Assessing incomplete neutralization of final devoicing in German

Röttger, T. B.[1], Winter, B.[2], Grawunder, S.[3], Kirby, J.[4] & Grice, M.[1]

[1] I*f*L Phonetik, University of Cologne, Herbert-Levin-Str. 6, D-50931 Köln, Germany

[2] Department of Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA

[3] Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6 D-04103 Leipzig, Germany

[4] School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Charles Street, Edinburgh, EH8 9AD, Scotland (U.K.)

Manuscript correspondence:

Timo B. Röttger

timo.roettger@uni-koeln.de

Tel.: +49 (0)221/4707047

Fax: +49 (0)221/470-5938

Herbert-Levin-Str. 6, D-50931 Köln, Germany

Abstract

It has been claimed that the long established neutralization of the voicing distinction in domain final position in German is phonetically incomplete. However, various studies leading to this claim have been criticized in terms of their methodology. 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. Here we bias the auditory stimuli against the phenomenon by manipulating the duration of the preceding vowel. While experiment 2 replicated our findings, experiment 3 failed to replicate incomplete neutralization statistically, even though we found numerical tendencies into 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 incomplete neutralization, the small effect sizes obtained further highlight the limits of investigating incomplete neutralization emphasizing the limited importance of this phenomenon for everyday speech communication. We argue that without postulating functional relevance, incomplete neutralization can be accounted for by recent models of lexical organization and is compatible with formal phonological models that entertain unpronounced projection relations.

Keywords: German; final devoicing; incomplete neutralization; obstruent voicing;

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. This asymmetrical distribution is commonly described in terms of final devoicing, a process that is sometimes viewed as purely phonological, and therefore as a discrete computation. Final devoicing in German[1] has been called the "universally recognized archetype of phonological neutralization" (Fourakis & Iverson, 1984: 141), and it has been described as a "classic example of a phonological rule" (Wiese, 1996: 204). In terms of generative phonology, this neutralization has been formalized in the following way:

(1) [+son] → [-voiced] / __ # (Vennemann 1968: 179)

In traditional formal theories of phonology, *Rad* ('wheel') and *Rat* ('advice; council') are thought to differ in their "underlying" lexical representation. Rules or constraints cause the voiced stop to become indistinguishable from the corresponding voiceless stop on the surface. In other words, neutralization of the final voicing distinction is assumed to be phonetically complete, such that the resulting two segments are surfacing as homophonous. However, numerous researchers have argued that German, in fact, shows *incomplete* neutralization. These studies suggest that words such as *Rad* and *Rat* are characterized by small acoustic and articulatory differences (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 evidence also suggests 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, ranging from classic works such as Jespersen (1913) and Trubetzkoy (1939), to modern descriptions such as Zifonun et al. (1997) and Wiese (1996). These phonological accounts assume abstract phonological categories and thus lack gradient phonetic information. Accounts based on this view have problems incorporating

[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".

[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 which corresponds 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.

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) have led to 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 state that incomplete neutralization poses "a threat to phonological theory" (see also Port & Leary, 2005).

Irrespective of its phonological implementation, the predominant response to incomplete neutralization studies has been skepticism. The numerically small effect sizes have raised serious methodological concerns (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.

Since some early studies found no evidence for incomplete neutralization (Fourakis & Iverson, 1984; Jassem & Richter, 1989), the debate has been taken to be settled (e.g., Kohler, 2007, 2012). Others, however, continue to collect evidence (e.g., Piroth & Janker, 2004), and yet others conduct follow up studies (e.g., Kleber, John, & Harrington, 2010). Moreover, research is 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)) and incomplete neutralization of other processes (Bishop, 2007; Braver & Kawahara, ms; de Jong, 2011; Dinnsen, 1985; Gerfen, 2002; Gerfen & Hall, 2001; Simonet et al., 2008). Thus, the debate surrounding incomplete neutralization is still very much ongoing. Our first and foremost aim is to provide methodologically sound evidence that will place the debate surrounding incomplete neutralization on a firmer footing. To do this we address the methodological and conceptual concerns raised against previous studies. Our second aim is to (re-)evaluate the potential consequences of incomplete neutralization for linguistic theory.

In section 2, we summarize previous empirical findings as well as their critiques, with a particular focus on Fourakis & Iverson (1984) and Jassem & Richter (1989). In section 3, 4, and 5 we discuss the results of three production experiments that were inspired by Fourakis & Iverson's study. Section 6 presents the results of a perception experiment. In section 7, we discuss implications of this work for an assessment of the status of incomplete neutralization in German, as well as implications for the relationship between phonetics and phonology.

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 across languages the duration of the preceding vowel has been shown to be the most reliable correlate of "voicing" in syllable final position. Thus in the present study we shall focus on this acoustic parameter.

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, the differences in 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, whilst Warner and colleagues (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 out by the participants, 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 auditorily presented 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. They measured the duration of the preceding vowel, voicing-into-the-closure/frication,

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 Röttger (2011).

closure/frication duration, and 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., Wiese, 1996; Kohler, 2007; Kohler, 2012). But do these studies, in fact, represent conclusive counter-evidence?

There are several issues surrounding these studies. For example, FI did not use minimal pairs. They compared words such as *mied* and *riet* (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 measured acoustic differences due to voicing are confounded with durational differences because of the initial consonants, ultimately leading to a reduction of statistical power.

Both studies furthermore take their null results as evidence for the absence of incomplete neutralization. There is a logical problem with "accepting the null", and most researchers believe 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 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 of the individual tests there were only a few data points. Indeed, an across-speaker re-analysis of the published data by FI conducted by Port and Crawford (1989) did find significant differences consistent with incomplete neutralization. Given the low statistical power (because there were no 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 (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 & Lahiri (1992) failed to find a subphonemic difference, but Warner et al. (2004), with more speakers, did find significant 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. While the studies by FI and JR certainly suggest that effect sizes for incomplete neutralization are small, they do not present *conclusive* counter-evidence against the phenomenon.

With regard to the perceptibility of incomplete neutralization, previous studies tested identification accuracies in forced choice tasks. In general, 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.

Warner et al. (2004) provided evidence for Dutch that within the context of a forced choice experiment, listeners can use an acoustic cue to discriminate voiced and voiceless stops, even if the cue does not actually occur in non-neutralized positions. This suggests that listeners locate the only possible perceptual cue available in the perception experiment and use this cue for their response decisions even though the cue does not appear in natural productions (see also Broersma, 2005). In other words, the results of perception studies seem to suggest a heavy influence of task demands.

Moreover, previous studies used auditory stimuli for the perception experiments which come from a small set of speakers (e.g., Port & Crawford, 1989), or even from only a single speaker (Kleber, John, & Harrington, 2010). This, together with many repetitions, gives participants ample opportunity to familiarize themselves with speaker characteristics and this in turn might make it very easy to detect those subtle cues for voicing in a neutralizing context and enhance the likelihood of participants using small and subtle cues that they would not use in listening situations outside of the laboratory.

So even though there is evidence that listeners are able to exploit small subphonemic cues to distinguish devoiced from voiceless stops in final position, one should be careful about its interpretation. While some see this as genuine evidence for incomplete neutralization as a perceptual phenomenon with potential real-world relevance, some are more inclined to view it as the result of task demands (e.g., Slowiascek & 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 large variability suggest that incomplete neutralization does not have a functional relevance in everyday communicative situations.

To sum up, previous studies reporting on incomplete neutralization have been criticized in terms of methodology. 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 the effect. The present study aims to circumvent those concerns as follows:

The present study

Our production studies are inspired by Fourakis and Iverson (1984)'s 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 (Perre, Midgley, & Ziegler, 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998). Given that FI and JR used real word stimuli, literate speakers inevitably know their written forms. Thus, the effect found in Port and Crawford's (1989) re-analysis of FI could be due to automatic activation of orthographic representations. In our follow-up study on FI, we used pseudowords, such as *Gobe* or *Gope,* instead of real words. In this study subjects are presented with a plural form in which the target consonant is intervocalic (/go:be/), and were instructed to produce the singular form (/go:p/) 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. Participants might merely imitate the acoustics of the stimulus they hear, i.e. the plural form (for an overview of the speech accommodation literature, see Babel, 2009). To address this issue we replicated our findings in two additional experiments, eliminating this potential confound. Moreover, given the small effect sizes, the functional relevance of incomplete neutralization for speech communication is at least questionable. To evaluate the perceptibility of incomplete neutralization in a more economically valid design, we replicated earlier perception studies utilizing a number of different voices.

