TABU Bulletin voor Taalwetenschap Groningen

Festschrift for Jack Hoeksema

WHEN LESS IS MORE

John Nerbonne

Abstract
The grammatical ambition to distinguish well-formed from ill-formed structures very often leads to more complicated analyses, which in turn can impede the use of analyses in further studies. We argue thus that less ambitious and less complicated analyses can often provide more scientific insight. Two concrete cases are presented where less discriminating analyses enabled interesting investigations. In one case it was fortunate that the researchers could appeal an external, quantitative standard of quality to justify the simplification. At the risk of platitude, we conclude by confirming that the pursuit of simplicity shouldn't be exaggerated.

Keywords: scientific simplicity, Ockham's razor, theoretical categories, broad vs. narrow phonetic transcription

1. Introduction

More reflective readers may be appalled by the title, but it suggests a pleasant apparent paradox in research about language. I chose it here, as the title in a *Festschrift* for Jack Hoeksema[1], because he remarked to me once that he'd seen countless excellent theoretical analyses in linguistics foiled by a very small number of counterexamples, which then stimulated more discriminating analyses, i.e., involving "more", i.e., more features, more rules or constraints, or at least more apologies about the data. At the same time, there are many reasons to resist this apparently ineluctable dynamic toward more on the side of the *explanans*.

University of Groningen, University of Freiburg and University of Tübingen
Corresponding author: John Nerbonne, j.nerbonne.work@gmail.com

ISSN 0165-9200. Published by University of Groningen Press, Copyright © by author
How to cite this article: Nerbonne, J. (2024). When less is more. *TABU Festschrift for Jack Hoeksema*. 180-187. https://doi.org/10.21827/tabu.2023.41267

Of course, the philosophically minded will interject, we've always been committed not only to accounting for data, but also to accounting for it as simply as possible. Ockham's razor seems apodictic, *Entia non multiplicanda sunt sine necessitate*.[2] Modern formulations of this sort of appeal to simplicity leave room for noting exceptions, even if disfavoring them (Rissanen 1978). Finally, as practicing linguists know, appeals to Ockham or to simplicity are often unsuccessful, witness the many contemporary schools of morphology (Carstairs-McCarthy 2002) or the debates about the explanatory value of deep structure (Huck & Goldsmith 1995).

This article does not pretend to be a research report, but a rather somewhat indulgent essay on the virtue of crying "enough", indeed in reversing the trend, even if only temporarily, in order to work with simpler explanatory apparatuses (or is it *appartūs*, Jack?).

2. Fewer phonetic categories

From 1997 on, my group in computational linguistics developed a novel line of research in dialectology, focusing on the application of edit distance measures to phonetic transcriptions (Nerbonne 2017). The techniques have now been applied to thirty or more languages and have been demonstrated to be reliable and valid (Heeringa et al. 2006; Wieling et al. 2014). But this doesn't mean that we were never surprised. The compilers of the Goeman-Taeldeman-van Reenen project (GTRP) once asked whether we shouldn't wish to analyze their very rich data, which was rich indeed, with 1.876 items transcribed from 613 collections sites in the Netherlands and Flanders. There isn't space to review the data analysis here, which is reported fully in Wieling et al. (2007), but the initial results suggested a novelty, namely that the Dutch spoken in the two countries were very different (Figure 1)! This contradicted earlier results (Heeringa 2004 and references in his literature review).

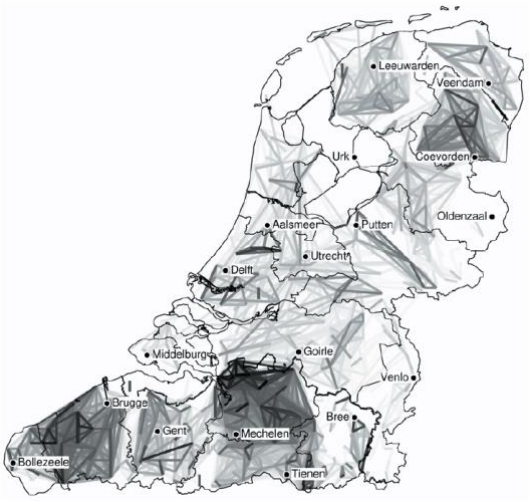

Figure 1: Initial results of the GTRP analysis. Phonetic similarity between sites is shown by darkness of lines. Belgian Dutch appears to have more cohesive areas than the Dutch in the north. Compare Figure 2.

Figure 2: This map displays the cohesiveness of the Belgian and Netherlandic dialect areas separately. Some Dutch dialect areas are more cohesive than in Figure 1 and may form dialect areas.

