JMORF — Morpho-Syntax

First attempts at a theory of grammar

Francis Bond
Palacký University
https://fcbond.github.io/
bond@ieee.org

Lecture 1b
Location: SV 2.39

Overview

- Some initial attempts to model grammar
  - Context Free Grammars
  - Central claims of CFG
-

What makes a good model?

- **generative**: license all grammatical sentences and only them
  ⇒ **precise**

- **explanatory**: can explain generalizations
  - *the cat chased the rat* ∼ *the rat was chased by the cat* (semantics)
  - phrases tend to act like one member of the phrase (headedness)
  - new information tends to come first/last (information theory)

- **concise**: the model is as simple as possible (elegant)
  ⇒ **universal** (minimal stipulations)

- **tractable**: the model can be modeled computationally

Our models are normally imperfect:
we aim for iteratively improved approximations

Identifying constituents

- Language structure is **hierarchical**, and is made up of identifiable constituents
- Some words go together more closely to form a constituent, e.g. noun phrases and postpositional phrases.
- Such strings form **immediate constituents** of units higher up in the hierarchy.
- Constituents are identified on the basis of **formal criteria ...**

Formal criteria for identifying constituents

Three things to consider:

- **Distribution**: If the same sequence of constituents occurs repeatedly, this sequence might be a constituent.
- **Substitution**: If we can substitute a sequence of words by a single word, keeping the reference more or less the same, then this sequence is probably a constituent.
- **Mobility**: If we can move a sequence of constituents around in a sentence, and they have to move together, then this is probably a constituent

Insufficient Theory #1

- A grammar is simply a list of sentences.
- What's wrong with this?

Insufficient Theory #2: Regular Expressions

(1) *the noisy dogs left*
D A N V

(2) *the noisy dogs chased the innocent cats*
D A N V D A N

- (D) A* N V ((D) A* N)

Regular expressions: a formal language for matching things.

Symbol	Matches
.	any single character
$*$	the preceding element zero or more times.
?	the preceding element zero or one time: OR just () = ()?.
$+$	the preceding element one or more times.
\| | either the expression before or after the operator. |

Context-Free Grammar

- A quadruple: $\langle C, V, P, S \rangle$

  C set of categories $(\alpha, \beta, \ldots)$
  V set of terminals (vocabulary)
  P set of rewrite rules $\alpha \rightarrow \beta_1, \beta_2, \ldots, \beta_n$
  S the start symbol $\mathbf{S} \in C$

- For each rule $\alpha \rightarrow \beta_1, \beta_2, \ldots, \beta_n \in P$

  - $\alpha \in C$
  - $\beta_i \in C \cup V; 1 \leq i \leq n$

A Toy Grammar

- RULES

S	$\rightarrow$	NP VP
NP	$\rightarrow$	(D) A* N PP*
VP	$\rightarrow$	V (NP) (PP)
PP	$\rightarrow$	P NP

