Symbolic and Statistical Approaches to Language

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Main Reference

- Judith L. Klavans, Philip Resnik (Eds.), *The balancing act: combining symbolic and statistical approaches to language*, MIT Press, 1996.

Syntactic Parameters and Models of Language Acquisition

• Linguistic parameter space $\mathcal{H}$ with $|\mathcal{H}| = 2^N$ for N parameters

• problem of locating a target grammar $\mathcal{G}$ in this parameter space on the basis of input text in the language $\mathcal{L}_{\mathcal{G}}$

• already seen some models (Markov Chain Model)

• *general idea*: combine linguistic (symbolic) and statistical techniques in constructing such models

- Shyam Kapur and Robin Clark, *The Automatic Construction of a Symbolic Parser via Statistical Techniques*, in "The balancing act: combining symbolic and statistical approaches to language", MIT Press, 1996, pp. 95–117

• the process of setting values of syntactic parameters also involves reorganizing the grammar to reflect changed parameter value (more linguistic input)

• self-modification process (realistic model of language acquisition/evolution)

• most commonly used learning algorithm (see previous lectures) moves one step in parameter space triggered by failure to parse an incoming sentence

• inefficient: basically amounts to a random walk in parameter space

• *different idea*: next step choice uses previously built structure (incrementally build the grammar, modifying it when some parameter needs to be reset)

- focus on a set of syntactic parameters

1. Relative order of specifier and head (place of determiner relative to noun, position of VP-modifying adverbs)
2. Relative order of head and complement (VO versus OV; prepositions versus postpositions)
3. Scrambling: (some amount of) free word order allowed
4. Relative order of negative markers and verbs (more than one parameter: English has not after first tensed auxiliary, French wraps around verb: *ne ... pas*, etc.)
5. Root word order changes: certain word order changes allowed in root clauses but not in embedded clauses (eg inversion in root questions in English)

6. Rightward dislocation (as in: *That this happens amazes me*)
7. Wh-movement: location of wh-questions in phrase (English only one in first position, French as English or *in situ*, Polish several wh-questions stacked at beginning)
8. Exceptional case marking, structural case marking: allows for structures like $V_{[+tense]}NPVP_{[-tense]}$ tensed verb, noun phrase, verb phrase headed by infinitive verb
9. Raising and Control: distinguishes raising and control verbs (eg *they seem to be trying*: seem is a raising-to-subject verb, takes a semantic argument that belongs to an embedded predicate; or *he proved them to be wrong*: prove is raising-to-object verb; control verbs: *he stopped laughing*, *they told me to go there*,...)
10. Long and short-distance anaphora: short-distance anaphor *himself* corefers to NP within same local domain; other languages have long distance...

• in Principles and Parameters theory trigger data (cues) force learner to set certain particular parameters

• where do *statistical properties* of the input text enter in parameter setting?

• Example: in English sentences can have *John thinks that Mary likes him*, where "him" is a local anaphor (for John), or sentences like *Mary likes him*, where "him" is not co-referential to anything else in the sentence $\Rightarrow$ by statistical (frequency) of occurrences "him" is not always an anaphor. (This will avoid erroneously setting Long-distance anaphor parameter for English; unlike "sig" in Icelandic that can only be used as anaphor, long or short distance)

• Idea: a model of parameter setting should involve statistical analysis of the input text

Parameter Setting Model

- Space with N binary parameters
- Random subdivision of parameters into m groups: $\mathcal{P}_1, \ldots, \mathcal{P}_m$
- first set all parameters in first group $\mathcal{P}_1$:
  1. no parameter is set at the start
  2. both values $\pm$ or each $\Pi_i \in \mathcal{P}_1$ are "competing"
  3. for each Π_i a pair of *hypotheses* $H^i_\pm$
  4. these hypotheses are tested on input evidence
  5. if H^i_- fails or H^i_+ succeeds set $\Pi_i = +$, else $\Pi_i = -$
- continue with $\mathcal{P}_2, \ldots, \mathcal{P}_m$

Window sizes

- for hypotheses testing, suitable window sizes during which algorithm is sensitive to occurrence/non-occurrence; failure to occur within specified window taken as negative evidence

