

Computational Linguistics

CSC 485/2501
Fall 2025

5A

5a. Extending grammars with features

Gerald Penn

Department of Computer Science, University of Toronto

Reading: Jurafsky & Martin: 12.3.4–6, 15.0–3;
[Allen: 4.1–5]; Bird et al: 9.

Copyright © 2017 Suzanne
Stevenson, Graeme Hirst
and Gerald Penn. All rights
reserved.

Agreement and inflection

- Problem: **Agreement** phenomena.

*Nadia {washes/***wash**} the dog.*

*The boys {***washes**/wash} the dog.*

*You {***washes**/wash} the dog.*

- **Morphological inflection** of verb must match subject noun in person and number.

Subject–verb agreement 1

Present tense

	Singular	Plural
1	<i>I wash</i>	<i>we wash</i>
2	<i>you wash</i>	<i>you wash</i>
3	<i>he/she/it washes</i>	<i>they wash</i>
1	<i>I am</i>	<i>we are</i>
2	<i>you are</i>	<i>you are</i>
3	<i>he, she, it is</i>	<i>they are</i>

Subject–verb agreement 2

Past tense

	Singular	Plural
1	<i>I washed</i>	<i>we washed</i>
2	<i>you washed</i>	<i>you washed</i>
3	<i>he, she, it washed</i>	<i>they washed</i>
1	<i>I was</i>	<i>we were</i>
2	<i>you were</i>	<i>you were</i>
3	<i>he, she, it was</i>	<i>they were</i>

Agreement features 1

- English agreement rules are fairly simple.
 - Subject : verb w.r.t. person and number.
 - No agreement required between verb and object.
- Many languages have other agreements.
 - *E.g.*, German: Article and adjective endings depend on noun gender and case:

Agreement features 2

Nominative Case (Subject Case)

Masculine

der

der neu**e** Wagen

the new car

Feminine

die

die schön**e** Stadt

the beautiful city

Neuter

das

das alte**n** Auto

the old car

Plural

die

die neu**en** Bücher

the new books

Masculine

ein

ein neu**er** Wagen

a new car

Feminine

eine

eine schön**e** Stadt

a beautiful city

Neuter

ein

ein alte**s** Auto

an old car

Plural

keine

keine neu**en** Bücher

no new books

Ask about.com: German language: Adjective endings I and II.
<http://german.about.com/library/weekly/aa030298.htm> and
[aa033098.htm](http://german.about.com/library/weekly/aa033098.htm)

Agreement features 2

Accusative Case (Direct Object)

Masculine	Feminine	Neuter	Plural
den	die	das	die

den neu**en** Wagen
the new car

die schön**e** Stadt
the beautiful city

das alt**e** Auto
the old car

die neu**en** Bücher
the new books

Masculine	Feminine	Neuter	Plural
einen	eine	ein	keine

einen neu**en** Wagen
a new car

eine schön**e** Stadt
a beautiful city

ein alt**es** Auto
an old car

keine neu**en** Bücher
no new books

Ask about.com: German language: Adjective endings I and II.
<http://german.about.com/library/weekly/aa030298.htm> and
[aa033098.htm](http://german.about.com/library/weekly/aa033098.htm)

Agreement features 3

E.g., Chinese: Numeral classifiers, often based on shape, aggregation, ...:

两条鱼	<i>liang tiao yu</i> ‘two CLASSIF-LONG-ROPELIKE fish’
两条河	<i>liang tiao he</i> ‘two CLASSIF-LONG-ROPELIKE rivers’
两条腿	<i>liang tiao tui</i> ‘two CLASSIF-LONG-ROPELIKE legs’
两条裤子	<i>liang tiao kuzi</i> ‘two CLASSIF-LONG-ROPELIKE pants’
两只胳膊	<i>liang zhi gebo</i> ‘two CLASSIF-GENERAL arms’
两件上衣	<i>liang jian shangyi</i> ‘two CLASSIF-CLOTHES-ABOVE-WAIST tops’
两套西装	<i>liang tao xizhuang</i> ‘two CLASSIF-SET suits’

Zhang, Hong (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, 16(1), 43–59. Thanks also to Tong Wang, Vanessa Wei Feng, and Helena Hong Gao.

Agreement features 1

- English agreement rules are fairly simple.
- Many languages have other agreements.
- Some languages have multiple grammatical genders.
 - E.g. Chichewa has genders for men, women, bridges, houses, diminutives, men inside houses, etc. Between 12-18 in total.
- Some languages overtly realize many of these distinctions.
 - E.g. some Hungarian verbs have as many as 4096 inflected forms.

Inflectional morphology

- Word may be inflected ...
 - ... to indicate paradigmatic properties, e.g. singular / plural, past / present, ...
 - ... to indicate some (other) semantic properties
 - ... to agree with inflection of other words.
- Each (open-class) word-type has a **base form / stem / lemma**.
- Each occurrence of a word includes inflection by a (possibly null) morphological change.

Rule proliferation 1

- **Problem:** How to account for this in grammar.
- **Possible solution:** Replace all NPs, Vs, and VPs throughout the grammar.

