Evidence in incomplete neutralisation: tonal data and Bayesian inference

I take incomplete neutralisation (IN) to be phonological alternations that result in seemingly indistinguishable outputs which nevertheless leave instrumentally detectable differences. German final devoicing is an example. Traditionally, voiced coda obstruents (e.g. in *rad*) are thought to be devoiced and neutralised with voiceless coda obstruents (e.g. in *rat*) in a “classic example of a phonological rule” (Wiese, 1996). However, numerous experimental studies have challenged this view, showing fine-grained acoustic and even perceptual difference between allegedly neutralised pairs like *rad* and *rat* (Port & Crawford, 1989; Port & O’Dell, 1985; Roettger, Winter, Grawunder, Kirby, & Grice, 2014).

Despite the significant amount of empirical literature (e.g. final devoicing in Catalan, Dinnsen & Charles-Luce, 1984; Dutch, Ernestus & Baayen, 2006; Russian, Kharlamov, 2014; American English intervocalic flapping, Braver, 2014; Japanese monomoraic lengthening, Braver & Kawahara, 2014), the field has yet to reach anything resembling a consensus as to whether IN effects are genuine (Roettger et al., 2014). This is partly due to persistent scepticism of IN studies on methodological grounds. These sceptics suspect a case of “bad data”, and they have good reasons to do so. IN presents a challenging task for experimentalists: we have to test for an effect which will be small (largely imperceptible) by definition, in the face of many potential confounds.

These confounds cited by critics primarily include laboratory speech effects and orthographic interference (Fourakis & Iverson, 1984; Manaster Ramer, 1996; Warner, Good, Jongman, & Sereno, 2006). Fourakis and Iverson find that IN in German final obstruents occur only if subjects read word lists that shows minimal pairs. When the experiment looks like morphological paradigm elicitation (e.g. conjugate strong verbs given the infinitives), IN effect becomes statistically insignificant. In another line of objection, Warner et al. (2006) attribute IN in Dutch final devoicing to “spelling pronunciation”, as only minimal pairs that differ in spelling (e.g. <d> vs <t>) elicit acoustic differences.

Another “bad data problem” in the IN literature, surprisingly under-acknowledged, concerns the use of frequentist statistical inference (usually resulting in a p value) for data evaluation. Frequentist tools are fundamentally unable to reject the alternative hypotheses or positively establish equivalence (Gallistel, 2009). For these models, a small p may indicate statistical significance, but a large p is uninformative – the absence of evidence does not constitute evidence of absence. The dominance of frequentist statistics in linguistics and in the IN literature creates a persistent bias towards rejecting the null hypothesis in favour of IN. Studies which do not find significant difference can simply be explained as having insufficient statistical power (e.g. not enough items and subjects; Roettger et al., 2014).

I argue that the study of tone sandhi and the use of Bayesian statistical inference provide alternatives that will expand the typology of IN (good data) and put the overall literature on firmer empirical ground (good data evaluation). I will present results from an ongoing study on Fuzhou, a Min Chinese language spoken in Southeast China with a rich tone sandhi system.

Tone sandhi refers to tonal alternations in connected speech. An example of sandhi IN is Mandarin T3 sandhi, turning Tone 3 /L/ into Tone 2 /LH/ before Tone 3. Lexical Tone 2 and sandhi T3 (from Tone 3) are perceptually indistinguishable (Wang & Li, 1967), but phonetic studies show they differ acoustically (Kratochvil, 1987; Xu, 1993; Zee, 1980; Peng, 2000).

The advantages gained from studying Fuzhou tone sandhi are manyfold: Fuzhou is a
largely unwritten language, and the stimuli (Chinese script) for eliciting speech do not mark tones, reducing potential orthographic influence to a minimum; my personal observation also suggests very low conscious awareness of tones for Fuzhou speakers. The sheer amount of neutralisation in Fuzhou allows for the testing of IN effects in sandhi tones of varying shapes, generated under different tones and from different underlying tones. For example, Tone AEFG (44, 212, 232, 23) apparently all turn into Tone A (high level 44) when followed by Tone ABC. Because of this property, Fuzhou tone sandhi contains neutralisations which I dub the “A/B > C” type (e.g. T232/T44 > T53), in contrast to the commonly found “A/B > A” type (e.g. voiced/voiceless > voiceless).

The use of Bayesian inference has at least two advantages (Gallistel, 2009; Roudet et al., Submitted): It is not prejudiced against the null hypothesis, as it evaluates the likelihood of each possible hypothesis. Further, it provides evidence on a graded scale rather than with a fixed cut-off point (e.g. alpha = 0.05). I argue that it is now both necessary and feasible to introduce Bayesian inference into IN studies. Necessary because of its ability to both confirm and reject IN and as its graded nature allows cross-study comparison and the accumulation of evidence. Feasible because recent advances are making Bayesian inference increasingly user-friendly (Masson, 2011; Rouder et al., Submitted; Sorensen & Vasishth, 2015).

In this talk, I will present production data gathered from 15 Fuzhou speakers on two neutralising pairs: T53/T44 > T44 before T53 and T232/T44 > T53 before T32, with methodology similar to Li (2015). I first compare the F0 realisations of sandhi T53/T44 and T232/T44 using Smoothing-Spline ANOVA (Gu, 2013). The 95% confidence intervals of T53 and T44 are largely overlapping, while those of T232 have slightly lower F0 onset than T44. I then calculate the Bayes Factors for possible hypotheses in a hierarchical model (Rouder et al., Submitted), which show the odds of the alternative hypothesis being true relative to the null hypothesis. In Figure 1, a model with underlying tone as a fixed factor and speaker as random factor (the fourth bar) outperforms the null hypothesis by a Bayes Factor of 13.26, providing “positive” evidence that sandhi T232 and T44 do differ in F0. In Figure 2, the same “tone.pair + speaker” model (the first bar) has a Bayes Factor of 1.25, a sign that the null hypothesis is very likely true. Using Bayesian inference, I reach the tentative conclusion that Fuzhou neutralisation of T232/T44 > T53 is incomplete, while T53/T44 > T44 is complete.

Figure 1: Bayes factors of alternative hypotheses against the null for neutralisation in T53

Figure 2: Bayes factors of alternative hypotheses against the null for neutralisation in T44
Reference