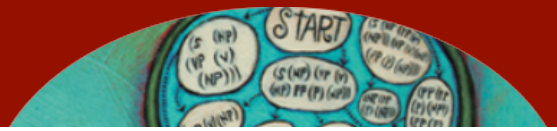
The Two Cultures: Linguistics in the Age of Data

James Pustejovsky
Brandeis University

January 20, 2017
CS 114

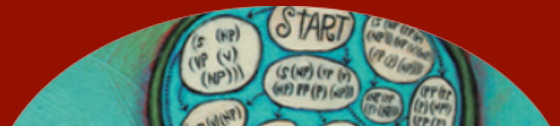


Talk Outline

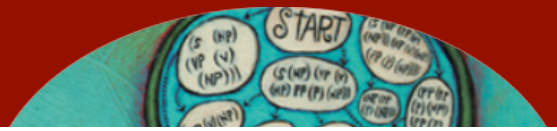
- How Linguists collect data

  Well-formedness in Linguistics
- A little History
- A History of Data –

  Corpora and Structural Discovery
- Learning with Big Data
- Revisiting Method and Theory

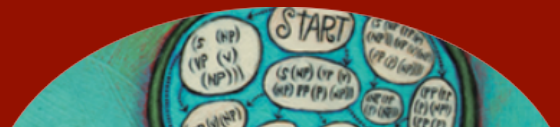


How Linguists Collect Data



Investigate hypotheses by consulting native speakers' intuitions

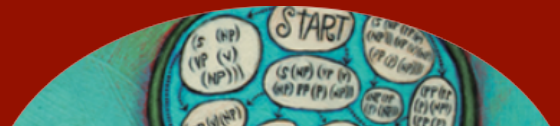
- Most linguists assume that people can distinguish strings of words that are sentences of their language from strings of words that are not sentences of their language.
  - So imagine that you are a machine or a classifier that takes a sentence as input, and returns "accept" or "reject" as output.



Native speakers as automata that accept and reject strings of words.

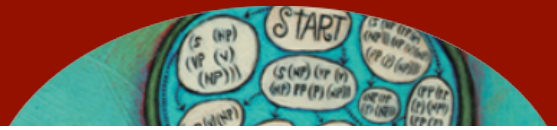
✓ The student read a book.

* Student the a read book.



Grammaticality

- A string of words that you recognize as a sentence in your native language is *grammatical*.
- A string of words that you do not recognize as a sentence in your native language is *ungrammatical*.
- When you decide whether a sentence is grammatical or ungrammatical, this is called giving a *grammaticality judgment*.
- Ungrammatical sentences are preceded by an asterisk or star (*). Sometimes they are called *starred sentences*.
- If native speakers can't decide whether the sentence is grammatical or ungrammatical, it is preceded by a combination of stars and question marks.



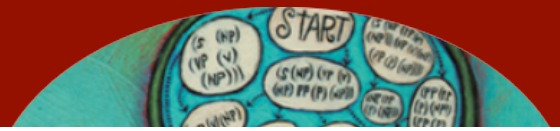
Grammatical ≠ meaningful

- It is unlikely that Pat will succeed.
- It is improbable that Pat will succeed.
- Pat is unlikely to succeed.
- *Pat is improbable to succeed.

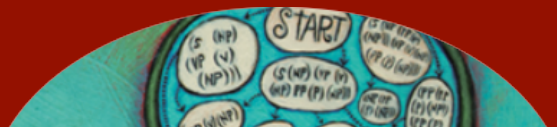
  This could be meaningful, but most people consider it to be ungrammatical.

- They saw Pat with Chris.
- They saw Pat and Chris.
- Who did they see Pat with?
- *Who did they see Pat and?

  Again, this could be meaningful, but it is ungrammatical.