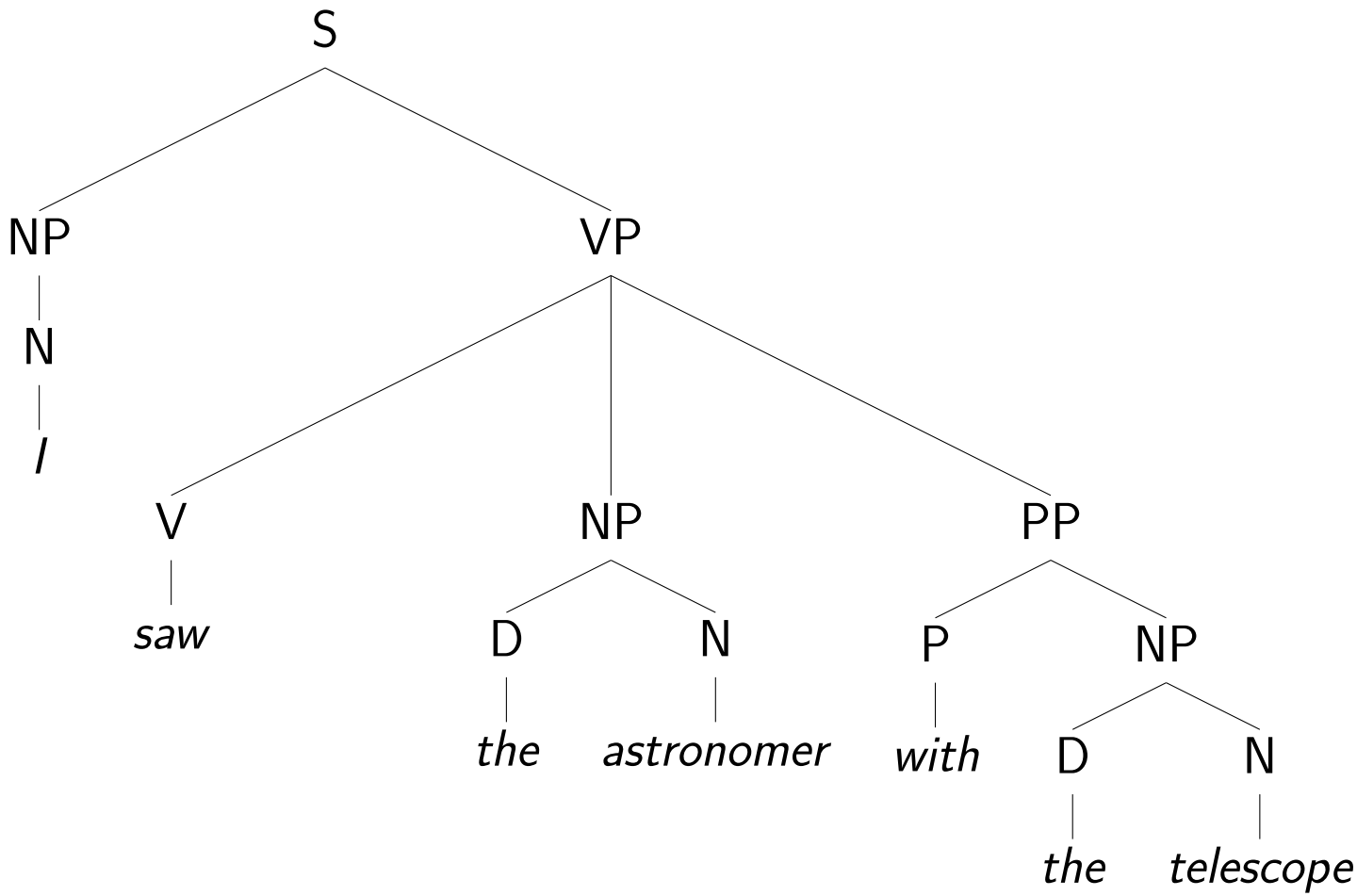
- VOCABULARY

  D: the, some
  A: big, brown, old
  N: birds, fleas, dog, hunter, I
  V: attack, ate, watched
  P: for, beside, with

