Grammar Complexity

Jurafsky and Martin, Chapter 16

Miriam Butt

October 2012

Complexity

- How Complex is a given Problem?
- What formal mechanisms best model this complexity?

Natural Language: used to be thought of as a sort of "code". That is hard, but regular.

Now: mind-bogglingly complex.

But: is it an unsolvable problem?

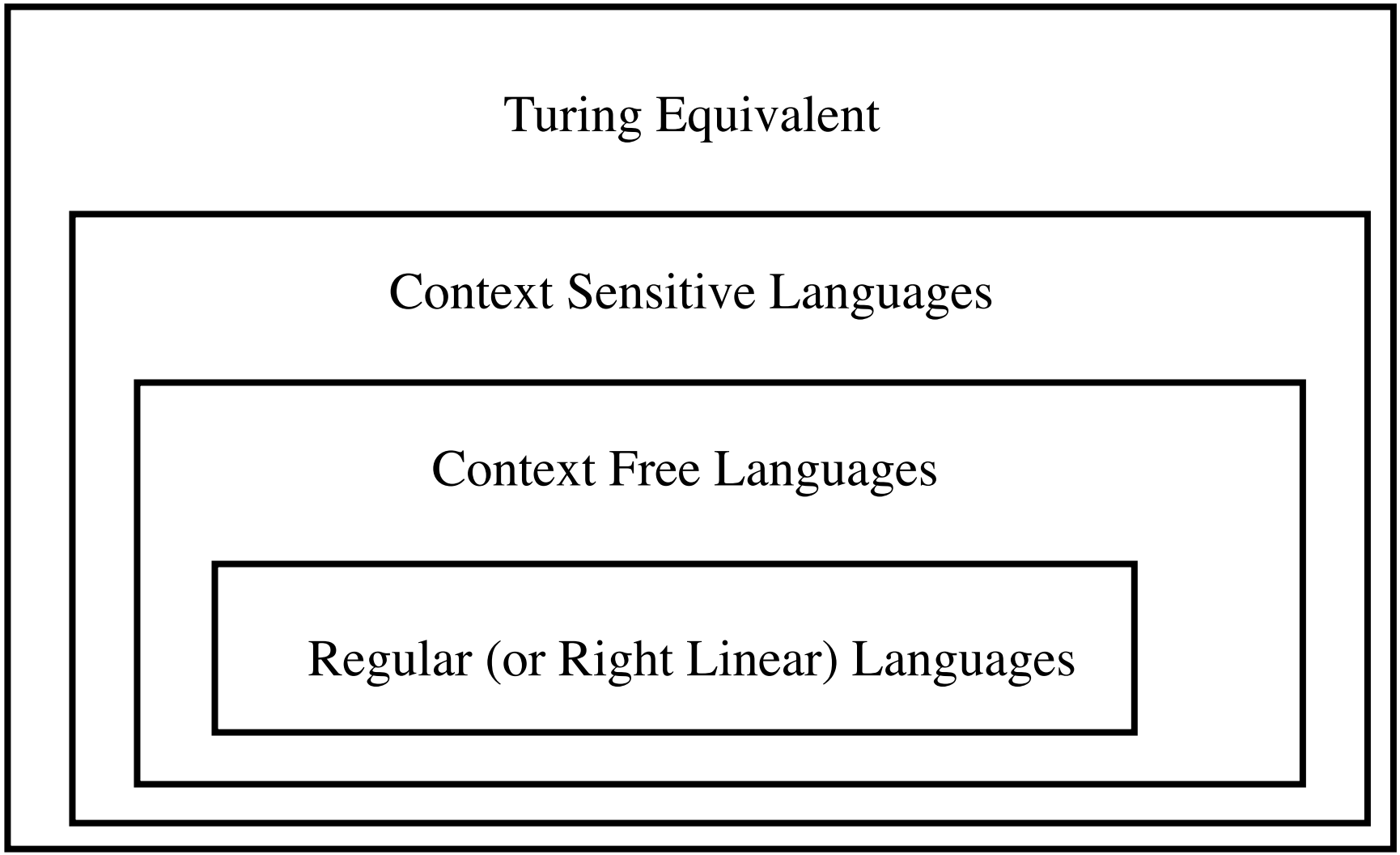
Generative Power

Chomsky defined a theory of language (syntax) in terms of **generative** linguistics.