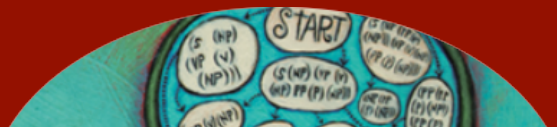
A Little History



Some Brief History

Foundational Insights: 1940s and 1950s

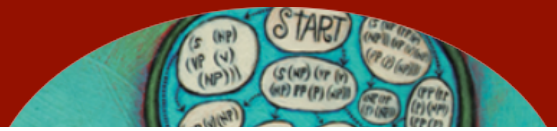
- Two foundational paradigms:
  - the automaton and
  - probabilistic or information-theoretic models
- Turing's work led first to the McCulloch-Pitts neuron (McCulloch and Pitts, 1943),
  - a simplified model of the neuron as a kind of computing element that could be described in terms of propositional logic,
- And then to the work of Kleene (1951) and (1956) on
  - finite automata and regular expressions.
- Shannon (1948) applied probabilistic models of discrete Markov processes to automata for language. (*continued*)



Some Brief History

Foundational Insights: 1940s and 1950s

- Chomsky (1956), drawing the idea of a finite state Markov process from Shannon's work, first considered finite-state machines as a way to characterize a grammar, and defined a finite-state language as a language generated by a finite-state grammar.
- These early models led to the field of formal language theory, which used algebra and set theory to define formal languages as sequences of symbols.
  - This includes the context-free grammar, first defined by Chomsky (1956) for natural languages but independently discovered by Backus (1959) and Naur et al. (1960) in their descriptions of the ALGOL programming language.



Some Brief History

Foundational Insights: 1940s and 1950s

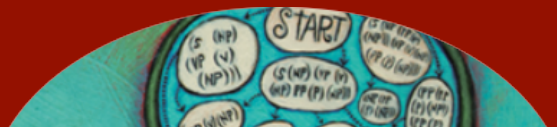
The second foundational insight of this period was the development of probabilistic algorithms for speech and language processing, which dates to Shannon's other contribution:

the metaphor of the **noisy channel** and **decoding** for the transmission of language through media like communication channels and speech acoustics.

Shannon also borrowed the concept of **entropy** from thermodynamics as a way of measuring the information capacity of a channel, or the information content of a language, and performed the first measure of the entropy of English using probabilistic techniques.

It was also during this early period that the sound spectrograph was developed (Koenig et al., 1946), and foundational research was done in instrumental phonetics that laid the groundwork for later work in speech recognition.

This led to the first machine speech recognizers in the early 1950s.



Some Brief History

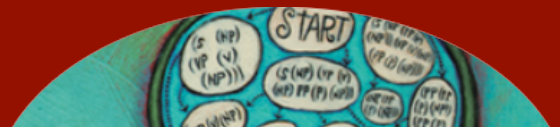
The Two Camps: 1957–1970

By the end of the 1950s and the early 1960s, SLP had split very cleanly into two paradigms: *symbolic* and *stochastic*.

The symbolic paradigm took off from two lines of research.

The **first** was the work of Chomsky and others on formal language theory and generative syntax throughout the late 1950s and early to mid 1960s, and the work of many linguistics and computer scientists on parsing algorithms, initially top-down and bottom-up and then via dynamic programming.

One of the earliest complete parsing systems was Zelig Harris's Transformations and Discourse Analysis Project (TDAP), which was implemented between June 1958 and July 1959 at the University of Pennsylvania (Harris, 1962).



Some Brief History

The Two Camps: 1957–1970

The second line of research was the new field of artificial intelligence.

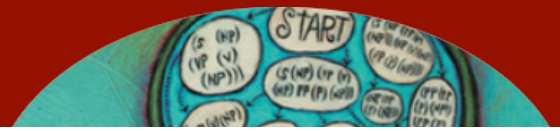
In the summer of 1956 John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester brought together a group of researchers for a two-month workshop on what they decided to call artificial intelligence (AI).

Although AI always included a minority of researchers focusing on stochastic and statistical algorithms (include probabilistic models and neural nets), the **major focus of the new field was the work on reasoning and logic** typified by Newell and Simon's work on the Logic Theorist and the General Problem Solver.

At this point **early natural language understanding systems were built**.

These were simple systems that worked in single domains mainly by a combination of pattern matching and keyword search with simple heuristics for reasoning and question-answering.

By the late 1960s more formal logical systems were developed.



