The Complexity of Human Language

Informatics 2A: Lecture 26

Adam Lopez

17 November 2015

1. Are natural languages regular?
2. Interlude: mathematical properties of context-freeness
3. Are natural languages context-free?
4. Mildly Context-Sensitive Grammars

Reading: J&M. Chapter 16.3–16.4.

Reminder: Essential epistemology

	Exact sciences	*Empirical sciences*	*Engineering*
Deals in...	Axioms and theorems	Facts and theories	Artifacts
Truth is...	Forever	Temporary	It works!
Examples...	Maths, CS theory, **formal language theory**	Physics, Biology, **Linguistics**	Many, inc. Applied CS and **NLP**

Essentially, all models are wrong, but some are useful.
— George Box

Recursion

The potential infiniteness of language has been recognized for centuries (by Galileo, Descartes, von Humboldt...)

Discrete Infinity

- Sentences are built up by discrete units
- There is no longest sentence

Mary thinks that John thinks that George thinks that Mary thinks that this course is boring!
I woke up and had a coffee and got dressed and checked facebook and walked in the park and ate lunch ...

Is Natural Language Regular?

Of course, many infinite languages are regular, e.g. $\{a^n | n \geq 0\}$ is regular. But what about natural languages?

1. yes
2. no

Question. How would you demonstrate this?

Possible answers. To show that a language is regular, write a finite automaton or regular expression that produces that grammar. To show that it is not regular, use the pumping lemma to demonstrate a contradiction (specifically, that you can pump some string in the language to produce a string not in the language.)

Does this seem hard?

Mathematical tools

We also have some other tools at our disposal.

Recall. (Lecture 20) If language L_1 is context-free and L_2 is regular, then $L_1 \cap L_2$ is a context-free language.

Claim. Context-free languages are closed under homomorphism (finite-state transduction). More precisely, if L_0 is a context-free language, and $\tau \in L_1 \times L_2$ is a context-free language, then $\tau(L_0)$ is context-free.

Sketch of proof. Use the Bar-Hillel construction (Lecture 20) on the grammar for L_0 and the transducer for τ, replacing the symbols in the alphabet of $L_0 \cap L_1$ for those of L_2 in the corresponding transitions.

Back to the main question...

Is Natural Language Regular?

Consider the set of sentences

Centre-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.
Assumption 1. *(the+noun)^n (transitive verb)^m likes tuna fish.* is ungrammatical for $m \neq n - 1$.
Assumption 2. n is unbounded. (Is this reasonable?)

Interlude: Context-free intersection

Recall two facts:

1. If languages L_1 and L_2 are regular, then $L_1 \cap L_2$ is a regular language.
2. If language L_1 is context-free and L_2 is regular, then $L_1 \cap L_2$ is a context-free language.

Question. If languages L_1 and L_2 are both context-free, what can we conclude about $L_1 \cap L_2$?

1. $L_1 \cap L_2$ is a regular language.
2. $L_1 \cap L_2$ is a context-free language.
3. $L_1 \cap L_2$ might not be a context-free language. (What then?)

Non-context-free languages

... Question. Are there non-context-free languages?

Intuition. If there is some property A such that any language with the property is context-free, and we can demonstrate that all possible languages have property A, then all languages must be context-free. Alternatively, if we can exhibit some language that *does not* have property A, then it is possible for languages to be non-context-free.

... Question. What is a fundamental property of context-free languages that we can test?

Non-context-free languages

Context-free languages can be pumped, much like regular languages

Claim. Suppose language L is context-free. Then L has the following property.
There exists an integer $k \geq 1$ (called the "pumping length") such that every string $s \in L$ with length of k or more symbols (i.e. $|s| \geq k$) can be written as $s = uvwxy$ with substrings u, v, w, x and y, such that:

1. $vx \geq 1$.
2. $vwx \leq k$.
3. $uv^n wx^n y \in L$ for all $i > 0$.

John will explain this in more detail next week.

