Formal Languages and Linguistics

Pascal Amsili

Sorbonne Nouvelle, Lattice (CNRS/ENS-PSL/SN)

Cogmaster, september 2022

Overview

Formal Languages

Regular Languages

Formal Grammars

Formal complexity of Natural Languages

Introduction

Are NL regular?

Are NL context-free?

Are NL context-sensitive?

Motivation

Why an inquiry into the formal complexity of Natural Language(s)?

- It gives us knowledge about the **structure** of natural languages,
- It helps us assess the **adequation** of linguistic formalisms,
- It gives bound for the **complexity** of NLP tasks,
- It provides us with **predictions** about human language processing.

Hypotheses

We assume that:

- We can talk about "natural language" in general: all languages have a similar structure, a similar power
- Natural languages are recursively enumerable, i.e. they are formal languages
- Natural languages are infinite

⇒ Under these hypotheses, it is possible to ask the question: what is the complexity of natural languages?

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A stranger arrived.

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A tall stranger arrived.

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A tall handsome stranger arrived.

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A dark tall handsome stranger arrived.

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A very dark tall handsome stranger arrived.

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A very very dark tall handsome stranger arrived.

An infinite number of sentences

Arbitrary long sentences can be built by adding new material:

(4) A very very dark tall handsome stranger arrived.

A very^n handsome stranger arrived $\in$ NL

An infinite number of sentences

More interestingly, arbitrary long sentences can be built through center-embedding. In this case, there is a dependancy between arbitrary far apart elements:

An infinite number of sentences

More interestingly, arbitrary long sentences can be built through center-embedding. In this case, there is a dependancy between arbitrary far apart elements:

(5) The cats hunt.

An infinite number of sentences

More interestingly, arbitrary long sentences can be built through center-embedding. In this case, there is a dependancy between arbitrary far apart elements:

(5) The cats the neighbor owns hunt.

An infinite number of sentences

More interestingly, arbitrary long sentences can be built through center-embedding. In this case, there is a dependancy between arbitrary far apart elements:

(5) The cats the neighbor who arrived owns hunt.

An infinite number of sentences

More interestingly, arbitrary long sentences can be built through center-embedding. In this case, there is a dependancy between arbitrary far apart elements:

(5) The cats the neighbor who arrived owns hunt.

center-embedding: embedding a phrase in the middle of another phrase of the same type

Overview

Formal Languages

Regular Languages

Formal Grammars

Formal complexity of Natural Languages

- Introduction
- Are NL regular?
- Are NL context-free?
- Are NL context-sensitive?

Chomsky's first attempt

Consider the 3 structures:

- If S_1, then S_2.
- Either S_1 or S_2.
- The man who said S_1 is coming today.

1. The colored items are *dependent* one from the other
2. It is possible to create nested sentences of arbitrary length:

(6) If either the man who said S_a is coming today, or S_b, then S_c.

« Since such sentences are instances of mirroring and since the mirror language is not regular, then English is not regular » *(Chomsky, 1957, p. 22)*.

erroneous claim: **a regular language may contain a non regular sub-language**

Classical argument I

Let's consider the sentence(s):

(7) A man fired another man.

Classical argument I

Let's consider the sentence(s):

(7) A man that a man hired fired another man.

Classical argument I

Let's consider the sentence(s):

(7) A man that a man that a man hired hired fired another man.

Classical argument I

Let's consider the sentence(s):

(7) A man that a man that a man hired hired fired another man.
A man (that a man)^2 (hired)^2 fired another man.

Classical argument I

Let's consider the sentence(s):

(7) A man that a man that a man hired hired fired another man.
A man (that a man)^2 (hired)^2 fired another man.

The sentences (8) are all well-formed sentences (for any n).

(8) A man (that a man)^n (hired)^n fired another man.

Discussion

(9) A man (that a man)^n (hired)^n fired another man.

(10) #A girl that the man that the doctor knows like was fired.

Good examples:

(11) A foreman that an employee who were recently hired talked with was fired.

Discussion: processing problems with nested structures

Psycholinguistic evidence that (12b) is more accepted than (12a) (Fodor, Frazier)

(12) a. The patient who the nurse who the clinic had hired admitted met Jack.
b. The patient who the nurse who the clinic had hired met Jack.

Other factors:

(13) a. The pictures which the photographer who I met yesterday took were damaged by the child.
b. ?The pictures which the photographer who John met yesterday took were damaged by the child.