- Example
  1. H^i_+: expect not to observe phenomena from a fixed set O^i_- supporting $\Pi_i = -$
  2. H^i_-: expect not to observe phenomena from a fixed set O^i_+ supporting $\Pi_i = +$
- testing H^i_+: two small numbers w_i, k_i
  1. input of sentences of size w_i: record occurrences of phenomena in O^i_-
  2. repeat this construction of window of size w_i for k_i times: fraction c_i of times that phenomena in O^i_- occurred at least once
  3. hypothesis H^i_+ succeeds if $c_i/k_i < 1/2$

• sets $O^i_{\pm}$ have to be such that parser is always capable of analyzing the input for occurrences

• Note: with this method some parameters get set quicker than others (those parameters that are expressed more frequently)

• Word order parameters, for example, are expressed in all sentences: first ones to be set

• but for example have languages like German that are SOV but with V2 parameter moving verb in second position in root clauses (making some sentences look SVO)

• know from previous discussion of Gibson–Wexler algorithm that the parameter space for these word order plus V2 parameters has local maxima problem

• what happens to V2 parameter setting in this model? Can it avoid the problem?

Word order and V2 parameter

- Entropy $S(X) = -\sum_{X=x} p(x) \log p(x)$ or random variable X
- Conditional Entropy

$$S(X \mid Y) = -\sum_{X=x, Y=y} p(x,y) \log p(x \mid y) = \sum_{X=x, Y=y} p(x,y) \log \frac{p(y)}{p(x,y)}$$

how much better first variable can be predicted when second known

- pin down word order by analyzing entropy of positions in the neighborhood of verbs
- observation: in a V2 language more entropy to the left of verb than to the right (position to the left is less predictable)

- in input text consider data (v, d, w) with v one of 20 most frequent verbs, d a position either to the left or to the right of v and w the word that occurs in that position
- then procedure for setting $V2$ parameter
  - Compute conditional entropies $H(W \mid V, D)$
  - if $H(W \mid V, D = \text{left}) > H(W \mid V, D = \text{right})$ set $V2 = +$
  - otherwise set $V2 = -$
- correct result obtained when algorithm testes on 9 languages
- What hypothesis $H_{\pm}^{V2}$ does this procedure correspond to?
- simply use $H_{+}^{V2} =$ expect not to observe lower entropy on the left of verbs
- window size used 300 sentences and 10 repetitions

The conditional entropy results

	$H(W \mid V, D = \text{left})$	$H(W \mid V, D = \text{right})$
English	4.22	4.26
French	3.91	5.09
Italian	4.91	5.33
Polish	4.09	5.78
Tamil	4.01	5.04
Turkish	3.69	4.91
Dutch	4.84	3.61
Danish	4.42	4.24
German	5.55	4.97

Another application of the same algorithm: Clitic Pronouns

• *Clitic Pronouns* ($\kappa\lambda\iota\tau\iota\kappa o\varsigma$ = inflexional): syntactically independent but phonologically associated to another word

• Example: in French *me, te* (object clitic), *je, tu* (subject clitic), *moi, toi* (non-clitic, free standing), *nous, vous* (ambiguous)