Non-context-free languages

The basic idea is that for any sufficiently long string, there is some nonterminal that must be repeated along a path from root to string. The material inside the ancestor copy of this symbol and outside the descendant copy can be pumped.

Non-context-free languages

Claim. $a^n c^n b^n$ is not a context-free language. (Try to show this as an exercise...)

Suppose we have two languages:
$\mathcal{L}_1 = \{a^n b^n c^m | n, m \geq 0\}$
$\mathcal{L}_2 = \{a^m b^n c^n | n, m \geq 0\}$

Both $\mathcal{L}_1$ and $\mathcal{L}_2$ are context-free. (**Why?**)
The language $\mathcal{L}_1 \cap \mathcal{L}_2$ is $a^n c^n b^n$. Hence the intersection of context-free languages is not always context-free.

Non-context-free languages

Question. Just how complex is the family containing the intersection language?

Consider the following problem: I give you a set of indexed pairs of strings in the form $i : (\alpha_i, \beta_i)$ where i is the index, α_i and β_i are strings. E.g.

$\{1 : (a, baa), 2 : (ab, aa), 3 : (bba, bb)\}$

Now I ask you the following yes/ no question: is there a sequence of indexes $\iota_1 ... \iota_n$ (with repetitions and omissions allowed) s.t. $\alpha_{\iota_1} ... \alpha_{\iota_n} = \beta_{\iota_1} ... \beta_{\iota_n}$?

In this example, (3,2,3,1) works:

$$\begin{aligned} \alpha_3\alpha_2\alpha_3\alpha_1 &= bba + ab + bba + a \\ &= bbaabbbaa \\ &= bb + aa + bb + baa \\ &= \beta_3\beta_2\beta_3\beta_1 \end{aligned}$$

Non-context-free languages

Claim. We can answer this question by intersecting a pair of context-free grammars $\mathcal{L}_1$ and $\mathcal{L}_2$.

$\mathcal{L}_1$:

$S \to aS1|a1$

$S \to abS2|ab2$

$S \to bbaS3|bba3$

$\mathcal{L}_2$:

$S \to baaS1|baa1$

$S \to aaS2|aa2$

$S \to bbS3|bb3$

If $\mathcal{L}_1 \cap \mathcal{L}_2$ is nonempty, there must be a sequence of indexes that produces equivalent strings.

Problem. This Post Correspondence Problem is undecidable (Lecture 31)

Consequence. There is no algorithm that can always answer the question: is the intersection of two context-free languages nonempty. (Many other properties are also undecidable)

Back to question: Is Natural Language Context Free?

In Swiss German, some verbs (e.g. *let*, *paint*) take an object in accusative form, while others (e.g. *help*) take an object in dative form. The nouns are case-marked even in subordinate clauses, which in Swiss-German, can exhibit cross-serial dependencies.

Cross-serial dependencies

. . . das mer	d'chind	em Hans	es huus	lönd	hälfe	aastriiche
. . . that we	the children	Hans	the house	let	help	paint
	NP-ACC	NP-DAT	NP-ACC	V-ACC	V-DAT	V-ACC

. . . that we let the children help Hans paint the house

Back to question: Is Natural Language Context Free?

Claim 1. Swiss German subordinate clauses can have a structure in which all the Vs follow all the NPs.

Claim 2. Among such sentences, those with all dative NPs preceding all accusative NPs, and all dative-subcategorizing Vs preceding all accusative-subcategorizing Vs are acceptable.

Claim 3. The number of Vs requiring dative objects must equal the number of dative NPs and similarly for accusatives.

Claim 4. An arbitrary number of Vs can occur in a subordinate clause. (cf. similar claim in our proof of English context-freeness)

Back to question: Is Natural Language Context Free?

Claim. Swiss-German is not context-free.

