Bayesian Models of Lexicon Learning

Timothy J. O'Donnell
McGill Linguistics

Language Learning and the Lexicon

Language learning consists mostly (entirely?) of learning the lexicon (words, morphemes, etc.)?

The Problem of Lexical Uncertainty

How do learners identify the lexical units in their language?

The Problem of Lexical Uncertainty

	Suffix
	-ness
	-th
	-ity

The Problem of Lexical Uncertainty

Nouns (`N`)

Attach to adjectives (`Adj`)

Mean abstract quality or state

	Suffix
goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, business, goodness, …	-ness
	-th
	-ity

The Problem of Lexical Uncertainty

Nouns (`N`)

Attach to adjectives (`Adj`)

Mean abstract quality or state

	Suffix
goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, business, goodness, …	-ness
truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …	-th
	-ity

The Problem of Lexical Uncertainty

Nouns (`N`)

Attach to adjectives (`Adj`)

Mean abstract quality or state

	Suffix
goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, business, goodness, …	-ness
truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …	-th
verticality, tractability, severity, seniority, inanity, electricity, parity, scarcity, reality, …	-ity

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness

pine-scented *pine-scentedness*

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness

goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, pretentiousness, business, goodness, greenness, …

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness

goodness, cheapness, forgiveness, circuitousness, grandness, orderliness, pretentiousness, business, greenness, …

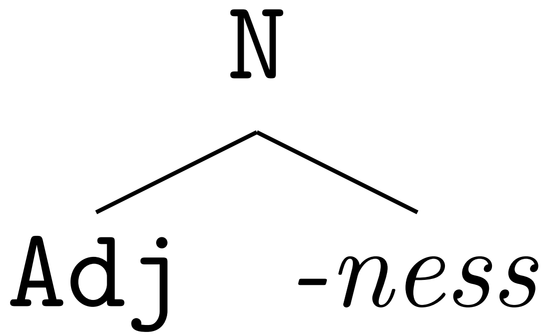
Phonological Regularities

Morpho-Syntactic Regularities

Semantic Regularities

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness



The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th

cool _*coolth_

The Problem of Lexical Uncertainty

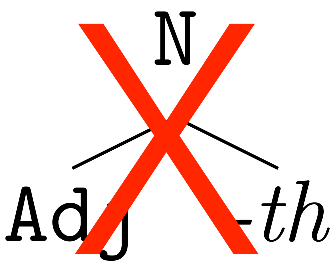
	Suffix
Productive	-ness
Unproductive	-th

truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th

truth, warmth, width, depth, filth, sloth, strength, death, dearth, wealth, length, youth, …



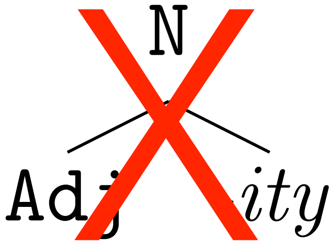
The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity

pine-scented	**pine-scentedity*
cool	**coolity*

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity



The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity

-ile, -al, -able, -ic, -(i)an

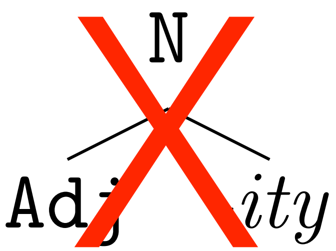
The Problem of Lexical Uncertainty

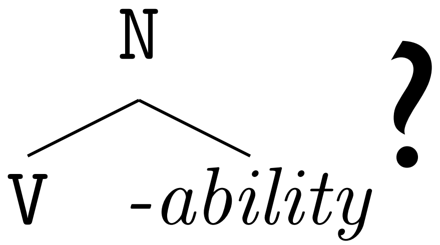
	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity

sequentiable ***sequentiability***

The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity





The Problem of Lexical Uncertainty

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity

The Problem of Lexical Uncertainty

What principles can the learner use to distinguish real lexical units like -*ness* from false lexical units like -*th*?

Productive	-ness
Unproductive	-th
Productive in Combination	-ity

Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.

Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.

Bayesian Approach

- Use probabilistic models to define prior distribution over possible lexicons.
- Find posterior distribution over lexicons given actual data using probabilistic conditioning.

$$P(L \mid D) \propto P(D \mid L)\, P(L)$$

The Framework

1. An underlying computational system that defines space of possible structures.
2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic model implementing a **tradeoff** between storage and computation (prior and likelihood).
3. Use probabilistic inference to derive language-specific predictions.

