker voices ($t(15)=3.79$, $p=0.0017$), with estimates of 0.25 (SE=0.036), 0.29 (SE=0.066) and 0.27 (SE=0.072), respectively.

In the mixed logistic regression analysis, there were no effects for TRIAL ORDER ($\chi^2(3)=5.53$, $p=0.14$) or REPETITION ($\chi^2(3)=5.01$, $p=0.17$). The absence of an effect of REPETITION indicates that participants were no more likely to respond correctly the second time they heard the same item spoken by the same voice. This suggests that there was no familiarization effect. The absence of an effect of TRIAL ORDER on accuracy indicates that there was no overall learning effect either. There was, however, a significant effect of RESPONSE TIME ($\chi^2(3)=8.88$, $p=0.03$). Although faster responses were less accurate, the decrease was very small, with only a 5% decrease in accuracy per SD of response times (log odds: −0.053, SE=0.027).

Crucially, the intercept of this analysis was positive and significant ($p<0.0001$), with an estimated overall accuracy of 55% (log odds: 0.35, SE=0.09). This indicates that listeners were, on average, more likely to respond correctly than incorrectly. If we add CORRECT VOICING as a predictor (whether the spoken word was the intervocalic counterpart of a voiced or voiceless stop), we can divide up the results according to whether tokens have voiced and voiceless counterparts and look at differences in accuracies for these two conditions. With this model, participants were not significantly above chance for devoiced stops (51.4%, log odds: 0.057, SE=0.07), but they were for voiceless stops ($p<0.0001$), with 58.66% (log odds: 0.29, SE=0.05), indicating that they were 1.3 times more likely to respond correctly when listening to a voiceless stop.

6.3. *Discussion*

The accuracy average of just 55% is barely above chance performance, and contrasts starkly with the near 100% accuracy averages obtained in the norming studies of Experiments 1–3. In addition, participants performed worse when responding to devoiced stops (average accuracy of just 51%), which was not significantly different from chance performance. In turn, the overall significant accuracy scores might be due to a ceiling effect, i.e., participants correctly identified voiceless stops more often than devoiced stops. Even though similar results were obtained in previous perception studies on incomplete neutralization (e.g., Port and O'Dell (1985) report mean accuracy values of 59%), the present results are the lowest accuracy scores reported in the literature.[9]

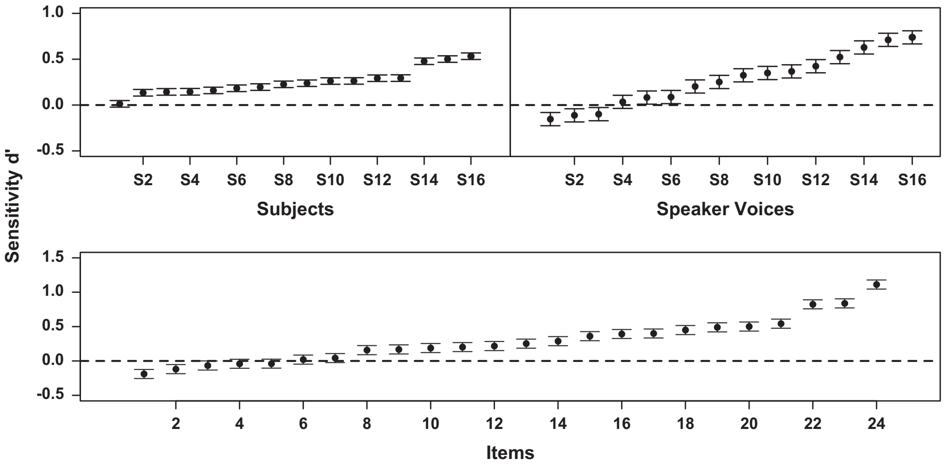

Fig. 4. Results of Experiment 4. *d'* sensitivity values arranged according to size for subjects (upper left plot), speaker-voices (upper right plot) and items (lower plot) with standard errors. Dashed lines indicate chance performance. There was no subject that scored below chance ($n=16$). There were only 5 items that scored below chance ($n=24$). And there were only 3 voices that scored below chance ($n=16$).

These low accuracy values naturally lead one to the question of whether incomplete neutralization plays any role in speech perception outside of the laboratory whatsoever. As there are only a handful of minimal pairs which are distinguished by the voicing specification of the final stop, and as these minimal pairs most often have different syntactic contexts which help to disambiguate them (e.g., the adjective *tot* 'dead' and the noun *Tod* 'death' would never appear in the same syntactic position), one could argue that the role of incomplete neutralization as a perceptual cue outside of a controlled laboratory context is negligible and unlikely to have a great deal of functional relevance in everyday speech communication. However, there is a great body of evidence demonstrating that fine phonetic detail (which may not be immediately perceptually detectable) is used in lexical access and spoken word recognition (e.g., Davis, Marslen-Wilson, & Gaskell, 2002; Hawkins & Nguyen, 2003). Thus, the present results should not necessarily be interpreted as evidence against the functional relevance of incomplete neutralization.

7. General discussion

A substantial number of experiments over the last three decades have reported minor acoustic differences between obstruents in a phonologically neutralizing context corresponding to voiced and voiceless counterparts intervocalically (which we have referred to as 'devoiced' and 'voiceless', respectively). The bulk of these studies were focused on German, although the finding has been advanced for other languages as well. As noted earlier, the findings of many of these studies have been called into question on methodological grounds, but studies purporting to provide counter-evidence such as Fourakis and Iverson (1984) or Jassem and Richter (1989) were at least as problematic (see Section 2). The aim of the present study was to put the debate surrounding incomplete neutralization on a firmer empirical footing by using an auditory design similar to that of Fourakis and Iverson, but with a larger sample of subjects and items.

We found vowel duration to be a robust acoustic correlate of devoiced and voiceless stops in syllable-final position: vowels were longer before devoiced stops than before voiceless stops. By using a different methodology from previous studies, this study contributes to the converging evidence that neutralization of German final stops is incomplete. Our finding that incomplete neutralization emerges even in a completely auditory task, and when using pseudowords instead of real words, was replicated in two additional experiments that mitigated the potential for accommodation to an intervocalic input stimulus.

We also conducted a perception experiment, in which we found that although participants were able to distinguish devoiced from voiceless final stops, their accuracy was barely above chance performance (55%, as opposed to 98–100% intervocalically in the norming studies). Moreover, this overall above-chance accuracy was largely driven by greater accuracy in correctly identifying voiceless stops; participants were at chance when identifying devoiced stops. This speaks to limitations with respect to the perception of incomplete neutralization, indicating that, at least within a forced choice paradigm, pre-stop vowel length is not a robust cue to voicing category in final position. Thus, while the present experiments provide robust evidence for incomplete neutralization in production, it remains unclear whether listeners actually use these small differences in perception.

The acceptance of any phenomenon should never be based on a single study, and several studies, such as Fourakis and Iverson (1984), have been overemphasized relative to the totality of incomplete neutralization studies (see Winter and Roettger, 2011, for discussion). Only by accumulating converging evidence from different methodologies can we be more certain about whether neutralization is complete or not. To date, studies finding evidence of incomplete neutralization (both for German and for other languages) outnumber those finding counter-evidence, suggesting that statements like those of Wiese (1996: 205) that "[t]hese results are rather tentative […] given that the recognition of non-neutralized devoicing was found in a minority of cases only" can safely be said to have been superseded. Positive results for incomplete neutralization characterize the majority of studies on this topic and several of the methodological concerns raised in earlier work have now been successfully addressed.

Since the body of evidence is in favor of incomplete neutralization, we now turn to how it can be accounted for. Accepting that neutralization was incomplete was previously thought to entail changes in phonological theory; early work assumed that the obtained differences had to be explained in terms of differences in abstract representations or different ordering of implementation rules (e.g., Brockhaus, 1995; Charles-Luce, 1985; Piroth & Janker, 2004; Port & O'Dell, 1985).

[9] It might be argued that the very low accuracy scores may also be due to the dialectal background of the subjects. Even though subjects reported speaking standard German, they came from the Central Franconian and Saxon dialect area. Dialects or regional varieties are still spoken in both areas, and most subjects are likely to have been exposed to them. For example in Saxon dialects, the Central German lenition rule operates, i.e., the voiced/voiceless contrast is neutralized (through fortis stop lenition) in all positions including intervocalically. As a result, the performance of Saxon listeners in perceiving the voicing contrast in intervocalic stops is generally lower (John, 2004). This interpretation, however, stands in contrast to the very high accuracy scores in intervocalic position we found in the norming studies of Experiments 1–3.

Given the subtle and perhaps barely detectable nature of incomplete neutralization, it is reasonable to question whether it is really necessary to capture such a small effect in terms of reflexes of abstract linguistic entities. However, there are alternative approaches of incomplete neutralization, to which we now turn.

Lexical representations are now commonly assumed to contain considerably fine-grained and redundant types of information, including phonetic detail of completely inflected forms (Alegre & Gordon, 1999; Baayen et al., 1997; Brown & McNeill, 1966; Bybee, 1994, 1995; Butterworth, 1983; Goldinger, 1996, 1997; Manelis & Tharp, 1977; Palmeri et al., 1993; Pisoni, 1997; Sereno & Jongman, 1997). Ernestus and Baayen (2006) propose the possibility of incomplete neutralization effects being due to the co-activation of paradigmatically related forms, i.e., when speakers pronounce *Rad*, they also activate the non-neutralized *Räder*. This co-activation of the related voiced forms could influence the speech production mechanism in subtle ways, leading to incomplete neutralization.

This hypothesis is based on the concept of spreading activation (e.g., see references in Collins & Loftus, 1975). With respect to morphological relations, there is evidence for the automatic activation of morphological “neighbors” in perception, whereby words with more or more frequent neighbors may be recognized more quickly and/or more accurately than those with fewer or less frequent neighbors (e.g., Andrews, 1989; Sears, Hino & Lupker, 1995). One might object that most previous evidence for activation of lexical neighbors comes from perception studies. However, recent studies have also demonstrated the effects of lexical neighbors on speech production. Baese-Berk and Goldrick (2009), for instance, showed how production of VOT is modulated on-line depending on whether or not a word is presented in context with its minimal pair neighbor, while Munson (2007) found greater vowel-space expansion for words with larger numbers of neighbors (see also Wright, 2004).

The co-activation account has two advantages over traditional accounts of incomplete neutralization. First, from a functionalist perspective, an account that treats incomplete neutralization as an artifact of lexical representations is more attractive than an account based on phonetic or phonological rules and/or representations that are extracted from auditory information. Under the co-activation account, speakers would not need to extract any subtle contrast from the signal (e.g. incomplete neutralization effects in *Rad* vs. *Rat*) so long as they can perceive the contrast between the corresponding paradigmatic neighbors (e.g. *Räder* vs. *Räte*). The presence of the contrast somewhere in the paradigm leads automatically to the prediction of incomplete neutralization. Thus the acoustic cues found in the neutralized position have no functional utility and are not reliably used in regular communication to differentiate between minimal pairs. This interpretation is in line with the low accuracy scores in perception experiments.

Second, the co-activation account makes testable predictions for future experiments. For one thing, it predicts recency effects: if the response is delivered immediately following the stimulus, the effect should be stronger than after a longer time interval. This is because spreading activation generally recedes over a relatively short time span. Furthermore, there should be frequency effects: words that have very frequent neighbors with voiced stops in intervocalic position should exhibit stronger incomplete neutralization effects than words with very infrequent neighbors (see e.g., Bybee, 2001, for the role of frequency in analogy). Moreover, it predicts incomplete neutralization effects to be dependent on lexical density. This means that a word with many voiceless lexical neighbors should surface with no (or at least weaker) incomplete neutralization effects compared to a word with many voiced lexical neighbors. Recall that in German, voiced and voiceless stops following tense/long vowels are unequally distributed depending on the place of articulation: tense/long vowels tend to precede voiced bilabial and velar stops (e.g., /liːbə/ ‘love’, /fliːgə/ ‘fly’), while lax/short vowels are very rare in this environment (e.g., /ɛbə/ ‘tide’). Given the co-activation hypothesis, forms with many lexical neighbors containing a voiced stop should show stronger incomplete neutralization effects. In line with that, we would expect that pseudoword pairs ending in alveolar stops (e.g., *Frade/Frad*) co-activate more lexical neighbors with voiced stops than pseudoword pairs ending in bilabial or velar stops (e.g., *Gobe/Gob, Schuge/Schug*). This predicts that in German the degree of incomplete neutralization may be modulated by place of articulation. This hypothesis, however, could not be confirmed; in none of our models did we find a statistically significant effect of place of articulation or even a numerical trend in the predicted direction.

However, we would like to point out once again that the experimental task employed in this and similar studies is subject to pragmatic limitations due to small effect sizes and considerable degrees of variation. Showing a statistically significant effect of incomplete neutralization is already difficult, and finding significant differences in effect sizes due to factors such as frequency asymmetries in the mental lexicon is more challenging still. This points to the practical limits of investigating incomplete neutralization, and of using incomplete neutralization as a test bed for investigating the cognitive architecture of the lexicon: in the absence of viable strategies of strengthening the effect, research on incomplete neutralization will always have to cope with high Type II error rates.

Table A1
Critical stimuli of E1.

[+voice]		[−voice]		Place	Δ devoiced – voiceless in ms
Blode	[bloːdə]	Blote	[bloːtʰə]	Alveolar	7.1
Drude	[dʁuːbə]	Drute	[dʁuːtʰə]	Alveolar	15.2
Flabe	[flaːbə]	Flape	[flaːpʰə]	Bilabial	−7.2
Frade	[fʁaːdə]	Frate	[fʁaːtʰə]	Alveolar	7.8
Froge	[fʁoːgə]	Froke	[fʁoːkʰə]	Velar	20.8
Frube	[fʁuːbə]	Frupe	[fʁuːpʰə]	Bilabial	4.3
Gage	[gaːgə]	Gake	[gaːkʰə]	Velar	3.7
Gaude	[gaʊ̯də]	Gaute	[gaʊ̯tʰə]	Alveolar	9.2
Gobe	[goːbə]	Gope	[goːpʰə]	Bilabial	4.6
Griede	[gʁiːdə]	Griete	[gʁiːtʰə]	Alveolar	7.3
Jiede	[jiːdə]	Jiete	[jiːtʰə]	Alveolar	29.9
Klabe	[klaːbə]	Klape	[klaːpʰə]	Bilabial	6.4
Mube	[muːbə]	Mupe	[muːpʰə]	Bilabial	13.1
Nauge	[naʊ̯gə]	Nauke	[naʊ̯kʰə]	Velar	8.1
Priege	[pʁiːgə]	Prieke	[pʁiːkʰə]	Velar	8.4
Pruge	[pʁuːgə]	Pruke	[pʁuːkʰə]	Velar	21.0
Quade	[kvaːdə]	Quate	[kvaːtʰə]	Alveolar	−0.7
Quobe	[kwoːbə]	Quope	[kwoːpʰə]	Bilabial	8.4
Roge	[ʁoːgə]	Roke	[ʁoːkʰə]	Velar	18.5
Schmaube	[ʃmaʊ̯bə]	Schmaupe	[ʃmaʊ̯pʰə]	Bilabial	2.4
Schriege	[ʃʁiːgə]	Schrieke	[ʃʁiːkʰə]	Velar	7.3
Schuge	[ʃuːgə]	Schuke	[ʃuːkʰə]	Velar	16.3
Stauge	[ʃtaʊ̯gə]	Stauke	[ʃtaʊ̯kʰə]	Velar	5.5
Wiebe	[viːbə]	Wiepe	[viːpʰə]	Bilabial	1.2

8. Conclusion

The primary goal of this paper was to assess whether or not neutralization in German final stop voicing is indeed incomplete. We demonstrated the robustness of an effect on production in three production experiments, ruling out a number of claims that the incompleteness is a purely methodological artifact, and arguing that even if non-functional, the robustness of incomplete neutralization warrants explanation. We would like to emphasize that our results are crucially independent of whatever mechanism actually explains incomplete neutralization. Phonologists have been justifiably skeptical of the previous evidence arguing for incomplete neutralization, but as we have reviewed above, incomplete neutralization does not necessarily have to be explained in terms of representational differences; more parsimonious accounts are suggested by existing experimental work on lexical co-activation. Such accounts seem to us to be fruitful avenues for further investigations (cf., discussion in Winter & Roettger, 2011). Manaster-Ramer (1996: 487) used the incomplete neutralization debate as a call for an increased collaboration between phonologists and phoneticians. In Manaster-Ramer's words (Manaster-Ramer, 1996: 487), “Phonologists cannot afford to be neutral” with respect to incomplete neutralization. We have shown that the phenomenon can be seen in a different light if psycholinguistic and cognitive evidence is taken into account. We would like to extend Manaster-Ramer's call in the hopes that we may gain new perspectives on old problems by engaging with work from related disciplines.

Acknowledgments

This work has benefited from feedback we received for presentation at the ICPhS 2011 and LabPhon 2012. We would like to thank all anonymous reviewers and the editors of Journal of Phonetics for useful comments and suggestions. Furthermore, we would like to thank Klaus Kohler for his remarks on our work. Finally, we thank all of our participants for their participation in our tiresome experiments. Remaining errors are our own.

Table B1
Critical stimuli of E2 and E3 (in bold).

[+voice]		[-voice]		Place	E2 Δ devoiced−voiceless in ms	E3 Δ devoiced−voiceless in ms
Bauge	**[baʊ̯gə]**	**Bauke**	**[baʊ̯kʰə]**	**Velar**	4.0	NA
Bege	[beːgə]	Beke	[beːkʰə]	Velar	11.2	
Blebe	**[bleːbə]**	**Blepe**	**[bleːpʰə]**	**Bilabial**	−0.4	4.5
Bloge	**[bloːdə]**	**Bloke**	**[bloːtʰə]**	**Velar**	4.8	5.6
Dage	[daːgə]	Dake	[daːkʰə]	Velar	−0.5	
Dabe	[daːbə]	Dape	[daːpʰə]	Bilabial	17.3	
Diege	[diːgə]	Dieke	[diːkʰə]	Velar	4.2	
Dobe	**[doːbə]**	**Dope**	**[doːpʰə]**	**Bilabial**	4.2	−0.7
Drube	**[dʁuːbə]**	**Drupe**	**[dʁuːpʰə]**	**Bilabial**	8.9	5.0
Duge	[duːgə]	Duke	[duːkʰə]	Velar	5.5	
Fage	**[faːgə]**	**Fake**	**[faːkʰə]**	**Velar**	9.8	4.0
Faube	[faʊ̯bə]	Faupe	[faʊ̯pʰə]	Bilabial	0.5	
Flabe	[flaːbə]	Flape	[flaːpʰə]	Bilabial	−0.2	
Flebe	**[fleːbə]**	**Flepe**	**[fleːpʰə]**	**Bilabial**	−0.2	2.0
Frebe	[fʁeːbə]	Frepe	[fʁeːpʰə]	Bilabial	4.7	
Froge	**[fʁoːgə]**	**Froke**	**[fʁoːkʰə]**	**Velar**	8.9	3.4
Frobe	[fʁoːbə]	Frobe	[fʁoːpʰə]	Bilabial	−0.5	
Frube	[fʁuːbə]	Frupe	[fʁuːpʰə]	Bilabial	6.0	
Gage	[gaːgə]	Gake	[gaːkʰə]	Velar	0.5	
Gaube	**[gaʊ̯bə]**	**Gaupe**	**[gaʊ̯pʰə]**	**Bilabial**	9.4	6.4
Gauge	[gaʊ̯gə]	Gauke	[gaʊ̯kʰə]	Velar	1.8	
Glege	**[gleːgə]**	**Gleke**	**[gleːkʰə]**	**Velar**	7.2	2.2
Gliebe	**[gliːbə]**	**Gliepe**	**[gliːpʰə]**	**Bilabial**	4.8	−0.3
Gobe	[goːbə]	Gope	[goːpʰə]	Bilabial	8.3	
Griebe	[gʁiːbə]	Griepe	[gʁiːpʰə]	Bilabial	8.4	
Hege	[heːgə]	Heke	[heːkʰə]	Velar	4.6	
Klabe	**[klaːbə]**	**Klape**	**[klaːpʰə]**	**Bilabial**	4.8	−0.4
Krobe	**[kʁoːbə]**	**Krope**	**[kʁoːpʰə]**	**Bilabial**	4.7	3.8
Miebe	**[miːbə]**	**Miepe**	**[miːpʰə]**	**Bilabial**	4.6	5.6
Naube	[naʊ̯bə]	Naupe	[naʊ̯pʰə]	Bilabial	5.8	
Nauge	**[naʊ̯gə]**	**Nauke**	**[naʊ̯kʰə]**	**Velar**	2.8	−0.1
Nuge	[nʊːgə]	Nuke	[nʊːkʰə]	Velar	1.8	
Priege	[pʁiːgə]	Prieke	[pʁiːkʰə]	Velar	1.2	
Pruge	**[pʁuːgə]**	**Pruke**	**[pʁuːkʰə]**	**Velar**	2.0	11.2
Roge	[ʁoːgə]	Roke	[ʁoːkʰə]	Velar	1.1	
Schlabe	**[ʃlaːbə]**	**Schlape**	**[ʃlaːpʰə]**	**Bilabial**	−0.7	2.5
Schmaube	**[ʃmaʊbə]**	**Schmaupe**	**[ʃmaʊ̯pʰə]**	**Bilabial**	8.8	−0.2
Schriege	**[ʃʁiːgə]**	**Schrieke**	**[ʃʁiːkʰə]**	**Velar**	7.6	2.8
Schuge	**[ʃuːgə]**	**Schuke**	**[ʃuːkʰə]**	**Velar**	7.9	2.0
Spage	**[ʃpaːgə]**	**Spake**	**[ʃpaːkʰə]**	**Velar**	5.0	−0.0
Stauge	[ʃtaʊ̯gə]	Stauke	[ʃtaʊ̯kʰə]	Velar	9.7	
Strege	**[ʃtʁeːgə]**	**Streke**	**[ʃtʁeːkʰə]**	**Velar**	4.7	4.4
Sube	**[zʊːbə]**	**Supe**	**[zʊːpʰə]**	**Bilabial**	0.2	2.3
Triege	**[tʁiːgə]**	**Trieke**	**[tʁiːkʰə]**	**Velar**	5.3	0.7
Wiebe	[viːbə]	Wiepe	[viːpʰə]	Bilabial	7.4	
Wube	[vuːbə]	Wupe	[vuːpʰə]	Bilabial	4.1	
Wuge	[vuːgə]	Wuke	[vuːkʰə]	Velar	0.8	
Zebe	[tseːbə]	Zepe	[tseːpʰə]	Bilabial	5.8	

Appendix A

See Table A1.

Appendix B

See Table B1.

References

Abboud, H. (1991). *SuperLab*. Wheaton, MD: Cedrus.

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, *40*, 41–61.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 802–814.

Baayen, R. H., Dijkstra, T., & Schreuder, S. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, *37*, 94–117.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*, 177–189.

Baese–Berk, M., & Goldrick, M. (2009). Mechanism of interaction in speech production. *Language and Cognitive Processes*, *24*, 527–554.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999–0.

Baumann, M. (1995). *The production of syllables in connected speech (Unpublished Ph.D. dissertation)*. University of Nijmegen.

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how?. *Journal of Clinical Epidemiology*, *54*, 343–349.

Bishop, J. B. (2007). Incomplete neutralization in Eastern Andalusian Spanish: Perceptual consequences of durational differences involved in s-aspiration. In *Proceedings of the 16th ICPhS*, Saarbrücken (pp. 1765–1768).

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer (Version 5.2) [Computer program].

Braver, A., & Kawahara, S. (2012). Complete and incomplete neutralization in Japanese monomoraic lengthening. Ms. Rutgers University.

Brockhaus, W. (1995). *Final Devoicing in the Phonology of German*. Max Niemeyer Verlag: Tübingen.

Brown, R., & McNeill, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.

Butterworth, B. (1983). Lexical representation. In: B. Butterworth (Ed.), *Language production*, vol. 2 (pp. 257–294). London: Academic Press.

Bybee, J. (1994). A view of phonology from a cognitive and functional perspective. *Cognitive Linguistics*, *5*, 285–305.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Charles-Luce, J. (1985). Word-final devoicing in German: Effects of phonetic and sentential contexts. *Journal of Phonetics*, *13*, 309–324.

Charles-Luce, J., & Dinnsen, D. (1987). A reanalysis of Catalan devoicing. *Journal of Phonetics*, *15*, 187–190.

Charles-Luce, J. (1993). The effects of semantic context on voicing neutralization. *Phonetica*, *50*, 28–43.

Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*, 466–485.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 218–244.

de Jong, K. J. (2011). Flapping in American English. In: M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *Blackwell Companion to Phonology* (pp. 2711–2729). Oxford: Wiley-Blackwell.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunez Filho, G., Jobert, A., et al. (2010). How learning to read changes the cortical networks for vision and language. *Science*, *330*, 1359–1364.

Dinnsen, D. A. (1985). A re-examination of phonological neutralization. *Journal of Linguistics*, *21*, 265–279.

Dinnsen, D. A., & Charles-Luce, J. (1984). Phonological neutralization, phonetic implementation and individual differences. *Journal of Phonetics*, *12*, 49–60.

Dinnsen, D. A., & Garcia–Zamor, M. (1971). The three degrees of vowel duration in German. *Papers in Linguistics*, *4*, 111–126.

Dmitrieva, O., Jongman, A., & Sereno, J. (2010). Phonological neutralization by native and non–native speakers: The case of Russian final devoicing. *Journal of Phonetics*, *38*, 483–492.

Ernestus, M., & Baayen, R. H. (2006). The functionality of incomplete neutralization in Dutch: The case of past-tense formation. In: L. M. Goldstein, D. H. Whalen, & C. T. Best (Eds.), *Laboratory Phonology*, 8 (pp. 27–49). Berlin: de Gruyter.

Fourakis, M., & Iverson, G. K. (1984). On the 'Incomplete Neutralization' of German final obstruents. *Phonetica*, *41*, 140–149.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132–138.

Fuchs, S. (2005). Articulatory correlates of the voicing contrast in alveolar obstruent production in German. *ZAS Papers in Linguistics*, 41.

Gerfen, C. (2002). Andalusian Codas. *Probus*, *14*, 247–277.

Gerfen, C., & Hall, K. (2001). Coda aspiration and incomplete neutralization in Eastern Andalusian Spanish. Manuscript, University of North Carolina at Chapel Hill. Retrieved from: ⟨http://www.unc.edu/∼gerfen/papers/GerfenandHall.pdf⟩.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In: K. Johnson, & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 33–65). San Diego: Academic Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Gregory, S. W., & Hoyt, B. R. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psychological Research*, *11*, 35–46.

Greisbach, R. (2001). Experimentelle Testmethodik on Phonetik und Phonologie. Untersuchungen zu segmentalen Grenzphänomenen im Deutschen. Frankfurt a. M.: Lang.

Hawkins, S., & Nguyen, N. (2003). Effects on word recognition of syllable-onset cues to syllable-coda voicing. In: J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology VI* (pp. 38–57). Cambridge: Cambridge University Press.

Jassem, W., & Richter, L. (1989). Neutralization of voicing in Polish obstruents. *Journal of Phonetics*, *17*, 317–325.

Jespersen, O. (1913). *Lehrbuch der Phonetik* (2nd ed.). Leipzig: G.B. Teubner.

John, T. (2004). *Eine akustische Analyse der Lenis/Fortis—Opposition in Varietäten des Sächsischen (Unpublished MA thesis)*. University of Kiel.

Jongman, A., Sereno, J. A., Raaijmakers, M., & Lahiri, A. (1992). The phonological representation of [voice] in speech perception. *Language and Speech*, *35*, 137–152.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, *60*, 286–319.

Kharlamov, V. (2012). *Incomplete neutralization and task effects in experimentally-elicited speech: Evidence from the production and perception of word-final devoicing in Russian (Ph.D. dissertation)*. University of Ottawa.

Kleber, F., John, T., & Harrington, J. (2010). The implications for speech perception of incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *38*, 185–196.

Kohler, K. J. (1984). Phonetic explanations in phonology: The feature fortis/lenis. *Phonetica*, *31*, 150–174.

Kohler, K. J. (2007). Beyond Laboratory Phonology. The phonetics of speech communication. In: M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 41–53). Oxford: Oxford University Press.

Kohler, K. J. (2012). Neutralization?! The phonetics–phonology issue in the analysis of word–final obstruent voicing. ⟨http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2012_3/⟩. Retrieved 14.02.13.

Manaster–Ramer, A. (1996). A letter from an incompletely neutral phonologist. *Journal of Phonetics*, *24*, 477–489.

Manelis, L., & Tharp, D. A. (1977). The processing of affixed words. *Memory and Cognition*, *5*, 690–695.

Mitleb, F. (1981). Temporal correlates of "voicing" and its neutralization in German. *Research in Phonetics*, *2*, 173–191 (Bloomington, Indiana: Indiana University).

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

Munson, B. (2007). Lexical access, lexical representation, and vowel production. In: J. S. Cole, & J. I. Hualde (Eds.), *Laboratory phonology*, 9 (pp. 201–228). Berlin: de Gruyter.

Natale, M. (1975a). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, *32*, 790–804.

Natale, M. (1975b). Social desirability as related to convergence of temporal speech patterns. *Perceptual Motor Skills*, *40*, 827–830.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142.

O'Dell, M., & Port, R. (1983). Discrimination of word-final voicing in German. *Journal of the Acoustical Society of America*, *73*(S1), S31 (A).

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309–328.

Perre, L., Midgley, K., & Ziegler, J. C. (2009). When beef primes reef more than leaf: Orthographic information affects phonological priming in spoken word recognition. *Psychophysiology*, *46*, 739–746.

Piroth, H. G., & Janker, P. M. (2004). Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics*, *32*, 81–109.

Pisoni, D. (1997). Some thoughts on 'normalization' in speech perception. In: K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego: Academic Press.

Port, R., & Crawford, P. (1989). Incomplete neutralization and pragmatics in German. *Journal of Phonetics*, *17*, 257–282.

Port, R., & Leary, A. (2005). Against formal phonology. *Language*, *81*, 927–964.

Port, R., & O'Dell, M. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*, 455–471.

Port, R., Mitleb, F. M., & O'Dell, M. (1981). Neutralization of obstruent voicing in German is incomplete. *Journal of the Acoustical Society of America*, *70*(S13), F10.

R Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna. 〈http://www.R-project.org〉.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency-effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 876–900.

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 546–554.

Sereno, J., & Jongman, A. (1997). Processing of English inflectional morphology. *Memory and Cognition*, *25*, 425–437.

Simonet, M., Rohena-Madrazo, M., & Paz, M. (2008). Preliminary evidence for incomplete neutralization of coda-liquids in Puerto Rican Spanish. In: L. Colantoni, & J. Steele (Eds.), *Selected proceedings of the 3rd conference on laboratory approached to spanish phonology* (pp. 72–86). Somerville, MA: Cascadilla Press.

Slowiaczek, L., & Dinnsen, D. (1985). On the neutralizing status of polish word final devoicing. *Journal of Phonetics*, *13*, 325–341.

Slowiaczek, L., & Szymanska, H. (1989). Perception of word-final devoicing in polish. *Journal of Phonetics*, *17*, 205–212.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck and Ruprecht.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub–phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, *32*, 251–276.

Weitzman, R. A. (1984). Seven treacherous pitfalls of statistics, illustrated. *Psychological Reports*, *54*, 355–363.

Whalen, D. H. (1991). Infrequent words are longer in duration than frequent words. *Journal of the Acoustical Society of America*, *90*(4), 2311.

Whalen, D. H. (1992). Further results on the duration of infrequent and frequent words. *Journal of the Acoustical Society of America*, *91*(4), 2339–2340.

Wiese, R. (1996). *The Phonology of German*. Oxford: Clarendon Press.

Winter, B., & Roettger, T. B. (2011). The nature of incomplete neutralization in German. *Grazer Linguistische Studien*, *76*, 55–74.

Wright, R. A. (2004). Factors of lexical competition in vowel articulation. In: J. J. Local, R. Ogden, & R. Temple (Eds.), *Laboratory phonology*, 6 (pp. 26–50). Cambridge: Cambridge University Press.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, *38*, 329–336.

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin and Review*, *5*, 683–689.

Zifonun, G., Hoffmann, L., Strecker, B., Ballweg, J., Brauße, U., Breindl, E., et al. (1997). *Grammatik der deutschen Sprache (Band 1)*. Berlin: de Gruyter.

Laboratory Phonology
Journal of the Association for Laboratory Phonology

Roettger, T. B. 2019 Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 1, pp. 1–27. DOI: https://doi.org/10.5334/labphon.147

JOURNAL ARTICLE

Researcher degrees of freedom in phonetic research

Timo B. Roettger
Department of Linguistics, Northwestern University, Evanston, IL, US
timo.b.roettger@gmail.com

The results of published research critically depend on methodological decisions that have been made during data analysis. These so-called 'researcher degrees of freedom' (Simmons, Nelson, & Simonsohn, 2011) can affect the results and the conclusions researchers draw from it. It is argued that phonetic research faces a large number of researcher degrees of freedom due to its scientific object—speech—being inherently multidimensional and exhibiting complex interactions between multiple covariates. A Type-I error simulation is presented that demonstrates the severe inflation of false positives when exploring researcher degrees of freedom. It is argued that combined with common cognitive fallacies, exploitation of researcher degrees of freedom introduces strong bias and poses a serious challenge to quantitative phonetics as an empirical science. This paper discusses potential remedies for this problem including adjusting the threshold for significance; drawing a clear line between confirmatory and exploratory analyses via preregistration; open, honest, and transparent practices in communicating data analytical decisions; and direct replications.

Keywords: false positive; methodology; preregistration; replication; reproducibility; speech production; statistical analysis

1. Introduction

Data analysis—that is, the path we chose from the raw data to the results section of a paper—is a complex process. We can look at data from different angles and each way to look at them may lead to different methodological and analytical choices. These potential choices are collectively referred to as researcher degrees of freedom (Simmons et al., 2011). Said choices, however, are often not specified in advance but are made in an ad hoc fashion, after having explored several aspects of the data and analytical choices. In other words, they are data-contingent rather than motivated on independent, subject-matter grounds. As is argued below, exploiting researcher degrees of freedom is often not an intentional process. However, it needs to be addressed because exploiting researcher degrees of freedom, intentionally or not, increases the chances of finding a false positive, i.e., finding a pattern that is incorrectly interpreted as rejecting the null hypothesis. This problem is shared by all quantitative scientific fields (Gelman & Loken, 2014; Simmons et al., 2011; Wicherts et al., 2016), but has not been extensively discussed for the specific characteristics of phonetic data analyses.

In this paper, I will argue that analyses in quantitative phonetics face a high number of researcher degrees of freedom due to the inherent multidimensionality of speech behavior, which is the outcome of a complex interaction between different functional layers. This article will discuss relevant researcher degrees of freedom in quantitative phonetic research, reasons as to why exploiting researcher degrees of freedom is potentially harmful for phonetics as a cumulative empirical endeavor, and possible remedies to these issues.

The remainder of this paper is organized as follows: In Section 2, I will review the concept of researcher degrees of freedom and how they can lead to misinterpretations in combination with certain cognitive biases. In Section 3, I will argue that researcher degrees of freedom are particularly prevalent in quantitative phonetics, focusing on the analysis of speech production data. In Section 4, I will present a simple simulation, demonstrating that chance processes lead to a large inflation of false positives when we exploit common researcher degrees of freedom. In Section 5, I will discuss possible ways to reduce the probability of false positives due to researcher degrees of freedom, discussing adjustments of the alpha level (Section 5.1), stressing the importance of a more rigorous distinction between confirmatory and exploratory analyses (Section 5.2), preregistrations and registered reports (Section 5.3), transparent reporting (Section 5.4), and direct replications (Section 5.5).

2. Researcher degrees of freedom

Every data analysis is characterized by a multitude of decisions that can affect its outcome and, in turn, the conclusions we draw from it (Gelman & Loken, 2014; Simmons et al., 2011; Wicherts et al., 2016). Among the decisions that need to be made during the process of learning from data are the following: What do we measure? What predictors and what mediators do we include? What type of statistical models do we use?

There are many choices to make and most of them can have an influence on the results that are obtained. Often these decisions are not made prior to data collection: Instead, we often explore the data and possible analytical choices to eventually settle on one ‘reasonable’ analysis plan which, ideally, yields a statistically convincing result. This paper argues that our statistical results are strongly affected by the number of hidden analyses performed.

In most scientific papers, statistical inference is drawn by means of null hypothesis-significance-testing (NHST, Gigerenzer, Krauss, & Vitouch, 2004; Lindquist, 1940). Because NHST is by a large margin the most common inferential framework used in quantitative phonetics, the present discussion is conceived with NHST in mind. Traditional NHST performs inference by assuming that the null hypothesis is true in the population of interest. For concreteness, assume we are conducting research on an isolated undocumented language and investigating the phonetic instantiation of phonological contrasts. We suspect the language has word stress, i.e., certain syllables are phonologically ‘stronger’ than other syllables within a word. We are interested in how word stress is phonetically manifested and pose the following hypothesis:

(1) There is a phonetic difference between stressed and unstressed syllables.

In NHST, we compute the probability of observing a result at least as extreme as a test statistic (e.g., *t*-value), assuming that the null hypothesis is true (the *p*-value). In our concrete example, the *p*-value tells us the probability of observing our data or more extreme data, if there was no difference between stressed and unstressed syllables (null hypothesis). Receiving a *p*-value below a certain threshold (commonly 0.05) is then interpreted as evidence to claim that the probability of the data, if the null hypothesis was in fact true (no difference between stressed and unstressed syllables), is sufficiently low. This is henceforth considered a positive result.

One common error within this framework that can occur is a false positive, i.e., incorrectly rejecting a null hypothesis (Type I error).[1] When undetected, false positives

[1] There are other errors that can happen and that are important to discuss. Most closely related to the present discussion are Type II errors (i.e., false negatives, Thomas et al., 1985), Type M(agnitude), and Type S(ign) errors (Gelman & Carlin, 2014). Within quantitative linguistics, these errors have recently been discussed by, for example, Kirby and Sonderegger (2018), Nicenboim, Roettger, and Vasishth (2018a), Nicenboim and Vasishth (2016), and Vasishth and Nicenboim (2016).

can have far reaching consequences, often leading to theoretical claims that may misguide future research (Smaldino & McElreath, 2016). These errors can be persistent through time because our publication system neither incentivizes publishing null results nor direct replication attempts, biasing the scientific record toward novel positive findings. As a result, there may be a large number of null results in the 'file drawer' that will never see the light of day (e.g., Sterling, 1959).

Within the NHST framework, any difference between conditions that yields a *p*-value below 0.05 is, in practice, considered sufficient to reject the null hypothesis and to claim that there is a difference. However, these tests have a natural false positive rate, i.e., given a *p*-value of 0.05, there is a 5% probability that our data accidentally suggest that the null hypothesis can be refuted.

Coming back to our hypothetical example, if, for example, we decide to measure only a single phonetic parameter (e.g., vowel duration) to test the hypothesis in (1), 5% would be the base rate of false positives, given a *p*-value of 0.05 (and assuming that the null hypothesis is true). However, this situation changes if we measure more than one parameter. For example, we could test, say, vowel duration, average intensity, and average f_0 (all common phonetic correlates of word stress, e.g., Gordon & Roettger, 2017), amounting to three null hypothesis significance tests. One of these analyses may yield a *p*-value of 0.05 or lower. We might proceed to write a paper based on this significant finding in which we argue that stressed and unstressed syllables are phonetically different in the language under investigation.

This procedure increases the chances of finding a false positive. If n independent comparisons are performed, the false positive rate would be $1 - (1 - 0.05)^n$ instead of 0.05. Three tests, for example, will produce a false positive rate of approximately 14% (i.e., $1 - 0.95 * 0.95 * 0.95 = 1 - 0.857 = 0.143$). Why is that? Assuming we could get a significant result with a *p*-value of 0.05 by chance in 5% of cases, the more often we look at random samples, the more often we will accidentally find a significant result (e.g., Tukey, 1953).

This reasoning can be applied to all researcher degrees of freedom. With every analytical decision, with every forking path in the analytical labyrinth, with every researcher degree of freedom, we increase the likelihood of finding significant results due to chance. In other words, the more we look, explore, and dredge the data, the greater the likelihood of finding a significant result. Exploiting researcher degrees of freedom until significance is reached has been called out as harmful practices for scientific progress (John, Loewenstein, & Prelec, 2012). Two often discussed instances of such harmful practices are HARKing (Hypothesizing After Results are Known, e.g., Kerr, 1998) and *p*-hacking (e.g., Simmons et al., 2011). HARKing refers to the practice of presenting relationships that have been obtained after data collection as if they were hypothesized in advance. *P*-hacking refers to the practice of hunting for significant results in order to ultimately report these results as if confirming the planned analysis. While such exploitations of researcher degrees of freedom are certainly harmful to the scientific record, there are good reasons to believe that they are, more often than not, unintentional.

People are prone to cognitive biases. Our cognitive system craves coherency and we are prone to seeing patterns in randomness (apophenia, Brugger, 2001); we weigh evidence in favor of our preconceptions more strongly than evidence that challenges our established views (confirmation bias, Nickerson, 1998); we perceive events as being plausible and predictable after they have occurred (hindsight bias, Fischhoff, 1975).[2] Scientists are no exception. For example, Bakker and Wicherts (2011) analyzed statistical errors in over

[2] See Greenland (2017) for a discussion of cognitive biases that are more specific to statistical analyses.

250 psychology papers. They found that more than 90% of the mistakes were in favor of the researchers' expectations, making a non-significant finding significant. Fugelsang, Stein, Green, and Dunbar (2004) investigated how scientists evaluate data that are either consistent or inconsistent with prior expectations. They showed that when researchers are faced with results that disconfirm their expectations, they are likely to blame the methodology while results that confirmed their expectations were rarely critically evaluated.

We work in a system that incentivizes positive results more than negative results (John et al., 2012), so we have the natural desire to find a positive result in order to publish our findings. A large body of research suggests that when we are faced with multiple decisions, we may end up convincing ourselves that the decision with the most rewarding outcome is the most justified one (e.g., Dawson, Gilovich, & Regan 2002; Hastorf & Cantril, 1954). In light of the dynamic interplay of cognitive biases and our current incentive structure in academia, having many analytical choices may lead us to unintentionally exploit these choices during data analysis. This can inflate the number of false positives.

3. The garden of forking paths in quantitative phonetics

Quantitative phonetics is no exception to the issues discussed above, and may in fact be particularly at risk because its very scientific object offers a considerable number of perspectives and decisions along the data analysis path. The next section will discuss researcher degrees of freedom in phonetics, and will, for exposition purposes, focus on speech production research. It turns out that the type of data we are collecting, i.e., acoustic or articulatory data, opens up many different forking paths (for a more discipline-neutral assessment of researcher degrees of freedom, see Wicherts et al., 2016). I discuss the following four sets of decisions (see **Figures 1** and **2**): choosing phonetic parameters (Section 3.1), operationalizing chosen parameters (Section 3.2), discarding data (Section 3.3), and choosing additional independent variables (Section 3.4). These distinctions are made for convenience and I acknowledge that there are no clear boundaries between these four sets. They highly overlap and inform each other to different degrees.

3.1. Choosing phonetic parameters

When conducting a study on speech production, the first important analytical decision to test a hypothesis is the question of operationalization, i.e., how to measure the phenomenon of interest. For example, how do we measure whether two sounds are phonetically identical, whether one syllable in the word is more prominent than others, or whether two discourse functions are produced with different prosodic patterns? In other words, how do we quantitatively capture relevant features of speech?

Speech categories are inherently multidimensional and vary through time. The acoustic parameters for one category are usually asynchronous, i.e., appear at different points of time in the unfolding signal and overlap with parameters for other categories (e.g., Jongman, Wayland, & Wong, 2000; Lisker, 1986; Summerfield, 1981; Winter, 2014). For example, the distinction between voiced and voiceless stops in English can be manifested by many different acoustic parameters such as voice onset time (e.g., Lisker & Abramson, 1963), formant transitions (e.g., Benkí, 2001), pitch in the following vowel (e.g., Haggard et al., 1970), the duration of the preceding vowel (e.g., Raphael, 1972), the duration of the closure (e.g., Lisker, 1957), as well as spectral differences within the stop release (e.g., Repp, 1979). Even temporally dislocated acoustic parameters correlate with voicing. For example, in the words led versus let, voicing correlates can be found in the acoustic

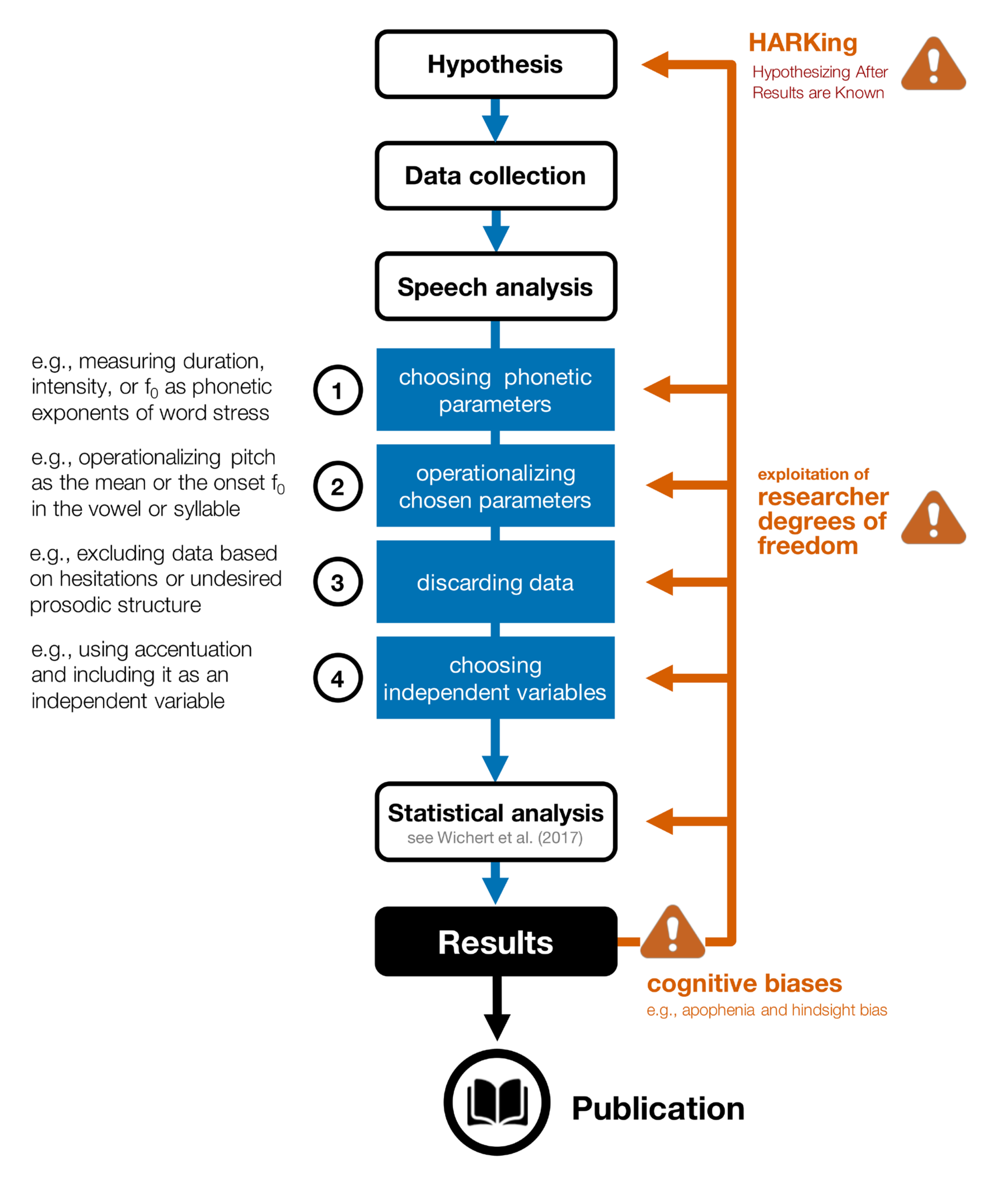
Hypothesis
Data collection
Speech analysis
choosing phonetic parameters
operationalizing chosen parameters
discarding data
choosing independent variables
Statistical analysis
see Wichert et al. (2017)
Results
Publication
1
e.g., measuring duration, intensity, or f_0 as phonetic exponents of word stress
2
e.g., operationalizing pitch as the mean or the onset f_0 in the vowel or syllable
3
e.g., excluding data based on hesitations or undesired prosodic structure
4
e.g., using accentuation and including it as an independent variable
HARKing
Hypothesizing After Results are Known
exploitation of
researcher degrees of freedom
cognitive biases
e.g., apophenia and hindsight bias

Figure 1: Schematic depiction of decision procedures during data analysis that can lead to an increased false positive rate. Along the first analysis pipeline (blue), decisions are made as to what phonetic parameters are measured (Section 3.1), how they are operationalized (Section 3.2), what data are kept and what data are discarded (Section 3.3), and what additional independent variables are measured (Section 3.4). The results are statistically analyzed (which comes with its own set of researcher degrees of freedom, see Wicherts et al., 2016) and interpreted. If the results are as expected and/or desired, the study will be published. If not, cognitive biases facilitate reassessments of earlier analytical choices (red arrow) (or a reformulation of hypotheses, i.e., HARKing), increasing the false positive rate.

manifestation of the initial /l/ of the word (Hawkins & Nguyen, 2004). These acoustic aspects correlate with their own set of articulatory configurations, determined by a complex interplay of different supralaryngeal and laryngeal gestures, coordinated with each other in intricate ways.

The multiplicity of phonetic cues grows exponentially if we look at larger temporal windows as is the case for suprasegmental aspects of speech. Studies investigating acoustic correlates of word stress, for example, have been using many different measurements including temporal characteristics (duration of certain segments or subphonemic intervals), spectral characteristics (intensity measures, formants, and spectral tilt), and measurements related to fundamental frequency (f_0) (Gordon & Roettger, 2017).

Looking at even larger domains, the prosodic expression of pragmatic functions can be expressed by a variety of structurally different acoustic cues which can be distributed throughout the whole utterance. Discourse functions are systematically expressed by multiple tonal events differing in their position, shape, and alignment (e.g., Niebuhr, D'Imperio, Gili Fivela, & Cangemi, 2011). They can also be expressed by global or local pitch scaling, as well as acoustic information within the temporal or spectral domain (e.g., Cangemi, 2015; Ritter & Roettger, 2014; van Heuven & van Zanten, 2005).[3] All of these phonetic parameters are potential manifestations of a communicative function of interest and therefore researcher degrees of freedom.

If we ask questions such as "is a stressed syllable phonetically different from an unstressed syllable?", any single measure has the potential to reject the corresponding null hypothesis (there is no difference). But which measurement should we pick? There are often many potential phonetic correlates for the relevant phonetic difference under scrutiny. Looking at more than one measurement seems to be reasonable. However, looking at many phonetic measurements to test a single global hypothesis increases the probability of finding a false positive. If we were to test 20 measurements of the speech signal repeatedly, on average one of these tests will, by mere chance, result in a spurious significant result (at a 0.05 alpha level). We obviously have justified preconceptions about which phonetic parameters may be good candidates for certain functional categories, informed by a long list of references. However, while these preconceptions can certainly help us make theoretically-informed decisions, they may also bear the risk for *ad hoc* justifications of analytical choices that happen *after* having explored researcher degrees of freedom.

