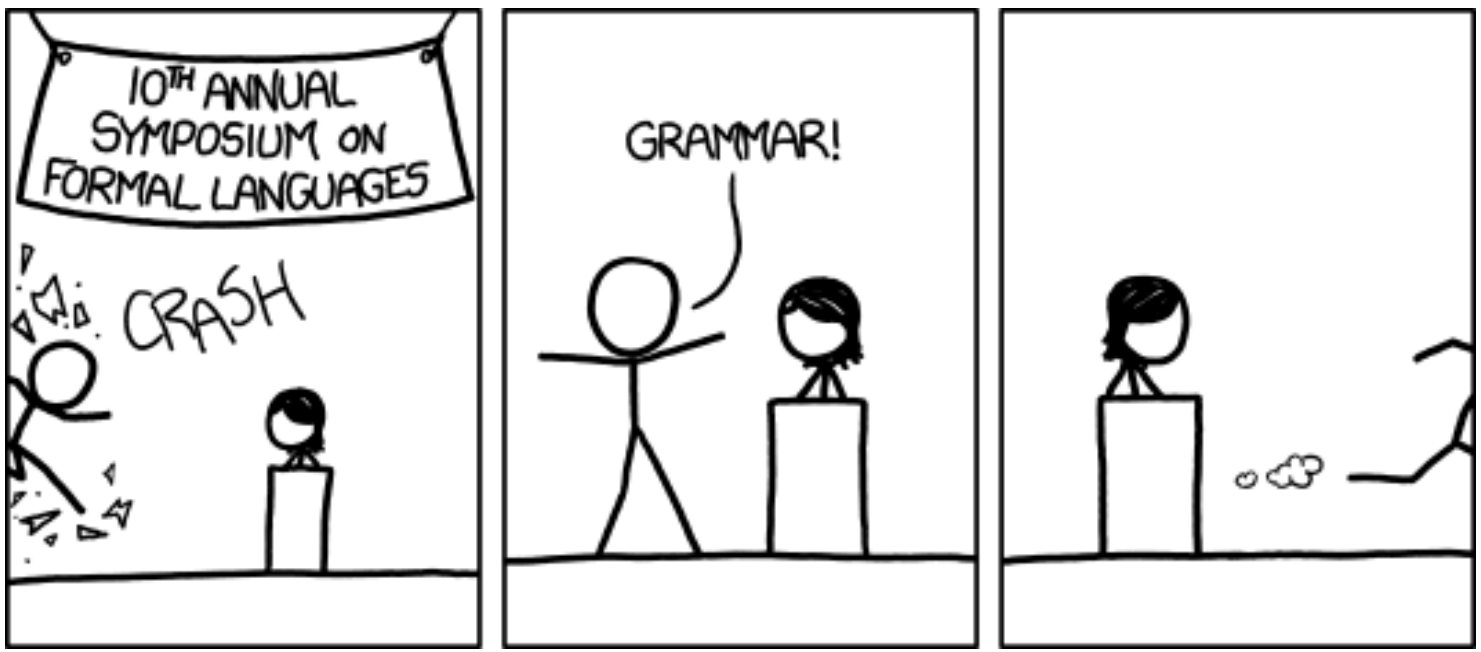
Context Free Grammars (& English)

Linguistics 409 · Computational Linguistics



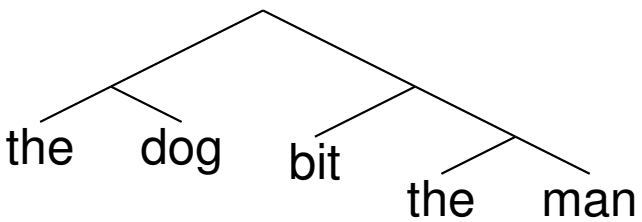
alt-text: [audience looks around] ‘What just happened?’ ‘There must be some context we’re missing.’

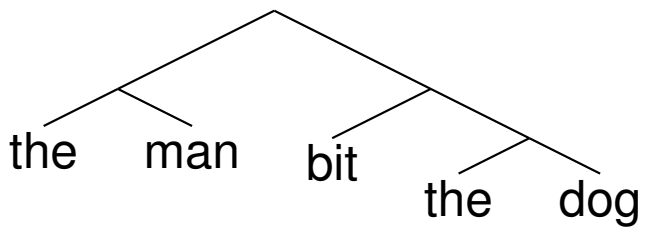
Sentences have syntactic structure.

were few hardest journey minutes of probe's the the the last

man dog the the bit

arf!