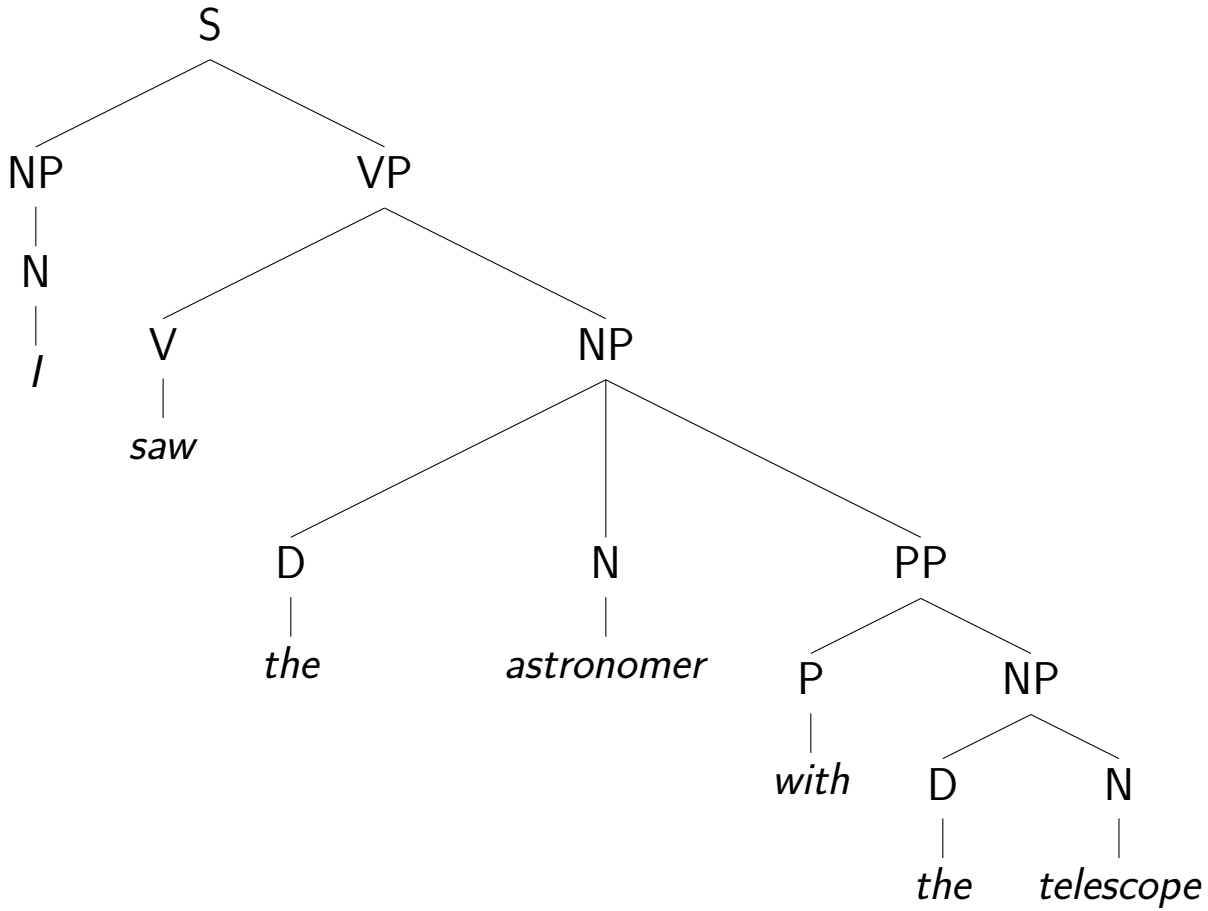
Structural Ambiguity

I saw the astronomer with the telescope.

Structure 1: PP under VP



Structure 2: PP under NP



Constituency Tests

- Recurrent Patterns

  (3) *The quick brown fox with the bushy tail jumped over the lazy brown dog with one ear.*

- Coordination

  (4) *The quick brown fox with the bushy tail and the lazy brown dog with one ear are friends.*

- Sentence-initial position

  (5) *The election of 2000, everyone will remember for a long time.*

- Cleft sentences

  (6) *It was a book about syntax that they were reading.*

General Types of Constituency Tests

- Distributional
- Intonational
- Semantic
- Psycholinguistic

... but they don't always agree.

Central claims implicit in CFG formalism:

1. Parts of sentences (larger than single words) are linguistically significant units, i.e. phrases play a role in determining meaning, pronunciation, and/or the acceptability of sentences.

2. Phrases are contiguous portions of a sentence (no discontinuous constituents).

3. Two phrases are either disjoint or one fully contains the other (no partially overlapping constituents).

4. What a phrase can consist of depends only on what kind of a phrase it is (that is, the label on its top node), not on what appears around it.

- Claims 1-3 characterize what is called **phrase structure grammar**

- Claim 4 (that the internal structure of a phrase depends only on what type of phrase it is, not on where it appears) is what makes it **Context-Free**.