One may further object that phonetic parameters are not independent of each other, i.e., many measurements covary systematically. Such covariation between multiple acoustic parameters can, for instance, result from the same underlying articulatory configurations. For example, VOT and onset f_0, the fundamental frequency at the onset of the vowel following the stop, are systematically covarying across languages, which has been argued to be a biomechanical consequence of articulatory and/or aerodynamic configurations (Hombert, Ohala, & Ewan, 1979). Löfqvist, Baer, McGarr, and Story (1989) showed that when producing voiceless consonants, speakers exhibit higher levels of activity in the cricothyroid muscle and in turn greater vocal fold tension. Greater vocal fold tension is associated with higher rates of vocal fold vibration leading to increased f_0. Since VOT and onset f_0 are presumably originating from the same articulatory configuration, one could argue that we do not have to correct for multiple testing when measuring these two acoustic parameters. However, as will be shown in Section 4, correlated measures lead to

[3] Speech is not an isolated channel of communication; it cooccurs in rich interactional contexts. Beyond acoustic and articulatory parameters, spoken communication is accompanied by non-verbal modalities such as body posture, eye gaze direction, head movement, and facial expressions, all of which have been shown to contribute to comprehension and may thus be considered relevant parameters to measure (Cummins, 2012; Latif, Barbosa, Vatiokiotis-Bateson, Castelhano, & Munhall, 2014; Prieto, Puglesi, Borràs-Comes, Arroyo, & Blat, 2015; Rochet-Capellan, Laboissière, Galván, & Schwartz, 2008; Yehia, Rubin, & Vatiokiotis-Bateson, 1998).

false positive inflation rates that are nearly as high as in independent multiple tests (see also von der Malsburg & Angele, 2017).

3.2. Operationalizing chosen parameters

The garden of forking paths is not restricted to choosing phonetic parameters. There are many different ways to operationalize the dimensions of speech that we have chosen. For example, when we want to extract specific acoustic parameters of a particular interval of the speech signal, we need to operationalize how to decide on a given interval. We usually have objective annotation procedures and clear guidelines that we agree on prior to the annotation process, but these decisions have to be considered researcher degrees of freedom and can potentially be revised after having seen the results of the statistical analysis.

Irrespective of the actual annotation, we can look at different acoustic domains. For example, a particular acoustic parameter such as duration or pitch can be operationalized differently with respect to its domain and the way it is assessed: In their survey of over a hundred acoustic studies on word stress correlates, Gordon and Roettger (2017) encountered dozens of different approaches how to quantitatively operationalize f_0, intensity and spectral tilt as correlates of word stress. Some studies took the whole syllable as a domain, others targeted the mora, the rhyme, the coda, or individual segments. Specific measurements for f_0 and intensity included the mean, the minimum, the maximum, or even the standard deviation over a certain domain. Alternative measures included the value at the midpoint of the domain, at the onset and offset of the domain, or the slope between onset and offset. The measurement of spectral tilt was also variegated. Some studies measured relative intensity of different frequency bands where the choice of frequency bands varied considerably across studies. Yet other studies measured relative intensity of the first two harmonics.

Another example of variable operationalizations can be found in time-series data such as pitch curves, formant trajectories, or articulatory movements. These time-series data have been analyzed as sequences of static landmarks (‘magic moments,’ Vatikiotis-Bateson, Barbosa, & Best, 2014), with large differences across studies regarding the identity and the operationalization of these landmarks. For example, articulatory studies looking at intragestural coordination commonly measure gestural onsets, displacement metrics, peak velocity, the onset and offset of gestural plateaus, or stiffness (i.e., relating peak velocity to displacement). Alternatively, time-series data can be analyzed holistically as continuous trajectories, differing with regard to the degrees of smoothing applied (e.g., Wieling, 2018).

In multidimensional data sets such as acoustic or articulatory data, there may be thousands of sensible analysis pathways. Beyond that, there are many different ways to process these raw measurements with regard to relevant time windows and spectral regions of interest; there are many possibilities of transforming or normalizing the raw data or smoothing and interpolating trajectories.

3.3. Discarding data

Setting aside the multidimensional nature of speech and assuming that we actually have *a priori* decided on what phonetic parameters to measure (see Section 3.1) and how to measure and process them (see Section 3.2), we are now faced with additional choices. Segmenting the acoustic signal may be difficult due to undesired speaker behavior, e.g., hesitations, disfluencies, or mispronunciations. There may be issues related to the quality of the recording such as background noise, signal interference, technical malfunctions, or interruptive external events (something that happens quite frequently during field work).

Another aspect of the data that may strike us as problematic when we extract phonetic information are other linguistic factors that could interfere with our research question. Speech consists of multiple information channels including acoustic parameters that distinguish words from each other and acoustic parameters that structure prosodic constituents into rhythmic units, structure utterances into meaningful units, signal discourse relations, or deliver indexical information about the social context. Dependent on what we are interested in, the variable use of these information channels may interfere with our research question. For example, a speaker may produce utterances that differ in their phrase-level prosodic make-up. In controlled production studies, speakers often have to produce a very restricted set of sentences. Speakers may for whatever reason (boredom, fatigue, etc.) insert prosodic boundaries or alter the information structure of an utterance, which, in turn, may drastically affect the phonetic form of other parts of the signal. For example, segments of accented words have been shown to be phonetically enhanced, making them longer, louder, and more clearly articulated (e.g., Cho & Keating, 2009; Harrington, Fletcher, & Beckman, 2000). Related to this point, speakers may use different voice qualities, some of which will make the acoustic extraction of certain parameters difficult. For example, if we were interested in extracting f_0, parts of the signal that are produced with a creaky voice may not be suitable; or if we were interested in spectral properties of the segments, parts of the signal produced in falsetto may not be suitable.

It is reasonable to 'clean' the data and remove data points that we consider as not desirable, i.e., productions that diverge from the majority of productions (e.g., unexpected phrasing, hesitations, laughter, etc.). These undesired productions may interfere with our research hypothesis and may mask the signal, i.e., make it less likely to find systematic patterns. We can exclude these data points in a list-wise fashion, i.e., subjects who exhibit a certain amount of undesired behavior (set by a certain threshold) may be entirely excluded. Alternatively, we can exclude trials on a pair-wise level across levels of a predictor. We can also simply exclude problematic tokens on a trial-by-trial basis. All of these choices change our final data set and, thus, may affect the overall result of our analysis.

3.4. Choosing independent variables

Often, subsetting the data or exclusion of whole clusters of data can have impactful consequences, as we would discard a large amount of our collected data. Instead of discarding data due to unexpected covariates, we can add these covariates as independent variables to our statistical model. In our example, we could include the factor of accentuation as an interaction term into our analysis to see whether the investigated effect may interact with accentuation. If we either find a main effect of syllable position on our measurements or an interaction with accentuation, we would probably proceed and refute the null hypothesis (there is no difference between stressed and unstressed syllables). This rationale can be applied to external covariates, too. For example, in prosodic studies, researchers commonly add the sex of their speakers to their models, either as a main effect or an interaction term, so as to control for the large sex-specific f_0 variability. Even though this reasoning is justified by independent factors, it represents a set of researcher degrees of freedom.

To summarize, the multidimensional nature of speech offers a myriad of different ways to look at our data. It allows us to choose dependent variables from a large pool of candidates; it allows us to measure the same dependent variable in alternative ways; and it allows us to preprocess our data in different ways by for example normalization or smoothing algorithms. Moreover, the complex interplay of different levels of speech phenomena introduced the possibility to correct or discard data during data annotation in a non-blinded fashion; it allows us to measure other variables that can be used as covariates, mediators, or moderators. These variables could also enable further exclusion of participants (see **Figure 2** for a schematic example).

While there are often good reasons to go down one forking path rather than another, the sheer amount of possible ways to analyze speech comes with the danger of exploring many of these paths and picking those that yield a statistically significant result. Such practices, exactly because they are often unintentional, are problematic because they drastically increase the rate of false positives.

It is important to note that these researcher degrees of freedom are not restricted to speech production studies. Within more psycholinguistically oriented studies in our field, we can for example examine online speech perception by monitoring eye movements (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) or hand movements (Spivey, Grosjean, & Knoblich, 2005). There are many different informative relationships between these measured motoric patterns and aspects of the speech signal.

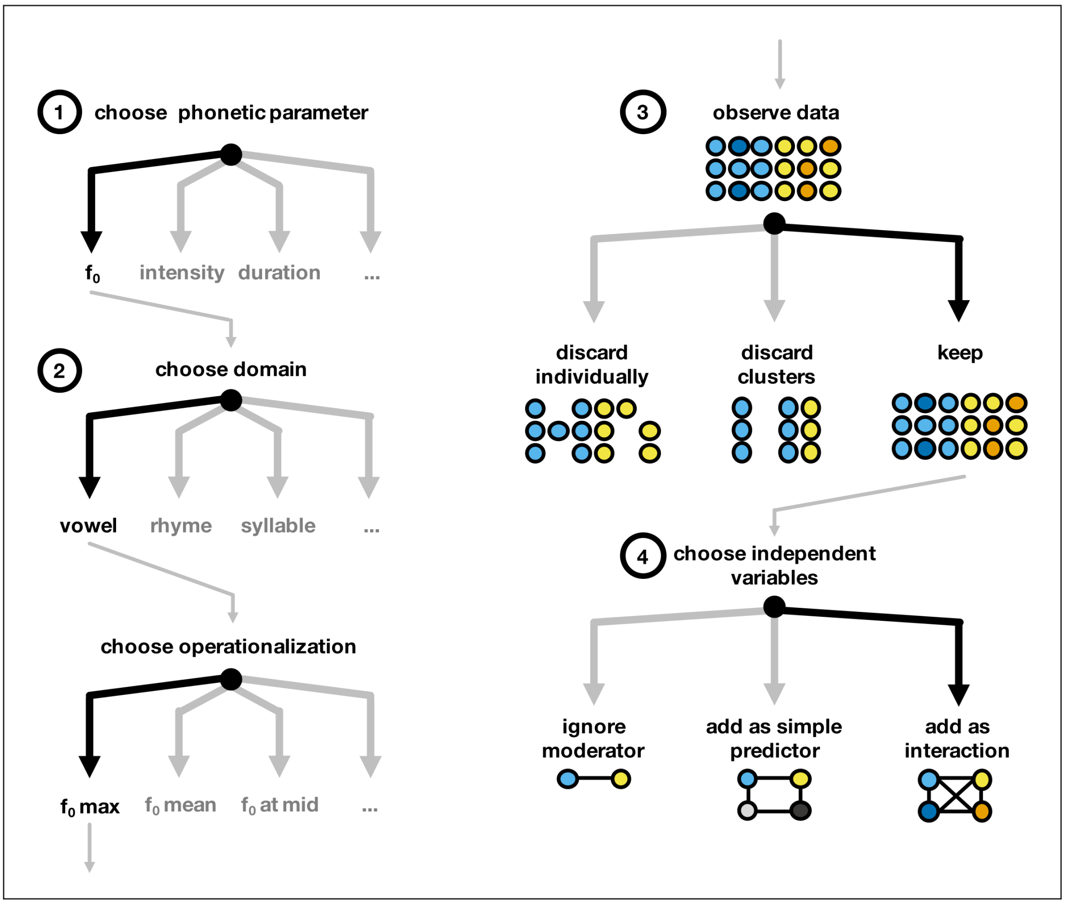

Figure 2: Schematic example of forking analytical paths as described in Section 3. First, the researchers choose from a number of possible phonetic parameters (1), then they choose in what temporal domain they measure this parameter and how they operationalize it (2). After extracting the data, the researchers must decide how to deal with aspects of speech that are orthogonal to what they are primarily interested in (say the difference between stressed and unstressed syllables: blue vs. yellow circles, [3]). For example, speakers might produce words either with or without a pitch accent (light vs. dark circles). They can discard undesired observations (e.g., discard all deaccented target words), discard clusters (e.g., discard all contrasting pairs that contain at least one deaccented target word), or keep the entire data set. If they decide to keep them, the question arises as whether to include these moderators or not and if so whether to include them as a simple predictor or add them to an interaction term (4). Note that these analytical decisions are just a subset of all possible researcher degrees of freedom at each stage.

For example, in eye tracking studies we can investigate different aspects of the visual field (foveal, parafoveal, peripheral), we can look at the number and time course of different fixations of a region (e.g., first, second, third), and the duration of a fixation or the sum of fixations before exiting/entering a particular region (see von der Malsburg & Angele, 2017, for a discussion).

More generally, the issue of researcher degrees of freedom is relevant for all quantitative scientific disciplines (see Wicherts et al., 2016). As of yet, we have not discussed much of the analytical flexibility that is specific to statistical modeling: There are choices regarding the model type, model architecture, and the model selection procedure, all of which come with their own set of researcher degrees of freedom. Having these choices is not problematic *per se*, but unintentionally exploring these choices before a final analytical commitment has been made may inflate false positive rates. Depending on the outcome and our preconception of what we expect to find, confirmation and hindsight bias may lead us to believe there are justified ways to look at the data in one particular way, until we reach a satisfying (significant) result. Again, this is a general issue in scientific practices and applies to other disciplines, too. However, given the multidimensionality of speech as well as the intricate interaction of different levels of speech behavior, the possible unintentional exploitation of researcher degrees of freedom is particularly ubiquitous in phonetic research. To demonstrate the possible severity of the issue, the following section presents a simulation that estimates false positive rates based on analytical decisions that speech production studies commonly face.

4. Simulating researcher degrees of freedom exploitation

In this section, a simulation is presented which shows that exploiting researcher degrees of freedom increases the probability of false positives, i.e., erroneously rejecting the null hypothesis. The simulation was conducted in R (R Core Team, 2016)[4] and demonstrates the effect of two different sets of researcher degrees of freedom: testing multiple dependent variables and adding a binomial covariate to the model (for similar simulations, see e.g., Barr, Levy, Scheepers, & Tily, 2013; Winter, 2011, 2015). Additionally, the correlation between dependent variables is varied with one set of simulations assuming three entirely independent measurements ($r = 0$), and one set of simulations assuming measurements that are highly correlated with each other ($r = 0.5$). The script to reproduce these simulations is publicly available here (http://osf.io/6nsfk).[5]

The simulation is based on the following hypothetical experiment: A group of researchers analyzes a speech production data set to test the hypothesis that stressed and unstressed syllables are phonetically different from each other. They collect data from 64 speakers producing words with either stress category and they measure one, two, or three acoustic parameters (e.g., vowel duration, intensity, f_0). Speakers vary regarding the intonational form of their utterances with approximately half of the speakers producing a pitch accent on the target word and the other half deaccenting the target word.

As opposed to a real-world scenario, we know the true underlying effect, since we draw values from a normal distribution around a mean value that we specify. In the present simulation, there is no difference between stressed and unstressed syllables in the 'population,' i.e., values for stressed and unstressed syllables are drawn from the same underlying distribution. However, due to random sampling (i.e., randomly picking a subset of values from the entirety of values), there are always going to be small differences

[4] The script utilizes the MASS package (Venables & Ripley, 2002) and the tidyverse library (Wickham, 2017).

[5] The simulation was inspired by Simmons et al.'s (2011) simulation which is publicly available here: https://osf.io/a67ft/.

between stressed and unstressed syllables in any given sample. Whether the word carried an accent or not is also randomly assigned to data points. In other words, there is no true effect of stress, neither is there an effect of accent on the observed productions.

We simulated 10,000 data sets and tested for the effect of stress on dependent variables under different scenarios. In our hypothetical scenarios, the following researcher degrees of freedom were explored:

(i) Instead of measuring a single dependent variable to refute the null hypothesis, the researchers measured three dependent variables. Measuring multiple aspects of the signal is common practice within the phonetic literature. In fact, measuring only three dependent variables is probably on the lower end of what researchers commonly do. However, phonetic aspects are often correlated, some of them potentially being generated by the same biomechanical mechanisms. To account for this aspect of speech, these variables were generated as either being entirely uncorrelated ($r = 0$) or being highly correlated with each other ($r = 0.5$).

(ii) The researchers explored the effect of the accent covariate and tested whether inclusion of this variable as a main effect or as an interacting term with stress yields a significant effect (or interaction effect). Accent (0,1) was randomly assigned to data rows with a probability of 0.5, that is, there is no inherent relationship with the measured parameters.

These two hypothetical scenarios correspond to researcher degrees of freedoms, discussed in Section 3.1 and Section 3.4 (see **Figures 1** and **2**). Based on these scenarios, the researchers ran simple linear models on respective dependent variables with stress as a predictor (and accent as well as their interaction in scenario ii). The simulation counts the number of times the researchers would obtain at least one significant result (p-value < 0.05) exploring the garden of forking paths described above.

The simulation (as well as the scenario it is based on) is admittedly a simplification of real-world scenarios and might therefore not be representative. As discussed above, the number of researcher degrees of freedom are manifold in phonetic investigations, therefore any simulation has to be a simplification of the real state of affairs. For example, the simulation is based on a regression model for a between-subject design, therefore not taking within-subject variation into account. It is thus important to note that the 'true' false positive rates in any given experimental scenario are not expected to be precisely as reported, or that the relative effects of the various researcher degrees of freedom are similar to those found in the simulation. Despite these necessary simplifications, the generated numbers can be considered informative and are intended to illustrate the underlying principles of how exploiting researcher degrees of freedom can impact the false positive rate.

We should expect about 5% significant results for an alpha level of 0.05 (the commonly accepted threshold in NHST). Knowing that there is no stress difference in our simulated population, any significant result is by definition a false positive. Given the proposed alpha level, we expect that—on average—1 out of 20 studies will show a false positive. **Figure 3** illustrates the results based on 10,000 simulations.

The baseline scenario (only one measurement, no covariate added) provides the expected base rate of false positives (5%, see row 1 in **Figure 3**). Departing from this baseline, the false positive rate increases substantially (rows 2–6). Looking at the results for the uncorrelated measurements first (light/red bars), results of the simulation indicate that as the researchers measure three dependent variables (rows 2–3), e.g., three possible acoustic exponents, they obtain a significant effect in 14.6% of all cases. In other words, by testing

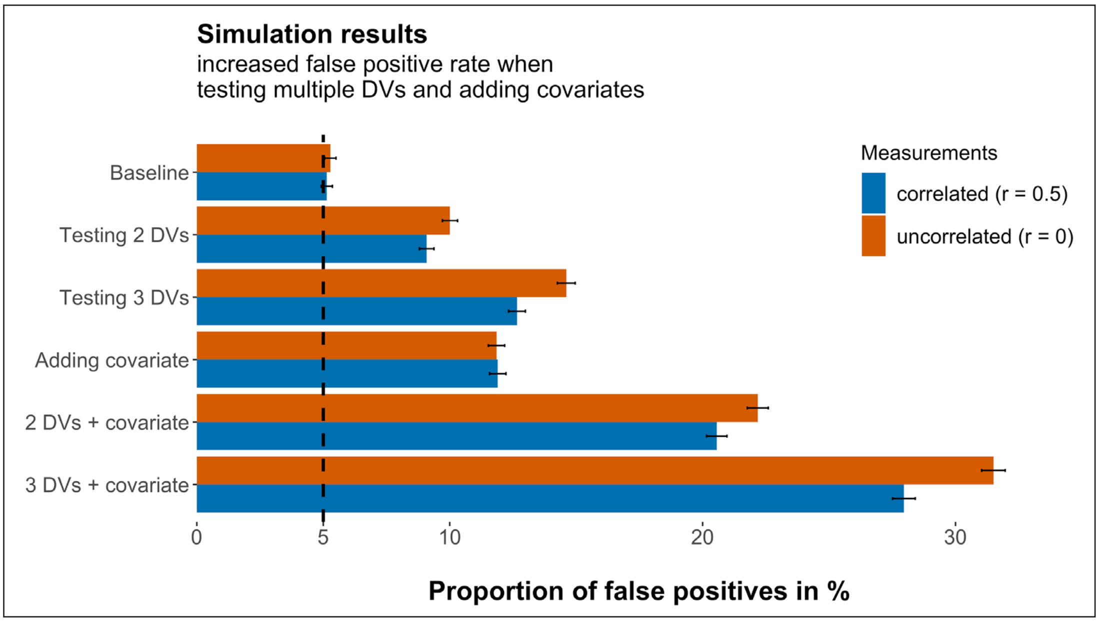

Figure 3: The x-axis depicts the proportion of simulations for which at least one attempted analysis was significant at a 0.05 level. Error bars correspond to standard deviations. Light/red bars indicate the results for uncorrelated measures, dark/blue bars indicate results for highly correlated measures (*r* = 0.5). Results were obtained for running three separate linear models for each dependent variable (row 1–3); for running one linear model with the main effect only, one model allowing for covariance with an accent main effect, and one model with an analysis of covariance with an accent interaction (row 4, a significant effect is reported if the effect of the condition or its interaction with accent was significant in any of these analyses). Row 5–6 combine multiple dependent variables with adding a covariate. The dashed line indicates the expected false positive base rate of 5%.

three different measurements, the false positive rate increases by a factor of almost 3. When researchers explore the possibility of the effect under scrutiny covarying with a random binomial covariate, they obtain a significant effect in 11.9% of cases, increasing the false positive rate by a factor of 2.3. These numbers may strike one as rather low but they are based on only a small number of researcher degrees of freedom. In real-world scenarios, some of these choices are not entirely random but potentially informed by the researcher's preconceptions and expectations, potentially clouded by cognitive biases.

Moreover, often we unintentionally exploit multiple different researcher degrees of freedom at the same time. If the researchers combine measuring three different acoustic exponents *and* run additional analyses for the added covariate, the false error rate increases to 31.5%, i.e., the false positive rate is more than six times larger than our agreed rate of 5%.

One might argue that the acoustic measurements we take are highly correlated, making the choice between them not as impactful since they 'measure the same thing.' To explore the impact of correlated measurements, the same simulations were run with highly correlated measurements (*r* = 0.5, see dark/blue bars in **Figure 3**). Although the false positive rate is slightly lower than for the uncorrelated measures, we still obtain a large number of false positives. If the researchers measure three different acoustic exponents and run additional analyses for the added covariate, the false error rate is 28%. Thus, despite the measurements being highly correlated, the false error rate is still very high (see also von der Malsburg & Angele, 2017 for a discussion of eye-tracking data). In order to put these numbers into context, imagine a journal that publishes 10 papers per issue and each paper reports only one result. In the worst-case scenario, i.e., a scenario in which the null hypothesis is true and researcher degrees of freedom are explored as described above, three to four papers of this issue would report a false positive.

Keep in mind that the presented simulation is necessarily a simplification of real-world data sets. The true false positive rates might differ from those presented here. However, even within a simplified scenario, the present simulation serves as a proof of concept that exploiting researcher degrees of freedom can impact the false positive rate quite substantially.

Another limitation of the present simulation (and of most simulations of this nature) is that it is blind to the direction of the effect. Here, we counted every *p*-value below our set alpha level as a false positive. In real-world scenarios, however, one could argue that we often have directional hypotheses. For example, we expect unstressed syllables to exhibit 'significantly' shorter durations/lower intensity/lower f_0, etc. It could be argued that finding a significant effect going in the 'wrong' direction is potentially less likely to be considered trustworthy evidence (i.e., unstressed syllables being longer). However, due to cognitive biases such as hindsight bias (Fischhoff, 1975) and overconfident beliefs in the replicability of significant results (Vasishth, Mertzen, Jäger, & Gelman, 2018) the subjective expectedness of effect directionality might be misleading.

The issues discussed above are general issues of data analysis operating in a particular inferential framework and they are not specific to phonetic sciences. However, we need to be aware of these issues and we need to have an open discourse about them. It turns out that one of the aspects that makes our scientific object so interesting—the complexity of speech—can pose a methodological burden.

5. Possible remedies

The false positive rates demonstrated above are not an inevitable fate. In the following, I will discuss possible strategies to elude the threat posed by researcher degrees of freedom, focusing on five different topics: I discuss the possibility of adjusting the alpha level as a way to reduce false positives (Section 5.1). Alternatively, I propose that drawing a clear line between exploratory and confirmatory analyses (Section 5.2) and committing to analytical decisions prior to data collection with preregistered protocols (Section 5.3) can limit the number of false positives. Additionally, a valuable complementary practice may lie in open, honest, and transparent reporting on how and what was done during the data analysis procedure (Section 5.4). While transparency cannot *per se* limit the exploitation of researcher degrees of freedom, it can facilitate their detection. Finally, it is argued that direct replications are our strongest strategy against false positives and researcher degrees of freedom (Section 5.5).

5.1. Adjusting the significance threshold

The increased false positive rates as discussed in this paper are closely linked to null hypothesis significance testing which, in its contemporary form, constitutes a dichotomous statistical decision procedure based on a preset threshold (the alpha level). At first glance, an obvious solution to reduce the likelihood of obtaining false positives is to adjust the alpha level, either by correcting it for relevant researcher degrees of freedom or by lowering the decision threshold for significance *a priori*.

First, one could correct the alpha level as a function of the number of exploited researcher degrees of freedom. This solution has mainly been discussed in the context of multiple comparisons, in which the researcher corrects the alpha level threshold according to the number of tests performed (Benjamini & Hochberg, 1995; Tukey, 1953). If we were to measure three acoustic parameters to test a global null hypothesis, which can be refuted by a single statistically significant result, we would lower the alpha level to account for three tests. These corrections can be done for example via the Bonferroni or the Šidák method.[6]

One may object that corrections for multiple testing are not reasonable in the case of speech production on the grounds that acoustic measures are usually correlated. In this case, correcting for multiple tests may be too conservative. However, as demonstrated in Section 4 and discussed by von der Malsburg and Angele (2017), multiple testing of highly correlated measures leads to false positive rates that are nearly as high as for independent multiple tests. Thus, a multiple comparisons correction is necessary even with correlated measures in order to obtain the conventional false positive rate of 5%. However, given the large analytical decision space we have discussed above, it remains unclear as to how much to correct the alpha level for other individual researcher degree of freedom. Moreover, given that speech production experiments usually yield a limited amount of data, strong alpha level corrections can drastically inflate false negatives (Type II errors, e.g., Thomas et al., 1985), i.e., erroneously failing to reject the null hypothesis.

Complementarily, one could pose a more conservative alpha level *a priori*. Benjamin et al. (2018) recently made such a proposal, recommending to lower our commonly agreed alpha level from $p \leq 0.05$ to $p \leq 0.005$. All else being equal, lowering the alpha level will reduce the absolute number of false positives (e.g., in the simulation above, we would only obtain around 0.5% to 3% of false positives across scenarios). While some researchers articulated their concerns that lowering the commonly accepted threshold for significance comes with important drawbacks such as an increase in false negatives and increased resource costs (Amrhein & Greenland, 2018; Lakens et al., 2018, but see de Ruiter, 2018), it can certainly help us reduce false positives by raising the bar as to what counts as significant and what does not.

In sum, correcting the alpha level or setting a lower alpha level threshold for significance can be helpful strategies to control for false positive rates in a conservative way. As with every practice, alpha level adjustments have their own drawbacks. It sometimes remains unclear as how to exactly adjust the alpha level in a non-conservative way. Moreover, alpha level adjustment can increase false negative rates.

5.2. Flagging analyses as exploratory vs. confirmatory

One important remedy to the issue of researcher degrees of freedom is to draw a clear line between exploratory and confirmatory analyses, two conceptually separate phases in scientific discovery (Box, 1976; Tukey, 1980; de Groot, 2014; see Nicenboim, Vasishth, Engelmann, & Suckow, 2018b, and Roettger, Winter, & Baayen, accepted, for recent discussions related to linguistic research). In an exploratory analysis, we observe patterns and relationships which lead to the generation of concrete hypotheses as to how these observations can be explained. These hypotheses can then be challenged by collecting new data (e.g., in controlled experiments). Putting our predictions under targeted scrutiny helps us revise our theories based on confirmatory analyses. Our revised models can then be further informed by additional exploration of the available data. This iterative process of alternating exploration and confirmation advances our knowledge. This is hardly news to the reader. However, quantitative research in science in general, as well as in our field in particular, often blurs the line between these two types of data analysis.

Exploratory and confirmatory analyses should be considered complementary to each other. Unfortunately, when it comes to publishing our work, they are not weighted equally.

[6] One approach is to use an additive (Bonferroni) inequality: For *n* tests, the alpha level for each test is given by the overall alpha level divided by *n*. A second approach is to use a multiplicative inequality (Šidák): For *n* tests, the alpha level for each test is calculated by taking 1 minus the n^{th} root of the complement of the overall alpha level.

Confirmatory analyses have a superior status, determining the way we frame our papers and the way funding agencies demand successful proposals to look like. This asymmetry can have harmful consequences which I have discussed already: *HARKing* and *p-hacking*. It may also incentivize researchers to sidestep clear-cut distinctions between exploratory and confirmatory findings. The publication apparatus forces us into a confirmatory mindset, while we often want to explore the data and generate hypotheses. For example, we want to explore what the most important phonetic exponents of a particular functional contrast are. We may not necessarily have a concrete prediction we want to test at this stage, but we want to understand patterns in speech with respect to their function. Exploratory analyses are necessary to establish standards as to how aspects of speech relate to linguistic, cognitive, and social variables. Once we have established such standards, we can agree to only look at relevant phonetic dimensions, reducing the analytical flexibility with regard to what and how to measure (Sections 3.1–3.2).

Researcher degrees of freedom mainly affect the confirmatory part of scientific discovery; they do not restrict our attempts to explore our data. But claims based on exploration should be cautious. After having looked at 20 acoustic dimensions, any seemingly systematic pattern may be spurious. Instead, this exploratory step should generate new hypotheses which we then can confirm or disconfirm using a new data set. In many experiments, prior to data collection, it may not be clear how a functional contrast may phonetically manifest itself. Presenting such exploratory analyses as confirmatory may hinder replicability and may give a false feeling of certainty regarding the results (Vasishth et al., 2018). In a multivariate setting, which is the standard setting for phonetic research, there are multiple dimensions to the data that can inform our theories. Exploring these dimensions may often be more valuable than just a single confirmatory test of a single hypothesis (Baayen, Vasishth, Kliegl, & Bates, 2017). These cases make the distinction between confirmatory and exploratory analyses so important. We should explore our data. Yes. Yet we should not pretend we are testing concrete hypotheses when doing so.

Although our academic incentive system makes drawing this line difficult, journals have started to become aware of this issue and have started to create incentives to explicitly publish exploratory analyses (for example Cortex, see McIntosh, 2017). One way of ensuring a clear separation between exploratory and confirmatory analyses are preregistrations and registered reports (Nosek, Ebersole, DeHaven, & Mellor, 2018; Nosek & Lakens, 2014).

5.3. Preregistrations and registered reports

A preregistration is a time-stamped document in which researchers specify exactly how they plan to collect their data and how they plan to conduct their confirmatory analyses. Such reports can differ with regard to the details provided, ranging from basic descriptions of the study design to detailed procedural and statistical specifications up to the publication of scripts.

Preregistrations can be a powerful tool to reduce researcher degrees of freedom because researchers are required to commit to certain decisions prior to observing the data. Additionally, public preregistration can at least help to reduce issues related to publication bias, i.e., the tendency to publish positive results more often than null results (Franco, Malhotra, & Simonovits, 2014; Sterling, 1959), as the number of failed attempts to reject a hypothesis can be tracked transparently (if the studies were conducted).

There are several websites that offer services and/or incentives to preregister studies prior to data collection, such as AsPredicted (AsPredicted.org) and the Open Science Framework (osf.io). These platforms allow us to time-log reports and either make them publicly available or grant anonymous access only to a specific group of people (such as reviewers and editors during the peer-review process).

A particular useful type of preregistration is a peer-reviewed registered report, which an increasing number of scientific journals have adopted already (Nosek et al., 2018; Nosek & Lakens, 2014; see cos.io/rr for a list of journals that have adopted this model).[7] These protocols include the theoretical rationale of the study and a detailed methodological description. In other words, a registered report is a full-fledged manuscript minus the result and discussion section. These reports are then critically assessed by peer reviewers, allowing the authors to refine their methodological design. Upon acceptance, the publication of the study results is in-principle guaranteed, no matter whether the results turn out to provide evidence for or against the researcher's predictions.

For experimental phonetics, a preregistration or registered report would ideally include a detailed description of what is measured and how exactly it is measured/operationalized, as well as a detailed catalogue of objective inclusion criteria (in addition to other key aspects of the method including all relevant researcher degrees of freedom related to preprocessing, postprocessing, statistical modelling, etc.; see Wicherts et al., 2016). Committing to these decisions prior to data collection can reduce the danger of unintentionally exploiting researcher degrees of freedom.

At first sight, there appear to be several challenges that come with preregistrations (see Nosek et al., 2018, for an exhaustive discussion). For example, after starting to collect data, we might realize that our preset exclusion criteria do not capture an important behavioral aspect of our experiment (e.g., some speakers may produce undesired phrase-level prosodic patterns which we did not anticipate). These patterns interfere with our research question. Deviations from our data collection and analysis plan are common. In this scenario, we could change our preregistration and document these changes alongside our reasons as to why and when we have made these changes (i.e., after how many observations). This procedure still provides substantially lower risk of cognitive biases impacting our conclusions compared to a situation in which we did not preregister at all.

Researchers working with corpora may object that preregistrations cannot be applied to their investigations because their primary data have already been collected. But preregistration of analyses can still be performed. Although, ideally, we limit researcher degrees of freedom prior to having seen the data, we can (and should) preregister analyses after having seen pilot data, parts of the study, or even whole corpora. When researchers generate a hypothesis that they want to confirm with a corpus data set, they can preregister analytic plans and commit to how evidence will be interpreted before analyzing the data.

Another important challenge when preregistering our studies is predicting appropriate inferential models. Preregistering a data analysis necessitates knowledge about the nature of the data. For example, we might preregister an analysis assuming that our measurements are generated by a Gaussian process. After collecting our data, we might realize the data have heavy right tales, calling for a log-transformation; thus, our preregistered analysis might not be appropriate. One solution to this challenge is to define data analytical procedures in advance that allow us to evaluate distributional aspects of the data and potential data transformations irrespective of the research question. Alternatively, we could preregister a decision tree. This may actually be tremendously useful for people using hierarchical linear models. When using appropriate random effect structures (see Barr et al., 2013; Bates et al., 2015), these models are known to run into convergence issues (e.g., Kimball, Shantz, Eager, & Roy, 2018). To remedy such convergence issues, a common strategy is to drop complex random effect terms incrementally (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Since we do not know whether a model will converge or not in advance, a concrete plan of how we reduce model complexity can be preregistered in advance.

[7] Note that there are only a few journals on quantitative linguistics that have adopted registered reports.

Preregistrations and registered reports help us draw a line between the hypotheses we intend to test and data exploration (see **Figure 4**). Any exploration of the data beyond the preregistered analysis have to be considered as hypotheses-generating only. If we are open and transparent about this distinction and ideally publicly indicate where we draw the line between the two, we can limit false positives due to researcher degrees of freedom exploitation in our confirmatory analyses and commit more honestly to subsequent exploratory analyses.

5.4. Transparency

The credibility of scientific findings is mainly rooted in the evidence supporting it. We assess the validity of this evidence by constantly reviewing and revising our methodologies, and by extending and replicating findings. This becomes difficult if parts of the process are not transparent or cannot be evaluated. For example, it is difficult to evaluate whether exploitation of researcher degrees of freedom is an issue for any given study if the authors are not transparent about when they made which analytical decisions. We end up having to *trust* the authors. Trust is good, control is better. We should be aware and open about researcher degrees of freedom and communicate this aspect of our data analyses to our peers as honestly as we can. For example, if we have measured and analyzed ten phonetic parameters to establish whether the final syllable of a word is more prominent than the prefinal syllable, we should provide this information: If we have run analyses with and without gender as a covariate, we should say so and discuss the results of these analyses. An open, honest, and transparent research culture is desirable. As argued above, public preregistration can facilitate transparency of what analytical decisions we made and when we have made them. Being transparent does not, of course, prevent *p*-hacking, HARKing or other exploitations of researcher degrees of freedom, but it makes these harmful practices *detectable*.

For our field, transparency with regard to our analysis has many advantages (e.g., Nicenboim et al., 2018b).[8] As analysis is subjective in the sense that it incorporates the researcher's beliefs and assumptions about a study system (McElreath, 2016) the only way to make analyses objectively assessable is to be transparent about this aspect of empirical work. Transparency then allows other researchers to draw their own conclusions as to which researcher degrees of freedom were present and how they may have affected the original conclusion.

In order to facilitate such transparency, we need to agree on how to report aspects of our analyses. While preregistrations and registered reports lead to better discoverability of researcher degrees of freedom, they do not necessarily allow us to systematically evaluate them. We need institutionalized standards, as many other disciplines have already developed. There are many reporting guidelines that offer standards for reporting methodological choices (see the Equator Network for an aggregation of these guidelines: http://www.equator-network.org/). Systematic reviews and meta analyses such as Gordon and Roettger (2017) and Roettger and Gordon (2017) can be helpful departure points to create an overview of possible analytical decisions and their associated degrees of freedom (e.g., what is measured; how it is measured, operationalized, processed, and extracted; what data are excluded; when and how are they excluded, etc.). Such guidelines, however,

[8] Beyond sharing data tables and analysis scripts, it would be desirable to share raw acoustic or articulatory files. However, making these types of data available relies on getting permission from participants in advance (as acoustic data are inherently identifiable). Making raw speech production data available to the community would greatly benefit evidence accumulation in our field. We can share these data on online repositories such as, for example, OSCAAR (the Online Speech/Corpora Archive and Analysis Resource: https://oscaar.ci.northwestern.edu/).

are only effective when a community agrees on their value and applies them including journals, editors, reviewers, and authors.

5.5. Direct replications

The above discussed remedies help us to either limit the exploitation of researcher degrees of freedom or make them more detectable. However, none of these strategies is a fool-proof protection against false positives. To ultimately avoid the impact of false positives on the scientific record, we should increase our efforts to directly replicate previous research, defined here as the repetition of the experimental methods that led to a reported finding.

The call for more replication is not original. Replication has always been considered a tremendously important aspect of the scientific method (e.g., Campbell, 1969; Kuhn, 1962; Popper, 1934/1992; Rosenthal, 1991) and in recent coordinated efforts to replicate published results, the social sciences uncovered unexpectedly low replicability rates, a state of affairs that has been coined the 'replication crisis.' For example, the Open Science Collaboration (2015) tried to replicate 100 studies that were published in three high-ranking psychology journals. They assessed whether the replications and the original experiments yielded the same result and found that only about one third to one half of the original findings (depending on the definition of replication) were also observed in the replication study. This lack of replicability is not restricted to psychology. Concerns about the replicability of findings have been raised for medical sciences (e.g., Ioannidis, 2005), neuroscience (Wager, Lindquist, Nichols, Kober, & van Snellenberg, 2009), genetics (Hewitt, 2012), cancer research (Errington et al., 2014), and economics (Camerer et al., 2016).

Most importantly, it is a very real problem for quantitative linguistics, too. For example, Nieuwland et al. (2018) recently tried to replicate a seminal study by DeLong, Urbach, and Kutas (2005) which is considered a landmark study for the predictive processing literature and which has been cited over 500 times. In their preregistered multi-site replication attempt (9 laboratories, 334 subjects), Nieuwland et al. were not able to replicate some of the key findings of the original study.

Stack, James, and Watson (2018) recently failed to replicate a well-cited study on rapid syntactic adaptation by Fine, Jaeger, Farmer, and Qian (2013). After failing to find the original effect in an extension, they went back and directly replicated the original study with appropriate statistical power. They found no evidence for the original effect.

Possible reasons for the above cited failures to replicate are manifold. As has been argued here, exploitation of researcher degrees of freedom is one of the reasons why there is a large number of false positives. Combined with other statistical issues such as low power (e.g., for recent discussion see Kirby & Sonderegger, 2018; Nicenboim et al., 2018a), violation of the independence assumption (Nicenboim & Vasishth, 2016; Winter, 2011, 2015), and the 'significance' filter (i.e., treating results publishable because $p < 0.05$ leads to overoptimistic expectations of replicability; see Vasishth et al., 2018), it is to be expected that there are a large number of experimental phonetic findings that may not stand the test of time.

The above replication failures sparked a tremendously productive discourse throughout the quantitative sciences and led to quick methodological advancements and best practice recommendations. For example, there are several coordinated efforts to directly replicate important findings by multi-site projects such as the ManyBabies project (Frank et al., 2017) and Registered Replication Reports (Simons, Holcombe, &, Spellman, 2014). These coordinated efforts can help us put theoretical foundations on a firmer footing. However, the logistic and monetary resources associated with such large-scale projects are not always pragmatically feasible for everyone in the field.

Replication studies are not very popular because the necessary time and resource investment are not appropriately rewarded in contemporary academic incentive systems (Koole & Lakens, 2012; Makel, Plucker, & Hegarty, 2012; Nosek, Spies, & Motyl, 2012). Both successful replications (Madden, Easley, & Dunn, 1995) and repeated failures to replicate (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012) are rarely published, and if they are published they are usually published in less prestigious outlets than the original findings. To overcome the asymmetry between the cost of direct replication studies and the presently low academic payoff for it, we as a research community must re-evaluate the value of direct replications. Funding agencies, journals, editors, and reviewers should start valuing direct replication attempts, be it successful replications or replication failures, as much as they value novel findings. For example, we could either dedicate existing journal space to direct replications (e.g., as an article type) or create new journals that are specifically dedicated to replication studies. For example, the Royal Society Open Science has recently initiated an interesting new publication model. For their Psychology and Cognitive Neuroscience section, they guarantee to publish any close replication of any article published in their own and other journals (https://blogs.royalsociety.org/publishing/reproducibility-meets-accountability/). Thus, if the journal agrees to publish a study, it becomes responsible for publishing direct replications of that study, too.

As soon as we make publishing replications easier, more researchers will be compelled to replicate both their own work and the work of others. Only by replicating empirical results and evaluating the accumulated evidence can we substantiate previous findings and extend their external validity.

6. Summary and concluding remarks

This article has discussed researcher degrees of freedom in the context of quantitative phonetics. Researcher degrees of freedom concern all possible analytical choices that may influence the outcome of our analysis. In a null-hypothesis-significance testing framework of inference, intentional or unintentional exploitation of researcher degrees of freedom can have a dramatic impact on our results and interpretations, increasing the likelihood of obtaining false positives. Quantitative phonetics faces a large number of researcher degrees of freedom due to its scientific object being inherently multidimensional and exhibiting complex interactions between many covarying layers of speech. A Type-I error simulation demonstrated substantial false error rates when combining just two researcher degrees of freedom such as testing more than one phonetic measurement, and including a speech-relevant covariate in the analysis. It has been argued that combined with common cognitive fallacies, unintentional exploitation of researcher degrees of freedom introduces strong bias and poses a serious challenge to quantitative phonetics as an empirical science.

Several potential remedies for this problem have been discussed (see **Figure 4**). When operating in the NHST statistical framework, we can reconsider our preset threshold for significance. We should draw an explicit line between confirmatory and exploratory analyses. One way to enforce such a clear line are preregistrations or registered reports, i.e., records of the experimental design and the analysis plan that are committed prior to data collection and analysis. While preregistration offers better detectability of researcher degrees of freedom, standardized reporting guidelines and transparent reporting might facilitate a more objective assessment of these researcher degrees of freedom by other

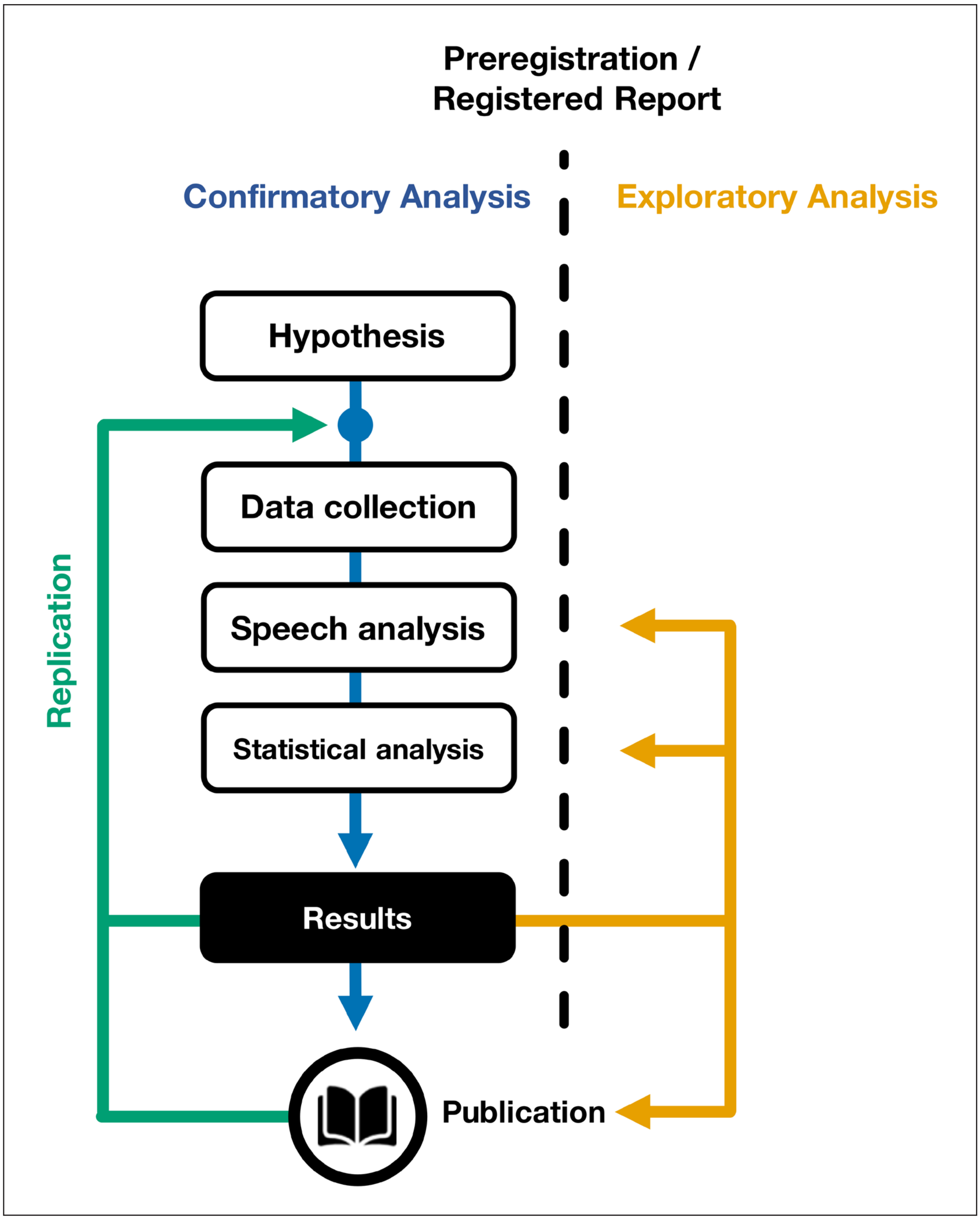

Figure 4: Schematic depiction of the decision procedure during data analysis that limits false positives: Prior to data collection, the researcher commits to an analysis pipeline via preregistration/registered reports, leading to a clear separation of confirmatory (blue arrows) and exploratory analysis (yellow arrows). The analysis is executed accordingly and the results are interpreted with regard to the confirmatory analysis. After the confirmatory analysis, the researcher can revisit the decision procedure and explore the data. The interpretation of the confirmatory analysis and potential insights gained from the exploratory analysis are published alongside an open and transparent track record of all analytical steps (preregistration, code, and data for both confirmatory and exploratory analyses). Finally, either prior to publication or afterwards, the study is directly replicated (green arrow) by either the same research group or independent researchers in order to substantiate the results.

researchers. Yet all of these proposals come with their own limitations and challenges. A complementary strategy to limit false positives lies in direct replications, a form of research that is unfortunately not well rewarded within the present academic system.

As a community, we need to openly discuss such issues and find feasible solutions to them. Possible solutions must not only be practical from a logistic perspective but should also avoid punishing rigorous methodology within our academic incentive system. Explicitly labeling our work as exploratory, being transparent about potential bias due to researcher degrees of freedom, or running direct replications may make it more difficult to be rewarded for our work (i.e., by being able to publish our studies in prestigious journals). Thus, authors, reviewers, and editors alike need to be aware of these methodological challenges. The present paper was conceived in the spirit of such an open discourse. Thus, the single most powerful solution to methodological challenges as described in this paper is engaging in a critical and open discourse about our methods and analyses.

Acknowledgements

I would like to thank Aviad Albert, Eleanor Chodroff, Emily Cibelli, Jennifer Cole, Tommy Denby, Matt Goldrick, James Kirby, Nicole Mirea, Bruno Nicenboim, Limor Raviv, Shayne Sloggett, Lukas Soenning, Mathias Stoeber, Bodo Winter, two anonymous reviewers, and the handling editor for valuable feedback on earlier drafts of this paper. All remaining errors are my own.

Competing Interests

The author has no competing interests to declare.

References

Amrhein, V., & Greenland, S. 2018. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, *2*(1), 4. DOI: https://doi.org/10.1038/s41562-017-0224-0

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. DOI: https://doi.org/10.1016/j.jml.2016.11.006

Bakker, M., & Wicherts, J. M. 2011. The (mis) reporting of statistical results in psychology journals. *Behavior research methods*, *43*(3), 666–678. DOI: https://doi.org/10.3758/s13428-011-0089-5

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. 2015. Parsimonious mixed models. *arXiv Preprint*. arXiv:1506.04967.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., & Cesarini, D. 2018. Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. DOI: https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Benkí, J. R. 2001. Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, *29*(1), 1–22. DOI: https://doi.org/10.1006/jpho.2000.0128

Box, G. E. P. 1976. Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. DOI: https://doi.org/10.1080/01621459.1976.10480949

Brugger, P. 2001. From haunted brain to haunted science. In: Houran, J., & Lange, R. (eds.), *Hauntings and poltergeists: Multidisciplinary perspectives*, 195–213. McFarland.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., & Heikensten, E. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. DOI: https://doi.org/10.1126/science.aaf0918

Campbell, D. T. 1969. Reforms as experiments. *American Psychologist*, *24*(4), 409–429. DOI: https://doi.org/10.1037/h0027982

Cangemi, F. 2015. *Prosodic detail in Neapolitan Italian*. Language Science Press: Berlin. DOI: https://doi.org/10.26530/oapen_533874

Cho, T., & Keating, P. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485. DOI: https://doi.org/10.1016/j.wocn.2009.08.001

Cummins, F. 2012. Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes*, *27*(10), 1525–1549. DOI: https://doi.org/10.1080/01690965.2011.615220

Dawson, E., Gilovich, T., & Regan, D. T. 2002. Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387. DOI: https://doi.org/10.1177/014616702236869

de Groot, A. D. 2014. The meaning of "significance" for different types of research [translated and annotated by E.-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, & H. L. J. van der Maas]. *Acta psychologica*, *148*, 188–194. DOI: https://doi.org/10.1016/j.actpsy.2014.02.001

DeLong, K. A., Urbach, T. P., & Kutas, M. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117–1121. DOI: https://doi.org/10.1038/nn1504

de Ruiter, J. 2018. Redefine or justify? Comments on the alpha debate. *Psychonomic bulletin & review*, 1–4. DOI: https://doi.org/10.3758/s13423-018-1523-9

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. 2012. Behavioral priming: It's all in the mind, but whose mind? *PloS one*, *7*(1), e29081. DOI: https://doi.org/10.1371/journal.pone.0029081

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. 2014. Science forum: An open investigation of the reproducibility of cancer biology research. *Elife*, *3*, e04333. DOI: https://doi.org/10.7554/eLife.04333

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. 2013. Rapid expectation adaptation during syntactic comprehension. *PloS one*, *8*(10), e77661. DOI: https://doi.org/10.1371/journal.pone.0077661

Fischhoff, B. 1975. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance*, *1*(3), 288–299. DOI: https://doi.org/10.1037/0096-1523.1.3.288

Franco, A., Malhotra, N., & Simonovits, G. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. DOI: https://doi.org/10.1126/science.1255484

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., & Lew-Williams, C. 2017. A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. DOI: https://doi.org/10.1111/infa.12182

Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. 2004. Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(2), 86–95. DOI: https://doi.org/10.1037/h0085799

Gelman, A., & Carlin, J. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. DOI: https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. 2014. Ethics and statistics: The AAA tranche of subprime science. *Chance*, *27*(1), 51–56. DOI: https://doi.org/10.1080/09332480.2014.890872

Gigerenzer, G., Krauss, S., & Vitouch, O. 2004. The null ritual. In: Kaplan, D. (ed.), *The Sage handbook of quantitative methodology for the social sciences*, 391–408. Thousand Oaks, CA: Sage. DOI: https://doi.org/10.4135/9781412986311.n21

Gordon, M., & Roettger, T. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1). DOI: https://doi.org/10.1515/lingvan-2017-0007

Greenland, S. 2017. Invited commentary: The need for cognitive science in methodology. *American journal of epidemiology*, *186*(6), 639–645. DOI: https://doi.org/10.1093/aje/kwx259

Haggard, M., Ambler, S., & Callow, M. 1970. Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, *47*(2B), 613–617. DOI: https://doi.org/10.1121/1.1911936

Harrington, J., Fletcher, J., & Beckman, M. 2000. Manner and place conflicts in the articulation of accent in Australian English. In: Broe, M. (ed.), *Papers in Laboratory Phonology*, *5*, 40–55. Cambridge University Press: Cambridge.