$S \rightarrow NP VP$

$NP \rightarrow you, dog, dogs, bear, bears, \dots$

$VP \rightarrow V NP$

$V \rightarrow washes, wash, washed, is, was, \dots$



$S \rightarrow NP_{3s} VP_{3s}$

$S \rightarrow NP_{3p} VP_{3p}$

$S \rightarrow NP_2 VP_2$

$S \rightarrow NP_{1s} VP_{1s}$

$S \rightarrow NP_{1p} VP_{1p}$

$NP_2 \rightarrow you$

\vdots

$VP_{3s} \rightarrow V_{3s} NP$

\vdots

$V_{3s} \rightarrow is, was, washes, washed, \dots$

$V_{3p} \rightarrow are, were,$

$wash, washed, \dots$

$V_{1s} \rightarrow am, was, wash, washed, \dots$

\vdots

$NP_{3s} \rightarrow dog, bear, \dots$

$NP_{3p} \rightarrow dogs, bears$

Rule proliferation 2

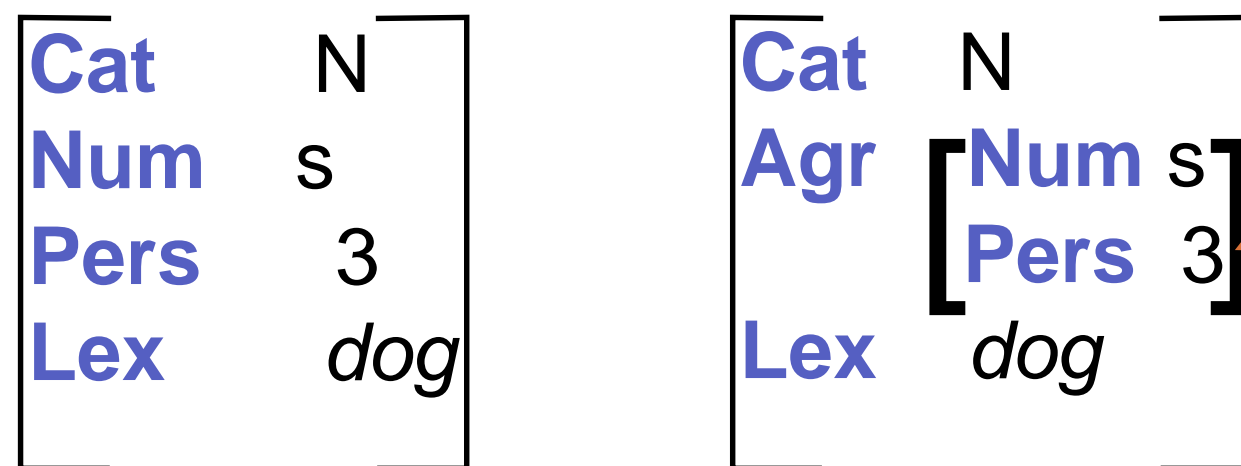
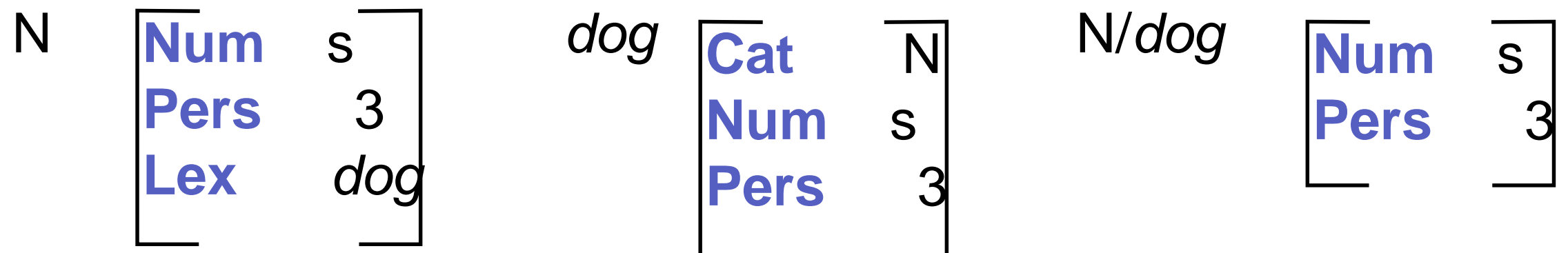
- *Drawback 1:* the result is big ... really big.
- *Drawback 2:* Losing the generalization:
 - All these Ss, NPs, VPs have the same structure.
 - Doesn't depend on particular verb, noun, and number.
- CF rules collapse together structural and featural information.
- All information must be completely and directly specified.
 - *E.g.*, can't just say that values must be equal for some feature without saying exactly what values.

Feature structures 1

- *Solution:* Separate feature information from syntactic, structural, and lexical information.
- A **feature structure** is a list of pairs:
 [*feature-name feature-value*]
- Feature-values may be atoms or feature structures.
- Can consider syntactic category or word to be bundle of features too.
- Can represent syntactic structure.

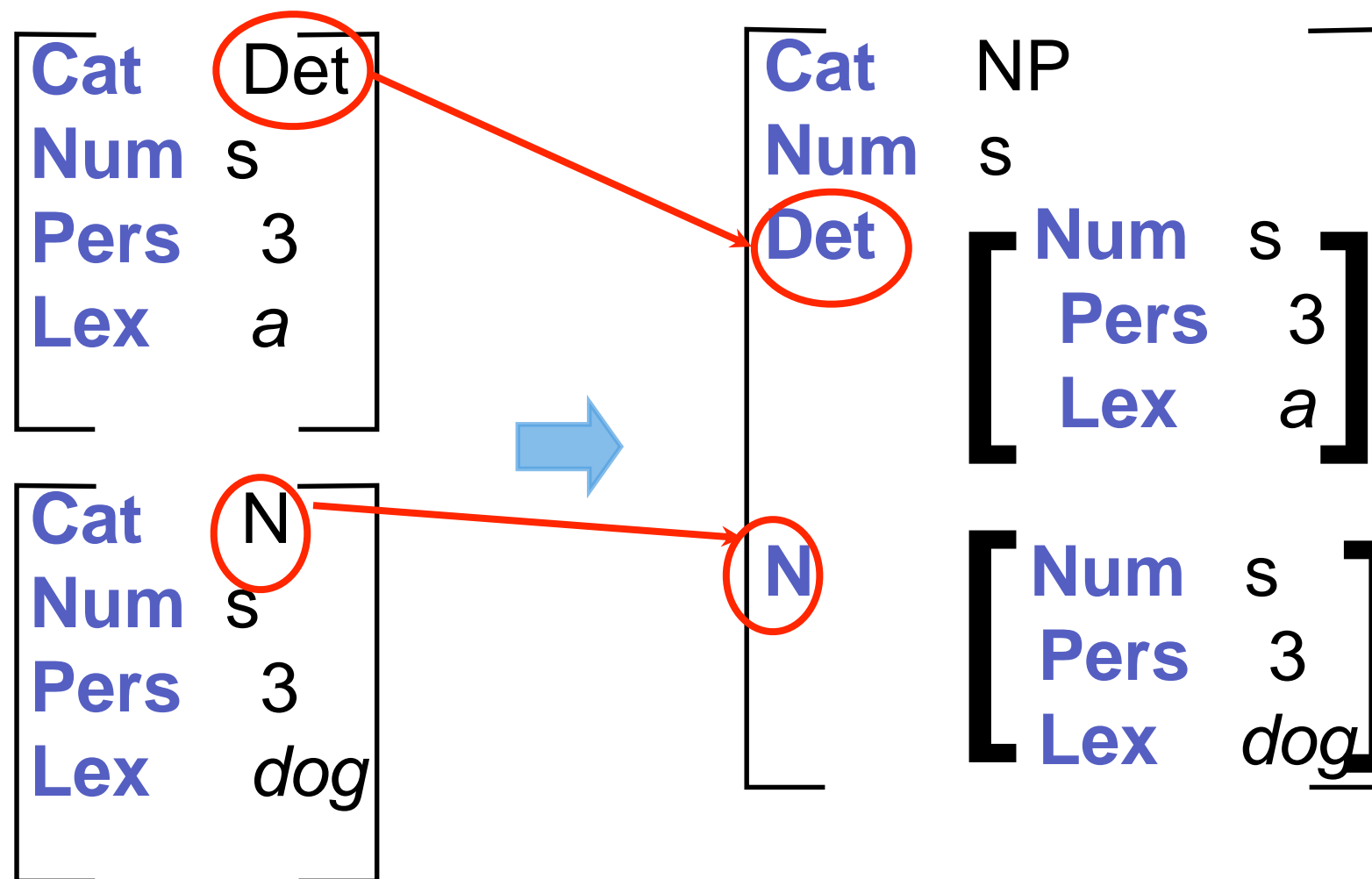
Feature structures 2

- *Drawback:* many equivalent notations.



