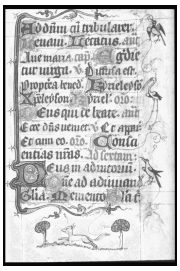
Ling 235: Quantitative and Probabilistic Explanation in Linguistics

Christopher Manning
and Roger Levy

Winter 04/05

Models for language

- Human languages are the prototypical example of a symbolic system
- From the beginning, logics and logical reasoning were invented for handling natural language understanding
- Logics and formal languages have a language-like form that draws from and meshes well with natural languages
- Where are the numbers?



Linguistic Facts vs. Linguistic Theories

- "Everyone knows that language is variable"
  - Sapir (1921: 147)
- Weinreich, Labov and Herzog (1968) see 20th century linguistics as having gone astray by mistakenly searching for homogeneity in language, on the misguided assumption that only homogeneous systems can be structured
- Probability theory provides a method for showing structure in variable systems

The quest for homogeneity: Bloch (1948: 7)

- "The totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker is an *idiolect*. … The phrase 'with one other speaker' is intended to exclude the possibility that an idiolect might embrace more than one *style* of speaking."

But variation is everywhere

- The definition fails, as variation occurs even internal to the usage of a speaker in one style.
- *As least as:*
  - Black voters also turned out *at least as* well as they did in 1996, if not better in some regions, including the South, according to exit polls. Gore was doing *as least as* well among black voters as President Clinton did that year.
    - (Associated Press, 2000)

Except in Linguistic Theory! Joos (1950: 701-702)

- Ordinary mathematical techniques fall mostly into two classes, the continuous (e.g., the infinitesimal calculus) and the discrete or discontinuous (e.g., finite group theory). Now it will turn out that the mathematics called "linguistics" belongs to the second class. It does not even make any compromise with continuity as statistics does, or infinite-group theory. Linguistics is a quantum mechanics in the most extreme sense. All continuities, all possibilities of infinitesimal gradation, are shoved outside of linguistics in one direction or the other.

Probabilistic models in areas related to grammar

- Human cognition has a probabilistic nature: we continually have to reason from incomplete and uncertain information about the world
- Language understanding is an example of this
  - P(*meaning* | *utterance*, *context*) [cf. NLP]
- Language acquisition is an example of this
  - Both early formal (e.g., Horning 1969) and recent empirical (e.g., Saffran et al. 1996) results demonstrate the effectiveness of probabilistic models in language acquisition
- What about for the core task of describing the grammar of a human language?

7

The need for frequencies / probability distributions

The motivation comes from two sides:

- Categorical linguistic theories claim too much:
  - They place a hard categorical boundary of grammaticality, where really there is a fuzzy edge, determined by many conflicting constraints and issues of conventionality vs. human creativity
- Categorical linguistic theories explain too little:
  - They say nothing at all about the *soft constraints* which explain how people choose to say things
    - Something that language educators, computational NLP people – and historical linguists and sociolinguists dealing with real language – usually want to know about

8

1. The hard constraints of categorical grammars

- Sentences must satisfy all the 'rules' of the grammar
  - One group specifies the arguments that different verbs take – lexical subcategorization information
    - Some verbs must take objects: **Kim devoured*
      [* means ungrammatical]
    - Others do not: **Dana's fist quivered Kim's lip.*
    - But there's a lot of squishyness
- The problem with this model was noticed early on:
  - "All grammars leak." (Sapir 1921: 38)
- In Intro Syntax classes, people complain
- In NLP systems, ungrammatical sentences don't parse

9

Bresnan and Nikitina (2003) on the Dative Alternation

- Pinker (1981), Krifka (2001): verbs of instantaneous force allow data alternation but not "verbs of continuous imparting of force" like *push*
  - As Player A **pushed him the chips**, all hell broke loose at the table.
    - www.cardplayer.com/?sec=afeature&art id=165
- Pinker (1981), Levin (1993), Krifka (2001): verbs of instrument of communication allow dative shift but not verbs of manner of speaking
  - "Hi baby." Wade says as he stretches. You just **mumble him an answer**. You were comfy on that soft leather couch. Besides …
    - www.nsyncbitches.com/thunder/fic/break.htm
- In context such usages are unremarkable!
  - It's just the productivity and context-dependence of language
  - Note that you need lotsa data for this game! (Here the web.) These things are rare → because these are *gradient* constraints

10

2. Explaining more: What *do* people say?

- Labov (1972: 207):
  - The basic sociolinguistic question is "why anyone says anything".
- This is too broad for what we want to cover … or what Labov has done
- Analogy from Natural Language Generation:
  - *strategic generation:* deciding which ideas to express and combine, how to interact, etc.
  - *tactical generation:* working out how to express those ideas felicitously in language

11

2. Explaining more: What *do* people say?

- What people *do* say has two parts:
  - Contingent facts about the world
    - People have talked a lot about tsunamis recently
  - The way speakers choose to express ideas using the resources of their language
    - People don't often put *that* clauses pre-verbally:
      - *That we will have to revise this program is almost certain*
- The latter is properly part of people's Knowledge of Language. Part of linguistics.

12

What *do* people say?