Given a set of rules and a lexicon, what well-formed expressions can we generate and do those adequately cover the empirical data we observe?

"One grammar is of greater generative power or complexity than another if it can define a language that the other cannot define." (J&M p. 564)

The Chomsky Hierarchy



Natural Language

Is it regular? Overall no.

But, subparts of it are: phonology and morphology (can be treated via FST which are known to be regular, Kaplan and Kay 1994, Karttunen 2002).

How can we tell if a language is not regular?

The Pumping Lemma

The Pumping Lemma

Let L be an infinite regular language. Then there are strings x, y, and z, such that $y \neq \varepsilon$ and $xy^nz \in L$ for $n \geq 0$.

If a language is regular, it can be modeled by a FSA.

If you have a string which is longer than the fixed number of, the FSA must have a loop.

$a^n\, b^n$ is not a part of this language (from J&M 16.2.1)

Trying to Solve $a^n b^n$

See if one can get to $a^n b^n$ via $xy^n z$.

1. Assume y is composed of as. Then x is all as as well, z all bs. But if so, then always have more as than bs!
2. Assume y is composed of bs. Then z is all bs as well, x all as. But if so, then always have more bs than as!
3. Assume y is composed of as and bs. Then z is all bs, x all as. But if so, then also allow some bs before as, so no good either.

So $a^n b^n$ is not a regular language.

Natural Language

Natural Language contains strings like:

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.

$a^n\ b^{n-1}$ so, not a regular language

Natural Language

Another famous case: Swiss cross-serial dependencies

Jan säit das,

… mer em Hans es huus hälfed aastriche.
we the Hans.Dat the house.Acc helped paint

… mer d'chind^{n} em Hans^{m} es huus haend wele laa^{n} hälfe^{m} aastriche.
we the children.Acc the Hans.Dat the house.Acc have wanted let helped paint

