Complexity and Character of Human Languages

Informatics 2A: Lecture 21

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

04 November 2011

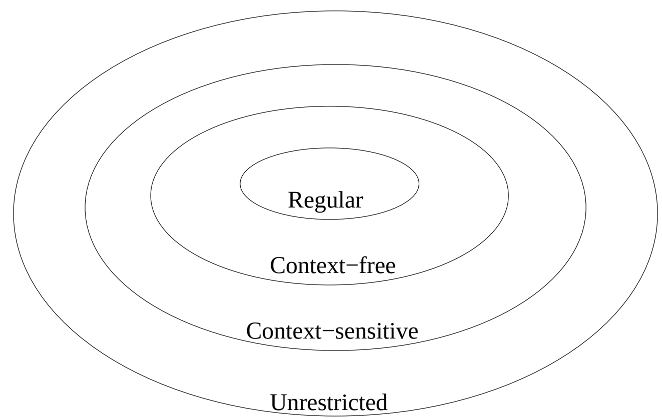
Reading: J&M. Chapter 16.3–16.4; Hauser, Chomsky, and Fitch (2002)

Review

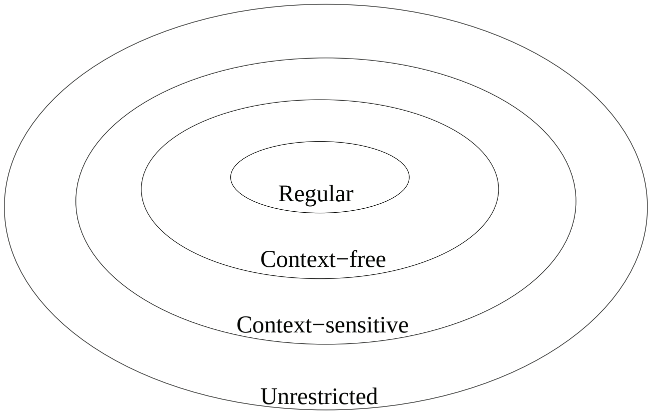
Chomsky Hierarchy: classifies languages on scale of complexity:

- Regular languages: those whose phrases can be ‘recognized’ by a finite state machine.
- Context-free languages: most programming languages, and many aspects of natural languages can be described at this level; the set of languages accepted by pushdown automata.
- Context-sensitive languages: equivalent with a linear bounded nondeterministic Turing machine, also called a linear bounded automaton.
- Unrestricted languages: *all* languages that can in principle be defined via mechanical rules.

Review

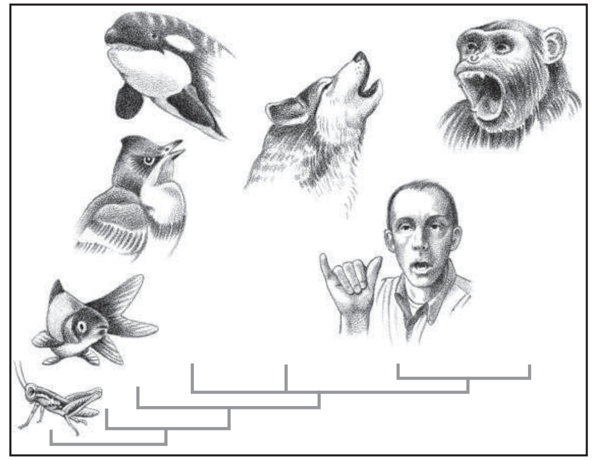


Review



Where do human languages fit within this complexity hierarchy?

The Faculty of Language



The "language faculty" has a broad sense and a narrow sense (Hauser, Chomsky, and Fitch 2002).

The Faculty of Language (Broad Sense)

Sensory-motor system

- for producing and perceiving linguistic communication
- spoken language: vocal track, auditory system
- sign language: gestural system, visual system
- written language: writing system, visual or tactile system

Conceptual-intentional system

- who to communicate with and what to communicate about
- generating mental states and attributing them to others;
- acquiring conceptual representations that are non-linguistic;
- referring to entities and events.

The Faculty of Language (Narrow Sense)

Abstract computational system

- one part of which is narrow syntax which generates representations internal to the mind/brain and maps them to:
- sensory-motor interface through phonological, gestural system;
- conceptual-intentional system through semantic (and pragmatic) systems.

A core property of narrow syntax is recursion: takes a fine set of elements and yields a potentially infinite array of discrete expressions.

Recursion

The potential infiniteness of the language faculty has been recognized by Galileo, Descartes, von Humboldt.

Discrete Infinity

- Sentences are built up by discrete units
- There are 6-word sentences, and 7-word sentences, but no 6.5 word sentences
- There is no longest sentence!
- There is no non-arbitrary upper bound to sentence length!

