Journal of Phonetics 70 (2018) 39–55

Contents lists available at ScienceDirect

Journal of Phonetics

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

Special Issue: Emerging Data Analysis in Phonetic Sciences, eds. Roettger, Winter & Baayen

Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German ☆

Bruno Nicenboim [a,*], Timo B. Roettger [b,c], Shravan Vasishth [a]

[a] University of Potsdam, Germany
[b] Northwestern University, Evanston, United States
[c] University of Cologne, Germany

ARTICLE INFO

Article history:
Received 5 October 2017
Received in revised form 22 May 2018
Accepted 16 June 2018

Keywords:
Meta-analysis
Incomplete neutralization
Final devoicing
German
Bayesian data analysis

ABSTRACT

Within quantitative phonetics, it is common practice to draw conclusions based on statistical significance alone. Using *incomplete neutralization* of final devoicing in German as a case study, we illustrate the problems with this approach. If researchers find a significant acoustic difference between voiceless and devoiced obstruents, they conclude that neutralization is incomplete; and if they find no significant difference, they conclude that neutralization is complete. However, such strong claims regarding the existence or absence of an effect based on significant results alone can be misleading. Instead, the totality of available evidence should be brought to bear on the question. Towards this end, we synthesize the evidence from 14 studies on incomplete neutralization in German using a Bayesian random-effects meta-analysis. Our meta-analysis provides evidence in favor of incomplete neutralization. We conclude with some suggestions for improving the quality of future research on phonetic phenomena: ensure that sample sizes allow for high-precision estimates of the effect; avoid the temptation to deploy researcher degrees of freedom when analyzing data; focus on estimates of the parameter of interest and the uncertainty about that parameter; attempt to replicate effects found; and, whenever possible, make both the data and analysis available publicly.

1. Introduction

Theories of speech communication and its cognitive underpinnings are increasingly shaped by experimental data and quantitative analyses. Ideally, our theories progressively grow and change with *accumulating* empirical evidence. The evidence provided by a single study, however, is limited to the applied method and the particular sample. Its results are prone to random statistical fluctuations and its interpretation is dependent on methodological and analytical choices. To assess the evidence that a single study can provide, we need a good understanding of statistical theory and inference. There are several specific aspects of statistical analysis, which, despite having received little attention in our field, researchers need to be aware of when carrying out statistical inference.

Beyond statistical assessments of a single study, we can assess the robustness of a phenomenon by synthesizing evidence across many studies. One technique that allows us to synthesize evidence is meta-analysis, which is a quantitative summary of the results of multiple studies. Here, we apply this technique to a representative phenomenon from the speech production literature which has already fueled fruitful discussions surrounding methodological and analytical practices in phonetics in the past: incomplete neutralization of final devoicing.

1.1. Final devoicing and incomplete neutralization

Final devoicing is a common phonological alternation in the world's languages. For example, languages such as Catalan, Dutch, Polish, Russian, Turkish, and German contrast voiced obstruents intervocalically but neutralize the contrast syllable or word finally in favor of voiceless obstruents, as in the following German examples (cf. 1–2):

☆ For partial support of this research, we thank the Volkswagen Foundation through grant 89 953 to Shravan Vasishth and the Deutsche Forschungsgemeinschaft for two grants: VA 482/8-1 to Shravan Vasishth and Frank Roesler, which funded Bruno Nicenboim, and the Q project, PIs Shravan Vasishth and Ralf Engbert, in the Sonderforschungsbereich 1287, Limits of Variability in Language. We thank Susanne Fuchs and Sven Grawunder for sharing their data with us and allowing us to make them available. We also thank Paul Bürkner for the support on the *R* package *brms*. The complete code, data, and supplementary material for this paper are available at https://osf.io/g5ndw/.

* Corresponding author.
E-mail address: bruno.nicenboim@uni-potsdam.de (B. Nicenboim).

https://doi.org/10.1016/j.wocn.2018.06.001
0095-4470/

(1) Rad [ʁaːt] ‘wheel’; Räder [ʁɛːdɐ] ‘wheels’
(2) Rat [ʁaːt] ‘council’; Räte [ʁɛːtə] ‘councils’

In intervocalic position, the voicing contrast of oral stops can be manifested by different acoustic dimensions, such as the preceding vowel duration, glottal pulsing during the closure, closure duration, and voice onset time (e.g., Lisker, 1986), with voiced stops exhibiting longer preceding vowels, more glottal pulsing during the closure, a shorter closure duration, and shorter (or negative) voice onset time. The term neutralization implies that the acoustic form of the alveolar stop in *Rad* [ʁaːt] *‘wheel’* is identical to the alveolar stop in *Rat* [ʁaːt] *‘council’*, resonating with ear-phonetic assessments of traditional linguistic descriptions (Jespersen, 1920; Trubetzkoy, 1939; Wiese, 1996).