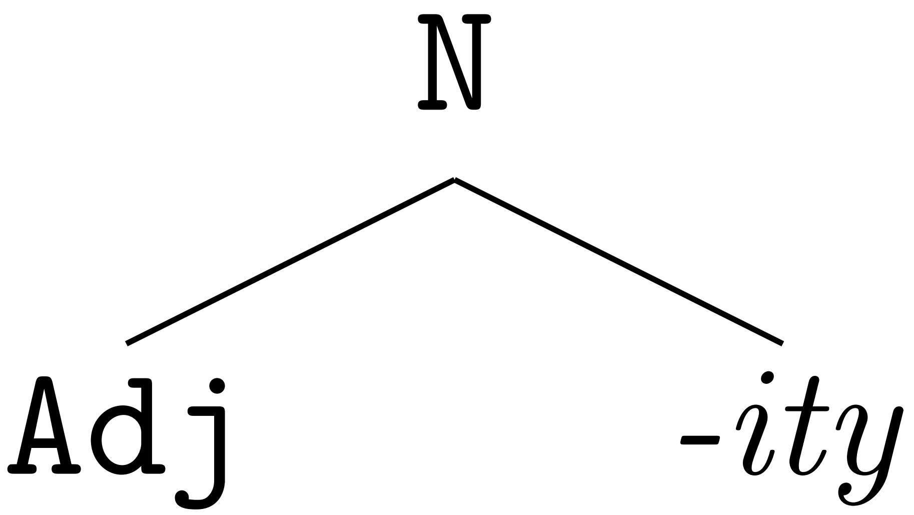
The Framework

1. An underlying computational system that defines space of possible structures.
2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic model implementing a **tradeoff** between storage and computation (prior and likelihood).
3. Use probabilistic inference to derive language-specific predictions.

Underlying Computational System

1. Inventory of structured units (lexicon).
2. Structure-building operations.

Lexical Items



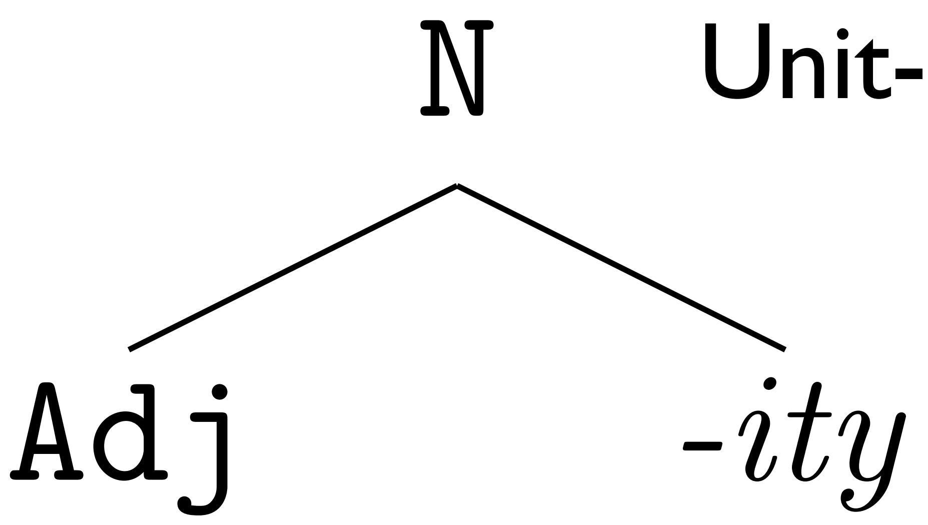
N

Adj

-ity

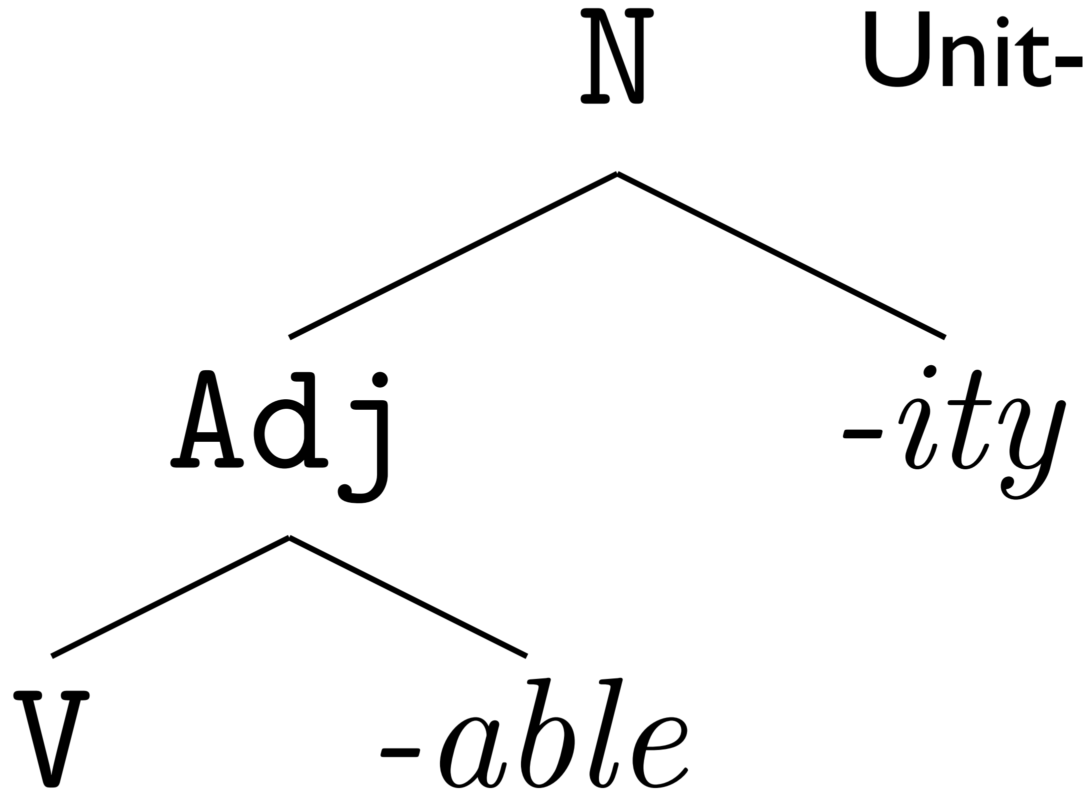
Stems, suffixes, prefixes with combinatorial restrictions.