Mary thinks that John thinks that George thinks that Mary thinks that this course is boring!
I ate lunch and slept and watched tv and went to the bathroom and had a coffee and got dressed . . .

Strong and Weak Adequacy

Questions about the formal complexity of language are about the computational power of syntax, as represented by a grammar that's adequate for it.

A strongly adequate grammar

- generates all and only the strings of the language;
- assigns them the "right" structures — ones that support a correct representation of meaning.

A weakly adequate grammar

generates all and only the strings of a language but assigns them "wrong" structures.

Is Natural Language Regular?

It is generally agreed that natural languages are not regular!

Center-embedding

[The cat$_1$ likes tuna fish$_1$].
[The cat$_1$ [the dog$_2$ chased$_2$] likes tuna fish$_1$].
[The cat$_1$ [the dog$_2$ [the rat$_3$ bit$_3$] chased$_2$] likes tuna fish$_1$].

Idea of proof

(the+noun)^n (transitive verb)$^{n-1}$ likes tuna fish.
A = { the cat, the dog, the rat, the elephant, the kangaroo . . . }
B = { chased, bit, admired, ate, befriended . . . }
Intersect `/A* B* likes tuna fish/` with English
$L = x^n y^{n-1}$ *likes tuna fish*$, x \in A, y \in B$
Use pumping lemma to show L is not regular

Is Natural Language Context Free?

It doesn't look like it is context free either! Evidence comes from a Swiss German dialect and Bambara, a language spoken in Mali.

Crossing dependencies

omdat Wim$_1$ Jan$_2$ Henk$_3$ de kinderen$_4$ zag$_1$ helpen$_2$ leren$_3$ zwemmen$_4$
because Wim$_1$ Jan$_2$ Henk$_3$ the children$_4$ saw$_1$ help learn swim$_4$
because Wim saw Jan help Henk teach the children to learn to swim

| *zag* | depends on | *Wim* |, and | *helpen* | depends on | *Jan* |, etc.

Idea of Proof

Languages $\{xx|x \in \{a,b\}^*\}$ are not context-free.
Related $a^n b^m c^n d^m$ language also not context -free.
Swiss German crossing dependencies equivalent to $a^n b^m c^n d^m$

Is Natural Language Context Free?

It doesn't look like it is context free either! Evidence comes from a Swiss German dialect and Bambara, a language spoken in Mali.

Crossing dependencies

omdat Wim$_1$ Jan$_2$ Henk$_3$ de kinderen$_4$ zag$_1$ helpen$_2$ leren$_3$ zwemmen$_4$
because Wim$_1$ Jan$_2$ Henk$_3$ the children$_4$ saw$_1$ help learn swim$_4$
because Wim saw Jan help Henk teach the children to learn to swim

| *zag* | depends on | *Wim* |, and | *helpen* | depends on | *Jan* |, etc.

Idea of Proof

Languages $\{xx|x \in \{a,b\}^*\}$ are not context-free.
Related $a^n b^m c^n d^m$ language also not context -free.
Swiss German crossing dependencies equivalent to $a^n b^m c^n d^m$

Is Natural Language Context Free?

It doesn't look like it is context free either! Evidence comes from a Swiss German dialect and Bambara, a language spoken in Mali.

Crossing dependencies

omdat Wim$_1$ Jan$_2$ Henk$_3$ de kinderen$_4$ zag$_1$ helpen$_2$ leren$_3$ zwemmen$_4$
because Wim$_1$ Jan$_2$ Henk$_3$ the children$_4$ saw$_1$ help learn swim$_4$
because Wim saw Jan help Henk teach the children to learn to swim

| *zag* | depends on | *Wim* |, and | *helpen* | depends on | *Jan* |, etc.

Idea of Proof

Languages $\{xx|x \in \{a,b\}^*\}$ are not context-free.
Related $a^n b^m c^n d^m$ language also not context -free.
Swiss German crossing dependencies equivalent to $a^n b^m c^n d^m$

Is Natural Language Context Free?

It doesn't look like it is context free either! Evidence comes from a Swiss German dialect and Bambara, a language spoken in Mali.

Crossing dependencies

omdat Wim$_1$ Jan$_2$ Henk$_3$ de kinderen$_4$ zag$_1$ helpen$_2$ leren$_3$ zwemmen$_4$
because Wim$_1$ Jan$_2$ Henk$_3$ the children$_4$ saw$_1$ help learn swim$_4$
because Wim saw Jan help Henk teach the children to learn to swim

| *zag* | depends on | *Wim* |, and | *helpen* | depends on | *Jan* |, etc.