Some Brief History

The Two Camps: 1957–1970

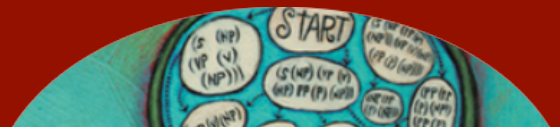
The stochastic paradigm took hold mainly in departments of statistics and of electrical engineering.

By the late 1950s the Bayesian method was beginning to be applied to the **problem of optical character recognition**.

Bledsoe and Browning (1959) built a Bayesian system for text-recognition that used a large dictionary and computed the likelihood of each observed letter sequence given each word in the dictionary by multiplying the likelihoods for each letter.

Mosteller and Wallace (1964) applied Bayesian methods to the problem of authorship attribution on *The Federalist* papers.

The 1960s also saw the rise of the first serious testable psychological models of human language processing based on transformational grammar, as well as the first on-line corpora: the Brown corpus of American English, a 1 million word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc.), which was assembled at Brown University in 1963–64 (Kučera and Francis, 1967; Francis, 1979; Francis and Kučera, 1982), andWilliam S. Y.Wang's 1967 DOC (Dictionary on Computer), an on-line Chinese dialect dictionary.



Some Brief History

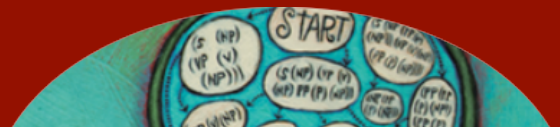
Four Paradigms: 1970–1983

The next period saw an explosion in research in SLP and the development of a number of **research paradigms** that still dominate the field.

The **stochastic** paradigm played a huge role in the development of *speech recognition* algorithms in this period,

particularly the use of the *Hidden Markov Model* and the metaphors of the noisy channel and decoding, developed independently by Jelinek, Bahl, Mercer, and colleagues at IBM's Thomas J. Watson Research Center, and by Baker at Carnegie Mellon University, who was influenced by the work of Baum and colleagues at the Institute for Defense Analyses in Princeton.

AT&T's Bell Laboratories was also a center for work on speech recognition and synthesis; see Rabiner and Juang (1993) for descriptions of the wide range of this work.



Some Brief History

Four Paradigms: 1970–1983

The **natural language understanding** field took off during this period,

beginning with Terry Winograd's SHRDLU system, which simulated a robot embedded in a world of toy blocks (Winograd, 1972a).

The program was able to accept natural language text commands *(Move the red block on top of the smaller green one)* of a hitherto unseen complexity and sophistication.

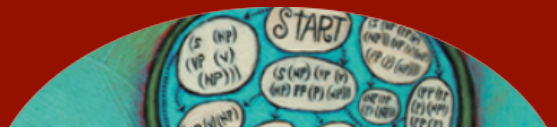
His system was also the first to attempt to build an extensive (for the time) grammar of English, based on Halliday's systemic grammar.

Winograd's model made it clear that the problem of parsing was well-enough understood to begin to focus on semantics and discourse models.

Roger Schank and his colleagues and students (in what was often referred to as the *Yale School*) built a series of language understanding programs that focused on human conceptual knowledge such as scripts, plans and goals, and human memory organization (Schank and Albelson, 1977; Schank and Riesbeck, 1981; Cullingford, 1981; Wilensky, 1983; Lehnert, 1977).

This work often used network-based semantics (Quillian, 1968; Norman and Rumelhart, 1975; Schank, 1972; Wilks, 1975c, 1975b; Kintsch, 1974) and began to incorporate Fillmore's notion of *case roles* (Fillmore, 1968) into their representations (Simmons, 1973).

The logic-based and natural-language understanding paradigms were unified on systems that used predicate logic as a semantic representation, such as the LUNAR question-answering system (Woods, 1967, 1973).



Some Brief History

Empiricism and Finite State Models Redux: 1983–1993

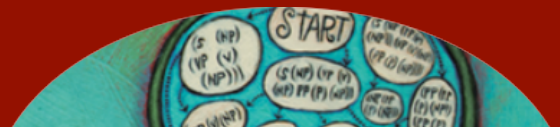
This next decade saw the return of two classes of models which had lost popularity in the late 1950s and early 1960s, partially due to theoretical arguments against them such as Chomsky's influential review of Skinner's *Verbal Behavior* (Chomsky, 1959b).