If we were able to show that speakers consistently fail to fully neutralize the voicing contrast in final position and listener are able to distinguish between these realizations (i.e.

those that match an intervocalic voiced consonant and those that match an intervocalic voiceless consonant), then the incompleteness of the neutralization warrants explanation. Even if it arguably has no functional utility for communication, we, nevertheless, have to account for such an effect. It might turn out that incomplete neutralization does not inform us about abstract phonological representations but more about the way phonological forms are stored or accessed by the speaker-hearer.

3. Production Experiment 1

3.1 Methodology

3.1.1 Participants

16 native speakers of German participated in the experiment (mean age: 25 years; nine women), all undergraduates or PhD students in the humanities that were 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. No participant had heard of incomplete neutralization before the debriefing at the end of the experiment.

3.1.2 Experimental procedure

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 consistent difference in final devoicing 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 & Röttger, 2011). After the written instructions, the remaining procedure was conducted auditorily. In each trial, participants first heard a stimulus sentence such as (2) via headphones and then produced a corresponding sentence such as (3).

(2) Plural stimulus

Aus	*Dortmund*	*kamen*	*die*	***Drude***.
From	Dortmund	come.3PL.PST	DET.1SG.M.NOM	NONCE-PL

'From Dortmund came the NONCE-PL.'

(3) Singular response

Ein	**Drud**	wollte	nicht	mehr.
DET.1SG.M.NOM	NONCE-SG	want.3SG.PST	NEG	longer

'One NONCE-SG refused 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 head phones. After an inter-stimulus-interval of 500ms 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.

Before 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 stop 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 minutes.

3.1.3 Stimuli

The experimental items consisted of 24 pseudoword pairs such as (4-6) (see Appendix A):

(4) Wiebe [viːbə] vs. Wiepe [viːpʰə]
(5) Gaude [gaʊ̯də] vs. Gaute [gaʊ̯tʰə]
(6) Gage [gaːgə] vs. Gake [gaːkʰə]

All pseudowords were trochaic and complied with German phonotactic rules. There were eight bilabial, seven alveolar and nine velar stimuli pairs, each containing one of the vowels /aː, oː, uː, iː, aʊ̯/. Each experimental item was introduced as a masculine noun inflected for

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' and *Stift*/*Stifte* 'pen/pens'). We explicitly 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*, 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 wanted to norm our stimuli with respect to their morphology. A list was given with the intended singular forms of the stimuli 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. To what extent a stimulus was identified as schwa-plural or not will be included into the analyses presented below.

To further alleviate the concern of hyperarticulation due to minimal pair awareness, there were 96 fillers (2/3 of the total stimulus set), 70% of which contained an umlaut vowel. These fillers introduce an additional choice that participants have to make, as plural forms with umlaut vowels sometimes do and sometimes do not require a vowel change (e.g., *Turm* > *Türme; Bär* > *Bären*). This makes the filler items more salient and defers attention away from the critical stimuli. Forty different city names were embedded in the carrier phrase (randomized over stimuli pairs) to introduce an additional distracting element. The rest of the carrier phrase was kept 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 male native speaker of German (second author) 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 stimuli pair were always within different blocks. At the beginning of the experiment, each participant was randomly assigned to one of eight block orders.

3.1.4 Acoustic analysis of stimuli

We performed an acoustic analysis of the plural stimuli that were presented to participants. Using Praat 5.2 (Boersma & Weenink, 2009), we analyzed 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 (f_0 is known to be higher after voiceless/fortis stops, Kohler, 1984).

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 about 42 ms (SE = 2.3) longer for voiceless stops ($\chi^2(1)=65.57$, $p<0.0001$). Closures were about 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 ($\chi^2(1)=0.37$, $p=0.99$). Also, there were no differences of mean f_0 ($\chi^2(1)=2.17$, $p=0.7$) and of 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. This means that there are at least three 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.

3.1.5 Acoustic analysis of responses

The responses of participants were digitized at a sampling rate of 44.1 kHz (16bit). The vowel durations of the vowels preceding the stops were annotated by the first and second 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 nasals (/m/ and /n/), palatal approximants (/j/) and liquids (/l/ and /ʁ/) (e.g., [muːp], [jiːt], or [fʁaːt]). 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 hyperarticulation, 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.

4 All of these analysis are based on Likelihood Ratio tests based on 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.

3.1.6 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 (Experiment 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., 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), and for this fixed effect, we included random slopes for participants and items (this quantifies by-participant and by-item variability in the effect of VOICING). The random effects were allowed to correlate.

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 would lead to a significant interaction with VOICING, this would indicate that the amount of incomplete neutralization depends on prosodic conditions and could thus be the result of hyperarticulation. VOWEL QUALITY and PLACE OF ARTICULATION (bilabial vs. alveolar vs. velar) were also included to explain residual variance. Since participants might perceive some singular forms as better matches to their corresponding plural forms than others, there might be processing differences that can be accounted for by including the results of our stimulus norming as an additional control variable, PLURAL ASSOCIATION.

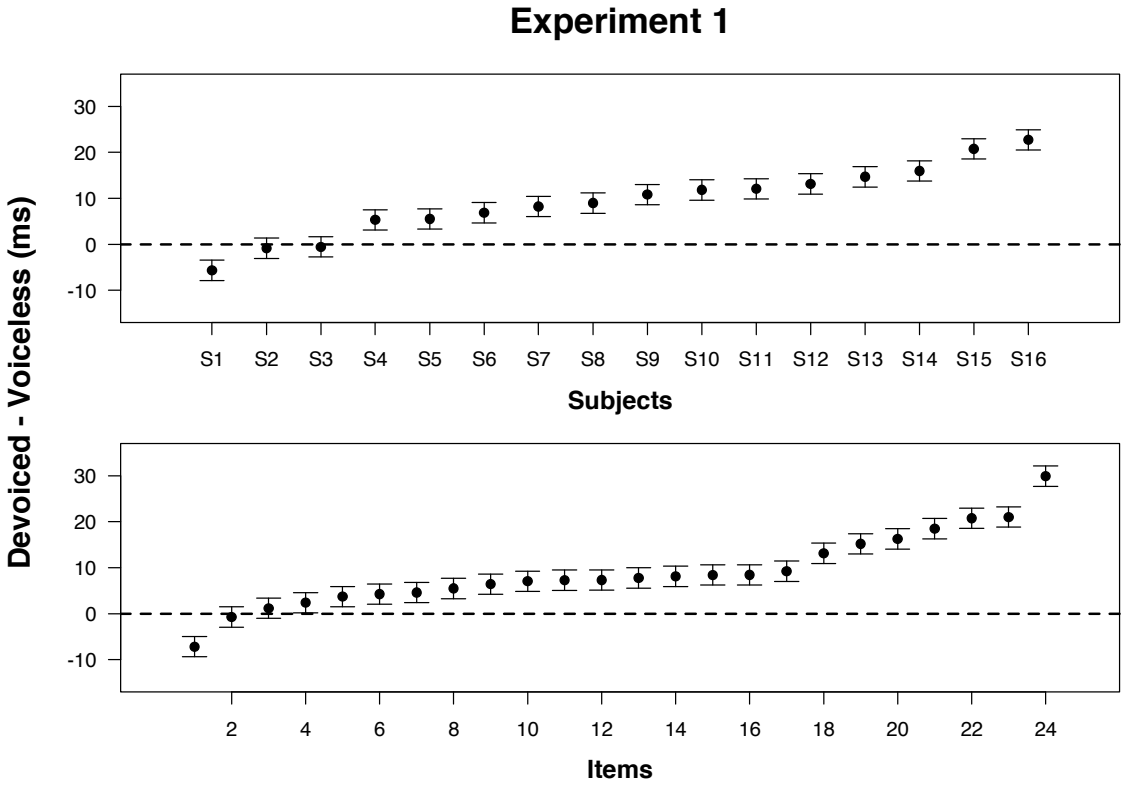
We first tested whether VOICING interacted with the control variables by performing a likelihood ratio test of the model with interactions against the model with only the main effects. We then excluded the interaction between VOICING and all control variables. P-values were generated using likelihood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

3.2 Results

VOICING had a significant effect on production of the vowel duration in the singular form ($\chi^2(1)=13.76$, $p<0.0002$), with vowels estimated to be 8.6 ms longer before stops corresponding to voiced stops in the plural form ('devoiced') rather than to the voiceless stops in the plural form ('voiceless') (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 a vowel duration difference in the predicted direction. 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 any of the prosodic variables (Accent Type and Prosodic Boundary)[5], Vowel quality, Place of Articulation, or the Plural Association variable.