Parse

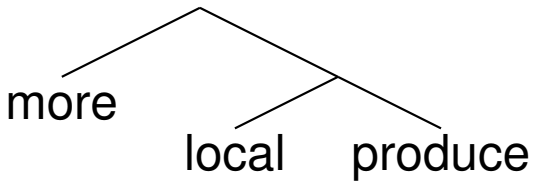
Buy more local produce.

Parse

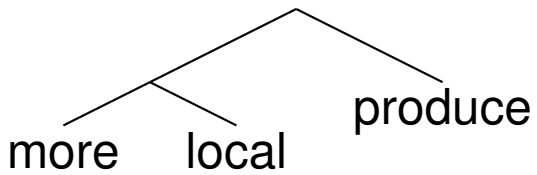
Buy more **local produce**.

Buy **more local** produce.

produce produce



“more produce that is locally grown”



“produce that is grown closer to here”

Constituency

Buy more local produce

- What matters for meaning is how *closely related* the words in this sentence are.
- Words that function together are called a **constituent**.
- We have a number of tests for constituency (some better than others)

Tests for constituency

constituent is *always* a contiguous string of words.

- No skipping words in a constituent.
- Syntactic operations can only be performed on constituents.
- This means we can use these same ‘syntactic operations’ to identify constituents.

Tests for constituency

1. Stand Alone Test
2. Substitution
3. Movement
4. Coordination

Tests for constituency: Stand Alone Test

A constituent can often stand alone as the answer to a question.

- What can stand alone as the answer to a question?
- What can a constituent do?
- Stand alone as the answer to what?

Tests for constituency: Stand Alone Test

A constituent can often stand alone as the answer to a question.

- What can stand alone as the answer to a question?
- What can a constituent do?
- Stand alone as the answer to what?

Tests for constituency: Stand Alone Test

A constituent can often stand alone as the answer to a question.

- What can stand alone as the answer to a question?
- What can a constituent do?
- Stand alone as the answer to what?

Tests for constituency:Substitution

A constituent can often be replaced by a single word.

- *They* can often be replaced by a single word.
- A constituent *does*.
- A constituent can often be replaced by *one*.

Tests for constituency:Substitution

A constituent can often be replaced by a single word.

- *They* can often be replaced by a single word.
- A constituent *does*.
- A constituent can often be replaced by *one*.

Tests for constituency:Substitution

A constituent can often be replaced by a single word.

- *They* can often be replaced by a single word.
- A constituent *does*.
- A constituent can often be replaced by *one*.

Tests for constituency: Movement

Constituents can sometimes be moved as units to change emphasis (but not meaning).

- To change emphasis, constituents can sometimes be moved as units.
- As units, constituents can sometimes be moved to change emphasis.
- Sometimes, constituents can be moved to change emphasis.
- It is constituents that can sometimes be moved as units to change emphasis.

Tests for constituency: Coordination

The coordinating conjunctions *and* and *or* must conjoin equal units. This means you can try balancing a unit you think is a constituent with something you know already to be a constituent. Try using *yelled* to test these potential constituents:

- The meerkats *invaded enemy territory*.
- The referee *called back the goal again*.
- Every man *kills the thing he loves*.

Tests for constituency: Coordination

The coordinating conjunctions *and* and *or* must conjoin equal units. This means you can try balancing a unit you think is a constituent with something you know already to be a constituent. Try using *yelled* to test these potential constituents:

- The meerkats *invaded enemy territory*.
- The referee *called back the goal again*.
- Every man *kills the thing he loves*.

Tests for constituency: Coordination

The coordinating conjunctions *and* and *or* must conjoin equal units. This means you can try balancing a unit you think is a constituent with something you know already to be a constituent. Try using *yelled* to test these potential constituents:

- The meerkats *invaded enemy territory*.
- The referee *called back the goal again*.
- Every man *kills the thing he loves*.

Tests for constituency: Are these constituents?

- Pam saw the boy with a telescope
- Johan's head feels better.
- Chris stopped all the *shots easily*.
- She said "eh, I know you and you cannot sing".
- I said, "that's nothing, you should hear me play piano".

We might notice that certain constituent types emerge in our testing.

Noun Phrases (aka NP) can be subjects, can be objects, can be replaced by pronouns, can answer questions like 'who?' or 'I drank what?'