(14) a. Isn't it true that example sentences [that people [that you know] produce] are more likely to be accepted? (De Roeck et al, 1982)
b. A book [that some Italian [I've never heard of] wrote] will be published soon by MIT Press (Frank, 1992)

(Gibson & Thomas, 1997)

Discussion (end)

- Obvious problems of performance
- however in writing, or with an appropriate intonation, there doesn't seem to be a hard-wired limit

Classical Argument II

$$\begin{aligned} \text{Let} \quad x &= \text{that a man} \\ y &= \text{hired} \\ w &= \text{a man} \\ v &= \text{fired another man} \end{aligned}$$

- wx^*y^*v is regular
- English $\cap$ $wx^*y^*v = wx^ny^nv$ (14)
- If English is regular, then wx^ny^nv must be regular (for the intersection of two regular languages is regular)
- But wx^ny^nv is not regular (pumping lemma).
  Contradiction $\Rightarrow$ English is not regular.

(Schieber, 1985)

Overview

Formal Languages

Regular Languages

Formal Grammars

Formal complexity of Natural Languages

- Introduction
- Are NL regular?
- Are NL context-free?
- Are NL context-sensitive?

Pumping lemma: intuition

1. If a word is long enough, then there is (at least) one non terminal symbol appearing several times in its derivation.

"long enough" ?

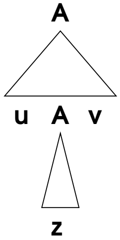
$$\begin{array}{lcl} S & \rightarrow & A\ B \\ A & \rightarrow & abaccabca \\ & | & abSba \\ B & \rightarrow & ccccc \end{array}$$

Minimal length : 14:

$S \rightarrow AB \rightarrow abaccabcaB \rightarrow abaccabcaccccc$

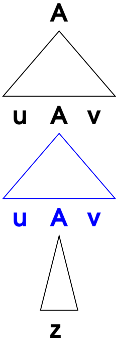
Pumping lemma: intuition

2 Let's call this non terminal symbol *A*.



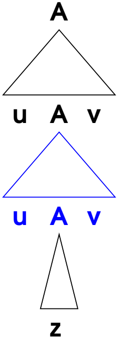
Pumping lemma: intuition

2 Let's call this non terminal symbol *A*.



Pumping lemma: intuition

2 Let's call this non terminal symbol A.



$$A \xrightarrow{*} uAv$$

$$A \xrightarrow{*} uAv \xrightarrow{*} uzv$$

$$A \xrightarrow{*} uAv \xrightarrow{*} uuAvv \xrightarrow{*} \underbrace{u \ldots u}_{n} z \underbrace{v \ldots v}_{n}$$

Pumping Lemma for CF languages

Def. 20 (Star lemma – CF languages)

If L is context-free, there exists $p \in \mathbb{N}$ such that:
$\forall w$ s.t. $|w| \geqslant p$,
w can be factorized $w = rstuv$,
with:

$$\begin{aligned} & |su| \geqslant 1 \\ & |stu| \leqslant p \\ \forall i \geqslant 0, \quad & rs^i tu^i v \in L \end{aligned}$$

(Bar-Hillel *et al.* , 1961)

Pumping lemma: Consequences

The pumping lemma gives us a tool to prove that a language is **not** context-free.

$\mathcal{L}$ context-free	$\Rightarrow$	pumping lemma $(\forall i, rs^i tu^i v \in \mathcal{L})$
pumping lemma	$\not\Rightarrow$	$\mathcal{L}$ context-free
NO pumping lemma	$\Rightarrow$	$\mathcal{L}$ **NOT** context-free

to prove that $\mathcal{L}$ is

context-free provide a type 2 grammar

not context-free show that the pumping lemma does not apply

Results: expressivity

- well-parenthetized words (dyck's language) is context-free
  $S \rightarrow (S)S \mid \varepsilon$
- $a^n b^n (n \geqslant 0)$ is a context-free language
  $S \rightarrow aSb \mid \varepsilon$
- $ww^R, w \in \Sigma^*$ (mirror language) is a context-free language
  $S \rightarrow aSa \mid bSb \mid \varepsilon$
- $ww, w \in \Sigma^*$ (copy language) is not context-free
  proof: pumping lemma
- $a^n b^n c^n$ is not context-free
  proof: pumping lemma
- $a^m b^n c^m d^n$ is not context-free
  proof: pumping lemma
- $xa^m b^n yc^m d^n z$ is not context-free
  proof: pumping lemma

Closure properties I

- CF languages are closed under rational operations

▶ union (gather all the rules, avoiding name conflicts, and adding a new start rule $S \rightarrow S_1|S_2$),

▶ product ($S \rightarrow S_1 S_2$),

▶ and Kleene star ($S \rightarrow S_1 S \mid \varepsilon$).

Closure properties II : intersection

- CF languages are not closed under intersection

Example

$L_1 = \{a^i b^i c^j \mid i, j \geq 0\}$ is context-free: $\begin{aligned} S &\rightarrow XY \\ X &\rightarrow aXb \mid \varepsilon \\ Y &\rightarrow cY \mid \varepsilon \end{aligned}$

$L_2 = \{a^i b^j c^j \mid i, j \geq 0\}$ is also context-free: $\begin{aligned} S &\rightarrow XY \\ X &\rightarrow aX \mid \varepsilon \\ Y &\rightarrow bYc \mid \varepsilon \end{aligned}$

But $L_1 \cap L_2 = \{a^n b^n c^n \mid n \geq 0\}$ is not contex-free.