Figure 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 §3.1.6. The dashed line indicates 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 did subset analyses where we excluded all responses where the pseudoword was either incorrectly remembered (e.g., substituting the vowel /i:/ for /e:/), or produced with a lot of hesitation. This led to a total 10.02 % data loss, quite a considerable reduction of the size of the dataset. Nevertheless, even if these responses were excluded, the main results still holds ($\chi^2(1)=13.34$, $p<0.0003$), with vowels being 8.6 ms longer (SE = 2.04 ms) before devoiced stops.

[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).

Our results indicate a successful extension of the FI study of incomplete neutralization in German. However, there is a potential confounding effect inherent in our design. As the analyses of the stimulus acoustics above show (§3.1.4), there were pronounced vowel duration differences between intervocalic voiceless and voiced stops in the stimuli that were presented to our participants (mean = 28ms). Could it be that participants were influenced by the acoustics of the intervocalic stimulus they heard, i.e. the plural form?[6] Experiment 2 and 3 were conducted to address this issue.

4. Production experiment 2

4..1 Material and methods

4.1.1 Participants and experimental procedure

16 speakers participated in Experiment 2 (mean age: 27 years; 9 women, 7 men). Background details of participants are as stated for E1. None of them had participated in E1. All details of the procedure were the same as in the previous experiment if not stated otherwise.

In Experiment 2, we used a different carrier sentence. For each trial, participants first heard a sentence such as (7) and then produced a corresponding sentence such as (8).

(7) Plural stimulus

Peter	weiß	nun,	wie	die	Bauge	aussehen.
Peter	know.3SG.PRS	now	how	DET.1SG.M.NOM	NONCE-PL	look

'Peter knows now what the NONCE-PL look like.'

(8) Singular response

Denn	nur	der	Baug	sieht	so	aus.
Because	only	DET.1SG.M.NOM	NONCE-SG	look.3SG.PRS-	like	-PART

'As only the NONCE-SG looks like this.'

[6] If we replace VOICING with vowel duration of the corresponding plural form, this becomes significant ($\chi^2(4)=49.6$, $p<0.0001$). For every additional millisecond in the stimulus, participants lengthened their vowels by 0.23 ms (SE = 0.05 ms). As VOICING and STIMULUS VOWEL DURATION are obviously correlated with each other, they cannot be put into the same analysis due to collinearity. However, we can compare the relative evidence between two models using evidence ratios (Richard, Whittingham, & Stephens, 2011). Using the R package *qpcR* (Spiess, 2012), the evidence ratio between the model with VOICING and the model with STIMULUS VOWEL DURATION was found to be 9.37. This can be interpreted as showing that the VOICING model has 9.37 more support than the STIMULUS VOWEL DURATION model (10 is commonly interpreted as analogous to "significance", Richard et al., 2011). This already points towards VOICING having a stronger influence than STIMULUS VOWEL DURATION, but we cannot firmly conclude this as the two factors are correlated with each other in our design.

4.1.2 Speech material, stimuli manipulation, and norming

The experimental items consisted of 48 pseudoword pairs (cf. Appendix B). There were 24 stimuli pairs with labial and 24 with velar stops, each of which followed one of the vowels /iː, eː, aː, aʊ, oː, uː/.[7] Stimuli were balanced for vowel quality. Each experimental item was introduced as a masculine noun inflected for plural as stated for E1. Similar to E1, a norming study shows that the schwa-plural (/-ə/) was the most frequent response pattern (82% of all responses). As opposed to E1, there were no fillers, thus the contrast between the corresponding members of a minimal pair was very obvious for the participants. This might lead 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 stimuli pairs were spoken by a trained native speaker of German (male) along with the demonstration and practice items in a sound-treated booth recorded with an AKG C420 III microphone.

For the manipulation, we calculated the percentage of the vowel duration preceding the voiceless stop in comparison to the vowels preceding the voiced stop for each minimal pair. The mean difference was ~16%. This value was taken as a baseline. We then manipulated the vowel durations using PSOLA resynthesis in Praat 5.2 (Boersma & Weenink, 2009). Minimal pairs were manipulated and grouped into four sets: Set A obtained a difference in vowel duration of 32% (henceforth enhanced), that is vowels preceding voiced stops were 32% longer than vowels preceding voiceless stops. Set B obtained a difference in vowel duration of 16% (henceforth original) resembling the original data. Set C obtained no difference at all (henceforth neutralized), so vowel duration as a cue for voicing was neutralized. Set D was manipulated so that the voicing distinction realized through vowel duration was reversed resulting in a mirror image of original (henceforth reversed). In other words after the manipulation set D contained stimuli with 16% longer vowels *preceding voiceless stops* than preceding voiced stops.

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 experiment. The norming study confirmed that the voicing contrast is very easy to perceive

[7] Acoustical analyses of the stimuli show that the vowel preceding the critical stop was 23.78 ms (SE = 2.62) shorter before voiceless stops ($\chi^2(1)=48.547$, $p<0.0001$). The closure duration was 13.7 ms (SE = 1.74) longer for voiceless stops ($\chi^2(1)=40.32$, $p<0.0001$). VOTs were on average 47.34 ms long for voiceless stops (SE = 1.75, $\chi^2(1)=208.28$, $p<0.0001$). All of the voiced stimuli had voicing during the closure. So as stated for E1, there were robust cues for the voicing status of the critical stop in intervocalic position. There was no interaction between manipulation condition and voicing for the parameters ($\chi^2(1)\leq 3.91$, $p\geq 0.27$), thus there were no differences of intervocalic voicing cues between conditions.

for both manipulation conditions: Participants did not make any errors in identifying the voicing category. Even though we manipulated one perceptual cue for the voicing distinction, participants relied on other cues like voicing of the closure, VOT, and closure duration.

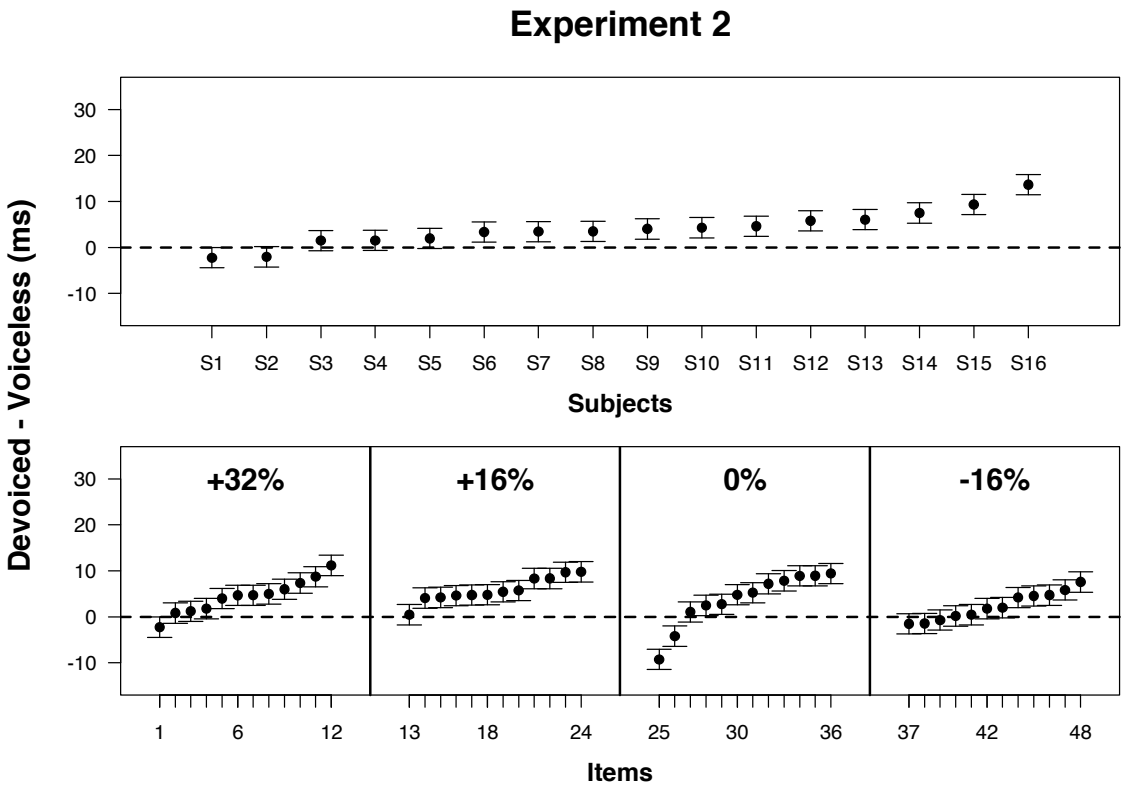
4.1.3 Stimuli presentation, acoustic analyses of responses and statistics