Verb Phrases (aka VP) seem to assign **grammatical relations** to the noun phrases (a subject is the SUBJECT because of the verb, an object is the OBJECT because of the verb!), can be replaced by a single known-verb (e.g. *yelled*), etc.

Prepositional Phrases start with a preposition, always seems to contain an NP, seem to be attachable to Noun Phrases or Verb Phrases depending on meaning, etc.

Many others...

Two types of structure immediately emerge when we classify constituents into types...

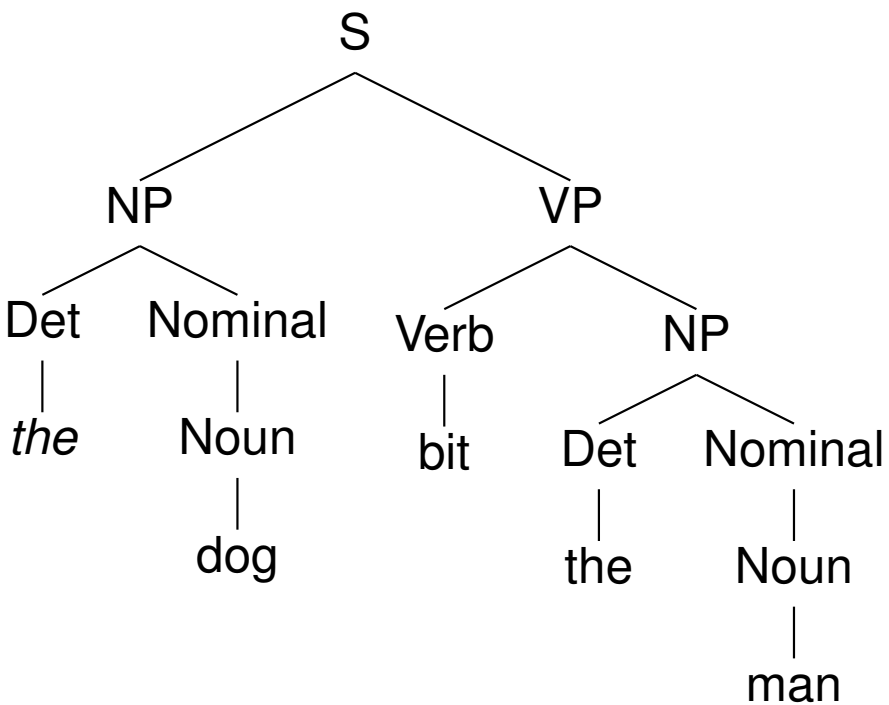
1. These constituents or phrases have internal structure!

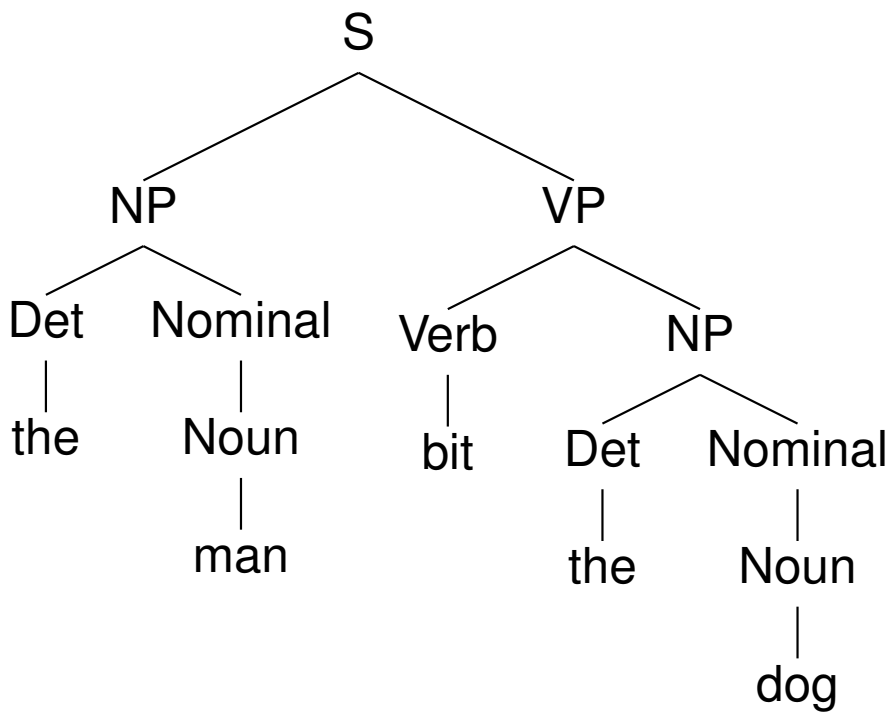
   NP $\rightarrow$ *Det Nominal*
   NP $\rightarrow$ *ProperNoun*
   NP $\rightarrow$ *Pronoun*
   Nominal $\rightarrow$ *Noun | Nominal Noun PP* $\rightarrow$ *Preposition NP*

2. And the constituents can only be combined in certain ways.

   e.g. English sentences seem to tend to be Subject first and Verb second
   So we'll want to be able to predict something like:
   Sentence $\rightarrow$ *NP VP*

arf!


