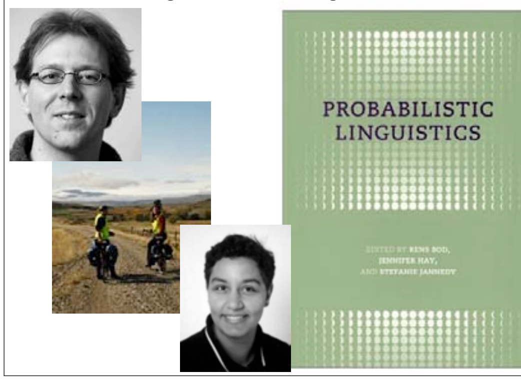
The Great Number Crunch

Charles Yang (2008)

Models of Language Class

Kenny Vaden

Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), Probabilistic linguistics. Cambridge, MA: MIT Press, 2003.



1. Introduction

Models of Probability Book:

1) Demonstrated uses of probabilistic language models
2) Superiority over Generative Grammar and Maxim of Categoricity

> "One of the foundations of modern linguistics is the maxim of categoricity: language is categorical. Numbers play no role, or, where they do, they are artifacts of nonlinguistic performance factors."

However, Yang points out:

Logical Structure of Linguistic Theory (Chomsky) was explicitly set up to include probability concepts

Variationist Analysis already developed as an approach to study the <u>distribution</u> of <u>discrete choices</u>.

2. Highlights from *Probabilistic Linguistics*:

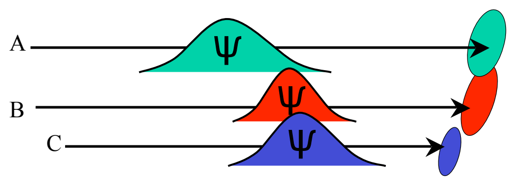
1. **Jurafsky:** broad summary of probabilistic effects & models in psycholinguistics, *plus potential objections*
2. **Pierrehumbert** models American English vowel variation
3. **Baayen** on morphological productivity as it calls on storage or computation in the mental lexicon
4. **Cohen** discusses usage for frequency adverbs (*always, sometimes, often*) while other applications not as well developed

… models of language change (Zuraw), phonological adaptation in Oprah Winfrey's speech (Mendoza-Denton, Hay, Jannedy), linguistic corpora and theory (Manning), etc.

3. Linguistic Probability and Linguistic Theory

Probabilistic effects do not automatically constitute a rebuttal of categorical linguistics

Probabilities may be performance factors, cognitive components acting on discrete linguistic components



3.1 Probabilistic Facts

One Concern:
Categorical Tasks vs. Gradient Tasks

Another: weakly substantiated claims

Zuraw: frequent words adopt automatic phonetic rules first

example: t/d deletion in word-final consonant clusters

Yang bashes Zuraw – actually a stable speech variation, seen in children and adults, however *lenition*-weakening seems to occur in high frequency words…

Yang indicates that linguistic facts founded only on probabilistic bases are **not trustworthy**. *Us vs. them??*

3.2 Variation & Grammar

Book Claim: Categorical Linguistics only study endpoints, ignoring *gradient middle*

X No categorical models prohibit the use of frequencies.

X Variationist Perspective (Labov, 1969) has long held statistical data as part of picture, just as categories are.

Yang:
The categorical linguist's interest in the endpoints … provide the very units of distribution that the probabilistic linguist works with.

3.3 The Locus of Linguistic Probability

Competence vs. performance issue in Grammar and its use – broader conflict

Yang focuses on probabilistic effects – not homogeneous, and may be accommodated by existing models.

1. **Linguistic levels:** distinct domains can produce joint products that blur phonemic distinctions.

2. **Interaction with Cognitive and Perceptual systems** means that probabilistic effects are difficult to attribute to specific language faculty, even language *per se*.

3.3 Example of Lexical Access and Frequency

More frequent items are recognized faster.

Forster's Bin Model: items ranked by freq. (hi = fast)

Model explains all manner of frequency effects, plus pseudowords are accessed more slowly.

Discrete linguistic model produces stochastic results.

With models that place language faculties into a context of cognitive/perceptual components – gradients will fall out even with categorical language processes

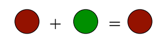
3.4 Probabilistic evidence for categorical linguistics

Natural sciences: variation provides compelling evidence for a discrete system

Evolution depends on statistical distribution of phenotypes:

brown eyes: 80/100
green eyes: 20/100

We infer underlying genotypes on the basis of distributions



Similarly, probabilistic variation in *phenotypes of language*: use, learning, and change – may reflect the underlying system of <u>discrete linguistic units</u>

3.4 Probabilistic evidence for categorical linguistics

The rise of periphrastic ***do*** in the history of English

- Change follows the normal pattern, probabilistically; takes place gradually, characterized by a mixture of linguistic forms whose distribution is in fluctuation.

"**<u>Do</u>** send the periphrastic email, Pernille."

Categorically: this use of *do* is huge departure from past!

Kroch (1989) provides statistical evidence that the uses of do in several seemingly unrelated constructions follow the same trajectory of change. A semantically empty "do" emerged as other inflective forms disappeared from Middle English.