- **Context-Sensitive Grammar** (CSG) gives up 4. That is, it allows the applicability of a grammar rule to depend on what is in the neighboring environment. So rules can have the form:
  $A \rightarrow X$ in the context of α_β ($\alpha A \beta \rightarrow \alpha X \beta$)

Possible Counterexamples

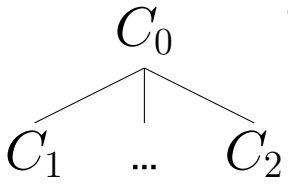
- To Claim 2 (no discontinuous constituents):
  A technician arrived who could solve the problem.

- To Claim 3 (no overlapping constituents):
  I read what was written about me.

- To Claim 4 (context independence):

  (7) *He arrives this morning.*
  (8) **He arrive this morning.*
  (9) **They arrives this morning.*
  (10) *They arrive this morning.*

Trees and Rules

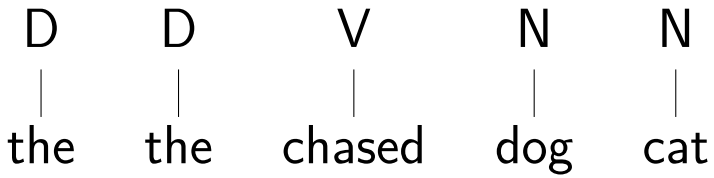


is a well-formed nonlexical tree if (and only if)

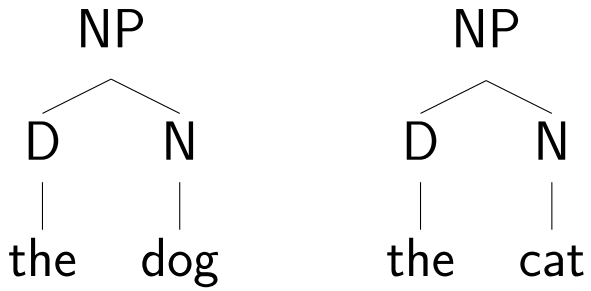
- $C_0, \ldots, C_n$ are well-formed trees
- $C_0 \rightarrow C_1 \ldots C_n$ is a grammar rule

