Advice on comparing languages and varieties

John Nerbonne[1,2]

[1] *University of Groningen,* [2]*University of Freiburg*

j.nerbonne@rug.nl

It is exciting that a DGfS workshop is to be held on "Empirical Studies of Word Sense Divergences across Language Varieties", and I look forward to hearing the papers. In the present contribution I relate my own experience as a computational linguist who has worked a great deal for over twenty years in dialectology, and I'll even presume to offer some, I hope not wholly unwanted advice to others interested in similar cross-disciplinary work.

Let's begin by noting that there is increasing interest in using computational methods to compare languages and varieties from several perspectives. COMPARATIVE LINGUISTICS used to designate (comparative-)historical linguistics (Wikipedia). In addition to work in dialectology and sociolinguistics, and indeed, some modern comparative-historical linguistics is computationally sophisticated (Dellert 2019). There's a good deal of work on multilingualism and contact (Gooskens & van Heuven 2017). I'll review these and others in my talk, I hope underscoring the potential interest in studying word-sense divergences.

Lots of data on linguistic variation are themselves variable. Not every speaker of Cockney glottalizes non-initial /t/'s, and not every New Englander negates elided VPs while maintaining a positive meaning (*She does, and so doesn't he*). This makes it essential to collect data in a way that yields representative samples and to observe a range of cases. Arm-chair work, however phenomenologically astute, is limited even if it may be useful initially.

It is very important in applying computational techniques to linguistic problems that the reliability of the computational measure be considered. In the case of WORD SENSES, the late lexicographer, Adam Kilgariff, often pointed out the conceptual problems adhering to them (1997). As he noted, these difficulties infect the word-sense disambiguation (WSD) problem, making it difficult to evaluate. This naturally has impact on detecting WSD historically.

Not only the reliability but also the validity of the measure often needs to be established. In dialectological work, my colleagues and I applied a modified edit-distance measure to phonetic transcriptions, and the work has come to be accepted (Nerbonne 2009), but it was important that the measure was validated in comparison to dialect speakers' judgments of similarity and in comparison to judgments of "how non-native" foreign accents sound.

Finally, as further encouragement for comparative work (in the broader sense) I'll note areas where the dialectological work has inspired forays into other linguistic sub-disciplines. A favorite of mine is Greg Kondrak and Bonnie Dorr's work on detecting potential confusing drug names using edit distance, which was used by the US Food and Drug Administration (Kondrak & Dorr 2006).

References: Wikipedia "Comparative Linguistics"; J.Dellert (2019) *Information-theoretic causal inference of lexical flow.* Lang.Sci.Press; C.Gooskens & V.van Heuven (2017) "Measuring X-linguistic intelligibility in Germanic, Romance and Slavic" *Speech Communication 25-36*. A.Kilgarriff (1997) "I don't believe in word senses" *Computers and the Humanities 91-113*; J. Nerbonne (2009) "Data-driven dialectology" *Language and Linguistics Compass, 175-198*. G.Kondrak & B. Dorr (2006) "Automatic identification of confusable drug names" *Artificial Intelligence in Medicine.*