Ling 236: Quantitative, Probabilistic, and Optimization-Based Explanation in Linguistics

Christopher Manning

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 variation

Seeking 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 including 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)

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.

Dominant answer in linguistic theory: Nowhere

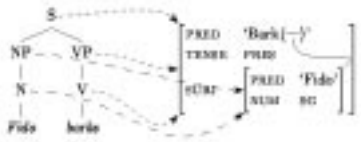
Chomsky (1969: 57; also 1956, 1957, etc.):

- "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." [cf. McCarthy in AI]
- Probabilistic models wrongly mix in world knowledge
  - New York vs. Dayton, Ohio
- They don't model grammaticality [also, Tesnière 1959]
  - *Colorless green ideas sleep furiously*
  - *Furiously sleep ideas green colorless*
- Don't meet goal of describing I-language vs. E-language
  - Perhaps. Perhaps not. But E-language is empirical.

7

Categorical linguistic theories (GB, Minimalism, CG, HPSG, LFG, ...)

- Systems of variously rules, principles, and representations is used to describe an infinite set of grammatical sentences (*types*) of the language
- Other sentences are deemed ungrammatical
- Word strings are given a (hidden) structure
- Only models free variation. No relative goodness, etc.



8

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 syntax – the grammar – of a human language?

9

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

10

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.*
    - Others take various forms of sentential complements
- In NLP systems, ungrammatical sentences don't parse
- But the problem with this model was noticed early on:
  - "All grammars leak." (Sapir 1921: 38)

11

Transitive/intransitive verbs

- How well can verb usage by captured by such sets of categorical subcategorization frames?
- Consider:
  - *quake*
  - *quiver*
- Intransitive only?
  - This is what all three of COBUILD, LDOCE and OALD say for *quake*, and what COBUILD and LDOCE say for *quiver* (OALD does allow transitive usages of *quiver* - interestingly, here, OALD is the one more based on intuition than corpora)

12

Atkins and Levin corpus study

- Atkins and Levin (1995; *Intl Jnl Lexicography*):
  - Find transitive usages of both verbs
    - Using 50 million word corpus – you need lotsa data to do empirical syntax!
  - *It quaked her bowels*
  - *The bird sat, quivering its wings*
- In context such usages are unremarkable
  - It's just the productivity of language

13

Example: verbal clausal subcategorization frames

- Some verbs take various types of sentential complements, given as subcategorization frames:
  - *regard:* __ NP[acc] *as* {NP, AdjP}
  - *consider:* __ NP[acc] {AdjP, NP, VP[inf]}
  - *think:* __ CP[*that*]; __ NP[acc] NP
- **Problem:** in context, language is used *more flexibly* than this model suggests
  - Most such subcategorization 'facts' are **wrong**

14

Standard subcategorization rules (Pollard and Sag 1994)

- We consider Kim to be an acceptable candidate
- We consider Kim an acceptable candidate
- We consider Kim quite acceptable
- We consider Kim among the most acceptable candidates
- *We consider Kim as an acceptable candidate
- *We consider Kim as quite acceptable
- *We consider Kim as among the most acceptable candidates
- ?*We consider Kim as being among the most acceptable candidates

15

Subcategorization facts from *The New York Times*

Consider as:

- The boys consider her as family and she participates in everything we do.
- Greenspan said, "I don't consider it as something that gives me great concern.
- "We consider that as part of the job," Keep said.
- Although the Raiders missed the playoffs for the second time in the past three seasons, he said he considers them as having championship potential.
- Culturally, the Croats consider themselves as belonging to the "civilized" West, …

16

More subcategorization facts: *regard*

Pollard and Sag (1994):

- *We regard Kim to be an acceptable candidate
- We regard Kim as an acceptable candidate

The New York Times:

- As 70 to 80 percent of the cost of blood tests, like prescriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.
- Conservatives argue that the Bible regards homosexuality to be a sin.

17

More subcategorization facts: *turn out* and *end up*

Pollard and Sag (1994):

- Kim turned out political
- *Kim turned out doing all the work

The New York Times:

- But it turned out having a greater impact than any of us dreamed.

Pollard and Sag (1994):

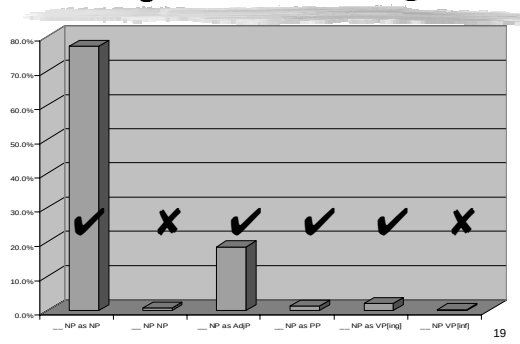
- Kim ended up political
- *Kim ended up sent more and more leaflets

The New York Times:

- On the big night, Horatio ended up flattened on the ground like a fried egg with the yolk broken.