Bottom-up Tree Construction

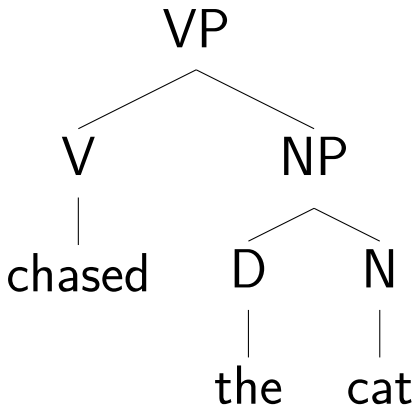
D: the
V: chased
N: dog, cat



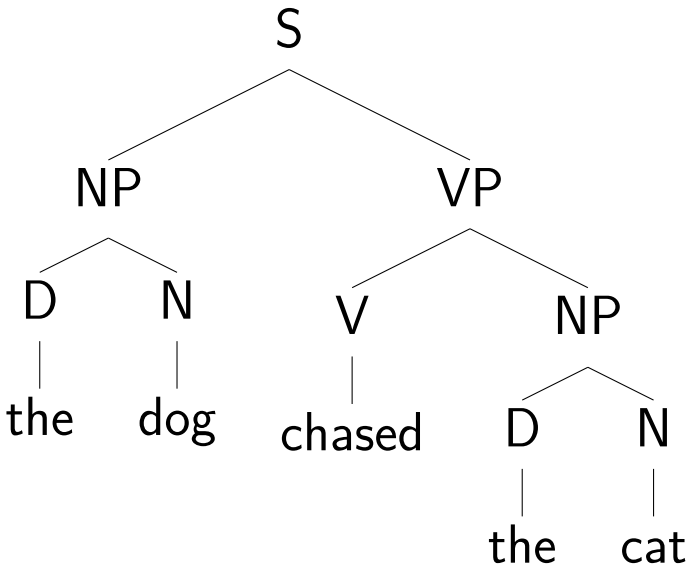
$NP \rightarrow D\ N$



$VP \rightarrow V\ NP$

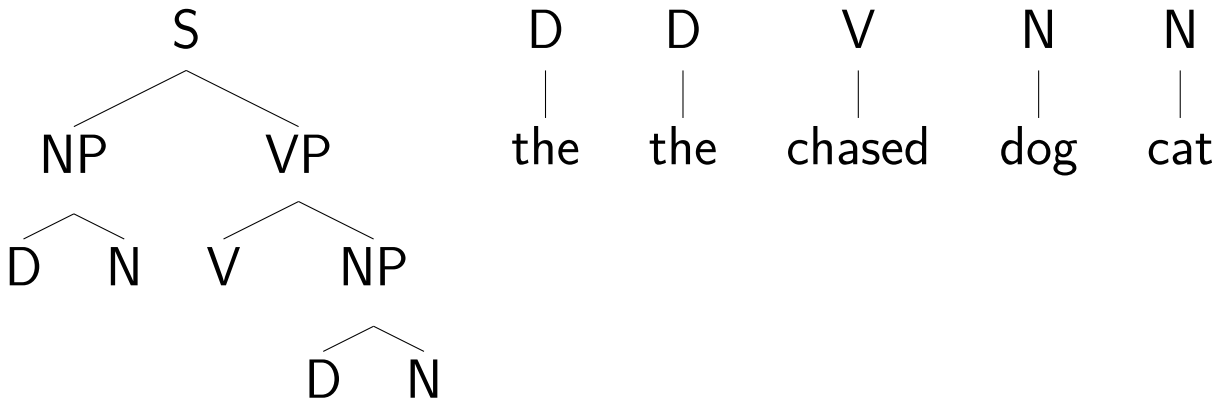


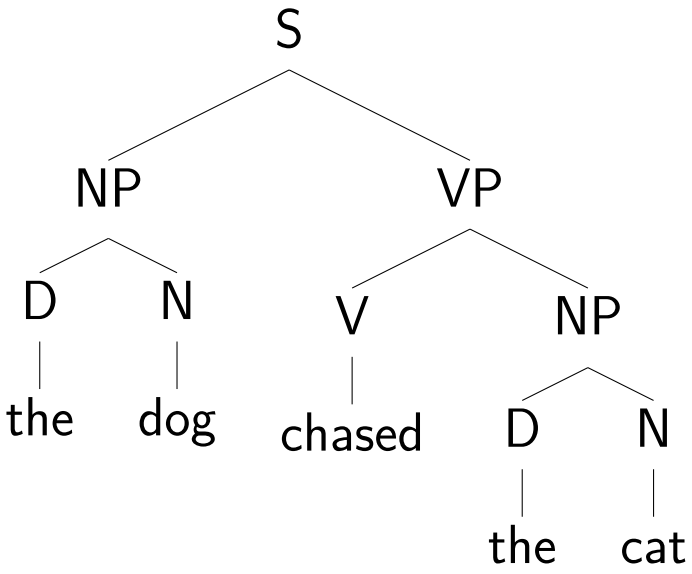
$S \rightarrow NP\ VP$



Top-down Tree Construction

$S \rightarrow NP\ VP$ $VP \rightarrow V\ NP$ $NP \rightarrow D\ N$ $NP \rightarrow D\ N$





- **Bottom-up**: string $\rightarrow$ tree
- **Top-down**: tree $\rightarrow$ string
- CFG is **declarative** so it is independent of order

Weaknesses of CFG (atomic node labels)

- It doesn't tell us what constitutes a linguistically natural rule
  - VP $\rightarrow$ P NP
  - NP $\rightarrow$ VP S
- Rules get very cumbersome once we try to deal with things like agreement and transitivity.
- It has been argued that certain languages (notably Swiss German and Bambara) contain constructions that are provably beyond the descriptive capacity of CFG.