Further investigation revealed that dialects in the two countries, when analyzed separately, were distributed in ways similar to those revealed in earlier research (Figure 2), corroborating the suspicion that something was wrong in the data, rather than the analysis. It turned out that, although the data collection efforts were coordinated, field workers in the Netherlands had consistently used a set of 86 IPA symbols, while the Belgians had restricted themselves to only 56. We then suspected that the Belgian-Netherlands dialect border (Figure 1) might turn out to be a field work isogloss!

Wieling and Nerbonne (2011) then undertook a project reducing the numbers of phonetic segments by identifying the segment pair that contrasted least in order to eliminate the contrast, a procedure that was applied iteratively in order to finally reduce the phonetic inventory in the data collections on both sides of the border. The procedure obviated appeals to researcher intuition by using edit-distance alignments (Nerbonne & Heeringa 2010) to isolate segment correspondences together with the POINTWISE-MUTUAL INFORMATION metric introduced by Church and Gale (1990) to identify least discriminating contrasts. The resulting site × site distance table correlated with the originals almost perfectly ($r = 0.97$), enabling a comprehensive analysis, shown in Fig. 3. We note as well that the opportunity to gauge the

effect of the simplifying step quantitatively (the correlation analysis) definitely strengthened the confidence in the reduction.

Figure 3: A comprehensive illustration of the dialect differences in the GTRP, once the phonetic system was simplified and multidimensional scaling was applied. See Wieling & Nerbonne (2011) for details.

Before saying more about this particular, less discriminating data categorization, let us anticipate the obvious complaint that information has been discarded. Information has most certainly been discarded! For example, the distinctions among all the palatal and alveopalatal fricatives, i.e., [s, z, ʃ, ʂ, ʒ, ɕ] are not represented in the analysis that led to Fig. 3 (Wieling & Nerbonne 2011: 156). In this case the loss of information enabled a comprehensive analysis that would otherwise have been impossible, but it is easy to imagine research questions for which this loss of information would be terminal. The point is that there are many research questions for which the less discriminating analysis *is* sufficient, e.g., questions about the distribution of aggregate linguistic variation, and questions about the relation of dialectal distribution to aggregate familial relatedness (Manni et al. 2008); questions about the role culture plays in mobility (Falck et al. 2012); or questions about the degree to which semantic relatedness shapes dialect distribution (Huisman et al. 2021).

If the degree of necessary discrimination depends on the research question, then Hoeksema's worry about the dynamic of grammatical thinking, which seems incessantly to require ever finer distinctions, can be appeased, at least a bit. We should note the data distinctions but also be willing to ignore them for some purposes.

Returning to the case at hand, the degree of phonetic discrimination worthwhile in analyzing aggregate patterns in variation, we should note that Wieling & Nerbonne (2011) also analyzed the effect of reducing the number of phonetic categories not only for the GTRP (Dutch) data, but also for datasets in German, Bulgarian, and Norwegian, always reducing them to 42 segments (as in fact was the Dutch set, a detail suppressed above), and resulting in simplified sets where the aggregate correlation to the original was near perfect ($0.96 \leq r \leq 0.995$). They likewise noted that the characteristic Seguy curves plotting linguisticdifferences as a function of geographic distance were systematically lower when the simpler phonetics were used, but that they retained their logarithmic form. This suggests that one may ignore a great deal of phonetic detail in many dialectological investigations.

3. Another case: Syntactic categories

Since the earlier 1990s natural language processing has pursued a strategy from engineering in which large and complex tasks such as understanding natural language are broken down into smaller and less complex ones. One such task is that of identifying the lexico-syntactic categories of the words in sentences, better known as PART-OF-SPEECH TAGGING (or POS tagging). This has proven very useful in any number of applications in natural language processing, and it stood to reason that the technique might be useful in theoretical work, too.

Wiersma et al. (2011) investigated the English of Finnish immigrants to Australia. Their plan was to tag the conversational transcripts of the immigrants and compare their syntax to that of (near-)natives, which involved collection sequences of POS-tags and comparing their frequencies. They chose to tag the conversations using the tag set of the International Corpus of English (ICE, https://www.ucl.ac.uk/english-usage/projects/ice.htm), which had been designed by linguists, and which included 270 different syntactic categories (Garside et al., 1997). Table 1 illustrates the result of tagging (tagged using Brant's (2000) tagger).

Table 1: Example sentence from the Finnish Australian English corpus tagged using Brants' TnT tagger using the ICE tagset (see text for further explanation).

We	'll	have	a	roast	leg	of	lamb	tomorrow
PRON 1.pl	aux modal pres encl	V trans inf	ART indef	CN sg	CN sg	PREP	CN sg	ADV

The n-grams of POS-tags were then analyzed to confirm the unsurprising hypothesis that the immigrants' syntax was different, but also to aid in detecting deviations (see Sec. 4.2 in Wiersma et al. 2011).