Correlations are not accidental; grammatical change must be attributed to the change in a SINGLE syntactic parameter.

3.4 Probabilistic evidence for categorical linguistics

Language learning in the child

How do changes in Grammar occur over time?

Quantitative analyses of child language shows variation that cannot be attributed to **a single potential grammar**.

BUT if we interpret LEARNING as probabilistic and GRAMMAR as categorical, then variation in child language is a statistical ensemble of possible grammars whose distribution changes over time.

3.4 Probabilistic evidence for categorical linguistics

Yang does a balancing act:

Probabilistic aspects of language learning, use, and change do raise a challenge – **to categorical models** of learning, use, and change, but **not** to the **categorical view of language itself.**

4. Data, Model, and Inference

Yang states that the book understated the difficulty of problems facing probabilistic models while overstating their accomplishments.

At the same time, he recognizes that probabilistic linguistics has a place in the study of language.

4.1 Probability, Reality, and Computation

Probabilistic models incorporate performance into their explanatory scope, so are more accountable for matching **reality**.

General Learning Algorithm

Formal properties, learnability, convergence time, etc., and unlearnable datasets (rigorous evaluations that are standard in natural language processing) not applied.

Unsupervised Cluster Analysis for Vowel Learning

Used k-means algorithm to find k clusters of vowel formant data, supposedly like infant learning; k is set by the programmer!

4.1 Probability, Reality, and Computation

Problems with probabilistic models in general:

Sparse data problem**:** as the model gets richer, number of parameters set increases exponentially

Independence Assumption**:** cleans up interactions to pretend components can just be summed or multiplied, *but in a linguistic expression, hardly any two items are ever independent*

Scalability issue: surprisal for word-parsing in sentences requires calculation of probabilities for infinite alternative strings – scaled down by appealing to … **categorical syntactic structures.**

Abstractions based on linguistic categories may hold the key to empirical progress for probabilistic models, then they may scale up to naturalistic data.

4.2 The Case of Missing Data

How to use linguistic data properly is a difficult question, particularly **corpus usage**.

'*Not every regularity in the use of language is a matter of grammar*' (Zwicky & Pullum, 1987).

VS . Baayen ~ *If Google can find it, it is linguistics.*

Baayen used examples of -th as a nominalizing suffix in gloomth, greenth, *and* coolth *from the Internet to make some point about language.*

- Yang thinks this is uncoolth.

4.2 The Case of Missing Data

There are good uses of corpora, CHILDES

(MacWhinney 1995)

Evidence for Productive vs. Unproductive Morphology (Categorical) Distinction from CHILDES:

Unproductive Irregular Past Tense Pattern:

wipe-wope, bring-brang, trick-truck, walk-has walken

(Xu & Pinker found these occur in 0.2% of feasible places.)

Overapplication of Productive Regular Past Tense Pattern:

Compounds: *mice-eater* 90% <u>versus</u> *rats-eater* 2%

(Pinker 1999 states these are very common)

4.2 The Case of Missing Data

Questions about corpora use

1. Why does only the corpus count as 'verifiable linguistic data'? (*as Manning allegedly asserted*)

   Yang then points out how corpora are "highly sensitive to genre and style" while GRAMMATICALITY JUDGMENTS *can be confirmed or rejected…*

X ☹ *WRONG, YANG*

4.2 The Case of Missing Data

Questions about corpora use

2. Goal of linguistic theory is to est. bounds of possible *and impossible* linguistic forms – corpora only show a handful of possibilities in the ocean of infinite generativity…

Yang screws it up here, disliking Pierrehumbert's statement: 'Statistical underrepresentation must do the job of negative evidence', then citing Saffran et al. (1996) as the example of ***presumed*** frequency learning over linguistic units.

"I don't see exactly how learning something in the data tells the learner what is not in the data." - Yang

IT DOESN'T! - Safran showed **phonemic patterns** are learnable during <u>brief exposures</u> to artificial language, **not rules**.

4.3 Probabilistic Learning & Language Learning

What is necessary to learn a Grammar?

Relaxation of discreteness can simply computations, and produce provably superior formal properties for language learning.

Comparing 2 related but distinct frameworks of learning:

1. Categorical framework of **Gold** (1967) requires exact identification of the target hypothesis

2. Probably Approximately Correct (PAC) framework only requires learner to get close to the target, but within reasonable bounds of computational resources.

4.3 Probabilistic Learning & Language Learning

With finite hypotheses, learnability is ensured.

One conclusion from both frameworks:

Learning is not possible unless the hypothesis space is tightly constrained by prior knowledge / Universal Grammar

TWO Big Questions:

How are grammatical hypotheses are *scattered* such that they are distinguished by data in a computationally tractable way?

Are the language learning models psychologically plausible; matching language development data?

5. Conclusion

Book presents a diverse range of probability applications in linguistic study

But it fails to respect real progress made by traditional generative approach, while adapting those methods for their own goals – something anticipated by Chomsky and Labov, but not acknowledged in the current volume.

Yang points out that probabilistic models are gaining ground, but that they cannot completely replace their own framework!