- Simply delimiting a set of grammatical sentences provides only a very weak description of a language, and of the ways people choose to express ideas in it
- Probability densities over sentences and sentence structures can give a much richer view of language structure and use
- In particular, we find that the *same* soft generalizations and tendencies of one language often appear as (apparently) categorical constraints in other languages
- Linguistic theory should be able to uniformly capture these constraints, rather than only recognizing them when they are categorical

13

Model

- People have some idea they want to express
- To express it, they are choosing between various forms, such as active, passive, topicalized:
  - I really like Izzy's bagels
  - Izzy's bagels, I really like
  - Izzy's bagels are really liked by me. [???]
- People choose a form on the basis of discourse, grammatical, and many other (soft) constraints

14

Is variation only for sociolinguists?

- On the ground, sociolinguists have dealt with variation
- Syntacticians mainly haven't
- One might conclude from this that most syntactic variation only has a sociological or stylistic basis.
- But I don't think this is true:
  - Most variation is probably sensitive to information structure, viewpoint, etc.
- Cf. NLP where people greatly want to model variability, but normally not sociolinguistically

15

Systemic functional grammar?

- This viewpoint is largely similar to the main idea of systemic functional grammar (as I understand it):
  - Choosing how to realize an idea using the resources provided by a language
- Any particular realization may not include all of the idea, and may have various other connotations (cf. machine translation)

16

Competence/performance

- Traditional linguistics has dealt with variation by merely defining the *types* that are considered grammatical
- The choice of a *token* from the *types* in a particular context is then a matter of performance
- But we need a theory of this choice
- This kind of supposed 'performance' has little to do with real processing limitations in production

17

Variation is part of competence: (Labov 1972: 125)

- "The variable rules themselves require at so many points the recognition of grammatical categories, of distinctions between grammatical boundaries, and are so closely interwoven with basic categorical rules, that it is hard to see what would be gained by extracting a grain of performance from this complex system. It is evident that [both the categorical and the variable rules proposed] are a part of the speaker's knowledge of language."

18

'The Conditional Model'

- *Context* + *Meaning* → *Form*

  We have an incomplete model of this, and there may just be inherent randomness in the mapping (neurons, etc.)

- Want:
  - P(*form* | *context, meaning*)
- This is hard to object to, but so far largely empty

19

Probabilistic models of Knowledge of language

- *That the sport has become popular enough to warrant a special workout program is no real surprise.* (*NYT*, 1994, discussing rollerblades)
- *It was no surprise that the nascent Palestinian authority would face enormous difficulties creating equitable social, legal and political systems from the chaos left behind when Israel's troops and military government departed in May, ending 27 years of occupation in Gaza and Jericho.* (*NYT,* 1994)

20

Some more examples (writ small)

- That he did so surprised some top administration officials, most of whom learned only Friday afternoon that Clinton had decided to send former President Jimmy Carter to Port-au-Prince.
- It should come as no surprise that in his foreign policy showdowns, President Clinton has made economic sanctions his weapon of choice, even though such measures have, at best, had a middling record of success.
- It isn't like that, and it surprises me that people don't know that.
- It doesn't surprise Moore that no one seems to be able to describe what this thing is, because it introduces an entirely new genre.

21

Building Models

- Categorical model:
  - Both are possible, grammatical sentences
- Simplest probabilistic model:
  - P([S [SBAR[*that*]] VP]) = 0.02
  - P([S *it* VP [SBAR[*that*]]]) = 0.98
- Mildly more complex model
  - P([S [SBAR[*that*]] VP] | C = written) = 0.04
  - P([S *it* VP [SBAR[*that*]]] | C = written) = 0.96
  - P([S [SBAR[*that*]] VP] | C = spoken) = 0.01
  - P([S *it* VP [SBAR[*that*]]] | C = spoken) = 0.99

22

Better models

- Length of SBAR clause
  - P([S [SBAR[*that*]] VP] | C = written, len(SBAR) = k) = $0.5e^{-0.5k}$
- Or maybe it's relative to the length of the VP?
  - P([S [SBAR[*that*]] VP] | C = written, len(SBAR) – len(VP) = k) = $0.5e^{-0.5k}$ ($k \geq 0$, 0 otherwise)
- Depends on information structure:
  - P([S [SBAR[*that*]] VP] | C = written, topic SBAR) = 0.08
- Whole aim is to identify most predictive features

23

Where do our models come from?

- Two possibilities:
  - Start from a theoretically/typologically-motivated model, and test it with data
  - Start with a lot of data, and see what seems to explain the patterning in the data (exploratory data analysis)
- The first is more traditional. The latter is more feasible now with computers and big data sets.
  - I think both can be valuable.
  - Both can be evaluated for empirical adequacy

24

What's achieved

- We can minimally give a better description of language
- We can rigorously evaluate how good our theory is
- By allowing the incorporation of soft functional pressures into formal models, we can hope to give better explanatory accounts of language as well. Cf. Prince and Smolensky…

25

Prince and Smolensky (1993: 198)

- "When the scalar and the gradient are recognized and brought within the purview of theory, Universal Grammar can supply the very substance from which grammars are built: a set of highly general constraints, which, through ranking, interact to produce the elaborate particularity of individual languages."

26

Conclusions

- There are many phenomena in language that cry out for non-categorical and probabilistic modeling and explanation
- Probabilistic models can be applied on top of one's favorite sophisticated linguistic representations!
- Frequency evidence can enrich linguistic theory by revealing soft constraints at work in language use
- Probabilistic syntactic models increase the interestingness and usefulness of theoretical syntax to neighboring academic communities

27