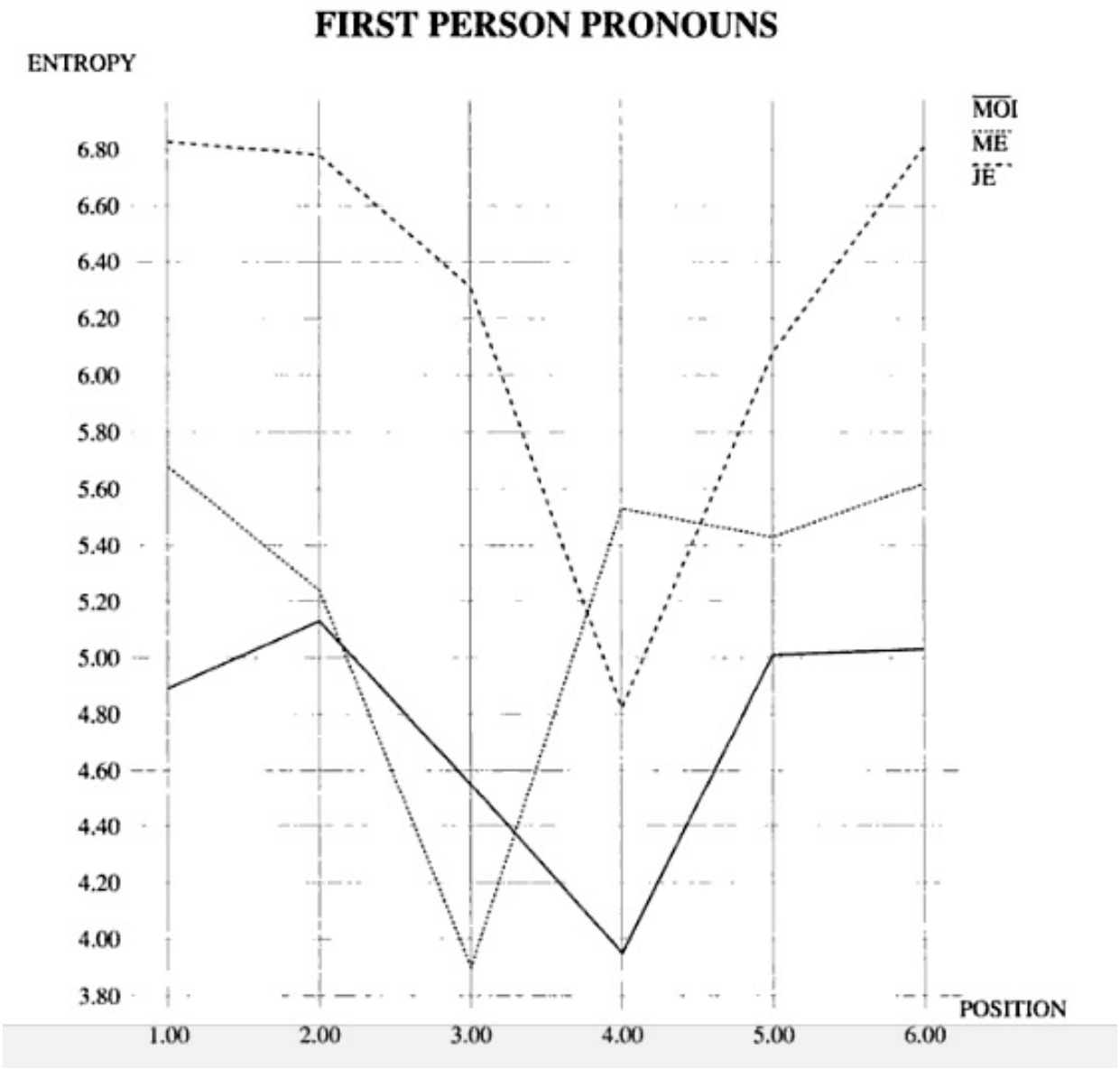
• Automatic identification and classification of clitic pronouns