All stimuli 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. A subset of these items was repeated twice in block 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, PLURAL ASSOCIATION, VOWEL QUALITY, PLACE OF ARTICULATION, 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 voiced and voiceless stops became 1.49 ms smaller (SE = 0.57 ms). Interestingly, even the neutralized and the reversed condition, that show either no or an unnatural duration cue for voicing, elicit an incomplete neutralization effect in the expected direction. There also was 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 vowels (pooled across different manipulation conditions, cf. Figure 2).

Figure 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. The dashed line indicates no difference between voiced and voiceless stops.



Similar to the results from 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, REPETITION, PLACE OF ARTICULATION, and VOWEL QUALITY did not have an effect.

These results 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, showing that participants were not solely influenced by accommodation.

However, since this experiment exhibited a between-item design, all speakers were prompted with items of 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"). Additionally, the manipulation conditions were not perfectly balanced,

i.e., there was an overall duration advantage for vowels preceding voiced vowels of +16% (=32% + 16% + 0% - 16%). This advantage is actually in favor of showing incomplete neutralization, as participants might have adapted to the overall 16% vowel duration difference. 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 Material and methods

5.1.1 Participants and experimental procedure

16 speakers participated in Experiment 3 (mean age: 24 years; 10 women, 6 men). Background details of participants are as stated for E1 and E2. None of them had participated in the previous experiments. All details of the procedure were the same as in E2 if not stated otherwise.

5.1.2 Speech material, manipulation, and norming

The 24 experimental item pairs consisted of a subset of the items used in E2 (cf. Appendix B).[8,9] Again, we manipulated the vowel durations via using PSOLA. Each minimal pair was manipulated twice resulting in two stimuli 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 binary forced-choice experiment. 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 both manipulation conditions: 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. And, as was done for E2, we also conducted a norming study with respect to plural association. For this subset, participants formed the schwa-plural (/-ə/) in 82% of cases.

[8] 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.

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

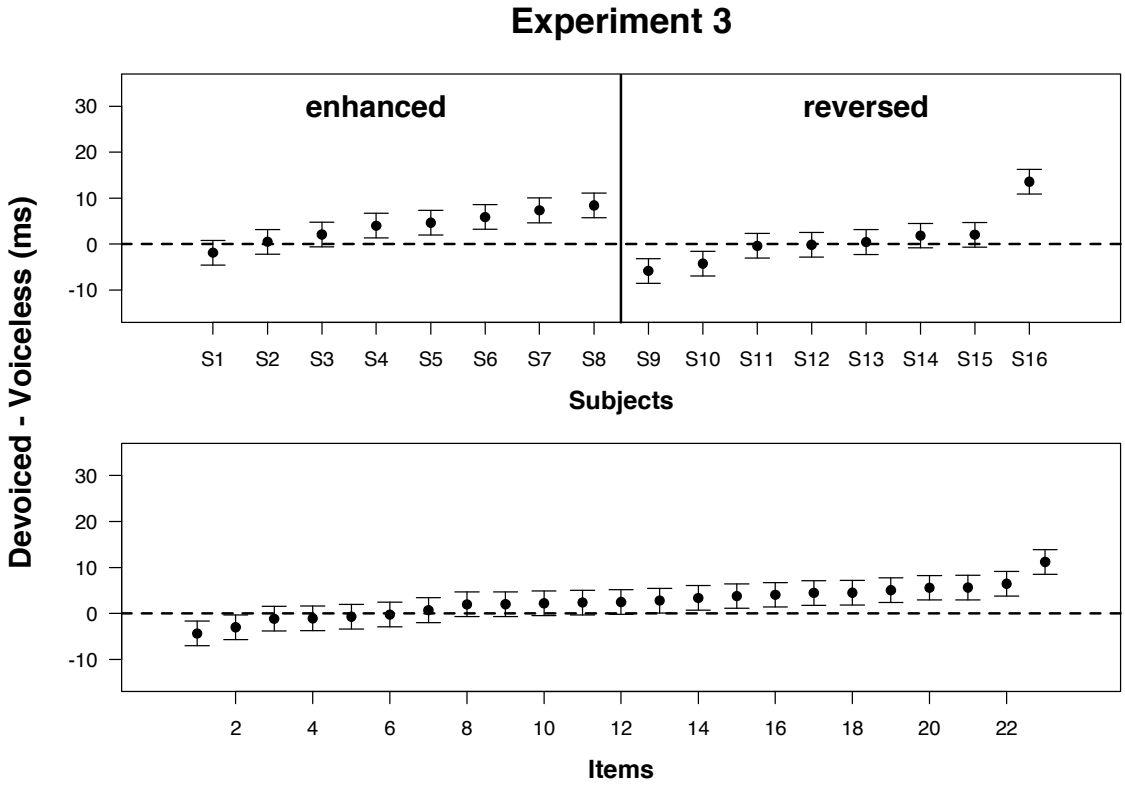
5.1.3 Stimuli presentation, acoustic analyses of responses, and statistics

Subjects were randomly assigned to group A and group B. 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, albeit not significant, resembles the accommodation effect in E2. Crucially, 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. Figure 3). Experiment 3 thus marks a failure to replicate the incomplete neutralization effect.

Figure 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 "enhanced" vs. "reversed" vowel duration. The dashed line indicates no difference between devoiced and voiceless stops. The lower plot only shows 23 items because one item had to be excluded due to coding error.



5.3 Discussion of production results

In Experiment 1, we found a difference in vowel duration depending on the intervocalic voicing of the presented plural 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 (*incomplete neutralization effect*) even in an experimental paradigm that minimizes the influence of orthography. This was achieved through auditory presentation (no visual presentation of orthography) and pseudowords (no pre-existing orthographic representation). Moreover, our analyses controlled for variation between different participants and items, and the pattern was found to be remarkably consistent across different individuals and stimuli. 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 obtained *incomplete neutralization effect* was not affected by experimental task demands such as repetition (participants did not produce a stronger or weaker effect with subsequent

repetitions). Moreover, the *incomplete neutralization effect* was not altered by prosodic characteristics or place of articulation, suggesting relative independence from these factors as well.

Finally, our debriefing indicated that participants were in fact thinking that the task was morphological – none were aware of the fact that we were looking specifically at minimal pairs such as *Gobe* and *Gope*. 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 large role. We can then safely conclude that we have replicated earlier incomplete neutralization experiments while avoiding the methodological shortcomings that may have impacted previous findings.

Experiment 2 replicated the findings of E1 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 though they were prompted with intervocalic cues running in the opposite direction. Although there was a statistically significant difference between the manipulation conditions, 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 E3. As has been observed repeatedly in the literature on final devoicing in Dutch and German, incomplete neutralization effects are brittle and difficult to detect with inferential statistics.