Closure properties III: other results

- CF languages are not closed under complement (since they are not closed under intersection)
- CF languages are closed under intersection with a regular language
- a sub-class of CF languages, *deterministic CF languages* are closed for set complement, but not for union (one can easily define an intrinsequely non deterministic language as the union of two "independant" languages)

Final argument I

After many attempts by various scholars, attempts which are severely critized and ruined in (Gazdar & Pullum, 1985), Schieber (1985) came up with a widely accepted answer:

1. In swiss-german, subordinate clauses can have a structure where all NPs precede all Vs. (15)

(15) Jan säit das mer NP* es huus haend wele V* aastrüche
Jan said that we NP* the house have wanted V* paint
'Jan said that we have wanted (that) V* NP* paint the house'

2. Among those subordinate clauses, those where all the dative NPs precede all the accusative NPs are well-formed. (16)

(16) ... das mer d'chind em Hans es huus haend wele laa hälfe aastrüche
... that we the_children.ACC Hans.DAT the house.ACC have wanted let help paint
'... that we have wanted to let the children help Hans to paint the house'

Final argument II

3. The number of verbs requiring a dative has to be equal to the number of dative NPs, the same for accusative.
4. The number of verbs in a subordinate clause is limited only by performance

Let R be the language:

R = {Jan säit das mer (d'chind)^h (em Hans)^i es huus haend wele (laa)^j (hälfe)^k aastrüche, $i, j, k, h \geqslant 1$}

Then let $L =$ Swiss-German $\cap\ R =$

{Jan säit das mer (d'chind)^m (em Hans)^n es huus haend wele (laa)^m (hälfe)^n aastrüche, $m, n \geqslant 1$}

L is not context-free, whereas R is regular.

⇒ Swiss-German is not context-free.

Overview

Formal Languages

Regular Languages

Formal Grammars

Formal complexity of Natural Languages

Introduction

Are NL regular?

Are NL context-free?

Are NL context-sensitive?

Joshi's proposal

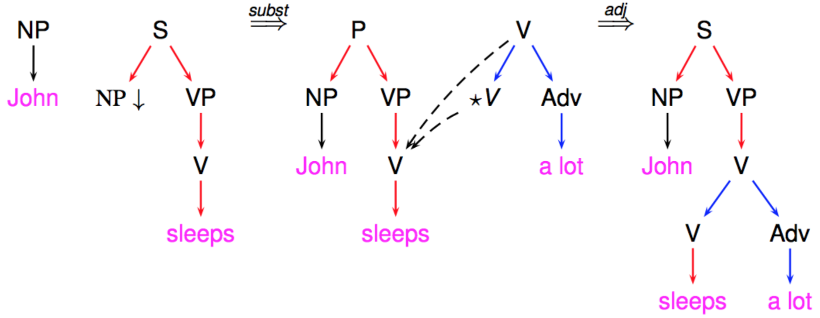
Joshi (1985): what's needed is a class of grammars/languages that are only slightly more powerfull than CFGs.
A class of mildly context-sensitive grammars should have the following properties:

- limited cross-serial dependencies (cf. Swiss-German)
- constant growth (a^{2^i} should not belong to the class)
- polynomial parsing

The class should of course also include all CFG languages.

Formal definitions still needed; note that parsing depends on the grammar rather than on the language

Tree Adjoining Grammars



TAG = MCSL

Tree Adjoining Grammars define the class of MCSL, which have the following properties (among others):

- ww is MCS
- $a^n b^n c^n$ is MCS
- $a^n b^n c^n d^n$ is MCS
- $a^i b^j c^i d^j$ is MCS
- $a^n b^n c^n d^n e^n$ is not MCS
- www is not MCS
- $ab^h ab^i ab^j ab^k ab^l, h > i > j > k > l \geqslant 1$ is not MCS
- a^{2^i} is not MCS

TAG = MCSL

Tree Adjoining Grammars define the class of MCSL, which have the following properties (among others):

- ww is MCS
- $a^n b^n c^n$ is MCS
- $a^n b^n c^n d^n$ is MCS
- $a^i b^j c^i d^j$ is MCS
- $a^n b^n c^n d^n e^n$ is not MCS
- www is not MCS
- $ab^h ab^i ab^j ab^k ab^l, h > i > j > k > l \geqslant 1$ is not MCS
- a^{2^i} is not MCS