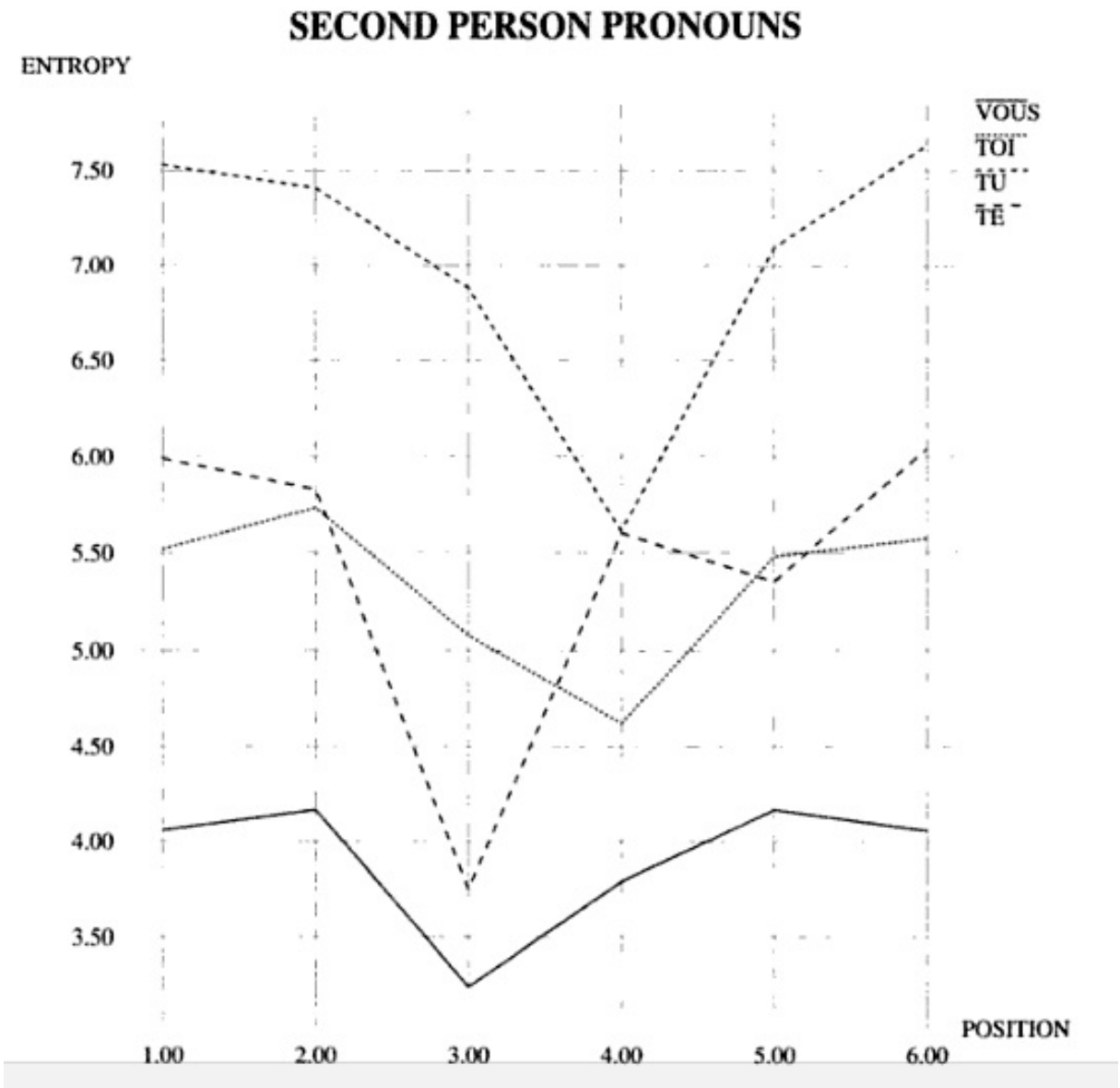
• Related to correctly setting syntactic parameters for syntax of pronominals

