

The certainty of chance

Review of Bod, R., Hay, J. and Jannedy, S. (2003)
Probabilistic Linguistics.

Robert Munro

This text describes the use of probabilistic models of language across a variety of linguistic fields, from variation in the wavelength of frontal vowels to the semantic contingency of statements like *birds fly*. Probabilities are described in terms of the motivational factors, theoretical models of probabilities/gradation, computational representations, and in the analysis of specific examples. The chapters are authored by leading and/or emerging linguists in the respective fields and are a useful reference for older supporting work, with an impressive bibliography. In this review I focus in particular on the computational modeling of probabilities and gradations.

The first two chapters give an introduction to probabilities in linguistics and to formal probability theory. The following general motivating factors for the inclusion of probabilities are discussed: variation, frequency, gradience, category membership, well-formedness, morphological productivity, morphological decomposition, the argument / adjunct distinction, acquisition and universals. If this isn't enough, more are given in the respective chapters, and remaining sceptics can turn to Chapter 3 where Jurafsky addresses common criticisms/queries of probabilistic models such as 'surely you don't think people are doing complex maths in their heads' (p. 89), 'are probabilistic models always non-modular' (p. 90), 'corpus frequencies don't match norming study frequencies' (p. 90), and 'maybe frequency is just an epiphenomenon of other structural factors' (p. 92). The last two also relate to the use of corpora to derive probabilities.

Affiliation

Robert Munro, Department of Linguistics, School of Oriental and African Studies, University of London
email: rmunro@soas.ac.uk

In Chapter 2, 'Introduction to elementary probability theory and formal stochastic language theory', Bod gives an account of Bayesian probability theory using language examples. It is as good an introduction to probability theory as can be found anywhere. The only criticism is that there is too much attention paid to the different methods of modelling the {A → B C} relationships in probabilistic grammars, such as Probabilistic Context-Free Grammars (PCFG) and Probabilistic Tree Substitution Grammars (PTSG). For a reader without a background in probability theory some of the proofs might be hard to follow, especially those showing that 'there exists a PTSG for which there is no strongly stochastically equivalent PCFG' (p. 34) and 'for every PCFG there exists a strongly stochastically equivalent PTSG' (p. 34). As a computer scientist I found the proofs themselves interesting, but as a linguist it wasn't clear to me that I needed to understand these for this text, or that they are applicable to many languages (although the fundamental concepts are implicit in the accessing of linguistic structure from the mental lexicon or grammar).

Missing from Chapter 2 is mention of the many popular methods for measuring and modelling gradation. This was presumably because good introductory accounts are given in later chapters. Chapters 3 and 4 give the best introductions to computational modelling and machine learning in the text and for this reason are useful reading, even if the particular fields are not of immediate interest.

In Chapter 3, 'Probabilistic modeling in psycholinguistics: linguistic comprehension and production', Jurafsky describes the role of probabilities in psycholinguistics. In language comprehension, probabilities are described as having three roles: accessing linguistic structure from the mental lexicon or grammar; disambiguation in interpretation; and in processing difficulty. The importance of lexical frequency is described in both comprehension and production. This extends to the frequency of semantic forms and lexical and morphological categories, such as the ambiguity between participle and preterite readings of terms such as *searched* and *selected*. Interestingly, some research reported here suggests that the frequency of semantic lemma is important in language comprehension but might not be important in language production.

Different metrics for representing the collocational tendencies of words are described, as is the importance of syntactic subcategorisation frequencies in comprehension. These lead to a discussion of probabilistic architectures, including constraint based models, rational and utility-based probabilistic models, Markov models, Stochastic Context-Free Grammars (including their use in predicting reading time), and Bayesian Belief Networks. One conclusion is that the relationship between connectionist and probabilistic architectures is an important area of future research.

In Chapter 4, 'Probabilistic sociolinguistics: beyond variable rules', Mendoza-Denton, Hay and Jannedy describe analysing language with the use of

VARBRUL, a set of analysis programs utilising binary logistic regression and Classification and Regression Trees (CART), a machine learning algorithm. While decision trees are notoriously unstable learners (small changes in the data can lead to large changes in the models learned, making analysis complicated) these authors state that the patterns found were fairly reliable. The description of how a decision tree is built from the data is given in plain language, making it accessible reading for those without a background in statistics. It includes explanations of pruning and cross-validation. Drawing on the empirical focus of sociolinguistics, they describe how linguistic innovation, nuances in speech patterns and variants of lexical choice are ‘manifestations of the subtle patterning and interaction of linguistic and social competence’ (p. 99), making it a prime candidate for analysis through machine learning.

The analysis describes the monophthongization of the diphthong /ay/ in the speech of Oprah Winfrey. Readers of this journal might be disappointed that the authors do not focus on more complex probabilistic relationships between social and language structures, but their choice makes the chapter accessible to a broad linguistic audience and the social structures affecting the phonological variation here are shown to be quite rich. Many readers should welcome their conclusion that ‘[an] important challenge for the future will be to move toward a more unified understanding of how subtle gradient patterns of variation affect and are affected by cognitive, linguistic, and social structures’ (p. 137).

In Chapter 5, ‘Probability in language change’, Zuraw explains how an ‘S-shaped’ curve can be mapped to the uptake of change: new variants appear rarely, followed by a quick increase in frequency, with the rate of change then slowing as the maximum frequency is reached. It is shown how this can be modelled with a logistic function such as that used in VARBRUL, whose slope can be explicitly measured and compared to a constant-change hypothesis. Also discussed is the shift in probabilities over time, with a word’s frequency affecting its susceptibility to phonological, morphological, and morphosyntactic change.