Hastorf, A. H., & Cantril, H. 1954. They saw a game: A case study. *The Journal of Abnormal and Social Psychology*, *49*(1), 129–134. DOI: https://doi.org/10.1037/h0057880

Hawkins, S., & Nguyen, N. 2004. Influence of syllable-coda voicing on the acoustic properties of syllable-onset/l/in English. *Journal of Phonetics*, *32*(2), 199–231. DOI: https://doi.org/10.1016/s0095-4470(03)00031-7

Hewitt, J. K. 2012. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior genetics*, *42*(1), 1–2. DOI: https://doi.org/10.1007/s10519-011-9504-z

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. 1979. Phonetic explanations for the development of tones. *Language*, *55*, 37–58. DOI: https://doi.org/10.2307/412518

Ioannidis, J. P. 2005. Why most published research findings are false. *PLoS medicine*, *2*(8), e124. DOI: https://doi.org/10.1093/bioinformatics/bti536

John, L. K., Loewenstein, G., & Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth tell. *Psychological Science*, *23*, 524–532. DOI: https://doi.org/10.1037/e632032012-001

Jongman, A., Wayland, R., & Wong, S. 2000. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263. DOI: https://doi.org/10.1121/1.1288413

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Kimball, A. E., Shantz, K., Eager, C., & Roy, J. 2018. Beyond maximal random effects for logistic regression: Moving past convergence errors. *Journal of Quantitative Linguistics*, 1–25. DOI: https://doi.org/10.1080/09296174.2018.1499457

Kirby, J., & Sonderegger, M. 2018. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics*, *70*, 70–85. DOI: https://doi.org/10.1016/j.wocn.2018.05.005

Koole, S. L., & Lakens, D. 2012. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608–614. DOI: https://doi.org/10.1177/1745691612462586

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., & Buchanan, E. M. 2018. Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. DOI: https://doi.org/10.1038/s41562-018-0311-x

Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M. S., & Munhall, K. G. 2014. Movement coordination during conversation. *PLoS One*, *9*(8), e105036. DOI: https://doi.org/10.1371/journal.pone.0105036

Lindquist, E. F. 1940. *Statistical Analysis in Educational Research*. Boston: Houghton Mifflin.

Lisker, L. 1957. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, *33*(1), 42–49. DOI: https://doi.org/10.2307/410949

Lisker, L. 1986. "Voicing" in English – A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*, 3–11. DOI: https://doi.org/10.1177/002383098602900102

Lisker, L., & Abramson, A. S. 1963. Crosslanguage study of voicing in initial stops. *The Journal of the Acoustical Society of America*, *35*(11), 1889–1890. DOI: https://doi.org/10.1121/1.2142685

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. 1989. The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America*, *85*(3), 1314–1321. DOI: https://doi.org/10.1121/1.397462

Madden, C. S., Easley, R. W., & Dunn, M. G. 1995. How journal editors view replication research. *Journal of Advertising*, *24*(4), 77–87. DOI: https://doi.org/10.1080/00913367.1995.10673490

Makel, M. C., Plucker, J. A., & Hegarty, B. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. DOI: https://doi.org/10.1177/1745691612460688

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. DOI: https://doi.org/10.1016/j.jml.2017.01.001

McElreath, R. 2016. *Statistical Rethinking*: *A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Press.

McIntosh, R. D. 2017. Exploratory reports: A new article type for Cortex. *Cortex*, *96*, A1–A4. DOI: https://doi.org/10.1016/j.cortex.2017.07.014

Nicenboim, B., Roettger, T. B., & Vasishth, S. 2018a. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, *70*, 39–55. DOI: https://doi.org/10.1016/j.wocn.2018.06.001

Nicenboim, B., & Vasishth, S. 2016. Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, *10*(11), 591–613. DOI: https://doi.org/10.1111/lnc3.12207

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. 2018b. Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive science*, *42*, 1075–1100. DOI: https://doi.org/10.1111/cogs.12589

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220. DOI: https://doi.org/10.1037//1089-2680.2.2.175

Niebuhr, O., D'Imperio, M., Gili Fivela, B., & Cangemi, F. 2011. Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category. *Proceedings of the 17th International Congress of Phonetic Sciences*, 120–123. Hong Kong.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu Wolfsthurn, S. V. G., Bartolozzi, F., Kogan, V., Ito, A., & Mézière, D. 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. DOI: https://doi.org/10.7554/eLife.33468

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. DOI: https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. DOI: https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. DOI: https://doi.org/10.1177/1745691612459058

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, *349*(6251). DOI: https://doi.org/10.1126/science.aac4716

Popper, K. 1992. *The logic of scientific discovery*. New York: Routledge. (Original work published 1934).

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. 2015. Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, *49*, 41–54. DOI: https://doi.org/10.1016/j.wocn.2014.10.005

Raphael, L. J. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303. DOI: https://doi.org/10.1121/1.1912974

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Repp, B. H. 1979. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, *22*(2), 173–189. DOI: https://doi.org/10.1177/002383097902200207

Ritter, S., & Roettger, T. B. 2014. Speakers modulate noise-induced pitch according to intonational context. In: *Proceedings of the 7th International Conference on Speech Prosody*, 890–893. Dublin.

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J. L. 2008. The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1507–1521. DOI: https://doi.org/10.1044/1092-4388(2008/07-0173)

Roettger, T. B., & Gordon, M. 2017. Methodological issues in the study of word stress correlates. *Linguistics Vanguard*, *3*(1). DOI: https://doi.org/10.1515/lingvan-2017-0006

Roettger, T. B., Winter, B., & Baayen, H. accepted. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Accepted for publication at Journal of Phonetics*.

Rosenthal, R. 1991. Replication in behavioral research. In: Neuliep, J. W. (ed.), *Replication research in the social sciences*, 1–39. Newbury Park, CA: Sage.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366. DOI: https://doi.org/10.1037/e519702015-014

Simons, D. J., Holcombe, A. O., & Spellman, B. A. 2014. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. DOI: https://doi.org/10.1177/1745691614543974

Smaldino, P. E., & McElreath, R. 2016. The natural selection of bad science. *Royal Society Open Science*, *3*(9). DOI: https://doi.org/10.1098/rsos.160384

Spivey, M. J., Grosjean, M., & Knoblich, G. 2005. Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398. DOI: https://doi.org/10.1073/pnas.0503903102

Stack, C. M. H., James, A. N., & Watson, D. G. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & cognition*, *46*(6), 864–877. DOI: https://doi.org/10.3758/s13421-018-0808-6

Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. DOI: https://doi.org/10.1080/01621459.1959.10501497

Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, *7*, 1074–1095. DOI: https://doi.org/10.1037//0096-1523.7.5.1074

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. DOI: https://doi.org/10.1126/science.7777863

Thomas, D. C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., & Armstrong, B. G. 1985. The problem of multiple inference in studies designed to generate hypotheses. *American Journal of Epidemiology*, *122*(6), 1080–1095. DOI: https://doi.org/10.1093/oxfordjournals.aje.a114189

Tukey, J. W. 1953. *The problem of multiple comparisons* [Mimeographed notes]. Princeton, NJ: Princeton University.

Tukey, J. W. 1980. We need both exploratory and confirmatory. *The American Statistician, 34*(1), 23–25. DOI: https://doi.org/10.2307/2682991

van Heuven, V. J., & van Zanten, E. 2005. Speech rate as a secondary prosodic characteristic of polarity questions in three languages. *Speech Communication*, 47(1–2), 87–99. DOI: https://doi.org/10.1016/j.specom.2005.05.010

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. DOI: https://doi.org/10.1016/j.jml.2018.07.004

Vasishth, S., & Nicenboim, B. 2016. Statistical methods for linguistic research: Foundational ideas–Part I. *Language and Linguistics Compass*, *10*(8), 349–369. DOI: https://doi.org/10.1111/lnc3.12201

Vatikiotis-Bateson, E., Barbosa, A. V., & Best, C. T. 2014. Articulatory coordination of two vocal tracts. *Journal of Phonetics*, *44*, 167–181. DOI: https://doi.org/10.1016/j.wocn.2013.12.001

Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S* (4th Edition). Springer, New York. DOI: https://doi.org/10.1007/978-0-387-21706-2

von der Malsburg, T., & Angele, B. 2017. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of memory and language*, *94*, 119–133. DOI: https://doi.org/10.1016/j.jml.2016.10.003

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., & van Snellenberg, J. X. 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, *45*, S210–S221. DOI: https://doi.org/10.1016/j.neuroimage.2008.10.061

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. DOI: https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package = tidyverse.

Wieling, M. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. DOI: https://doi.org/10.1016/j.wocn.2018.03.002

Winter, B. 2011. Pseudoreplication in phonetic research. In: *Proceedings of the International Congress of Phonetic Science*, 2137–2140. Hong Kong.

Winter, B. 2014. Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967. DOI: https://doi.org/10.1002/bies.201400028

Winter, B. 2015. The other N: The role of repetitions and items in the design of phonetic experiments. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: The University of Glasgow.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1–2), 23–43. DOI: https://doi.org/10.1016/s0167-6393(98)00048-x

How to cite this article: Roettger, T. B. 2019 Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 1, pp. 1–27. DOI: https://doi.org/10.5334/labphon.147

Submitted: 20 March 2018 **Accepted:** 15 November 2018 **Published:** 04 January 2019

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS

Timo Roettger[1] / Matthew Gordon[2]

Methodological issues in the study of word stress correlates

[1] University of Cologne, 50923 Cologne, Germany
[2] University of California, Santa Barbara, CA 93106, USA, E-mail: mgordon@linguistics.ucsb.edu

Abstract:
The investigation of acoustic correlates of word stress is a prominent area of research. The literature is rife with studies of the acoustic exponents of what is often referred to as stress but the methodological diversity of this research has created an unclear picture of the properties robustly associated with it. The present paper explores the methodological issues involved in examining word stress correlates with the goal of proposing a set of recommendations for future research. Based on a survey of 110 (sub-) studies on 75 languages, desiderata for research on the acoustics of stress are identified: descriptions of employed methods should be as detailed as possible, speech material should be designed to allow for teasing apart word level stress from phrase level prominence, and sample sizes should be chosen according to statistical considerations.

Keywords: word stress, phonetics, phonology, typology

DOI: 10.1515/lingvan-2017-0006

Received: March 7, 2017; **Accepted:** April 26, 2017

1 Introduction

The investigation of acoustic correlates of word stress is a prominent area of research dating back to pioneering work by Fry (1955, 1958). The literature is rife with studies of the acoustic exponents of stress but the methodological diversity of this research has created an unclear picture of the properties reliably associated with it. The present paper explores the methodological issues involved in examining word stress correlates with the goal of proposing a set of recommendations for future research.

2 Methodology

The present survey is gleaned from multiple sources, including papers found in phonetic and areal studies journals, working papers, and assorted books and dissertations. The study included only works containing quantitative results specifically targeting the acoustic correlates of stress (as opposed to acoustic correlates of other prosodic properties, e. g. intonation, prosodic constituency, boundary phenomena, rhythm and timing). Only studies involving adult speakers without reported speech impairments were considered.[1]

The corpus included a total of 110 (sub-)studies on 75 languages (or language varieties), which are plotted geographically in Figure 1(see Gordon and Roettger this volume for a list of languages surveyed).

Matthew Gordon is the corresponding author.

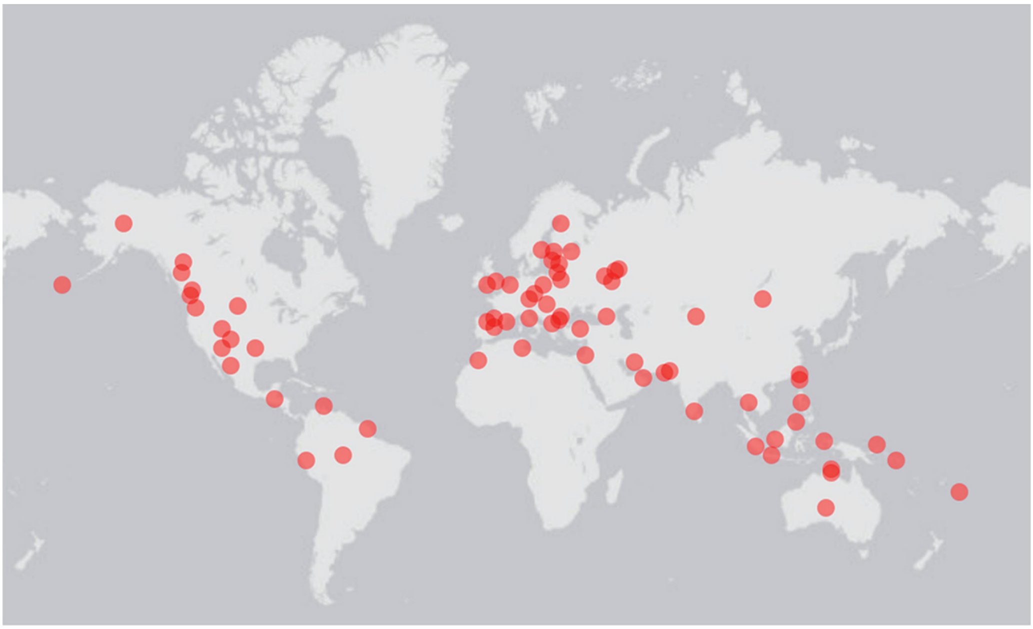

Figure 1: Geographical distribution of languages included in the survey of acoustic correlates of stress plotted via the "lingtypology" package (Moroz 2017) for R (2017).

For each surveyed source, besides the acoustic parameters signaling word stress (for discussion, see Gordon and Roettger this volume), we logged several methodological aspects.

The first was the corpus type examined, either lab speech or spontaneous speech (see Beckman 1997; Xu 2010). The former refers to speech that is usually recorded in the form of reading aloud pre-composed stimuli. Studies using lab speech examined target words in isolation, in a context phrase that varied depending on the target word, or in a fixed metalinguistic phrase, e. g. "Say ________ again" or "I say ________". We also included larger pre-composed texts in this category. Spontaneous speech is here defined as "speech that is not read to script" (Beckman 1997: 7) and includes narratives, conversations, and interviews. In addition to the corpus type, the position of the target word is also provided, where options include phrase-final (including both final words in a phrase and words in isolation), phrase-non-final, or varied, in cases where the phrase position was not systematically varied. Cases in which final and non-final were crossed were labeled as 'controlled'. If the position of the target word was unclear, a common situation for spontaneous speech, it was labeled as 'unspecified'.

Also tracked was whether the target words occurred in a position that controlled for post-lexical tonal events such as accents. To operationalize encoding, five settings were adopted: accented, unaccented, accent controlled, implicit contrast or unspecified (in the case of spontaneous speech and studies employing carrier phrases that are not presented). The accented category includes contexts in which target words are either stated to or likely to carry a phrasal accent because they are explicitly focused, e. g. "I said ________ not *car*." The unaccented category comprises cases in which the target word was either explicitly stated to be unaccented or in which focus fell on another word in the phrase, e. g. "I said ________ *slowly*, not *quickly*". 'Accented controlled' reflects cases in which both the accented and unaccented conditions were crossed. Implicit contrast refers to cases in which focus was implicitly on the target word because it was the only varying element in the corpus, e. g. in a list of isolated words or words in an invariant (metalinguistic) phrase. It is impossible to know whether the target words in these cases carried a phrasal accent or not.

Additionally, we tracked the word stress levels examined (primary stress (1S), secondary stress (2S) and unstressed (US)) and the number of speakers, words (or alternative operationalizations for the corpus size), and repetitions of each targeted word under a given condition (applicable only for lab speech). For most studies, the most easily recoverable information was the number of words containing targeted syllables (or vowels). Depending on the corpus, each target word potentially consisted of one measured syllable (or vowel), either stressed or unstressed in studies employing cross-word comparisons, or more than one in intraword comparisons. For certain studies, the total number of data points rather than the number of targeted words is given since it was the only information provided.

The statistical test(s) employed and results for the examined acoustic parameters were also logged. Final notes columns provide additional information about the methodology and/or results. In coding statistics, descriptive studies are labeled as such and, in the case of those also using inferential statistics (which characteristically implies the simultaneous presentation of descriptive statistics), the type of test(s) employed is provided. Even though they are of interest, the statistical results of studies in our corpus are difficult to compare, since there is considerable variation in the statistical models employed and the population over which inferences are drawn. Discussing the diversity of statistical choices and their potential impact on the reported results goes beyond the scope of this paper. Therefore, we have limited the present survey to encode the effect direction (e. g. greater duration of stressed syllables) of differences reported to be statistically significant regardless of the magnitude of the difference. Of course, given the diversity of statistical practices represented in the corpus, the reported significant differences should be regarded with caution. Furthermore, even differences that are genuinely statistically significant may not be perceptually relevant. In the absence of direct perception experiments, this distinction is difficult to assess, as even just-noticeable-differences reported in the literature and based on a small set of languages may not reflect difference limens for speakers of all languages. Although relying on statistical results to assess robustness runs the risk of attaching unwarranted importance to differences that may turn out not to be meaningful, we view this as a preferable alternative to dismissing results that could potentially be meaningful.

3 Results

3.1 Teasing apart word-level stress from phrase-level prominence

One challenge in studying word stress is teasing apart word-level stress from properties attributed to other sources. A salient potential source of confound stems from phrase-level prominence (for a recent discussion, see Gordon 2014). In many languages, certain elements of the utterances are intonationally marked by tonal movements. In languages with word stress, pragmatically highlighted words often co-occur with post-lexical tonal events (pitch accents, phrasal accents, edge tones), which are usually realized on or near a syllable bearing word stress. Research conducted mainly on languages that exhibit pitch accents (e. g. English, French, Italian, and German) reports that the co-occurrence of an accent with a syllable results in temporal and spatial expansion of the articulatory gestures involved, which, in turn, leads to detectable acoustic differences (e. g. Cho 2005; Cho and Keating 2009; Cho and McQueen 2005; Harrington et al. 2000).

Languages also commonly modify temporal and spatial phonetic parameters at prosodic edges. There are two major phenomena falling under the umbrella of boundary-induced strengthening: “initial strengthening” and “final lengthening”. The former includes the phrase-initial enhancement of acoustic parameters that are phonemically contrastive. For instance, in languages with aspirated stops, voice onset time (VOT) is typically longer phrase-initially than phrase-medially (Cooper 1991; Pierrehumbert and Talkin 1992; Jun 1993; Cho and Jun 2000; Choi 2003; Cole et al. 2003). At the right edge, prosodic constituents are longer phrase-finally than phrase-medially in many languages, including Arabic (De Jong and Zawaydeh 1999), Dutch (Gussenhoven and Rietveld 1992), English (e. g. Beckman and Edwards 1990; Edwards et al. 1991; Wightman et al. 1992; Cho 2005; Turk and Shattuck-Hufnagel 2007), and Hebrew (Berkovits 1991) among many others.

These properties make it imperative to methodologically tease apart phonetic factors attributed to genuine word stress from those caused by phrasal phenomena such as prosodic boundaries and accentual prominence. Probably most crucial in avoiding these confounds is the phrasal context in which the target occurs. Figure 2 shows the number of studies in the database employing words elicited in lab speech (either in isolation or in a carrier phrase), and words elicited in spontaneous speech.

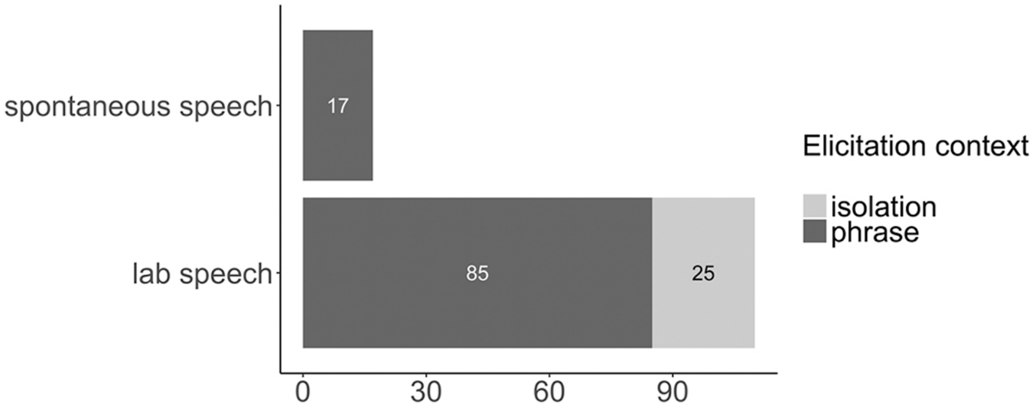

Figure 2: Number of studies as a function of corpus type (y-axis) and elicitation context in which target words appeared (color coded) for all (sub-)studies in the database. Note that studies using more than one type of context are included in the counts of all relevant categories.

As is evident in Figure 2, many studies in the database examined words in isolation, where word stress is straightforwardly confounded with phrase-level prominence. Studies based on spontaneous speech also run the risk of conflating word-level and phrase-level properties unless the target words are carefully chosen based on their role in the overall phrasal prosody. Many prominence asymmetries found in the survey thus might ultimately be attributed to phrase-level properties.

The majority of studies examine words in a phrase. In most of these, the target appeared in non-final position in at least one condition (which in some studies was crossed with final position). Riviera-Castillo and Pickering (2004) on Papiamentu, Astruc and Prieto (2006), Lehiste et al. (2005) on Meadow Mari, Lehiste et al. (2008) on Livonian, and Yakup and Sereno (2016) on Uyghur are exceptional in positioning the word only in phrase-final position, making it difficult to tease apart word stress from boundary-induced effects.[2]

There are several examples of potential interference from phrase-level effects in the database. Starting with the left-edge, Pycha (2006) finds that intensity in Turkish is *not* higher on a stressed final syllable than on an unstressed penult, contradicting results in Levi (2005). Pycha (2006) suggests that the lack of a difference in her study may be due to initial strengthening of the penult, which was also the first syllable in her database employing disyllabic words. Similarly, the lengthening of the initial stressed syllable in Estonian (Lehiste 1966) could be attributed to domain-initial strengthening.

Turning to the right edge, in their study of Indonesian stress in isolated words, Adisasmito-Smith and Cohn (1996) find that the vowel in the stressed penult for one of two speakers is longer than the pretonic but not the post-tonic (and final) vowel, the latter of which the authors suggest is lengthened due to its position, thereby obscuring a lengthening effect due to word stress. Cho (2006) observes that F0 in Lakhota is higher on the stressed syllable of a disyllabic word when the stress is initial but not when the stress is on the second syllable. The absence of higher F0 on a stressed second syllable is potentially due to final lowering, i. e. a phrase-final low boundary tone.

Similarly, higher F0 on a stressed syllable may be attributed to post-lexical tonal events (e. g. Bolinger 1958, 1961; Beckman 1986; Ladd 2008 inter alia). Several studies, e. g. Gonzalez (1970) on Tagalog, Adisasmito-Smith and Cohn (1996) on Indonesian, Chian and Chiang (2005) on Saisiyat, Hargus and Beavert (2006) on Yakima Sahaptin, Gordon and Applebaum (2010) on Kabardian, and Simard et al. (2014) on Savosavo, find raised F0 to be a correlate of stress in isolated words. It is unclear whether F0 in these cases is attributed to word stress or to a post-lexical tonal event.[3]

Looking at lab speech stimuli embedded into a context phrase, we encounter difficulties in disentangling word prominence from phrasal prominence that is orthogonal to the position of the target word in the phrase. Intonation commonly expresses discourse relationships such as information structure (e. g. focus or topic) and information status (e. g. given vs. new information) (see, for example, Pierrehumbert 1980 on English, Grice et al. 2005a on Italian, Grice et al. 2005b on German, and Meyer and Mleinek 2006 on Russian). Many languages flag explicitly or implicitly focused constituents as well as new information with post-lexical tonal events which themselves might carry additional acoustic prominence. Thus, the speech material uttered in context phrases should carefully control for these discourse effects.

Vogel et al. (2016) demonstrate the potential influence of focus on the realization of prominence in a study of four languages (Spanish, Greek, Turkish, and Hungarian) that teases apart lexical prominence and phrasal prominence induced by focus through a series of dialogues. All target words in their study are non-final in a phrase but vary in whether they appear before another word that is narrowly focused or not. For example, the word *rope* is new focused information in the sentence “Ayse said ‘**rope**’ in the afternoon” uttered in response to the question “What did Ayse say in the afternoon?”. On the other hand, *rope* is background information that is not focused in the sentence “No. Ayse **said** ‘rope’ in the afternoon, she didn’t **write** it” in response to the question “Did Ayse write ‘rope’ in the afternoon?”. Vogel et al. (2016) find that stressed syllables in Hungarian have higher F0 only when focused and that the effect of stress on duration and F0 in stressed syllables is enhanced under focus in Greek. On the other hand, focus in Turkish induces a lowering effect on F0 in the stressed final syllable relative to its non-focused counterpart. They suggest that this pattern reflects a phrase boundary following a focused word in Turkish. The Vogel et al. (2016) results demonstrate the interaction of focus with both post-lexical prominence and prosodic phrasing, both of which make the evaluation of the acoustic correlates of word stress difficult (see also, for example, Beckman and Pierrehumbert 1986 on Japanese, Kanerva 1990 on Chicheŵa, Hayes and Lahiri 1991 on Bengali, Jun 2005 on Korean).

Teasing apart phrasal prominence from word-level prominence is more subtle than merely ensuring the target word is not contrastively focused. Although metalinguistic carrier phrases like “I say ______ again” shield the target word from the right edge of the phrase, the targets in such phrases are invariably new information and implicitly contrasted with each other, making a post-lexical prominence on the word likely. In fact, in their study of English stress, Plag et al. (2011) classify the condition in which the target is embedded in the carrier phrase “She said ______ again” as the ‘accented’ condition, which they contrast with an ‘unaccented’ condition in sentences with narrow focus on another word. The same context that is employed to elicit the focused condition in the Plag et al. study is thus the one used to trigger the non-focused condition in many other studies.

There are only seven studies in the database that employ a condition in which the target is clearly non-focused. These studies are not uniform, however, in the strategy they adopt to ensure the target is not focused; it is conceivable that the different defocusing constructions could induce their own prominence-enhancing or suppressing effects, e. g. post-focal compression (e. g. Bruce 1982 on Swedish, Cooper et al. 1985 on English, Féry and Kügler 2008 on German, Sadat-Tehrani 2008 on Persian and Tibetan, Lee and Xu 2010 on Korean, and Wang et al. 2011 on Uyghur).[4]

The potential conflation of post-lexical tonal events with word-level stress does not mean that studies that confound the two sources of prominence are completely uninformative. At least in languages that exhibit a post-lexical tonal event that is referred to as a 'pitch accent', the tonal event (per definition) docks on syllables carrying word stress. In these cases, prominence on a particular syllable in isolation is suggestive of primary word stress on that syllable.

In spite of the fact that post-lexical tonal events often co-occur with word stress in many well-described languages, one should nevertheless exercise caution when assuming a particular mapping of phrase-level prominence and word stress. Although tonal events such as pitch accents are typically assigned in “bottom-up” fashion docking on the primary stressed syllable of a word, there are many prosodic systems that exhibit tonal events that look like pitch accents but do not show the same temporal co-occurrence with stressed syllables. In Chickasaw, phrasal accent assignment operates orthogonally to word-level stress, leading to cases in question phrases in which a syllable that is unstressed at the word-level carries a tonal event (Gordon 2003). “Top-down” tonal placement in Chickasaw thus demonstrates the potential pitfalls of using phrasal accents to diagnose word stress. More generally, the link between phrasal accent and word stress commonly assumed relies on a particular interpretation of intonational events that may at least in part be an artifact of theoretical biases stemming from familiarity with more thoroughly studied (typically European) languages. Certain cases may be open to re-interpretation in different terms much like some “stress” languages, e. g. Indonesian (Goedemans and van Zanten 2007), have been re-analyzed as intonation-only languages.

Furthermore, even in languages in which phrasal accents diagnose word stress location, cues to word stress are not necessarily equivalent to those associated with phrasal accent. Word stress and phrasal accent may diverge acoustically, as shown in studies of several languages in which F0 is a more reliable correlate of phrasal accent than word stress when phrasal position is controlled for, e. g. Huss (1978) on American English, Sluijter and Van Heuven (1996) on Dutch, Hintz (2006) on South Conchucos Quechua (2006), Ortega-Llebaria (2006) on Spanish, and Guion et al. (2010) on Balsas Nahuatl.

The upshot is that without detailed knowledge of how a language prosodically encodes discourse relations and the nature of its tonal events, the assessment of the source of acoustic prominence remains difficult.

Figure 3 illustrates the distribution of studies using lab speech according to the phrasal position of the target word and the accentual context. Most studies provided information on the target word position (all but 10). Most studies either controlled for phrase position by using stimuli with target words in both final and non-final positions (11) or used stimuli in non-final position (53). However, of these 64 studies, 12 studies did not provide any further information on potential confounds due to tonal events associated with information structure (green bars). Another 33 studies used either target words in accented positions or contexts in which the target words were explicitly contrasted (red bars). In these cases, the possibility that the observed prominence patterns are confounded with post-lexical prominence cannot be ruled out. In turn, this leaves us with 19 out of 85 studies (22 %), that both controlled for phrasal position and assumed post-lexical prominence due to accent. Thus, a substantial number of studies in our survey are characterized by experimental design choices that make an interpretation of their results with regard to word stress difficult.

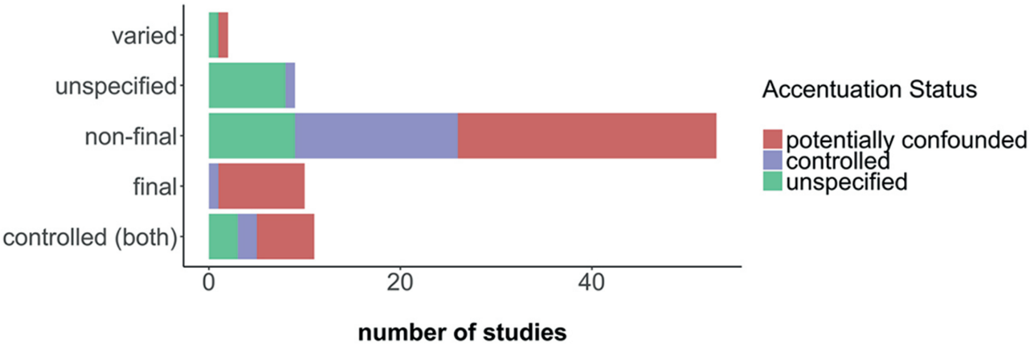

Figure 3: Number of studies as a function of target word position (y-axis) and accent information (color coded) within (sub-) studies using lab speech. Note that studies using more than one context are included in counts of all relevant categories.

3.2 Sample size

A crucial desideratum for any behavioral study, including those of acoustic correlates of stress, is the use of a large enough sample to confidently infer that the results reflect the broader population of speakers of a language beyond those contributing the data for the particular study. Additionally, a sufficient sample of lexical items is also important to ensure the generalizability of the obtained findings beyond the recorded words (e. g. Clark 1973). Another aspect of the design that affects the confidence with which we are able to estimate a representative effect is the number of recorded instances (referred to as repetitions) of a particular word produced by a particular speaker.

As Figure 4 shows, studies in the database vary considerably in the number of speakers, lexical items, and repetitions.

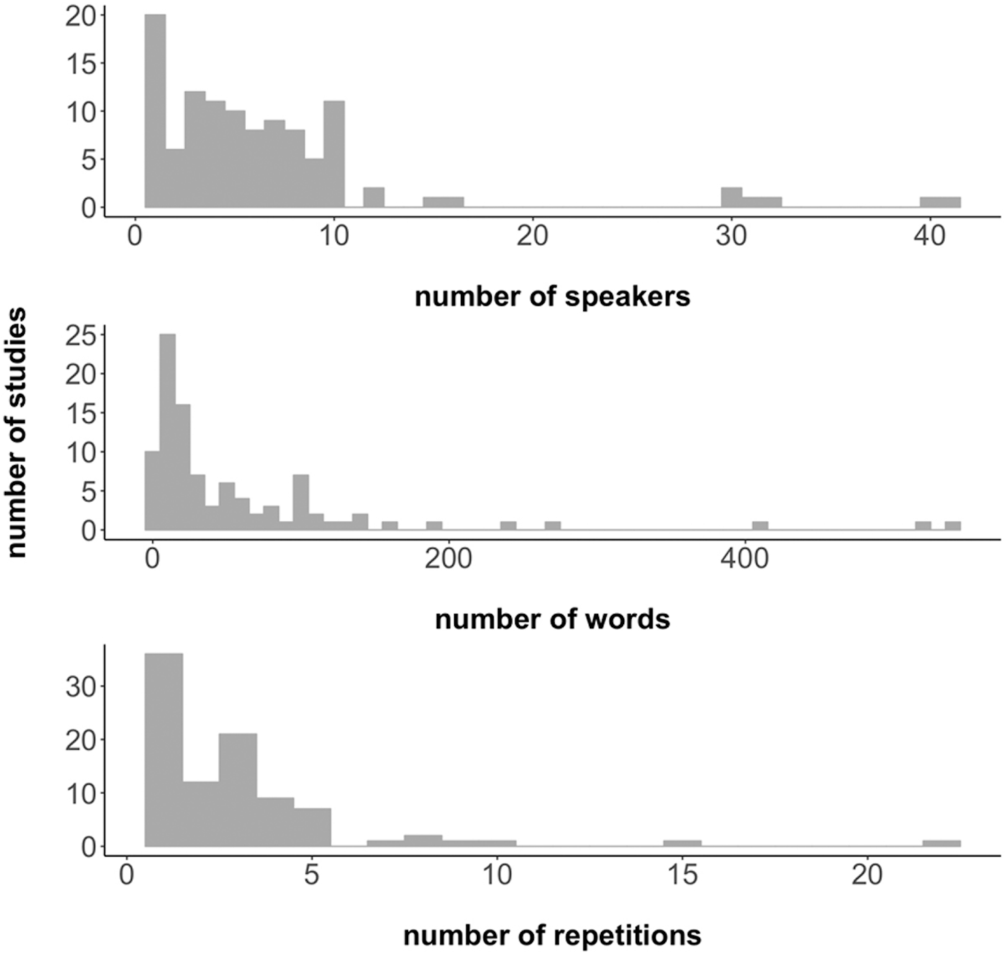

Figure 4: Number of speakers (top panel, bin width = 1), words (middle panel, bin width = 10), and repetitions (bottom panel, bin width = 1) represented in (sub-) studies in the database.

The majority of studies employ data from between one and ten speakers but vary widely within this range. On the other hand, the mode (20 studies) is a single speaker. The eight studies including data from more than 10 speakers range from 12 speakers in the study by Fry (1955) on American English and in the study by Everett (1998) on Pirahã to 41 in the work by Lesho (2013) on Cavite Chabacano.

It is interesting to note that only four studies reach the desired target of six speakers of each gender suggested by Ladefoged (1997: 140) in his recommendations for phonetic fieldwork. Roughly 46 % (51 of 110) of the surveyed studies involve at least six speakers, which Ladefoged regards as the absolute minimum number necessary.

It should be emphasized, however, that Ladefoged's recommendations are not based on statistical considerations. The minimum sample size allowing for statistical inference beyond the tested sample depends largely on statistical power, a dimension independent from significance (for a recent overview with regard to linguistic research, see Vasishth and Nicenboim 2016 or Kirby and Sonderegger). Assessing power is more difficult than assessing significance because it depends on multiple factors including the true (or expected) effect size, the sample size, and the degree of variability. For one, the effect size and its variability are strong determinants of how large a sample should be to enable statistically robust inferences over a speaker population or a language's lexicon (or both). The variability of an effect, say a durational difference, is further dependent on the variability of the measurement. Speech data are very noisy and dependent upon many factors. Controlling for confounding factors is one way of reducing variability. Spontaneous speech is thus generally expected to yield more variable data due to the large number of confounding factors (higher level prosodic structure, intonation, syntax, etc.).

Another way to reduce variability and to achieve a better estimate of the effect is to increase observations for a single speaker. Researchers usually include many lexical items for a more precise estimate of the acoustic patterns of a particular speaker. This, in fact, might reduce variability attributed to the large variability across words, thereby enabling a better estimate of the behavior of the speech community. Looking at Figure 4, the majority of studies in our corpus use between 1 and 40 different lexical items. Depending on how a lexical item was defined (word type or word form) and the corpus (lab speech vs. spontaneous speech), some studies look at many hundreds of words.

Similarly, multiple repetitions of the same lexical item allow for a more precise estimate of the representative acoustic form of a specific word and thus reduces variability, again enabling a better estimate of the true effect in the population. Studies vary widely in repetition count ranging from one (where one repetition refers to only one instance of a word) in many studies to 22 in the Caldecott (2009) study of St'át'imcets. It is apparent in Figure 4 that the majority of studies use between 1 and 6 repetitions with a single repetition being most common (35 studies). An advantage to recording more than one token is that it allows for the possibility of discarding tokens associated with dysfluency or other localized problems with the recording. On the other hand, a large corpus consisting of many different lexical items ensures that the examined words are representative of those in the language.

3.3 Corpus composition

The issue of lexical breadth in a corpus is an important one that has received short shrift in the stress literature. In addressing the potential source of discrepancies between the results of their study of Tashlhiyt and those of Gordon and Nafi (2012), Roettger et al. (2015; see also Roettger accepted for a detailed analysis) suggest that a skewing in favor of words consisting of a light syllable followed by heavy syllable may have contributed to the evidence for final prominence in the Gordon and Nafi study. In a related vein, studies of Turkish (Levi 2005; Pycha 2006; Vogel et al. 2016) indicate a difference in the acoustic exponents of stress between words with exceptional penultimate stress vs. those with default final stress. These studies suggest that corpora should include sufficiently diverse word structures to uncover potential interactions between word type and the manifestation of prominence.[5]

Another related issue is whether comparisons between stressed and unstressed syllables are syntagmatic, i. e. involve comparisons within words, or paradigmatic, i. e. involve comparisons across words. The effects of a confounding variable potentially create differences between the two types of comparisons. For example, in their study of Hebrew, Silber-Varod et al. (2016) observe an interaction between position of the syllable and the realization of stress (see also Tuomainen et al. 1999 for a similar interaction in Finnish). Disyllabic words with final stress have higher F0 on the stressed final syllable, whereas disyllabic words with penultimate stress have *lower* F0 on the stressed penult. This result amounts to an effect of syllable position on F0 values such that the final syllable has higher values than the penult.[6] By comparing stressed and unstressed syllables across words with different stress positions, Silber-Varod et al. (2016) identified this effect, which would have otherwise escaped notice if their study had included only intraword comparisons.

4 Recommendations for best practice

In this section, we use the methodological variation observed in a survey of 110 (sub-) studies as a springboard for proposing recommendations for future research on word stress. These recommendations surround three key design aspects: the transparency of the experimental design, the speech material, and the sample size. For recommendations concerning instrumentation, e. g. microphones, recording equipment, file format, etc., the reader is referred to Ladefoged (1997, 2003), Bowern (2008), Butcher (2013), Chelliah and De Reuse (2011), and Maddieson (2001) .

First, the survey contains many studies lacking sufficient methodological details to allow for either evaluation or replication. Particularly crucial omissions in several studies include the phrasal position and intonational context in which target words were uttered. As a minimum requirement, future studies of stress should thus make explicit (to the extent possible) the phrasal and accentual structure of their speech material (in addition, of course, to other relevant methodological details, e. g. corpus size, number of repetitions and speakers, etc.).

Second, it is imperative to control for context to ensure that the results reflect genuine word stress rather than phrasal prominence. Target words uttered in isolation make it difficult to disentangle word stress from phrase-level prominence. Moreover, target words should be removed from phrasal boundaries to avoid confounds attributed to either boundary-associated strengthening or tonal events. Finally, intonational structure should be controlled for, by either putting target words in an unaccented position or by examining both accented and unaccented positions. Only 19 lab speech studies in our corpus avoided these confounds altogether by controlling for phrasal and accentual structure. Another consideration is the comparability of the segments targeted for measurement since properties such as vowel quality, consonant type, surrounding sounds, and syllable structure all potentially impact common acoustic diagnostics of stress.

Third, a sample must be sufficiently large to infer that results reflect the broader population of speakers. A sufficient sample of lexical items is also important to ensure the generalizability of the obtained findings beyond the recorded words (e. g. Clark 1973; Winter 2011; Judd et al. 2012). The choice of the sample size is ultimately a question of statistical power (Vasishth and Nicenboim 2016 or Kirby and Sonderegger in press). Properties that determine statistical power are the (true) effect size, sample size, and degree of variability. Reduction of variability can be achieved by controlling for other factors interfering with measurements and by increasing the number of observations through additional words and/or repetitions. If the effect size and its variability can be estimated from previous studies, statistical power can be estimated in order to determine required sample size (see Kirby and Sonderegger in press). Consequently, even if some studies in our corpus remain anecdotal, i. e. do not allow for statistical inference over a broader population, they might allow us to generate hypotheses and represent valuable departure points for further investigations.

In summary, we hope that these recommendations based on evaluation of a survey of acoustic studies will encourage more careful experimental designs in future studies. Following these recommendations as much as possible given the constraints of a particular study will likely lead to more generalizable results within a language and to typological data that are easier to compare. It is hoped that this paper brings the relevance of experimental design to the attention of typologists, phoneticians, and phonologists, and that it encourages a renewed interest in experimental methods and statistical analyses.

Acknowledgments

The authors thank Shigeto Kawahara, an anonymous reviewer, and the Department of Phonetics in Cologne for their helpful comments and feedback on an earlier version of this manuscript. Any remaining errors or misconceptions are our own.

Notes

[1] The corpus (in form of a table) is publically available online at https://osf.io/9r2cd/ alongside a script to reproduce respective counts presented in this manuscript. To establish a reliable and informative corpus that can be used in the future, cited authors are encouraged to submit corrections, if we have interpreted respective aspects of their method and/or results incorrectly. Further, we would like to invite scholars that have published work on word stress that is not logged in the present corpus to share their results with us for inclusion in the database.

[2] The controlled comparison of target words in Riviera-Castillo and Pickering (2004) involves words in final position, although they also investigate all syllables in another set of sentences. Astruc and Prieto (2006) examine accented and unaccented words in final position. The two Lehiste et al. studies (2005, 2008) target words in both phrase-final and sentence-final position.

[3] A similar interference of phrasal pitch accent also potentially obtains for words in final position of a phrase, a position toward which pitch accents tend to gravitate in phrases lacking narrow focus (Ladd 2008). There are relatively few studies in the database that do not explicitly control for accent but position words in phrase-final position. Among these, only Lehiste et al. (2005) on Meadow Mari and Lehiste et al. (2008) on Livonian find F0 to be a correlate of prominence.

[4] In two studies (Crosswhite 2003 on Bulgarian and Polish and Sadeghi 2011 on Persian), the target word appears in a carrier phrase in which another word is explicitly focused, e. g. "I say ______ **again**" where the target appears in prefocal position. The target in such phrases could still be implicitly focused because it is systematically varied while the phrase is held constant. These are categorized as deaccented to distinguish them from studies in which the carrier phrase lacks any explicit focus, a context in which the risk of implicit focus on the target is greater.

[5] If, however, speakers produce multiple items, the statistical analysis must account for this introduced non-independence (e. g. Clark 1973) to avoid statistically misguided interpretations (Judd et al. 2012; see also Winter 2011, for a brief discussion of this issue in phonetic experiments).

[6] Lieberman (1960), Crosswhite (2003), and Yakup and Sereno (2016) also report different results between intraword and interword comparisons of stressed and unstressed syllables.

References

Adisasmito-Smith, Niken and Abigail C. Cohn. 1996. Phonetic correlates of primary and secondary stress in Indonesian: A preliminary study. *Working papers of the Cornell phonetics laboratory* 11. 1–16.

Astruc, Lluïsa & Pilar Prieto. 2006. Acoustic cues of stress and accent in Catalan. *Proceedings of 3rd International Conference on Speech Prosody, Dresden, Germany*.

Beckman, Mary. 1986. *Stress and non-stress accent*. Dordrecht: Foris.

Beckman, Mary. 1997. A typology of spontaneous speech. In Yoshinori Sagisaka, Nick Campbell & Norio Higuchi (eds.), *Computing prosody: Computational models for processing spontaneous speech*, 7–26. New York: Springer Science & Business Media.

Beckman, Mary and Jan Edwards. 1990. Lengthenings and shortenings and the nature of prosodic constituency. In John Kingston & Mary Beckman (eds.), *Papers in laboratory phonology I*, 179–200. Cambridge: Cambridge University Press.

Beckman, Mary & Janet Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3. 255–309.

Berkovits, Rochele. 1991. The effect of speaking rate on evidence for utterance-final lengthening. *Phonetica* 48. 57–66.

Bolinger, Dwight L. 1958. A theory of pitch accent in English. *Word* 14. 109–149.

Bolinger, Dwight L. 1961. Contrastive accent and contrastive stress. *Language* 37(1). 83–96.

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. London: Palgrave Macmillan.

Bruce, Gösta. 1982. Developing the Swedish intonation model. *Lund University, Dept. of Linguistics Working Papers* 22. 51–116.

Butcher, Andy. 2013. Research methods in phonetic fieldwork. In Mark Jones & Rachael-Anne Knight (eds.), *The bloomsbury companion to phonetics*, 57–78. New York: Bloomsbury Publishing.

Caldecott, Marian. 2009. *Non-exhaustive parsing: Phonetic and phonological evidence from St'át'imcets*. University of British Columbia. Ph.D. dissertation.

Chelliah, Shobhana & Willem De Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. New York: Springer.

Chian, Wen-yu & Fang-mei Chiang. 2005. Saisiyat as a pitch accent language: Evidence from acoustic study of words. *Oceanic Linguistics* 44. 404–426.

Cho, Taehong. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/in English. *Journal of the Acoustical Society of America* 117. 3867–3878.

Cho, Taehong. 2006. An acoustic study of the stress and intonational system in Lakhota: A preliminary report. *Speech Sciences* 13. 23–42. (Published by The Korean Association of Speech Sciences).

Cho, Taehong and Sun-Ah Jun. 2000. Domain-initial strengthening as featural enhancement: Aerodynamic evidence from Korean. *Chicago Linguistics Society* 36. 31–44.

Cho, Taehong and Patricia Keating. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics* 37. 466–485.

Cho, Taehong and James McQueen. 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics* 33. 121–157.

Choi, Hansook. 2003. Prosody-induced acoustic variation in English stop consonants. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*, 2661–2664.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12. 335–359.

Cole, Jennifer, Hansook Choi, Heejin Kim & Mark Hasegawa-Johnson. 2003. The effect of accent on the acoustic cues to stop voicing in Radio News speech. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*. 2665—2668.

Cooper, A. 1991. Glottal gestures and aspiration in English. PhD dissertation Yale University .

Cooper, William E., Stephen J. Eady & Pamela R Mueller. 1985. Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America* 77. 2142–2156.

Crosswhite, Katherine. 2003. Spectral tilt as a cue to word stress in Polish, Macedonian, and Bulgarian. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*, 767–770.

De Jong, Kenneth & Bushra Adnan Zawaydeh. 1999. Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics* 27. 3–22.

Edwards, Jan, Mary Beckman & Janet Fletcher. 1991. The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America* 89. 369–382.

Everett, Keren. 1998. The acoustic correlates of stress in Pirahã. *Journal of Amazonian Languages* 1(2). 104–162.

Féry, Caroline & Frank Kügler. 2008. Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics* 36. 680–703.

Fry, Dennis B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* 27. 765–768.

Fry, Dennis B. 1958. Experiments in the perception of stress. *Language and Speech* 1. 120–152.

Goedemans, Rob & Ellen van Zanten. 2007. Stress and accent in Indonesian. *LOT Occasional series* 9. 35–62.

Gonzalez, Andrew. 1970. Acoustic correlates of accent, rhythm, and intonation in Tagalog. *Phonetica* 22. 11–44.

Gordon, Matthew. 2003. The phonology of pitch accent placement in Chickasaw. *Phonology* 20. 173–218.

Gordon, Matthew. 2014. Disentangling stress and pitch accent: Toward a typology of prominence at different prosodic levels. In Henry van der Hulst (eds.), *Word stress: Theoretical and typological issues*, 83–118. Cambridge: Cambridge University Press.

Gordon, Matthew & Ayla Applebaum. 2010. Acoustic correlates of stress in Turkish Kabardian. *Journal of the International Phonetic Association* 40. 35–58.

Gordon, Matthew & Latifa Nafi. 2012. The acoustic correlates of stress and pitch accent in Tashlhiyt Berber. *Journal of Phonetics* 40. 706–724.

Gordon, Matthew & Timo B. Roettger. This issue. *Acoustic correlates of word stress: A cross-linguistic survey*.

Grice, Martine, Stefan Baumann & Ralf Benzmüller. 2005b. German intonation in autosegmental-metrical phonology. In Sun-Ah Jun (eds.), *Prosodic typology – The phonology of intonation and phrasing*, 55–83. New York: Oxford University Press.

Grice, Martine, Mariapaola D'Imperio, Michelina Savino & Cinzia Avesani. 2005a. A strategy for intonation labelling varieties of Italian. In Sun-Ah Jun (eds.), *Prosodic typology – The phonology of intonation and phrasing*, 362–389. New York: Oxford University Press.

Guion, Susan, Jonathan D. Amith, Christopher S. Doty & Irina A. Shport. 2010. Word-level prosody in Balsas Nahuatl: The origin, development, and acoustic correlates of tone in a stress accent language. *Journal of Phonetics* 38. 137–166.

Gussenhoven, Carlos & A. C. M. Rietveld. 1992. Intonation contours, prosodic structure and preboundary lengthening. *Journal of Phonetics* 20. 283–303.

Hargus, Sharon & Virginia Beavert. 2006. A note on the phonetic correlates of stress in Yakima Sahaptin. *University of Washington Working Papers in Linguistics* 24. 64–95.

Harrington, Jonathan, Janet Fletcher & Mary Beckman. 2000. Manner and place conflicts in the articulation of accent in Australian English. In Michael B. Broe & Janet Pierrehumbert (eds.), *Papers in laboratory phonology V: Language acquisition and the lexicon*, 40–51. Cambridge: Cambridge University Press.

Hayes, Bruce & Aditi Lahiri. 1991. Bengali intonational phonology. *Natural Language and Linguistic Theory* 9. 47–96.

Hintz, Diane. 2006. Stress in South Conchucos Quechua: A phonetic and phonological study. *International Journal of American Linguistics* 72. 477–521.

Huss, Volker. 1978. English word stress in the post-nuclear position. *Phonetica* 35. 86–105.