18

Probability mass functions: subcategorization of *regard*



19

Grammaticality and probability

- Sampson (1987) argues against a grammatical ungrammatical distinction on the basis of finding a continuous gradual drop-off of different phrase structures for noun phrases in corpora, rather than a sudden cliff.
- But this isn't sufficient (and the argument is countered in papers by Briscoe et al. and Culy)
- In part they argue for a lack of articulated syntactic structure in the analysis.

20

Grammaticality and probability

- Probabilistic CFG:
  - P(S → *this is* AP) = 1
  - P(AP → *interesting*) = 0.5
  - P(AP → *really interesting*) = 0.25
  - P(AP → *really really interesting*) = 0.125
  - P(AP → *really really really interesting*) = 0.0625
  - …
- But this grammar still has a clear grammatical/ ungrammatical distinction

21

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 I 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

22

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

- What people *do* say has two parts:
  - Contingent facts about the world
    - People in the Bay Area have talked a lot about electricity, housing prices, and stocks lately
  - 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.

23

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
- A syntactic theory should be able to uniformly capture these constraints, rather than only recognizing them when they are categorical

24

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

25

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

26

Variables

- (Socio)linguistic variables
  - are normally taken to be variable language features that covary with sociological or stylistic features
- Variables
  - things that can take on various values
- Random variables
  - Probabilistic things with *numeric* values
- We're interested in all uncertainties, even if no social significance

27

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 largely empty
  - Joint model: P(*form, context, meaning*)

28

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)

29

Halliday (1991) on 1950s

[in Aijmer and Altenberg (eds)]

- "It had always seemed to me that the linguistic system was inherently probabilistic, and that frequency in text was the instantiation of probability in the grammar."
- "To the 'instance' observer, the system is the potential, with its set of probabilities attached; each instance by itself is unpredictable, bu the system appears constant through time. To the 'system' observer, each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other (as each moment's weather at every point on the globe redefines the global climate)."

30

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
- Optionality is not a problem to standard syntactic theories. Probabilistic or frequentistic application of rules depending on features is.

31

Soft constraints on variation

- The central advance of the variable rules framework was the discovery of noncategorical grammatical constraints on variation
- Variation is not random.
- To the extent that people know about frequencies of forms, there are compelling reasons to see this as Knowledge of Language, that is, of linguistic competence
- If performance is the real processing load of sentence production, some of it is processing, but much of it clearly isn't

32

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."

33

Bender (2000)

- Systematic yet violable constraints are well-studied in sociolinguistics.
- Should they be part of competence grammar or are they just systematic performance factors?
- She aims to show that speakers have knowledge of non-categorical constraints.
- More work needed (I think) but her one (matched guise) experiment suggests that speakers have knowledge of these constraints.

34

Bender (2000)

- A more inclusive model of grammatical competence
- "The rate of copula absence in AAVE is sensitive to the part of speech of the predicate. While it is possible that these effects are a matter of (universal, functional) performance factors, the results of the experiment reported in Chapter 4 indicate that they are also a matter of linguistic knowledge. I will suggest in chapter 6 that they may also be a matter of linguistic competence."

35

Expanding competence?

- Undeniably, when one heads in this direction, the domain of competence expands
- This has been happening a lot in linguistics anyway (semantics, information structure)
- But I sense there is still a distinction between what is now in competence and genuine performance issues (hesitations, restarts, false resumptives), etc.
- However, it's a difficult issue

36

Langacker (1987: 36)

- "The notion of syntax as an autonomous formal system has encouraged the expectation that speakers should be capable of simple categorical judgments (grammatical/ ungrammatical) on the well-formedness of sentences, out of context and without regard for semantic considerations: either a sentence meets all formal specifications, or it does not."
- But it isn't this simple … as we saw before

37

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)

38

Some more examples (writ small)

- That this trip is proving so difficult is not so much a surprise as it is a sobering reminder of 1991, when the first-place Dodgers went 2-9 on their first East Coast trip after the break and let the Atlanta Braves climb from 9 out to 3 out.
- 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.

39

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

40

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

41

What's achieved

- We can minimally give a better description of language
- But by allowing us to incorporate soft functional pressures into formal models, we can hope to give better explanatory accounts of language as well. Cf. Prince and Smolensky…

42

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."

43

Conclusions

- There are many phenomena in syntax 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

44