Feature paths:
features of
features; e.g.,
(Agr Pers 3)

Feature structures 3



NP formed from Det and N.
Feature values in components become
feature names in new constituent.

Components of feature use

- 1. **Lexical specification:**

Description of *properties* of a word:
morphological, syntactic, semantic, ...

dog: $\begin{bmatrix} \text{Cat} & \text{N} \\ \text{Agr} & 3\text{s} \end{bmatrix}$

sleeps: $\begin{bmatrix} \text{Cat} & \text{V} \\ \text{Agr} & 3\text{s} \end{bmatrix}$

dogs: $\begin{bmatrix} \text{Cat} & \text{N} \\ \text{Agr} & 3\text{p} \end{bmatrix}$

sleep: $\begin{bmatrix} \text{Cat} & \text{V} \\ \text{Agr} & \{1\text{s}, 2\text{s}, 1\text{p}, 2\text{p}, 3\text{p}\} \end{bmatrix}$

Or: $\wedge 3\text{s}$

Or: $\text{N} \rightarrow \text{dog}$
 $(\text{N Agr}) = 3\text{s}$

$\text{V} \rightarrow \text{sleeps}$
 $(\text{V Agr}) = 3\text{s}$

$\text{N} \rightarrow \text{dogs}$
 $(\text{N Agr}) = 3\text{p}$

$\text{V} \rightarrow \text{sleep}$
 $(\text{V Agr}) = \{1\text{s}, 2\text{s}, 1\text{p}, 2\text{p}, 3\text{p}\}$

Components of feature use

- **2. Agreement:**
 - **Constraints** on co-occurrence in a rule — within or across phrases.
 - Typically are equational constraints.

NP \rightarrow Det N
(Det **Num**) = (N **Num**)

S \rightarrow NP VP
(NP **Agr**) = (VP **Agr**)

Components of feature use

- **3. Projection:**

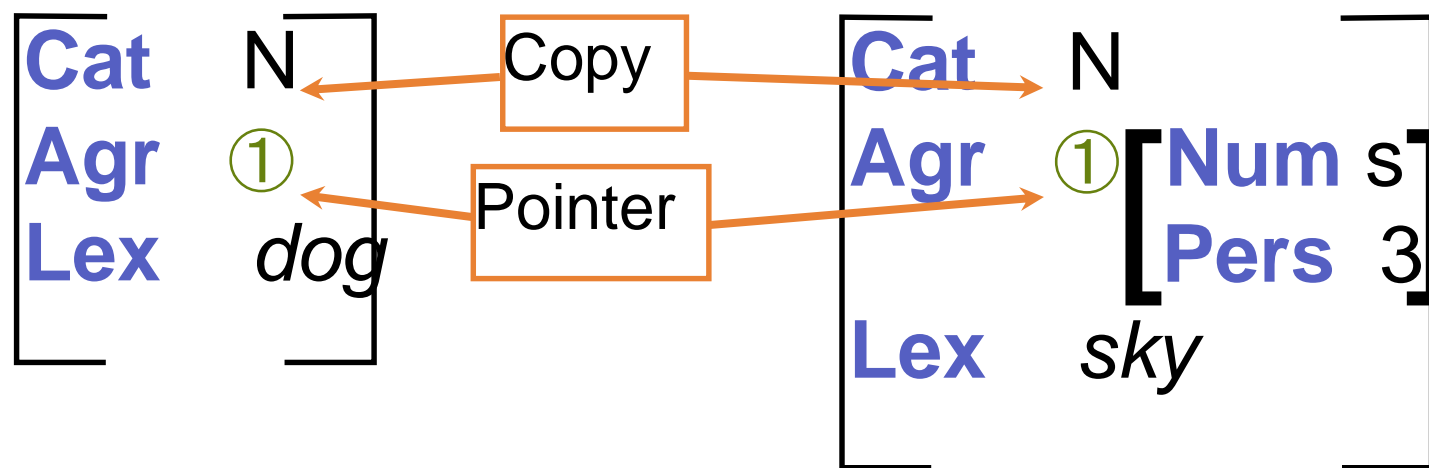
- ***Sharing of features*** between the head of a phrase and the phrase itself.

$$\begin{aligned} \text{VP} &\rightarrow \text{V} \dots \\ (\text{VP } \text{Agr}) &= (\text{V } \text{Agr}) \end{aligned}$$

- Head features:
 - **Agr** is typical, but so is the head-word itself as a feature.
(Common enough that there's usually a mechanism for “declaring” head features and omitting them from rules.)

Constraints on feature values 1

- What does it mean for two features to be “equal”?
- A *copy* of the value or feature structure, or a *pointer* to the same value or feature structure (re-entrancy, shared feature paths).



Constraints on feature values 2

- But: It may be sufficient that two features are not equal, just *compatible* — that they can be *unified*.

- *E.g.*,

Cat	N
Pers	3
Num	s

 and

Cat	N
Pers	3
Gndr	F

Subsumption of feature structures 1

- Feature structure X *subsumes* feature structure Y if Y is consistent with, and at least as specific as X .
 - Also say that Y *extends* X .
 Y can add (non-contradictory) features to those in X .
- **Definition:** X *subsumes* Y ($X \sqsubseteq Y$) iff there is a *simulation* of X inside Y , i.e., a function s.t.:
 - $\text{sim}(X) = Y$
 - If X is atomic, so is Y and $X = Y$
 - Otherwise, for all feature values $X.f$: $Y.f$ is defined, and sim simulates $X.f$ inside $Y.f$.