Conjecture : NL $\in$ MCSL

Categorial Combinatorial Grammars

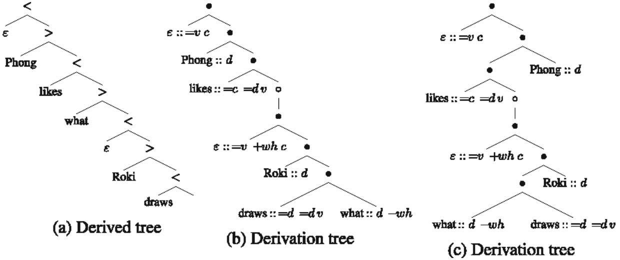
A formalism introduced by Steedman (see (Steedman *et al.* , 2012))

$$\dfrac{\dfrac{\dfrac{\textbf{the}}{NP/N} \qquad \dfrac{\textbf{dog}}{N}}{NP}> \qquad \dfrac{\dfrac{\textbf{bit}}{(S\backslash NP)/NP} \qquad \dfrac{\textbf{John}}{NP}}{S\backslash NP}>}{S}<$$

Vijay-Shanker & Weir (1994) proved the équivalence between CCG and TAG

Other formalisms

From the minimalist programme Chomsky (1995), a formalism called Minimalist Grammars was introduced by Stabler (2011).



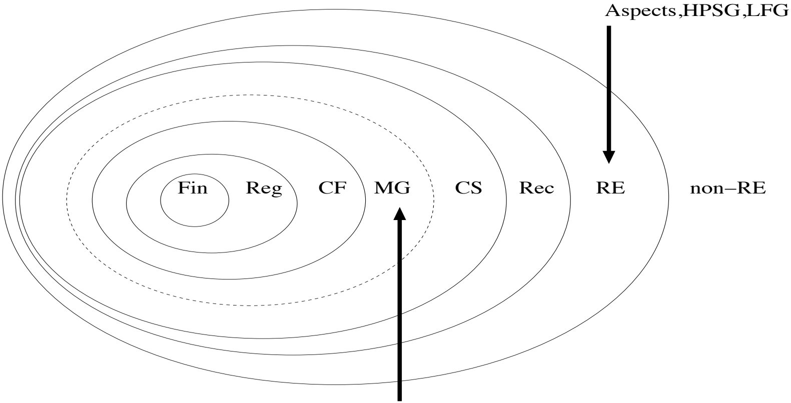
(a) Derived tree

(b) Derivation tree

(c) Derivation tree

It has been demonstrated that the class of languages definable by MGs is exactly the class definable by multiple CFG (MCFGs), linear context-free rewrite systems (LCFRSs), and other formalisms.

Big picture (Stabler, 2011)



Th: $CE \subset \boxed{TAG \equiv CCG} \subset \boxed{MCFG \equiv LCFRS \equiv MG} \subset CS$

References I

Bar-Hillel, Yehoshua, Perles, Micha, & Shamir, Eliahu. 1961. On formal properties of simple phrase structure grammars. *STUF-Language Typology and Universals*, 14(1-4), 143–172.

Chomsky, Noam. 1957. *Syntactic Structures*. Den Haag: Mouton & Co.

Chomsky, Noam. 1995. *The Minimalist Program*. Vol. 28. Cambridge, Mass.: MIT Press.

Gazdar, Gerald, & Pullum, Geoffrey K. 1985 (May). *Computationally Relevant Properties of Natural Languages and Their Grammars*. Tech. rept. Center for the Study of Language and Information, Leland Stanford Junior University.

Gibson, Edward, & Thomas, James. 1997. The Complexity of Nested Structures in English: Evidence for the Syntactic Prediction Locality Theory of Linguistic Complexity. *Unpublished manuscript, Massachusetts Institute of Technology*.

Joshi, Aravind K. 1985. *Tree Adjoining Grammars: How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions?* Tech. rept. Department of Computer and Information Science, University of Pennsylvania.

Langendoen, D Terence, & Postal, Paul Martin. 1984. *The vastness of natural languages*. Basil Blackwell Oxford.

Mannell, Robert. 1999. *Infinite number of sentences*. part of a set of class notes on the Internet. http://clas.mq.edu.au/speech/infinite_sentences/.

Schieber, Stuart M. 1985. Evidence against the Context-Freeness of Natural Language. *Linguistics and Philosophy*, 8(3), 333–343.

Stabler, Edward P. 2011. Computational perspectives on minimalism. *Oxford handbook of linguistic minimalism*, 617–643.

References II

Steedman, Mark, *et al.* . 2012 (June). *Combinatory Categorial Grammars for Robust Natural Language Processing*. Slides for NASSLLI course `http://homepages.inf.ed.ac.uk/steedman/papers/ccg/nasslli12.pdf`.

Vijay-Shanker, K., & Weir, David J. 1994. The Equivalence of Four Extensions of Context–Free Grammars. *Mathematical Systems Theory*, **27**, 511–546.