עיבוד שפות טבעיות

שולי וינטנר

Complexity of natural languages

Computational complexity: the expressive power of (and the resources needed in order to process) classes of languages

Linguistic complexity: what makes individual constructions or sentences more difficult to understand

> This is the dog, that worried the cat, that killed the rat, that ate the malt, that lay in the house that Jack built.
>
> This is the malt that the rat that the cat that the dog worried killed ate.

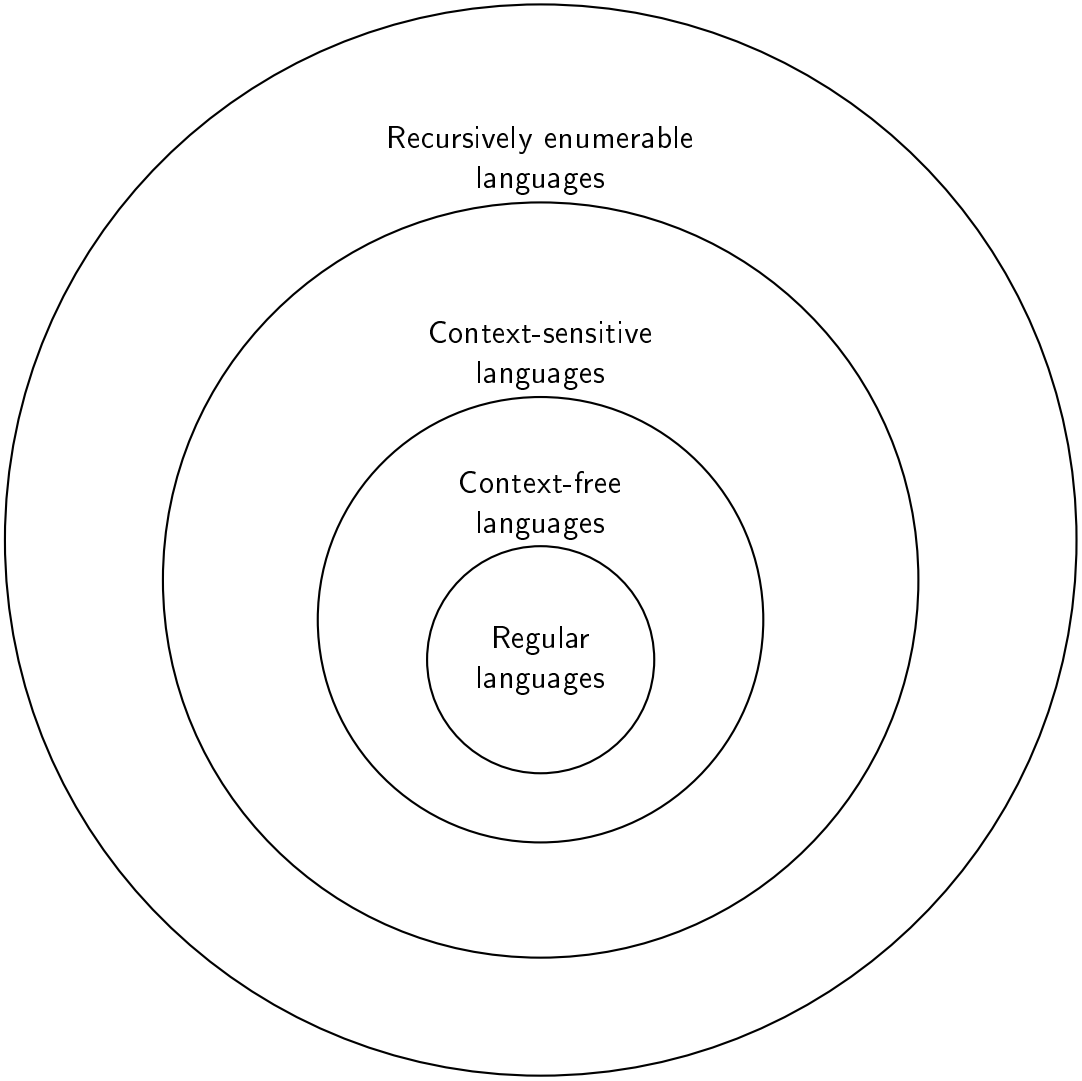
The Chomsky hierarchy of languages

A hierarchy of *classes* of languages, viewed as sets of strings, ordered by their "complexity". The higher the language is in the hierarchy, the more "complex" it is.