Sketch of proof. Using the above claims and a transduction mapping dative NPs, accusative NPs, dative-subcategorizing Vs, and accusative-subcategorizing Vs to a, b, c, and d, respectively. This produces the sublanguage $a^n b^m c^n d^m$. If we intersect this language with $a^* b^* c^* d^*$, this sublanguage becomes the intersected language.

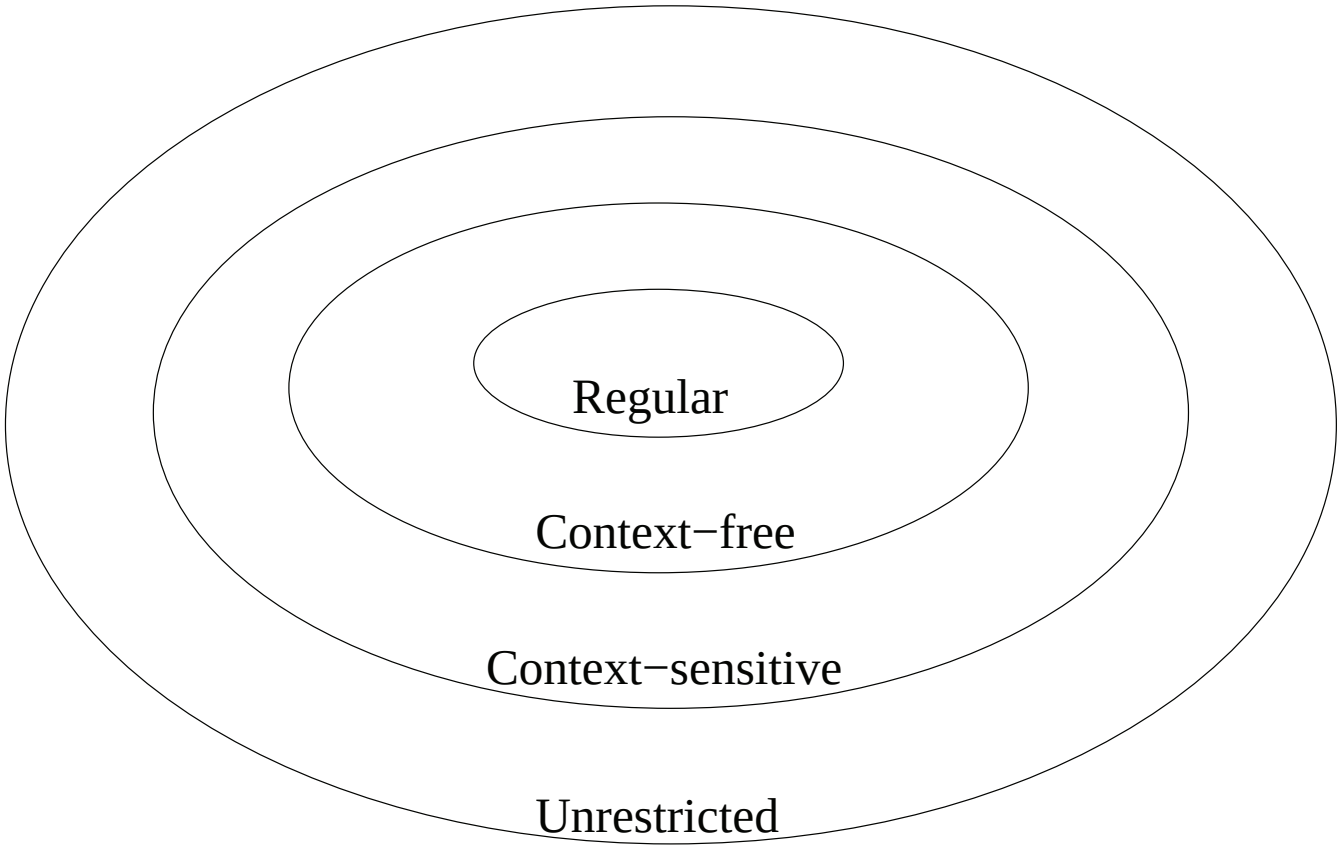
This language is not context-free (prove this as an exercise). But since context-free languages are closed under both finite-state transduction and intersection with regular languages, this means that Swiss-German cannot be context-free either!