Subsumption of feature structures 2

- Examples:

$$\begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \end{bmatrix} \sqsubseteq \begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \\ \text{Gndr} & \text{F} \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \\ \text{Num} & \text{s} \end{bmatrix} \not\sqsubseteq \begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \\ \text{Gndr} & \text{F} \end{bmatrix}$$

$$\begin{bmatrix} \text{Cat} & \text{VP} \\ \text{Agr} & \textcircled{1} \\ \text{Subj} & [\text{Agr} \textcircled{1}] \end{bmatrix} \sqsubseteq \begin{bmatrix} \text{Cat} & \text{VP} \\ \text{Agr} & \textcircled{1} \\ \text{Subj} & \text{Agr} \textcircled{1} [\text{Pers} \textcircled{3} \\ & \text{Num} \textcircled{\text{s}}] \end{bmatrix}$$

Unification 1

- The *unification* of X and Y ($X \sqcup Y$) is the most general feature structure Z that is subsumed by both X and Y .
 - Z is the smallest feature structure that extends both X and Y .
- Unification is a constructive operation.
 - If any feature values in X and Y are incompatible, it fails.
 - Else it produces a feature structure that includes all the features in X and all the features in Y .

Unification 2

$$\begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \\ \text{Num} & \text{s} \\ _ & _ \end{bmatrix} \sqcup \begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \\ \text{Gndr} & \text{F} \\ _ & _ \end{bmatrix} = \begin{bmatrix} \text{Cat} & \text{N} \\ \text{Pers} & 3 \\ \text{Num} & \text{s} \\ \text{Gndr} & \text{F} \\ _ & _ \end{bmatrix}$$

Features in chart parsing

- Each constituent has an associated feature structure.
 - Constituents with children have a feature structure for each child.
- Arc addition:
 - The feature structure of the new arc is initialized with all known constraints.
- Arc extension:
 - The feature structure of the predicted constituent must unify with that of the completed constituent extending the arc.