Judd, Charles M., Jacob Westfall & David A. Kenny. 2012. Treating stimuli as random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology* 103. 54–69.

Jun, Sun-Ah. 1993. The phonetics and phonology of Korean Prosody. PhD Dissertation Ohio State University.

Jun, Sun-Ah. 2005. Korean intonational phonology and prosodic transcription. In Sun-Ah Jun (ed.), *Prosodic typology – The phonology of intonation and phrasing*, 201–229. New York: Oxford University Press.

Kanerva, J. 1990. *Focus and phrasing in Chichewa phonology*. New York: Garland.

Kirby, James & Morgan Sonderegger. in press. Model selection and phonological argumentation. In Diane Brentari & Jackson Lee (eds.), *Shaping phonology*. Chicago: University of Chicago Press.

Ladd, D. Robert. 2008. *Intonational phonology*. New York: Cambridge University Press.

Ladefoged, Peter. 1997. Instrumental techniques for phonetic fieldwork. In William Hardcastle & John Laver (eds.), *The handbook of phonetic sciences*, 137–166. Malden, MA: Blackwell Publishing.

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Malden, MA: Blackwell Publishing.

Lee, Yong-cheol & Yi Xu. 2010. Phonetic realization of contrastive focus in Korean. *Proceedings of 5th International Conference on Speech Prosody, Chicago*.

Lehiste, Ilse. 1966. *Consonant quantity and phonological units in Estonian*. Bloomington: Indiana University Press.

Lehiste, Ilse, Pire Teras, Valts Ernštreits, Pärtel Lippus, Karl Pajusalu, Tuuli Tuisk & Tiit-Rein Viitso. 2008. *Livonian Prosody*. Helsinki: Suomalais-Ugrilainen Seura (*Mémoires de la Société Finno-Ougrienne* 255).

Lehiste, Ilse, Pire Teras, Toomas Help, Pärtel Lippus, Einar Meister, Karl Pajusalu & Tiit-Rein Vittso. 2005. *Meadow Mari Prosody* (*Linguistica Uralic Supplementary Series* 2) Tallinn: Teaduste Akadeemia Kirjastus.

Lesho, Marivic. 2013. The sociophonetics and phonology of the cavite chabacano vowel system The Ohio State University PhD dissertation.

Levi, Susannah V. 2005. Acoustic correlates of lexical accent in Turkish. *Journal of the International Phonetic Association* 35. 73–97.

Lieberman, Philip. 1960. Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America* 32. 451–454.

Maddieson, Ian. 2001. Phonetic fieldwork. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 211–229. New York: Cambridge.

Meyer, Roland & Ine Mleinek. 2006. How prosody signals force and focus: A study of pitch accents in Russian yes–no questions. *Journal of Pragmatics* 38. 1615–1635.

Moroz, George. http://CRAN.R-project.org/package=lingtypology 2017. *Lingtypology: Linguistic typology and mapping*.

Ortega-Llebaria, Marta. 2006. Phonetic cues to stress and accent in Spanish. *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology*, 104–118.

Pierrehumbert, Janet. 1980. The phonology and phonetics of English intonation. PhD Dissertation MIT.

Pierrehumbert, Janet and David Talkin. 1992. Lenition of /h/and glottal stop. In Gerard Docherty & D. Robert Ladd (eds.), *Papers in laboratory phonology II: Gesture, segment, prosody*, 90–117. Cambridge: Cambridge University Press.

Plag, Ingo, Gero Kunter & Mareile Schramm. 2011. Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics* 39. 362–374.

Pycha, Anne. 2006. A duration-based solution to the problem of stress realization in Turkish. UC Berkeley Phonology Lab Annual Reports.

R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riviera-Castillo, Yolanda & Lucy Pickering. 2004. Phonetic correlates of stress and tone in a mixed system. *Journal of Pidgin and Creole Languages* 19(2). 261–284.

Roettger, Timo B. accepted. *Tonal placement in Tashlhiyt Berber – How an intonation system accommodates to adverse phonological environments*. Studies in Laboratory Phonology. Berlin: Language Science Press.

Roettger, Timo B., Anna Bruggeman & Grice. Martine. 2015. Word stress in Tashlhiyt – Postlexical prominence in disguise. *Proceedings of the 18th International Congress of Phonetic Sciences*. Hong Kong.

Sadat-Tehrani, Nima. 2008. The structure of Persian intonation. *Proceedings of 4th International Conference on Speech Prosody, Campinas, Brazil*, 249–252.

Sadeghi, Vahid. 2011. Acoustic correlates of lexical stress in Persian. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, 1738–1741.

Silber-Varod, Vered, Hagit Sagi & Noam Amir. 2016. The acoustic correlates of lexical stress in Israeli Hebrew. *Journal of Phonetics* 56. 1–14.

Simard, Candide, Claudia Wegener, Albert Lee & Connor Youngberg. 2014. Savosavo word stress: A quantitative analysis. *Proceedings of 7th International Conference on Speech Prosody, Dublin, Ireland*.

Sluijter, Agaath M. C. & Vincent J. Van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100. 2471–2485.

Tuomainen, Jyrki, Stefan Werner, Jean Vroomen & Beatrice De Gelder. 1999. Fundamental frequency is an important acoustic cue to word boundaries in spoken Finnish. *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco*, 921–923.

Turk, Alice & Stefanie Shattuck-Hufnagel. 2007. Phrase-final lengthening in American English. *Journal of Phonetics* 35. 445–472.

Vasishth, Shravan & Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas – part I. *Language and Linguistics Compass* 10(8). 349–369. DOI: 10.1111/lnc3.12201. DOI:10.1111/lnc3.12201.

Vogel, Irene, Angeliki Athanasopoulou & Nadia Pinkus. 2016. Prominence, contrast, and the functional load hypothesis: An acoustic investigation. In Jeffrey Heinz, Rob Goedemans & Harry van der Hulst (eds.), *Dimensions of phonological stress*, 123–167. Cambridge: Cambridge University Press.

Wang, Bei, Ling Wang & Tursun Kadir. 2011. Prosodic encoding of focus in six languages in China. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf & Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 92. 1707–1717.

Winter, Bodo. 2011. Pseudoreplication in phonetic research. *Proceedings of the 17th International Congress of Phonetic Science, Hong Kong*. 2137—2140.

Xu, Yi. 2010. In defense of lab speech. *Journal of Phonetics* 38(3). 329–336.

Yakup, Mahire & Joan Sereno. 2016. Acoustic correlates of lexical stress in Uyghur. *Journal of the International Phonetic Association* 46. 61–77.

Mapping prosody onto meaning
– the case of information structure in American English[1]

Timo B. Roettger[1], Tim Mahrt[2] & Jennifer Cole[1]

[1]*Northwestern University*
[2]*Wovn Technologies, Inc.*

Prosody is a central part of human speech, with prosodic modulations of the signal expressing important communicative functions. Yet, the exact mechanisms of how listeners map prosodic aspects of the speech signal onto speaker-intended discourse functions are only poorly understood. Here we present three perception experiments that test the mapping between the prosodic form of a heard utterance and possible information structural categories (here: focus and givenness) determined by a discourse context. Results suggest varying degrees of accuracy dependent on the specific information structure categories that are presented to the listener in the experiment (the target and the competitor). Moreover, listeners are sometimes biased towards or against certain used discourse contexts. These biases are compatible with the idea that listeners infer speaker intentions based not only on bottom-up processing of acoustic cues but also on probabilistic knowledge about how likely prosodic forms co-occur with specific discourse contexts.

Keywords: prosody; intonation; focus; information structure, American English

1. Introduction

The prosodic form of a linguistic expression is an integral part of signalling meaning in human language. Prosody can not only encode emotions, speaker involvement, and attitude, it also plays a crucial role in expressing linguistic meaning: It conveys the intended illocutionary act, structures the utterance into smaller meaningful units, and allows the speaker to emphasize certain units while deemphasizing less important information. Given the importance of all of these dimensions of meaning for successful communication, our knowledge about how prosody guides listener's interpretation of utterance meaning is surprisingly small.

A central concern for a theory of prosodic meaning is how intonational form maps onto discourse functions. For example, information structure (the division of sentences into focus and background) and information status (the degree of activation of a referent in the current discourse

[1] TM and JC conceived the idea for this research and developed the experimental design. TM generated stimuli, carried out the experiment, and assembled the data. TR analyzed the data statistically and archived materials, data, and scripts in online repositories. TR and JC wrote the manuscript.

model) can be expressed by certain prosodic parameters. Some authors have proposed a direct mapping of acoustic parameters onto information structural categories (e.g. Cooper, Eady, & Mueller, 1985; Fry, 1955), others have proposed that phonological categories mediate acoustics and discourse functions (e.g. Pierrehumbert, 1980; Ladd, 2008). Regardless of its phonological interpretation, it has been argued that information structure and information status can be expressed through the assignment of phrasal prominence (i.e. positioning the word in a strong position in metrical structure, such as the head of the prosodic phrase) and the association of pitch accents (i.e. tonal events co-occurring with lexically stressed syllables) in English (e.g., Brown & Yule, 1983; Büring, 2006; Chafe, 1987; Ladd, 2008; Rooth, 1992; Selkirk, 1995). Different pitch accents have been described to express different types of discourse relations. For instance, a pitch accent with a late (and high) fundamental frequency (f_0) peak and a rising onglide (L+H* in the ToBI annotation) is described as signalling contrastive focus; A pitch accent with a medial peak and shallow rising onglide (H*) is described as signalling new information (cf. Pierrehumbert & Hirschberg, 1990, Watson, Tanenhaus, & Gunlogson, 2008).

A challenge for theories of prosodic meaning is seen in detailed empirical studies on several languages showing that implicitly assumed one-to-one-mappings between pitch contours and discourse function do not hold for all speakers of a language, or even for one speaker all of the time (German: Cangemi, Krüger, & Grice, 2015; Grice, Ritter, Niemann, & Roettger, 2017; English: Cruttenden, 1986; Peppé, Maxim, & Wells, 2000; Turnbull, 2017; Tashlhiyt: Roettger, 2017). For example, Grice et al. (2017) present evidence from a German speech production experiment. Prompted by discourse setting questions. Speakers had to produce utterances with different focus structures (Broad, Narrow, Contrastive, No focus). Some speakers produced different pitch accents for the same focus category and other speakers produced one and the same pitch accent for different focus categories. Similarly, Roettger (2017) shows that speakers of Tashlhiyt Berber can prosodically encode questions and contrastive statements with a rise-fall in pitch on the phrase-final word. This tonal event can either occur on the final or prefinal syllable. Both questions and contrastive statements can occur with either final or prefinal rise-falls. However, questions are probabilistically more likely to be produced with a final rise-fall (see also Grice, Ridouane, & Roettger, 2015; Roettger & Grice, 2015).

These studies suggest that there is no one-to-one-mapping between intonational events and speaker intentions; any assumed mapping is probabilistic at best (systematic but not deterministic). More recent work takes such variability into account and provides information as to the statistical distribution of alternative realisations of a given function (e.g. Yoon, 2010 for English; Grice et al., 2017, Baumann, 2006 and Baumann, Röhr, & Grice for German; Cangemi & Grice, 2016 for Italian).

Despite this large amount of variability, psycholinguistic work has shown that in some contexts listeners can rapidly anticipate speaker intentions based on intonational information even before disambiguating lexical material is heard (e.g., Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008; Roettger & Franke, 2018a,b; Roettger & Stoeber, 2017; Watson, Tanenhaus, & Gunlogson, 2008; Weber, Braun, & Crocker, 2006). These studies have demonstrated that listeners show anticipatory eye movements (or hand movements) when hearing an intonational event that allows them to predict an upcoming word based on its status as, e.g., new or given relative to the prior discourse context. This predictive behaviour is not only informed by bottom-up acoustic cues but also by dynamically adaptable probabilistic expectations about likely intonational contours in a given context (Kurumada et al., 2014; Roettger & Franke, 2018a,b).

These latter findings are in line with a rational analysis approach (Anderson, 1990) to speech perception (e.g. Clayards, Tanenhaus, Aslin, & Jacobs., 2008; Kleinschmidt & Jaeger, 2015; Kleinschmidt, Weatherholtz, & Jaeger, 2018; Norris, McQueen, & Cutler, 2003), assuming that speech perception can be thought of as a process of *inference under uncertainty*: listeners know that each linguistic unit is realised as a distribution of acoustic cues. The listener probabilistically infers how likely each possible linguistic unit is, taking into account their knowledge of these cue distributions within a given context. This inference process is informed by many different information sources, including information associated with the speaker and the discourse context. Prosodic processing as inference under uncertainty can account for successful perception of prosodic information despite its ubiquitous variability. It simultaneously allows for the integration of top-down information. This account contrasts with traditional models of perception of prosody that implicitly or explicitly assume a simple mapping of acoustic cues onto respective discourse functions.

Taking the systematic but probabilistic nature of mapping prosodic form onto discourse function into account, listeners should in principle have some ability to distinguish discourse functions based on only prosodic information, even in contextually impoverished contexts (e.g. in a controlled experiment). At the same time, listeners' performance should be poor when their task is devoid of communicative context and they are not able to adapt to a given situation, because expectations from prior discourse are impoverished or missing, decreasing the influence of top-down processing on perceiving prosody.

The present paper tests to what extent prosodic form-function relationships can be detected on the basis of prosodic cues. To that end, we test how well listeners detect and distinguish prosodic forms expressing different types of information structural relations: Givenness and Focus distinctions, which have been prominently discussed in the literature as important discourse functions expressed by prosody, most notably, in West Germanic languages (Büring, 2006; Ladd, 2008; Rooth, 1992; Selkirk, 1995). We define focus here according to an alternative semantics account as proposed by Rooth (1992). Focus is a semantic attribute of a word or phrase signalling that the proposition or parts of it have discourse-relevant alternatives. Focus can differ with respect to the location and scope of its domain.

Focus types can be marked by morphosyntactic devices such as word order or focus particles. Alternatively, in English and German, focus is often described as being signalled only by intonation with the position and type of pitch accents differentiating between focus type and scope. Acoustic correlates of focus and information status distinctions have been identified from experimental and corpus studies of English. In English, the nuclear prominence is located by default on the rightmost (content) word in the prosodic phrase (Chafe, 1987; Pierrehumbert, 1980; Selkirk, 1995). Nuclear prominence can be assigned to a word in an earlier position in the phrase if that word is focused and if the phrase-final word is lexically or referentially given. Speakers often distinguish a focus-marking prominence from a non-focus-marking prominence through scaling and alignment of the pitch contour (Breen, Fedorenko, Wagner, & Gibson, 2010). Such differences are analysed by some authors as differences in the tonal specification of the pitch accent, with high rising pitch accents (L+H* within the ToBI annotation) being the preferred pitch accent for focused words (Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980; Pierrehumbert & Hirschberg, 1990), while others consider the scaling and alignment differences as gradual in nature (Calhoun, 2006, 2012; Ladd & Schepman, 2003). Given that focus and information status distinctions are reflected in production in the form of measurable differences in acoustic parameters, there is a basis for experimental hypotheses that listeners use the same acoustic parameters as cues to recover focus and information status of words in comprehending speech.

There are several empirical studies that have investigated the perceptual detectability of prosodic focus marking: Gussenhoven (1983) asked listeners to determine whether the question and answer of a question-answer pair came from the same or a different conversation. He compared broad to narrow focus and reports that at least for certain structures there is a perceptible difference between narrow and broad focus, but listeners cannot use this information to reliably tell in which context the sentence was uttered, suggesting that listeners cannot easily associate focus types with respective acoustic forms.

In Welby (2003), English listeners rated a sentence like “I read the DISPATCH” with a pitch accent on ‘dispatch’ as similarly acceptable to questions with either narrow focus (i.e. “What newspaper do you read?”), or broad focus (i.e. “How do you keep up with the news?”), suggesting that listeners cannot easily tease focus types apart based on the acoustic form of the utterance only.

In Rump and Collier (1996), Dutch listeners judged which of four focus structures (neutral, double focus, focus on subject, focus on object) was most likely signalled by resynthesised intonation contours. Listeners were not consistent with respect to how they matched contour and focus structure and some pitch contours remained ambiguous with respect to focus. Other contours were more consistently classified as signalling a particular focus structure.

Breen et al. (2010) asked English listeners to match a recorded statement presented auditorily to a question that sets the discourse context for the statement. Their results indicate that listeners were generally accurate in identifying the focus position (subject focus, verb focus, object focus), but were often not able to differentiate different types of focus on the same constituent. In their experiment, listeners had to choose between seven different response options, making the task particularly difficult.

Cangemi, Krüger, and Grice (2015) asked German listeners to identify four different focus types. Stimuli were taken from a production corpus, in which five speakers produced utterances with different focus conditions (broad, narrow, contrastive, no focus) on the same sentential argument, where each focus condition was prompted by a preceding question. In the perception study, listeners heard these sentences and had to select in a four-alternative forced choice task, which among the four prompting questions provided an appropriate discourse context for the heard sentence. They report on categorisation accuracy above chance performance for all focus categories. Their experimental design, however, allowed for an exceptional high degree of accommodation to the stimuli: Speaker productions occurred in separate blocks, i.e. speakers were not interspersed with each other, giving listeners ample opportunity to 'tune' into speaker idiosyncrasies. Moreover, the speech material was segmentally very homogenous. Utterances only differed with respect to the quality of the stressed vowel of the target noun (Bieber, Bahber, Bohber), calling listeners' attention to prosodic differences expressed in that region. Nevertheless, this study provides evidence that German listeners can detect focus types based on prosodic form, at least in some conditions.

All in all, the literature on intonation-based focus perception is characterised by a wide variety of methodologies employed. Studies mainly differ in the type of task (acceptability judgements: Welby, 2003; naturalness judgement: Gussenhoven, 1983; or question-answer congruence: Breen et al., 2010; Cangemi et al., 2015; Rump & Collier, 1996). The latter studies utilising question-answer matching tasks differed also with respect to the number of response options (four response alternatives in Rump & Collier, 1996 and Cangemi et al., 2015 and seven in Breen et al., 2010). The results of these studies reveal an empirically mixed picture and its methodological diversity makes accumulation of evidence difficult. With the exception of Cangemi et al. (2015) on German, none of the above studies was able to clearly show that listeners can detect focus type based on prosodic information. For American English in particular, there is no compelling evidence to date that listeners perceive a difference between focus types such as broad, narrow, and contrastive focus. Whether listeners can use prosody to recognise speaker-intended focus structures remains, however, an important empirical question: Given the inherent probabilistic nature of mapping prosodic form onto communicative function, it is important to test if listeners can make use of prosodic cues to speaker intentions nevertheless and if so to what extent they use these cues.

The present study is an effort to reveal empirical evidence for the relationship between the prosodic signal and information structure as perceived by a listener. Experimental results are presented here to investigate the prosodic form-function mapping in perception by asking (i) how well listeners can identify the focus condition of an utterance based on its prosodic form (form-to-function mapping), and (ii) how well they can identify an appropriate prosodic form to match the focus condition specified by the discourse context (function-to-form mapping). Similar to several prior studies, the present study uses question-answer congruence, which provides a detailed view of the form-function mapping perceived by listeners. However, our studies differ from prior studies in reducing the complexity of the experimental task, towards the goal of minimizing task effects on the listeners' judgments of form-function association.

2. Methods

This paper presents a series of experiments exploring listeners' perception of the relationship between the prosodic form of an utterance and its focus conditions as established from the immediate discourse context.[2] Here we describe the methodology and statistical analysis employed. Section 2.1 presents the focus categories tested as they relate to theories of information structure and information status; Section 2.2 describes the experimental stimuli; Section 2.3 describes the design and procedures; Section 2.4 discusses the statistical methods we use to model the data.

2.1 Information structure categories

In the following experiments, listeners reacted to short question-answer dialogues in which the question provides the discourse context that establishes one of four information structure conditions of the answer: broad focus, narrow focus, and contrastive focus on the sentential subject, and the sentence subject as discourse-given. We adopt the question-answer congruence paradigm and operationalise focus following Büring (2012): in an answer, focus marks that constituent which can be construed as corresponding to a wh-phrase in a preceding question. Consider the following example.

(1) Damon fried the omelet.

a.	Do you know what happened yesterday?	[Damon fried the omelet]$_F$.
b.	Do you know who fried the omelet?	[Damon]$_F$ fried the omelet.
c.	Do you know what Damon fried?	Damon fried [the omelet]$_F$.
d.	Did Pam fry the omelet?	[Damon]$_F$ fried the omelet.
e.	Did Damon fry the omelet?	Damon fried the omelet.

The statement in (1) is a suitable morphosyntactic construction to answer all questions in (a-e), they only differ in their focus structure. Question (a) elicits whole-sentence focus (also referred to as 'broad focus'). A sentence has broad focus when it is uttered in an out-of-the-blue context, in the absence of a preceding discourse context, or with no particular correspondence to a preceding context. For a sentence with broad focus, the entire proposition expressed by the sentence is in focus and all constituents constitute new information. A common question used to elicit broad focus is "What happened?" or "What is new?" (e.g., question-answer pair 1a).

[2] The findings from these experiments are also discussed in Mahrt (2018), with qualitative comparisons across experimental conditions.

Questions (b-c) elicit ‘narrow focus’ either on the subject (b) or the object (c). A narrow focus sentence is one that contains a constituent that introduces relevant, new information to the discourse. The constituent with narrow focus may provide the answer to a wh-question, or it may highlight new information that is relevant to the discourse context, e.g., as an elaboration of information already given. In example (b) “Damon” contrasts with an open set of alternatives to the experiencer subject (all entities that could have eaten cheese), while in (c) “the omelet” contrasts with an open set of alternatives to the thematic object (all possible things Damon could have fried).

In (d), the focused constituent is explicitly contrasted with the alternative in the question (*Pam*), and constitutes a specific type of narrow focus, which is referred to as ‘contrastive focus’ (or ‘corrective focus’). Similar to a narrow focus sentence, a sentence with contrastive focus contains a constituent (here *Damon*) that relates specifically to an element of the preceding discourse (here *Pam*). Contrastive focus marks the referent of the constituent as singled out from a set of possible alternatives made salient by the discourse context (Rooth, 1992).

A sentence that cannot be construed as providing an answer to a wh-question or as specifying a contrastive referent may lack focus altogether. Such an example is illustrated in (1e), where all elements in the sentence are discourse-given, both lexically (the words are explicitly mentioned in the preceding question) and referentially (the referent of each word and phrase is established in the preceding discourse). We refer to such sentences as ‘given’ in what follows.

2.2 Stimuli

The stimuli used in the perception experiments were selected from productions of nine different English sentences (see 2). Each sentence was produced four times, once for each of the four focus categories described in the preceding section (Broad, Narrow, Contrastive, Given). The stimuli were recorded in a soundproof booth with a high-quality, head-mounted microphone. One informed female speaker of American English produced all of the stimuli (see osf.io/4qxmh/Stimuli). To make her productions as natural as possible, the sentences were produced in a live dialogue enacted with the experimenter who asked questions (see 1) that prompted the speaker to produce an appropriate full sentence response for each of the four focus conditions described in Section 2.1.

(2) (a) Daisy warned the owner.
(b) Damon fried the omelet.
(c) Dorah filmed the movie.
(d) Harry raised the window.
(e) Jamie dyed the laundry.
(f) Jonny helped the warden.
(g) Jonah burned the onion.

(h) Maddie found the TV.
(i) Mary rolled the barrel.

Except for the discourse-given context, the full sentence responses were not read aloud from text but were formulated by the speaker as appropriate full sentence responses to the experimenter's question (for a full list, see Appendix 1), inserting in subject position a name that was presented in written form for each sentence trial. The full sentence responses for the given condition (no focus) were written out and read aloud by the speaker. This was done to avoid the inadvertent production of a sentence with pronominal elements (e.g., "*Yes, she warned him*" for 2a), which are also acceptable responses to a polar question (e.g., "Did Daisy warn the owner?" for 2a). The recorded sentences were normalised for amplitude based on the peak amplitude of the entire recording session, using the normalise function in Audacity (Audacity Team, 2015) with the DC offset removed and peak amplitude normalised to -1.0 db. The result of this process resulted in 36 utterances (9 utterances for 4 different focus categories) of roughly equal amplitude.

Auditory inspection of stimuli by two native speakers (TM and JC) determined that the focus categories were produced with intonation patterns that sounded natural and congruent with the matched discourse prompt (i.e. the context question). Qualitative characterisation of f_0 contours, based on ToBI criteria, reveals that the answers exhibit expected intonational contours, with distinct contours for each of the four focus conditions. Figure 1 shows time-normalised f_0 contours and ToBI labels for each of the 9 sentences (grey) alongside the mean contour (in color), with productions grouped by the intended focus condition (i.e., focus conditions determined by the question prompt for each production).[3] The differences among the four focus categories can be seen in the f_0 contours on the subject and object positions. In the subject position, the broad and contrastive focus conditions show a noticeable rise-fall contour, the narrow focus condition has a shallower rise-fall (and in some tokens just a shallow fall), and the given condition exhibits a relatively low f_0 with an even shallower fall. In the object position, the given and narrow focus conditions show a flat or mildly falling f_0 excursion that extends with a nearly even slope across the interval of the object noun. The broad focus condition shows a noticeable rise-fall f_0 contour on the object, while the contrastive focus condition exhibits a low plateau that ends in a sharp fall (or in one instance, a rise) to the end of the utterance.

[3] The ToBI labels represent the one or two most frequent pitch accents produced on the subject and object nouns, over the nine sentence stimuli in each focus condition.

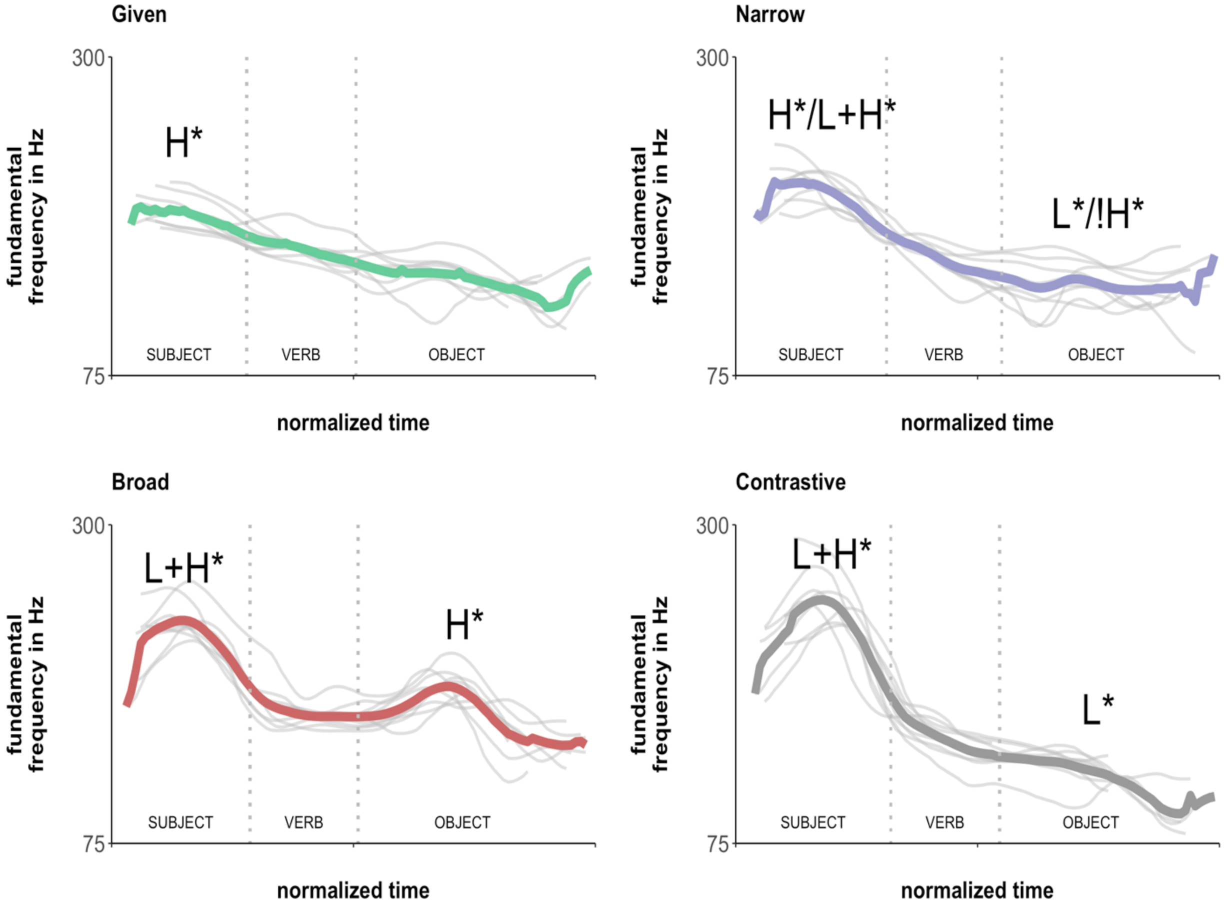

Figure 1: Smoothed and interpolated f_0 contours of the acoustic stimuli (grey) alongside the average f_0 contour (coloured) for all four focus conditions.

Figure 2 shows the raw values for the maximum f_0 values of the sentence subject and the sentence object. The f_0 max values for the subject overlap substantially between the broad and contrastive focus conditions, and between the narrow and given conditions. While the overlap between broad and contrastive is resolved when looking at f_0 maxima on the sentence object (clear separation), given and narrow remain highly overlapping. We will come back to these different degrees of overlap later.

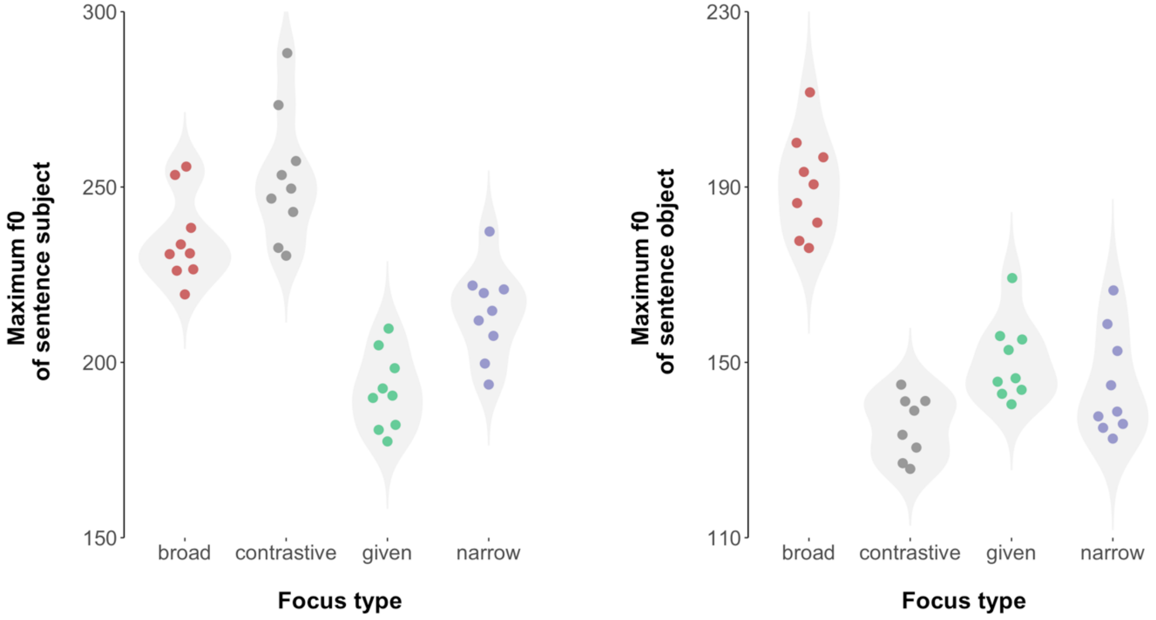

Figure 2: Raw f_0 maximum values of the sentence subject (left panel) and sentence object (right panel) for all target sentences. The grey shapes in the background indicate kernel density curves of these raw values in order to allow for a better visual assessment of overlap between categories.

To establish that the stimuli were acoustically differentiable into four prosodically distinct classes, we submitted the stimuli to linear discriminate analysis (LDA). We first inspected a variety of measures of pitch, intensity, and duration that were extracted from the subject, verb, and object positions of the sentence utterances. These were fed into an LDA analysis in the R *MASS* package (Venables & Ripley, 2002). The function indicated that there was collinearity among some of the measures. To determine which variables exhibited collinearity, the correlation of all possible feature pairs was taken and, for each correlated pair, only one of the features was chosen to be in the final set of acoustic features. With the final selection of acoustic features chosen, the LDA was run again. Using a leave-one-out (LOO) analysis, the LDA was able to discriminate the four focus categories with high accuracy (91.4%).

Note that we do not take the LDA results as proxy for human judgments of perceptual distinctiveness. Rather, the LDA analysis serves to independently verify that there is a basis for perceptual distinction in measurable acoustic distinctions that are sufficient for classification by statistical methods. It is important to note that the LDA analysis indicates some degree of overlap between the four focus categories, which suggests a degree of acoustic ambiguity in some of the productions for the acoustic measures. It is possible that using other acoustic measures, the distinction between the four focus categories might have been better captured, leading to higher accuracy. Our concern here is not primarily about how the four focus categories are differentiated from one another by the stimulus speaker, but rather to demonstrate that the four focus categories are acoustically differentiated.

2.3 Study design

The study was conducted to evaluate listeners' perception of four focus categories in relation to the prosodic form of an utterance. The perception test was operationalised through two tasks: one where listeners had to select which of two prosodic patterns best signalled a specified focus category (1 context - 2 prosodic forms, henceforth 1C-2P); and the other where listeners had to select which of two focus categories was signalled by the prosody of an utterance (2 contexts - 1 prosodic form, henceforth 2C-1P). Rather than presenting participants with all four focus categories in a single (lengthy) experiment, a between-subjects design was chosen, exposing individual participants to only one pair of focus conditions (broad focus vs. given, broad vs. contrastive focus, broad vs. narrow focus, given vs. contrastive focus, given vs. narrow focus, and contrastive vs. narrow focus). A further distinction in the response option given to the participant was introduced: One group of participants used a two-alternative forced-choice response, while another group of participants used a 5-point scale response. In what follows, these experimental conditions are grouped into 3 experiments, Experiments 1 and 2 use the two-alternative forced choice response option, and Experiment 3 uses the 5-point scale response option.

2.3.1 Tasks

In each experiment, auditory stimuli in the form of mini question-answer dialogues were presented to participants. There were two such dialogues on each trial, which differed in the prosodic congruence of the question and answer. Participants were presented with two play buttons on opposite sides of the screen, one for each mini dialogue (Q-A pairing). Participants were allowed to listen to the audio files as many times as they wanted before responding. Participants proceeded through the experiment at a self-selected pace. In one dialogue, the question-answer pair was matched in their focus condition (i.e., the answer was produced by the model speaker as a response to the question appearing in the dialogue), while in the other dialogue the question-answer pair was mismatched (the answer was produced by the model speaker in response to a different question than the one appearing in the dialogue). Participants were instructed to either choose the dialogue that sounded the most appropriate or natural (Experiment 1 and 2) or to use a 5-point scale to indicate which of the two dialogues they prefer.

The 1C-2P task tests the mapping from discourse function to prosodic form. In this task, the two dialogues in a trial had the same question, but the question was paired with answers that were prosodically distinct, i.e., from two different categories shown in Fig. 1. This task examined whether listeners could identify a preferred acoustic prosodic signal for the particular focus condition specified by the discourse context. Note that the proposition of the answers was always the same for both dialogues and was textually appropriate as a response to the prompting questions. An example of the 1C-2P task is shown in (3), contrasting broad and narrow focus conditions. If narrow focus is prosodically encoded and perceptually detectable by the listener, the dialogue in (3a) with narrow focus prosody (as in Fig. 1 above) should sound more natural than the dialogue in (3b) with broad focus prosody (as in Fig. 1 above).

(3) Dialogue pair from the 1C-2P task

a. *Incongruous*
Q: Do you know who ripped the ledger? [Narrow focus prompt]
A: Yes, [Mary ripped the ledger]$_F$. [Broad focus prosody]

b. *Congruous*
Q: Do you know who ripped the ledger? [Narrow focus prompt]
A: Yes, [Mary] $_F$ ripped the ledger. [Narrow focus prosody]

The 2C-1P task tests the mapping from prosodic form to discourse function. In this task, the two dialogues had textually different questions, each of which set up different focus conditions for the answer, while the answers were the same in both dialogues (i.e., the same audio file). This task examined whether listeners could identify the discourse context that matched the focus condition of the answer, perceived on the basis of its prosodic form. An example of the 2C-1P task is shown in (4), contrasting broad and narrow focus.

(4) Dialogue pair from the 2C-1P task

a. *Incongruous*
Q: Do you know what happened yesterday? [Broad focus prompt]
A: Yes, [Mary]$_F$ ripped the ledger. [Narrow focus prosody]

b. *Congruous*
Q: Do you know who ripped the ledger? [Narrow focus prompt]
A: Yes, [Mary] $_F$ ripped the ledger. [Narrow focus prosody]

The two experimental tasks (1C-2P, 2C-1P) were designed to explore possible sources of ambiguity stemming from overlap in the range of acoustic patterns that a speaker may produce in a certain focus condition, and which a listener may judge as acceptable acoustic cues for perceiving distinctions in focus-related meaning. If listeners can successfully detect prosody-focus mappings, then participants in the 2C-1P and 1C-2P tasks should be equally accurate in identifying the most natural sounding dialogue in each trial. If participants perform poorly in the 1C-2P task (choosing from two prosodically distinct answers), that would suggest that a range of acoustic cues can signal the same meaning. If participants perform poorly in the 2C-1P task (choosing from two textually distinct context questions), it would suggest that a certain acoustically specified prosodic pattern may be congruent with multiple focus-related meanings (as specified by the discourse context).

The same dialogues with the same acoustic stimuli were used for all three experiments. In experiment 1, participants performed only the 1C-2P task in a forced choice design. In experiment 2, participants performed only the 2C-1P task in a forced choice design. Experiment 3 comprised both 1C-2P and 2C-1P tasks but with more nuanced response options on a 5-point Likert scale:

- *only left*: only the dialogue on the left side of the screen sounded natural
- *left preferred:* both dialogs sound natural, but the left dialogue is preferred
- *equally good:* both dialogs sound equally natural and acceptable
- *right preferred*: both dialogs sound natural, but the right dialogue is preferred
- *only right*: only the dialogue on the left side of the screen sounded natural

Only in Experiment 3, 1C-2P and 2C-1P trials were presented in different trials to the same participants.[4] Participants in Experiment 3 were instructed that sometimes the answers would vary in the way they were said and sometimes the questions would vary. None of the experiments gave participants any explicit instructions or training regarding the information structure or prosody of the dialogs they would hear.

Experiment 1 and 2 consisted of 18 trials presenting a pair of Q-A dialogues for one of the six different focus condition pairs (broad-given, broad-contrastive, broad-narrow, given-contrastive, given-narrow, contrastive-narrow). Each of the 9 stimulus sentences in (2) was presented as the answer in two trials that differed in which of the two focus categories was specified in the congruent dialogue. For example, the dialogue pairs in (3) and (4) are taken from the group testing broad vs. narrow focus prosody. In the trials shown in (3) and (4), it is the (b) dialogues that are congruent—the focus condition prompted by the question matches the focus condition that is expressed by the prosody of the answer. These same experiments (testing the broad vs. narrow focus categories) included another trial with "Mary ripped the ledger" in which the congruent dialogue matches the question and answer in the broad focus condition. The order of the stimuli was pseudorandomised, i.e. shuffled by hand such that no two consecutive items contained the same lexical content (i.e., the same answer sentence).

Experiment 3 consisted of 36 trials, presenting a pair of Q-A dialogues for one of six different focus condition pairs for both 2C-1P and 1C-2P tasks.

[4] Our decision to administer both tasks (1C-2P, 2C-1P) to the same participants in Experiment 3, rather than run two separate experiments with the expanded, scalar response set, were driven by practical constraints of time and money. We were also interested in testing the feasibility of combining both tasks in one experiment, to approximate the design of earlier experiments testing categorical perception of phoneme contrasts, in which identification and discrimination experiments are administered to the same participants. Mahrt (2018: 30) reports on a pilot experiment using both tasks, involving 45 subjects recruited from Mechanical Turk. In the pilot experiment, a third of the participants first did the identification task followed by the discrimination task, another third did the tasks in the opposite order, and the final third did them with the tasks interleaved in random sequence over trials. The same set of stimuli were presented to all participants. The results of the pilot showed that the task done second resulted in higher accuracy than when it was performed first, with intermediate accuracy for participants in the interleaved tasks condition. On the basis of these findings, Experiment 3 adopted the interleaved task design.

2.3.2 Participants

Participants were recruited online via the crowd-sourcing website Amazon Mechanical Turk (AMT). Participants were restricted to people from the United States via filtering of IP address by the AMT system, and were restricted to be at least 18 years old. The participants came from all around the U.S. Participants who reported themselves as being non-native speakers or indicated that they were born and grew up outside of the U.S. were excluded from the study. To prevent incentivising dishonesty, they were not told that they had to be a native speaker of American English to participate and they were still compensated for their time, regardless of whether or not their data was used. Data that was excluded due to these circumstances was replenished by running additional participants. Data analysis was not initiated before the complete data set was available. Participants were only allowed to participate in one of the three experiments. These constraints were managed by LMEDS (Mahrt, 2016), the web platform used to run all of the experiments.

Experiments 1, 2, and 3 used data from 180 participants each, for a total of 540 participants. Participants in each Experiment (1-3) were randomly assigned to one of six groups testing different pairs of focus conditions. Experiments 1 and 2 took about 15 minutes to complete, while Experiment 3 took about 25 minutes. Participants were compensated at a rate of $10/hour.

2.4 Statistical analysis

We submitted participants' responses to Bayesian hierarchical models using the Stan modelling language (Carpenter et al., 2016) and the `R` (R Core Team, 2018) package `brms` (Bürkner, 2016). We operate within the Bayesian inferential framework (rather than within a frequentist framework) due to two reasons:

First, Bayesian methods allow us to directly answer the primary question: *How plausible is our hypothesis given the data?* We can answer this question by quantifying our uncertainty about the parameters of interest, which frees us from committing to hard cut-off points for statistical significance (such as the arbitrary 0.05 alpha level).

Second, it is easier to flexibly define hierarchical models (also known as mixed effects or multilevel models) in the Bayesian framework than in the frequentist framework. The frequentist linear mixed model standardly used in quantitative linguistics is generally fit with the lme4 package (Bates et al., 2015b) in R. However, the linear mixed effects models for categorical data that also include the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013; Schielzeth & Forstmeier, 2009) tend not to converge or to give unrealistic estimates of the correlations between random effects (Bates, Kliegl, Vasishth, & Baayen, 2015). Such non-convergence issues are particularly severe for logistic regression models (Kimball et al., 2016). In contrast, the maximal random effects structure can be fit without problems using Bayesian hierarchical models.

We used different statistical models for Experiments 1 and 2 than for Experiment 3. For Experiments 1 and 2 we fit a hierarchical logistic regression model to response accuracy (binomial: correct vs. incorrect) predicted by the target focus category in the congruent dialogue (4 levels: Given, Broad, Narrow, Contrastive), the competitor focus category (3 levels, e.g. for the target category Broad, the competitor focus would be Narrow, Contrastive, or Given) and their two-way interaction. The models included a maximal random-effect structure, including a random intercept for subjects (since it is a between-subject design), and a random slope allowing the predictor interaction to vary by experimental items (the 9 sentences comprising the experimental stimuli).

We used weakly informative Gaussian priors centred around zero with σ = 5 for the Intercept and with σ = 5 for all population-level regression coefficients. Four sampling chains with 2000 iterations each were run for each model, with a warm-up period of 1000 iterations. We report, for each parameter of interest, 95% credible intervals and the posterior probability that a coefficient parameter β is bigger than zero $Pr(\beta > 0)$. A 95% credible interval demarcates the range of values that comprise 95% of the probability mass of our posterior beliefs, such that no value inside the CI has a higher probability than any point outside of it (see, e.g., Jaynes & Kempthorne, 1976; Morey et al., 2016). We judge there to be substantial evidence for an effect if zero is (by a reasonably clear margin) not included in the 95% CI and Pr($\beta > 0$) is close to zero or one.

For Experiment 3, where the same participants performed both tasks (1C-2P and 2C-1P) we ran two subset analyses on the data, one that models the 1C-2P trials and one that models the 2C-1P trials. Recall that Experiment 3 used an elaborated set of five response options. For both tasks, we fitted Bayesian hierarchical ordinal logistic models to the ordered response options predicted by the target focus category of the congruent dialogue (4 levels, as for Exps. 1 and 2), the competitor focus category (3 levels, as for Experiments 1 and 2) and their two-way interaction. The five responses were re-labelled as follows: If the congruent question-answer pair was on the right side of the screen, we binned "always right" as "always", "right preferred" as "preferred", "equally good" as "equal", "left preferred" as "dispreferred", and "always left" as "never". Responses were similarly re-labelled for trials in which the congruent question-answer pair was on the left side of the screen, by swapping "left" for "right" in the re-labelling scheme. The re-labelled responses were rank ordered: never > dispreferred > equal > preferred > always. The models for Experiment 3 included a random intercept for subjects (since main effects of focus conditions were tested in a between-subject design), and a random slope allowing the predictor interaction to vary by experimental items. We used weakly informative student-t priors centred around zero with σ = 1 and dfs = 5 for all population-level regression coefficients. The inferential criteria are the same as discussed for Experiments 1 and 2.

Posterior probabilities tell us the probability that the parameter has a certain value (given the data and model); note that these probabilities are not frequentist *p*-values. Note also that there is no notion of Type I or II error in Bayesian statistics because the inference does not depend on hypothetical repetitions of the experiment; the data are evaluated on their own merits, and no supposition is made about the replicability of the effect. In order to present statistics as close to widely used frequentist practices, we chose to define an inferential criterion that seems familiar (95%), but the strength of evidence should not be taken as having clear cut-off points (such as in a null-hypothesis significance testing framework). In line with standards of reproducible research, the data tables and the scripts for the statistical analyses are made available and can be retrieved here osf.io/4qxmh.

3. Results for experiment 1: One context – two prosodic forms

Figure 3 and Table 1 summarise the posterior distribution across conditions for experiment 1. Instead of interpreting regression coefficients, we directly calculate the posterior distribution and accompanying credible intervals for each condition (given the data and the model). We can further directly calculate the probability of respective accuracy estimates being above chance (log odds > 0).

Looking at the estimates, overall, listeners performed well in the task. However, there are obvious interactions between target (henceforth X^T) and competitor categories (henceforth X^C), with varying accuracy estimates for different combinations of categories. Dependent on the competitor category for a trial, listener performance differs tremendously: Except for Given^T competing with Narrow^C, and Narrow^T competing with Contrastive^C, all conditions show evidence for above chance accuracy. Listeners thus seem to be able to infer the intended prosodic information in the signal based on the discourse setting question.

Listeners' performance differed, however, as a function of which categories were compared. For Broad^T (upper left panel), listeners exhibit higher accuracies when the competitor is Contrastive^C (β = 0.90 [0.83,0.95]) or Narrow^C (β = 0.86 [0.77,0.92]) than when the competitor is Given^C (β = 0.67 [0.55,0.77)]; For Contrastive^T (upper right), listeners exhibit higher accuracies when the competitor is Broad^C (β = 0.92 [0.86,0.96]) or Given (β = 0.94 [0.83,0.98]) than when it is Narrow^C (β = 0.66 [0.47,0.82]); For Given^T (lower left), listeners exhibit higher accuracies when the competitor is Contrastive^C (β = 0.84 [0.72,0.91) than when it is Narrow^C (β = 0.56 [0.41,0.70]); For Narrow^T (lower right), listeners exhibit higher accuracies when the competitor is Broad^C (β = 0.85 [0.76,0.92]) than when it is Contrastive^C (β = 0.46 [0.30,0.62]).

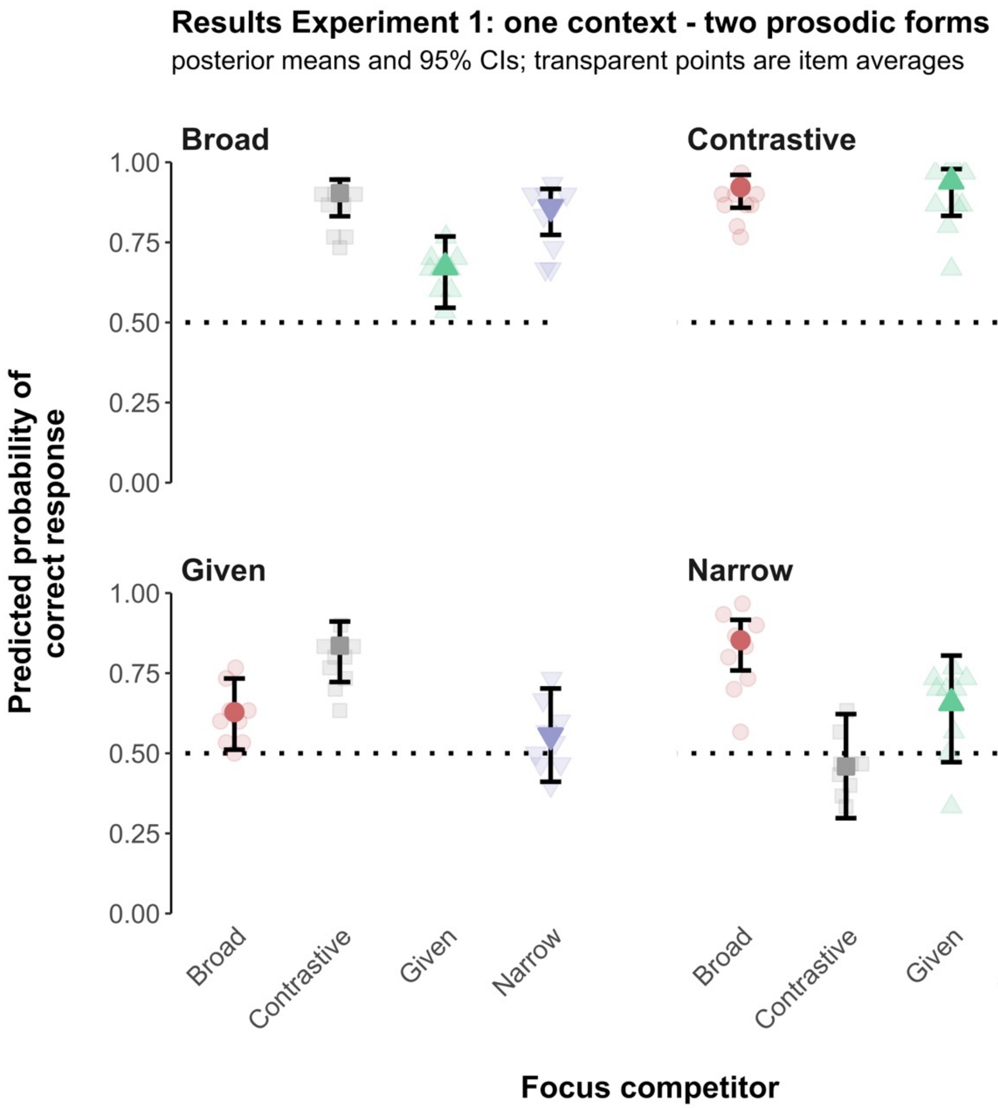

Figure 3: Mean posteriors and 95% credible intervals for the results of Experiment 1, showing predicted accuracy across target focus conditions (in the four panels), and their accompanying focus competitors (x-axis). Semi-transparent small points are average values for each experimental item (sentence). The dotted line indicates chance performance.

target	**competitor**	**estimate**	***P(β* > 0)**
Broad	Contrastive	0.9 (0.83,0.95)	1.00
Broad	Given	0.67 (0.55,0.77)	1.00
Broad	Narrow	0.86 (0.77,0.92)	1.00
Contrastive	Broad	0.92 (0.86,0.96)	1.00
Contrastive	Given	0.94 (0.83,0.98)	1.00
Contrastive	Narrow	0.66 (0.47,0.82)	0.96
Given	Broad	0.63 (0.51,0.73)	0.98
Given	Contrastive	0.84 (0.72,0.91)	1.00
Given	Narrow	0.56 (0.41,0.7)	0.77
Narrow	Broad	0.85 (0.76,0.92)	1.00
Narrow	Contrastive	0.46 (0.3,0.62)	0.31
Narrow	Given	0.66 (0.47,0.8)	0.96

Table 1: Summary of posterior distributions for Experiment 1: Posterior means (95% credible intervals in brackets) for all focus combinations alongside the probability that the estimate is above chance level (log odds > 0) given the data and the model. Shaded rows indicate conditions that did not perform above chance according to set inferential criteria.