$wa^{n}\, b^{m} x c^{n}\, d^{m} y$ so, not a regular language

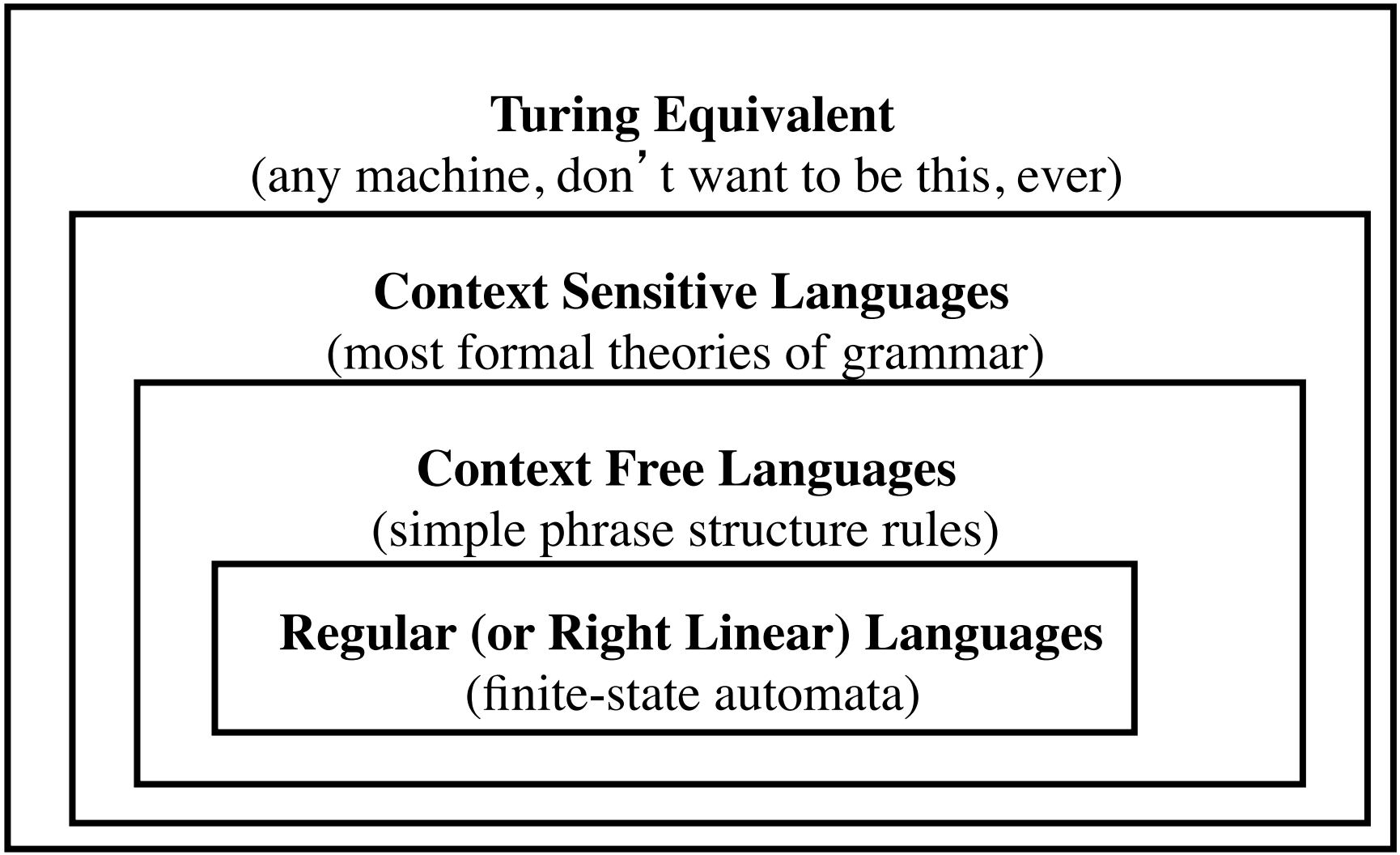
Natural Language

Is it context-free? No.

So, Natural Language turns out to be a very hard problem: an **NP-complete** problem (term from computer science).

Should we give up? No --- there are still ways to make things computable.