Sample grammar fragment

$S \rightarrow NP \ VP$

$(NP \text{ Agr}) = (VP \text{ Agr})$

$NP \rightarrow Det \ N$

$(NP \text{ Agr}) = (N \text{ Agr})$

$(Det \text{ Agr}) = (N \text{ Agr})$

$VP \rightarrow V$

$(VP \text{ Agr}) = (V \text{ Agr})$

$Det \rightarrow a$

$[Agr \ 3s]$

$Det \rightarrow all$

$[Agr \ 3p]$

$Det \rightarrow the$

$[Agr \ \{3s, 3p\}]$

$N \rightarrow dog$

$[Agr \ 3s]$

$N \rightarrow dogs$

$[Agr \ 3p]$

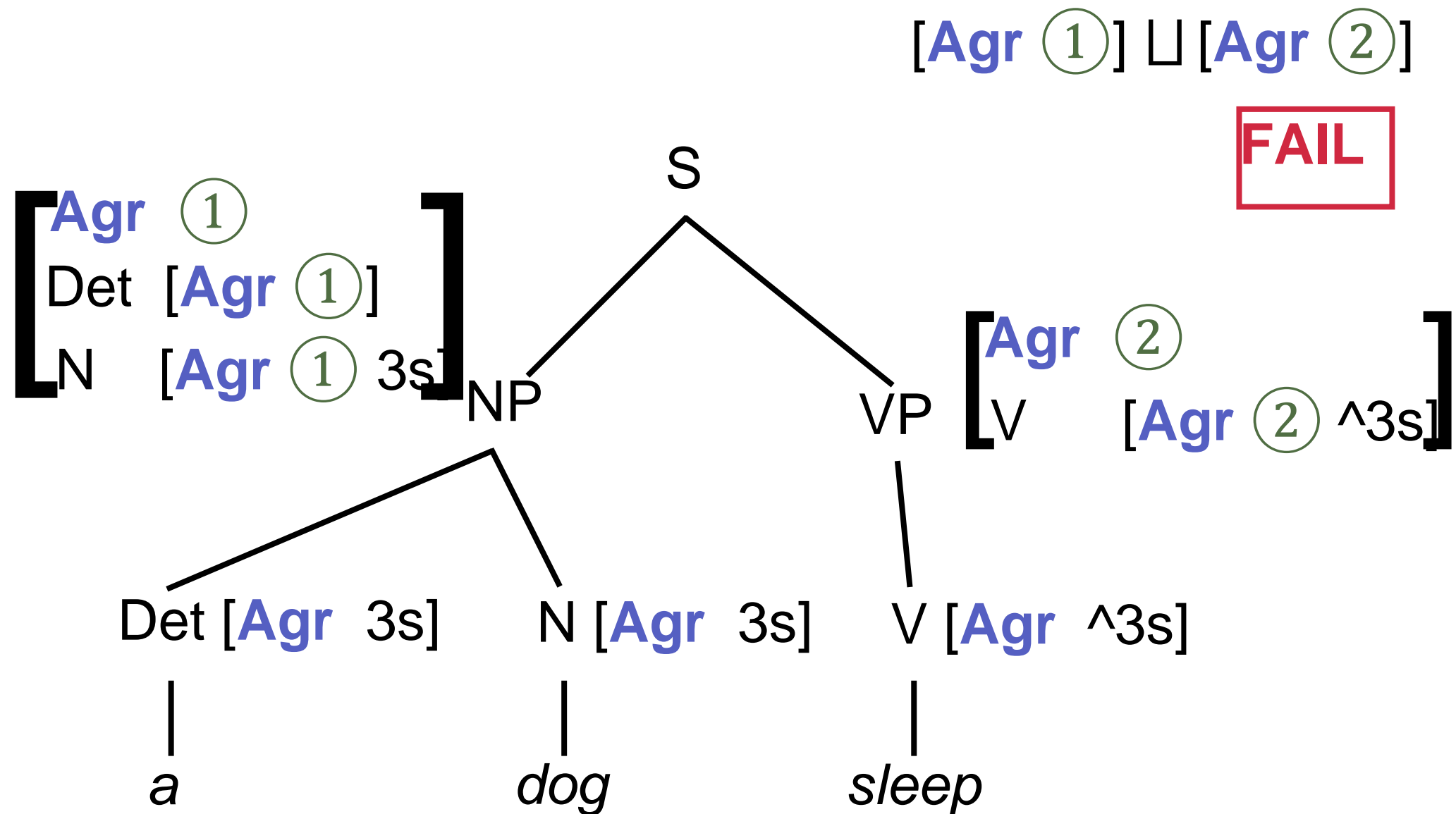
$V \rightarrow sleep$

$[Agr \ ^3s]$

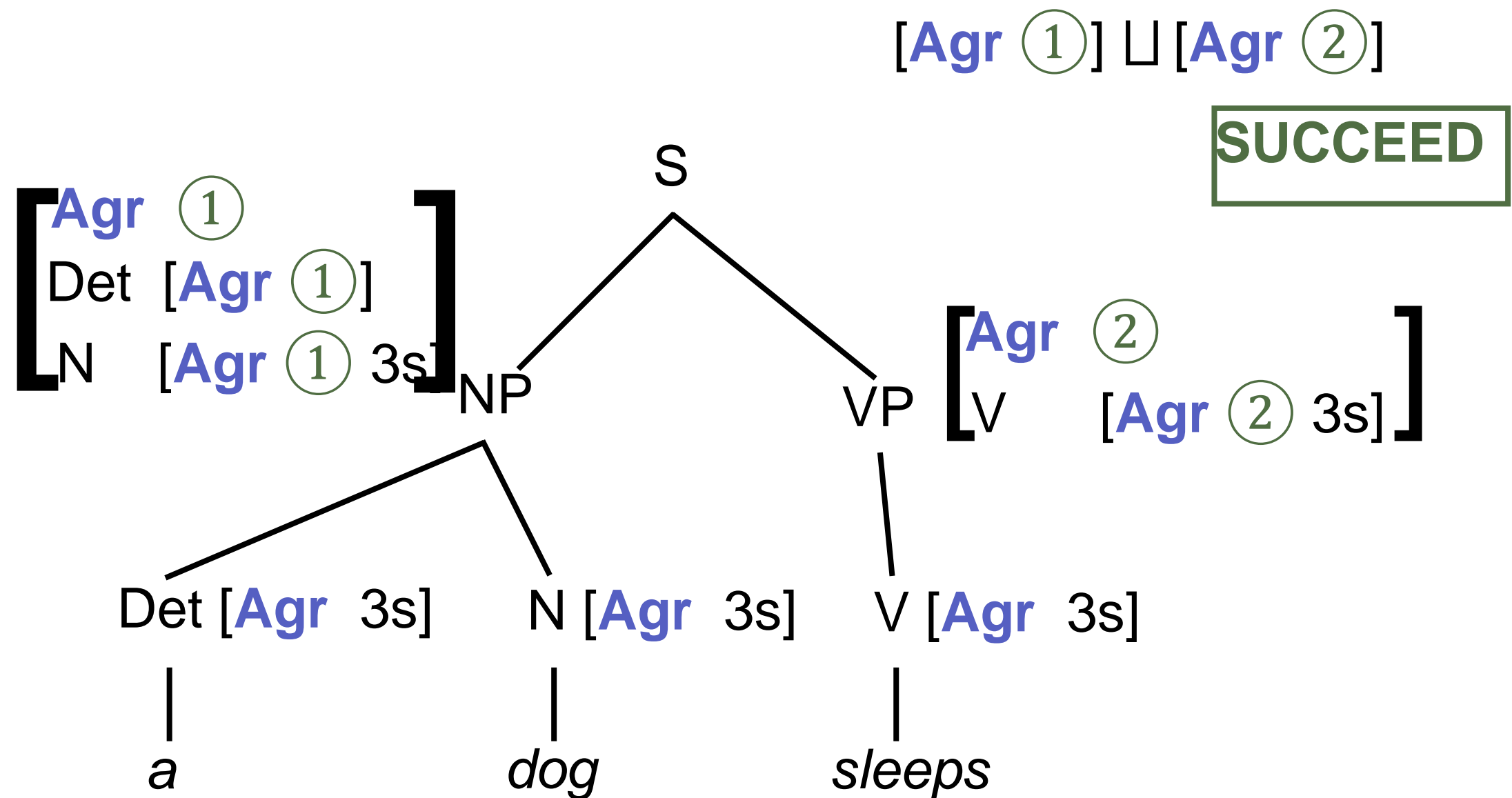
$V \rightarrow sleeps$

$[Agr \ 3s]$

Mismatched features fail




Unifiable features succeed



Advantages of this approach

- Distinguishes structure from "functional" info.
- Allows for economy of specification:
 - Equations in rules:
$$S \rightarrow NP \ VP$$
$$(NP \text{ Agr}) = (VP \text{ Agr})$$

Must unify with


 - Sets of values in lexicon:
$$N \rightarrow \textit{fish}$$
$$(N \text{ Agr} \{3s, 3p\})$$
- Allows for indirect specification and transfer of information, e.g., head features.

Features and the lexicon

- Lexicon may contain each inflected form.
 - Feature values and base form listed.
- Lexicon may contain only base forms.
 - Process of *morphological analysis* maps inflected form to base form plus feature values.
 - Time–space trade-off, varies by language.
- Lexicon may contain *semantics* for each form.

Morphological analysis

- Morphological analysis is simple in English.
- Reverse the rules for inflections, including spelling changes.

<i>dogs</i> → <i>dog</i> [Agr 3p]	<i>eats</i> → <i>eat</i> [Agr 3s, Tns pres]
<i>dog</i> → <i>dog</i> [Agr 3s]	<i>ripped</i> → <i>rip</i> [Tns past]
<i>berries</i> → <i>berry</i> [Agr 3p]	<i>tarried</i> → <i>tarry</i> [Tns past]
<i>buses</i> → <i>bus</i> [Agr 3p]	<i>running</i> → <i>run</i> [Tns pp]

- Irregular forms will always have to be explicitly listed in lexicon.

children → *child* [**Agr** 3p] *sang* → *sing* [**Tns** past]

Morphology in other languages

- Rules may be more complex in other (even European) languages.
- Languages with compounding (e.g., German) or agglutination (e.g., Finnish) require more-sophisticated methods.
 - E.g., *Verdauungsspaziergang*, a stroll that one takes after a meal to assist in digestion.

Semantics as a lexical feature

- Add a **Sem** feature:

Cat	N
Num	s
Pers	3
Lex	<i>dog</i>
Sem	dog

Typewriter font
for semantic objects

- The meaning of *dog* is dog.
The meaning of *chien* and *Hund* are both dog.
The meaning of *dog* is G52790.
Not uncommon, but pretty unimpressive...

Verb subcategorization 1

- **Problem:** Constraints on verbs and their complements.

*Nadia told / instructed / *said / *informed Ross to sit down.*

*Nadia *told / *instructed / said / *informed to sit down.*

*Nadia told / *instructed / *said / informed Ross of the requirement to sit down.*

Nadia gave / donated her painting to the museum.

*Nadia gave / *donated the museum her painting.*

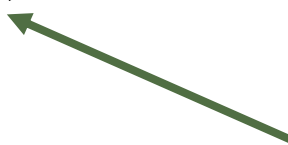
Nadia put / ate the cake in the kitchen.

*Nadia *put / ate the cake.*

Verb subcategorization 2

- VPs are much more complex than just V with optional NP and/or PP.
 - Can include more than one NP.
 - Can include clauses of various types:
that Ross fed the marmoset
to pay him the money
- **Subcat**: A feature on a verb indicating the kinds of verb phrase it allows:
_np, _np_np, _inf, _np_inf, ...