We now turn to the role of orthography. Given that literate adult speakers constantly and habitually associate phonological with orthographic forms (Perre, Midgley, & Ziegler, 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998), participants might have mentally generated orthographic representations "on the fly". Thus, a given participant that has just heard a pseudoword such as [goːbə] might have created 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 *had* orthographic representations as the input.

Although we minimized the role of orthography at least to the same extent as Fourakis and Iverson (1984) did, our use of pseudowords comes with its own set of problems. For example, as pseudowords are necessarily unknown and unfamiliar (and thus have a frequency of 0), they might be even more hyperarticulated (cf. Whalen, 1991, 1992, for frequency effects on segmental duration). 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 156ms (SD = 44ms), whereas Port and O'Dell (1985: 459) reported 202-305ms, and Charles-Luce (1985: 315) reported 184-211ms (1985: 315), suggesting that relative to these other experiments, our vowels were *less* hyperarticulated.

Furthermore, pseudowords always introduce the possibility of formal analogy to real words. The lexical asymmetry of different places of articulation and vowel qualities is the main source of potential analogical asymmetries. We addressed this issue by adding those effects to our statistical models and found no noteworthy interaction. This suggests that potential item-specific effects due to the make-up of individual pseudowords are marginal. Further, we checked whether certain stimuli are more difficult in terms of singular-plural formation than others causing any confound. We used data collected in a norming study to predict production results and found no effect of plural preferences either. We conclude that any processing difficulties due to idiosyncratic properties of the stimuli are of minor importance for our results.

6. Perception experiment

Our production experiments confirmed that neutralization of the voicing contrast in final position is incomplete. We have ruled out a number of potential reasons for this incompleteness. However, as it is discussed above, the fact that the neutralization is incomplete might not play any functional role in speech communication. To further our understanding of the perceptibility of incomplete neutralization, our fourth experiment sets out to replicate and extend earlier incomplete neutralization effects in perception. Previous studies used auditory stimuli which came from a small set of speakers, or even just a single speaker (e.g., Port & Crawford, 1989; Kleber et al., 2010). 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 perception experiment confronted listeners with all speaker voices of E1.

6.1 Material and method

6.1.1 Participants and experimental procedure

16 listeners participated in the experiment, none of which had participated in the preceding experiments. All participants were native speakers of German with no hearing deficits (mean age: 30 years; five woman, eleven men). Two of the participants were authors of this study (the first and the second author, both from the Cologne/Rhine region), neither of which 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., *Drud* vs. *Drut*). These were presented on the left and the right side of the screen (counterbalanced), and participants had to press a left or right button. 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 devoiced/voiceless distinction as well as long-term success over many trials and repetitions. To this end, we randomly sampled subsets (192 items) of the items in the production study until we gained a subset in which the incomplete neutralization effect for vowel duration was significant. We then constructed four lists. In each of the lists, each stimulus pair (e.g., *Wieb* vs. *Wiep*) appeared once. Also, each speaker appeared at least once. In order to make the task not too difficult, each voiced/voiceless combination came from the same speaker (e.g., *Wieb* and *Wiep* in list 1 were both from speaker 4). Given that there were 24 item pairs but only 16 speakers, 8 speakers appeared twice per block. Only target stimuli were included in the perception experiment (see Appendix A).

6.1.3 Statistics

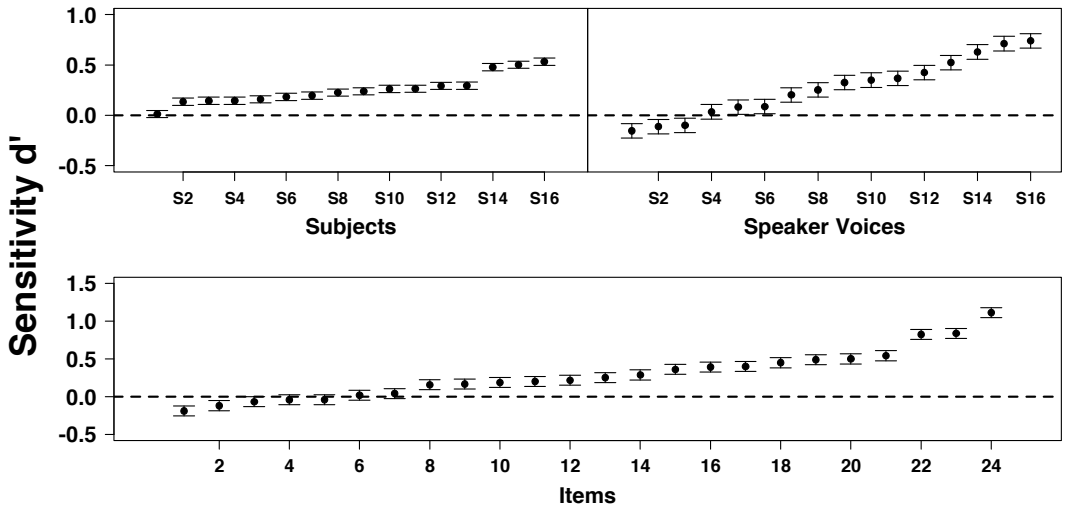
There are several ways of analyzing the present data. The traditional way works with d', a sensitivity index derived from Signal Detection Theory (Green & Sweets, 1966). d' takes into

account a subject's bias; good sensitivity is thought to be above 1. 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 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 dependent measure. As fixed effects we included mean-centered response times, trial order and repetition. As random effects, we included subject, item and speaker voice. If the *intercept* of this model is significantly above zero, we can conclude that people are able to perceive the voicing contrast in the neutralization context with above chance accuracy.

6.2 Results

Figure 4 displays d' per subject, item, and speaker voice. Overall, d' was fairly low and barely above 0. This indicates that recognition of voiced and voiceless final stops, if controlled for response bias, 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 voice ($t(15)=3.79$, $p=0.0017$), with estimates of 0.25 (SE = 0.036), 0.29 (SE = 0.066) and 0.27 (SE = 0.072).

Figure 4: Results of Experiment 4. d' sensitivity values arranged according to size for subjects (upper plot), speaker-voices (middle plot), and items (lower plot) with standard errors. The dashed line indicates 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).



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 *not* more likely to respond more accurately 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 TIMES ($\chi^2(3)=8.88$, $p=0.03$). 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 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

With just 55%, the accuracy averages are very low and barely above chance performance. In comparison to 99-100% accuracy averages for the intervocalic contrast obtained in the norming studies of E1-E3, this is a rather poor performance. Moreover, participants performed worse responding to devoiced stops (51%), which turned out not to be significantly different from chance performance. In turn, the overall significant accuracy scores might be due to a ceiling effect, i.e. participants were biased towards the voiceless category. Even though similar results were obtained in previous perception studies on incomplete neutralization (e.g., Port & O'Dell (1985) report mean accuracy values of 59%), these results show the lowest accuracy scores reported in the literature.[10]

[10] It might be argued that the very low accuracy scores might 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. In both areas, dialects or regional varieties are still present, and most subjects are exposed to them. In Saxony 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, 2006). This interpretation, however, stands in contrast to the very high accuracy scores in intervocalic position we found in the norming studies of E1-E3.

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 where voiced and voiceless stops are distinguished in final position, and as these minimal pairs most often have different syntactic contexts which help to disambiguate them (e.g., the adjective *tod* and the noun *Tot* would never appear in the exact same syntactic position), the cognitive effort necessary to detect these minute acoustic differences may not be justified.

The fact that accuracies are so low indicates that the role of incomplete neutralization as a perceptual phenomenon outside of a controlled laboratory context is negligible and unlikely to have a great deal of functional relevance in everyday speech communication.

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 (recall that we call those segments 'devoiced' and 'voiceless', respectively). Although there were studies investigating obstruent devoicing in German and other languages, the main focus of attention was on the German. As noted earlier, many of these studies have been criticized on methodological grounds. However, apparent counter evidence such as that put forward by Fourakis and Iverson (1984) or Jassem and Richter (1989) were at least as problematic as they might have been subject to a lack of statistical power (see §2). The aim of the present study was to put incomplete neutralization on a firmer footing by using a similar auditory design as Fourakis and Iverson 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, we have provided converging evidence that neutralization of German final stops is incomplete. Our finding that incomplete neutralization occurs even in a completely auditory task that uses pseudowords instead of real words was replicated in two additional experiments that mitigated the potential for accommodation to an intervocalic input stimulus. However, the failure to replicate the effect inferentially in Experiment 3 is indicative that this phenomenon is barely detectable with a small amount of speakers, especially if the experimental design biases against finding such an effect.