Idea of Proof

Languages $\{xx|x \in \{a,b\}^*\}$ are not context-free.
Related $a^n b^m c^n d^m$ language also not context -free.
Swiss German crossing dependencies equivalent to $a^n b^m c^n d^m$

Is Natural Language Context Free?

It doesn't look like it is context free either! Evidence comes from a Swiss German dialect and Bambara, a language spoken in Mali.

Crossing dependencies

omdat Wim$_1$ Jan$_2$ Henk$_3$ de kinderen$_4$ zag$_1$ helpen$_2$ leren$_3$ zwemmen$_4$
because Wim$_1$ Jan$_2$ Henk$_3$ the children$_4$ saw$_1$ help learn swim$_4$
because Wim saw Jan help Henk teach the children to learn to swim

| *zag* | depends on | *Wim* |, and | *helpen* | depends on | *Jan* |, etc.

Idea of Proof

Languages $\{xx|x \in \{a,b\}^*\}$ are not context-free.
Related $a^n b^m c^n d^m$ language also not context -free.
Swiss German crossing dependencies equivalent to $a^n b^m c^n d^m$

Linear Indexed Grammars

A linear indexed grammar (LIG) is more powerful than a CFG, but much less powerful than an arbitrary CSG; it is a "mildly CS" grammar.

Definition

An indexed grammar has three disjoint sets of symbols: terminals, non-terminals and indices.

An index is a stack of symbols that can be passed from the LHS of a rule to its RHS, allowing counting and recording what rules were applied in what order.

Linear Indexed Grammars

$\mathrm{S} \rightarrow \mathrm{D}_f$	pushes an f onto the index on D
$\mathrm{D} \rightarrow \mathrm{D}_g$	pushes a g onto the index on D
$\mathrm{D} \rightarrow \mathrm{ABC}$	passes the index on D to A, B and C
$\mathrm{g} = \langle\, \mathrm{A} \rightarrow \mathrm{Aa} \mid \mathrm{B} \rightarrow \mathrm{Bb} \mid \mathrm{C} \rightarrow \mathrm{Cc} \,\rangle$	pops g from an index
$\mathrm{f} = \langle\, \mathrm{A} \rightarrow \mathrm{a} \mid \mathrm{B} \rightarrow \mathrm{b} \mid \mathrm{C} \rightarrow \mathrm{c} \,\rangle$	pops f from an index

Derivation in an Indexed Grammar

$\mathrm{S} \rightarrow \mathrm{D}_f$ $\quad \mathrm{g} = \langle\, \mathrm{A} \rightarrow \mathrm{Aa} \mid \mathrm{B} \rightarrow \mathrm{Bb} \mid \mathrm{C} \rightarrow \mathrm{Cc} \,\rangle$
$\mathrm{D} \rightarrow \mathrm{D}_g$ $\quad \mathrm{f} = \langle\, \mathrm{A} \rightarrow \mathrm{a} \mid \mathrm{B} \rightarrow \mathrm{b} \mid \mathrm{C} \rightarrow \mathrm{c} \,\rangle$
$\mathrm{D} \rightarrow \mathrm{ABC}$

S

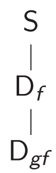
Derivation in an Indexed Grammar

$\mathrm{S} \rightarrow \mathrm{D}_f$ $\quad \mathrm{g} = \langle\, \mathrm{A} \rightarrow \mathrm{Aa} \mid \mathrm{B} \rightarrow \mathrm{Bb} \mid \mathrm{C} \rightarrow \mathrm{Cc} \,\rangle$
$\mathrm{D} \rightarrow \mathrm{D}_g$ $\quad \mathrm{f} = \langle\, \mathrm{A} \rightarrow \mathrm{a} \mid \mathrm{B} \rightarrow \mathrm{b} \mid \mathrm{C} \rightarrow \mathrm{c} \,\rangle$
$\mathrm{D} \rightarrow \mathrm{ABC}$

S
|
D_f

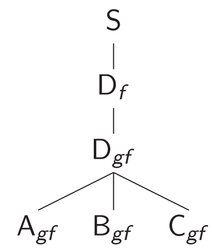
Derivation in an Indexed Grammar

$S \to D_f$ $\quad g = \langle\, A \to Aa \mid B \to Bb \mid C \to Cc \,\rangle$
$D \to D_g$ $\quad f = \langle\, A \to a \mid B \to b \mid C \to c \,\rangle$
$D \to ABC$



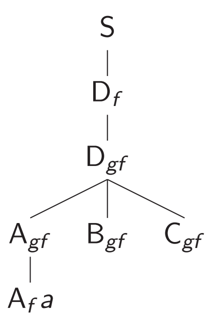
Derivation in an Indexed Grammar

$S \to D_f$ $\quad g = \langle\, A \to Aa \mid B \to Bb \mid C \to Cc \,\rangle$
$D \to D_g$ $\quad f = \langle\, A \to a \mid B \to b \mid C \to c \,\rangle$
$D \to ABC$