Write this way to distinguish from constituents.



Verb tense and aspect 1

- **Tense and aspect** markings on verb:
 - Locate the event in time (relative to another time).
 - Mark the event as complete/finished or in progress.

Nadia rides the horse. — In progress now.

Nadia rode the horse. — Completed before now.

Nadia had ridden the horse. — Completed before before now.

Nadia was riding the horse. — In progress before now.

⋮

Verb tense and aspect 2

- Tense: past or present
- Aspect: simple, progressive, or perfect

<i>Nadia ...</i>		Auxiliary verb		In progress		Complete	
	Simple	Progressive		Perfect			
Present	<i>rides</i>	<i>is</i>	<i>riding</i>	<i>has</i>	<i>ridden</i>		
Past	<i>rode</i>	<i>was</i>	<i>riding</i>	<i>had</i>	<i>ridden</i>		
		Present participle		Past participle		<i>... the horse</i>	

Verb tense and aspect 3

- Tense: past or present
- Aspect: simple, progressive, or perfect

<i>Nadia ...</i>		Auxiliary verbs	
	Simple		Perfect progressive (continuous)
Present	<i>rides</i>	<i>has been</i>	<i>riding</i>
Past	<i>rode</i>	<i>had been</i>	<i>riding</i>
<i>... the horse</i>			

Modal verbs

- **Modal verbs:** Auxiliary verbs that express degrees of certainty, obligation, possibility, prediction, etc.

Nadia

{could, should, must, ought to, might, will, ...}

{ride, be riding, have ridden, have been riding}

the horse.

English auxiliary system

- Structure (so far):
[MODAL] [HAVE] [BE] MAIN-VERB
- General pattern:
VP \rightarrow AUX VP
AUX \rightarrow MODAL | HAVE | BE
 - Use features to capture necessary agreements.

Voice 1

The goalie kicked the ball.

ACTIVE

Event: *kicked*

Role: Agent
(doer)

Role: Theme
(thing affected)

Thing: *the goalie* Thing: *the ball*

kick (agent=goalie, theme=ball)

Voice 2

The ball was kicked.

PASSIVE

Event: *kicked*

Role: Theme
(thing affected)

Thing: *the ball*

kick (agent=?, theme=ball)

Voice 3

The ball was kicked by the goalie. **PASSIVE**

Event: *kicked*

Role: Theme
(thing affected)

Role: Agent
(doer)

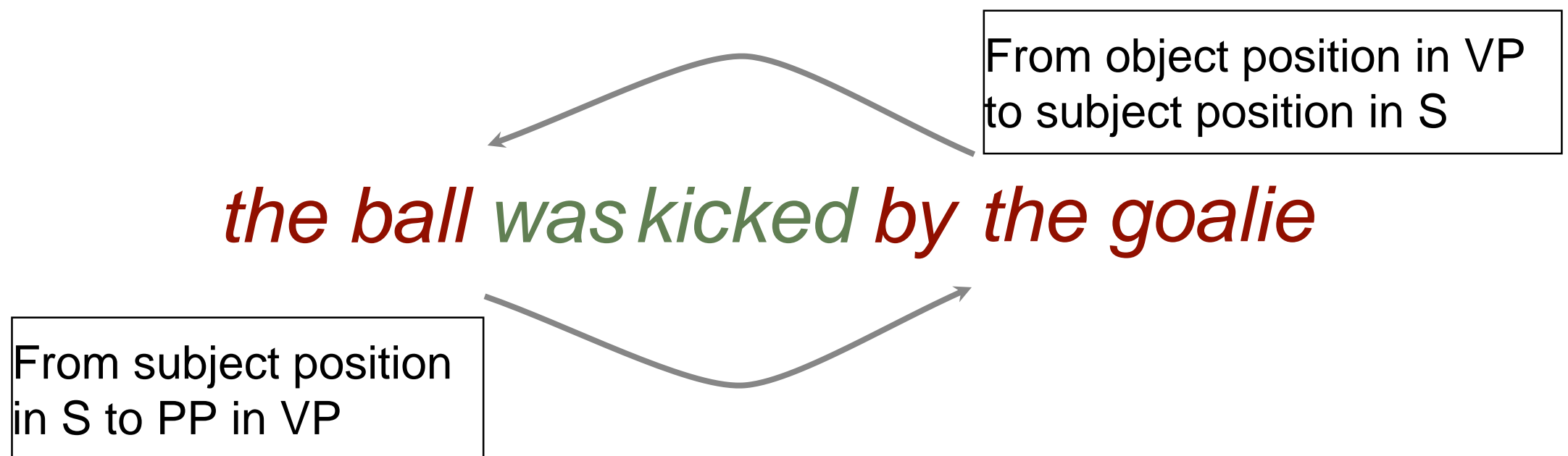
Thing: *the ball* Thing: *the goalie*

kick (agent=goalie, theme=ball)

Passive as *Diathetic alternation*

the goalie *kicked* *the ball*

Passive as *Diathetic alternation*



But the semantic representation doesn't change

Voice 4

- **Voice:** System of assigning thematic roles to syntactic positions.
 - English has **active** and **passive** voices.
- Passive expressed with *be*+past participle.
Other auxiliaries may also apply, including progressive *be*.
- *Nadia was kissed.* *Nadia was being kissed.*
Nadia had been kissed. *Nadia had been being kissed.*
Nadia could be kissed. *Nadia could have been being*
kissed.
- Structure:
[MODAL] [HAVE] [BE₁] [BE₂] MAIN-VERB

Some useful features

- **VForm**: The tense/aspect form of a verb:
passive, pastprt, ...
- **CompForm**: The tense/aspect form of the
complement of an auxiliary.