The first class was finite-state models, which began to receive attention again after work on finite-state phonology and morphology by Kaplan and Kay (1981) and finite-state models of syntax by Church (1980).

The second trend in this period was what has been called the "return of empiricism"; most notably here was the rise of probabilistic models throughout speech and language processing, influenced strongly by the work at the IBM Thomas J. Watson Research Center on probabilistic models of speech recognition.

These probabilistic methods and other such data-driven approaches spread into part-of-speech tagging, parsing and attachment ambiguities, and connectionist approaches from speech recognition to semantics.

This period also saw considerable work on natural language generation.



Some Brief History

The Field Comes Together: 1994–2015

By the last five years of the millennium it was clear that the field was vastly changing.

First, probabilistic and data-driven models had become quite standard throughout natural language processing.

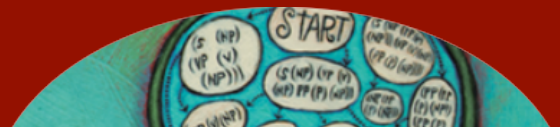
Algorithms for parsing, part-of-speech tagging, reference resolution, and discourse processing all began to incorporate probabilities, and employ evaluation methodologies borrowed from speech recognition and information retrieval.

Second, the increases in the speed and memory of computers had allowed commercial exploitation of a number of subareas of speech and language processing, in particular

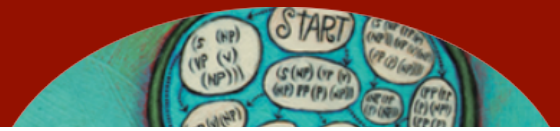
speech recognition and spelling and grammar checking.

Speech and language processing algorithms began to be applied to Augmentative and Alternative Communication (AAC).

Finally, the rise of the Web emphasized the need for language-based information retrieval and information extraction.



A History of Data



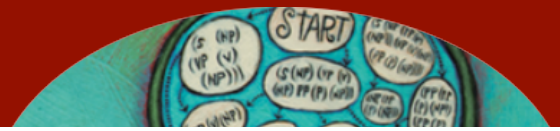
Rationalism and Empiricism

- Rationalism:

  the source of knowledge is reason

- Empiricism:

  the source of knowledge is data



Rationalists vs. empiricists

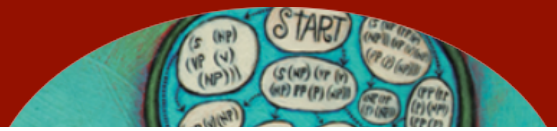
Chomsky emphasized "creativity" of language, as manifested in recursive generative rules

S → NP VP, VP → VP NP

Empiricists emphasize common language patterns (e.g. collocations) and predictability of language

Warren Weaver, pioneer of MT (1949)

about half of the letters or words we choose in writing or speaking (although we are not ordinarily aware of it) are really controlled by the statistical structure of the language.



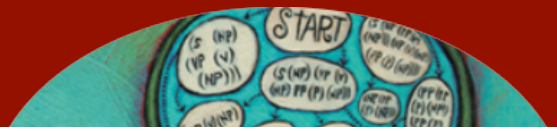
Contrastive viewpoints

Chomsky (1957):

> *I think that we are forced to conclude that grammar is autonomous and independent of meaning*

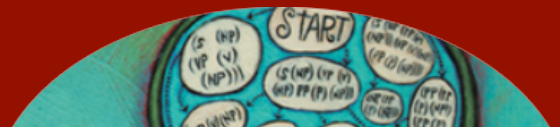
Corpus linguist John Sinclair (1991):

> *it is folly to decouple lexis and syntax, or either of those and semantics. The realization of meaning is far more explicit than is suggested by abstract grammars. The model of a highly generalized formal syntax, with slots into which fall neat lists of words, is suitable only in rare uses and specialized texts. By far the majority of text is made of the occurrence of common words in co˜mmon patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text.*



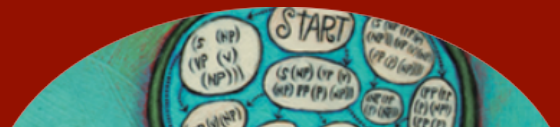
Rationalist view of linguistic data

- Language is something in people's minds – a set of rules and principles that allows them to make grammaticality judgments and produce and understand sentences that they have never heard before
  - i-language or internal language
- We study i-language asking people to give grammaticality judgments.
- A corpus (a collection of texts or speech) is e-language, or external language. It is not the object of study.