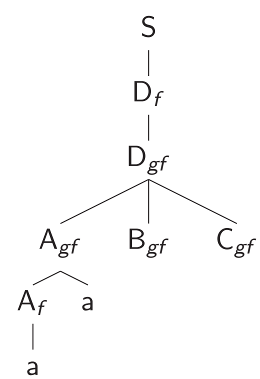
Derivation in an Indexed Grammar

$S \to D_f$ $\quad g = \langle\, A \to Aa \mid B \to Bb \mid C \to Cc \,\rangle$
$D \to D_g$ $\quad f = \langle\, A \to a \mid B \to b \mid C \to c \,\rangle$
$D \to ABC$



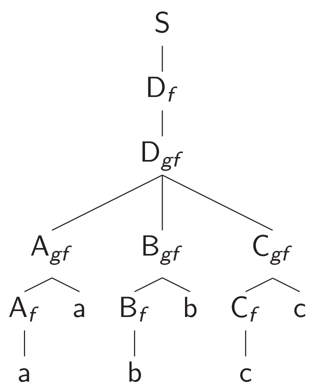
Derivation in an Indexed Grammar

$S \to D_f$ $\quad g = \langle\, A \to Aa \mid B \to Bb \mid C \to Cc \,\rangle$
$D \to D_g$ $\quad f = \langle\, A \to a \mid B \to b \mid C \to c \,\rangle$
$D \to ABC$



Derivation in an Indexed Grammar

$$S \to D_f \qquad g = \langle\, A \to Aa \mid B \to Bb \mid C \to Cc \,\rangle$$
$$D \to D_g \qquad f = \langle\, A \to a \mid B \to b \mid C \to c \,\rangle$$
$$D \to ABC$$



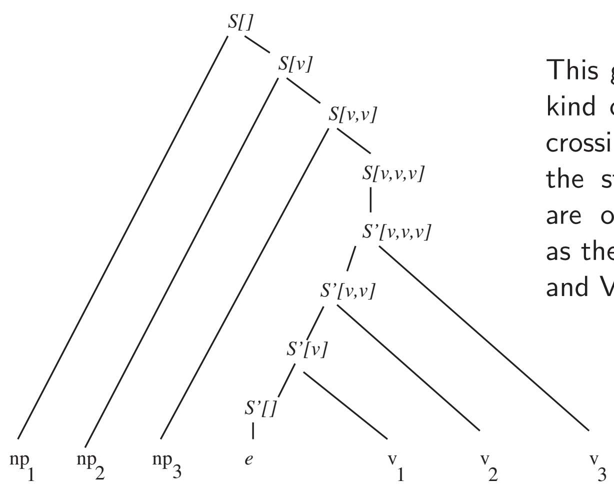
Linear Indexed Grammars

Linear Indexed Grammars (LIGs) allow an index to pass to only one non-terminal on the RHS (not three, as in previous example).

Here we'll push numbers onto an index.
An LIG for crossing dependencies in $np^k v^k$:

$S_{[\ldots]}$	$\to$	$np_i\ S_{[i,\ldots]}$	emit NP, push a number
$S_{[\ldots]}$	$\to$	$S'_{[\ldots]}$	switch to verb sequence rule
$S'_{[i,\ldots]}$	$\to$	$S'_{[\ldots]}\ v_i$	pop a number, emit a verb
$S'_{[\,]}$	$\to$	ϵ	stop if stack is empty

Example: LIG derivation for $np^3 v^3$



This grammar produces the kind of strings we want for crossing dependencies, but the structures it generates are only weakly adequate, as they don't associate NPs and Vs directly.

Linear Indexed Grammars

As a consequence of the weak adequacy of LIGs, other "mildly CS" grammar formalisms have been developed that are strongly adequate for NL:

- Tree Adjoining Grammar (TAG): a system of tree re-writing rules (ie, not string re-writing rules) in which elementary trees are combined by substitution and adjunction;
- Combinatory Categorial Grammar (CCG): a system that links words to complex categories that specify how adjacent words fit together, in terms of combinators like apply a functor to an argument, compose two functors, etc..

Summary

- The faculty of language contains a computational system that generates syntactic representations that can be mapped onto meanings.
- This raises the question of the complexity of this system (its position in the Chomsky hierarchy).
- A weakly adequate grammar generates the correct strings, while a strongly adequate one also generates the correct structures.
- There are structures in NLs which can be mapped on formal languages which are not context-free.
- NL probably belongs to the class of mildly context-sensitive languages, whose least powerful member (LIGs) is weakly adequate for NL.

Next Lecture: models of human parsing.