Augmenting rules for passive voice

- For all rules of the form:

$VP \rightarrow V \ NP \ X$

$(V \text{ Subcat}) = _y$

ADD
→

$VP \rightarrow V \ X$

$(V \text{ Subcat}) = _y$

$(V \text{ VForm}) = \text{passive}$

$(VP \text{ VForm}) = \text{passive}$

Metarule to ease grammar coding

- Augment Aux+VP rules:

$VP \rightarrow AUX \ VP$

$(AUX \text{ Root}) = \text{Be2}$

$(AUX \text{ CompForm}) = (VP_2 \text{ VForm})$

$(VP_2 \text{ VForm}) = \text{passive}$

The GAP feature for passive voice

$S \rightarrow NP \ VP$

- $_1 (NP \text{ Agr}) = (VP \text{ Agr})$
- $_2 (VP \text{ VForm}) = \text{passive}$
- $_3 (VP \text{ Gap Cat}) = NP$
- $_4 (VP \text{ Gap Agr}) = (NP \text{ Agr})$
- $_5 (VP \text{ Gap Sem}) = (NP \text{ Sem})$

$VP \rightarrow AUX \ VP$

- $_1 (VP_1 \text{ Agr}) = (AUX \text{ Agr})$
- $_2 (VP_1 \text{ VForm}) = (VP_2 \text{ VForm})$
- $_3 (VP_1 \text{ Gap}) = (VP_2 \text{ Gap})$
- $_4 (AUX \text{ Lex}) = \text{be2}$
- $_5 (VP_2 \text{ VForm}) = \text{passive}$

$V \rightarrow \text{kicked}$

- $_1 (V \text{ VForm}) = \{\text{pastprt, passive}\}$
- $_2 (V \text{ Subcat}) = \text{_np}$
- $_3 (V \text{ Lex}) = \text{kick}$
- $_4 (V \text{ Sem}) = \text{kick}$

$VP \rightarrow V \ NP$

- $_1 (VP \text{ VForm}) = (V \text{ VForm})$
- $_2 (VP \text{ Gap}) = (NP \text{ Gap})$
- $_3 (V \text{ Subcat}) = \text{_np}$

Empty string

$NP \rightarrow \varepsilon$

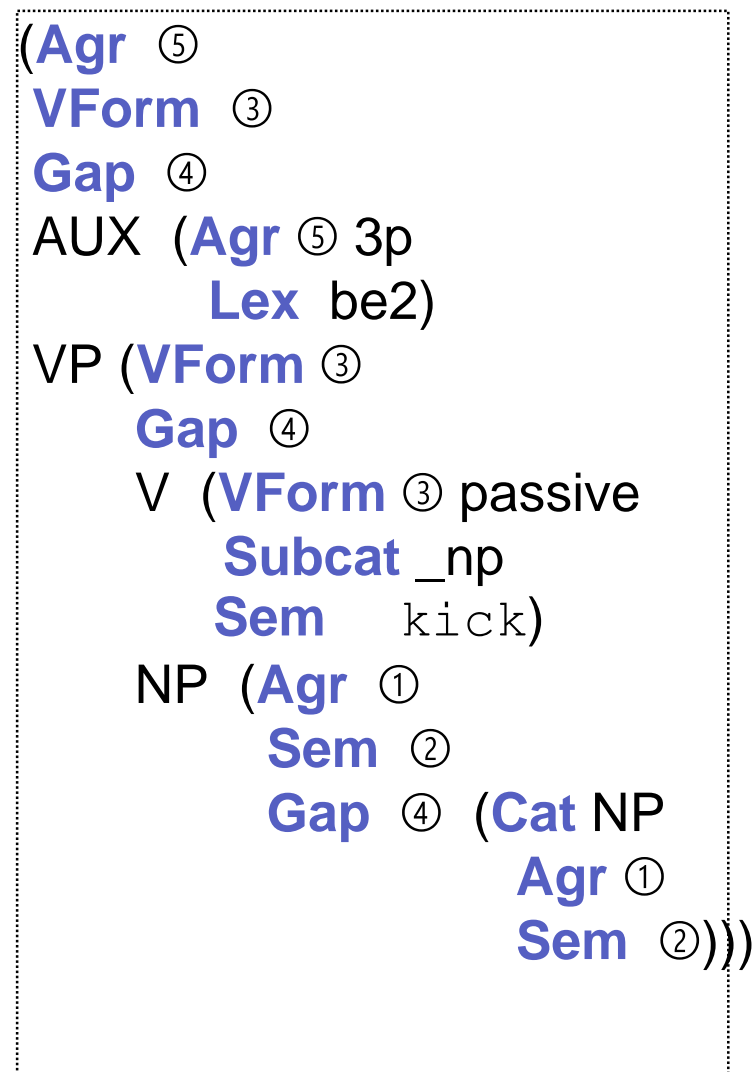
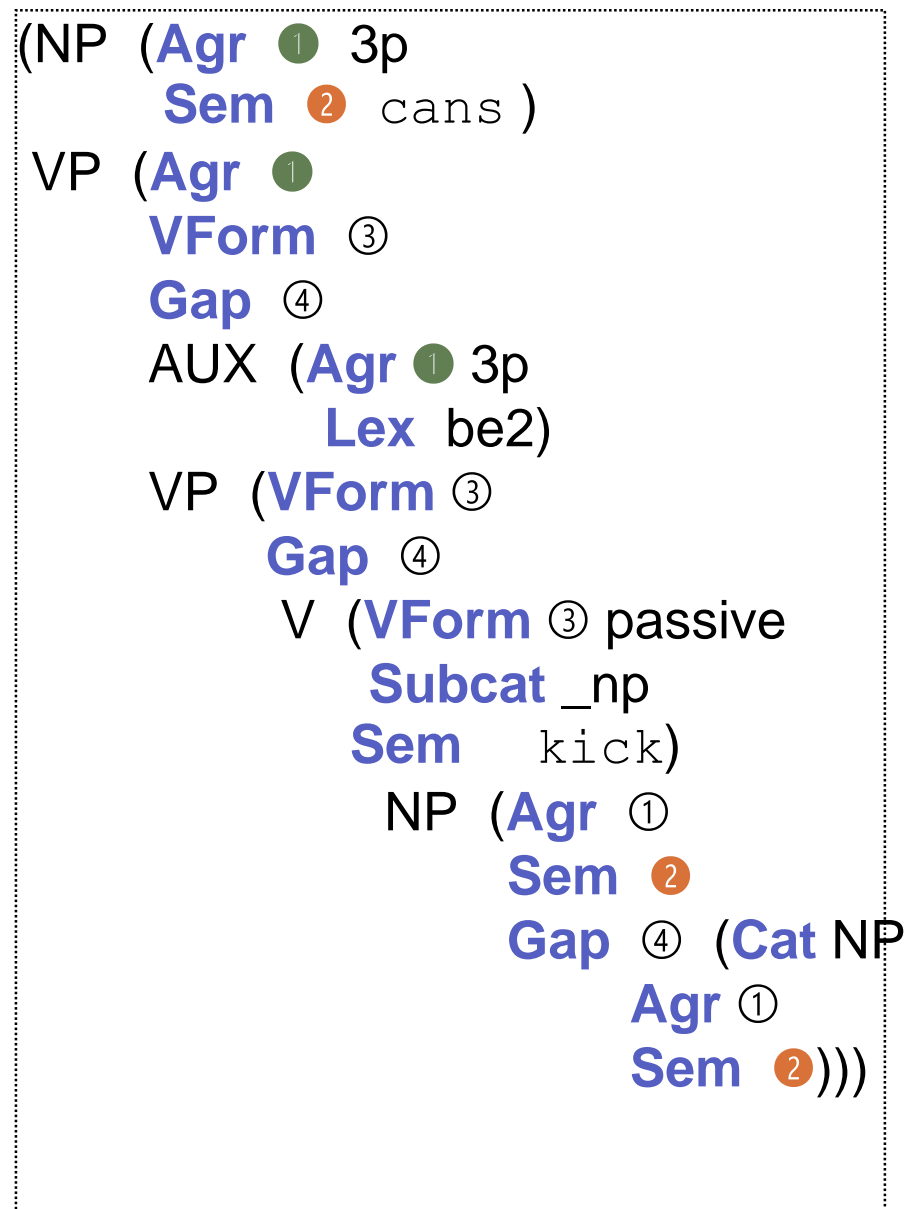
- $_1 (NP \text{ Gap Cat}) = NP$
- $_2 (NP \text{ Gap Agr}) = (NP \text{ Agr})$
- $_3 (NP \text{ Gap Sem}) = (NP \text{ Sem})$