On the other hand ...

- It's a simple formalism that can generate infinite languages and assign linguistically plausible structures to them.
- Linguistic constructions that are beyond the descriptive power of CFG are rare.
- It's computationally tractable and techniques for processing CFGs are well understood.

So ...

- CFG is the starting point for most types of generative grammar.
- The theory we develop in this course is an extension of CFG.

Transitivity and Agreement

- Consider the following transitivity examples

  (11) *The bird arrives*
  (12) *The bird devours the worm*
  (13) **The bird arrives the worm*
  (14) **The bird devours*

- Consider the following agreement examples

  (15) *The bird sings*
  (16) *The birds sing*
  (17) **The bird sing*
  (18) **The birds sings*

- Can we deal with them with a CFG?

Basic Concepts of Morphology: Morphological Models

- **Item-and-Arrangement (IA)**: list morphemes + specify their linear arrangement. *spiral + -ize + -er*.
- **Item-and-Process (IP)**: items undergo processes/rules (*make-verb*, *instrument-noun*).
- **Word-and-Paradigm (WP)**: words as **word-forms** in **paradigms**; realizational rules fill paradigm slots.
- Each model fits different language properties, ...

Based on Panocová (2021, Chapter 3), Czech examples added

IA: The Morpheme List + Arrangement

- Hockett: utterances = minimal meaningful elements (**morphemes**) arranged linearly.
- Focus: **free** vs. **bound** forms; segment-and-label.
- IA Worked Example
  - *spiralizer* ⇒ *spiral* + *-iz(e)* + *-er*.
  - Inventory entries: *spiral* (free), *-ize* (V-forming), *-er* (N-forming: instrument/agent).

IA Strengths and Weaknesses

- IA Strength: Transparent Affixation
  - Agglutinative-style sequences: **one function one form**.
  - Hungarian: suffix allomorphs by stem vowels (back vs. front).

olvas-	'read'	**yl-**	'sit'
olvas-unk	'we read'	**yl-ynk**	'we sit'
olvas-tok	'you (pl) read'	**yl-tyk**	'you (pl) sit'
olvas-nak	'they read'	**yl-nek**	'they sit'

- IA Challenges: When Segments Don't Behave
  - **Ablaut** pasts: *take* → *took*—no neat "past" morpheme.
  - **Cumulative exponence**: Czech *-u* in *ženu* = [accusative singular feminine].
  - Such patterns motivate IP/WP analyses.

IP: Items Undergo Processes

- Lexicon stores bases; morphology applies **processes** to derive outputs
- *spiral* → *spiralize* (*make-verb*)
  spiralize → *spiralizer* *instrument-noun*)
- IA vs. IP: What Counts as an "Item"?
  - **IA**: morphemes + concatenation.
  - **IP**: bases + processes (rules for alternations, allomorphy)

WP: Words and Paradigms

- Central unit: **word-form** (member of a lexeme's paradigm).
- Goal: **realizational rules** filling paradigm slots.
- Great for **syncretism**, rich inflection.
- WP in Action: Czech Masculine Nouns
  - Genitive plural is *-ů* across many declension types.

Lexeme	Gen. pl.
pán 'lord'	*pánů*
hrad 'castle'	*hradů*
muž 'man'	*mužů*
stroj 'machine'	*strojů*

  - Rule (simplified): $x_{N:Mc}$[+pl, +gen] $\Rightarrow$ x-*ů*$_N$.

Mini-paradigm Tables (Czech)

- *pán* 'lord' (paradigm: nominative–accusative–genitive plural):

Nom. sg.	*pán*
Acc. sg.	*pána*
Gen. pl.	*pánů*

- *hrad* 'castle' (hard-type masculine inanimate):

Nom. sg.	*hrad*
Acc. sg.	*hrad*
Gen. pl.	*hradů*

There can be many sub rules and irrgularities

Choosing the Right Lens

- **IA**: clean concatenation, transparent affixation, listable allomorphy (also influential in **Distributed Morphology**).
- **IP**: rules/processes for alternations, non-linear morphology, construction-based generalizations.
- **WP**: paradigm-centered patterns, syncretism, realizational statements over lexemes.
- Use the model that best fits the language property you're analyzing.