In particular, the class of languages in one class properly includes the languages in lower classes.

There exists a correspondence between the class of languages and the format of phrase-structure rules necessary for generating all its languages. The more restricted the rules are, the lower in the hierarchy the languages they generate are.

The Chomsky hierarchy of languages



The Chomsky hierarchy of languages

Regular (type-3) languages:

Grammar: right-linear or left-linear grammars
Rule form: $A \rightarrow \alpha$ where $A \in V$ and $\alpha \in \Sigma^* \cdot V \cup \{\epsilon\}$.
Computational device: finite-state automata

The Chomsky hierarchy of languages

Context-free (type-2) languages:

Grammar: context-free grammars
Rule form: $A \rightarrow \alpha$ where $A \in V$ and $\alpha \in (V \cup \Sigma)^*$
Computational device: push-down automata

The Chomsky hierarchy of languages

Context-sensitive (type-1) languages:

Grammar: context-sensitive grammars

Rule form: $\alpha \rightarrow \beta$ where $|\beta| \geq |\alpha|$

Computational device: linear-bounded automata (Turing machines with a finite tape, linearly bounded by the length of the input string)

The Chomsky hierarchy of languages

Recursively-enumerable (type-0) languages:

Grammar: general rewriting systems
Rule form: $\alpha \rightarrow \beta$ where $\alpha \neq \epsilon$
Computational device: Turing machines

Where are natural languages located?

Why is it interesting?

The hierarchy represents some informal notion of the complexity of natural languages

It can help accept or reject linguistic theories

It can shed light on questions of human processing of language

Where are natural languages located?

What exactly is the question?

When viewed as a set of strings, is English a regular language? Is it context-free? How about Hebrew?

Competence vs. Performance

This is the dog, that worried the cat, that killed the rat, that ate the malt, that lay in the house that Jack built.
This is the malt that the rat that the cat that the dog worried killed ate.

Where are natural languages located?

Chomsky (1957):

"English is not a regular language"

As for context-free languages, "I do not know whether or not English is itself literally outside the range of such analyses"

How *not* to do it

An introduction to the principles of transformational syntax (Akmajian and Heny, 1976):

> Since there seem to be no way of using such PS rules to represent an obviously significant generalization about one language, namely, English, we can be sure that phrase structure grammars cannot possibly represent *all* the significant aspects of language structure. We must introduce a new kind of rule that will permit us to do so.

How *not* to do it

Syntax (Peter Culicover, 1976):

> In general, for any phrase structure grammar containing a finite number of rules like (2.5), (2.52) and (2.54) it will always be possible to construct a sentence that the grammar will not generate. In fact, because of recursion there will always be an infinite number of such sentences. Hence, the phrase structure analysis will not be sufficient to generate English.

How *not* to do it

Transformational grammar (Grinder & Elgin, 1973)

the girl saw the boy

$*$*the girl kiss the boy*

this well-known syntactic phenomenon demonstrates clearly the inadequacy of ... context-free phrase-structure grammars...

How *not* to do it

the defining characteristic of a context-free rule is that the symbol to be rewritten is to be rewritten without reference to the context in which it occurs... Thus, by definition, one cannot write a context-free rule that will expand the symbol V into *kiss* in the context of being immediately preceded by the sequence *the girls* and that will expand the symbol V into *kisses* in the context of being immediately preceded by the sequence *the girl*. In other words, any set of context-free rules that generate (correctly) the sequences *the girl kisses the boy* and *the girls kiss the boy* will also generate (incorrectly) the sequences *the girl kiss the boy* and *the girls kisses the boy*.

How *not* to do it

The grammatical phenomenon of Subject-Predicate agreement is sufficient to guarantee the accuracy of: "English is not a context-free language".

How *not* to do it

Syntactic theory (Bach 1974):

> These examples show that to describe the facts of English number agreement is literally impossible using a simple agreement rule of the type given in a phrase-structure grammar, since we cannot guarantee that the noun phrase that determines the agreement will precede (or even be immediately adjacent) to the present-tense verb.