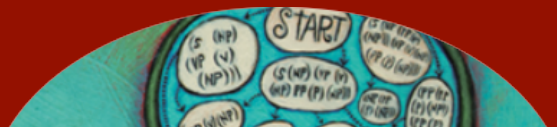
Empiricist view of linguistic data

- Corpora are the objects of study.
- We study language by examining patterns in corpora (collections of texts or speech).



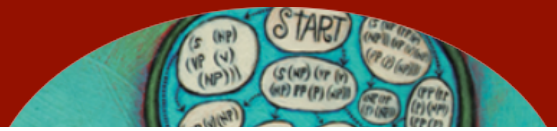
Strong points of rationalism

- Infinite, creative capacity: People can produce and understand sentences that have never been uttered before. They are not repeating memorized patterns, but applying productive rules.
- Leads people to wonder about things that don't exist in a corpus: *Who did you see Pat and?
- Probability is not grammaticality: grammatical sentences may have very low probability.
- Probability reflects facts about the world, but grammaticality is independent of context.
  - Clyde is an African elephant.
  - Clyde is a pink elephant



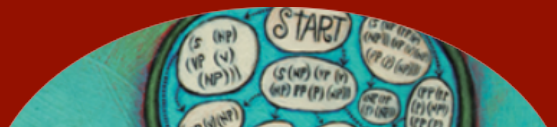
Strong points of empiricism

- Frequency of occurrence in a corpus is easier to measure reliably than a grammaticality judgment.
- Many ungrammatical sentences turn out to be acceptable in the right context.
  - Identifying the right context turns out to be an interesting question that does not arise in the rationalist approach.
  - I gave her the book.
  - I gave the book to her.



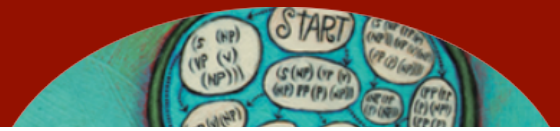
Structuralist Tradition

- Bloomfield, others, 1940s – 1950s
  - Language can be explained in probabilistic, behaviorist terms
  - Languages are diverse systems learned from the environment
  - The aim was to describe the diversity of linguistic behavior; analyze linguistic structure in formal terms – producing formal descriptions of grammar (including phonetics, morphology, syntax, etc.)



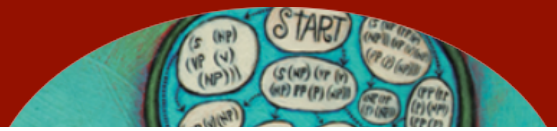
Early Computational Models

- Empirical and statistical methods were popular the 1950s
- Shannon's information-theoretic approach to language
  - *All of us were convinced that speech, in English or any other language, was a Markov process. From this to the conviction that ... the set of all English sentences can be generated by a Markov source was only a small step.* (Bar-Hillel, 1975)
- Early machine translation efforts of 1950s and 1960s



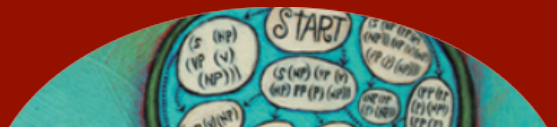
Chomsky vs. Corpus Linguistics

- Popularity of empirical and statistical methods faded in the 1960s under the 'cognitive revolution'
- Chomsky's mentalistic, generative approach to language revolutionized linguistics and cognitive science in the 20th century
- Influential Events
  - Chomsky's *Syntactic Structures* (1957) and *Aspects of the Theory of Syntax* (1965)
  - Chomsky & Miller's critiques of statistical language models
  - Chomsky's critique of Skinner's *Verbal Behavior*



Chomsky

I had no personal interest in the experimental studies and technological advances. [...] As for machine translation and related enterprises, this seemed to me pointless, as well as probably quite hopeless. [...] I could not fail to be aware of the ferment and excitement [in the early 1950s]. But I felt myself no part of it. (Chomsky, 1975)

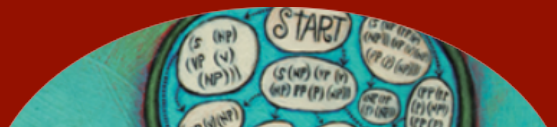


Syntactic Structures (Chomsky 1957)

From now on I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements.