We also conducted a perception experiment, in which we found that although participants were able to distinguish devoiced versus voiceless final stops with above chance accuracy,

this was only barely so (55% as opposed to 98-100% intervocalically). It is quite possible that the overall effect found was in fact due to a bias towards voiceless stops. Thus, while the present experiments provide evidence for incomplete neutralization in production, it is highly questionable whether listeners can 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 (for a discussion of this, see also Winter & Röttger, 2011). 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 across languages) outnumber those finding counter-evidence. Wiese (1996: 205) commented on incomplete neutralization experiments as follows: "These results are rather tentative […] given that the recognition of non-neutralized devoicing was found in a minority of cases only". By now, we can – with relative confidence – say that this statement has been superseded. Positive results for incomplete neutralization characterize the majority of studies on this topic and several of the methodological issues have been successfully addressed.

Assuming that 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 on incomplete neutralization was motivated by the assumption that there are, in fact, voiced "underlying" segments and that the obtained differences have to be explained in terms of differences of these abstract representations. Recall that studies on incomplete neutralization, including the present experiments, have obtained very small effects elicited in a laboratory setting which could simply be the result of careful, hyperarticulated speech. Outside of the lab, in everyday communication, the obtained differences should be even smaller and thus even less perceivable. To assume that these small, barely noticeable acoustic differences are features of the abstract linguistic system strikes us as rather unparsimonious. However, as there are alternative approaches of incomplete neutralization, the phenomenon does not necessarily force one to make assumptions about the presence of underlying segments or to propose functional relevance of the phenomenon.

Orthography

Previous work has adduced evidence for a connection between orthography and incomplete neutralization effects. For example, items such as *seid* 'to be.PRS.2PL' and *seit* 'since' have been argued to exhibit incomplete neutralization effects (Port & Crawford, 1989; Port &

O'Dell, 1985), even though *seid* does not have any morphologically related forms where the /d/ surfaces in a non-neutralized position. The fact that there is no phonological evidence for a voiced stop in these cases has been interpreted by Fourakis and Iverson (1984) and Manaster-Ramer (1996) as showing that people base their pronunciations on the orthographic difference between the two words. Further evidence comes from Warner and colleagues (Warner, Good, Jongman, & Sereno, 2006), who demonstrated that phonologically identical words in Dutch were pronounced differently in accordance with voicing distinctions indicated by the orthography.

Even when using a completely auditory design, as in the present experiments, orthographic representations are problematic because a wealth of experimental studies showed that these become automatically activated in auditory tasks (Perre et al., 2009; Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998). Therefore, the possible influence of orthographic mental representations cannot be excluded as long as one works with a literate speaker population.

One approach to address the role of orthography is to compare incomplete neutralization effects in languages *without* an orthographic distinction of the neutralized elements to those languages *with* such an orthographic distinction. Dinnsen and Charles-Luce (1984) reported incomplete neutralization for Catalan, a language in which the neutralized contrast in question is not indicated by orthography, but the direction of the differences was different for the five speakers. Kopkallı (2007) and Wilson (2003) found no significant duration differences for word-final stops in Turkish (another language that makes no orthographic distinction) but there was a numerical tendency for longer vowels before devoiced stops. Although these effects did not reach statistical significance, this might have been overlooked due to small sample sizes. Kim and Jongman (1996) investigated the coda neutralization of /t, t^h, s/ to [t] in Korean, and found no evidence for incomplete neutralization *despite* an orthographic difference. Finally, Yu (2007) found incomplete neutralization for tone neutralization in Cantonese where there is *no* orthographic correlate. These studies paint a complicated picture, however, the numerical tendencies for Turkish, and the statistical effects for Catalan and Cantonese might suggest that at least some instances in which neutralization is incomplete are independent from considerations of orthography.

Ultimately, this debate does not appear to be about methodology regarding the stimulus material but more about the population tested. Given the automatic activation of orthographic representations, an experiment that presents stimuli to literate participants always has the orthography bias as a potential methodological confound. Recently, Grawunder and Lancia (2012) indeed tested preschoolers that have not yet acquired literacy. Preliminary results show

a significant incomplete neutralization effect, i.e. longer vowels before devoiced stops. This suggests that incomplete neutralization is unlikely to be a purely orthographic effect. Thus, we conclude that there is converging evidence for the relative independence of orthography and the incompleteness of neutralization.

Moreover, orthography might have long-term effects on the linguistic system (for an example, see Blevins, 2006). If speakers occasionally hypercorrect linguistic forms based on the written language (e.g., in reading situations), this might be enough to keep slightly altered representations within a speaking community. The incomplete neutralization effect in experiments such as ours would then not be a *direct* result of orthography (as an methodological artifact), but an indirect one.

Formal accounts of incomplete neutralization

Most formal attempts to incorporate incomplete neutralization have led to a proliferation of post-hoc assumptions. They either propose an extra phonological derivation process or assume an alternative rule ordering to account for the phonetic difference between voiceless and devoiced stops (e.g., Charles-Luce, 1985; Port & O'Dell 1985). Alternatively, Port and O'Dell (1985) propose a specific phonological feature (VOICE-F) for devoiced stops, which in turn leads to three different voicing categories: voiced, voiceless, and devoiced. But does incomplete neutralization really forces us to implement new rules and/or categories?

In a more recent account in the framework of Optimality Theory, van Oostendorp (2008) argues that incomplete neutralization can be captured by tubid representations of phonological outputs (Goldrick 2001). In this account output structures can be characterized in terms of structural relationships (*Projection*) and audible surface relationships (*Pronunciation)*. Projection relations are abstract relationships between segments and their features. Pronunciation relations are output relationships between the feature and the segment. They describe the actual phonetic output of a structure. This results in three different output categories: there are segments that are linked to the feature [+VOICE] and pronounced voiced (e.g., [d] in [ʁæ:dɐ] 'wheels'), segments that are linked to the feature [-VOICE] and pronounced voiceless (e.g., [t] in [ʁæ:tɐ] 'councils' or [ʁa:t] 'council'), and segments that are linked to the feature [+VOICE] but pronounced as voiceless in final position (e.g., [t] in [ʁa:t] 'wheel'). Because of their structural differences, [ʁa:t] 'council' and [ʁa:t] 'wheel' will phonetically surface as different from each other. In contrast to Port and O'Dell (1985), van Oostendorp's account assumes only two abstract phonological representations, a voiced and a voiceless category. The phonetic realization of the output results through an abstract link from

the phonological level of representation to the phonetic output. This account is able to leave phonology phonetics-free and captures incomplete neutralization in a parsimonious way.

Co-activation of paradigmatic neighbors

Van Oostendorp's account states that incomplete neutralization is a result of an output link of the feature specification of neutralized (i.e. [-VOICE]) segments, to the feature specification in non-neutralized (i.e. [+VOICE]) segments in intervocalic position. The nature of the link is, however, rather abstract. A similar way to account for incomplete neutralization is based on concrete links between lexical forms: In the framework of exemplar based models of speech processing it is assumed that the lexicon is rich and full of detailed interconnected lexical forms. There is evidence suggesting that lexical representations contain a lot of information, including detailed phonetic information of individual word forms (Brown & McNeill, 1966; Bybee, 1994; Goldinger, 1996, 1997; Palmeri et al., 1993; Pisoni, 1997) and completely inflected forms (e.g., Alegre & Gordon, 1999; Baayen, Dijkstra, & Schreuder, 1997; Butterworth, 1983; Bybee, 1995; Manelis & Tharp, 1977; Sereno & Jongman, 1997). These models of lexical organization and access assume that German speakers have inflected forms such as *Räder* 'wheels' in their mental lexicon. Due to its morphological relation 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 morphologically 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 minute incomplete neutralization effects that we observed. One might object that most previous evidence for activation of morphological neighbors comes from perception studies. However, Wright (2004), Munson (2007), and Baese-Berke and Goldrick (2009) demonstrate the effects of neighborhood density (i.e., the number of lexical forms in a speakers lexicon phonologically related to a target word) on speech production. In historical linguistics analogical effects from morphological paradigms are commonplace (e.g., Trask, 2007) and these ideas have been extended to synchronically observable phenomena (Benua, 1995; Burzio, 1994). One striking example was demonstrated by Yu (2007), where tonal near mergers in Cantonese were shown to be facilitated by interactions with their morphological neighbors.

