Computational Linguistics: Chomsky Hierarchy

Raffaella Bernardi

e-mail: raffaella.bernardi@unitn.it

1. Credits

Remark Some of the slides (on Chomsky Hierarchy, etc.) are from Gerhard Jaeger course given at ESSLLI '04.

2. Generative Power

Every (formal) grammar generates a unique language. However, one language can be generated by several different (formal) grammars.

Formal grammars differ with respect to their **generative power**.

One grammar is of a greater generative power than another if it can recognize a language that the other cannot recognize.

Two grammars are said to be

- **weakly** equivalent if they generate the same string language.
- **strongly** equivalent if they generate both the same string language and the same tree language.

3. Hierarchy of Grammars and Languages

A hierarchy of grammars: the set of languages describable by grammars of grater power subsumes the set of language describable by grammars of less power.

The most commonly used hierarchy is the **Chomsky Hierarchy of Languages** introduced in 1959.

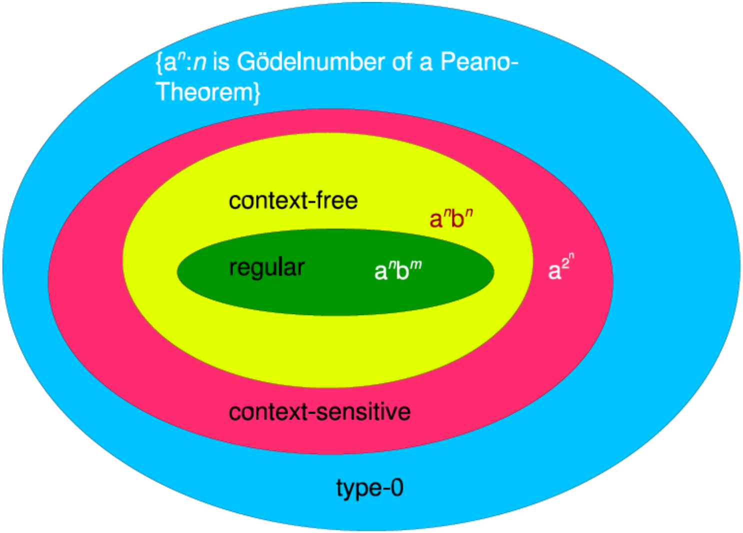
Hence, the two questions to ask are:

- Where does Natural Language fit in the Chomsky Hierarchy?
- Which is the generative power of the different Formal Grammars?

If we know where NL fit, we would know

- which formal language can represent NL;
- which rules to use in writing formal grammars for NL.

4. Chomsky Hierarchy of Languages



5. Dissenting Views

Claim: NL are not RL.

- all arguments to this effect use center-embedding
- humans are extremely bad at processing center-embedding
- notion of competence that ignores this is dubious
- natural languages are regular after all.

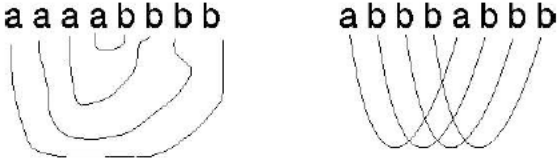
5.1. Are NL Context Free (CF)?

History of the problem:

1. Chomsky 1957: conjectures that natural languages are not CF
2. sixties, seventies: many attempts to prove this conjecture
3. Pullum and Gazdar 1982:
    - all these attempts have failed
    - for all we know, natural languages (conceived as string sets) might be context-free
4. Huybregts 1984, Shieber 1985: proof that Swiss German is not context-free
5. Culy 1985: proof that Bambara is not context-free

5.2. Nested and Crossing Dependencies

- CFLs—unlike regular languages—can have unbounded dependencies
- however, these dependencies can only be **nested**, not **crossing**
- example:
  - $a^n b^n$ has unlimited nested dependencies → context-free
  - the copy language has unlimited crossing dependencies → not context-free



5.3. Cross-serial dependencies Swiss German

Today many theorists believe natural languages are not context-free. (Huybregts 1984, Shieber 1985).

Evidences are given by **cross-serial dependencies** found in Swiss German where verbs and their argument can be orderded cross-serially.

A sentence can have a string of

dative nouns followed by a string of accusative nouns, followed by a string of dative-taking verbs, followed by a string of accusative-taking verbs. E.g.

mer em Hans es huus halfed aastriiche
we Hans/Dat the house/Acc helped paint.

tr. we helped Hans paint the house.

The number of verbs requaring dative objects must equal the number of dative NPs and similarly for accusatives, and this number can be arbitrary. Hence, the language representing this phenomenon is $wa^nb^mxc^nd^my$ which is not Context Free (CF).

However, notice that those construction types used to prove that NLs is not CF appear to be hard to understand for humans too.

6. Where does NL fit?

How large NL are continues to be a less simple matter. There are two main non-compatible views:

1. Natural Language forms a class of languages that **includes the CF** family, but is larger than it.
2. Natural Language occupies a position eccentric with respect to that hierarchy, in such a way that it does not contain any whole family in the hierarchy but is **spread along all of them**

The first view gave rise to a new family of languages which is of clear linguistic interest, **Mildly Context-sensitive Languages**.

7. Mildly Context-sensitive Languages (MSC)

A concept motivated by the intention of characterizing a narrow class of formal grammars which are **only slightly more powerful than CFGs**, and which nevertheless allow for descriptions of natural languages in a linguistically significant way (Joshi 1985).

According to Joshi (1985, p. 225) a mildly context-sensitive language, L, has to fulfill three criteria, to be understood as a rough characterization. Somewhat paraphrased, these are:

1. the parsing problem for L is solvable in **polynomial time**,
2. L has the constant **growth property** (i.e. the distribution of string lengths should be linear rather than supralinear.), and
3. there is a finite **upper bound** for L limiting the number of different instantiations of factorized **cross-serial dependencies** occurring in a sentence of L.

8. Where do the different Formal Grammars stand?

The interest in the frameworks is tied to their generative power, . . . as well as their **destiny**.

Chomsky's formal language theory made it possible to ask for the generative strength of a grammar.

After the discovery of languages which require cross-serial dependencies, grammars that were proved to be Context Free **lost their appeal**. Since CFGs were shown to be inadequate to model those natural languages.

We are interested in the problem of determining whether a string is in the language generated/recognized by a grammar of a certain type.

- For Context Free Language the problem is polynomial.
- the same holds for Mildly CFL.
- whereas, for Context Sensitive Languages the problem is PSPACE-complete

If NLs were CSL, this would be a bad news for CL!

9. Next class

On the 5th of October, will discuss the content of this video: where do you stand? What is your opinion?

During the online meeting (all in zoom 17:00-18:00), we will do pen-and-paper exercises on:

- GFG and Formal Languages
- CFG and Natural Language