The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical sequences.

On what basis do we actually go about separating grammatical sequences from ungrammatical sequences?



Syntactic Structures (Chomsky 1957)

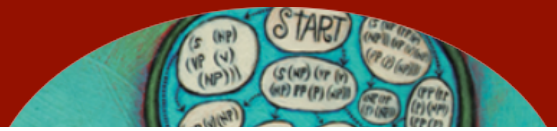
First, it is obvious that the set of grammatical sentences cannot be identified with any. . . finite and somewhat accidental corpus of observed utterances. . .

Second, the notion "grammatical" cannot be identified with "meaningful" or "significant" in any semantic sense.

Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.

(1) Colorless green ideas sleep furiously.

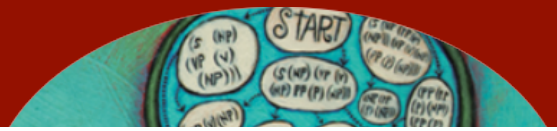
(2) Furiously sleep ideas green colorless.



Syntactic Structures (Chomsky 1957)

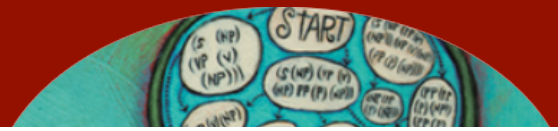
Third, the notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English." It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse.

Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.



Syntactic Structures (Chomsky 1957)

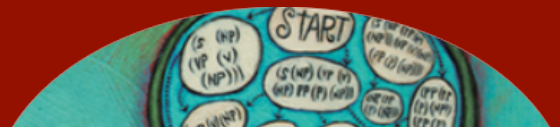
Evidently, one's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like. . . I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure.



Carnap's *Logical Structure of Language* (1934)

"Pirots karulize elatically."

Umformungsregeln von gleicher Art sind, und daß beide formal gefaßt werden können. Daß z.B. die Wortreihe "Piroten karulieren elatisch" ein Satz ist, kann, wenn eine geeignete Regel aufgestellt ist, festgestellt werden, sofern nur bekannt ist, daß "Piroten" ein *Substantivum* (im Plural), "karulieren" ein *Verbum* (in der 3. Pers. Plur, Ind.) und "elatisch" ein *Adverbium* ist (was übrigens in einer gut konstruierten Wortsprache, wie z.B. in Esperanto, aus der *Form* der Wörter zu ersehen sein würde); die *Bedeutung* der Wörter braucht hierfür nicht bekannt zu sein. Ferner kann, wenn eine geeignete Regel aufgestellt ist, aus dem genannten Satz und dem Satz "A ist ein Pirot" der Satz "A karuliert elatisch" erschlossen werden, wenn nur wieder die Wortarten der einzelnen Wörter bekannt sind; auch hierfür braucht ihre Bedeutung und der Sinn der drei Sätze nicht bekannt zu sein.

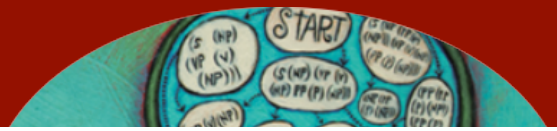


Chomsky's Claim

(1) Colorless green ideas sleep furiously.

(2) Furiously sleep ideas green colorless.

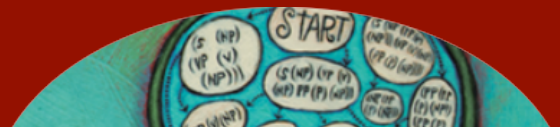
- Neither of these sentences has ever appeared before in human discourse. None of the substrings has either.
- Hence the probabilities P(W_1) and P(W_2) cannot be relevant to a native English speaker's knowledge of the difference between these sentences.