The results of this experiment suggest that, in general, listeners can use the prosodic cues available in the signal to distinguish between focus types. Some categories are perceived better than others and accuracy is very much dependent on the competing category. Accuracy was highest for the pairs Contrastive and Given as well as Contrastive and Broad. This is not surprising considering the very distinct f_0 patterns in the stimuli (see Fig. 1). Given referents were produced with a high pitch accent (H*), the most frequently occurring pitch accent type, and the smallest f_0 excursion on the subject, while contrastive referents were produced with a high rising pitch accent (L+H*), arguably the most prominent pitch accent type, and the greatest magnitude f_0 excursion on the subject. Broad focus utterances exhibited two prominent pitch accents on the subject and the object, a noticeably distinct utterance-wide pattern.

The other focus pairs were not as well distinguished, including contrastive and narrow, and given and narrow. The observation that the accuracy between narrow and every other category is low might be attributed to the acoustic form of narrow focus utterances. Narrow focus stimuli exhibit an intonational form that greatly overlaps with the other categories. For instance, in Narrow focus stimuli, the subject exhibits a rise-fall contour that is variously labelled as H* or L+H*, but the difference between that and the f_0 contour of the subject in Contrastive focus stimuli, all of which are labelled as L+H* can be characterised as a difference in pitch scaling. Likewise, for the Narrow focus stimuli, the f_0 excursion of the less prominent H* of the subject appears to partially overlap with the f_0 excursion of the Broad focus subject in some instances, and with that of the L* pitch accent of stimuli in the Given category.

In sum, some focus categories elicit lower accuracies while others elicit higher accuracies. These differences may be a reflection of different degrees of acoustic overlap. However, overall, listeners seem to be able to match the intended focus type of an utterance to its respective discourse setting question above chance level. The 1C-2P task taps into the question of which acoustic form best conveys the focus condition selected by a particular discourse context, while the 2C-1P task taps into which discourse context best matches the focus condition conveyed by a particular acoustic form

4. Results for experiment 2: Two contexts – one prosodic form

Figure 4 and Table 2 summarise the posterior distribution across conditions for experiment 2. Looking at the estimates, the 2C-1P results differ from the results of experiment 1. Overall, listeners' accuracy is not as high as in the 1C-2P task. This effect is mainly driven by two factor levels: Broad^T and Narrow^C. When Broad focus is the target, listeners were systematically *below* chance, i.e. identifying the utterance as indicating the competitor focus category (Contrastive^C: $\beta = 0.22$ [0.14,0.32]; Given^C: $\beta = 0.30$ [0.22,0.41]; Narrow^C: $\beta = 0.17$ [0.12,0.23]). Beyond showing poor performance in identifying Broad^T, listeners consistently picked the wrong response alternative, suggesting a *bias against* Broad focus. Similarly, when Narrow is the competitor, listeners were systematically below chance, i.e. incorrectly identifying the utterance as indicating Narrow focus (Broad^T: $\beta = 0.17$ [0.12,0.23]; Contrastive^T: $\beta = 0.36$ [0.25,0.48]; Given^T: $\beta = 0.35$ [0.21,0.51]). Again, beyond having difficulty in identifying the target category, listeners were zealous in consistently identifying utterances as Narrow, suggesting a *bias towards* Narrow focus.

In addition to these two biases, listeners had difficulties identifying Given^T and Narrow^T when paired with Contrastive^C, although, in both cases, there is weak evidence that listeners perform above chance (Given^T: $\beta = 0.66$ [0.48,0.84]; Narrow^T: $\beta = 0.65$ [0.48,0.80]).

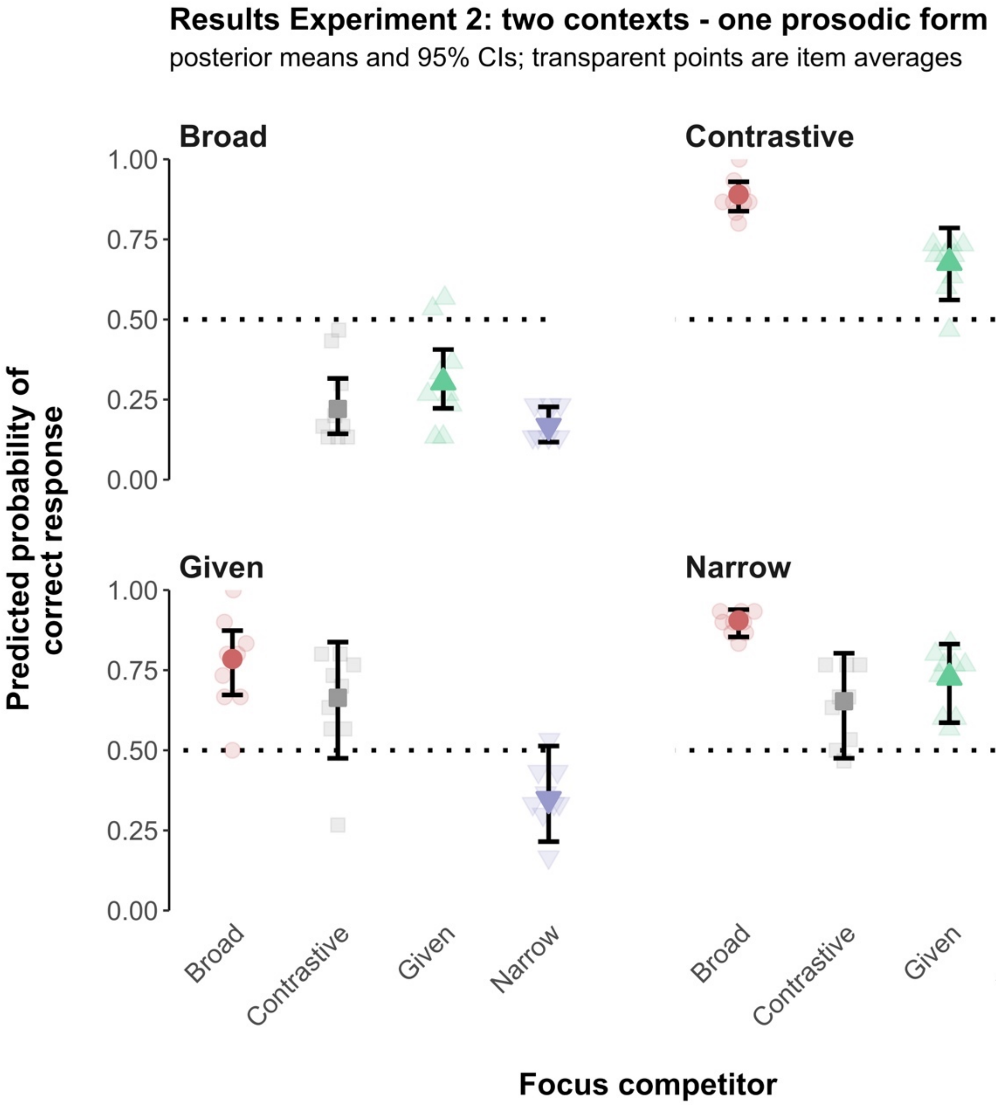

Figure 4: Mean posteriors and 95% credible intervals for the results of Experiment 2, showing predicted accuracy across target focus conditions (in the four panels), and their accompanying focus competitors (x-axis). Semi-transparent small points are descriptive average values for each experimental item (sentence). The dotted line indicates chance performance.

target	**competitor**	**estimate**	$Pr(\beta > 0)$
Broad	Contrastive	0.22 (0.14,0.32)	0.00
Broad	Given	0.3 (0.22,0.41)	0.00
Broad	Narrow	0.17 (0.12,0.23)	0.00
Contrastive	Broad	0.89 (0.84,0.93)	1.00
Contrastive	Given	0.68 (0.56,0.79)	1.00
Contrastive	Narrow	0.36 (0.25,0.48)	0.01
Given	Broad	0.78 (0.67,0.87)	1.00
Given	Contrastive	0.66 (0.48,0.84)	0.95
Given	Narrow	0.35 (0.21,0.51)	0.03
Narrow	Broad	0.91 (0.85,0.94)	1.00
Narrow	Contrastive	0.65 (0.48,0.8)	0.96
Narrow	Given	0.73 (0.59,0.83)	1.00

Table 2: Summary of posterior distribution for Experiment 2: Posterior means (95% credible intervals in brackets) for all focus combinations alongside the probability that the estimate is above chance level (log odds > 0) given the data and the model. Shaded rows indicate conditions that did not perform above chance according to set inferential criteria.

Our results indicate that when having to identify a discourse context on the basis of the focus condition conveyed by the prosodic form, listeners have substantial difficulties. The observed biases against pairing the Broad^T focus prosody with its matched discourse context and in favour of pairing any prosodic form with the Narrow^C focus discourse context suggests that listeners are influenced by aspects of the context other than the prosodic information in the signal. (Note that listeners were clearly able to use acoustic prosodic cues in the 1C-2P task with the same stimuli). The acoustic prosodic expression used in the Narrow focus stimuli in this study are apparently congruent with a variety of information structure contexts, and similarly, any type of prosodic form is judged as congruent with a Broad focus context. As opposed to that, for Given and Contrastive focus contexts, listeners showed a preference for one prosodic form—the congruent one in this experiment.

The results of experiment 1 and 2 suggest both overlap and differentiation in the association of prosodic form and focus condition. Differentiation is seen in the finding that listeners show above-chance accuracy in associating prosodic forms with the focus conditions intended by the speaker, for at least some of the distinctions in focus conditions. The associations between form and meaning are far from being one-to-one, and there appear to be ambiguities in both directions of the form-function mapping. This pattern of results may stem from ambiguity in the prosodic encoding of focus that leaves listeners uncertain about the intended focus condition. An alternative account of the results involves listener bias, as suggested in the findings from the 2C-1P task, where given a choice in meaning listeners lean towards or away from inferring certain focus-related meanings.

Experiment 3 seeks to further explore the ambiguity (or bias) in the mapping between prosodic form and focus-related meaning, by offering participants five response options that differ in the strength of association for each of the two form-function mappings presented in each trial.

5. Experiment 3 – Scalar endorsement ratings

The data from Experiment 3 differed from that of Experiments 1 and 2 with respect to available response options. Experiment 3 also differed in presenting participants with 36 trials, 18 per task (9 sentences in two different congruent pairings, as in Experiments 1 and 2). Thus, participants in Experiment 3 produced 18 responses in the same 1C-2P task as those in Experiment 1, and they produced 18 responses in the same 2C-1P task as those in Experiment 2. The data from each task in Experiment 3 was modelled in separate subset analyses, as described in Section 2.5.

5.1 Results for experiment 3: One context – two prosodic forms

Figure 5 and Appendix 2 summarise the posterior distributions across conditions for the 1C-2P task. Overall, participants tend to select the responses "equal", "preferred" and "always" above chance (= 0.2), suggesting that listeners have a general tendency to rate the match between prosodic pattern and focus condition as acceptable, even when the match is incongruent. This is illustrated by the asymmetry of stacked bar plots in Figure 5, which show a greater probability mass in the green bars ("preferred", "always") compared to the red bars ("dispreferred", "never"). (If there was no bias towards either the negative or positive end of the response scale, the stacked bar plots would be symmetrically centred around the horizontal line.)

These general patterns are in line with the results from Experiment 1. Listeners can use the prosodic information in the signal to discriminate intended focus categories above chance levels. However, there is a great amount of variability in how listeners match prosody and focus conditions. Listeners generously endorse utterances as belonging to focus categories other than the one intended by the speaker, indicated here by the large amount of "equal" ratings (both dialogues in the trial rated as equally acceptable). Beyond these general patterns, and in line with our earlier findings, there are also clear differences among responses for different pairings of target focus and competitor focus category.

For Broad^{T}, there is evidence that listeners are more likely to endorse a broad focus prosody correctly paired to a broad focus discourse context when the competitor pairs Contrastive^{C} prosody with the broad focus context than when the competitor pairing has Given^{C} prosody. This asymmetry is seen in the comparison of "equal" and "preferred" responses for Broad^{T} when paired with Contrastive^{C} ("equal": $\beta = 0.29$ [0.21,0.37]; "preferred": $\beta = 0.58$ [0.52,0.64]) compared to when paired with Given^{C} ("equal": $\beta = 0.47$ [0.39,0.54]; "preferred": $\beta = 0.4$ [0.3,0.5]) or Narrow^{C} ("equal": $\beta = 0.45$ [0.37,0.53]; "preferred": $\beta = 0.43$ [0.33,0.53]).

This asymmetry in response pattern suggests, again, that listeners find it easier to correctly endorse a Broad^T focus prosody when Contrastive^C focus is the competitor, than with other competitor categories.

In line with that, there is evidence that when paired with Broad^C, Contrastive^T elicits fewer "equal" and more "preferred" responses ("equal": β = 0.38 [0.28,0.48]; "preferred": β = 0.5 [0.41,0.6]) than when Contrastive^T is paired with Narrow^C ("equal": β = 0.54 [0.48,0.58]; "preferred": β = 0.26 [0.13,0.39]). Again, Broad and Contrastive focus categories elicit the strongest endorsements.

For Given^T, there is substantial evidence that Contrastive^C elicits fewer "equal" and more "preferred" and always responses ("equal": β = 0.19 [0.12,0.27]; "preferred": β = 0.63 [0.59,0.66]; "always": β = 0.16 [0.1,0.24]) than Broad^C ("equal": β = 0.54 [0.5,0.58]; "preferred": β = 0.28 [0.2,0.36]; "always": β = 0.02 [0.01,0.03]) and Narrow^C ("equal": β = 0.52 [0.46,0.57]; "preferred": β = 0.32 [0.22,0.42] ; "always": β = 0.16 [0.1,0.24]), suggesting that listeners are most likely to endorse a Given^T prosody as correctly paired to its discourse context when the competitor pairing has Contrastive^C prosody.

Interestingly, Narrow^T did not elicit different responses across competitor categories. All three conditions seem to behave similarly and exhibit predominantly "equal" responses. This response pattern indicates that listeners endorse all prosodic patterns conditions as equally acceptable in pairings with the Narrow focus discourse context.

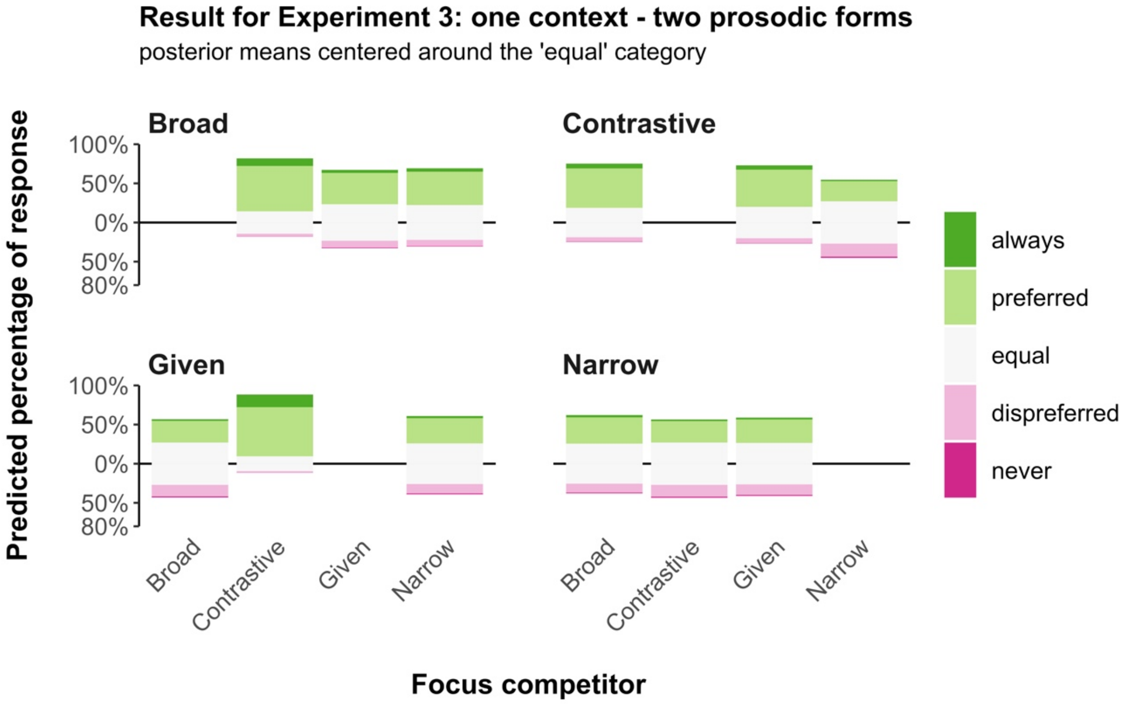

Figure 5: Stacked bar plots for the predicted probability of choosing one response over the others across target focus conditions and their accompanying focus competitors. Stacked bar plots are centred around the middle category ("equal") indicated by the solid horizontal line. Visual mass

above the line indicates tendency to prefer the match between prosody and focus condition, mass below the line indicates tendency to not prefer the match.

5.2 Results for experiment 3: Two contexts – one prosodic form

Figure 6 and Appendix 3 summarise the posterior distribution across conditions for the 2C-1P task. As opposed to the 1C-2P task, listeners do not show an overall tendency to endorse the stimuli pairings. There are generally stronger differences between focus conditions, with some eliciting responses predominantly on the negative end of the scale and others eliciting responses predominantly on the positive end of the scale. The generally weaker performance of listeners in this task compared to the 1C-2P task is in line with the results from Experiment 2.

For Broad^{T}, there is some evidence that Narrow^{C} elicits more "dispreferred" and "equal" and less "preferred" ratings ("dispreferred": β = 0.46 [0.39,0.53]; "equal": β = 0.33 [0.27,0.39]; "preferred": β = 0.11 [0.07,0.16]) than Contrastive^{C} ("dispreferred": β = 0.24 [0.16,0.32]; "equal": β = 0.41 [0.39,0.44]; "preferred": β = 0.29 [0.21,0.38]) and Given^{C} ("dispreferred": β = 0.21 [0.15,0.27]; "equal": β = 0.41 [0.38,0.43]; "preferred": β = 0.32 [0.24,0.4]). A general bias against Broad^{T} cannot be observed here (remember, in the forced choice 2C-1P task, Broad^{T} was systematically avoided as a possible response, whether congruent or incongruent on the trial).

The general bias in favor of the Narrow^{C} competitor remains apparent in experiment 3. When Narrow^{C} is available as a response option, listeners tend to prefer it over Broad^{T}. The bias towards Narrow responses is also seen for Contrastive^{T}. When paired with Narrow^{C}, listeners selected more "dispreferred" and "equal" responses as well as less "preferred" responses ("dispreferred": β = 0.36 [0.26,0.46]; "equal": β = 0.39 [0.34,0.43]; "preferred": β = 0.18 [0.11,0.25]) than when Contrastive^{T} was paired with Given^{C} ("dispreferred": β = 0.07 [0.04,0.11]; "equal": β = 0.25 [0.17,0.33]; "preferred": β = 0.55 [0.49,0.61]) and Broad^{C} ("dispreferred": β = 0.06 [0.04,0.09]; "equal": β = 0.23 [0.17,0.3]; "preferred": β = 0.56 [0.51,0.6]).

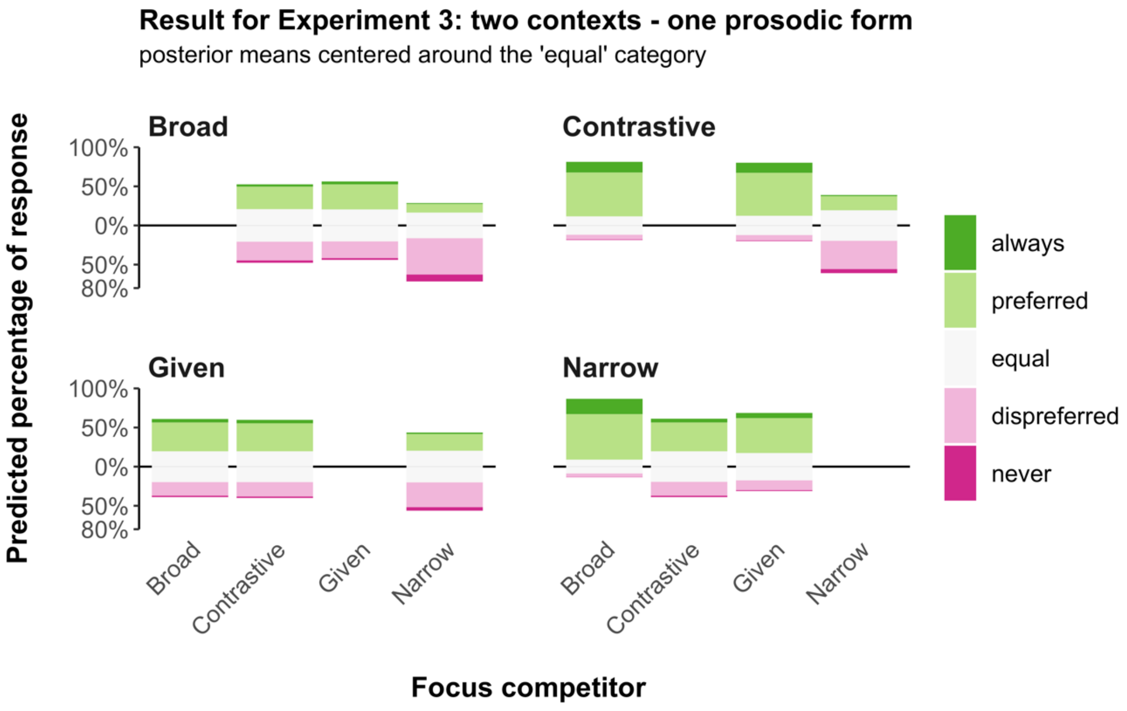

Figure 6: Stacked bar plots for the predicted probability of choosing one response over the others across target focus conditions and their accompanying focus competitors. Stacked bar plots are centred around the middle category ("equal") indicated by the solid horizontal line. Visual mass above the line indicates tendency to prefer the match between prosody and focus condition, mass below the line indicates the tendency to not prefer the match.

Overall, Experiment 3 confirms the results from Experiments 1 and 2. Listeners can match different prosodic realisations to their speaker-intended focus categories, but listeners' performance differed across focus category pairs. In the 1C-2P task, endorsement was highest for the pairs {Contrastive, Given} as well as {Contrastive, Broad}. The acoustic overlap of their prosodic realisations explains some of these differences. The other pairs were not as well endorsed, including {Contrastive, Narrow} and {Given, Narrow}.

In the 2C-1P tasks, endorsement rates were generally more variable. In Experiment 3, where listeners are given more nuanced response options, the bias against matching a prosodic pattern to a Broad^T focus condition is not apparent anymore, but we do find evidence for the bias favouring matches to a Narrow focus condition, with weaker endorsement rates for Narrow^C and strong endorsement rates for Narrow^T.

In sum, the experiment with a 5-point response option qualitatively confirmed most of the results from Experiments 1 and 2. It also becomes clear that given more nuanced response dimensions, listeners turn out to be very liberal when it comes to acceptable matches between prosodic form and focus-related meaning established by discourse context.

6. General Discussion

6.1 Summary

We have reported on three experiments to answer the question whether listeners perceive focus-related meaning on the basis of the prosodic form of an utterance. In Experiment 1, listeners had to decide which of two acoustic realisations matches a particular focus-related meaning established by the immediate discourse context (1C-2P). Listeners were able to distinguish different prosodic forms to match a certain focus category with above chance accuracy. Although listeners were able to match acoustic form and intended focus context, accuracy was rather low and performance varied strongly across different focus pairs. While some pairs of prosodically encoded focus categories seem to be more accurately distinguished (e.g. Contrastive vs. Broad, Contrastive vs. Given), other pairs elicited substantially worse performance, sometimes even failing to show above-chance accuracy (e.g. Given vs. Narrow).

In Experiment 2, listeners had to decide which of two focus categories specified by different discourse contexts is the best match to a particular acoustic prosodic form (2C-1P). This experiment uncovered interesting divergent results from Experiment 1, with listeners having greater difficulty matching question-answer pairs. We observed biases against selecting Broad focus as a match to any prosodic form, and favouring matches to Narrow focus to any prosodic form. These results suggest that listeners are influenced by other aspects of the stimuli than just the prosodic information in the signal (which they were clearly able to use during 1C-2P). As opposed to observed biases with Broad and Narrow focus prosody, listeners were able to assign Given and Contrastive prosodic realisations to their congruent discourse contexts.

Experiment 3 conceptually replicated Experiments 1 and 2 but using a 5-point scalar response option instead of a two-alternatives forced choice task. The results confirm what we have observed for the other experiments with one notable exception. The strong bias towards Broad focus vanishes in the 2C-1P task, suggesting that the bias towards Broad focus contexts only surfaces when listeners have to categorically decide for or against a context. When they have less restricted decision options, e.g. being able to choose that neither of the offered question-answer pairs is a better match, no bias against matches to the Broad focus context manifests anymore.

This is not true for the Narrow focus bias. Experiment 3 shows that listeners have a clear bias towards Narrow focus contexts, confirming that this focus type allows for a large variety of different prosodic realisations.

While we can confirm our hypothesis that listeners are sensitive to the acoustic prosodic expression of focus categories, there are two groups of questions that arise from our results: First, why are listeners' accuracies generally so low and why are some categories better distinguished than others? Second, why are listeners generally biased to match utterances with certain contexts but not with others?

6.2 Perceptual sensitivity is dependent on target and competitor category

The prosodic realisations of our stimuli are acoustically distinct (i.e. an LDA analysis can tease them apart with very high accuracy), so why do listeners have difficulties in mapping the speech signal onto speaker intentions?

One could argue that the low accuracy might be an artefact of the task, being artificial to some extent and devoid of (linguistic) functionality. Acoustic cues are more pronounced when the interlocutor is present (Breen et al., 2010; Buxo-Lugo, Toscana, & Watson, 2018; Turnbull, Royer, Ito, & Speer, 2017), when the speaker believes that the listener is distracted (Rosa, Finch, Bergeson, & Arnold, 2015), and when there is ambiguity in the context (Snedeker & Trueswell, 2003). Our model speaker produced her utterances in a context which is largely devoid of a communicative context, and listeners might not be able to access their entire knowledge about possible form-function mappings within the experiment. However, even in the experiment by Breen et al. (2010) which took great care in creating a functional communication situation between speakers and listeners, listeners still had difficulties mapping acoustic form onto intended focus type. This suggests that the low accuracy that we obtained is not necessarily an artefact of the task. Any explanation hinging on the artificial nature of the task also does not account for the fact that listeners can assign some prosodic forms to their intended focus context, in fact, with very high accuracy.

An alternative interpretation might be related to the amount of acoustic overlap between prosodic realisations of focus types. Concentrating on the f_0 maxima on the subject and the object constituent (as strongly related to the phonological pitch accent placement and pitch accent choice, see above), we can already see that some focus categories overlap more than others. For the f_0 maxima of the subject, the focus categories fall into two groups. Both the Broad and Contrastive groups, and the Given and Narrow groups overlap substantially. For the f_0 max of the object, Broad and Contrastive are actually well separated. Given and Narrow remain highly overlapping. These patterns reflect some of our 1C-2P results. Accuracy for Contrastive competing with Broad and Given was high, much higher than accuracy for Contrastive competing with Narrow. However, Given and Broad exhibit very well separated distributions in f_0 max, but elicit weaker accuracy, suggesting that there may be factors affecting their performance that go beyond simple acoustic overlap between categories.

Linguistic meaning is signalled by many temporally distributed cues throughout the discourse (e.g. Winter, 2014). Breen et al. (2010) showed that listeners' accuracies went up when the target sentences were preceded by the phrase "I heard that", suggesting that speakers signal focus categories prosodically on preceding syntactic material. Similarly, Xu and Xu (2005) found that focus categories are differentiated by both expanded pitch range on the focused constituents as well as post-focal compression on the lexical items following the focused constituent. Beyond distributed redundancy in the speech signal, non-verbal context might provide important disambiguating information. Speech communication does not happen in a void, but is accompanied by changes in body posture, head position, gaze, facial expressions, and manual gestures (e.g. Kendon, 2004; McNeill, 1992). For example, Krahmer and Swerts (2007) showed that Dutch speakers place more acoustic emphasis on words if their production is accompanied by a visual cue (eyebrow movement or head nod) and that subjects are more likely to perceive a word as prominent if accompanied by a visual cue. In an experiment such as the one that is the focus of the present analysis, in which speech stimuli are presented with very limited context, listeners have only a subset of information channels to make decisions about prosodically encoded meaning related to focus, leading to less certainty about their decisions. Some categories might benefit more or less from these contextual effects, accounting for category-specific performances.

Yet another aspect to consider is the inherent probabilistic nature of form-function mappings in prosody. One could argue that focus categories, as many other discourse functions, may not be discretely signalled by prosody in a deterministic way. In other words, listeners may be sensitive to prosodic cues, while recognising ambiguity in the mapping back to the speaker-intended meaning. Accumulating evidence reveals that intonation is characterised by a many-to-many-mapping between prosodic form and discourse function (Cangemi et al., 2015; Chodroff & Cole, 2018; Cruttenden, 1986; Grice et al., 2017; Peppé et al., 2000; Roettger, 2017, Turnbull, 2017). Specific prosodic forms are probabilistically associated with certain discourse functions. Language users have access to this knowledge which is reflected in (discretely) variable speech production patterns (one and the same speaker uses discretely different phonological forms to signal the same meaning) which results in observed flexibility in the comprehension of prosodically encoded discourse meaning in the lab (e.g. Grice et al., 2017; Roettger, 2017; Roettger & Grice, 2015).

In order to avoid making mappings that are different from those intended by the speaker, listeners need to adapt with respect to a given speaker (or a given context). The response data analysed here come from a series of experiments in which listeners rated as few as 18 utterances from a single speaker, offering only a slim basis for adaptation. A failure to adapt means that the listener's prior knowledge plays a greater role in speech perception. If listeners' prior beliefs of the form-function mapping for prosody is characterised by stochastic distributions rather than deterministic one-to-one relationships, that could account for some of the variability in the response patterns analysed here.

6.3 Listeners have biased expectations about suitable contexts

The rather low and inconsistent performance in mapping between prosodic form and discourse meaning might be a natural disposition of language users. The present study, like older studies on the perception of prosodic meaning, suggests that mapping an utterance onto a pragmatic meaning in the absence of a genuine communicative context is a difficult task and one that elicits highly variable performance from listeners. Nonetheless, and despite the inherent stochasticity of intonational form-function mappings, there are several studies showing that listeners rapidly integrate intonational information to anticipate speaker intentions (e.g., Dahan et al., 2002; Ito & Speer, 2008; Roettger & Stoeber, 2017; Watson et al., 2008; Weber et al., 2006). Listeners' ability to make use of bottom-up acoustic cues may be complemented by probabilistic knowledge about speaker production likelihoods, i.e. how likely the speaker is to use a particular prosodic form in order to express a particular discourse function (Buxo-Lugo, 2017; Buxo-Lugo & Watson, 2016; Kurumada et al. 2014, b, Roettger & Franke, 2018a,b).

For example, in Roettger and Franke (2018a,b), listeners were exposed to two intonation contours. These contours exhibited early intonational cues to speaker intentions, i.e. cues that become available before the lexical content disambiguates between competing interpretations of discourse meaning. Roettger and Franke showed that the assumed production likelihood of a prosodic cue predicted listeners' anticipatory behaviour at the beginning of the experiment as well as its development through exposure to confirming or disconfirming observations. In other words, when exposed to stochastically confirming or disconfirming form-function mappings, listeners adapt to what extent they predictively use an intonational cue. If listeners learn that an intonational cue (e.g. a particular pitch accent) is uninformative, they appear to weigh the informational value of that cue less heavily (see also Kurumada et al., 2014).[5] Roettger and Franke's results are in line with the assumption that language users have probabilistic knowledge about the stochastic co-occurrence of prosodic form and discourse function.

Coming back to the present findings, the above insights may offer an explanation as to why listeners are biased to (erroneously) reject broad focus and to (erroneously) accept narrow focus in the 2C-1P task. The broad focus question was a question like "What has happened?". This (or similar questions) are often used to elicit broad focus in the experimental literature. Semantically, this question does not pre-activate any discourse relations and allows for an out-of-the-blue interpretation. However, this discourse context is pragmatically very rare. We rarely encounter out-of-the-blue scenarios without any prior knowledge about the discourse, thus the likelihood of a speaker expressing (truly) Broad focus is arguably very low. As opposed to that, the Given context, i.e. repeating the previously heard proposition, and Contrastive focus, i.e. correcting the previously heard proposition, are very common discourse scenarios, albeit occurring in very specific discourse contexts. Finally, narrowly focusing a constituent is arguably a very general pragmatic function that applies to many different discourse contexts. We encounter a Narrow focus context very often, thus the likelihood of a speaker expressing Narrow focus is arguably very high. For exposition purposes, let us assume that Narrow and Broad focus are not prosodically differentiated (so any intonational cue (I) has the same probability (P) of expressing Narrow (N) and Broad (B) focus, i.e. $P(I|B) = P(I|N)$). If the prior assumption about how likely a discourse function is expressed is asymmetric, i.e. $P(B) < P(N)$, listeners would believe that Narrow focus is more likely to be expressed by any given prosodic form, i.e. the probability of a Narrow focus interpretation, given any intonational cue would be higher than the probability of a Broad focus interpretation, via Bayes Rule, cf. (1):

[5] This adaptive behavior is in line with language users adapting readily to their immediate local context in syntax (e.g. Fine, Jaeger, Farmer, & Qian, 2013; Jaeger & Snider, 2013), pragmatics (e.g. Grodner & Sedivy, 2011; Yildirim et al., 2016), and, most importantly, in speech (e.g. Kleinschmidt & Jaeger, 2015; Norris et al., 2003).

$$\frac{P(N|I)}{P(B|I)} = \frac{P(I|N)}{P(I|B)}\frac{P(N)}{P(B)} > \frac{P(I|B)}{P(I|N)}\frac{P(B)}{P(N)} = \frac{P(B|I)}{P(N|I)} \quad (1)$$

These types of assumptions are in line with an rational analysis approach (Anderson, 1990) to speech perception (e.g. Clayards et al., 2008; Kleinschmidt & Jaeger, 2015), assuming that prosodic perception and processing can be conceptualised as a process of *inference under uncertainty*: listeners know that certain discourse functions are realised as a distribution of acoustic cues and the listener probabilistically infers how likely any given speaker intention is, taking into account both their knowledge about stochastic cue distributions as well as their knowledge about speaker and context. We want to emphasize that this is an ad-hoc explanation that remains speculative until further investigations. We believe, however, that this explanation offers an insightful perspective on previous findings in general and our findings in particular.

7. Conclusion

The prosodic modulation of speech is a tremendously important aspect of human language. However, our knowledge as to how language users interpret prosody to guide intention recognition is still surprisingly small. The present paper contributes to this knowledge. We have presented evidence that listeners can use prosodic information to infer the intended information structure of an utterance, even in a laboratory setting that is devoid of contextual information. These results complement the existing literature on American English in that they clearly show listener's ability to discriminate prosodic forms intended by the speaker to signal focus types (e.g. Breen et al., 2010; Gussenhoven, 1983; Welby, 2003). Our study further contributes to research on prosody and meaning in general in that our 2C-1P tasks allow us to uncover certain meaning biases in how listeners associate prosodic forms with focus-related discourse meaning. The experimental tasks used here may tap into comprehension processes that are not only informed by acoustic information but also by listeners' prior knowledge of the contextual probability of a prosodic form.

More clearly than in previous studies, the experiments presented in this study suggest a high degree of overlap in the pairing of prosodic form and information structure categories, with some prosodically encoded focus types being more accurately associated with discourse contexts than others. These differences may be related to different degrees of acoustic / perceptual overlap between the prosodic categories. Although we did not investigate a representative sample of production data, we have discussed idiosyncratic patterns of our model speaker for whom some categories may be more or less overlapping with regard to relevant phonetic dimensions.

In addition to acoustic prosodic properties of the intended focus types, our data suggests additional factors contributing to listeners' mappings of form onto function: Listeners appear to be influenced by their probabilistic knowledge about how likely a speaker is to produce a certain prosodic form and how likely this form will be used as intended by a speaker to communicate a certain discourse function. The latter explanation can account for the observed meaning biases and is in line with recent studies on intonational processing (Buxo-Lugo & Watson, 2016;

Kurumada et al. 2014, Roettger & Franke, 2018a,b) and speech perception in general (e.g. Clayards et al., 2008; Kleinschmidt & Jaeger, 2015; Kleinschmidt et al., 2018; Norris et al., 2003). This explanation, although grounded in recent experimental studies, remains speculative and should merely serve as a departure point for future studies working on the mapping of prosody and meaning.

We conclude that listeners infer speaker intentions based on both bottom-up acoustic cues and top-down probabilistic expectations about likely intonation contours.

8. References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Earlbaum.

Audacity Team (2015). *Audacity (r): Free Audio Editor and Recorder* [computer program]." Version 2.1.0.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv* Preprint *ArXiv*:1506.04967.

Baumann, S. (2006). *The Intonation of Givenness - Evidence from German*. PhD thesis, Saarland University. Linguistische Arbeiten 508.Tübingen: Niemeyer.

Baumann, S., Röhr, C. T., & Grice, M. (2015). Prosodische (De-) kodierung des informationsstatus im Deutschen. *Zeitschrift für Sprachwissenschaft*, 34(1), 1-42.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology*, 3(1), 255–309.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7–9), 1044–1098.

Brown, G., & Yule, G. (1983). Intonation, the categories given/new and other sorts of knowledge. In A. Cutler & D. R. Ladd (Eds.)*, Prosodic Function and Prosodic Representation,* Cambridge: Cambridge University Press.

Büring, D. (2006). Intonation und Informationsstruktur. In H. Blühdorn, E. Breindle, & U. H. Waßner, (Eds.): *Text - Verstehen. Grammatik und darüber hinaus* (pp. 144–163). Berlin:

Büring, D. (2012). Focus and intonation. In G. Russell and D. Graff Fara, (eds*.) The Routledge Companion to the Philosophy of Language* (pp. 103-115). London: Routledge.

Bürkner, P.-C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.

Buxó-Lugo, A. F., & Watson, D. G. (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of Memory and Language*, 90, 1-13.

Buxo-Lugo, A. F. (2017). *Communicative context, expectations, and adaptation in prosodic production and comprehension* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Buxó-Lugo, A. F., Toscano, J. C., & Watson, D. G. (2018). Effects of participant engagement on prosodic prominence. *Discourse Processes*, 55(3), 305-323.

Calhoun, S. (2006). *Intonation and information structure in English*. (Doctoral dissertation, Ph. D. thesis, University of Edinburgh)

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics*, 40(2), 329–349.

Cangemi, F., & Grice, M. (2016). The importance of a distributional approach to categoriality in Autosegmental-Metrical accounts of intonation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1), 1–20.

Cangemi, F., Krüger, M., & Grice, M. (2015). Listener-specific perception of speaker-specific production in intonation. In S. Fuchs, D. Pape, C. Petrone, P. Perrier (Eds.): *Individual Differences in Speech Production and Perception* (pp. 123–145), Frankfurt a. M.: Peter Lang.

Chafe, W. (1987). Cognitive constraints on information flow. *Coherence and Grounding in Discourse*, 11, 21–51.

Chodroff, E., & Cole, J. (2018). Information Structure, Affect, and Prenuclear Prominence in American English. *Proceedings of Interspeech* 2018, 1848-1852.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804-809.

Cooper, W., Eady, S. & Mueller, P. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of Acoustical Society of America*, 77(6), 2142-2156.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, 8(10), e77661.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, 27(4), 765-768.

Grice, M., Ridouane, R., & Roettger, T. B. (2015). Tonal association in Tashlhiyt Berber: Evidence from polar questions and contrastive statements. *Phonology*, *32*(2), 241–266.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90–107.

Grodner, D., & Sedivy, J. C. (2011). The Effect of Speaker-Specific Information on Pragmatic Inferences. In N. Pearlmutter & E. Gibson (Eds.): *The processing and acquisition of reference* (pp. 239-272). Cambridge, MA.: MIT Press.

Gussenhoven, C. (1983). Testing the reality of focus domains. *Language and Speech*, 26(1), 61–80.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83.

Jaynes, E. T., & Kempthorne, O. (1976). Confidence Intervals vs. Bayesian Intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of Probability Theory, Statistical Inference,* and Statistical Theories of Science (Vol. 6b, pp. 175–257). Dordrecht: Springer.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2), 148.

Kleinschmidt, D. F., Weatherholtz, K., & Florian Jaeger, T. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in cognitive science*, 10, 818–834.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396-414.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In *Proceedings of the Annual Meetinf of the Cognitive Science Society*, 36. Austin, TX: Cognitive Science Society.

Ladd, D. R. (2008). *Intonational Phonology (2nd)*. Cambridge: Cambridge University Press.

Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics*, *31*(1), 81–112.

Mahrt, T. (2016). *LMEDS: Language markup and experimental design software*. Retrieved from https://github.com/timmahrt/LMEDS

Mahrt, T. (2018). *Acoustic cues for the perception of the information status of words in speech*. University of Illinois at Urbana-Champaign, Urbana, Illinois.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The Fallacy of Placing Confidence in Confidence Intervals, *Psychonomic bulletin & review,* 23(1), 103–123.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204–238.

Peppé, S., Maxim, J., & Wells, B. (2000). Prosodic variation in southern British English. *Language and Speech*, 43(3), 309–334.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. MIT, Bloomington, Indiana.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in Communication* (pp. 271–311). Cambridge, MA: MIT Press.

R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Roettger, T. B. (2017). *Tonal placement in Tashlhiyt: How an intonation system accommodates to adverse phonological environments*. Berlin: Language Science Press.

Roettger, T. B., & Grice, M. (2015). The role of high pitch in Tashlhiyt Tamazight (Berber): Evidence from production and perception. *Journal of Phonetics*, 51(1), 36–49.

Roettger, T., & Stoeber, M. (2017). Manual Response Dynamics Reflect Rapid Integration of Intonational Information during Reference Resolution. In G. Gunzelmannn, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of Annual Meeting of the Cognitive Science Society*, 39 (pp. 3010–3015). Austin, TX: Cognitive Science Society.

Roettger, T. B., & Franke, M. (2018a). Dynamic speech adaptation to unreliable cues during intonational processing. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of Annual Meeting of the Cognitive Science Society*, 40. Austin, TX: Cognitive Science Society.

Roettger & Franke (2018b). Evidential strength of intonational cues and rational adaptation to (un-)reliable intonation. Preprint at PsyArXiv: https://psyarxiv.com/awp87.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1), 75–116.

Rosa, E. C., Finch, K. H., Bergeson, M., & Arnold, J. E. (2015). The effects of addressee attention on prosodic prominence. *Language, Cognition and Neuroscience*, 30(1-2), 48-56.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech*, 39(1), 1–17.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, 20(2), 416–420.

Selkirk, E. (1995). Sentence prosody: Intonation, stress, and phrasing. In: J. A. Goldsmith (Ed.): *The handbook of phonological theory*, (pp. 550–569). Cambridge MA & Oxford: Blackwell.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, 48(1), 103-130

Turnbull, R. (2017). The role of predictability in intonational variability. *Language and speech*, 60(1), 123-153.

Turnbull, R., Royer, A. J., Ito, K., & Speer, S. R. (2017). Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Language, Cognition and Neuroscience*, 32(8), 1017-1033.

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+ H. *Cognitive Science*, 32(7), 1232–1244.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, 49(3), 367–392.

Welby, P. (2003). Effects of pitch accent position, type, and status on focus projection. *Language and Speech*, 46(1), 53–81.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, 36(10), 960-967.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, 33(2), 159-197

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, 87, 128–143.

Yoon, T.-J. (2010). Speaker consistency in the realization of prosodic prominence in the Boston University Radio Speech Corpus. In *Proceeding of 5th Speech Prosody, Chicago*.

9. Appendix

Appendix 1: All target statements and corresponding context questions used in experiments 1-3.

Broad Focus	Do you know what happened yesterday?

(a) Daisy warned the owner.

Narrow Focus	Do you know who warned the owner?
Contrastive Focus	Did Jennifer warn the owner?
Given	Did Daisy warn the owner?

(b) Damon fried the omelet.

Narrow Focus	Do you know who fried the omelet?
Contrastive Focus	Did Pam fry the omelet?
Given	Did Damon fry the omelet?

(c) Dorah filmed the movie.

Narrow Focus	Do you know who filmed the movie?
Contrastive Focus	Did Susy film the movie?
Given	Did Dorah film the movie?

(d) Harry raised the window.

Narrow Focus	Do you know who raised the window?
Contrastive Focus	Did Tom raise the window?
Given	Did Harry raise the window?

(e) Jamie dyed the laundry.

Narrow Focus	Do you know who dyed the laundry?
Contrastive Focus	Did Colin dye the laundry?
Given	Did Jamie dye the laundry?

(f) Jonny helped the warden.

Narrow Focus	Do you know who helped the warden?
Contrastive Focus	Did Liz help the warden?
Given	Did Jonny help the warden?

(g) Jonah burned the onion.

Narrow Focus	Do you know who burned the onion?
Contrastive Focus	Did Mark burn the onion?
Given	Did Jonah burn the onion?

(h) Maddie found the TV.

Narrow Focus	Do you know who found the TV?
Contrastive Focus	Did Jennifer find the TV?
Given	Did Maddie find the TV?

(i) Mary rolled the barrel.

Narrow Focus	Do you know who rolled the barrel?
Contrastive Focus	Did Trisha roll the barrel?
Given	Did Mary roll the barrel?

Appendix 2: Summary of posterior distributions for 1C-2P trials in Experiment 3: Posterior means (95% credible intervals in brackets) for all focus combinations for all five possible responses, given the data and the model. Dark shaded cells indicate conditions in which listeners selected given responses below chance and light shaded cells indicate conditions in which listeners selected given responses above chance according to set inferential criteria (systematically chosen)

Target	**Competitor**	**estimated probability of responses**				
		never	**dispreferred**	**equal**	**preferred**	**always**
broad	contr.	0 (0,0)	0.03 (0.02,0.05)	0.29 (0.21,0.37)	0.58 (0.52,0.64)	0.1 (0.06,0.14)
broad	given	0.01 (0,0.01)	0.09 (0.05,0.13)	0.47 (0.39,0.54)	0.4 (0.3,0.5)	0.04 (0.02,0.06)
broad	narrow	0.01 (0,0.01)	0.08 (0.05,0.11)	0.45 (0.37,0.53)	0.43 (0.33,0.53)	0.04 (0.02,0.06)
contr.	broad	0 (0,0.01)	0.05 (0.03,0.08)	0.38 (0.28,0.48)	0.5 (0.41,0.6)	0.06 (0.03,0.09)
contr.	given	0.01 (0,0.01)	0.06 (0.03,0.11)	0.4 (0.27,0.52)	0.48 (0.34,0.6)	0.06 (0.02,0.09)
contr.	narrow	0.02 (0.01,0.03)	0.17 (0.07,0.27)	0.54 (0.48,0.58)	0.26 (0.13,0.39)	0.02 (0.01,0.03)
given	broad	0.01 (0.01,0.02)	0.15 (0.09,0.2)	0.54 (0.5,0.58)	0.28 (0.2,0.36)	0.02 (0.01,0.03)
given	contr.	0 (0,0)	0.02 (0.01,0.03)	0.19 (0.12,0.27)	0.63 (0.59,0.66)	0.16 (0.1,0.24)
given	narrow	0.01 (0.01,0.02)	0.12 (0.07,0.18)	0.52 (0.46,0.57)	0.32 (0.22,0.42)	0.03 (0.01,0.04)
narrow	broad	0.01 (0.01,0.02)	0.11 (0.07,0.16)	0.51 (0.45,0.56)	0.34 (0.24,0.42)	0.03 (0.02,0.04)
narrow	contr.	0.01 (0.01,0.02)	0.15 (0.09,0.22)	0.54 (0.49,0.58)	0.27 (0.18,0.38)	0.02 (0.01,0.03)
narrow	given	0.01 (0.01,0.02)	0.13 (0.08,0.2)	0.53 (0.47,0.57)	0.3 (0.2,0.41)	0.02 (0.01,0.04)

Appendix 3: Summary of posterior distributions for 2C-1P trials in Experiment 3: Posterior means (95% credible intervals in brackets) for all focus combinations for all five possible responses, given the data and the model. Dark shaded cells indicate conditions in which listeners selected given responses below chance and light shaded cells indicate conditions in which listeners selected given responses above chance according to set inferential criteria.

Target	**Competitor**	**estimated probability of responses**				
		never	**dispreferred**	**equal**	**preferred**	**always**
broad	contr.	0.03 (0.02,0.04)	0.24 (0.16,0.32)	0.41 (0.39,0.44)	0.29 (0.21,0.38)	0.03 (0.02,0.05)
broad	given	0.02 (0.01,0.03)	0.21 (0.15,0.27)	0.41 (0.38,0.43)	0.32 (0.24,0.4)	0.04 (0.02,0.05)
broad	narrow	0.09 (0.05,0.13)	0.46 (0.39,0.53)	0.33 (0.27,0.39)	0.11 (0.07,0.16)	0.01 (0.01,0.01)
contr.	broad	0.01 (0,0.01)	0.06 (0.04,0.09)	0.23 (0.17,0.3)	0.56 (0.51,0.6)	0.14 (0.09,0.19)
contr.	given	0.01 (0,0.01)	0.07 (0.04,0.11)	0.25 (0.17,0.33)	0.55 (0.49,0.61)	0.13 (0.07,0.19)
contr.	narrow	0.05 (0.03,0.08)	0.36 (0.26,0.46)	0.39 (0.34,0.43)	0.18 (0.11,0.25)	0.02 (0.01,0.02)
given	broad	0.02 (0.01,0.03)	0.18 (0.12,0.23)	0.39 (0.35,0.43)	0.37 (0.29,0.45)	0.05 (0.03,0.07)
given	contr.	0.02 (0.01,0.03)	0.18 (0.11,0.27)	0.39 (0.34,0.44)	0.36 (0.24,0.47)	0.04 (0.02,0.07)
given	narrow	0.04 (0.02,0.07)	0.32 (0.2,0.43)	0.41 (0.36,0.44)	0.21 (0.12,0.31)	0.02 (0.01,0.03)
narrow	broad	0 (0,0.01)	0.04 (0.03,0.06)	0.18 (0.13,0.23)	0.58 (0.56,0.61)	0.19 (0.13,0.26)
narrow	contr.	0.02 (0.01,0.03)	0.18 (0.1,0.25)	0.39 (0.34,0.43)	0.37 (0.27,0.47)	0.05 (0.02,0.07)
narrow	given	0.01 (0.01,0.02)	0.13 (0.08,0.18)	0.35 (0.29,0.41)	0.45 (0.35,0.53)	0.07 (0.04,0.1)

Matthew Gordon[1] / Timo Roettger[2]

Acoustic correlates of word stress: A cross-linguistic survey

[1] University of California, Santa Barbara, CA 93106, USA, E-mail: mgordon@linguistics.ucsb.edu
[2] University of Cologne, 50923 Cologne, Germany

Abstract:
The study of the acoustic correlates of word stress has been a fruitful area of phonetic research since the seminal research on American English by Dennis Fry over 50 years ago. This paper presents results of a cross-linguistic survey designed to distill a clearer picture of the relative robustness of different acoustic exponents of what has been referred to as word stress. Drawing on a survey of 110 (sub-) studies on 75 languages, we discuss the relative efficacy of various acoustic parameters in distinguishing stress levels.