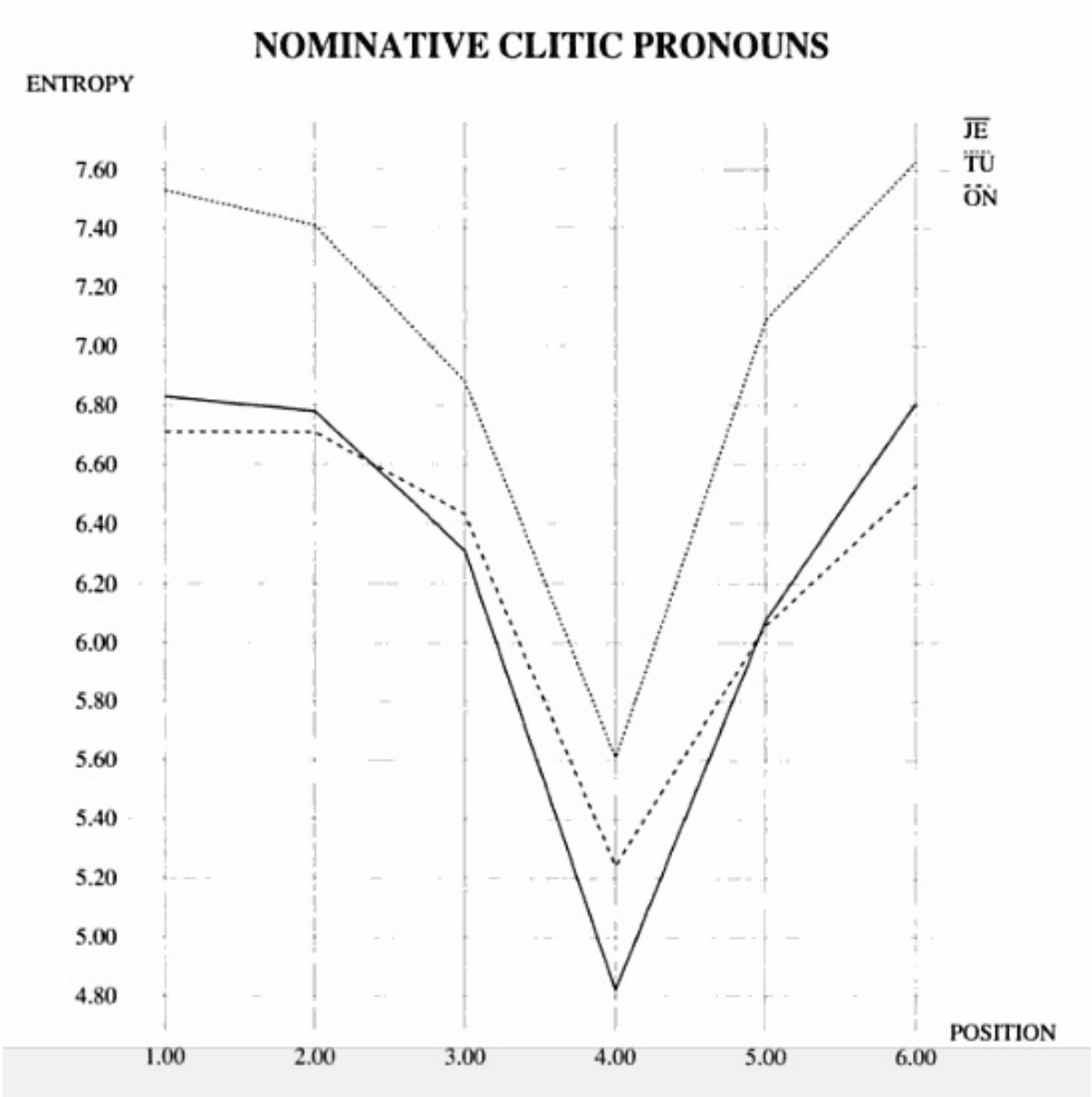
• also use method based on *entropies of positions*

• algorithm computes *entropy profiles* three positions to the left and to the right of each pronoun $H(W \mid P = p)$

• cluster together pronouns that have similar entropy profiles: find this gives the correct syntactic grouping


FIRST PERSON PRONOUNS
ENTROPY
MOI
ME
JE
6.80
6.60
6.40
6.20
6.00
5.80
5.60
5.40
5.20
5.00
4.80
4.60
4.40
4.20
4.00
3.80
POSITION
1.00
2.00
3.00
4.00
5.00
6.00


SECOND PERSON PRONOUNS
ENTROPY
VOUS
TOI
TU
TE
7.50
7.00
6.50
6.00
5.50
5.00
4.50
4.00
3.50
POSITION
1.00
2.00
3.00
4.00
5.00
6.00


NOMINATIVE CLITIC PRONOUNS
ENTROPY
7.60
7.40
7.20
7.00
6.80
6.60
6.40
6.20
6.00
5.80
5.60
5.40
5.20
5.00
4.80
JE
TU
ON
POSITION
1.00
2.00
3.00
4.00
5.00
6.00

Some Concluding Remarks on Linguistics and Statistics

• Statistical methods in Computational Linguistics (especially Hidden Markov Models) have come to play a prominent role in recent years

• Statistical methods are also the basis for Natural Language Processing and Machine Translation techniques

• Theoretical Linguistics, on the other hand, is focused on understanding syntactic structures of languages, generative grammars, models of how the human mind acquires and processes language, and of how languages change and evolve in time

• Is there a Linguistics versus Statistics tension in the field?