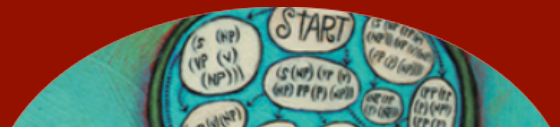
Chomsky's Claim

. . . in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English.

This claim is false—all modern statistical models of language can assign probabilities to previously-unseen utterances.



Learning with Data



Refuting Chomsky

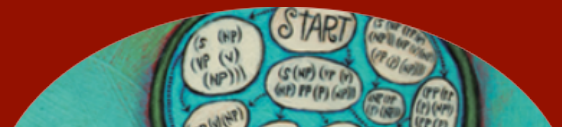
Language users *must* be able to generalize beyond their input.

Statistical inference ***is*** the study of such generalization

Pereira (2000) used a *class-based* bigram model to model (1) and (2) before:

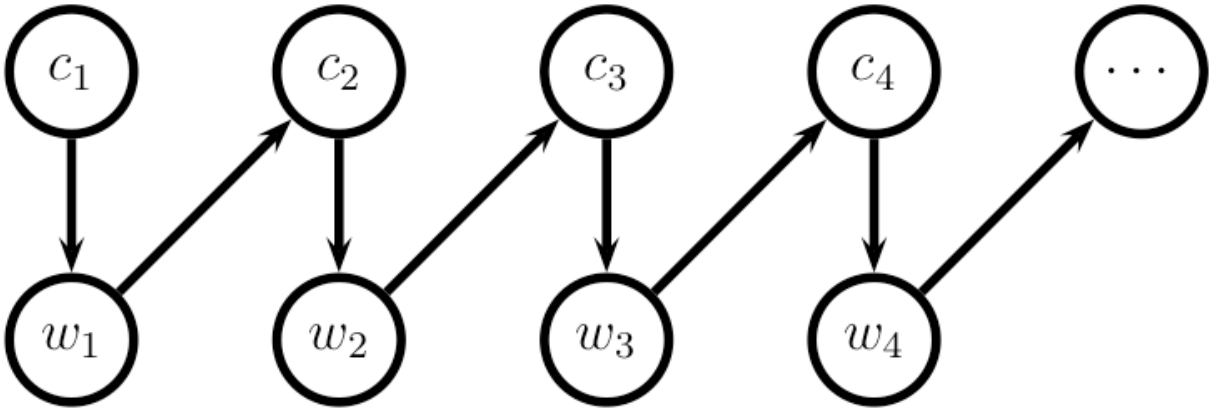
$$\frac{P(\text{Colorless green ideas sleep furiously})}{P(\text{Furiously sleep ideas green colorless})} \approx 2 \times 10^5$$

Maybe probabilities aren't such a useless model of language knowledge after all.



Language and probability

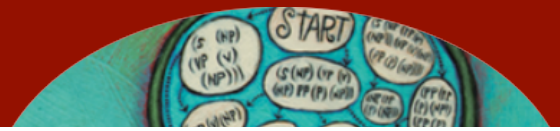
A *class-based* model looks something like this:



The c_i are ***unseen variables*** that have to be introduced into the model, either through:

- Annotation
- Unsupervised learning

For now, focus on the basic problem of the ***language model:*** P(*W*)



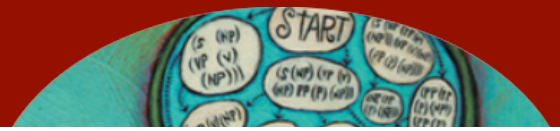
Creating a language model

- Language models are core component of many applications where task is to ***recover an utterance*** from an input. This is done through ***Bayes' rule:***

$$\mathrm{P}(\text{utterance}|\text{input}) = \frac{\mathrm{P}(\text{input}|\text{utterance})\mathrm{P}(\text{utterance})}{\mathrm{P}(\text{input})}$$

$$\propto \mathrm{P}(\text{input}|\text{utterance})\mathrm{P}(\text{utterance})$$

Language Model



We want to predict a sentence given an input sequence:

$$s^* = \arg\max_s P(s \mid a)$$

The noisy channel approach:

Build a generative model of production (encoding)