S
NP
VP
Det
Nominal
Verb
NP
the
Noun
bit
Det
Nominal
dog
the
Noun
man


S
NP
VP
Det
Nominal
Verb
NP
the
Noun
bit
Det
Nominal
man
the
Noun
dog

Sentence Structure

Grammar Rules			Examples
S	→	*NP VP*	I + want a morning flight
NP	→	*Pronoun*	I
	|	*Proper-Noun*	Los Angeles
	|	*Det Nominal*	a + flight
Nominal	→	*Nominal Noun*	morning + flight
	|	*Noun*	flights
VP	→	*Verb*	do
	|	*Verb NP*	want + a flight
	|	*Verb NP PP*	leave + Boston + in the morning
	|	*Verb PP*	leaving + on Thursday
PP	→	*Preposition NP*	from + Los Angeles

Jurafsky & Martin 12.3

Lexicon (Non-terminals $rightarrow terminals)

Noun	$\rightarrow$	flights | breeze | trip | morning
Verb	$\rightarrow$	is | prefer | like | need | want | fly
Adjective	$\rightarrow$	cheapest | non-stop | first | latest | other | direct
Pronoun	$\rightarrow$	me | I | you | it | they
ProperNoun	$\rightarrow$	Alaska | Baltimore | Los Angeles | Chicago | Southwest | Morrissey
Determiner	$\rightarrow$	the | a | an | this | these | that
Preposition	$\rightarrow$	from | to | on | near | above | through
Conjunction	$\rightarrow$	and | but | or

As with FSAs and FSTs, you can view these rules as either analysis or synthesis machines:

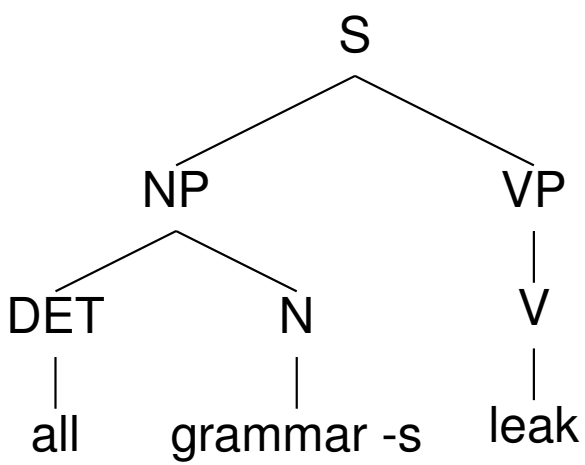
1. Generate (ideally all of) the strings in the language
2. Reject (ideall all of the) strings **not** in the language
3. Impose hierarchical structures (trees) on strings in the language

Derivations

A derivation is a sequence of rules applied to a string that accounts for that string

- Covers all the elements in the string
- Covers only the elements in the string

"All grammars leak." (Sapir 1921)



For next time:

For next time:

1. Next time we'll talk about English, TreeBanks and some more about CFGs
2. Keep reading chapter 12 of Jurafsky and Martin

English!