We discuss this example here because the work might have been simpler if a smaller tag set had been used, i.e., a linguistically less discriminating one. In fact, 75 of the TOSCA-ICE tags were not instantiated in the data at all. If we had used the 20-element reduced ICE tag set, the training needed for the tagger would have been shortened, the number of trigrams would have been reduced massively, and the need to discard infrequently encountered POS trigrams would have been mitigated.

4. Discussion and conclusions

We discussed only two cases above where it was argued that less is more, or less flippantly formulated, that reduced theoretical discrimination can support better analyses, but there are more general considerations that should be mentioned. It is sensible to note that we have not presumed to suggest how simple analyses should be, acknowledging that the general question is complex.

The wish to work with many categories is well motivated. Studies that fail to include influential factors are regarded as confounded, and, even if there is no way to guard against this in general, a natural reaction is to attempt to include as much as possible. This explains the

frequent question after talks in corpus linguistics, as to whether the researcher controlled for various demographic properties of the texts' authors and intended audiences, e.g., age, gender and sexual orientation, income or class, profession, genre, … There can always be yet another factor to be considered.

Analysis is often complicated when additional influences are considered, so much so that some speak of "the curse of dimensionality", a phrase which Wikipedia attributes to a technical report by Richard Bellman.[3] Including an additional potential influence (*explanans*) in a study with *n* potential factors does not increase the number of examinations of the data needed merely from *n* to *n*+1, because each subset of the potentially influential variables needs to be considered, resulting in an increase of $2^n$to 2^{n+1}, an exponential growth. This means that the analysis of data involving too many potential *explantia* is not just strenuous, but often infeasible.

As the introduction stated, we should resist the apparently ineluctable dynamic toward more on the side of the *explanans*. The virtue in keeping explanations simple lies in keeping analyses comparable (the case of the phonetic inventory), in avoiding gaps in analysis (the case of the overambitious tag set), and in avoiding infeasible analytical tasks (the curse of dimensionality). Simplicity has its rewards beyond abject obedience to Brother William of Ockham.

Endnotes

[1] Jack Hoeksema came to the Linguistics department at Ohio State University about 40 years ago, where he filled in for David Dowty, who was on sabbatical. I was a late grad student there, but we shared an interest in Categorial Grammar and for the "bigger questions" in linguistics, which smoothed the way for a sustained, collegial relationship in Groningen (1993-2017). Jack always seemed to enjoy wider-ranging conversations.

[2] Maybe overeagerly attributed to Ockham: https://plato.stanford.edu/entries/ockham/#OckhRazo

[3] en.wikipedia.org/wiki/Curse_of_dimensionality (consulted 2023.06.23).

References

Brants, T. (2000). TnT-a statistical part-of-speech tagger. *arXiv preprint cs/0003055*.

Carstairs-McCarthy, A. (2002). *Current Morphology*. London & New York: Routledge.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Falck, O., Heblich, S., Lameli, A., & Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, *72*(2-3), 225-239.

Garside, R., Leech, G. N., & McEnery, A. M. (1997). *Corpus Annotation: Linguistic information from computer text corpora*. London/New York: Longman.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In *Proc. Workshop on Linguistic Distances*. Shroudsburg, PA: Assoc. Computational Linguistics. 51-62.

Huck, G. J., & Goldsmith, J. A. (1995). *Ideology and Linguistic Theory: Noam Chomsky and the deep structure debates*. London: Psychology Press.

Huisman, J. L., Franco, K., & van Hout, R. (2021). Linking linguistic and geographic distance in four semantic domains: computational geo-analyses of internal and external factors in a dialect continuum. *Frontiers in Artificial Intelligence*, *4*, 668035.

Manni, F., Heeringa, W., Toupance, B., & Nerbonne, J. (2008). Do surname differences mirror dialect variation. *Human Biology*, *80*(1), 41-64.

Nerbonne, J. (2017). *Humanities, exactly!/Letteren, exact!* U. Groningen. Faculty of Arts.

Nerbonne, J., & Heeringa, W. (2010). Measuring dialect differences. In: Schmidt, E. & Auer, P. (eds.) *Language and Space.* Berlin: Mouton De Gruyter, 550-566.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465-471.

Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., & Nerbonne, J. (2014). Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change 4*(2), 253-269.

Wieling, M., Heeringa, W., & Nerbonne, J. (2007). An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data. *Taal en Tongval*, *59*(1), 84-116.

Wieling, M., & Nerbonne, J. (2011). Measuring linguistic variation commensurably. *Dialectologia: Revista Electrònica*, 141-162.

Wiersma, W., Nerbonne, J., & Lauttamus, T. (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, *26*(1), 107-124.