The co-activation account makes testable predictions for future experiments. First, it predicts recency effects: if the response is delivered right away after 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. Second, there should be frequency effects: words that have very frequent morphological neighbors with voiced stops in intervocalic position should exhibit stronger incomplete neutralization effects than words with very infrequent morphological neighbors (see e.g., Bybee, 2001, for the role of frequency in analogy). Moreover, it predicts incomplete neutralization effects to be dependent on lexical/paradigmatic density. So a language with many voiceless paradigmatic neighbors of a target word, should surface with no or at least weaker incomplete neutralization effects than a language with many voiced paradigmatic neighbors (as is the case in German).

Interestingly the proposed co-activation account is not entirely incompatible with existing formal accounts: van Oostendorp (2008) attributes incomplete neutralization to an abstract link between the underlying voicing specification of a segment and its corresponding "neutralized" counterpart. This voicing specification is based on the paradigmatic relationship between two words (e.g., *Räder* and *Rad*). The co-activation account attributes incomplete neutralization to direct links between those word forms in a fully specified interconnected network. Obviously those accounts differ considerably in the way the lexicon is conceptualized, however, the basic idea appears to be compatible.

Recall, however, that the experimental modulation of these minute acoustic differences has its pragmatic limits due to the small effect sizes. While manipulations that predict a strengthened effect (e.g., because of recency or a lexical neighborhood filled with many voiced neighbors for devoiced stops) have the potential to replicate the effect, all manipulations that bias against incomplete neutralization will have problems finding anything at all. This points to the limits of investigating incomplete neutralization, and for using incomplete neutralization as a test bed for looking at the cognitive architecture of the lexicon: Without finding viable ways of strengthening the effect, research on incomplete neutralization will always have to cope with high Type II error rates.

From a functionalist perspective, an account that treats incomplete neutralization as an artifact of paradigmatic representations is more attractive than an account based on phonetic or phonological rules and/or representations that are extracted from auditory information. Under both the co-activation account and van Oostendorp's turbidity account relation account, speakers would not need to extract any subtle contrast from the signal, at least as long as they do extract the contrast between the corresponding paradigmatic neighbors. The acquisition of the contrast within the paradigm would automatically cause interference resulting in incomplete neutralization. This is in line with the low accuracy scores in perception

experiments. The acoustic cues in the neutralized position have no functional utility and are not reliably used in regular communication to differentiate between minimal pairs.

To conclude, our results are crucially independent of whatever mechanism actually explains incomplete neutralization. The first and foremost goal of this paper was to assess whether or not neutralization is indeed incomplete. We demonstrated the robustness of the effect in three production experiments ruling out a number of claims that the incompleteness is a purely methodological artifact. We have thus shown what incomplete neutralization warrants explanation. As there is no point in incorporating a nonexistent phenomenon into phonological theory, phonologists were justified in being skeptical of the previous evidence for incomplete neutralization. However, although incomplete neutralization does not necessarily have to be explained in terms of formal phonology there are parsimonious accounts in existing phonological frameworks such as Optimality Theory (van Oostendorp, 2008). Moreover, "non-phonological" accounts that make recourse to existing experimental work on co-activation in the mental lexicon seem to be fruitful avenues for further investigations (cf. discussion in Winter & Röttger, 2011). Manaster-Ramer (1996: 487) uses the incomplete neutralization debate as a call for an increased collaboration between phonologists and phoneticians. In Manaster-Ramer's words (ibid. 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 to argue that we can gain new perspectives on old problems by engaging with work from other disciplines.

Acknowledgements

[ACKNOWLEDGMENTS ABOUT HERE]

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.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

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. E. (2008). Phonetic and social selectivity in speech accomodation. Ph.D. dissertation, University of California, Berkeley.

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.

Benua, L. (1995). *Transderivational identity: Phonological relations between words*. Ph.D. dissertation, University of Massachusetts, Amherst.

Benus, S., & Gafos, A. I. (2007). Articulatory characteristics of Hungarian 'transparent' vowels. *Journal of Phonetics*, 35, 271-300.

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

Blevins, J. (2006). New perspectives on English sound patterns: 'Natural' and 'Unnatural' in Evolutionary Phonology. *Journal of English Linguistics*, 34, 6–25.

Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/.

Braver, A., & Kawahara, S. (manuscript). Complete and Incomplete Neutralization in Japanese Monomoraic Lengthening.

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

Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *Journal of the Acoustical Society of America*, 117, 3890–3901.

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

Burzio, L. (1994). *Principles of English stress*. Cambridge: Cambridge University Press.

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.

Chongsuvivatwong, V. (2011). *epicalc: Epidemiological calculator*. R package version 2.12.2.0.

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.

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

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*. Oxford, UK: Wiley-Blackwell, pp. 2711-2729.

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

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

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

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.

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, 2, 8–11.

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.

Gafos, A. I., & Benus, S. (2006). Dynamics of phonological cognition. *Cognitive Science*, 30, 905-943.

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 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.

Goldrick, M. (2001). Turbid output representations and the unity of opacity. *Papers from the Annual Meeting of the North East Linguistic Society*, 30(1), 231–245.

Grawunder, S., & Lancia, L. (2012). What does a Frad look like? – Targeting the effect of incomplete neutralization. Oral presentation at Phonetik & Phonologie 8, Jena.

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

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

Hay, J., Pierrehumbert, J., & Beckman, M. (2003). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Odgen, & R. Temple (Eds.), *Papers in laboratory phonology VI* (pp. 58–74). Cambridge: Cambridge University Press.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2), 187–211.

Jaeger, F. T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434–446.

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

Jespersen, O. (1913). *Lehrbuch der Phonetik* (second edition). Leipzig: G. B. Teubner.

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

Kim, H., & Jongman, A. (1996). Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean. *Journal of Phonetics*, 24, 295–312.

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.

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

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. Manuscript, retrieved February 14th, 2013, from http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2012_3/

Kopkallı, H. (2007). *A phonetic and phonological analysis of final devoicing in Turkish*. Ph.D. dissertation, University of Michigan.

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.

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.

Nielsen, K. (2005). Specificity and Generalizability of Spontaneous Phonetic Imitation. *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, USA.

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, 309–28.

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.

Pitt, M., & McQueen, J. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39, 347–370.

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. (1984). Neutralization of syllable final voicing in German. *Research in Phonetics*, 4, 93–134. Bloomington, Indiana: Indiana University.

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. (1984). Neutralization of obstruent voicing in German is incomplete. In A. Koutsoudas (Ed.), *The Application of and Ordering of Grammatical Rules* (pp. 55–73). The Hague: Mouton.

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

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

Repp, B. H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial consonants. *Language and Speech*, 22, 173–189.

Richards, S. A., Whittingham, M. J., & Stephens, P. A. (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and Sociobiology*, 65, 77–89.

Röttger, T., Winter, B., & Grawunder, S. (2011). The robustness of incomplete neutralization in German. In *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong: City University of Hong Kong (pp. 1342-1345).

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

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*. Somerville, MA: Cascadilla Press, pp. 72-86.

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.

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually Accommodating: Speech rate accommodation to a virtual interlocutor. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 127–132). Austin, TX: Cognitive Science Society.

Spiess, A.-N. (2012). qpcR: Modelling and analysis of real-time PCR data. R package version 1.3-7.

Trask, R. L. (2007). *Historical Linguistics*. (2nd edition, revised by McColl Millar, R). London: Arnold.

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

van Oostendorp, M. (2008). Incomplete devoicing in formal phonology. *Lingua*, 118, 1362–1374.