The Chomsky Hierarchy

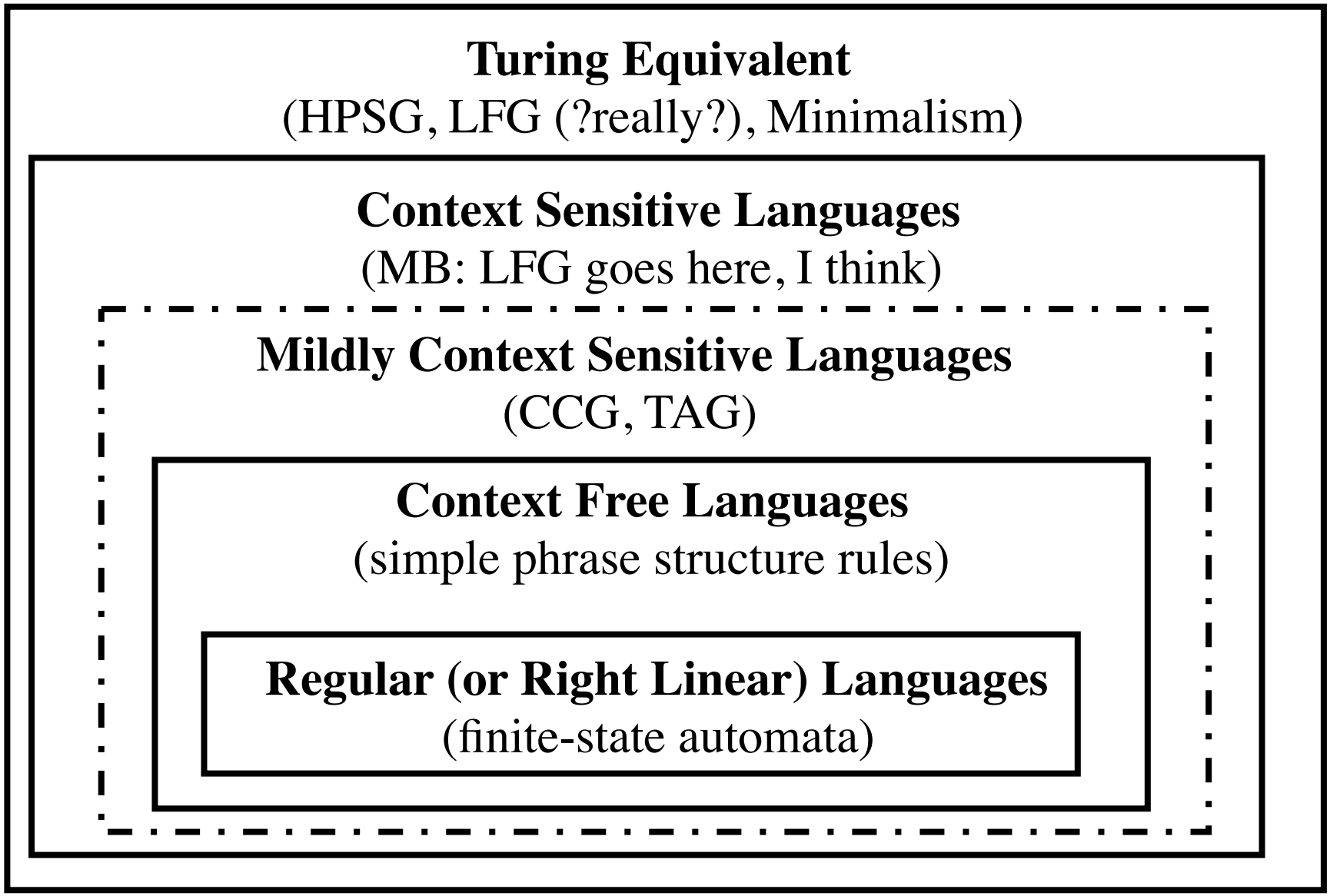


Chomsky Hierarchy via Rules
(cf. J&M p. 565)

Type	Rule Skeleton
Turing Equivalent	$\alpha \rightarrow \beta$, s.t. $\alpha \neq \varepsilon$
Context Sensitive	$\alpha A \beta \rightarrow \alpha \gamma \beta$, s.t. $\gamma \neq \varepsilon$
Context Free	$A \rightarrow \gamma$
Regular	$A \rightarrow \chi B$ or $A \rightarrow \chi$

Where A is a single non-terminal,
α, β, γ are arbitrary strings of terminal and non-terminal symbols

The Chomsky Hierarchy
(amended Version, cf. J&M)



Decidability

The more you know about the formal properties of an underlying syntactic theory, the better.

Montonicity: this basically means you do not overwrite information once you've got it as part of your analysis.

Mathematical Proofs: based on the properties of one's formal theory, one can prove whether it is *decidable* or not.

Decidability

The more you know about the formal properties of an underlying syntactic theory, the better.

GB/Minimalism: couched in a very formal way, but includes unconstrained movements, which makes it *non-monotonic* and puts it into the space of a Turing Machine.

HPSG: formal properties still under debate and an active area of research (e.g., Lexical Rules).

LFG: formal properties well understood and has been proven to be decidable (Kaplan and Bresnan 1982, Backofen 1993).

Decidability

"First, an explanatory linguistic theory undoubtedly will impose a variety of substantive constraints on how our formal devices may be employed in grammars of human languages. ... It is quite possible that the worst case computational complexity for the subset of lexical-functional grammars that conform to such constraints will be plausibly sub-exponential." [Kaplan and Bresnan 1982]

In practice, one can (and does) also come up with smart computational techniques that avoide the worst-case scenario.