For Chapter 6, ‘Probabilistic phonology: discrimination and robustness’, Pierrehumbert has the easiest case to make for probabilistic distributions. Here, another popular method for modelling gradience is used: assuming a normal (Gaussian) distribution over measured values, giving an example of the observed distributions of the wavelengths (frequency) of two front vowels. The point is made that discrimination is not significance, that is, we can be certain that two distributions are independent, even when they intersect. Because of this, we may not be able to predict whether a given instance belongs to a certain distribution. Pierrehumbert outlines how the outcome of two overlapping categories can be either a ‘sharpening up’ to reduce the overlap, or that one will be ‘eaten up’ by the other. She cites work that indicates that overlap varies

according to context and, interestingly, where two distinct categories are so close in wavelength that the speakers themselves cannot distinguish the two.

In Chapter 7, 'Probabilistic approaches to morphology', Baayen begins with a succinct account of the co-growth of computing power and the use of statistics in linguistics. An interesting critique of Artificial Neural Networks (ANNs) is given here. It is noted that to understand how an ANN decides upon a particular representation requires an understanding of multivariate statistical techniques. The difficulty in analysing ANNs, relative to other machine learners, has been a serious limitation in their use for analytical purposes. The most interesting question in the text is posed here: 'should [machine learners] be able to generalise outside the scope of their training space?' (p. 232).

In Chapter 8, 'Probabilistic syntax', Manning describes the relatively unexplored area of probabilistic approaches to syntax. He begins with an account of how categorical approaches to subcategorisation frames that make only grammatical / ungrammatical distinctions either fail to identify some rare complements or fail to fully describe the limitations of their use. This is extended to an explanation of how probabilities can be calculated over more complex hidden structures, how probabilities capture the notion of grammaticality, and, importantly, how stochastic grammars can retain formal models of learnability.

A description of 'continuous categories' is given with examples of words functioning as both verbs and prepositions (e.g. *concerning*, *regarding* and *following*), noting that the bulk of statistical Natural Language Processing (NLP) has simply calculated the probabilities over discrete structures and values, that is, used probabilities as confidence rather than gradation.

The discussion moves to constraint-based theories, with the motivation for probabilistic constraints being: 'categorical phenomena that are attributed to hard grammatical constraints in some languages continue to show up as soft constraints in other languages' (p. 316). A brief introduction to optimality theory (OT) is given. OT provides an interesting alternative to probabilistic modelling for some phenomena as it allows constraint violation within a discrete categorical framework through the modelling of a hierarchy of constraints. This leads into a description of stochastic optimality theory; constraint distributions at certain distances within a continuous space. Manning notes that stochastic OT is probably a better model for production (generation) than comprehension (parsing / interpretation), as its use of a hierarchy disallows the ganging-up by low-ranked constraints that could represent a wide variety of possible contextual influences. A good outline is also given of popular statistical approaches to modelling probabilities, such as log-linear modelling and generalised linear modelling.

In Chapter 9, 'Probabilistic approaches to semantics', Cohen makes arguments for probabilities in formal semantic theory. He explains how the philosophical

theory of probabilities can be directly modelled in statements such as ‘birds fly’, and that indirect use can still be made of mathematic models. The content is mostly outside my area of knowledge so it’s difficult to comment on, except to say that computational modelling seems to be only loosely relevant to the theory described here: ‘the *mathematical* theory of probability provides no semantics for probability judgements’ (p. 343).

Overall, *Probabilistic Linguistics* gives a very broad spread of the use of probabilities in different linguistic subfields, but the focus of the text is certainly probabilities, not gradations. A reader looking for formalisms for modelling continuous categories and gradient patterns in linguistics will be disappointed – surely the topologies of gradations are much richer than a smooth indeterminate space between categories? Many of the authors here flag new formalisations of gradational modelling as important future work, but none attempt it. Jurafsky notes that 77 per cent of the Association for Computational Linguistics (ACL) conference papers in 2000 used some form of probabilistic modelling. Within the same papers, I only found a couple that focused on probabilities as a target concept and none that made modelling gradations the focus of a paper, so while 77 per cent of researchers were happy to incorporate probabilities into their models, most evaluated them through categorical accuracy measures. In short, the computational practice of focusing on improved empirical outcomes may have driven a move towards probabilistic modelling, but could the need to easily judge the empirical quality of outcomes ultimately hold back probabilistic linguistics?

Gaps are inevitable in collections, and none were critical to the arguments made by the respective authors, but I would also liked to have learned more about the fields that have a longer history of modelling probabilities. Unfortunately, the exceptions were not detailed. For example, we’re told only in passing that probabilistic models of constraints are established in typological and functionalist theories, and that it’s desirable to follow the current functionalist practice of attempting to predict the overall rate of the different systemic choices for a given input. Of course, the text seems to be addressed to a linguist from a categorical background, so arguing for probabilities in the exceptions isn’t necessary: they’re already there. In any case, the computational models are just as relevant and linguists from any background can use this text as a valuable introduction to computational modelling and machine learning.

Book reviewed

Bod, R., Hay, J., Jannedy, S. (2003) *Probabilistic Linguistics*. Cambridge, MA: MIT Press. xii + 451 pp. ISBN 0262025361.