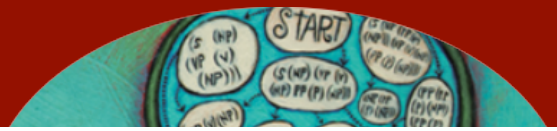
$$P(a, s) = P(s)\,P(a \mid s)$$

To decode, we use Bayes' rule to write

$$\begin{aligned} s^* &= \arg\max_s P(s \mid a) \\ &= \arg\max_s P(s)P(a \mid s)/P(a) \\ &= \arg\max_s P(s)P(a \mid s) \end{aligned}$$

Now, we have to find a sentence maximizing this product

Why is this progress? This predicts what to expect as input.



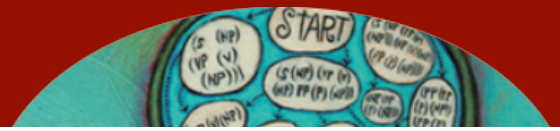
Corpora for linguistic research

It is quite typical for researchers to use *any collection of texts* for linguistic analysis.

- Often proceed opportunistically: whatever data comes in handy is used.

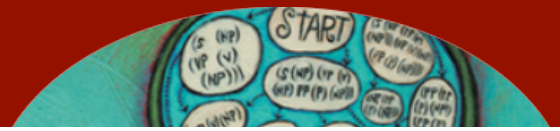
However, the term *corpus* usually implies the following characteristics:

- sampling/representativeness
- finite size
- machine-readable form
- a standard reference
- (time-bound)

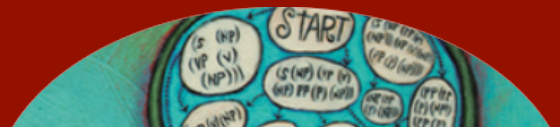


Corpora

Name of corpus	Year published	Size	Collection contents
British National Corpus (BNC)	1991–1994	100 million words	Cross section of British English, spoken and written
American National Corpus (ANC)	2003	22 million words	Spoken and written texts
Corpus of Contemporary American English (COCA)	2008	425 million words	Spoken, fiction, popular magazine, and academic texts



Revisiting Method and Theory



Method in Linguistics

- Sapir, Bloomfield, Hockett, Wells (Structural analysis)

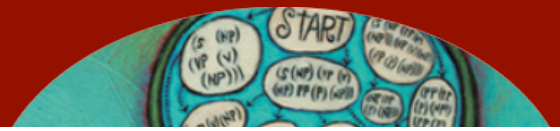
  Discovery procedure allows for the emergence of grammatical patterns and constructions in a dataset.

- Chomsky, Bar-Hillel (Transformational Grammars)

  Descriptive procedure allows for the generation of grammatical patterns.

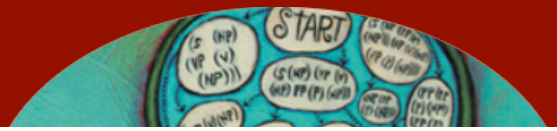
- Chomsky (Generative Grammar 1962- present)

  Explanatory model allows for the generation of best grammatical patterns.



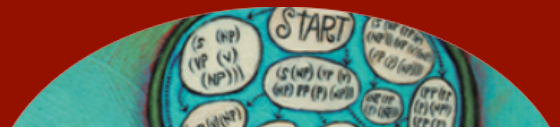
1950-1990 - The Absence of Data

- Chomsky liberated the field of linguistics in the 1950s
- Generation through recursive functions allows one to create your own corpus
- Experiment with new datasets that are not attested in actual “found data”



1990-2016 - The Absence of Theory

- Big Data and statistical modeling has largely dominated the fields of CL and AI, both theoretical and applied.
- Deep Learning seems positioned to obviate theory completely.
- This will not happen: machine learning and deep learning make theoretical assumptions in both the data preparation and feature selection and engineering phase of training.
- Theory is more relevant than ever before.



Conclusion

- Linguistics is now both a theoretical and experimental discipline
- The scope of observed data for language study and theorizing is richer and broader than ever.
- Linguistic Corpora and captured media datasets will enable contextualized and embodied interpretation of linguistic utterances
- This will enable the development of more expressive and broader theories of language and communication