1. Sentences
2. Noun Phrases (agreement)
3. Verb Phrases (subcategorization)

Sentence Types

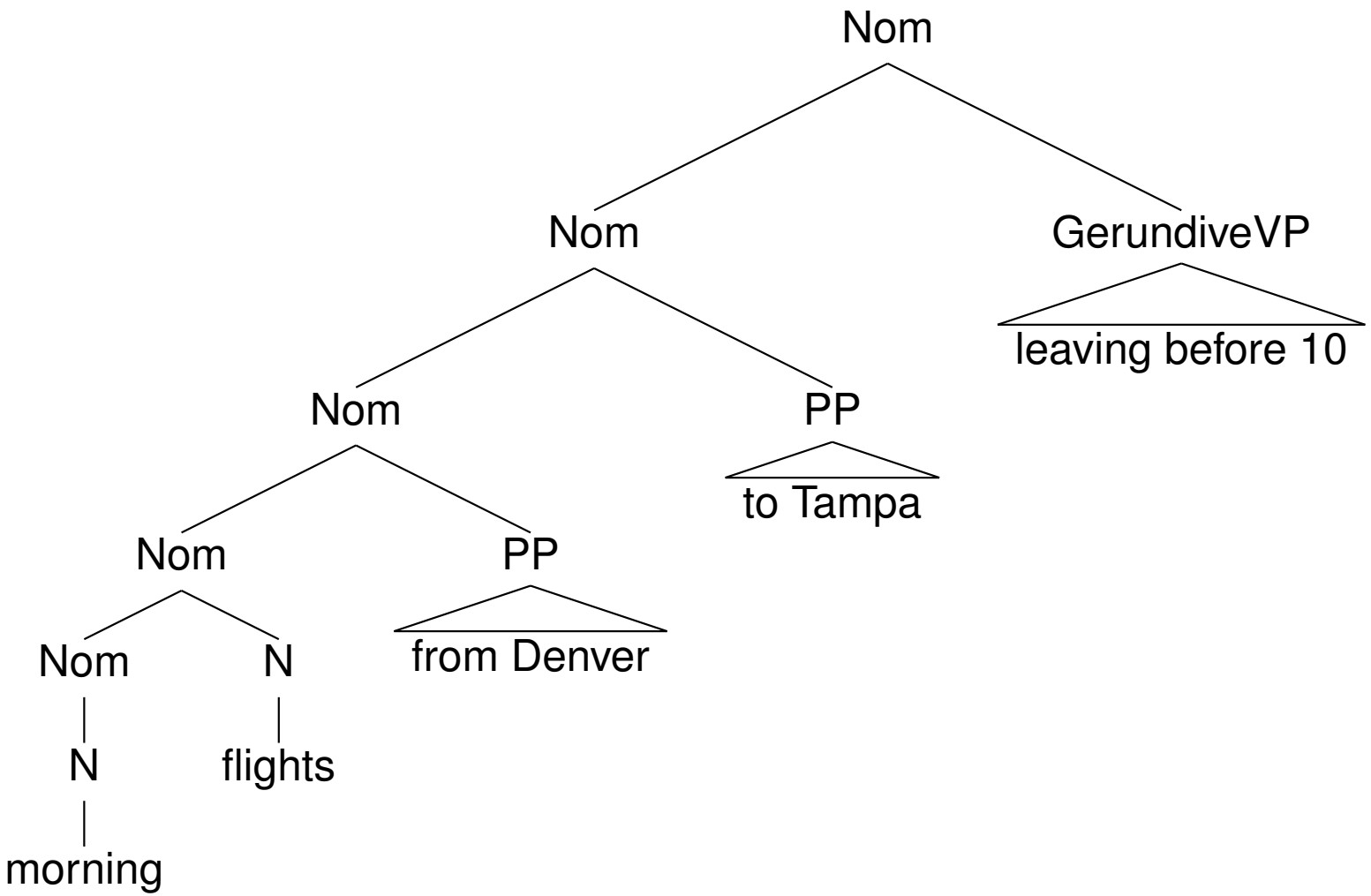
Declaratives	A plane left. S $\rightarrow$ NP VP
Imperatives	Leave! S $\rightarrow$ VP
Yes-No Qs	Did the plane leave? S $\rightarrow$ Aux NP VP
WH Qs	When did the plane leave? S $\rightarrow$ WH-NP Aux NP VP

Noun Phrases

1. NP $\rightarrow$ Det Nominal
2. Most of the complexity of English noun phrases is hidden in this rule.
3. Consider the derivation for the following example:

All the morning flights from Denver to Tampa leaving before 10.

Just one NP: [.NP [.Det All the [.Nom ...]]



Noun Phrase Structure

- Clearly this NP is really about *flights*. That's the central critical noun in this NP. Let s call that the **head**.
- We can dissect this kind of NP into the stuff that can come before the head, and the stuff that can come after it.