Keywords: word stress, phonetics, phonology, typology

DOI: 10.1515/lingvan-2017-0007

Received: March 7, 2017; **Accepted:** April 26, 2017

1 Introduction

The study of the acoustic correlates of word stress has been a fruitful area of phonetic research since the seminal research on American English by Fry (1955; 1958) over 50 years ago. This paper presents results of a cross-linguistic survey designed to distill a clearer picture of the relative robustness of different acoustic exponents of word stress. The present paper will not attempt to address the complex issue of situating word stress within the broader taxonomy of prosodic systems (see Beckman 1986; Hyman 2006; 2014; inter alia). Rather, we assume word stress (or simply 'stress') to be the phonological marking of one or more prominent syllables within the phonological word. In practice, for many of the languages surveyed in this paper, the classification of the prosodic system is not conclusive. In order to be as inclusive as possible, studies of languages whose prosodic systems are open to alternative interpretations were included in the present study. Although future consensus might suggest that these languages are better classified as lacking stress, their inclusion in the present study at least allows for contextualizing their phonetic properties relative to the broader literature on acoustic correlates of prominence.

2 Methodology

Several different primary sources were consulted, including a number of phonetics and areal studies journals, working papers volumes and books and dissertations. The corpus (in the form of a table) is publically available online at https://osf.io/9r2cd/ alongside a script to reproduce respective counts presented in this manuscript. To establish a reliable and informative corpus that can be used in the future, cited authors are encouraged to submit corrections, if we have interpreted respective aspects of their method and/or results incorrectly. Further, we invite scholars that have published work on word stress that is not logged in the present corpus to share their results with us for inclusion in the database.

Although the database was intended to be as comprehensive as possible, many works that dealt with stress were excluded from the present study on various methodological grounds. First, papers in which methodological description was too sparse or vague to allow for replication were excluded. Likewise excluded were studies that did not present quantitative results. Also omitted were papers not explicitly focused on stress. Papers on stress were included, however, even if experimental design created confounds that could render definitive interpretation of results impossible. For example, several studies were based on words uttered in isolation where word-level stress is conflated with phrase-level prominence, while many others employed carrier phrases in which the target word was (either likely or explicitly) focused, thereby creating a potential confound between

Matthew Gordon is the corresponding author.

phrase-level prominence and word-level stress (see Roettger and Gordon this volume). Finally, we included only studies on populations consisting of adult speakers without reported speech impairments.

The corpus encompassed a total of 110 (sub-)studies on 75 languages or language varieties, e. g. Jordanian and Tunisian Arabic, American and British English.[1] Languages in the survey are plotted geographically in Figure 1 and listed in Table 1 along with their genetic affiliation according to the 19th edition of the Ethnologue (Lewis et al. 2016) and the sources consulted in the survey.

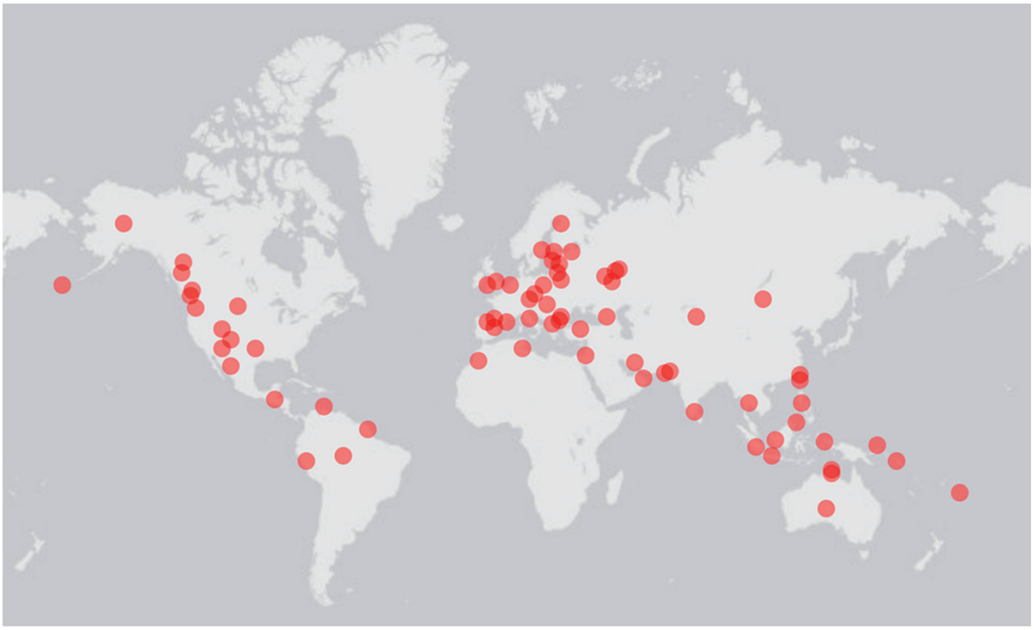

Figure 1: Geographical distribution of languages included in the survey of acoustic correlates of stress plotted via the "lingtypology" package (Moroz 2017) for R (2017).

Table 1: Languages included in the survey of acoustic correlates of stress.

Language	Genetic affiliation	Source(s)
Aleut	Eskimo-Aleut	Rozelle (1997)
Apache, Jicarilla	Na Dene	Tuttle (2005)
Apache, San Carlos	Na Dene	Tuttle (2005)
Arabic, Jordanian	Afro-Asiatic	De Jong and Zawaydeh (1999, 2002)
Arabic, Tunisian	Afro-Asiatic	Bouchhioua (2008)
Basque, Goizueta	Isolate	Hualde et al. (2008)
Belarusian	Indo-European	Borise (2015)
Besemah	Austronesian	McDonnell (2014)
Bininj Gun-wok	Australian	Bishop (2002), Fletcher and Evans (2002)
Bulgarian	Indo-European	Crosswhite (2003)
Catalan	Indo-European	Astruc and Prieto (2006), Ortega-Llebaria and Prieto (2010)
Chabacano, Cavite	Creole	Lesho (2013)
Chickasaw	Muskogean	Gordon (2004)
Chuvash	Turkic	Dobrovolsky (1999)
Czech	Indo-European	Duběda (2006)
Dalabon	Australian	Fletcher and Evans (2002)
Dutch	Indo-European	Sluijter and van Heuven (1996), Rietveld et al. (2004)
Émérillon	Tupi-Guarani	Gordon and Rose (2006)
English, American	Indo-European	Fry (1955), Lieberman (1960), Huss (1978), Plag et al. (2011)
English, British	Indo-European	Bouchhioua (2003), Eriksson and Heldner (2015)
Estonian	Uralic	Lehiste (1966), Gordon (1995), Lippus et al. (2006)
Finnish	Uralic	Tuomainen et al. (1999), Suomi et al. (2001)
Finnish, Ingrian	Uralic	Gordon (2009)
German	Indo-European	Dogil (1999), Kleber and Klipphahn (2006)
Greek	Indo-European	Vogel et al. (2016)
Hebrew	Afro-Asiatic	Silber-Varod et al. (2016)

Hungarian	Uralic	Vogel et al. (2016)
Indonesian, Javanese	Austronesian	Goedemans and van Zanten (2007)
Indonesian, non-Javanese	Austronesian	Adisasmito-Smith and Cohn (1996), Goedemans and van Zanten (2007)
Italian	Indo-European	Eriksson et al. (2016)
K'ekchi	Mayan	Berinstein (1979)
Kabardian	North Caucasian	Gordon and Applebaum (2010)
Kuot	Isolate	Lindström and Remijsen (2005)
Lakhota	Siouan	Cho (2006)
Latvian	Indo-European	Bond (1991)
Lithuanian	Indo-European	Dogil (1999)
Livonian	Uralic	Lehiste et al. (2008)
Ma'ya	Austronesian	Remijsen (2002)
Macedonian	Indo-European	Crosswhite (2003)
Meadow Mari	Uralic	Lehiste et al. (2005)
Mongolian	Mongolic	Harnud (2003)
Mordvin, Ezrya	Uralic	Lehiste et al. (2003)
Mordvin, Moksha	Uralic	Aasmäe et al. (2013)
Nahuatl, Balsas	Uto-Aztecan	Guion et al. (2010)
Paiwan	Austronesian	Chen (2009)
Papiamentu	Creole	Remijsen and van Heuven (2002), Rivera-Castillo and Pickering (2004)
Persian	Indo-European	Sadeghi (2011)
Pirahã	Mura	Everett (1998)
Pitjantjatjara	Australian	Tabain et al. (2014)
Polish	Indo-European	Dogil (1999), Crosswhite (2003), Newlin-Łukowicz (2012)
Portuguese, Brazilian	Indo-European	Barbosa et al. (2013)
Quechua, Conchucos	Quechua	Hintz (2006)
Saisiyat	Austronesian	Chiang and Chiang (2005)
Savosavo	Central Solomons	Simard et al. (2014)
Sekani	Na Dene	Hargus (2005)
Sindhi	Indo-European	Abbasi (2015)
Spanish	Indo-European	Ortega-Llebaria (2006), Ortega-Llebaria and Prieto (2010)
Squamish	Salish	Tamburri-Watt et al. (2000)
St'át'imcets	Salish	Caldecott (2009)
Swedish	Indo-European	Barbosa et al. (2013)
Tagalog	Austronesian	Gonzales (1970)
Tamil	Dravidian	Keane (2006)
Tanana, Minto	Na Dene	Tuttle (1998)
Tanana, Salcha	Na Dene	Tuttle (1998)
Tarahumara	Uto-Aztecan	Caballero and Carroll (2015)
Tashlhiyt	Afro-Asiatic	Gordon and Nafi (2012), Roettger et al. (2015)
Thai	Tai-Kadai	Potisuk et al. (1996)
Tongan	Austronesian	Garellek and White (2015)
Turkish	Turkic	Levi (2005), Pycha (2006), Vogel et al. (2016)
Urdu	Indo-European	Hussain (1997)
Ute, Southern	Uto-Aztecan	Oberly (2008)
Uyghur	Turkic	Yakup and Sereno (2016)
Welsh	Indo-European	Williams (1983, 1999)
Witsuwit'en, Babine	Na Dene	Hargus (2005)
Yakima Sahaptin	Sahaptian	Hargus and Beavert (2006)

For each of the studies (and sub-studies within a single work) that satisfied the criteria for inclusion in the survey, several pieces of information were logged, including the name of the language, whether the language is tonal (which includes languages often regarded as having lexical "pitch accent" rather than canonical tone), the word stress levels examined (primary stress (1S), secondary stress (2S) and unstressed (US)), the acoustic parameters used to express word stress, as well as other methodological aspects (see Roettger and Gordon this volume, for discussion).

3 Acoustic correlates of stress

Studies in the database differed in the acoustic dimension(s) investigated. These can be coarsely broken down into four categories: duration, fundamental frequency, intensity, and spectral characteristics.

In most studies, duration values were taken of only the vowel. Also attested were measures of the syllable rime (labeled "R" in the database), the nucleus (labeled "N") if the nucleus could be a consonant, the entire syllable (labeled "syll"), and consonant durations, typically of the constriction for syllable onsets (labeled "O") and, more rarely, of voice-onset-time (labeled "VOT") values for onsets or duration values for syllable coda consonants (labeled "C"). Most duration measurements were absolute measures calculated over a given domain, although some studies employed measures relative to another segment.

The most common fundamental frequency measurement (unlabeled in the corpus) was the mean for the vowel. Other fundamental frequency measurements (typically from the vowel) included peak F0, F0 at vowel midpoint or at the intensity peak, variability of F0 (calculated as F0 standard deviation) as well as time varying characteristics such as F0 slope, or values taken at regular intervals of either a fixed absolute length or a fixed proportion of a segment or syllable, e.g. quarter-length intervals.

The most frequent measure of intensity in the database was also the mean (usually calculated over the vowel), sometimes taken as a relative measure between stressed and unstressed syllables, which helps to mitigate fluctuations in intensity attributed to differences in the distance between the mouth and the microphone (if not worn on the head). Less common were measurements of peak intensity and intensity at the midpoint of the vowel and the intensity integral, the overall intensity aggregated over the entire duration of the target. This integration of intensity over time captures the increased perceptual loudness of a longer stimulus relative to a shorter one, at least over relatively short durations characteristic of vowels (see Moore 2013).

The final macro-category of measurements comprised various spectral measures. The most common frequency-sensitive measure consisted of formant values, typically for the first two formants. The other type of spectral measure observed in the database reflects the tendency for stressed vowels to display relatively less attenuation of energy at higher frequencies relative to unstressed values. Measurements of spectral tilt were quantified in various ways depending on the study, including frequency-bounded intensity bands, relative amplitude of the first and second harmonic (H1-H2), amplitude values of harmonics proximal to formants, slope of intensity declination as a function of frequency, and frequency-adjusted loudness scales such as the phon.

The stacked bar plot in Figure 2 graphically depicts both the number of (sub-) studies (out of a total of 110) which identified a given acoustic parameter as a marker of stress (dark bars) vs. the number of studies for which a given parameter was examined but found not to signal stress (grey bars). A parameter is identified as a successful marker of stress if it distinguishes at least two levels of stress, i. e. primary stressed vs. unstressed, primary stressed vs. secondary stressed, or secondary stressed vs. unstressed. The two frequency-sensitive measures, formant frequencies and spectral tilt, are separated due to the inclusion of data on both in many studies.

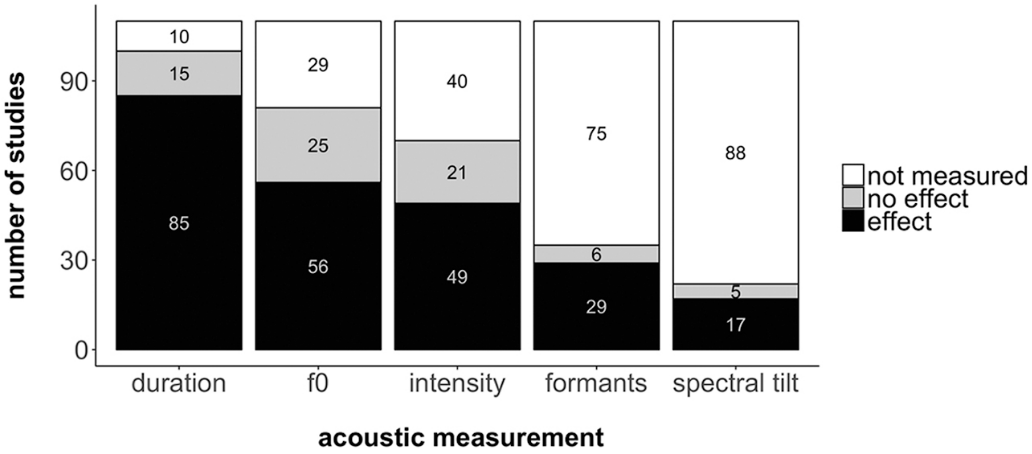

Figure 2: Number of (sub-) studies for which various acoustic correlates of stress were successful (black bars), unsuccessful (grey bars), and not measured (white bars) in differentiating stress level.

3.1 Duration

Duration was by far the most frequently measured property in the database (100 of 110 (sub-)studies ranging over 72 languages) and also the most successful marker of stress, distinguishing stress in 85 of 100 (sub-) studies and 65 of 72 languages.

Interestingly, for several languages, only consonant and not vowel duration successfully distinguished stress level. Thus, in Estonian (Gordon 1995; Lehiste 1966) and Peninsular Spanish Ortega-Llebaria (2006) onsets were lengthened in stressed syllables, while in Welsh (Williams 1999) stressed codas were durationally enhanced. Vowels in Lakhota (Cho 2006) were also not lengthened under stress, although VOT values for aspirated stops were greater in onset position of stressed syllables relative to their unstressed counterparts. Vowels in Yakima Sahaptin (Hargus 2005) were also not durationally distinct as a function of stress. However, contrary to the Lakhota results, VOT values in Yakima Sahaptin were *shorter* for stops in the onset of stressed syllables. The divergence between Lakhota and Yakima Sahaptin can be understood in terms of contrast enhancement: the phonemically aspirated stops of Lakhota are enhanced by lengthening VOT, whereas the unaspirated stops of Yakima Sahaptin are enhanced by shortening VOT values.

Finally, there are a few further studies in which an overall lengthening effect on either the syllable rime (Bond 1991; on Latvian, Chiang and Chiang 2005; on Saisiyat) or the entire syllable (Lehiste et al. 2005; on Meadow Mari, Sadeghi 2011; on Persian) emerged under stress, leaving the segmental source of the additional length unclear.

3.2 Fundamental frequency

Looking at F0 to examine word stress is notoriously difficult due to the common co-occurrence of word level prominence and post-lexical tonal events such as pitch accents (e. g. Bolinger 1958; 1961; Beckman 1986; Huss 1978; Ladd 2008; inter alia). As discussed in Roettger and Gordon, many studies in our corpus do not allow for teasing these levels apart. The following discussion about F0 as a marker of word stress should thus be interpreted with caution, a caveat that also applies to the other acoustic parameters to a certain extent.

F0 measures marked alleged word stress in 73 % (46 of 63) of the languages for which it was targeted for investigation in at least one study. The success rate of F0 as a correlate of stress becomes even higher if one excludes the five tone languages that fail to use F0 to mark stress.

Even if the languages for which studies demonstrated F0 to be a signal of stress but which relied on isolation forms are excluded, this still leaves a strong majority of languages in the database that used F0 to distinguish stress level. Most studies in which F0 was used to differentiate stress employed a static measure, typically the mean, but in some languages, only a dynamic and not a static measure of F0 was diagnostic of stress, e. g. Estonian (Liiv 1985; Gordon 1995), Thai (Potisuk et al. 1996), and Italian (Eriksson et al. 2016).

Of the nine tone languages in the database in which F0 was examined, it was reliably used to cue stress in only two, both of which lack the canonical profile of a tone language.[2] In Goizueta Basque (Hualde et al. 2008), a language in which tone is limited to certain lexical items (a property characteristic of traditional "pitch accent" languages), F0 distinguishes stress level only in words lacking lexically-specified tone. In Balsas Nahuatl (Guion et al. 2010), F0 has been retained as a diagnostic of stress even in dialects that have developed incipient tone distinctions while still retaining vestiges of the original penultimate stress system.

In most studies in which fundamental frequency diagnosed stress, F0 values were greater in stressed than unstressed syllables, although there were a pair of studies in which lowered F0 was symptomatic of stress: the speakers from Lahore (but not the one from Karachi) in the Hussain (1997) study of Urdu and the isolation words (but not those in context) in Eriksson's et al.'s (2016) research on Italian.

Although many of the F0 effects observed in the database could be attributed to post-lexical prominence (see Roettger and Gordon this volume for discussion), certain languages in the database still display an effect of stress on F0 when these factors are apparently controlled for by placing the target word in an utterance in which another word is explicitly focused, e. g. Finnish (Suomi et al. 2001), Greek, Hungarian, and Peninsular Spanish (Vogel et al. 2016). On the other hand, the possibility that target words are still associated with a phrasal accent cannot be definitely excluded even in cases where another constituent is explicitly focused. Vogel et al. (2016:134) allude to this possibility, which exists any time the target word is systematically varied in a metalinguistic carrier phrase while the rest of the phrase is held constant, as in their study.

3.3 Overall intensity

Non-frequency-dependent measures of intensity (e. g. mean, peak, midpoint) had similar success to F0 in their capacity to diagnose stress, functioning as a marker of stress in 75 % (39 of 52) of languages. In three studies encompassing two languages, Dobrovolsky (1999) on Chuvash, Lieberman (1960), Beckman (1986) on American English, the relevant intensity measure was the intensity integral, which incorporates duration.

A finding that casts doubt on the efficacy of overall measures of intensity, however, is the observation that few of the studies that controlled for phrase-level prominence found intensity to be a robust exponent of stress. Of the four languages (Greek, Hungarian, Spanish, and Turkish) in the Vogel et al. (2016) study, only Hungarian used mean intensity to distinguish stress in non-focused target words. Otherwise, only in Papiamentu (Remijsen and Van Heuven 2002) was mean intensity reliably associated with stress in a clearly defocused condition. Notably, though, Papiamentu is a tone language, in which, as mentioned earlier, F0 is less readily available for conveying stress distinctions. It is likely no coincidence that six of the seven tone languages in the database for which an overall measure of intensity was taken (all except Thai) employed intensity as a marker of stress.

3.4 Frequency-sensitive intensity

Of the 19 languages for which at least one study targeted a frequency-dependent intensity measure, 16 (84 %) used such a measure to differentiate stress levels, where the intensity of stressed vowels was weighted (in virtually all cases) toward higher frequencies in comparison to unstressed vowels. The exceptional languages in which spectral tilt was not an exponent of stress were Pitjantjatjara (Tabain et al. 2014), Peninsular Spanish (Ortega-Llebaria and Prieto 2010), and Brazilian Portuguese (Barbosa et al. 2013).

Studies differ considerably in how they quantify spectral tilt. Most studies (the unmarked case in the corpus) compare the relative intensity of different frequency bands in the spectrum as an index of stress, where the frequency of these bands varies across studies potentially contributing to differences between studies in results for the same language, e. g. Prieto and Ortega-Llebaria (2006) vs. Ortega-Llebaria and Prieto (2010) on Peninsular Spanish. On the other hand, Hussain (1997), Guion et al. (2010) ,[3]Garellek and White (2015), Caballero and Carroll (2015)[4] examine the relative intensity of the first two harmonics (H1-H2), which is typically analyzed as an index of voice quality (Gordon and Ladefoged 2001). Synthesizing the H1-H2 results across these studies suggests a pattern of increased breathiness in unstressed vowels relative to their stressed counterparts.

In summary, although spectral tilt is certainly a promising correlate of stress, the diversity of implementations makes it difficult to definitively establish its reliability relative to other potential markers of stress.

3.5 Formant frequency

The final measure assessed in several studies was formant frequency, most commonly the first (F1) and second (F2) formant, which can be interpreted as indices of centrality along the height dimension in the case of F1, reflecting degree of jaw opening (Erickson and Kawahara 2016), and backness in the case of F2, reflecting tongue dorsum advancement/retraction (Erickson 2002) dimensions. Typically, stressed vowels tend to be more peripheral than unstressed vowels, although there is a contrary effect observed in some languages whereby stressed vowels may be lower in the acoustic space (reflecting a lowered jaw position) than their unstressed counterparts *even if this entails a more central articulation*, e. g. in the case of high vowels (see Crosswhite 2004 on the typology of stress-related effects on vowel quality).

In interpreting the database results, formant frequency was classed as a reliable correlate of stress in a language if *either* the first or second formant reliably distinguished any phonemic vowels as a function of stress in one or more studies of the language. In 86 % (25 of 29) of languages for which formant data appeared vowel quality differed as a function of stress, though it should be mentioned that the database did not include studies of certain languages in the database that have been demonstrated in other work to have stress-induced vowel reduction, e. g. English (Lindblom 1963), Russian (Padgett and Tabain 2005), and Finnish (Wiik 1965).

In many of those languages in which vowel quality differed as a function of stress, the effect was limited to certain vowels and/or only one formant. The formant(s) and vowel qualities differentiated by stress varied from language to language making it difficult to draw any salient cross-linguistic generalizations about the phonetic nature of reduction other than the well-known tendency for stressed vowels to occupy a more peripheral vowel space than their unstressed counterparts. It is also noteworthy that none of the studies that controlled for phrase-level prominence found reliable differences in the first and second formant between stressed and unstressed syllables.

3.6 Relative efficacy of different cues

It is possible in principle to evaluate the relative effectiveness of different acoustic cues not only in aggregate across languages but also in languages that use multiple properties to distinguish stress levels. There are eight studies representing eleven different languages in the database that use statistical analyses, either logistic regression or linear discriminant analysis, to assess the relative capacity of different acoustic dimensions to predict stress level. Crucially, the estimations are based purely on production data and do not imply any perceptual weighting. In six of the eleven languages, an F0 property (either mean or change) was the most reliable predictor of stress level: Berinstein (1979) on K'ekchi, Garellek and White (2015) on Tongan, and Vogel et al. (2016) on Greek, Hungarian, Spanish, and Turkish. In the remaining five languages, duration emerged as the most predictive of stress: Potisuk et al. (1996) on Thai, Sluijter and van Heuven (1996) on Dutch, Remijsen (2002) on Ma'ya, Remijsen and Van Heuven (2002) on Papiamentu, and Silber-Varod et al. (2016) on Hebrew. The edge in favor of F0 becomes even greater if one excludes the three tone languages among the eleven, Ma'ya, Thai and Papiamentu, in all of which duration is a better predictor of stress.

The results of the linear discriminant analyses in the Vogel et al. (2016) study of Greek, Hungarian, Spanish, and Turkish demonstrate overall intensity and vowel quality to be relatively unreliable predictors of stress. It should be noted, however, that only two of the five studies that directly compared cues, Sluijter and van Heuven (1996) on Dutch and Remijsen (2002) on Ma'ya, incorporated a measure of spectral tilt, making it difficult to assess the efficacy of spectral tilt as a marker of stress relative to other acoustic properties.

4 Acoustic evidence for secondary stress

Although most studies in the survey evaluated only the acoustic distinction between primary stressed and unstressed syllables, there were 21 papers that also considered the acoustic evidence for secondary stress, a contentious issue in the stress literature for many languages, e. g. Polish and Estonian (see Hayes 1995 for these and other cases). Perhaps not surprisingly, evidence for secondary stress as distinct from both primary stress and lack of stress was less compelling in the survey than evidence for a distinction between primary stressed syllables and unstressed ones. In most studies, secondary stress was distinguished from other levels using only a subset of properties that were used to distinguish primary stress from lack of stress. Only two studies, Gordon (2004) on Chickasaw and Rietvald et al. (2004) on Dutch, distinguished secondary stressed syllables from both their primary stressed and unstressed counterparts along all the dimensions that differentiated primary stressed and unstressed syllables. Otherwise, secondary stressed syllables were neutralized with either primary stressed or unstressed syllables for at least one parameter that marked the contrast between primary stress and lack of stress. The most tenuous distinction in most cases was between secondary stress and lack of stress. Vowels claimed in the phonological literature to carry secondary stress were not different from unstressed vowels along any dimension in Erzya Mordvin (Lehiste et al. 2003), Pitjantjatjara (Tabain et al. 2014), Polish (Dogil 1999; Newlin-Łukowicz 2012), and Brazilian Portuguese (Barbosa et al. 2013). Similarly, the distinction between secondary stress and lack of stress in German (Kleber and Klipphahn 2006) was only evident for duration for only two (of six) vowel qualities and only for one or two (of six) speakers. In Ingrian Finnish (Gordon 2009), only slight lengthening of voiced onsets emerged as a potential cue to secondary stress as distinct from lack of stress, while F0, intensity, and lengthening of all onsets differentiated primary stressed syllables from unstressed syllables. Garellek and White (2015) find a similar pattern of stronger acoustic evidence for primary stress relative to secondary stress in their study of Tongan: in a linear discriminant analysis, they observe much higher classification rates for the primary stress vs. unstressed distinction than the secondary stress vs. unstressed difference (89.1 % vs. 64.5 %).

In summary, the search for secondary stress as a distinct level of prominence proved generally more elusive in the database than the diagnosis of primary stress, a finding that is consistent with the existence of several disputed cases of secondary stress in the phonological literature.

5 Acoustic correlates of stress and prosodic taxonomy

The database provides acoustic evidence for stress in a prosodically diverse set of languages. Evidence for stress emerged for languages with predictable phonological stress, both weight-sensitive stress, e. g. Chickasaw, Squamish, and strictly deliminative (primary) stress, e. g. Polish, Finnish, as well as those with robust phonemic stress distinctions, e. g. Russian, Hebrew, and with mixtures of phonemic and predictable stress, e. g. English, Spanish. For a few languages generally accepted to have stress, the consulted studies were too preliminary to offer compelling acoustic evidence of stress. For example, small studies of stress in Czech (Duběda 2006) and Lakhota (Cho 2006) failed to provide definitive corroboration of stress potentially due either to their confinement to a single potential correlate of stress, e. g. in Czech, or their small sample size, e. g. the single speaker examined in the Lakhota study. Presumably, future studies of these languages will provide more convincing evidence of stress.

Evidence for stress also emerged for tone languages, ranging from those with more canonical one-to-one mappings between syllables and tones, e. g. Thai and Pirahã, to those with more limited tone, i. e. pitch accentual, systems, e. g. Basque and Swedish. Not surprisingly, in languages with lexical tone contrasts, F0 typically played a subservient role in signaling stress.

Also included in the database were studies of some languages whose relationship to the tone-stress continuum is less clear. Recent literature has revealed the existence of some languages lacking evidence for either lexical tone or word-level stress. In these "intonation-only" systems, the most salient prosodic events are attributed to the intonation system in the form of phrasal tones realized at or near edges of prosodic phrases. Languages fitting this profile of having phrasal prosody rather than word-level stress include both some not appearing in the database, e. g. Korean (Jun 1993) and French (Jun and Fougeron 1995), as well as a few examined in studies considered here. One relevant case is Indonesian, which has traditionally been regarded as a language with word-level stress but whose membership in this prosodic category has more recently been questioned (see Goedemans and van Zanten 2007 for discussion). Goedemans and van Zanten (2007) show that the acoustic correlates of stress in Indonesian, which functions as a lingua franca for speakers with diverse native language backgrounds, diverge sharply based on the substrate language of the speaker. Thus, their speaker of Toba Batak, a language with clearly discernible stress distinctions in the acoustic domain, marks stress in Indonesian along multiple dimensions (duration, F0, and intensity), whereas their speaker of Javanese, another language lacking robust word-level stress, fails to signal stress through any of these acoustic properties. The results for their Toba Batak speaker parallel those for the non-Javanese speakers of Indonesian in the earlier Adisasmito-Smith and Cohn (1996) study, suggesting that Indonesian potentially lacks acoustic evidence for word-level stress independent of transfer effects associated with speakers from other languages with word stress.

Another language in the database that plausibly lacks both tone and word-level stress is Tashlhiyt. When controlling for phrase-level confounds, Roettger et al. (2015; see also Roettger; for a detailed analysis) find no evidence for consistent stress on the final syllable contra earlier results from Gordon and Nafi (2012). Yet another language in the survey lacking compelling evidence for word-level stress is Tamil, in which none of the potential acoustic correlates (duration, intensity, and F0) of stress emerged as reliable in Keane (2006) study.

In summary, although stress appears to be an acoustically manifested phonological property in both stress languages as well as in tone languages, its universal status (even in languages lacking lexical tone) remains to be corroborated.

6 Summary

Results of a survey of 110 studies of 75 languages indicate that a large number of parameters potentially signal stress, including duration (not just of the vowel but also the onset consonant), various F0 features, overall intensity, assorted frequency-weighted measures of intensity, and vowel formant frequencies. Studies vary considerably in which subset of these potential stress correlates are examined, making it difficult to establish which ones are most consistently cues to stress. Statistically, duration was the most reliable exponent of stress across languages, although all of the measured parameters succeeded in differentiating stress in the majority of languages for which they were assessed. In most studies that investigated secondary stress, it was distinguished from primary stress and/or lack of stress through only a subset of parameters differentiating primary stress from no stress.

This study thus offers a first cross-linguistic assessment of the relative robustness of different potential acoustic exponents of word stress. However, as remarked throughout the manuscript, the findings need to be considered in light of the methodology employed in the studies comprising the survey. Carefully evaluating experimental design choices and statistical analyses of the discussed studies (see Roettger and Gordon this volume) leads to a more conservative view of what the results can genuinely tell us about the phonetic manifestation of word stress.

Acknowledgments

The authors thank Shigeto Kawahara, an anonymous reviewer, and the Department of Phonetics in Cologne for their helpful comments and feedback on an earlier version of this manuscript. Any remaining errors or misconceptions are our own.

Notes

[1] The survey conflates as a single variety, non-Javanese Indonesian, the results of the Adisasmito-Smith and Cohn (1996) study of Indonesian based on the speech of a non-Javanese substrate speaker and the results for the Toba Batak substrate speaker of Indonesian in the Goedemans and van Zanten (2007) study (which also includes results for a Javanese substrate speaker).

[2] In a third tone language, Minto Tanana (Tuttle 1998), F0 has a marginal status as a stress correlate, only used to differentiate stress for short but not long vowels.

[3] Guion et al. (2010) also analyze H1-A2 (the intensity of the harmonic closest to the second formant).

[4] Garellek and White (2015) also take a measure of cepstral peak prominence (CPP) to assess the degree of periodicity in the signal.

References

Aasmäe, Niina, Pärtel Lippus, Karl Pajusalu, Nele Salveste, Tatjana Zirnask & Tiit-Rein Viitso. 2013. *Moksha Prosody*. (Mémoires de la Société Finno-Ougrienne 268), Helsinki: Suomalais-Ugrilainen Seura.

Abbasi, Abdul Malik. 2015. *Phonetic analysis of lexical stress in Sindhi*. Lahore, Pakistan: University of Management and Technology. PhD dissertation.

Adisasmito-Smith, Niken & Abigail C. Cohn. 1996. Phonetic correlates of primary and secondary stress in Indonesian: A preliminary study. *Working papers of the Cornell phonetics laboratory* 11. 1–16.

Astruc, Lluïsa & Pilar Prieto. 2006. Acoustic cues of stress and accent in Catalan. *Proceedings of 3rd International Conference on Speech Prosody, Dresden, Germany*.

Barbosa, Plínio, Anders Eriksson & Joel Åkesson. 2013. Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese. In E. L. Asu & Pärtel Lippus (eds.), *Nordic Prosody, Proceedings of the 6th Conference, Tartu 2012*, 97–106. Frankfurt am Main: Peter Lang.

Beckman, Mary. 1986. *Stress and non-stress accent*. Dordrecht: Foris.

Berinstein, Ava. 1979. A cross-linguistic study on the perception and production of stress. *UCLA Working Papers in Phonetics* 47. Los Angeles: UCLA.

Bishop, Judith. 2002. 'Stress accent' without phonetic stress: Accent type and distribution in Bininj Gun-wok. *Proceedings of 1st International Conference on Speech Prosody, Aix-en-Provence, France*.

Bolinger, Dwight L. 1958. A theory of pitch accent in English. *Word* 14. 109–149.

Bolinger, Dwight L. 1961. Contrastive accent and contrastive stress. *Language* 37(1). 83–96.

Bond, Dzintra. 1991. Vowel and word duration in Latvian. *Journal of Baltic Studies* 22. 133–144.

Borise, Lena. 2015. Prominence redistribution in the Aŭciuki dialect of Belarusian. *Formal Approaches to Slavic Linguistics* 24. (Accessed 19 May 2017 at https://www.nyu.edu/projects/fasl24/proceedings/borise_fasl24.pdf).

Bouchhioua, Nadia. 2008. Duration as a cue to stress and accent in Tunisian Arabic, Native English, and L2 English. *Proceeding of 4th International Conference on Speech Prosody, Campinas, Brazil*.

Caballero, Gabriela & Lucien Carroll. 2015. Tone and stress in Choguita Rarámuri (Tarahumara) word prosody. *International Journal of American Linguistics* 81. 459–493.

Caldecott, Marian. 2009. *Non-exhaustive parsing: Phonetic and phonological evidence from St'át'imcets*. University of British Columbia. PhD dissertation.

Chen, Chun-Mei. 2009. The phonetics of Paiwan word-level prosody. *Language and Linguistics* 10. 593–625.

Chiang, Wen-yu & Fang-mei Chiang. 2005. Saisiyat as a pitch accent language: Evidence from acoustic study of words. *Oceanic Linguistics* 44. 404–426.

Cho, Taehong. 2006. An acoustic study of the stress and intonational system in Lakhota: A preliminary report. *Speech Sciences* 13. 23–42. (Published by The Korean Association of Speech Sciences)

Crosswhite, Katherine. 2003. Spectral tilt as a cue to word stress in Polish, Macedonian, and Bulgarian. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*, 767–770.

Crosswhite, Katherine. 2004. Vowel reduction. In Bruce Hayes, Donca Steriade & Robert Kirchner (eds.), *Phonetically based phonology*, 191–231. New York: Cambridge University Press.

De Jong, Kenneth & Bushra Adnan Zawaydeh. 1999. Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics* 27. 3–22.

De Jong, Kenneth & Bushra Adnan Zawaydeh. 2002. Comparing stress, lexical focus, and segmental focus patterns of variation in Arabic vowel duration. *Journal of Phonetics* 30. 53–75.

Dobrovolsky, Michael. 1999. The phonetics of Chuvash stress: Implications for Phonology. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*, 539–542.

Dogil, Grzegorz. 1999. The phonetic manifestation of word stress in Lithuanian, Polish, German, and Spanish. In Harry van der Hulst (eds.), *Word prosodic systems in the languages of Europe*, 273–311. New York: Mouton de Gruyter.

Dubĕda, Tomáš. 2006. Intensity as a macroprosodic variable in Czech. *Proceedings of 3rd International Speech Prosody, Dresden, Germany*.

Erickson, Donna. 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* 59. 134–149.

Erickson, Donna & Shigeto Kawahara. 2016. Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistics Vanguard* 2(1).

Eriksson, Anders, Pier Marco Bertinetto, Mattias Heldner, Rosalba Nodari & Giovanna Lenoci. 2016. The acoustics of lexical stress in Italian as a function of stress level and speaking style. *Proceedings of Interspeech* 17. 1059–1063.

Eriksson, Anders & Mattias Heldner. 2015. The acoustics of word stress in English as a function of stress level and speaking style. *Proceedings of Interspeech* 16. 41–45.

Everett, Keren. 1998. The acoustic correlates of stress in Pirahã. *Journal of Amazonian Languages* 1(2). 104–162.

Fletcher, Janet & Nicholas Evans. 2002. An acoustic phonetic analysis of intonational prominence in two Australian languages. *Journal of the International Phonetic Association* 32. 123–140.

Fry, Dennis B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* 27. 765–768.

Fry, Dennis B. 1958. Experiments in the perception of stress. *Language and Speech* 1. 120–152.

Garellek, Marc & James White. 2015. Phonetics of Tongan stress. *Journal of the International Phonetic Association* 45. 13–34.

Goedemans, Rob & Ellen van Zanten. 2007. Stress and accent in Indonesian. *LOT Occasional series* 9. 35–62.

Gonzalez, Andrew. 1970. Acoustic correlates of accent, rhythm, and intonation in Tagalog. *Phonetica* 22. 11–44.

Gordon, Matthew. 1995. Acoustic properties of primary and secondary word-level stress in Estonian. *Poster presented at the 130th meeting of the Acoustical Society of America*. St. Louis.

Gordon, Matthew. 2004. A phonetic and phonological study of word-level stress in Chickasaw. *International Journal of American Linguistics* 70. 1–32.

Gordon, Matthew. 2009. Prominence and gemination in Ingrian. *Linguistica Uralica* 45. 81–100.

Gordon, Matthew & Ayla Applebaum. 2010. Acoustic correlates of stress in Turkish Kabardian. *Journal of the International Phonetic Association* 40. 35–58.

Gordon, Matthew & Peter Ladefoged. 2001. Phonation types: A cross-linguistic overview. *Journal of Phonetics* 29. 383–406.

Gordon, Matthew & Latifa Nafi. 2012. The acoustic correlates of stress and pitch accent in Tashlhiyt Berber. *Journal of Phonetics* 40. 706–724.

Gordon, Matthew & Françoise Rose. 2006. Émérillon stress: A phonetic and phonological study. *Anthropological Linguistics* 48. 132–168.

Guion, Susan, Jonathan D. Amith, Christopher S. Doty & Irina A. Shport. 2010. Word-level prosody in Balsas Nahuatl: The origin, development, and acoustic correlates of tone in a stress accent language. *Journal of Phonetics* 38. 137–166.

Hargus, Sharon. 2005. Prosody in two Athabaskan languages of Northern British Columbia. In Sharon Hargus & Keren Rice (eds.), *Athabaskan Prosody*, 393–423. Amsterdam: John Benjamins Publishing Company.

Hargus, Sharon & Virginia Beavert. 2006. A note on the phonetic correlates of stress in Yakima Sahaptin. *University of Washington Working Papers in Linguistics* 24. 64–95.

Harnud, Huha. 2003. Stress on Mongolian disyllabic words. *Proceedings of the XVth International Congress of Phonetic Sciences*, 2433–2436. Barcelona, Spain.

Hayes, Bruce. 1995. *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Hintz, Diane. 2006. Stress in South Conchucos Quechua: A phonetic and phonological study. *International Journal of American Linguistics* 72. 477–521.

Hualde, José Ignacio, Oihana Lujanbio & Francisco Torreira. 2008. Lexical tone and stress in Goizueta Basque. *Journal of the International Phonetic Association* 38. 1–24.

Huss, Volker. 1978. English word stress in the post-nuclear position. *Phonetica* 35. 86–105.

Hussain, Sarmad. 1997. *Phonetic correlates of lexical stress in Urdu*. Northwestern University. PhD Dissertation.

Hyman, Larry. 2006. Word-prosodic typology. *Phonology* 23. 225–257.

Hyman, Larry. 2014. Do all languages have word accent. In Harry van der Hulst (ed.), *Word stress: Theoretical and typological issues*, 56–82. New York: Cambridge University Press.

Jun, Sun-Ah. 1993. *The phonetics and phonology of Korean prosody*. The Ohio State University. PhD Dissertation.

Jun, Sun-Ah & Cécile Fougeron. 1995. The accentual phrase and the prosodic structure of French. *Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm*, 722–725.

Keane, Elinor. 2006. Prominence in Tamil. *Journal of the International Phonetic Association* 36. 1–20.

Kleber, Felicitas & Nadine Klipphahn. 2006. An acoustic investigation of secondary stress in German. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel* 37. 1–18.

Ladd, D. Robert. 2008. *Intonational phonology*. New York: Cambridge University Press.

Lehiste, Ilse. 1966. *Consonant quantity and phonological units in Estonian*. Bloomington: Indiana University Press.

Lehiste, Ilse, Niina Aasmäe, Einar Meister, Karl Pajusalu, Pire Teras & Tiit-Rein Viitso. 2003. *Erzya Prosody*. (Mémoires de la Société Finno-Ougrienne 245), Helsinki: Suomalais-Ugrilainen Seura.

Lehiste, Ilse, Pire Teras, Valts Ernštreits, Pärtel Lippus, Karl Pajusalu, Tuuli Tuisk & Tiit-Rein Viitso. 2008. *Livonian Prosody. (Mémoires de la Société Finno-Ougrienne* 255), Helsinki: Suomalais-Ugrilainen Seura.

Lehiste, Ilse, Pire Teras, Toomas Help, Pärtel Lippus, Einar Meister, Karl Pajusalu & Tiit-Rein Vittso. 2005. *Meadow Mari Prosody*. (*Linguistica Uralic Supplementary Series* 2). Tallinn: Teaduste Akadeemia Kirjastus.

Lesho, Marivic. 2013. *The sociophonetics and phonology of the Cavite Chabacano vowel system*. The Ohio State University. PhD dissertation.

Levi, Susannah V. 2005. Acoustic correlates of lexical accent in Turkish. *Journal of the International Phonetic Association* 35. 73–97.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2016. *Ethnologue: Languages of the world*, 19th edn. Dallas, TX: SIL International.

Lieberman, Philip. 1960. Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America* 32. 451–454.

Liiv, G. 1985. Akusticheskie korreliaty estonskogo slovesnogo udarenii v sootnoshenii s differentsial′noi dolgotoi. *Sovestskoe Finno-Ugrovedenie* 21(1). 1–13.

Lindblom, Björn. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35. 1773–1781.

Lindström, Eva & Bert Remijsen. 2005. Aspects of prosody of Kuot, a language where intonation ignores stress. *Linguistics* 43. 839–870.

Lippus, Pärtel, Karl Pajusalu & Pire Teras. 2006. The temporal structure of penta- and hexasyllabic words in Estonian. *Proceedings of* 3^{rd} *International Conference on Speech Prosody, Dresden, Germany*.

McDonnell, Brad. 2014. Acoustic correlates of prominence in Besemah (Malayic, Indonesia). Poster presented at the 5^{th} *Joint Meeting of the Acoustical Society of American and the Acoustical Society of Japan*. Honolulu, Hawai′i.

Moore, Brian C. J. 2013. *An introduction to the psychology of hearing*. Boston, MA: Brill.

Moroz, George. 2017. Lingtypology: linguistic typology and mapping. http://CRAN.R-project.org/package=lingtypology.

Newlin-Łukowicz, Luiza. 2012. Polish stress: Looking for phonetic evidence of a bidirectional system. *Phonology* 29(02). 271–329.

Oberly, Stacey. 2008. *A phonetic analysis of Southern Ute with a discussion of Southern Ute language policies and revitalization*. University of Arizona. PhD dissertation.

Ortega-Llebaria, Marta. 2006. Phonetic cues to stress and accent in Spanish. *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology*, 104–118.

Ortega-Llebaria, Marta & Pilar Prieto. 2010. Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Language and Speech* 54. 73–97.

Padgett, Jaye & Marija Tabain. 2005. Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica* 62. 14–54.

Plag, Ingo, Gero Kunter & Mareile Schramm. 2011. Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics* 39. 362–374.

Potisuk, Siripong, Jackson Gandour & Mary P. Harper. 1996. Acoustic correlates of stress in Thai. *Phonetica* 53. 200–220.

Pycha, Anne. 2006. A duration-based solution to the problem of stress realization in Turkish. *UC Berkeley Phonology Lab Annual Reports*.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Remijsen, Bert. 2002. Lexically contrastive stress accent and lexical tone in Ma`ya. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology VII*, 585–614. Berlin: Mouton de Gruyter.

Remijsen, Bert & Vincent Van Heuven. 2002. Stress, tone, and discourse prominence in the Curacao dialect of Papiamentu. *Phonology* 22. 205–235.

Rietvald, Toni, Joop Kerkhof & Carlos Gussenhoven. 2004. Word prosodic structure and vowel duration in Dutch. *Journal of Phonetics* 32. 349–371.

Riviera-Castillo, Yolanda & Lucy Pickering. 2004. Phonetic correlates of stress and tone in a mixed system. *Journal of Pidgin and Creole Languages* 19(2). 261–284.

Roettger, Timo B. accepted. Tonal placement in Tashlhiyt Berber – How an intonation system accommodates to adverse phonological environments. *Studies in Laboratory Phonology*. Berlin: Language Science Press.

Roettger, Timo B., Anna Bruggeman & Grice Martine. 2015. Word stress in Tashlhiyt – Postlexical prominence in disguise. *Proceedings of the 18th International Congress of Phonetic Sciences*. Hong Kong.

Roettger, Timo B. & Matthew K. Gordon. This issue. *Methodological issues in the study of word stress correlates*.

Rozelle, Lorna. 1997. The effect of stress on vowel length in Aleut. *UCLA Working Papers in Phonetics* 95. 91–101.

Sadeghi, Vahid. 2011. Acoustic correlates of lexical stress in Persian. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong*, 1738–1741.

Silber-Varod, Vered, Hagit Sagi & Noam Amir. 2016. The acoustic correlates of lexical stress in Israeli Hebrew. *Journal of Phonetics* 56. 1–14.

Simard, Candide, Claudia Wegener, Albert Lee & Connor Youngberg. 2014. Savosavo word stress: A quantitative analysis. *Proceedings of 7th International Conference on Speech Prosody, Dublin, Ireland*.

Sluijter, Agaath M. C. & Vincent J. van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100. 2471–2485.

Suomi, Kari, Juhani Toivanen & Riikka Ylitalo. 2001. On distinguishing stress and accent in Finnish. *Lund University Department of Linguistics. Working Papers* 49. 152–155.

Tabain, Marija, Janet Fletcher & Andrew Butcher. 2014. Lexical stress in Pitjantjatjara. *Journal of Phonetics* 42. 52–66.

Tamburri-Watt, Linda, Michael Alford, Jen Cameron-Turley & Carrie Gillon. 2000. Skwxwú7mesh (Squamish Salish) stress: A look at the acoustics of /a/and /u/. *International Conference on Salish (and Neighbo(u)ring) Languages* 35 (UBC Working Papers in Linguistics 3). 199–217.

Tuomainen, Jyrki, Stefan Werner, Jean Vroomen & Beatrice De Gelder. 1999. Fundamental frequency is an important acoustic cue to word boundaries in spoken Finnish. *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco*, 921–923.

Tuttle, Siri. 1998. *Metrical and tonal structures in Tanana Athabaskan*. University of Washington. PhD dissertation.

Tuttle, Siri. 2005. Duration, intonation and prominence in Apache. In Sharon Hargus & Keren Rice (eds.), *Athabaskan Prosody*, 319–344. Amsterdam: John Benjamins Publishing Company.

Vogel, Irene, Angeliki Athanasopoulou & Nadia Pinkus. 2016. Prominence, contrast, and the functional load hypothesis: An acoustic investigation. In Jeffrey Heinz, Rob Goedemans & Harry van der Hulst (eds.), *Dimensions of Phonological Stress*, 123–167. Cambridge: Cambridge University Press.

Wiik, Kalevi. 1965. *Finnish and English vowels: A comparison with special reference to the learning problems met by native speakers of finnish learning English*. Turku, Finland: Turun yliopisto.

Williams, Briony J. 1983. *Stress in modern Welsh*. University of Cambridge. PhD dissertation.

Williams, Briony J. 1999. The phonetic manifestation of stress in Welsh. In Harry van der Hulst (ed.), *Word prosodic systems in the languages of Europe*, 311–334. New York: Mouton de Gruyter.

Yakup, Mahire & Joan Sereno. 2016. Acoustic correlates of lexical stress in Uyghur. *Journal of the International Phonetic Association* 46. 61–77.

Journal of Phonetics 73 (2019) 1–7

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Special Issue: Emerging Data Analysis in Phonetic Sciences, eds. Roettger, Winter & Baayen

Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility

Timo B. Roettger[a,*], Bodo Winter[b], Harald Baayen[c]

[a] *Northwestern University, Department of Linguistics, United States*
[b] *University of Birmingham, Department of English Language and Applied Linguistics, United Kingdom*
[c] *University of Tübingen, Department of Linguistics, Germany*

ARTICLE INFO

Article history:
Received 6 April 2018
Received in revised form 10 December 2018
Accepted 11 December 2018

Keywords:
Data analysis
Statistics
Reproducibility
Open science
Null hypothesis significance testing
Bayesian modeling

ABSTRACT

This special issue introduces a series of papers that make available new methods to the phonetic and linguistic community and reflect upon existing data analysis practices. In our introduction, we highlight three themes that we consider pressing issues in data analysis and that run across the contributions to this special issue: the difference between exploratory and confirmatory analyses, different approaches to statistical inference, and the analysis of multidimensional multivariate speech data. Moreover, we provide a call for considering the importance of open and reproducible research practices, such as publishing one's data and analysis code. Rather than being dogmatic about particular statistical methods, the pluralism of analysis approaches in linguistics should excite debate and discussion, to which this special issue is an invitation. In addition, the co-existence of multiple ways of analyzing the same data (each with its own advantages and disadvantages and different analysis goals) makes it all the more important for researchers to make their research process open and accessible to other researchers.

Published by Elsevier Ltd.