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: the set of languages accepted by pushdown automata. Many aspects of PLs and NLs can be described at this level;
- Context-sensitive languages: equivalent with a linear bounded nondeterministic Turing machine, also called a linear bounded automaton. Need this to capture e.g. *typing rules* in PLs.
- Unrestricted languages: *all* languages that can in principle be defined via mechanical rules.

Review



Where do human languages fit within this complexity hierarchy?

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. (See previous lecture.)

A weakly adequate grammar

generates all and only the strings of a language but doesn't necessarily give a correct (insightful) account of their structures.

Weaker examples

These 'crossing dependencies' are non-context-free in a very strong sense: no CFG is even weakly adequate for modelling them.
Other phenomena can *in theory* be modelled using CFGs, though it seems unnatural to do so. E.g. a versus an in English.

a banana	an apple
a large apple	an exceptionally large banana

Over-simplifying a bit: a before consonants, an before vowels.

In theory, we could use a context-free grammar:

$$\begin{array}{rcl} \text{NP} & \to & \textbf{a}\ \text{NP1}^c \\ \text{NP1}^c & \to & \text{N}^c \mid \text{AP}^c\ \text{NP1} \\ \text{AP}^c & \to & \text{A}^c \mid \text{Adv}^c\ \text{AP} \end{array} \qquad \begin{array}{rcl} \text{NP} & \to & \textbf{an}\ \text{NP1}^v \\ \text{NP1}^v & \to & \text{N}^v \mid \text{AP}^v\ \text{NP1} \\ \text{AP}^v & \to & \text{A}^v \mid \text{Adv}^v\ \text{AP} \end{array}$$

But more natural to use context-sensitive rules, e.g.

$$\begin{array}{rcl} \text{DET [c-word]} & \to & \textbf{a}\ \text{[c-word]} \\ \text{DET [v-word]} & \to & \textbf{an}\ \text{[v-word]} \end{array}$$

Mild context sensitivity

A set $\mathcal{L}$ of languages is mildly context-sensitive if:

- $\mathcal{L}$ contains all context-free languages.
- $\mathcal{L}$ can describe cross-serial dependencies. There is an $n \geq 2$ such that $\{w^k | w \in T^*\} \in \mathcal{L}$ for all $k \leq n$.
- The languages in $\mathcal{L}$ are polynomially parsable.
- The languages in $\mathcal{L}$ have the constant growth property.

Let X be an alphabet and $L \subseteq X^*$. L has constant growth property iff there is a constant $c_0 > 0$ and a finite set of constants $C \subset \mathbb{N} \setminus \{0\}$ such that for all $w \in L$ with $|w| > c_0$, there is a $w' \in L$ with $|w| = |w'| + c$ for some $c \in C$

Example: the language $\{a^{2^n} | n \in \mathbb{N}\}$ does not have the constant growth property.

Combinatory Categorial Grammars

CCGs are more powerful than CFGs, but less powerful than arbitrary CSGs.

They satisfy the criteria for mildly context-sensitive languages, i.e. the set of languages defined by CCGs is mildly context-sensitive.

The set of categories (nonterminals) in CCG is compositional, defined by a set of atomic units such as S, NP and PP.

There are combination rules that tell us how to generate new categories from older ones in a derivation.

Linear Indexed Grammars

Linear indexed grammars (LIGs) are more powerful than CFGs, but much less powerful than an arbitrary CSGs. Think of them as mildly context sensitive grammars. These seem to suffice for NL phenomena.

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.

Summary

- The 'narrow' language faculty involves 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.
- NLs appear to surpass the power of context-free languages, but only just.
- The mild form of context-sensitivity captured by LIGs seems weakly adequate for NL structures.

Next Lecture: Models of human parsing.