$NP \rightarrow \text{cans}$

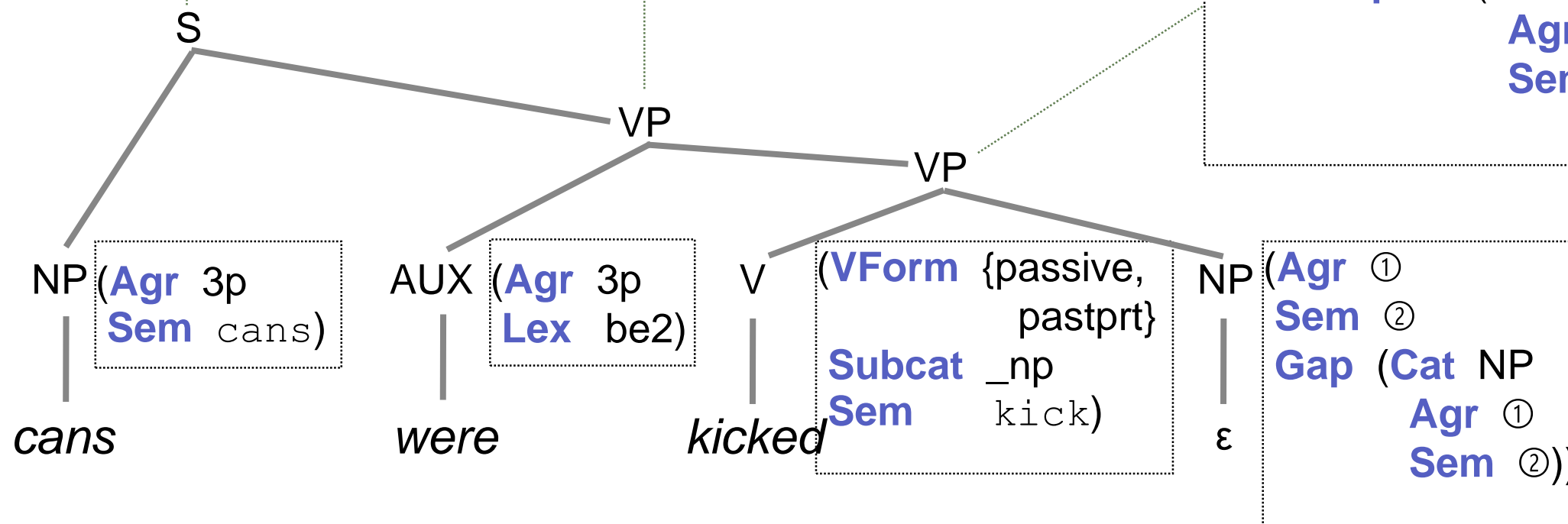
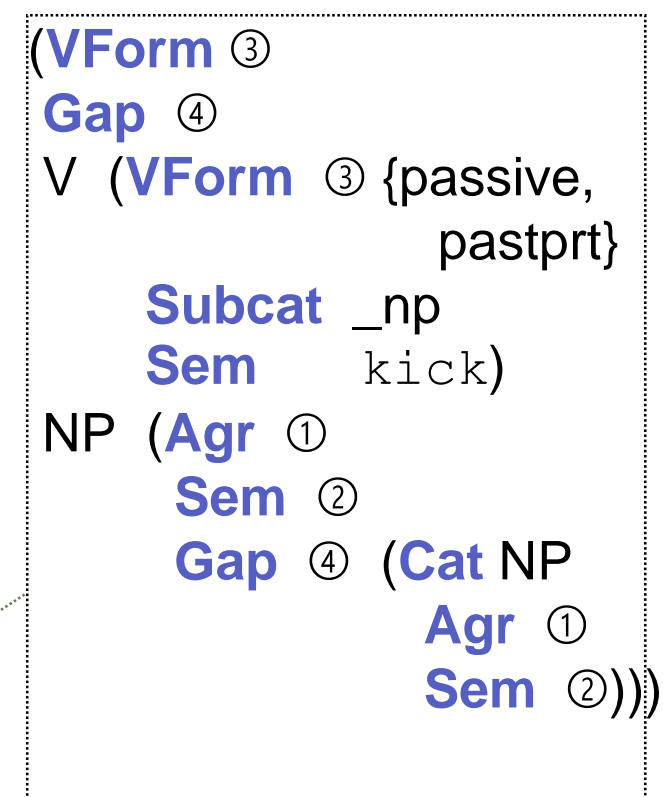
- $_1 (NP \text{ Agr}) = 3p$
- $_2 (NP \text{ Lex}) = \text{can}$
- $_3 (NP \text{ Sem}) = \text{cans}$

$AUX \rightarrow \text{were}$

- $_1 (AUX \text{ Agr}) = 3p$
- $_2 (AUX \text{ Lex}) = \text{be2}$



Note: The green ①'s of the S were ⑤'s until the 4th constraint of the rule $S \rightarrow NP VP$. The 5th constraint fills in the **Sem** of the **Gap** ②.



Other cases of *gap percolation*

- Other constructions involve NPs in syntactic configurations where they would not get the right thematic roles using linear order alone.

Nadia seems to like Ross.

Nadia seems to be liked.

Nadia is easy to like.

Who did Nadia like?

I fed the dog that Nadia likes to walk.

- Can use grammar rules with gap features to ensure correct structure/interpretation of these as well.

Summary

- Features help capture syntactic constructions in a general and elegant grammar.
- Features can encode the compositional semantics of a sentence as you parse it.
- Features can accomplish mapping functions between syntax and semantics that simplify the interpretation process.