1. Introduction

The landscape of data analysis in linguistics and other fields is constantly changing. Advances in computational power have made new analytical approaches possible, and the use of open access software such as R (R Core Team, 2013) increases the speed with which new statistical methods are shared both within our field and across disciplines. As accessibility to these methods increases, more and more people within linguistics employ increasingly complex analytical techniques. Parallel to the ever-growing toolkit of statistical methods, there are shifts in methodological traditions and statistical philosophies, with an array of differing views about how data should be analyzed, how it should be reported, and how it should be shared. In sum, the field of data analysis is in flux. Amidst the backdrop of changing practices, it is important to critically assess past practices, to reflect upon present practices, and to look out for what new developments will affect our future practices.

We approach data analysis with George Box's famous quote in mind, "all models are wrong, but some are useful" (Box, 1979, p. 2). This often-repeated quote embodies a fundamental truth about data analysis: We perform analyses to gain a better understanding of our world and the phenomena we investigate. Statistical models are thus supposed to be "useful". However, all models are also necessarily "wrong" to some extent, with each model providing only a snapshot of the underlying complexity of the phenomena to be modeled. Models can be "useful" in different ways and to differing degrees, and models can be more or less "wrong" as well. There is no single model that is the best model and that is equally useful across theories and phenomena. This very fact necessarily creates a *plurality* of analytical approaches, within and across disciplines. Even expert statisticians reach different conclusions when given the same dataset (Silberzahn et al., 2018). Rather than trying to provide gold standards and recipes, we endorse the plurality of approaches and highlight that pluralism calls for comparison, reflection, and a critical discourse about methods. We should not try to elevate any one method to the status of a "best" method or a canonical way of analyzing particular datasets; instead we should discuss the advantages and disadvantages of particular approaches openly.

In line with the idea of plurality, data analysis varies along important dimensions. We would like to highlight a few of these

[*] Corresponding author.
E-mail address: timo.roettger@uni-koeln.de (T.B. Roettger).

https://doi.org/10.1016/j.wocn.2018.12.001
0095-4470/Published by Elsevier Ltd.

dimensions to not only introduce the contributions to our special issue, but also to review what we conceive as important topics for data analysis in quantitative fields such as phonetics. In the following, we will discuss the distinction between exploratory and confirmatory analysis (Section 2); the differences between null hypothesis significance testing and Bayesian inference (Section 3); and analytical choices surrounding the multidimensionality of phonetic data (Section 4). Beyond reflecting on past and future methods it is also important to think about how data analyses are communicated and shared with the community. To this end, we will discuss the relevance of reproducibility and the benefits of an open and transparent phonetic community (Section 5), exemplified by the contributions in this special issue.

2. Exploratory vs. confirmatory data analysis

It is important to recognize that data analysis includes two stages which are more or less conceptually distinct, although they may overlap to considerable degrees in practice. In an *exploratory* stage, a researcher observes patterns and relationships leading to the generation of new hypotheses as to how these observations can be explained. This stage is a hypothesis-generating process. Many breakthroughs in science originate from the serendipity of researchers observing an unexpected pattern while exploring their data. In a *confirmatory* stage, novel hypotheses as well as hypotheses extending or challenging established theories are then pitted against new data, obtained in, for example, controlled experimental studies. This stage is a hypothesis-testing process. Putting our hypotheses under targeted scrutiny via confirmatory tests helps us to accumulate evidence in order to challenge, substantiate or revise established theories. The revised theories can then be further informed by additional exploration of the available data, leading to an iterative process that alternates between exploration and confirmation. Exploratory and confirmatory research should be considered complementary; both are necessary components of scientific progress. Moreover, both exploratory and confirmatory research should be guided by theory. An exploratory analysis does not have to be exclusively descriptive, but can, and often should be, tied in with specific linguistic theories.

The distinction between confirmation and exploration has large-scale consequences for research in the language sciences. It is important to realize that in an exclusively confirmatory setting, researchers have only one shot (Harrell, 2014), allowing for only a single theoretically motivated model to be fitted to the data. Subsequently, model criticism is carried out to clarify whether the resulting model is actually appropriate for the data. In a genuinely confirmatory analysis, there is no place for repeated modeling during data collection, no place for adding or removing interactions, and no place for including or removing control variables. As soon as a second model is fitted to a given dataset, the analysis is no longer confirmatory, but exploratory (see Baayen, Vasishth, Kliegl, & Bates, 2017, for further discussion).

Unfortunately, when it comes to publishing work, exploration and confirmation are not weighted equally. Confirmatory analyses have a superior status within the academic incentive system, determining the way funding agencies demand what proposals should look like, and shaping how we frame our papers. The prestige of confirmatory statistics is so high that occasionally the review process can force authors to recast the reporting of exploratory analyses in the format of the reporting of confirmatory analyses (see, e.g., Pham & Baayen, 2015, footnote 1). Whether due to publication pressure or not, the results of what has actually been an exploratory analysis are often presented as if they were the results of a confirmatory analysis. The prevalent expectation that the main results of a study should be predicted based on a priori grounds has led to harmful practices for scientific progress (John, Loewenstein, & Prelec, 2012).

Moreover, each analysis is characterized by a “garden of forking paths” (Gelman & Loken, 2013) or what Simmons, Nelson, and Simonsohn (2011) call “researcher degrees of freedom”. Some relevant researcher degrees of freedom for phonetic studies include what phonetic parameters are measured, how they are operationalized, what data is kept and what data is discarded and what additional independent variables are measured (for a discussion of researcher degrees of freedom in phonetics, see Roettger, 2019). This flexibility in conducting studies and analyzing data can, intentionally or unintentionally, lead to harmful practices such hunting for significant *p*-values, also known as *p*-hacking (see also Simmons et al., 2011) or HARKing “Hypothesizing After Results are Known” (e.g., Kerr, 1998).

Rather than discouraging exploratory analyses, they should be encouraged. The complexity of speech naturally means that we do not always have specific directed hypotheses for all aspects of the data. There are many interesting patterns to be discovered, and later confirmed on separate datasets. It is often the exploratory part of the analysis that we can learn the most from, especially with highly multidimensional data (see Section 4). However, while exploration is necessary, it has to be separated from confirmation. Each analysis needs to be clear about where it stands, i.e., the degree to which an analysis is confirmatory or exploratory needs to be explicitly stated. In particular, exploratory studies should be treated as such, rather than being re-framed as the results of a confirmatory analysis. More and more papers in our field acknowledge this important distinction and discuss confirmatory and exploratory analyses in different sections of their manuscripts, with the latter stressing the caveat that any generated hypotheses are waiting to be confirmed on new data (e.g., Baumann & Winter, 2018; Grice, Savino, & Roettger, 2018 for recent examples)

Researchers carrying out exploratory data analysis can to some extent protect themselves and their colleagues against spurious results by setting much more stringent alpha-levels when evaluating whether there are signals in the noise. In exploratory analysis, it is the researcher’s duty to launch adversarial attacks on potential effects, and to then only report those effects which survived such attacks consistently. If a strict null hypothesis significance testing approach is followed, confirmation cannot happen on the same dataset that was previously used as the basis for exploration. To the extent that confirmatory and exploratory analyses may blend into each other in actual practice, the researcher needs to be aware of this and report results accordingly.

3. Inferential frameworks: Frequentist and Bayesian inference

An important aspect of data analysis is making generalizable statements about observations. Inferential statistics is the process of using samples to make "inferences" for parameters of a population of interest. For example, a study may contain a subset of speakers from a linguistic community, and the sample is used to make inferences about all speakers of the language. Or a study may contain a subset of words from the language, and the sample is used to make inferences about all words of the language (see Clark, 1973). In statistics, there are various different approaches to making this inference, including frequentist and Bayesian statistics (e.g. Fisher, 1955; Gigerenzer, Krauss, & Vitouch, 2004; Dienes, 2008; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008; McElreath, 2016). Each approach has different analysis goals and makes different assumptions. Our special issue includes several papers that discuss aspects of different inferential frameworks as well as papers that make use of techniques and methods developed within each of these frameworks.

Classical methods for statistical inference (analysis of variance, discriminant analysis) are grounded in the work of Sir Ronald Fisher (1925). These methods, which are widely used in phonetics and many other fields of inquiry, are known as frequentist, as they are grounded in a particular understanding of the concept of probability, namely, the idea that the probability of an event is given by the limit of its relative frequency across a large number of trials. Fisher's method was later combined with Neyman and Pearson's approach to hypothesis testing (1928) to create what is now known as null hypothesis significance testing (NHST) (Gigerenzer et al., 2004; Lindquist, 1940). This framework became an extremely useful tool at a time in which computers did not exist, and has been used ever since across scientific disciplines. In traditional NHST, a researcher starts by assuming a null hypothesis (such as the absence of an effect) and gathers evidence against that initial assumption. The *p*-value measures the incompatibility of the data with the null hypothesis. It is often used as a hard cutoff, where an effect is accepted as "significant" if its associated *p*-value falls below a preset threshold probability. NHST provides a simple and specific decision procedure (using a particular threshold, such as $p < 0.05$) which will assure low error rates in the long run, across a series of repeated experiments.

The practice of NHST has been much criticized by researchers in many different disciplines (Gigerenzer, 2004; Goodman, 1999; Hubbard & Lindsay, 2008; Kline, 2004; Krantz, 1999; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Nickerson, 2000; Sterne & Smith, 2001; and many others). These criticisms surround, among other things, the practice of relying on an arbitrarily defined hard-cut threshold for "significance" (rather than taking the continuous strength of evidence into account), the practice of overly emphasizing point estimates (such as means) over interval estimates (such as confidence intervals or credible intervals) (e.g., Cumming, 2012, 2014), and the practice of not incorporating any prior knowledge into one's models and inferences.[1]

Frequentist inference, as introduced by Fisher, differs in many ways from what is now known as Bayesian inference. The field of statistics has a long history of a deep divide between classical frequentist statistics and Bayesian statistics, each camp having its own philosophical foundations and methodological goals (e.g., Fisher, 1955; Gigerenzer et al., 2004; Dienes, 2008; Wagenmakers et al., 2008).

There are different classes of Bayesian models, but one defining feature is that they quantify the degree to which a researcher needs to adjust their beliefs as a function of the researcher's prior beliefs and the data and model at hand. That is, Bayesian inference critically differs from other inferential approaches by incorporating so-called "priors", which are either defined by a priori assumptions about the measurement system, or estimated from previous research. For example, when estimating the difference in duration between two vowel categories, the researcher can incorporate priors which reasonably rule out durational values below zero and above one hour. As opposed to that, in standard frequentist inference, all parameter values are assumed to be equally likely. Bayesian inference makes it also possible for the analyst to include knowledge outside of the present data when modeling new data, e.g. estimated from previous research. The Bayesian paradigm has, intrinsically, a much more cumulative perspective on the gathering of scientific evidence and offers much more sophisticated tools for integrating knowledge across multiple studies.

That the need for clear 'gold standards' is still felt today is exemplified by the paper by Barr, Levy, Scheepers, and Tily (2013) on how to fit mixed models. With the wide spectrum of analytical techniques currently available, which will also increasingly include methods from machine learning, it is not possible nor desirable to enforce rules by means of which significance can be assessed mechanically. A spirit of plurality is needed that creates space for realizing that there are problems and applications that might be handled more easily by either Bayesian or frequentist approaches, and that the analyst is far better off having both tools in their toolbox. In particular, researchers need to be familiar with both approaches, since there is an increasing number of papers in quantitative linguistics that uses Bayesian approaches.

There are three papers in our special issue that focus on the merits and pitfalls of different statistical philosophies (such as NHST versus Bayesian inference). Vasishth et al. (2018, this collection) give an extended overview of the logic and benefits of standard Bayesian analyses and walk the reader through a concrete standard Bayesian analysis of an acoustic study, investigating whether and how voice onset time measurements discriminate different stop series across three different languages. Their paper provides a useful introduction to Bayesian data analysis in linguistics and offers annotated code to facilitate the implementation of Bayesian modeling.

In a second paper, Nicenboim, Roettger, and Vasishth (2018, this collection) investigate the phenomenon of incomplete neutralization of German final devoicing using Bayesian meta-analysis. Incomplete neutralization is a particularly valuable phenomenon to discuss methodologically, because the available evidence has been the subject of heated methodological debates (Fourakis & Iverson, 1984; Port & O'Dell, 1985; Roettger, Winter, Grawunder, Kirby, & Grice, 2014;

[1] We note here that classical 'frequentist' inference is not necessarily or intrinsically focused on hard cut-offs, which is only a particular interpretation of this framework (see Perezgonzalez, 2015).

Winter & Roettger, 2011). According to some researchers, the German voicing contrast is completely devoiced in final position; according to others, the devoicing is phonetically “incomplete”. A number of studies in this literature do not allow adequate statistical inferences because the sample sizes are too small, and hence accumulating the evidence *across* studies in a meta-analysis becomes crucial to establish whether incomplete neutralization effects are robust.

The controversial topic of incomplete neutralization is also explored in another paper that addresses issues in statistical inference. Kirby and Sonderegger (2018, this collection) look at the role of sample size in being able to estimate the incomplete neutralization effect accurately. Their numerical simulations suggest that linguists need to pay more attention to statistical power (the probability that a significance test will correctly reject a false null hypothesis) in designing experiments. Small sample sizes come with unrealistic expectations of replicability of the effect direction and magnitude (e.g. Vasishth, Mertzen, Jäger, & Gelman, 2018, for a recent discussion). Besides making important points about experimental design in phonetics, Kirby and Sonderegger (2018, this collection) demonstrate the utility of performing power simulations.

We want to stress here that including papers on either NHST or Bayesian inference is not intended to suggest that analyzing data within either of these frameworks is right or wrong. Given the prevalence of the decision procedure of NHST within phonetics, we think that it is most prudent at this stage to be aware of the opportunities offered by Bayesian and frequentist approaches. Moreover, learning about standard Bayesian methods may also help clarify misunderstandings about NHST (see Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Wagenmakers, Morey, & Lee, 2016; Morey et al., 2016; Nicenboim et al., 2018).

4. Dealing with the multidimensionality of speech communication

Our choice of data analysis varies tremendously as a function of the way phenomena are observed and measured. Depending on how observations are operationalized, certain analytical tools may or may not apply. Speech is inherently multidimensional and varies across time, as is the case with pitch curves, formant trajectories or articulatory gestures. These time-series data can be analyzed as a sequence of static landmarks (“magic moments”, Vatikiotis-Bateson, Barbosa, & Best, 2014) or as continuous trajectories, depending on how relevant the dynamic nature of speech behavior is for any given theory (Mücke, Grice, & Cho, 2014).

This special issue includes an introduction to Generalized Additive Models (GAMs), which are an extension of the classical generalized linear model (GLM) that enjoys wide use within phonetics (e.g., multiple regression, logistic regression, linear mixed effects models). Even traditional tests, such as *t*-tests and ANOVAs are approaches that can be re-expressed in a regression framework, in which case they yield equivalent results (if appropriately specified). GAMs extend GLMs with methods for modeling smooth nonlinear functions between a response and one or more predictors (Winter & Wieling, 2016; Wood, 2006). They also offer tools for addressing autocorrelations in the residual error, which are often present in time-series data, i.e. when observations are ordered in time, current observations may depend on previous observations.

Wieling (2018, this collection) introduces GAMs and offers a step-by-step tutorial based on an analysis of articulatory data (for other introductions, see, e.g., Winter & Wieling, 2016, and Baayen et al., 2017). As with any new tool, it is not always clear what the best approach to using this tool is from the outset. This was the case with linear mixed effects models, which incited a prominent debate about what the best random effects structure for the analysis of experimental designs is (see Barr et al., 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The flexibility inherent to statistical modeling is amplified in the case of GAMs, which provide many more options to their users. Wieling discusses some of these options.

An important complementary aspect of multidimensionality is tackled by Tomaschek, Hendrix, and Baayen (2018, this collection). They deal with a common problem in regression analyses (and by extension mixed models, GAMs etc.), namely, the issue of collinearity. When predictor variables in a model are highly correlated, estimates of parameters may become unstable and researchers can easily draw the wrong conclusions based on their data. Collinearity is an important problem that is often overlooked. As stated by Zuur, Leno, and Elphick (2010, p. 9): “If collinearity is ignored, one is likely to end up with a confusing statistical analysis in which nothing is significant, but where dropping one covariate can make the others significant, or even change the sign of estimated parameters.” Tomaschek et al. provide a critical discussion of three methods developed specifically for the analysis of data sets with many correlated predictors - regularization with the elastic net, regularization with supervised component regression, and random forests - each of which has its strengths and weaknesses, depending on the goals of the analysis.

Plummer and Reidy (2018, this collection) discuss another issue related to the multidimensionality of phonetic data analysis. They discuss a method for computing low-dimensional representations of speech which centers on the use of Laplacian Eigenmaps to build structures over data points from which low-dimensional representations of speech are learned. This technique enables researchers to reduce the multidimensional acoustic signal to lower dimensionality, which, as they argue, is a better proxy of cognitive and social speech categories.

Another aspect of multidimensionality is tackled by Danner, Barbosa, and Goldstein (2018, this collection), discussing topics related to the non-verbal context in which speech occurs. Speech communication is accompanied by changes in body posture, head position, gaze, facial expressions, and manual gestures (Goldin-Meadow, 2003; Kendon, 2004; McNeill, 1992). Danner et al. invite the reader to rethink how to characterize multimodal speech by applying dynamic approaches already used in speech research to multimodal communication. They discuss both the problem of automatically identifying visual gestures in video images, as well as the problem of correlating a gestural data stream with an acoustic data stream.

The papers discussed so far are either focused on the merits and pitfalls of different statistical philosophies (such as NHST versus Bayesian inference), or they discuss various new methods that are useful for different phonetic applications. Another strand that runs across the entire special issue is the issue of *reproducibility*. Reflecting on methods does not end with choosing a particular method, but it also includes thinking about how data analyses are communicated and shared with the community.

5. Towards reproducible phonetic sciences

To assess the strength of evidence for a theory, one needs to consider how the data were collected and how they were analyzed. Evaluating the strength of evidence becomes very difficult if part of the research process is not transparent. Reproducible research involves the capacity of other researchers (who have not conducted the original study) to repeat the analysis that is presented in a published study (see Peng, 2011; Munafò et al., 2017). Reproducibility minimally necessitates that both the data (either raw data or data tables) and the analysis code are made available to the community (if this is possible). Following recent calls for more transparent scientific practices (e.g. the Open Science Framework, see Nosek, 2017), we want to reiterate the plea for more reproducibility within phonetics in particular, and within linguistics more generally.

For our field, reproducible research has numerous advantages. First, as mentioned above, even expert data analysts will perform different analyses based on the same dataset (Silberzahn et al., 2018). Naturally, different analysis choices yield different conclusions (Roettger, 2019; Gelman & Loken, 2013; Simmons et al., 2011). McElreath (2016) emphasizes that statistical modeling is subjective, in the sense that it incorporates the researcher's beliefs and assumptions about a study system. Because of its inherent flexibility and subjectivity, the only way to allow evaluation of the process of statistical modeling by outsiders is to make it open.[2] Transparency then allows other researchers to draw their own conclusions based on the same dataset, reanalyze other aspects of them etc.

From a practical stand point, sharing materials, data, and code publicly has several applied advantages (Houtkoop et al., 2018). For example, data sharing has been associated with a citation benefit (Piwowar & Vision, 2013). Moreover, sharing data on online repositories can be a safeguard against 'scooping' (Houtkoop et al., 2018) since a researcher can claim precedence for a dataset or an analysis before a paper is published. In addition, permanently accessible repositories protect against data loss and link rot. Open research practices have furthermore shown to increase visibility, as well as to increase the number of opportunities for funding, jobs, and collaborations (McKiernan et al., 2016). If we make our materials and code available, the next research group (or our own) might have an easier time to replicate our experiment or extend our findings without duplicating efforts. This saves valuable resources and allows for a more rapid advancement of our field.

Publishing the data and code also facilitates knowledge transfer: Other researchers can learn from the ways a particular dataset was analyzed, and how the analysis was implemented in actual software code. It is within the spirit of sharing knowledge and being transparent, that all authors of this special issue make their code and data available on public repositories, allowing the readership of the special issue to readily implement the methods, as well as to actively participate in the discourse that surrounds the methods presented here. Reproducibility runs as a prominent thread through all of the papers in this special issue. All papers in this special issue contain links to publicly available repositories.

Many of the papers are written in a tutorial-like way, inviting the reader to reproduce and extend the offered analyses (Jadoul, Thompson, & de Boer, 2018; Vasishth et al., 2018; Wieling, 2018). For example, Politzer-Ahles and Piccinini (2018, this collection) discuss ways to visualize the results of hierarchical models that allows one to communicate the population-level estimates alongside the random variation associated with crossed random effects. Data visualization is an important aspect of communicating research findings and has been the subject of ongoing debates across scientific fields (e.g., Tufte, 1990; Kosslyn, 2006; Weissgerber, Milic, Winham, & Garovic, 2015). Politzer-Ahles and Piccinini's paper not only serves as a reminder of the importance of data visualization in communicating data and the results of statistical models; the inclusion of their scripts allows other users to apply them to new datasets.

The topic of reproducibility is also a prominent theme for Jadoul et al. (2018, this collection). As argued by many proponents of reproducible research, *all* aspects of the research workflow interact with reproducibility, not just the "final" data analysis stage. For example, in acoustic analyses, there are many degrees of freedom as to what acoustic parameters to extract and how, such as the settings used for the measurements of a particular speaker's fundamental frequency. We usually perform these analyses in available software such as Praat (Boersma & Weenink, 2018). However, data extraction in Praat is usually detached from subsequent statistical analyses. To streamline these processes, automated techniques can be used, for which Jadoul et al. (2018) propose a new toolkit, Parselmouth, which integrates the extraction of Praat-based acoustic analysis into a Python-based workflow. For users of Python, this allows the combination of acoustic and statistical analyses within one and the same script and may make acoustic analysis using Praat functionalities accessible. For those who currently use Praat, Parselmouth may provide a useful alternative to streamline the process of acoustic analysis and integrate it into a more reproducible workflow.

Taken together, the papers in this volume contribute to our mutual resources by introducing new tools, novel ways of analyzing our data, and by critically evaluating past, present and future analytical practices. Because all authors publicly share their materials, data, and code, they significantly contribute to our shared knowledge and facilitate future research. Aiming at increasing reproducibility has not only practical benefits for individual researchers, but it also benefits us as a collective scientific field, enabling us to access new methods and helping us to substantiate our findings.

[2] At present, many of the descriptions of statistical methods found in phonetics papers do not allow reproducing the performed analysis; in some cases, it is not even clear what general analysis was conducted (e.g., *p*-values may be listed without a detailed description of the associated statistical models these values are based on). For example, Winter (2011) tried to assess how often the independence assumption is violated in speech production data and found that many publications in phonetics do not provide enough information to allow such an assessment. This issue, common in all quantitative sciences, prevents the statistically minded readers to reproduce the analysis and does not allow proper evaluation of the presented evidence.

6. Conclusions

To conclude, we want to emphasize the spirit with which this special issue was conceived. As statistics is constantly evolving within and outside of linguistics and phonetics, there is a plurality of different analysis approaches. Many analytical philosophies alongside methodological tools and techniques co-exist alongside each other at any given point. In many ways, this is advantageous, as this creates the opportunity for discovery of new methods, many of which come from other fields, as well as the opportunity for honest discussion of the advantages and disadvantages of existing approaches. We are in no position, and nor is it our intention, to "police" any existing practices, or to provide recipes or guidelines that everybody should adhere to. Any strict rule will prove to be obsolete in the constantly changing landscape of statistical analysis. Instead, we want to invite the community to reflect on existing practices, as well as to look ahead to incorporate new analysis methods. Instead of accepting any of these techniques as absolute, we have to continue the methodological debate as a community. Moreover, by becoming increasingly reproducible, we can ensure that this plurality of methods benefits our common scientific goal, to understand the physical, cognitive, and social aspects of human speech communication.

References

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv*, preprint arXiv:1506.04967.

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics, 70*, 20–38.

Boersma, P. and Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics, 1*, 201–236.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics, 71*, 268–283.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London: Palgrave Macmillan.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B (Methodological), 17*(1), 69–78.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical proceedings of the Cambridge philosophical society* (Vol. 22, No. 5, pp. 700–725). Cambridge: Cambridge University Press.

Fourakis, M., & Iverson, G. K. (1984). On the 'incomplete neutralization'of German final obstruents. *Phonetica, 41*(3), 140–149.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine, 130*(12), 995–1004.

Grice, M., Savino, M., & Roettger, T. B. (2018). Word final schwa is driven by intonation – The case of Bari Italian. *The Journal of the Acoustical Society of America, 143*(4), 2474–2486.

Harrell, F. E. (2014). *Regression modeling strategies* as implemented in R package 'rms' version, 3(3). Berlin: Springer.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164.

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E. J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 70–85.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*(1), 69–88.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics, 71*, 1–15.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Kirby, J., & Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics, 70*, 70–85.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford: Oxford University Press.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*(448), 1372–1381.

Lindquist, E. F. (1940). *Statistical Analysis in Educational Research*. Boston, MA: Houghton Mifflin.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman & Hall/CRC Press.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., & Spies, J. R. (2016). Point of view: How open science helps researchers succeed. *Elife, 5*, e16800.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103–123.

Mücke, D., Grice, M., & Cho, T. (2014). More than a magic moment–Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics, 44*, 1–7.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20*, 175–240.

Nicenboim, B., Roettger, T. B., & Vasishth, Shravan (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics, 70*, 39–55.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.

Nosek, B. A. (2017). Center for Open Science: Strategic Plan. Open Science Framework. August 1. doi:10.17605/osf.io/x2w9h.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334* (6060), 1226–1227.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 223.

Pham, H., & Baayen, H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience, 30*(9), 1077–1095.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1* e175.

Plummer, A. R., & Reidy, P. F. (2018). Computing low-dimensional representations of speech from socio-auditory structures for phonetic analyses. *Journal of Phonetics, 71*, 355–375.

Politzer-Ahles, S., & Piccinini, P. (2018). On visualizing phonetic data from repeated measures experiments with multiple random effects. *Journal of Phonetics, 70*, 56–69.

Port, R. F., & O'Dell, M. L. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics, 13*, 455–471.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., & Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics, 43*, 11–25.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic sciences. *Journal of the Association for Laboratory Phonology.* Accepted for publication.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359–1366.

Sterne, J. A., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? *Physical Therapy, 81*(8), 1464–1469.

Tomaschek, F., Hendrix, P., & Baayen, H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71*, 249–267.

Tufte, E. R. (1990). *Envisioning information*. Cheshire: Graphics Press.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175.

Vasishth, S., Nicenboim, B., Fangfang, L., Kong, E., Beckman, M. E., & Edwards, J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics, 71*, 147–161.

Vatikiotis-Bateson, E., Barbosa, A. V., & Best, C. T. (2014). Articulatory coordination of two vocal tracts. *Journal of Phonetics, 44*, 167–181.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Springer.

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25*(3), 169–176.

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biology, 13*(4) e1002128.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116.

Winter, B. (2011). Pseudoreplication in phonetic research. In *Proceedings of the international congress of phonetic science* (pp. 2137–2140).

Winter, B., & Roettger, T. (2011). The nature of incomplete neutralization: Implications for laboratory phonology. *Grazer Linguistische Studien, 76*, 55–74.

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution, 1*(1), 7–18.

Wood, S. (2006). *Generalized additive models: An introduction with R*. CRC Press.

Zuur, A. F., Leno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14.

5 The intonation of questions and contrastive focus in Tashlhiyt

5.1 Introduction

Cross-linguistically, the flagging of questions and the marking of contrastive elements represent communicative functions that are prototypically expressed by means of morphosyntactic devices such as word order or morphological marking. In addition to morphosyntax, these functions can be expressed via prosodic structure and/or intonation. The present chapter will investigate the intonational marking of questions and contrastive statements in Tashlhiyt. We will show that under certain circumstances, both functions are expressed by similar phonetic parameters: questions and contrastive statements are characterised by a rise-fall in pitch. However, even though the pitch movements are very similar in some contexts, there are clear distributional properties and acoustic correlates that systematically distinguish the tonal events in questions from the ones in contrastive statements.

The chapter is structured as follows. After a brief introduction to linguistic aspects of flagging questions and marking contrastive elements (§5.2), common cross-linguistic strategies employed to express these functions prosodically and intonationally will be discussed (§5.3). Morphosyntactic constructions and the intonational expression of flagging questions and marking contrastive elements in Tashlhiyt are then described based on qualitative observations (§5.4). Subsequently, a production study will be presented that provides evidence for both global pitch parameters and local tonal events being used to distinguish questions from corresponding contrastive statements. Moreover, evidence will be presented that the location of tonal events is determined by several interacting factors (§5.5). A perception study will be presented that demonstrates the perceptual relevance of the identified cues (§5.6).

5.2 Theoretical background: questions and focus

5.2.1 Questions: a working definition

The distinction between questions and statements is usually defined by a bundle of properties associated with different linguistic levels of description. As a result, the term 'question' is used ambiguously in the literature. First, question can refer to a syntactically defined interrogative structure, irrespective of its pragmatic function. An example is the English yes-no question which is morphosyntactically distinguished from a corresponding statement by auxiliary fronting (cf. 1 and 2 below). Second, the term question can also refer to an utterance that functions as a request for information, regardless of its morphosyntactic form. An example is the declarative question in English, which exhibits a declarative syntactic structure (e.g. (3b) below). Following Bartels (1999) (after Lyons 1977, and Jacobs 1991), the term 'question' shall be used here as a purely functional term referring to utterances that "convey perceived relative lack of information […] regarding a relevant aspect of propositional content" (Bartels 1999: 9). In turn, statements are defined as utterances that lack such speaker uncertainty. The following study is limited to two question types: yes-no questions and declarative questions.

One of the most common and most often discussed question types is the 'yes-no question' (henceforth: y/n question, also known as 'polar question'). It is considered the most basic question type and a near universal across languages (Sadock & Zwicky 1985). Compared to a corresponding statement (cf. 1), y/n questions in English can be marked by auxiliary fronting (cf. 1 and 2).

(1) Helen ate the chocolate.

(2) Did Helen eat the chocolate?

(3) a. I ate the chocolate.
    b. You ate the chocolate?

Another closely related type of question is the declarative question. This question type is morphosyntactically identical to corresponding statements and is often reported to be distinguished from statements by intonation only (Haeseryn et al. 1997). Henceforth the term declarative question will be used to refer to requests for information with declarative syntactic structure. This includes both declarative questions and echo questions. The latter are repetitions of either an entire preceding utterance or parts of it, and express either surprise, disbelief or lack of comprehension, indicating that the proposition the question refers to is unexpected or inappropriate. Echo questions often ask for confirmation or clarification, and can be described as having a bias towards a negative response (cf. 3).[1]

5.2.2 Focus: a working definition

A pragmatic notion which is orthogonal to the distinction between questions and statements and which is cross-linguistically frequently expressed by intonation is 'focus'. Krifka (2008) defines focus as the marking of elements to indicate that alternatives to these elements are relevant for the interpretation of the utterance. Consider examples (4) and (5).

(4) Helen ate [the milk chocolate].

(5) [Helen] ate the milk chocolate.

The statements are morphosyntactically identical and denote the same proposition but differ with regard to which argument is in focus. In (4) the milk chocolate is in focus, implying that there are alternatives to the milk chocolate that are relevant for the interpretation of the utterance. The part of the utterance that is not in focus is often referred to as 'background' (Lambrecht 1996). In (5) Helen is in focus, implying that there are alternatives to Helen that are relevant for the interpretation of the utterance. Here, the milk chocolate is in the background.

In (4) a friend and I may have been talking about Helen's eating habits and her recent diet, which is mainly based on vegetables and fruits. Yesterday, Helen cheated and, instead of eating a banana, she ate the milk chocolate that I had bought for myself. The milk chocolate and the explicitly or implicitly mentioned alternatives are relevant for interpreting this particular statement here. I contrast the milk chocolate with fruits, for example, by means of focusing the milk chocolate. Alternatively, we may have been talking about the milk chocolate bars in my cupboard. One package disappeared and I ask my friends who took it. The answer in (5) is a compatible response here, since it signals that Helen did it and not someone else. Focus can, moreover, differ with respect to the size or scope of the focus domain. Consider example (5) again, here repeated as (6). It is a compatible answer to all of the following questions in (7):

(6) Helen ate the milk chocolate.

[1] Another common question type is the wh-question, which will be briefly discussed in Chapter 7 (see Bruggeman, Roettger & Grice 2017, for a first exploration of wh-questions in Tashlhiyt).

(7) a. What happened? [Helen ate the milk chocolate].
b. What did Helen do? Helen [ate the milk chocolate].
c. What did Helen eat? Helen ate [the milk chocolate].
d. Did Helen eat the banana? Helen ate [the milk chocolate].
e. Did Helen eat the white chocolate? Helen ate the [milk] chocolate.

Question (a) elicits whole-sentence focus (also referred to as 'broad focus') without pragmatically singling out a specific element in the utterance. Questions (b-e) elicit 'narrow focus' either on the verb phrase (b), the entire compound in object position (c and d), or the modifier of this compound (e). The answers differ with respect to the set of alternatives. An answer to (a) and (b) can either be described as lacking a juxtaposition of alternatives, or as being an alternative to another holistic proposition. In (c) the milk chocolate, on the other hand, contrasts with an open set of alternatives (all possible things Helen could have eaten). In an answer to questions (d) or (e), the focused constituent is explicitly contrasted with the alternative in the question (banana / white chocolate). The examples in (d) and (e) are a specific type of narrow focus, which is referred to as 'corrective focus' or 'contrastive focus'. Focus types such as contrastive focus can be marked by morphosyntactic devices such as word order or focus particles. They can also be marked by intonation or prosodic structure only.

5.3 The intonation of questions and focus

Many languages express both questions and contrastive focus by means of intonation. In some cases, the intonational parameters used to mark these functions look very similar, differing only in subtle ways. Thus, comparing these two functions is a promising departure point from which to gain an understanding of the intonation system of Tashlhiyt. The following section will give an overview of possible intonational devices employed to express these functions in other languages.

5.3.1 The intonation of questions

Languages have been frequently reported to distinguish questions from corresponding statements by means of 'global' or 'local pitch scaling' or certain tonal events. Global scaling, also referred to as 'pitch register' (Ladd 2008), involves the lowering or raising of phrase-length contours. If both the lowest and highest points of a contour are raised, there is an increase in 'pitch level'; if the lowest point is lowered and the highest point is raised, there is an increase in 'pitch span'. Local scaling involves a specific tonal event, such as a rise or a fall, and typically involves a local increase in pitch span also referred to as 'pitch excursion', i.e. there are lower lows and higher highs for the respective tonal event (Ladd 2008).

Cross-linguistically, both local and global pitch scaling have often been reported to be significantly different for questions and statements with questions generally exhibiting higher pitch (among others, American English: Hirst & Di Cristo 1998; Hausa: Inkelas & Leben 1990; Hawaiian: Murphy 2013; Mandarin Chinese: Shen 1990; Moroccan Arabic: Benkirane 1998; and Vietnamese: Brunelle, Phuong Ha & Grice 2012; cf. Haan 2002, for an overview). This difference in pitch scaling may be expressed in terms of different parameters: Finnish has been reported to exhibit higher initial pitch values in questions than in statements (Iivone 1998). Some languages have higher pitch peaks in questions than in corresponding statements, e.g. Swedish and Moroccan Arabic (Hadding-Koch & Studdert-Kennedy 1964; Gårding 1983; Benkirane 1998). In Hausa, the last lexical high tone in the utterance is raised in questions (Inkelas & Leben 1990). Bengali has been reported to have both raised pitch peaks as well as greater pitch excursions for the corresponding rises in questions (Hayes & Lahiri 1991).

Scaling differences have also been found to be relevant in perception. In their seminal study on English, Hadding-Koch & Studdert-Kennedy (1964) showed that listeners were more likely to rate a sentence as a question where there was higher f0 at three reference points (accent peak, post accentual low, and end of the phrase). Subsequent studies have consistently shown that greater pitch excursion in a rise-fall is more frequently perceived as a question than as a statement for a variety of languages, such as Hungarian (Gósy & Terken 1994), Swedish (House 2003), Russian (Makarova 2007), and Bari Italian (Savino & Grice 2007).

In addition to distinguishing questions from statements, scaling differences may cue different question types, too. In Bari Italian, y/n questions and echo questions are expressed by a rise-fall in pitch at the end of the phrase. Savino & Grice (2011) compared y/n questions with echo questions, which serve as an objection towards the repeated proposition. They found that echo questions exhibit higher pitch peaks than y/n questions.[2] This difference has been attributed to the pragmatically different functions associated with these question types. While echo questions exhibit a strong negative bias towards the questioned proposition in their corpus (listeners do not expect the proposition to be true), y/n questions

[2] Typologically, Italian is particularly remarkable because it lacks interrogative morphosyntax in y/n questions.

are rather neutral with regard to the expected answer. In a subsequent perception study, Savino and Grice showed that listeners can reliably make a categorical identification of these two intonational forms.

Apart from differences in scaling, a cross-linguistically common pattern setting questions apart from statements includes a sharp final rise at the end of the utterance in questions (reported for different types of questions, including y/n questions and echo questions; cf. Bolinger 1978). Another pattern that has been found to be relatively common across languages is a final rise-fall. One important parameter in this contour is the 'timing' of the pitch peak and, importantly, the rise up to this peak. In Palermo Italian, for example, the pitch peak occurs at a position in the intonation phrase that is structurally salient: the head of the intonation phrase, where it is part of the nuclear pitch accent that marks the element as pragmatically relevant (Grice 1995). Thus, the rise is on the syllable with the highest metrical strength, rather than at the edge of the phrase. The subsequent fall is realised at the edge. Comparable contours have been observed in other varieties of Italian (for an overview see Grice et al. 2005; Savino & Grice 2011), as well as in other languages such as Bengali (Hayes & Lahiri 1991), Bulgarian (Grice et al. 1995), Greek (Arvaniti 2001; Arvaniti & Ladd 2009), Hungarian (Ladd 1983; Gósy & Terken 1994; Varga 2002), Russian (e.g., Makarova 2007), and Moroccan Arabic (Benkirane 1998).

5.3.2 The intonation of focus

Rise-fall contours flagging questions often resemble tonal events marking focus. A number of discrete as well as continuous prosodic and intonational mechanisms have been found to be used cross-linguistically for marking focus. These include the presence, type, and timing of tonal events, phrasing as indicated by non-tonal boundary phenomena such as pauses and final lengthening, and spatio-temporal expansion of the segments involved (e.g. Buring 2009, for a typological overview of grammatical devices to mark focus). Many languages mark focused constituents using pitch accent type and alignment, including English (Pierrehumbert 1980; Pierrehumbert & Hirschberg 1990; Jun 2005), Italian (Grice et al. 2005), Russian (Meyer & Mleinek 2006), and German (Grice, Baumann & Benzmülller 2005; Grice et al. 2017). For example, Grice et al. (2017) showed that German speakers predominantly use early peak falling pitch accents to mark broad focus, while contrastive focus is predominantly marked by late peak rising pitch accents. In addition to pitch accents position and type, other intonational channels may be exploited to signal focus. Some researchers report on pitch excursion differences (e.g. Pierrehumbert 1980; Liberman & Pierrehumbert 1984;

Braun 2006; Grice et al. 2017) resembling the phonetic parameters described for questions vs. statements. For example, it has been shown that, even when expressed by the same pitch accent type, contrastive focus is realised with tonal movements exhibiting greater pitch excursions than narrow focus (Grice et al. 2017).

Other languages such as Japanese (Beckman & Pierrehumbert 1986), Chicheŵa (Kanerva 1990), Bengali (Hayes & Lahiri 1991), and Korean (Jun 2005) do not use pitch accents to mark focus but encode focus through phrasing. In these languages, phrase boundaries are inserted to the left or the right of focused constituents setting them apart from non-focused constituents. These boundaries may or may not go hand in hand with a perceivable pause, as well as boundary-related spatio-temporal expansion of the segmental material.

Moreover, a focused constituent may come with additional strengthening of the segmental material involved. Research conducted mainly on languages that exhibit pitch accents reports that focused constituents exhibit temporal and spatial expansion. This prosodic strengthening can extend beyond the accented syllable and can affect unaccented syllables within a word. Vowels tend to be longer and hyperarticulated in accented syllables. Harrington, Fletcher & Beckman (2000) show that English high vowels are generally raised when accented. Also for English, Cho (2005) presents data showing that /a/ is consistently lowered and /i/ is consistently fronted when accented. He further shows that vowels in accented position are articulated with faster and longer jaw opening gestures than vowels in unaccented syllables.

In sum, a number of prosodic and intonational mechanisms have been cross linguistically reported to signal that a constituent is focused. These include certain tonal events, phrasing, and spatio-temporal expansion of the segments involved.

5.3.3 Intonational differences between marking questions and marking focus

Some of the phonetic parameters employed to mark focused constituents (especially contrastive focus) resemble those marking questions: rising-falling tonal events with locally increased pitch excursion. Despite their similarities, questions often exhibit greater pitch scaling of the whole contour than corresponding (contrastive) statements. Moreover, focused constituents may exhibit spatio-temporal expansion of the segments involved, setting them locally apart from other elements of the utterance.

In certain constructions, a question and a statement containing a phrase-final contrasted element may have a similar surface form, although there are subtle differences in timing (Gósy & Terken 1994; D'Imperio & House 1997; Makarova 2007). For example, in Neapolitan Italian, y/n questions are characterised by a rise on the accented vowel, followed by a fall marking the end of the phrase. The contour of a statement with narrow focus also has a rise to a pitch peak on the accented vowel followed by a fall. Even though these contours appear to be very similar in certain contexts (i.e. exhibiting the same characteristic rise(-fall)), they have been described as differing in the alignment of the pitch peak. In questions, the pitch peak reaches its target later in the accented vowel than in narrow focus statements.

Most of the above-cited studies have shown that in perception studies listeners are able to use the timing of the pitch peak as a cue to sentence modality. D'Imperio & House (1997) show that in Neapolitan Italian the timing of the peak is an important cue to the underlying function (questions vs. narrow focus statements), when the accented word is final in the phrase. Similarly, Gósy & Terken (1994) showed that in Hungarian later peaks are more often perceived as questions than as narrow focus statements (see also House 2003, for Swedish and Makarova 2007, for Russian).

To sum up, both the pitch scaling and the timing of rising(-falling) tonal events are common phonetic parameters signalling questions and contrastivity across languages. These parameters are exploited by listeners to disambiguate questions from statements. The following sections explore intonational devices employed to mark questions and contrastive constituents in Tashlhiyt. More specifically, y/n questions, taken as a prototypical question type, will be compared to contrastive statements. Based on the cross-linguistic observations discussed above, these sentence modalities may exhibit similar intonation contours, but may differ in pitch scaling and the timing of tonal targets. Furthermore, echo questions will be compared to y/n questions. Since echo questions are usually considered to request confirmation or clarification rather than request new information, they may be distinguished intonationally from y/n questions as for example in Bari Italian (Savino & Grice 2011). Again, pitch scaling and the timing of tonal targets may be relevant parameters that distinguish these two question types.

5.4 Questions and contrastive focus in Tashlhiyt: qualitative observations

The following sections will discuss morphosyntactic devices used to mark questions and contrastive statements in Tashlhiyt. Moreover, the intonational marking of these functions will be discussed based on qualitative observations.

5.4.1 Questions in Tashlhiyt

Tashlhiyt marks questions morphosyntactically. Y/n questions are canonically characterised by a morphological marker (is-) which cliticises to the verb in clause-initial position (cf. 8).

(8) is=t-sʁa t-fruχ-t ddisk?
INT=3F.SG-bought F-child.BS-F.SG record
'Did the girl buy a record?'

(9) t-sʁa t-fruχ-t ddisk
3F.SG-bought F-child.BS-F.SG record
'The girl bought a record.' or 'The girl bought a record?'

In spontaneous speech, declarative questions are used more frequently to request information than y/n questions. Declarative questions do not differ from corresponding declarative statements morphosyntactically (cf. 9).

Both question types (y/n, declarative) are characterised by similar intonation contours: a rise to a pitch peak followed by a fall usually occurring on the last word of the phrase (cf. Figure 5.1).[3]

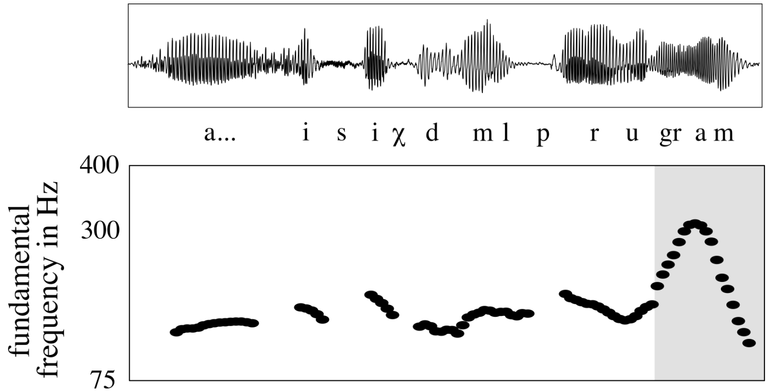

Figure 5.1: Representative waveform and f0 contour of the y/n question /a is iχdm lprugram/ 'Uh...does the program work?' Final syllable highlighted in grey.

Often the phrase-final fall does not reach a low target resulting in different degrees of truncation. In some cases, there is no fall at all. The degree of truncation is prone to both inter-speaker and intra-speaker variation. Figures 5.1 and 5.2

[3] This contour is also used for tag questions, in which the rise-fall is located on the tag.

illustrate two instances of questions. It is evident from the figures that most of the rise-fall trajectory takes place on the final syllable of the phrase (highlighted in grey).

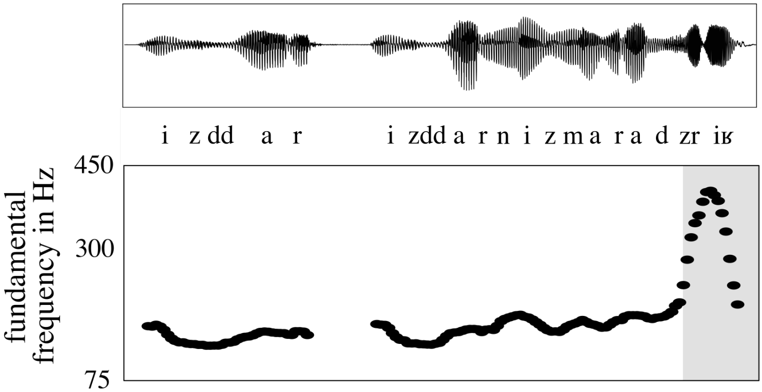

Figure 5.2: Representative waveform and f0 contour of the declarative question /izddar izddar n izm aʁ rad zriʁ/ 'The bottom… the bottom of the lion is it that I will go?' Final syllable highlighted in grey.

5.4.2 Contrastive focus in Tashlhiyt

Contrastive focus can be morphosyntactically expressed using either left dislocation or clefting. In both cases, the contrasted or emphasised element appears in clause-initial position preceding the verb (left dislocation, cf. 10). In clefted constructions the clause-initial constituent is additionally marked morphologically (cleft, cf. 11).

(10) ddisk t-sʁa=t t-fruχ-t
record 3F.SG-bought=it F-child.BS-F.SG
'A record, the girl bought it.'

(11) ddisk ad t-sʁa t-fruχ-t
record AD 3F.SG-bought F-child.BS-F.SG
'It is a record that the girl bought.'

Due to the frequent use of morphosyntactic constructions expressing focus, examples of contrastive focus expressed by intonation are only seldom found. In these cases, speakers use a rise-fall to mark contrasted constituents in-situ as illustrated in Figure 5.3. Here, the speaker contrasts /izddar/ 'underneath' with

/iggi/ 'above' via rise-falls at the end of the respective words. The rise is on the final syllable of each preposition.

This type of contrastive focus can be isolated in elicited speech (cf. Figure 5.4). The rise-fall is located at the right edge of the focused constituent. Similar to the question tune, pitch suddenly falls after reaching the high target and stays low until the end of the utterance. The entire movement is often restricted to one syllable, here the final syllable of the word (highlighted in grey).

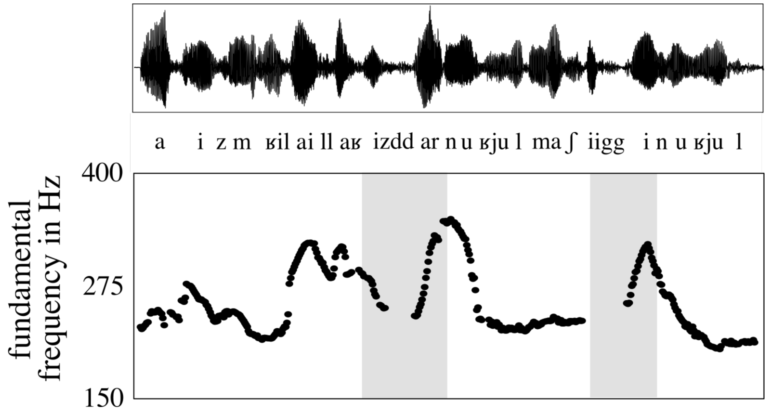

Figure 5.3: Representative waveform and f0 contour of the statement containing contrasted constituents /a izm ʁila illa ʁ izddar n uʁjul maʃi iggi n uʁjul/ 'Uh, the lion, now, he is underneath the donkey, not above the donkey'. Contrasted words highlighted in grey.

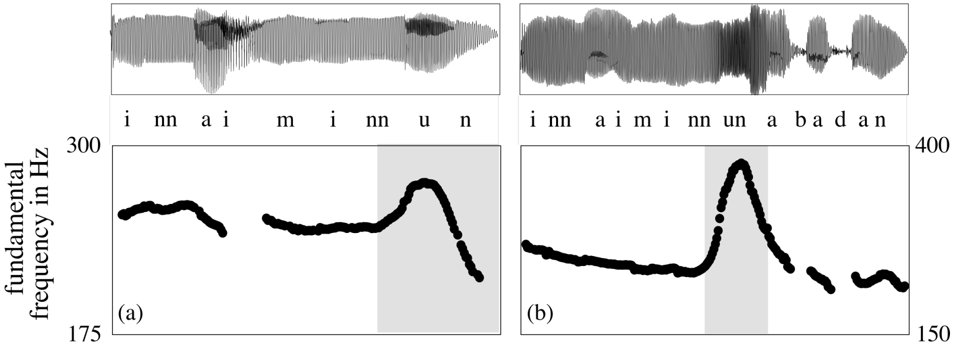

Figure 5.4: Representative waveforms and f0 contours of contrastive statements (a) /inna iminnun/ 'He said 'your mouths'.' and (b) /inna iminnun abadan/ 'He said 'your mouths' always.'. Final syllables of contrasted words highlighted in grey.

5.4.3 Intonational differences between flagging questions and marking contrastive focus in Tashlhiyt

The intonation of questions and contrastive statements are qualitatively similar. Both tunes are characterised by a local rise-fall in pitch. The sentence modalities differ, however, in the position of the rise-fall within the utterance. In contrastive statements, the rise-fall co-occurs with the right edge of the contrasted element. In questions, it co-occurs with the right edge of the phrase. Additionally, the question tune exhibits auditory qualities that distinguish it from the contrastive tune. Comparing the f0 range in questions to the one in statements, it becomes apparent that questions exhibit a higher pitch register, i.e. overall higher f0 values and higher pitch excursion. In fact, speakers frequently change phonation type towards the pitch peak producing a falsetto-like phonation. This is acoustically manifested by strikingly high f0 values of up to 700 Hz and a sudden shift into low vibrational amplitude (Laver 1994). The intonation contours of declarative questions and y/n questions resemble each other in terms of both tonal placement and f0 range.

We now set out to evaluate these qualitative observations quantitatively in a controlled reading experiment comparing questions and corresponding contrastive statements, as well as y/n questions and echo questions.