How Humans Handle Morphology

- We combine **rules** with **memory**.
- **Rules / Productivity**: apply general patterns (*talk* → *talked*, Cz. *hrad* → *hradů*).
- **Lexical Memory**: store irregulars (*go* → *went*, Cz. *oko* → *oči*) and regular high-frequency forms!
- **Dual-route model**: both operate in parallel.

Evidence: Rules + Exceptions Together

- **Analogy**: children overgeneralize (*goed*, Cz. *?hradové*).

- **Frequency effects**: common forms often stored whole, retrieved faster.
  - Even **regular** forms can show whole-word frequency effects at high frequency.
  - **English** *V+ed*: very frequent pasts (e.g., *looked*, *called*) behave like stored units rather than rule-built on the fly. (Alegre and Gordon, 1999).
  - **Dutch** *N+en*: high-frequency regular plurals show evidence of storage alongside parsing. (Baayen et al., 1997).

- **Statistical sensitivity**: humans track distribution of endings; in Czech, *-ů* becomes the **default** genitive plural despite exceptions.

- Result: humans learn to generalize where possible, memorize where needed, and adjust as experience grows.

Classification of grammatical categories

Languages do not differ so much in what you can say, but rather in *how* you must say it.

- **Inherent categories**: show a property related to the word class they are attached to. For example, plural in English.

- **Agreement categories**: show a property related to another word in the sentence. For example in English, an *-s* on the verb shows that the subject of the verb is 3rd person singular.

- **Relational categories**: show how a word fits into a larger structure. For example English *I* for nominative case or *me* for accusative case. The form is determined by the grammatical relation (subject or object) of the argument.

Different languages have different grammatical categories

There is lots of variation across the languages of the world.
Examples of nominal grammatical categories:

- person: 1st/2nd/3rd; inclusive exclusive distinctions
- number: singular/dual/plural
- noun class or gender
- case: core versus oblique
- definiteness/specificity

Verbal grammatical categories

- person
- number and gender of arguments as agreement categories
- temporal deixis
  1. tripartite systems: past/present/future
  2. binary systems: past/non-past; non-future/future
  3. metrical tense systems

- Aktionsart (lexical aspect): static, telic, punctual

- Aspect:

  1. perfective and imperfective
  2. progressive
  3. perfect

- Mood and modality: Realis versus irrealis; indicative, subjunctive, interrogative, imperative;

- Evidentiality

Czech verb derivational & inflectional morphology

	Czech	English gloss
a.	dát	'to give'
b.	dávat	'to keep giving'
c.	dám	'I will give'
d.	předáš	'you will hand over'
e.	prodáme	'we will sell (hand in exchange)'
f.	vyprodá	'she will sell off'
g.	vydám	'I will hand out, give out'
h.	povyprodáme	'we will sell out gradually'
i.	dáváš	'you keep giving'
j.	dopovyprodávášmi	'you will gradually finish selling (it) out for me'
k.	dopovyprodávámeti	'we will gradually finish selling (it) out for you'

What kind of information is encoded in the morphology?

Position-class diagrams: Czech verbs

	p4	p3	p2	p1	root	s1	s2	s3	s4
a.					dá		-t		
b.					dá	-va	-t		
c.					dá		-m		
d.				pře-	dá		-š		
e.				pro-	dá		-m	-e	
f.			vy-	pro-	dá				
g.				vy-	dá		-m		
h.		po-	vy-	pro-	dá		-m	-e	
i.					dá	-vá	-š		
j.	do-	po-	vy-	pro-	dá	-vá	-š		-mi
k.	do-	po-	vy-	pro-	dá	-vá	-m	-e	-ti
	ASP	ASP	ASP	DIR	root	ASP	S	NUM	O