How *not* to do it

A realistic transformational grammar (Bresnan, 1978):

> in many cases the number of a verb agrees with that of a noun phrase at some distance from it... this type of syntactic dependency can extend as far as memory or patience permits... the distant type of agreement... cannot be adequately described even by context-sensitive phrase-structure rules, for the possible context is not correctly describable as a finite string of phrases.

How *not* to do it

What is the source for this confusion?

The notion of "context-freeness"

How to do it right

Proof techniques:

- The pumping lemma for regular languages
- Closure under intersection
- Closure under homomorphism

English is not a regular language

Center embedding:

The following is a sequence of grammatical English sentences:

A white male hired another white male.

A white male – whom a white male hired – hired another white male.

A white male – whom a white male, whom a white male hired, hired – hired another white male.

Therefore, the language L_{trg} is a subset of English:

$L_{trg} = \{$ *A white male (whom a white male)*^{n} *(hired)*^{n} *hired another white male* $\mid n > 0\}$

English is not a regular language

- L_{trg} is *not* a regular language

- L_{trg} is the intersection of the natural language English with the regular set

  $L_{reg} = \{$ *A white male (whom a white male)** *(hired)** *hired another white male* $\}$

- L_{reg} is regular, as it is defined by a regular expression.

Since the regular languages are closed under intersection, and since L_{reg} is a regular language, then if English were regular, its intersection with L_{reg}, namely L_{trg}, would be regular. Since L_{trg} is trans-regular, English is *not* a regular language.

English is not a regular language

Similar constructions:

The cat likes tuna fish

The cat the dog chased likes tuna fish

The cat the dog the rat bit chased likes tuna fish

The cat the dog the rat the elephant admired bit chased likes tuna fish

$$(the + N)^n \; Vt^{n-1} \; likes\ tuna\ fish$$

English is not a regular language

Another construction:

If S_1 then S_2

Either S_3 or S_4

Through a homomorphism that maps *if, then* to a and *either, or* to b, and all other words to ϵ, English can be mapped to the trans-context-free language $\{xx^R \mid x \in \{a+b\}^*\}$

Is English context-free?

The common proof technique for showing that a language is not context-free is the pumping lemma for context-free languages, and two closure properties: closure under homomorphisms and under intersection with regular languages.

Some languages that are *not* context-free:

$$\{xx \mid x \in \{a+b\}^*\}$$
$$\{a^m b^n c^m d^n\}$$

Natural languages are not context-free

Data from Swiss-German:

Jan säit das (Jan said that)

mer	em Hans	es	huus	hälfed	aastriiche
we	*Hans-DAT*	*the*	*house-ACC*	*helped*	*paint*

"we helped Hans paint the house"

mer	d'chind	em Hans	es	huus	haend	wele
we	*the kids-ACC*	*Hans-DAT*	*the*	*house-ACC*	*have*	*wanted to*

laa	hälfe	aastriiche	
let	*help*	*paint*	

"we have wanted to let the kids help Hans paint the house"

Natural languages are not context-free

Dative NP's must precede accusative NP's, and dative-taking verbs must precede accusative-taking verbs:

Jan säit das mer (d'chind)* (em Hans)* es huus haend wele laa* hälfe* aastriiche

However, the number of verbs requiring dative objects (*hälfe*) must equal the number of dative NP's (*em Hans*), and similarly for accusatives. Intersecting the language defined by the above regular expression with Swiss-German yields

Jan säit das mer (d'chind)^m (em Hans)^n es huus haend wele laa^m hälfe^n aastriiche

which is trans-context-free.

Linguistic complexity

Why are some sentences more difficult to understand than others?

The cat the dog the rat bit chased likes tuna fish

Limitations on stack size?

Weak and strong generative capacity

If formal language theory, the natural equivalence relation on grammars is $G_1 \equiv G_2$ iff $L(G_1) = L(G_2)$.

When grammars for natural languages are involved, we say that G_1 and G_2 are *weakly equivalent* if their string languages are identical.

Two grammars are *strongly equivalent* if they are weakly equivalent and, in addition, assign the same structure to each sentence.

Weak and strong generative capacity

The *strong generative capacity* (or power) of a linguistic formalism is its ability to associate structure to strings.

Even if context-free grammars are weakly sufficient for generating all natural languages, their strong generative capacity is probably not sufficient for natural languages.