However, numerous experimental studies have argued that there are small acoustic and/or articulatory differences between words such as *Rad* and *Rat*, suggesting that in German, this neutralization is in fact incomplete (Charles-Luce, 1985; Dinnsen & Garcia-Zamor, 1971; Fuchs, 2005; Greisbach, 2001; Grawunder, 2014 Mitleb, 1981; Port & O’Dell, 1985; Port & Crawford, 1989; Roettger, Winter, Grawunder, Kirby, & Grice, 2014; Smith, Hayes-Harb, Bruss, & Harker, 2009; Taylor, 1975). Importantly, the direction of the difference resembles the non-neutralized contrast; for example, vowels preceding voiceless stops tend to be shorter than vowels preceding devoiced stops. The magnitude of the difference, however, is much smaller. For example, Port and Crawford (1989) report a vowel duration difference of approximately 1–6 ms between devoiced and voiceless stops in German, while Warner, Jongman, Sereno, and Kemps (2004) report a difference of 3.5 ms in Dutch (in comparison to substantially larger vowel duration differences found in non-neutralized contexts in German ranging from 24–41 ms; see Mitleb, 1981; Fuchs, 2005; Roettger et al., 2014). Beyond subtle differences in production, these acoustic differences can be perceptually recovered by listeners with above-chance accuracy (e.g., Kleber, John, & Harrington, 2010; Port & O’Dell, 1985; Port & Crawford, 1989; Roettger et al., 2014).

Many scholars have acknowledged the evidence for incomplete neutralization and proposed several ways to implement this phenomenon in formal models of phonological representations (e.g., Charles-Luce, 1985; Dinnsen & Charles-Luce, 1984; Van Oostendorp, 2008; Port & O’Dell, 1985). These formal accounts challenged several assumptions of contemporary phonological models, leading Port and Crawford (1989, pp. 10–15) to claim that incomplete neutralization poses “a threat to phonological theory” (see also Port & Leary, 2005). More recent accounts to incomplete neutralization are rooted in psycholinguistic models of lexical organization, suggesting that incomplete neutralization is an artifact of lexical co-activation (Ernestus & Baayen, 2006; Kleber et al., 2010; Roettger et al., 2014; Winter & Roettger, 2011).

Others scholars have remained skeptical regarding incomplete neutralization, crucially fueled by a few studies that did not find evidence for it (Fourakis & Iverson, 1984; Inozuka, 1991; Jessen & Ringen, 2002; Piroth & Janker, 2004). Studies on incomplete neutralization have also attracted serious criticism on methodological grounds (Kohler, 2012; Manaster-Ramer, 1996; Roettger et al., 2014), leading some researchers to disregard it as a methodological artifact (e.g., Kohler, 2007, 2012). For example, it has been argued that incomplete neutralization is an orthographically induced contrast, where speakers are thought to perform an “artificial” hypercorrection based on the written language (e.g., Fourakis & Iverson, 1984; Manaster-Ramer, 1996). This concern has been tackled by more recent studies, showing that incomplete neutralization is also obtained when participants do not encounter orthographic input (e.g., Roettger et al., 2014).

It has also been argued that early studies on incomplete neutralization have recorded German-speaking populations with high proficiency in English, which is a potential problem because English preserves the final voicing contrast (e.g., *bad* vs. *bat*, *bed* vs. *bet*) (Kohler, 2007; Winter & Roettger, 2011). However, many later studies used German speakers living in Germany and report similar effect sizes (Grawunder, 2014; Roettger et al., 2014).

It is safe to say that incomplete neutralization is a polarizing phonetic phenomenon. One camp of scholars interpret the available evidence in favor of incomplete neutralization, with important implications for models of speech production and linguistic representations, while others interpret the available evidence as either insufficient or pointing towards incomplete neutralization being a methodological artifact. The latter position has led to productive methodological debates, not only raising awareness for important aspects of experimental design, but also drawing attention to important conceptual issues regarding statistical inference beyond the observed data.

Incomplete neutralization is a prime example to discuss statistical misinterpretations due to several reasons. First, incomplete neutralization effects have been reported to be rather small, making an accurate estimate of the effect particularly important for scientific conclusions. Second, incomplete neutralization studies commonly use multiple acoustic and/or articulatory measures to test one (alternative) hypothesis, namely, devoiced stops are different from voiceless stops. However, the results from statistical tests are generally not corrected for multiple comparisons (using, for example, the Bonferroni correction). And third, the incomplete neutralization literature has a history of publishing null results, which led to several (conceptual) replication attempts.

All in all, the literature on incomplete neutralization is a representative area of phonetic research which has already been a source of methodological debates. We aim at continuing this tradition and use incomplete neutralization to discuss important aspects of statistical analyses and misconceptions that need to be taken into account when drawing inferences that go beyond the observed data. It is important to emphasize that incomplete neutralization only serves as a representative example for common practices in phonetic research. Both the misconceptions we discuss and the strategies to avoid potential analytical pitfalls generalize towards other areas of phonetics as well as the sciences in general. We further use the available evidence in the literature to assess the robustness of the phenomenon via a meta-analysis, a powerful statistical procedure for combining data from multiple studies that is standard in other fields. Our meta-analysis suggests that (i) incomplete neutralization is robust across the available data in the literature, (ii) there is insufficient evidence supporting the claim that previously mentioned potential confounds