5.5 Production study

The objective of the present study is to investigate whether the difference between questions and contrastive statements is reflected in global and local scaling differences and/or in the timing of the rise-fall in pitch. Moreover, echo questions are compared to y/n questions to investigate the potential differences between questions requesting information and questions requesting confirmation. In the following sections, production data from a read speech corpus is analysed. Data has been collected on a field trip in Agadir in November 2013.[4]

5.5.1 Method

5.5.1.1 Participants

Ten native speakers of Tashlhiyt (five male, five female, mean age = 22 (20-27)) were recorded. All live in Agadir, Morocco, and are fluent in Moroccan Arabic

[4] Parts of the analysis presented in this section have been published in Roettger & Grice (2015).

and have basic command of French. All of them had normal or corrected-to-normal vision. None reported on any hearing impairments. Subjects were paid for their participation (cf. Appendix A2.2 for speaker information).

5.5.1.2 Speech material

The present production data is part of a larger corpus of read speech. The read corpus consisted of short mock dialogues containing target words in four different sentence modalities (y/n question, negation, contrastive statement, and echo question) based on the simple sentence: inna TARGET 'he said TARGET'. Sentences differed with respect to the distance of the target word to the right phrase edge (phrase medial and phrase final). Examples (12a-d) present the layout of the dialogues with the targets in phrase-final position (no adverb).

(12) a. is inna **baba**?
'Did he say 'father'?'

b. ur inna **baba**.
'He did not say 'father'.'

c. inna **dari**.
'He said 'in my house'.'

d. manik? inna **dari**? irwas.
'How? He said 'in my house'? It seems like it.'

In (12a) the target word is in a y/n question. In (b) the same target word is in a negative assertion. Because of the preceding negation, a different target word is explicitly corrected in a contrastive statement in (c). Finally, in (d), the proposition in the contrastive statement is called into question in the counter-expectational echo question. In (13), an example dialogue with the target word in phrase-medial position followed by an adverb is given.[5]

(13) a. is inna **baba** abadan?
'Did he say 'father' then?'

b. ur inna **baba** abadan.
'He did not say 'father' then.'

[5] The author is aware of the metalinguistic nature of these context sentences. Exploration of semi-spontaneous and spontaneous speech, however, does not indicate any patterns diverging from the controlled corpus described above. Since target words of different parts of speech were compared, a 'more natural' context sentence for the present question was not available.

c. inna **dari** abadan.
'He said 'in my house' then.'

d. manik? inna **dari** abadan? irwas.
'How? He said 'in my house' then? It seems like it.'

As described above, speakers are expected to produce a rise-fall at the right edge of the phrase in questions (on the target in 12a and 12d, and on the adverb in 13a and 13d). In statements, speakers are expected to produce a rise-fall marking contrastive focus on the target word (phrase final in 12c and phrase medial in 13c). This means that the location of the pitch peak is always either on the target word or on the adverb, in both sentence modalities. The word co-occurring with the pitch peak is henceforth referred to as the 'tone bearing word'. Productions of the negative assertion in (12b) and (13b) are not subject to the present analysis.

The corpus contained 18 different target words. There were ten fully voiced target words and eight target words containing voiceless segments only. For the present analysis only fully voiced target words are considered (cf. Table 5.1). Each target word appeared in each context at least once. Several items appeared twice. This resulted in 54 fully voiced target words for each participant (total 540). The data with target words containing obstruents only will be subject to discussion in Chapter 6.

Table 5.1: Target words and translations of production study.

Word	Translation
ba.ba	'my father'
da.ri	'in my house'
di.ma	'always'
il.di	'he pulls'
i.min.nun	'your mouths'
ma.na.gu	'when'
ʁi.la	'now'
u.dm	'face'

5.5.1.3 Procedure

Participants were seated in front of a computer screen and read out orthographically presented material containing the target words as presented in carrier sentences in (12) to (13) (i.e. in mock dialogues). Participants were asked to enact these dialogues. The materials were presented in a version of the Latin script speakers are used to reading and writing in (see Chapter 3).

Recordings were made in a quiet room at the Ibn Zohr University in Agadir. The production data was recorded using a Marantz PMD 670 solid-state recorder at a sampling rate of 44.1 kHz, and an AKG C420 III head-mounted microphone. Before recordings began, participants were asked to read aloud a word list containing all of the target words to ensure that they were familiar with the words and their meanings. Dialogues were presented in random order.

5.5.1.4 Analyses

All acoustic data was manually annotated employing the following labelling criteria: segment boundaries (and, in turn, syllable and word boundaries) were identified in the acoustic waveform by means of an oscillogram and a wide-band spectrogram. All segmental boundaries of vowels and consonants were labelled at abrupt changes in the spectra at the time at which the closure was formed or released: this was the case for nasals, laterals (especially in the spectra for the intensity of higher formants), and obstruents (at random noise patterns in the higher frequency regions). All acoustic information was automatically extracted via Praat version 5.4 (Boersma & Weenink 2015).

F0 tracks for all utterances were extracted, manually corrected, and smoothed using the Praat script ‘mausmooth’ (Cangemi 2015). The smoothing algorithm levelled out strong microprosodic effects and enabled the inspection of uninterrupted contours. The smoothed contours were used for automatic extraction of the f0 mean of the word /inna/. Since /inna/ is expected to exhibit a relatively flat f0 over the course of the word in the investigated sentence modalities (cf. Figure 5.4), the mean f0 of /inna/ (in Hz) is taken as a reference level to operationalise ‘pitch level’. Any tonal movement following /inna/ can be assessed in relation to this reference level. Additionally, minimum and maximum f0 values of the utterance were extracted. The difference between minimum and maximum was calculated in semitones (ST) to operationalise ‘pitch range’. Finally, the timing of the rise-fall was investigated. Due to difficulties in reliably measuring low turning points, it is abstracted away from the actual f0 trajectory and focused on the high turning point for both pragmatic functions. Henceforth the high target is referred to as the pitch peak. Peak timing was calculated as the time lag between the acoustic onset of the final syllable within the word and the f0 maximum in seconds (cf. 14a-c):

(14) a. PITCH LEVEL: mean f0 of the reference word /inna/.
b. PITCH RANGE: difference between maximum and minimum f0 values in ST.
c. F0 MAX LAG: lag between the onset of the final syllable of the tone bearing word and the f0 maximum.

5.5.1.5 Statistics

Sentences produced with hesitation or unnatural phrasing patterns, mispronunciations of the segmental material, or instances exhibiting list intonation were excluded from the analyses. Moreover, most echo questions from speaker F5 were excluded. She produced echo questions with a monotonous contour. Native speakers who did not participate in the experiment judged these to resemble bored statements and thus inappropriate realisations of the intended context.

Generally, speakers did well in naturally enacting the dialogues, but they had difficulties with reading aloud. This resulted in an unusually large amount of hesitations and mispronunciations. Overall, acoustic parameters for 471 utterances were submitted to statistical analysis (= 7.8% data loss) and were analysed with generalised linear mixed models, using R (R Core Team 2015), the lme4 package (Bates et al. 2015), and the multcomp package (Hothorn, Bretz & Westfall 2008). Fixed effect specification will be given in the relevant paragraphs below. A term for varying intercepts for speakers and for target words was included. Terms for varying slopes were not included, since the data set is rather small and the factorial design was not balanced (asymmetric exclusion of data points frequently leading to converging issues). Speaker-specific tables will be provided to allow for inspection of consistency across speakers. To determine p-values for the main effects / interactions between factors, a model including the main effect / interaction of interest was compared to the same model with no main effect / no interaction via Likelihood Ratio Tests (LRT).

5.5.2 Results: pitch scaling

First, we will discuss the scaling differences between sentence modalities. Generally, questions exhibit a higher reference pitch level in /inna/ than statements (cf. Figure 5.5, Table 5.2). This is true for the comparison between contrastive statements (CS, 189 Hz) and both echo questions (EQ, 235 Hz) and y/n questions (Y/N, 264 Hz).

The difference between sentence modalities is consistent across all speakers. Interestingly, echo questions reveal an intermediate status, i.e. they are consis-

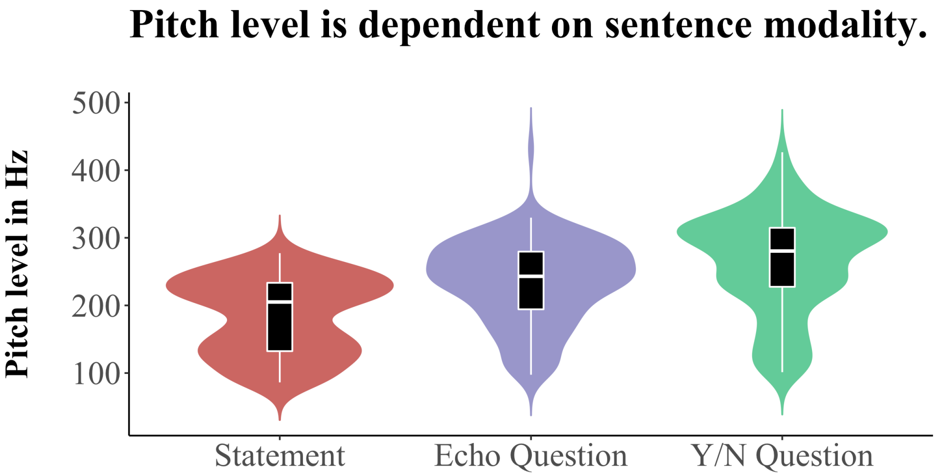

Figure 5.5: Violin plots of the pitch level values of the reference word /inna/ as a function of sentence modality. Inside each plot, the black boxes indicate the inter-quartile range (IQR), the range between the first and third quartile. The solid horizontal line indicates the median. The whiskers indicate the range, up to 1.5 times the IQR away from the median. The overall shape of the violin plots represent kernel density curves of the raw data distribution.

Table 5.2: Mean pitch level (and standard deviation, in Hz) of the reference word /inna/ as a function of sentence modality for each speaker individually averaged over words.

Speaker	Contrastive Statement	Echo Question	Y/N Question
F1	245 (10)	303 (13)	317 (20)
F2	241 (13)	298 (40)	336 (38)
F3	217 (5)	269 (14)	337 (39)
F4	211 (9)	249 (22)	280 (25)
F5	251 (16)	NA	313 (22)
M1	120 (4)	161 (20)	168 (19)
M2	129 (11)	194 (12)	226 (17)
M3	152 (11)	254 (23)	293 (35)
M4	164 (14)	230 (14)	237 (16)
M5	90 (5)	110 (8)	116(13)
overall	**189 (55)**	**235 (59)**	**264 (74)**

tently higher than statements and consistently lower than y/n questions. A linear mixed effects model was performed including sentence modality as a fixed effect. The model estimated the main effect of sentence modality to be significant (χ^2(2)=496, p<0.00001). Post-hoc Tukey tests reveal that the difference is in fact significant across all comparisons (statement vs. echo question: β=55.7, SE=2.9, z=19.1, p<0.00001; statement vs. y/n question: β=82.1, SE=2.8, z=29.7, p<0.00001; echo question vs. y/n question: β=26.4, SE=2.8, z=9.5, p<0.00001).

We now turn to the pitch range measurements, i.e. the difference between maximum and minimum f0 values within the utterance. Pitch range is numerically rather large with an average of 9 STs across all sentence modalities. In fact, in questions, speakers frequently changed phonation type towards the pitch peak producing a falsetto-like phonation. This is acoustically manifested by very high f0 values of up to 700 Hz and sudden shifts into low vibrational amplitude.

Overall, questions exhibit a greater pitch range than statements. This is true for the comparison between statements (6.8 ST) and both echo questions (10.7 ST) and y/n questions (9.5 ST) (cf. Figure 5.6 and Table 5.3).

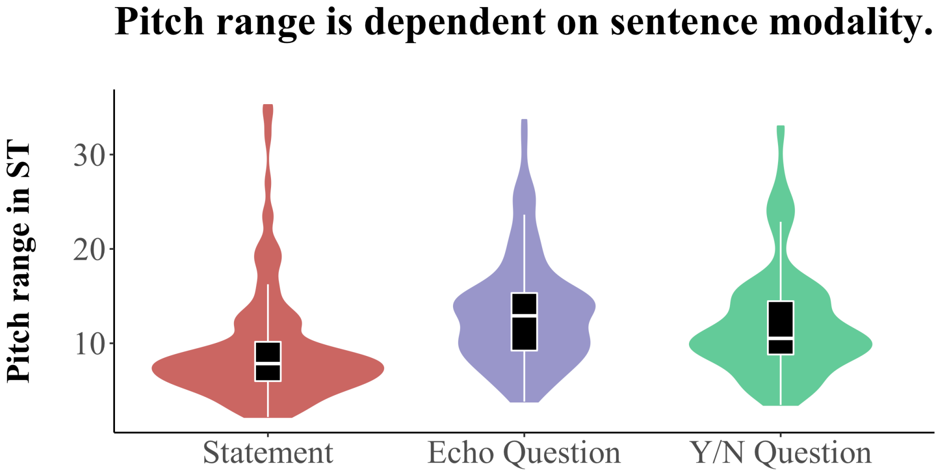

Figure 5.6: Violin plots of the mean pitch range values as a function of sentence modality. Inside each plot, the black boxes indicate the inter-quartile range (IQR), the range between the first and third quartile. The solid horizontal line indicates the median. The whiskers indicate the range, up to 1.5 times the IQR away from the median. The overall shape of the violin plots represent kernel density curves of the raw data distribution.

Table 5.3: Mean pitch range (and standard deviations, in ST) as a function of sentence modality for each speaker individually averaged over words.

Speaker	Contrastive Statement	Echo Question	Y/N Question
F1	6.6 (1.7)	8.4 (1.6)	7.3 (1.5)
F2	5.7 (1.6)	13.3 (2.6)	13.2 (2.2)
F3	6.7 (2.7)	8.2 (2.3)	10.8 (2.7)
F4	5.9 (1.2)	10 (2)	7.8 (1.3)
F5	9.1 (3.8)	NA	10.6 (2)
M1	3.1 (1.3)	15.5 (2.5)	7.6 (3.1)
M2	6.9 (1.9)	11.6 (2.1)	9 (2.8)
M3	9.8 (1.5)	14.6 (3.6)	13.3 (3.6)
M4	6.5 (3)	8 (2.3)	9 (3.2)
M5	5 (1.4)	5 (1.8)	5.8 (1.7)
overall	**6.8 (2.8)**	**10.7 (3.9)**	**9.5 (3.4)**

For some speakers, the distinction between statements and questions is very clear (e.g. F2 and M1), while for other speakers, the distinction is more subtle and characterised by heavy overlap (e.g. F1 and M5). With regard to the distinction between echo questions and y/n questions, there is a large amount of individual variation. Some speakers exhibit a greater pitch range in y/n questions (e.g. F3, M4, M5), while others exhibit a greater pitch range in echo questions (F1, F2, F4, M1, M2, M3) (cf. Table 5.3).

A linear mixed effects model was performed including sentence modality as a fixed effect. The model estimated the main effect of sentence modality to be significant ($\chi^2(2)$=143.1, p<0.00001). Post-hoc Tukey tests reveal that the difference is in fact significant across all comparisons (statement vs. echo question: β=4.2, SE=0.3 z=12.6, p<0.0001; statement vs. y/n question: β=2.6, SE=0.6, z=4.5, p<0.0001; echo question vs. y/n question: β=1.2, SE=0.3, z=3.9, p=0.0003). Statistically speaking, those speakers that exhibit a greater pitch range in echo questions show a rather strong effect (e.g. speaker M1) which potentially drives the overall mean differences. In light of these inter-individual difference and the fact that the statistical models did not account for varying speaker slopes, generalisations based on the inferential results need to be considered critically here.

5.5.3 Results: pitch peak timing

We now turn to the timing of the pitch peak. On average, pitch peaks occurred around 113 ms after the onset of the final syllable of the tone bearing word (utterance final in questions and focused constituent final in contrastive statements). However, the pitch peak is aligned later in y/n questions (144 ms) and echo questions (159 ms) than in statements in which the pitch peak is reached, on average, close to the syllable boundary between the penult and the final syllable (20 ms). A linear mixed effects model was performed including sentence modality as the critical fixed effect. Additionally, the presence of a coda consonant in the final syllable was added in an interaction with sentence modality. This was done because the presence of a voiced coda consonant might enable the pitch peak to be aligned later while still allowing for the full realisation of the subsequent fall (e.g. Mücke et al. 2009; Niemann & Mücke 2015). Here, the presence of a coda consonant actually confounds with the factor sentence modality because there are more cases of tone bearing words with a coda consonant in questions than in statements. Where target words were phrase-medial, the tonal event for questions is always found on the phrase-final adverb /abadan/.

There is no apparent interaction between the effect of sentence modality and the presence of a coda consonant ($\chi^2(2)=5.49$, p=0.064). The model estimates the main effect of sentence modality to be significant ($\chi^2(2)=120.41$, $p<0.00001$). Post-hoc Tukey tests reveal that there is a significant difference between statements and questions, but there is no significant difference between question types (statement vs. echo question: $\beta=0.11$, SE=0.01 z=10.9, $p<0.0001$; statement vs. y/n question: $\beta=0.10$, SE=0.01, z=10.3, $p<0.0001$; echo question vs. y/n question: $\beta=0.01$, SE=0.01, z=1.1, p=0.5).

It can thus be concluded that the pitch peak in statements is reached earlier in the word than in questions. Echo and y/n questions, on the other hand, appear to have a similar distribution of pitch peak alignment (cf. Figure 5.7).

Looking at the actual distributions of the peak in relation to the onset of the final syllable, it becomes apparent that the averaged values are somewhat misleading (cf. Figure 5.7). First, the alignment of pitch peaks in statements is more variable than in questions, indicated by the wider spread in the distribution. Moreover, the distribution for statements is not unimodal, but bimodal as indicated by the occurrence of two peaks in the distribution. Looking at questions, there is some indication of bimodality here, too. There is a small bump in the distribution left of the syllable boundary (more prominent in y/n questions). These bimodal distributions reflect the auditory impressions of other researchers (e.g. Dell & Elmedlaoui 1985, Ridouane p.c.), as well as the impression of the author.

Although the pitch peak is most often located on the final syllable, it is possible for speakers to produce the pitch peak on the penult as well. The resulting tonal patterns give rise to discretely different auditory impressions rather than to the impression of a continuously variable position of the tonal event.

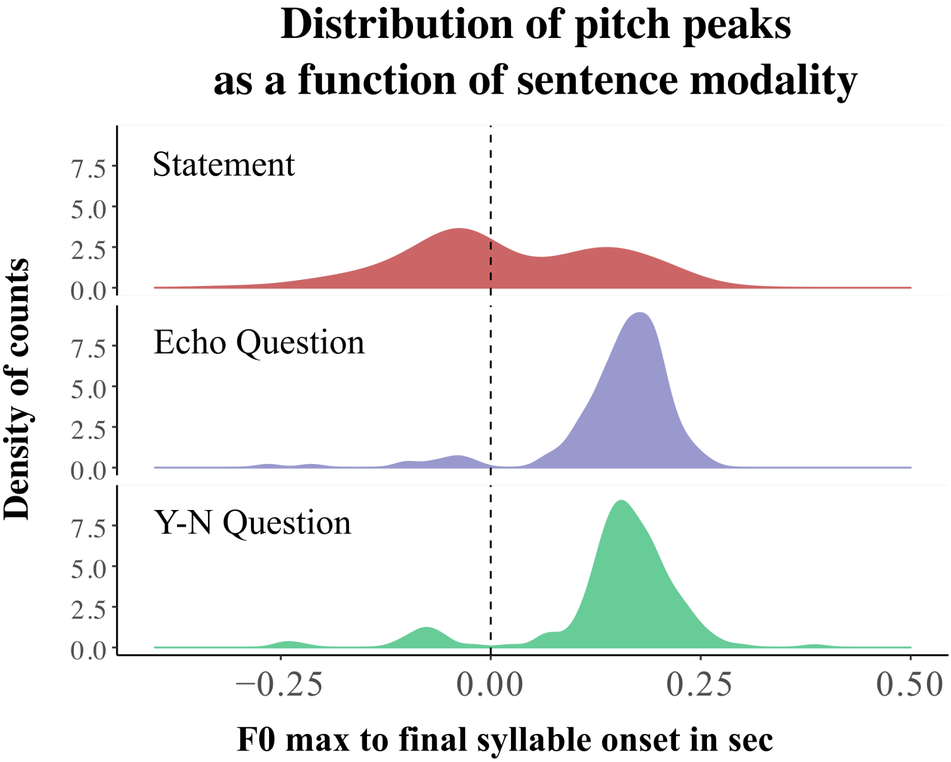

Figure 5.7: Kernel density curves for pitch peak measurements relative to the onset of the final syllable (at final-to-f0 max = 0) for (a) statements, (b) echo questions, and (c) y/n questions. Positive values indicate that the pitch peak occurs in the final syllable, negative values indicate that the pitch peak occurs in the penult. The dashed line marks the onset of the final syllable.

Table 5.4 illustrates the high degree of inter-speaker variability. Generally, there is a strong tendency for individual speakers to produce the pitch peak on the final syllable (most mean values are positive). This trend is stronger for questions than for statements. In fact, some speakers prefer placing the pitch peak on the penult in statements (cf. negative values in M2, M3, and M4). Generally, the pitch peak in statements is found on the penult more frequently than in questions.

Table 5.4: Alignment of mean final-to-f0 max (in ms) with respect to final syllable onset (and standard deviations) as a function of sentence modality for each speaker individually averaged over words.

Speaker	Contrastive Statement	Echo Question	Y/N Question
F1	17 (127)	153 (33)	162 (31)
F2	36 (111)	154 (73)	124 (80)
F3	53 (118)	175 (44)	208 (46)
F4	30 (105)	170 (76)	143 (67)
F5	23 (157)	NA	149 (117)
M1	66 (69)	126 (54)	75 (134)
M2	-3 (124)	147 (22)	154 (22)
M3	-31 (55)	147 (71)	146 (82)
M4	-17 (160)	177 (67)	167 (91)
M5	72 (121)	111 (133)	101 (109)
overall	**20 (119)**	**159 (75)**	**144 (89)**

This discretely formulated observation is also reflected in more gradual trends in phonetic alignment. Even if only pitch peaks on the final syllable are considered (f0 max lag > 0), the difference between statements and questions holds with statements reaching the pitch peak 129 ms after the onset of the final syllable while echo questions and y/n questions reach the pitch peak later in the syllable (166 and 167 ms, respectively).

Looking at word-specific distributions, it becomes clear that the mobility of the tonal event appears to be dependent on word-specific properties. Figure 5.8 illustrates the distribution of peak alignment for three representative target words: /i.min.nun/, /ba.ba/, and /u.dm/.

In Figure 5.8a, there is a subtle bimodality but the majority of peaks appears to be on the final syllable (positive values). In Figure 5.8b, there is still a strong bias towards the final syllable but also a noticeable amount of peaks occurring on the penult. In Figure 5.8c, the target /u.dm/ exhibits a strong bimodality with a clear bias towards peaks on the penult. There appears to be something specific about the lexical items that leads speakers to be more likely to produce pitch peaks on the penult or the final syllable.

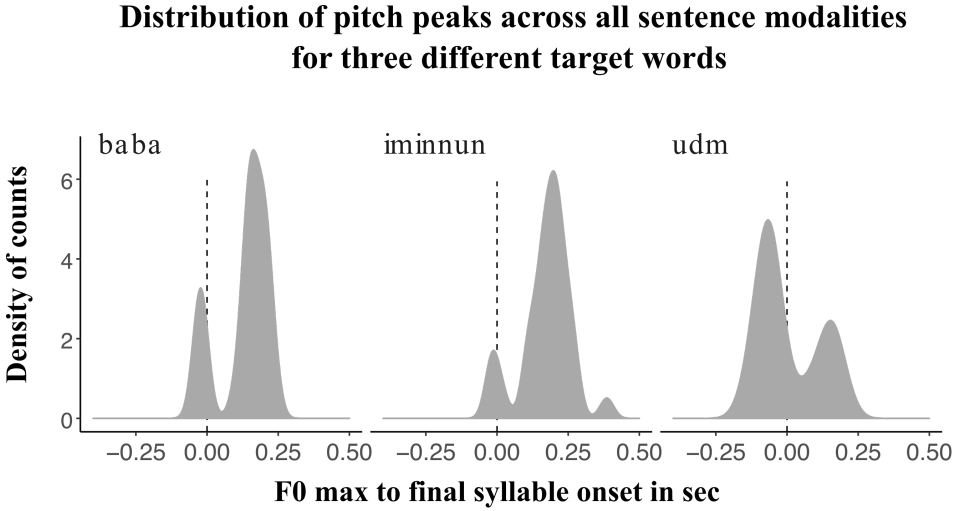

Figure 5.8: Kernel density curve for pitch peak measurements relative to the onset of the final syllable (at final syllable onset to f0 max = 0) for (a) /iminnun/ 'your mouths', (b) /baba/ 'father', and (c) /udm/ 'face'. Positive values indicate that the pitch peak occurs in the final syllable, negative values indicate that the pitch peak occurs in the penult. Values are averaged over subjects and sentence modalities. The dashed line marks the onset of the final syllable.

In addition to potential lexical effects, in many cases, tonal alignment appears to be prone to some degree of free alternation. Figure 5.9 illustrates examples for the pitch peak on the final syllable and the penult, respectively. In Figure 5.9a, the penult of /il.di/ is low before f0 suddenly rises towards the peak on the final syllable. In Figure 5.9b, f0 suddenly drops after reaching its high target on the penult leaving the final syllable low. In most cases, the rise-fall appears to be located exclusively on one syllable.

5.5.4 Discussion

To recapitulate the production results, questions in Tashlhiyt are distinguished from corresponding contrastive statements by all three investigated parameters: compared to statements, questions have an overall higher pitch level, a greater pitch range, and the pitch peak is realised later within the word. The alignment pattern can be described in both discrete as well as continuous terms. In questions, the pitch peak is aligned more often with the final syllable than in statements. Statements have many more instances of pitch peaks aligned to the penult

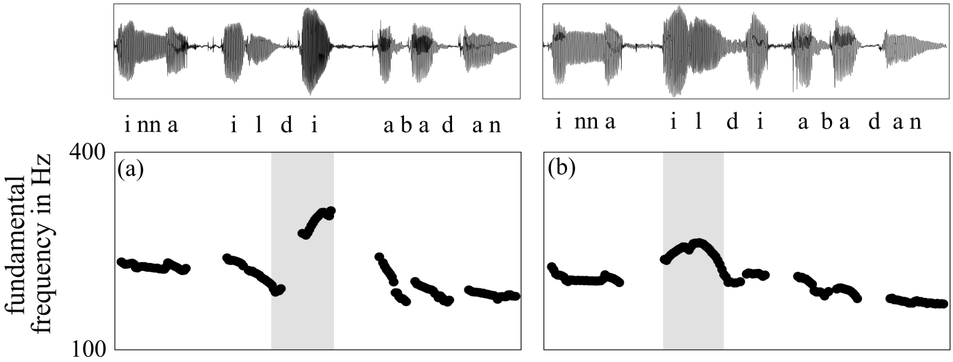

Figure 5.9: Representative waveforms and f0 contours of the contrastive statement /inna ildi abadan/ 'Did he say 'he pulls' always?' with the pitch peak (a) on the final syllable and (b) on the penult. Syllables co-occurring with the pitch peak are highlighted in grey. Both productions are from the same speaker.

than questions. In continuous terms, even if only pitch peaks on the final syllable are considered, questions have later peaks than corresponding statements.

In addition to the distinction between statements and questions, echo questions are distinguished from y/n questions. This distinction is mainly manifested in pitch level with echo questions being lower in pitch than y/n questions. Evidence for echo questions having a greater pitch range is weak at best and the two question types exhibit comparable alignment patterns of the pitch peak.

Generally, there is a high degree of variability both across and within speakers with regard to the position of the pitch peak. While some variability can be explained by functional factors like sentence modality, there remains a substantial amount of unexplained variance. A recent study by Grice, Ridouane & Roettger (2015) shed more light on the factors that determine tonal placement in Tashlhiyt. Their results will be reviewed here briefly.

Grice et al. recorded Tashlhiyt speakers living in Paris. Four native speakers of Tashlhiyt, originally from Morocco but permanently living in Paris, participated in the experiment. Even though the speakers did not live in Morocco anymore, they were frequently in contact with friends and family there. According to the author's impression (and according to Rachid Ridouane, p.c.), their intonational patterns did not diverge from those speakers recorded in Agadir.

Speakers read out short mock dialogues (15-17) similar to the corpus described in the production study above (§5.4.1). Twenty-eight pairs of disyllabic target words were recorded. Target words varied in the sonority of the syllable nucleus and in the weight of the final syllable.

(15) is inna **tugl**?
‘Did he say ‘she hung’?’

(16) ur inna **tugl**.
‘He did not say ‘she hung’.’

(17) inna **tmdl**.
‘He said ‘she buried’.’

In line with the findings presented above, Grice et al. observed that the phrase-final word bears a pitch peak in both y/n questions and contrastive statements. They described the distribution of the pitch peak as a bimodal pattern, with the pitch peak being aligned either with the penult or with the final syllable of the target word. This discrete pattern coincided with the auditory impression of prominence, i.e. the syllable on which the peak occurred sounded louder and longer to the authors. They explicitly state that the position of the pitch peak was unambiguously identifiable when listening to the utterances. In the following, we will focus on the discrete placement patterns of the pitch peak reported by Grice, Ridouane & Roettger (2015).

In target words with only one sonorant nucleus (i.e. a sonorant consonant, e.g. /r/ in /tr̩.kz̩, tb̩.dr̩/ ‘she danced, she mentioned’), the pitch peak was almost exclusively located on that syllable. When both syllables had a sonorant nucleus (i.e. a sonorant consonant or vowel, e.g. /tiri, tm̩.dl̩/ ‘she wanted, she buried’) there was no clear preference for a peak on the penult or final syllable. Here, the placement of the peak for both statements and questions was highly complex and subject to the influence of a number of interacting factors. Compared to statements, they found a preference for pitch peaks on the final syllable in questions. Orthogonal to that, they identified two independent factors relevant for determining the location of the pitch peak:

First, the pitch peak was more likely to co-occur with more sonorous syllable nuclei than with less sonorous syllable nuclei. For example, in a word like /tu.gl̩/ ‘she hung’, the vowel, which has a higher sonority than the liquid, was more likely to attract the pitch peak to the penultimate syllable. Conversely, in a word like /tn̩.za/ ‘it was solved’, the final syllable was more likely to attract the pitch peak. The sonority asymmetry was not only relevant for the distinction between vowels and sonorant consonants, but for the distinction between consonants with differing degrees of sonority, e.g. a liquid was preferred over a nasal (e.g. /tm̩.dl̩/ ‘she buried’ vs. /tr̩.km̩/ ‘she rotted’). Overall, the effect of sonority was stronger for vowels than for sonorant consonants, i.e. vowels were stronger attractors for the pitch peak than consonants.

Second, the pitch peak was more likely to co-occur with heavy syllables than with light syllables. In /tu.gl̩t/, the pitch peak was more likely to co-occur with the final syllable than in /tu.gl̩/.

The weighting and interaction of these factors was to some degree speaker-specific, although they generally appeared to be additive and systematic throughout the sample (cf. Tables 5.5–5.9). There was a general preference for the pitch peak to occur on the final syllable indicated by the high overall proportional values. Questions were generally more likely than statements to be produced with a final pitch peak, resulting in ceiling effects for several cells. Heavy syllables systematically attracted a final pitch peak more often than corresponding light syllables and more sonorous nuclei attracted a pitch peak more often than less sonorous nuclei.

The above-identified factors do not explain, however, the entire variance. In many cases, tonal placement was prone to some degree of free alternation. There were numerous cases where the same speaker produced pitch peaks in different locations within the same target word across different repetitions. Figure 5.10 illustrates a minimal quadruplet with statements and y/n questions.

Table 5.5: Results of Grice, Ridouane & Roettger (2015: 250f.): mean proportion (in %) of pitch peak location on the final syllable in words containing vowels in contrastive statements for each speaker separately. Results are ordered according to the syllable nuclei of the penult and final syllable (V = vowel, S = sonorant consonant) and syllable weight of the final syllable (light or heavy)

	Contrastive Statements Words Containing Vowels)				
	F1	F2	M1	M2	Overall
V.S	58.3	0	66.7	75	51.1
V.V	75	0	100	10	47.8
S.V	100	81.8	100	100	95.8
light	61.1	29.4	79	55.6	56.9
heavy	94.4	23.5	100	75	73.9

Table 5.6: Results of Grice, Ridouane & Roettger (2015: 250f.): mean proportion (in %) of pitch peak location on the final syllable in words containing consonants only in contrastive statements for each speaker separately. Results are ordered according to the syllable nuclei of the penult and final syllable (L = Liquid, N = Nasal) and syllable weight of the final syllable (light or heavy).

	Contrastive Statements (Words Containing no Vowels)				
	F1	F2	M1	M2	Overall
L.N	41.7	0	58.3	50	37.5
N.N / L.L	66.7	0	75	100	61.7
N.L	100	41.7	100	100	84.8
light	55.6	5.9	64.7	83.3	52.9
heavy	82.4	22.2	88.9	83.3	69

Table 5.7: Results of Grice, Ridouane & Roettger (2015: 250f.): mean proportion (in %) of pitch peak location on the final syllable in words containing vowels in y/n questions for each speaker separately. Results are ordered according to the syllable nuclei of the penult and final syllable (V = vowel, S = sonorant consonant) and syllable weight of the final syllable (light or heavy).

	Y/N Questions (Words Containing Vowels)				
	F1	F2	M1	M2	Overall
V.S	76.9	63.6	50	91.7	70.8
V.V	100	100	100	100	100
S.V	100	100	100	100	100
light	84.2	76.5	66.7	94.4	80.6
heavy	100	100	100	100	100

Table 5.8: Results of Grice, Ridouane & Roettger (2015: 250f.): mean proportion (in %) of pitch peak location on the final syllable in words containing consonants only in y/n questions for each speaker separately. Results are ordered according to the syllable nuclei of the penult and final syllable (L = Liquid, N = Nasal) and syllable weight of the final syllable (light or heavy).

	Y/N Questions (Words Containing no Vowels)				
	F1	F2	M1	M2	Overall
L.N	72.7	45.5	75	75	67.4
N.N / L.L	83.3	83.3	83.3	91.7	85.4
N.L	100	100	100	100	100
light	72.2	61.1	72.2	77.8	70.8
heavy	100	94.1	100	100	98.6

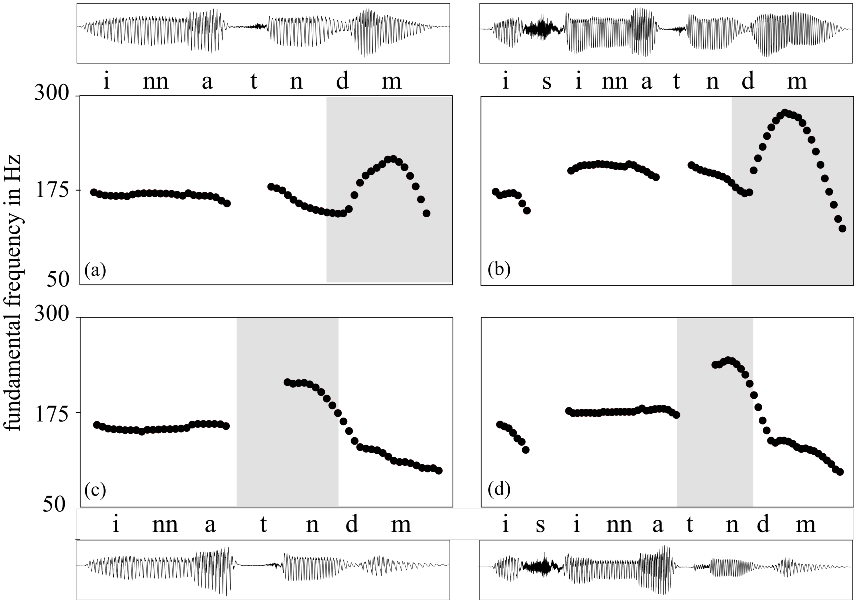

Figure 5.10: Representative waveforms and f0 contours of two realisations of the contrastive statement /inna tndm/ ‘he said ‘she regretted’’ (a,c) and the y/n question /is inna tndm/ ‘did he say ‘she regretted’?’ (b,d): the two different realisations for each sentence modality illustrate variation in pitch peak placement. All utterances are from the same male speaker. Tone bearing syllable is highlighted in grey.

Grice et al.'s findings put the results of the production study presented in this chapter into perspective. As opposed to Grice et al., there was an even stronger preference for final pitch peaks in the present data set. This might be due to the nature of stimuli employed, i.e. most words had a light final syllable with a vowel in syllable nucleus position. In line with Grice et al., these structures generally attract the pitch peak to the final syllable. Across sentence modalities, the word /u.dm̩/ 'face' exhibits an exceptionally large number of pitch peaks on the penult (overall 64% of all cases across sentence modalities). This may be a reflex of the sonority asymmetry in udm. The vowel in the penult is more sonorous than the nasal in the final syllable and attracts the pitch peak to the penult syllable. The word /i.min.nun/ 'your mouths', instead, exhibits mainly final pitch peaks. This is possibly a reflex of the final syllable being heavy. So even though Grice, Ridouane & Roettger (2015) and the present production study had different speaker samples and different speech materials, comparable tonal placement regularities can be observed.

To sum up, Tashlhiyt exhibits a remarkable amount of variability in tonal placement. However, this variability has a particular structure, i.e. it is bimodal. The occurrence of a particular tonal event can only be stated as a probabilistic distribution affected by multiple interacting factors. The question arises if, and if so, how, these probabilistic patterns are used to distinguish sentence modalities perceptually.

5.6 Perception study

In the preceding section, it has been shown that statements and questions differ in terms of pitch register according to pitch level and pitch range. Moreover, sentence modalities differ with respect to the timing of the tonal event involved. The timing of the pitch peak common to both statements and questions, showed variation in both discrete and continuous terms. The pitch peak co-occurred either with the penult or with the final syllable of the tone bearing word, with more instances of the pitch peak on the final syllable in questions. Moreover, even if only pitch peaks aligned with the final syllable are considered, questions still exhibited a later peak alignment within the syllable. The following perception study investigates whether, and if so, how, Tashlhiyt listeners use these pitch parameters to interpret morphosyntactically ambiguous sentences.

5.6.1 Method

5.6.1.1 Participants

Nine native speakers of Tashlhiyt (four male, five female, mean age = 21 (20–23)) participated in the experiment. None of them had participated in the previous production experiment. All live in Agadir, Morocco, are fluent in Moroccan Arabic, and have basic command of French. All of them had normal or corrected-to-normal vision. None reported on any hearing impairments (cf. Appendix A2.3 for participant information).

5.6.1.2 Speech materials and procedure

In order to control for pitch register and pitch peak placement, stimuli were resynthesised. As base stimuli, four fully voiced phrases were recorded: /inna baba/, / inna bibi/, / inna dima/, and / inna ʁila/ 'He said ('father, turkey, always, now')' all produced by a phonetically trained native speaker of Tashlhiyt (Rachid Ridouane). For each phrase, the speaker produced two contours corresponding to two discretely different pitch peak positions resulting in two sets of stimuli: one set contained the pitch peak on the final syllable (F) of the target word and the other set contained the pitch peak on the penult (PU). The speaker was instructed to produce the two sets in the same register. Subsequent inspections of the contours confirmed that this was the case.

Both sets were resynthesised using PSOLA in Praat version 5.4 (Boersma & Weenink 2015). F0 was manipulated resulting in two different pitch register conditions: the low register condition started with a baseline of 130 Hz, the high register condition started 4 semitones higher (~164 Hz). The difference is comparable to values obtained for male speakers in the production study above.

Generally, f0 was manipulated such that the start of the pitch rise was located at the offset of /inna/ towards two different f0 maximum locations for each set. In the early peak condition, f0 reached its maximum at 1/3 of the way into the vowel (penult vowel in set PU and final vowel in set F). In the late peak condition, f0 reached its maximum at 2/3 of the way into the respective vowel. Note that the alignment differences exceed those typically found in the production study above in order to maximise a potential effect of alignment within the syllable. The maximum f0 value was set to be four semitones higher than the baseline (164 Hz for low register and 206 Hz for high register, respectively). These values are comparable to the typical rise excursions obtained for male speakers in the

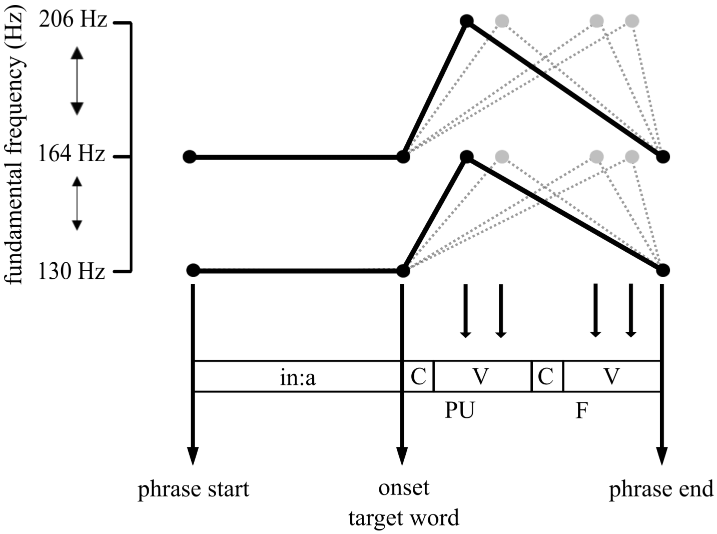

Figure 5.11: Schematised representation of the manipulation conditions displaying the differences in pitch register, discrete peak alignment (penultimate = PU, final = F), and gradual peak alignment: small arrows indicate early and late alignment within the syllable, respectively.

production study above. After reaching its maximum, f0 fell towards the baseline located at the end of the target word.[6]

These manipulations resulted in 32 stimuli (4 target words * 2 pitch registers (low vs. high) * 2 peak alignments in discrete terms (penult vs. final) * 2 peak alignments in gradual terms (early vs. late)) (cf. Figure 5.11).

Participants were seated in front of a computer screen in a quiet room at the Ibn Zohr University in Agadir. They were told that they were going to listen to a robot that speaks Tashlhiyt reasonably well, but has difficulties with producing the difference between statements and questions. Participants were asked to decide whether they would consider the sentences produced as statements or questions by pressing one of two buttons.

The experiment was run using Superlab (Haxby et al. 1993). At the beginning of each trial, a fixation stimulus consisting of a '+' was presented in the centre of

[6] Since the start of the rise and the end of the fall were not identifiable in the production data in a reliable way, the start of the rise and the end of the fall have been fixed to the offset of /inna/ and the end of the utterance accordingly. The potential implications of this methodological choice will be discussed below.

the screen for 1500 ms during which participant heard the stimulus. Following this, two sentences appeared on the right and left side of the screen. On one side the statement was displayed in blue (e.g. inna baba !), on the other side the question was displayed in red (e.g. inna baba ?). Both were presented in Latin script. The position of the question and the statement was kept constant within participants, but was counterbalanced across participants. Participants had to press the left or right button on the computer keyboard matched with the respective sentence modalities displayed on the screen. After response delivery, a blank screen appeared for 500 ms.

Each participant started with a training session, in which all combinations of pitch registers and peak alignments (discrete and continuous) were presented once. In subsequent test blocks, each target word in each of the manipulation conditions was repeated five times and presented in randomised order resulting in 160 data points per participant.

5.6.1.3 Statistics

All data was analysed with generalised linear mixed models, using R (R Core Team 2015) and the lme4 package (Bates et al. 2015). To analyse responses categorically, mixed logistic regression models were used with rating (question or statement) as the dependent measure. Pitch register (low vs. high), discrete peak alignment (PU vs. F), and gradual peak alignment (early vs. late), word, and mean-centred repetition were included as fixed effects. Additionally, a term for random intercepts for participants was included, which quantifies by-participant variability, as well as random slopes for the fixed effects pitch register, discrete peak alignment, and gradual peak alignment for each participant. Models including the main effect / interaction of interest were compared to the same models with no main effect / no interaction via Likelihood Ratio Tests (LRT) to determine p-values.

5.6.2 Results and discussion

Overall, participants rated the stimuli as corresponding to questions in 43% of the cases indicating a slight bias towards rating the stimuli as statements. This may be due to the declarative syntactic structure of the utterance (no interrogative marker). Regardless of this bias, there was a significant effect of pitch register ($\chi^2(1)=7.5$, p=0.006), such that items with a high pitch register were significantly more often rated as questions (58% vs. 28%). There was a significant effect of discrete peak alignment ($\chi^2(1)=8.9$, p=0.003), as well, such that items with the f0 peak on the final syllable were significantly more often rated as questions than statements (61% vs. 25%). Gradual peak alignment did not have a significant effect on ratings. F0 peaks early in a respective syllable were rated as corresponding to questions comparably as often as f0 peaks late in the respective syllable (44% vs. 41%) (χ^2(1)=1.16, p=0.28, cf. Table 5.9).

Table 5.9: Mean proportions of question ratings as a function of pitch register (low vs. high) and peak alignment (discrete and gradual).

Register	Discrete Alignment	Gradual Alignment	Responses		
Low	PU	early	0.16	0.15	0.28
		late	0.14		
	F	early	0.46	0.40	
		late	0.33		
High	PU	early	0.31	0.36	0.58
		late	0.41		
	F	early	0.82	0.79	
		late	0.77		

Figure 5.12 illustrates the overall results (left panel) and the listener-specific results (right panels) for pitch register and discrete peak alignment. As can be seen, the overall effects of register and discrete alignment appear to be additive, with a final pitch peak in a high register being the preferred question type, and a penultimate pitch peak in a low register being the least preferred question type.

However, there appears to be no clear-cut distinction between questions and statements. Even the least preferred intonational pattern for questions (low register and pitch peak on PU) shows a considerable amount of question ratings (14%). Even though these general trends are statistically generalisable, there is a considerable amount of variation across listeners. Consider the listener-specific patterns in Figure 5.12 (right panels). Most listeners show a clear bias towards rating the high register condition as more likely to be a question (black lines are above the grey lines). In fact, some listeners show almost no question ratings when the register is low (listeners 2 and 5). Listeners 6, 8, and 9, however, seem to show a much weaker effect of pitch register. With regard to timing differences, most listeners show a clear bias towards rating sentences with the pitch peak

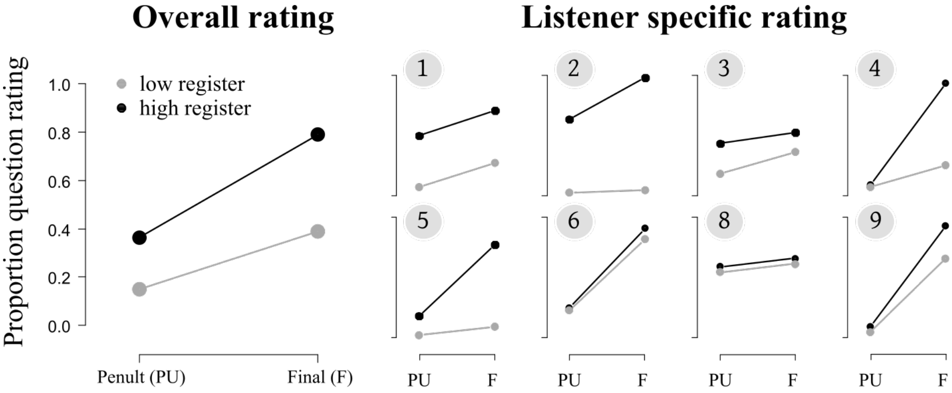

Figure 5.12: Ratings as a function of pitch register (low vs. high) and discrete peak alignment (in PU and F, respectively). Left panel displays overall results. Right panels display results for each listener individually (listener 7 was excluded as mentioned above).

aligned within the final syllable more likely to be a question than statements. Some listeners show almost no question ratings for pitch peaks aligned with the penult and, conversely, almost no statement ratings for pitch peaks aligned with the final syllable (listeners 4, 5, and 9). This effect appears to interact with register: the preference for questions with a pitch peak in the final syllable is stronger for the high register condition, as is very clearly illustrated by the patterns displayed by listeners 2, 4, and 5.

To sum up, two main factors are identified that affect the perception of morphosyntactically ambiguous sentences. First, contours in a high pitch register are perceived more frequently as questions than contours in a low register. This matches the strong pitch register differences found in the production study discussed in §5.4. Second, contours with pitch peaks on the final syllable are perceived more frequently as questions than contours with pitch peaks on the penult. However, even contours with a pitch peak on the penult appear to be acceptable contours for questions, reflecting the variation in pitch peak placement found in the production experiment. More gradual differences in tonal alignment within the syllable did not affect ratings. This null result could be due to the nature of the experimental design. Since both pitch register and discrete alignment are perceptually very prominent and represent sufficient cues to perform the task, listeners may not pay attention to subtle alignment differences. The present results are clear evidence for the rather discrete nature of peak alignment in our data. The relative peak delay is not as relevant as the position of the peak in a particular syllable.

As has been acknowledged, the low pitch target preceding the peak was set at the offset of /inna/, and the low pitch target after the peak was set to the end of the utterance. This alignment results in asymmetries in the steepness of rises and falls, across alignment conditions, with shallow rises and steep falls in peaks on the final syllable and steep rises and shallow falls in peaks on the penult. This potentially confounds the manipulation of the actual peak position (cf. Figure 5.11). Thus, the results could be interpreted as shallow rises and steep falls being more likely to be interpreted as questions than steep rises and shallow falls. The present investigation cannot rule out that the shape of the rise-fall affects listener ratings, but there are two arguments counter to this interpretation. This interpretation is not in line with the production results: rises in questions are consistently produced with substantially larger pitch excursions than statements. At the same time, the rise in pitch appears to start roughly at the same time across questions and statements. Taken together, these patterns result in substantially steeper rises in questions than in statements. Moreover, since falls are frequently truncated, the perceptual relevance of the fall is generally questionable.

5.7 Summary

The present chapter has explored the acoustic parameters associated with the distinction between contrastive statements and questions on the one hand, and information-requesting y/n questions and confirmation-seeking echo questions on the other. Production data revealed that, compared to statements, questions (a) had a higher pitch level and a greater pitch range, and (b) were more often realised with the pitch peak on the final syllable. In statements, the pitch peak occurred more often on the penult. Furthermore, there was a tendency for (c) the pitch peak in questions to be realised later within the syllable than in statements. Comparing questions types, echo questions were found to have a lower pitch level than y/n questions and a higher pitch level than corresponding statements. Other than that, the two question types revealed comparable pitch ranges and comparable peak alignment patterns.

The pitch register differences were consistent within and across speakers and appear to be a robust cue for disambiguating questions from statements, in both production and perception. In terms of pitch peak alignment there was a significant difference across sentence modalities in production. The perception results showed that listeners use pitch peak alignment in discrete terms (i.e. syllable-based alignment) to guide their perception of sentence modality, although pitch peaks in both syllable positions (on penult or final syllable) are acceptable locations for the pitch peak in both questions and statements.

Apart from this systematic correlation of tonal placement and sentence modality, there was considerable variation in peak alignment, both within and across speakers. Additional evidence presented by Grice, Ridouane & Roettger (2015) revealed that lexically determined segmental factors such as syllable weight of the final syllable and the sonority of syllable nuclei were relevant determinants of tonal placement. The investigated tones prefer to be realised on heavy syllables and sonorous elements. While syllables in other languages such as German or English usually contain at least a sonorant consonant, Tashlhiyt allows any type of segment in syllable nucleus position. The question arises as to how tones align with the segmental material when there are no sonorants available in the tone bearing word. The following chapter will explore these cases in detail.