When less is more

John Nerbonne[i]

j.nerbonne.work@gmail.com

Groningen, Freiburg and Tübingen

1. Introduction

The more reflective among the 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 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, we'd like to resist the apparently ineluctable dynamic toward more on the side of the *explanans*.

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

[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, and I hope this essay may lead to another.

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

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 (Wieling et al. 2007). 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. From Wieling & Nerbonne (2011)

Further investigation revealed that the 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 the Belgian-Netherlands dialect border (Figure 1) might turn out to be a (collective) field worker 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.

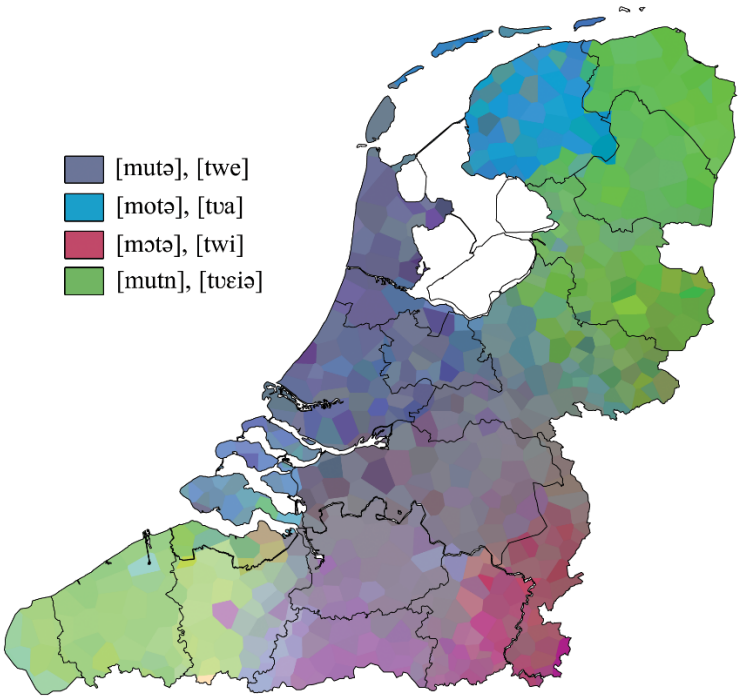

Figure 3 A comprehensive illustration of the dialect differences in the GTRP, once the phonetic system was simplified and multidimensional scaling was applied. The legend indicates by color the pronnuciation of moeten 'must' and twee 'two' . See Wieling & Nerbonne (2011) for details.

Before saying more about this particular less discriminating data categorization, let's anticipate the obvious complaint that information might have been discarded. Has reducing the phonetic inventory of the transcription failed to preserve information? Yes, guilty! 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 verbal 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 linguistics differences 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.

A second case where progress has built on ignoring phonetic detail comes from typology. Søren Wichmann and colleagues created the Automated Judgment of Similarity Program (AJSP) database (2022) around 2008 (Brown et al. 2008), which now contains word pronunciation transcriptions of over 5.500 of the world's nearly 8.000 languages (according to the ISO 639-3[3] list). The database has been used to date splits in language families (Holman et al. 2011), to determine the likely homelands of language families (Wichman et al. 2010b), and to investigate sound symbolism (Wichman et al. 2010a). And all this is accomplished using an alphabet of only forty-one sound segments![4] What is lost in phonetic discrimination is compensated by the simplicity of transcription and the degree to which comparison is facilitated. Less can be more!

3. Another case: Grammatical categories and the like

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. The strategy is undoubtedly risky, but the consensus in the field is that it has supported progress. 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

[3] See https://en.wikipedia.org/wiki/List_of_ISO_639-3_codes (consulted 3. July 2023).
[4] Allowing, however co-articulation combining two or three segments.

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 using Watson's (1996) Finnish-Australian English corpus. 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. The authors chose to tag the conversations using one of the tag sets of the International Corpus of English (TOSCA-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 one sentence from the Watson corpus (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 bring this example up here because the work might have been simpler if a smaller tag set had been sued, i.e., a linguistically less discriminating one. In fact, 75 of the TOSCA-ICE tags were not instantiated at all in the 3×10^5 words of the Watson corpus. 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.

But we might add that more ambitious collections of syntactic data have also insisted on a restricted theoretical inventory. Joakim Nivre and colleagues in several other places began this effort separately (see Nivre et al. 2016 for a very brief history), which require coordinating not only on a theoretic scheme (dependency grammar), but also on a tag set, and a scheme for morphological annotation. This required an enormous effort in grammatical research and in coordination, but there are now treebanks (collections of grammatically annotated texts) in over one hundred languages (de Marneffe et al. 2021). De Marneffe et al. (2021) discuss not only work in application-oriented natural language processing that has

exploited the resource, but also work in psycholinguistics and word-order typology (p.304). The authors emphatically wish to provide a resource for a wide range of pure and applied research topics, which understandably makes them less inclined to simplification. Indeed, they describe the work as adhering to the "Goldilocks" principle, using neither very simple nor very complex subsystems, like the girl in the children's fable, whose porridge had to be neither too hot nor too cold, but just the right temperature. The analysis scheme strives toward an emphasis on what's common among languages without "obscuring genuine differences" (p.302). Given their goal of providing a resource of general utility, supporting both practically useful language technology as well as cross-linguistic research of various sorts, it wouldn't be sensible to overdo the wish for simplicity that has been emphasized here. It is clear that, in aiming at great generality, one mustn't simplify too much.

4. Discussion and conclusions

We discussed only four 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.

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.[5] 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 the 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.

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

As the introduction stated, we should resist the apparently ineluctable dynamic toward more on the side of the *explanans,* in particular, the dynamic toward ever finer distinctions. The virtue in keeping explanations simple lies in keeping analyses comparable (the case of the phonetic inventories), 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.

References

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

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world′ s languages: a description of the method and preliminary results. *Language Typology and Universals*, *61*(4), 285-308.

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.

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255-308.

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. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. Dissertation, University of Groningen. Available at https://research.rug.nl/en/publications/

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.

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., ... & Egorov, D. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, *52*(6), 841-875.

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.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proc. 10th Int. Conf. Language Resources and Evaluation (LREC'16)*, 1659-1666.

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

Watson, G. J. (1996). The Finnish-Australian English corpus. *ICAME JOURNAL*, *20*, 41-70.

Wichmann, S., Holman, E.W., & Brown, C.H. (eds.). (2022). The ASJP Database (version 20). Available at https://asjp.clld.org/

Wichmann, S., Holman, E.W., & Brown, C.H. (2010a). Sound symbolism in basic vocabulary. *Entropy* 12.4: 844-858.

Wichmann, S., Müller, A., & Velupillai, V. (2010b). Homelands of the world's language families: A quantitative approach. *Diachronica*, *27*(2), 247-276.

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.