• the "sociological" answer is *yes*, but the scientific answer should be *no*

Language Learning

• if only a *discrete* setting where some parameters are switched on and off would expect abrupt changes in a learner's acquisition process

• in experimental observation of children learning a language, grammar changes happen as changes in frequencies of use of different possibilities, over a stretch of time

• more consistent with the idea that the language learner is dealing with *probabilistic grammars* and "trying out rules" for a time

• a probabilistic grammar is a combination of a theoretical linguistic substrate (context-free grammars, tree-adjoining grammars, etc.) with a probabilistic datum associated to the production rules

• Discrete (algebraic) structures + (continuous) probabilities

Language Evolution

• both language change by dialect diversification and by interaction with other languages require frequencies/probabilities describing spreading of change and proportions of different language speakers among a population

• even assuming every individual adult speaker uses a fixed (non-probabilistic) grammar, probabilistic methods are intrinsic in the description of language change over a population

• linguistics theories formulated before computational methods (like the wave theory model of language change) and already naturally compatible with the probabilistic approach

• even setting of syntactic parameters in a given language can be seen as probabilistic (see the head-initial/head-final subdivision)

Parsing Ambiguities

• even completely unremarkable and seemingly unambiguous sentences can have lots of different parsings (just most of them would be considered very unusual)

• a lot of these (grammatical) parsings would be accepted by a grammar but not in agreement with human perception

• Example: English sentence *The cows are grazing in the grass* seems completely unambiguous, but *are* is also a noun, a measure of size: *a hectare is a hundred ares*... it would be grammatical, but very unlikely ... *probabilistically suppressed*

Natural versus Computer Languages

• separating out the functioning of natural languages into grammar and compiler (that uses grammar to produce and parse sentences) is convenient for theoretical understanding, but does not correspond to an actual distinction (e.g. to different structures in the human mind)

• analogy with computer languages: grammars (formal languages) also work for describing computer languages... they provide an abstract description of the structure of the computation being performed

• ... but in the actual compiler operations grammar and parsing work simultaneously and not as separate entities

• grammar is an abstract idealization of linguistic data, which has the power of simplicity (like algebraic structures)

Autonomy of Syntax

• Chomsky's famous example: sentences
- *revolutionary new ideas appear infrequently*
- *colorless green ideas sleep furiously*

• syntactically equally well structured (same structure); the second is grammatical but would be discarded by any statistical analysis

• syntax is in itself an interesting (algebraic) structure, but it is autonomous only as long as it is not interfaced with semantics

• syntax as algebraic grammar is one (very important) aspect of linguistics, but not the only one

The Goals of Linguistics

• Describe language: how it is produced, comprehended, learned, and how it evolves over time

• Goal of Generative Linguistics: produce a model (grammar) that generates sentences in a given language $\mathcal{L}$ that reflect the structure as recognized by a human speaker of language $\mathcal{L}$

• Turing Test for Linguistics: a model passes the test if the sentences it generates are recognized as grammatical and "natural" by a human speaker... grammatical is not enough, "natural" is a matter of degrees... both algebraic and probabilistic aspects contribute (test cannot be passed by an unweighted grammar)

Criticism of Markov Models

• already in his early paper "Three models for the description of Language", Chomsky criticized Shannon's n-gram models and statistical approximations to English

• main point of criticism: it is impossible to choose n and ϵ so that $P_n(s) > \epsilon$ iff sentence s is grammatical

• this already pointed out by Shannon: at order n approximation there will be some more elaborate dependences affecting grammaticality that approximation does not capture

• ... but inadequacy of Markov model lies in their being finite-state automata not in being statistical: probabilistic context-free grammars or probabilistic tree-adjoining grammars are more sophisticated statistical models than Shannon's n-grams

- Reference for these conclusive remarks:
  - Steven Abney, *Statistical Methods and Linguistics*, in "The balancing act: combining symbolic and statistical approaches to language", MIT Press, 1996, pp. 1–26.