Determiners

- Noun phrases can start with determiners...
- Determiners can be
  - Simple lexical items: the, this, a, an, etc.
    A car
  - Or simple possessives
    John's car
  - Or complex recursive versions of that
    John's sister's husband's son's car

Nominal

- Contains the head and any pre- and post- modifiers of the head.
  - Quantifiers, cardinals, ordinals...
    Three cars
  - Adjectives and APs
    large cars
  - Ordering constraints
    Three large cars
    ?large three cars

Post Modifiers

- Three kinds:
  1. Prepositional phrases
     From Seattle
  2. Non-finite clauses
     Arriving before noon
  3. Relative clauses
     That serve breakfast
- Same general (recursive) rule to handle these:

  Nominal $\rightarrow$ Nominal PP
  Nominal $\rightarrow$ Nominal GerundiveVP
  Nominal $\rightarrow$ Nominal RelClause

Agreement

- Constraints that hold among various constituents that take part in a rule or set of rules
- For example, in English, determiners and the head nouns in NPs have to agree in their number.

Sg N head	Pl N head
A flight	*A flights
This koala	*This koalas
*Those banana	those bananas
*Five cow	Five cows

Problem

- Our earlier NP rules are clearly deficient – they don't capture this constraint!
  NP $\rightarrow$ Det Nominal
- Accepts, and assigns correct structures to, grammatical examples (this flight)
- But it's also happy with incorrect examples (*these flight)
- Such a rule is said to **overgenerate**.

Verb Phrases

English VPs consist of a head verb along with 0 or more following constituents which we'll call arguments.

$VP \rightarrow Verb$ disappear

$VP \rightarrow Verb\ NP$ prefer a morning flight

$VP \rightarrow Verb\ NP\ PP$ leave Boston in the morning

$VP \rightarrow Verb\ PP$ leaving on Thursday

Subcategorization

But, even though there are many valid VP rules in English, not all verbs are allowed to participate in all those VP rules.
We can **subcategorize** the verbs in a language according to the sets of VP rules that they participate in.
This is a modern take on the traditional notion of transitive/intransitive.
Modern grammars may have 100s of such classes.

Subcategorization

Sneeze	John sneezed
Find	Please find [a flight to NY]NP
Give	Give [me]NP[a cheaper fare]NP
Help	Can you help [me]NP[with a flight]PP
Prefer	I prefer [to leave earlier]TO-VP
Told	I was told [United has a flight]S
...	

Subcategorization

- *John sneezed the book
- *I prefer United has a flight
- *Give with a flight

As with agreement phenomena, we need a way to formally express the constraints.

Overgeneration

Right now, our various rules for VPs overgenerate.
They permit the presence of strings containing verbs and arguments that don't go together.

For example:

$$VP \rightarrow V\ NP$$

therefore Sneezed the book is a VP since 'sneeze' is a verb and 'the book' is a valid NP

Possible CFG Solution

Possible solution for agreement.

Can use the same trick for all the verb/VP classes.

SgS $\rightarrow$ SgNP SgVP
PlS $\rightarrow$ PlNp PlVP
SgNP $\rightarrow$ SgDet SgNom
PlNP $\rightarrow$ PlDet PlNom
PlVP $\rightarrow$ PlV NP
SgVP $\rightarrow$ SgV Np
...

Possible CFG Solution

1. It works and stays within the power of CFGs
2. But it's ugly
3. And it doesn't scale all that well because of the interaction among the various constraints –explodes the number of rules in our grammar!

The Takeaway Message

- CFGs appear to be *just about* what we need to account for a lot of basic syntactic structure in English.
- But there are problems
- These can be dealt with adequately, albeit inelegantly, within the CFG framework. There are simpler, more elegant, solutions that take us out of the CFG framework (beyond its formal power)
- LFG, HPSG, Construction grammar, XTAG, etc.

Treebanks

Treebanks are corpora in which each sentence has been paired with a parse tree (presumably the right one).

- These are generally created by first:
- parsing the collection with an automatic parser
- And then having human annotators correct each parse as necessary.
- This generally requires detailed annotation guidelines that provide a POS tagset, a grammar and instructions for how to deal with particular grammatical constructions.

Treebanks

Treebanks implicitly define a grammar for the language covered in the treebank.

- Simply take the local rules that make up the sub-trees in all the trees in the collection and you have a grammar.
- Not complete, but if you have decent size corpus, you'll have a grammar with decent coverage.

Treebank Uses

- Treebanks are particularly critical to the development of statistical parsers.
- The Penn Treebank's Wall Street Journal section (1,000,000 words from 1987 - 1989) is often the gold standard in statistical parsing.
- Also valuable to Corpus Linguistics. Investigating the empirical details of various constructions in a given language

For next time:

For next time:

1. Next time we'll start statistical parsing.
2. Start reading chapter 13 of Jurafsky and Martin