Position-class diagrams: Czech verbs

prefix.4	prefix.3	prefix.2	prefix.1	root	suffix.1	suffix.2	suffix.3	suffix.4
ASP do- 'to'	ASP po- 'along'	ASP vy- 'out'	DIRECTION pro- 'for' pře- 'over' vy- 'out'	root dá 'give'	ITERATIVE va-∼-vá	SUBJ -m (1) -š (2sg) -0 (3sg)	NUMBER -e (PL)	OBJ =mi (1SG.DAT) =ti (2SG.DAT)

Morphology is full of special cases

- Not every root can take every affix
- The meaning changes are unpredictable
- There are often sound/spelling changes
- But you can see patterns everywhere!

How do we do linguistic analysis?

1. Learn the Fundamentals
2. Investigate
3. Find out some stuff
4. Break our theory
5. Try to fix it.
6. Break it again.
7. Lather, rinse, repeat: we'll do that until we run out of time.

Jorge Hankamer's outline of a syntax course, but it's pretty applicable to everything we do. More formally: Successive Approximation.

Chapter 2, Problem 1

RULES			VOCABULARY
S	$\rightarrow$	NP VP	D: a, the
NP	$\rightarrow$	(D) NOM	N: cat, dog, hat, man, woman, roof
VP	$\rightarrow$	V (NP) (NP)	V: admired, disappeared, put, relied
NOM	$\rightarrow$	N	P: in, on, with
NOM	$\rightarrow$	NOM PP	CONJ: and, or
VP	$\rightarrow$	VP PP	
PP	$\rightarrow$	P NP	
X	$\rightarrow$	X+ CONJ X	

Chapter 2, Problem 1

A Make a well-formed English sentence unambiguous according to this grammar

B Make a well-formed English sentence ambiguous according to this grammar: draw trees

C Make a well-formed English sentence not licensed by this grammar (using V)

D Why is this (C) not licensed?

E Make a string licensed by this grammar that is not a well-formed English sentence

F How can we stop licensing the string in E (stop over-generating)

G How many strings does this grammar license?

H How many strings does this grammar license without conjunctions?

Shieber 1985

- Swiss German example:

(19) *...mer d'chind em Hans es huus lönd hälfe aastriiche*
...we the children-acc Hans-dat the hous-acc let help paint
we let the children help Hans paint the house

- Cross-serial dependency:
  - *lönd* "let" governs case on *d'chind* "children"
  - *hälfe* "help" governs case on *Hans* "Hans"
  - *aastriiche* "paint" governs case on *huus* "house"

- This cannot be modeled in a context free language

Strongly/weakly CF

- A language is weakly context-free if the set of strings in the language can be generated by a CFG.
- A language is strongly context-free if the CFG furthermore assigns the correct structures to the strings.
- Shieber's argument is that SW is not weakly context-free and therefore not strongly context-free.
- Bresnan et al (1983) had already argued that Dutch is strongly not context-free, but the argument was dependent on linguistic analyses.

Overview

- Formal definition of CFG
  - Constituency, ambiguity, constituency tests
  - Central claims of CFG
  - Order independence
  - Weaknesses of CFG
- Next Week: Feature structures

Acknowledgments and References

- Course design and slides borrow heavily from Emily Bender's course: *Linguistics 566: Introduction to Syntax for Computational Linguistics*
  http://courses.washington.edu/ling566

- Thanks to Alex Coupe for some inspirational slides

- Stuart M. Shieber. (1985) Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333-343

References

Maria Alegre and Peter Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40(1):41–61. URL https://www.tc.columbia.edu/faculty/pg328/faculty-profile/files/AlegreGordonJML99RegInfl.pdf.

R. Harald Baayen, Ton Dijkstra, and Robert Schreuder. 1997. Singulars and plurals in dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1):94–117. URL https://www.mpi.nl/publications/item3005386/singulars-and-plurals-dutch-evidence-parallel-dual-route-model.