Vennemann, T. (1968). German Phonology. Ph.D. dissertation, University of California, Los Angeles.

Warner, N., Good, E., Jongman, A., & Sereno, J. (2006). Orthographic versus morphological incomplete neutralization effects. *Journal of Phonetics*, 34, 285–293.

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.

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.

Wilson, S. M. (2003). A phonetic study of voiced, voiceless and alternating stops in Turkish. *CRL Newsletter*, Volume 15 No.1, April 2003.

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

Winter, B. (2011). Pseudoreplication in phonetic research. *Proceedings of the International Congress of Phonetic Science* (pp. 2137-2140). Hong Kong, August 2011.

Winter, B., & Röttger, T. (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.

Yu, A. C. L. (2007). Understanding near mergers: the case of mophological tone in Cantonese. *Phonology*, 24, 187–214.

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., Engel, U., Frosch, H., Hoberg, U., & Vorderwülbecke, K. (1997). *Grammatik der deutschen Sprache* (Band 1). Berlin: de Gruyter.

Appendix A: critical stimuli of E1				
[+voice]		**[-voice]**		**PLACE**
Blode	[blo:də]	Blote	[blo:t^{h}ə]	alveolar
Drude	[dʁu:də]	Drute	[dʁu:t^{h}ə]	alveolar
Flabe	[fla:bə]	Flape	[fla:p^{h}ə]	bilabial
Frade	[fʁa:də]	Frate	[fʁa:t^{h}ə]	alveolar
Froge	[fʁo:gə]	Froke	[fʁo:k^{h}ə]	velar
Frube	[fʁu:bə]	Frupe	[fʁu:p^{h}ə]	bilabial
Gage	[ga:gə]	Gake	[ga:k^{h}ə]	velar
Gaude	[gaʊ̯də]	Gaute	[gaʊ̯t^{h}ə]	alveolar
Gobe	[go:bə]	Gope	[go:p^{h}ə]	bilabial
Griede	[gʁi:də]	Griete	[gʁi:t^{h}ə]	alveolar
Jiede	[ji:də]	Jiete	[ji:t^{h}ə]	alveolar
Klabe	[kla:bə]	Klape	[kla:p^{h}ə]	bilabial
Mube	[mu:bə]	Mupe	[mu:p^{h}ə]	bilabial
Nauge	[naʊ̯gə]	Nauke	[naʊ̯k^{h}ə]	velar
Priege	[pʁi:gə]	Prieke	[pʁi:k^{h}ə]	velar
Pruge	[pʁu:gə]	Pruke	[pʁu:k^{h}ə]	velar
Quade	[kva:də]	Quate	[kva:t^{h}ə]	alveolar
Quobe	[kwo:bə]	Quope	[kwo:p^{h}ə]	bilabial
Roge	[ʁo:gə]	Roke	[ʁo:k^{h}ə]	velar
Schmaube	[ʃmaʊ̯bə]	Schmaupe	[ʃmaʊ̯p^{h}ə]	bilabial
Schriege	[ʃʁi:gə]	Schrieke	[ʃʁi:k^{h}ə]	velar
Schuge	[ʃu:gə]	Schuke	[ʃu:k^{h}ə]	velar
Stauge	[ʃtaʊ̯gə]	Stauke	[ʃtaʊ̯k^{h}ə]	velar
Wiebe	[vi:bə]	Wiepe	[vi:p^{h}ə]	bilabial

Appendix B: critical stimuli of E2 and E3 (in bold)				
[+voice]		**[-voice]**		**PLACE**
Bauge	**[baʊ̯gə]**	**Bauke**	**[baʊ̯kʰə]**	**velar**
Bege	[beːgə]	Beke	[beːkʰə]	velar
Blebe	**[bleːbə]**	**Blepe**	**[bleːpʰə]**	**bilabial**
Bloge	**[bloːdə]**	**Bloke**	**[bloːtʰə]**	**velar**
Dage	[daːgə]	Dake	[daːkʰə]	velar
Dabe	[daːbə]	Dape	[daːpʰə]	bilabial
Diege	[diːgə]	Dieke	[diːkʰə]	velar
Dobe	**[doːbə]**	**Dope**	**[doːpʰə]**	**bilabial**
Drube	**[dʁuːbə]**	**Drupe**	**[dʁuːpʰə]**	**bilabial**
Duge	[duːgə]	Duke	[duːkʰə]	velar
Fage	**[faːgə]**	**Fake**	**[faːkʰə]**	**velar**
Faube	[faʊ̯bə]	Faupe	[faʊ̯pʰə]	bilabial
Flabe	[flaːbə]	Flape	[flaːpʰə]	bilabial
Flebe	**[fleːbə]**	**Flepe**	**[fleːpʰə]**	**bilabial**
Frebe	[fʁeːbə]	Frepe	[fʁeːpʰə]	bilabial
Froge	**[fʁoːgə]**	**Froke**	**[fʁoːkʰə]**	**velar**
Frobe	[fʁoːbə]	Frobe	[fʁoːpʰə]	bilabial
Frube	[fʁuːbə]	Frupe	[fʁuːpʰə]	bilabial
Gage	[gaːgə]	Gake	[gaːkʰə]	velar
Gaube	**[gaʊ̯bə]**	**Gaupe**	**[gaʊ̯pʰə]**	**bilabial**
Gauge	[gaʊ̯gə]	Gauke	[gaʊ̯kʰə]	velar
Glege	**[gleːgə]**	**Gleke**	**[gleːkʰə]**	**velar**
Gliebe	**[gliːbə]**	**Gliepe**	**[gliːpʰə]**	**bilabial**
Gobe	[goːbə]	Gope	[goːpʰə]	bilabial

Griebe	[gʁi:bə]	Griepe	[gʁi:p^{h}ə]	bilabial
Hege	[he:gə]	Heke	[he:k^{h}ə]	velar
Klabe	**[kla:bə]**	**Klape**	**[kla:p^{h}ə]**	**bilabial**
Krobe	**[kʁo:bə]**	**Krope**	**[kʁo:p^{h}ə]**	**bilabial**
Miebe	**[mi:bə]**	**Miepe**	**[mi:p^{h}ə]**	**bilabial**
Naube	[naʊ̯bə]	Naupe	[naʊ̯p^{h}ə]	bilabial
Nauge	**[naʊ̯gə]**	**Nauke**	**[naʊ̯k^{h}ə]**	**velar**
Nuge	[nʊ:gə]	Nuke	[nʊ:k^{h}ə]	velar
Priege	[pʁi:gə]	Prieke	[pʁi:k^{h}ə]	velar
Pruge	**[pʁu:gə]**	**Pruke**	**[pʁu:k^{h}ə]**	**velar**
Roge	[ʁo:gə]	Roke	[ʁo:k^{h}ə]	velar
Schlabe	**[ʃla:bə]**	**Schlape**	**[ʃla:p^{h}ə]**	**bilabial**
Schmaube	**[ʃmaʊ̯bə]**	**Schmaupe**	**[ʃmaʊ̯p^{h}ə]**	**bilabial**
Schriege	**[ʃʁi:gə]**	**Schrieke**	**[ʃʁi:k^{h}ə]**	**velar**
Schuge	**[ʃu:gə]**	**Schuke**	**[ʃu:k^{h}ə]**	**velar**
Spage	**[ʃpa:gə]**	**Spake**	**[ʃpa:k^{h}ə]**	**velar**
Stauge	[ʃtaʊ̯gə]	Stauke	[ʃtaʊ̯k^{h}ə]	velar
Strege	**[ʃtʁe:gə]**	**Streke**	**[ʃtʁe:k^{h}ə]**	**velar**
Sube	**[zʊ:bə]**	**Supe**	**[zʊ:p^{h}ə]**	**bilabial**
Triege	**[tʁi:gə]**	**Trieke**	**[tʁi:k^{h}ə]**	**velar**
Wiebe	[vi:bə]	Wiepe	[vi:p^{h}ə]	bilabial
Wube	[vu:bə]	Wupe	[vu:p^{h}ə]	bilabial
Wuge	[vu:gə]	Wuke	[vu:k^{h}ə]	velar
Zebe	[tse:bə]	Zepe	[tse:p^{h}ə]	bilabial