Structure Building



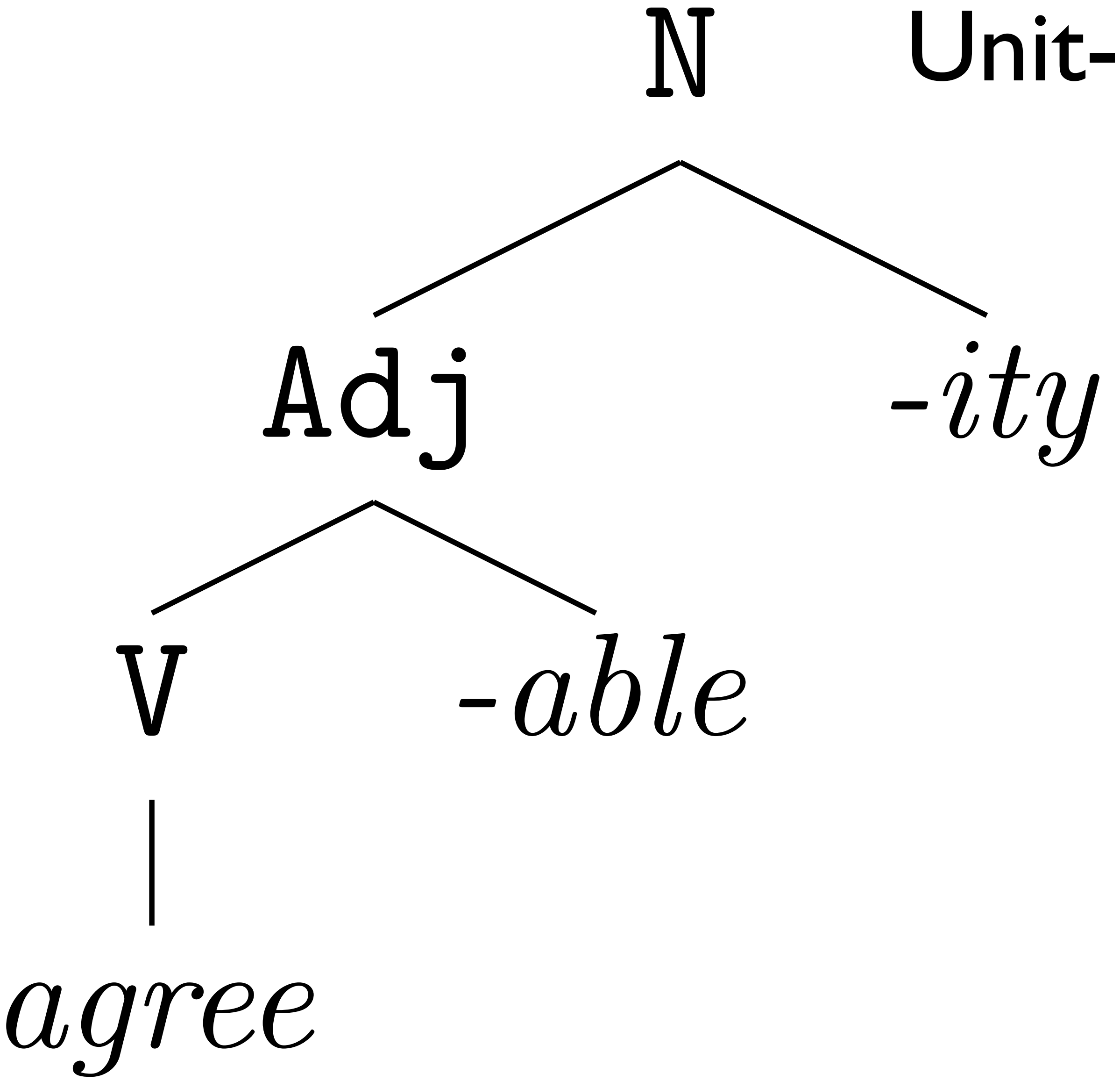
Unit-driven selection.

Structure Building



Unit-driven selection.

Structure Building



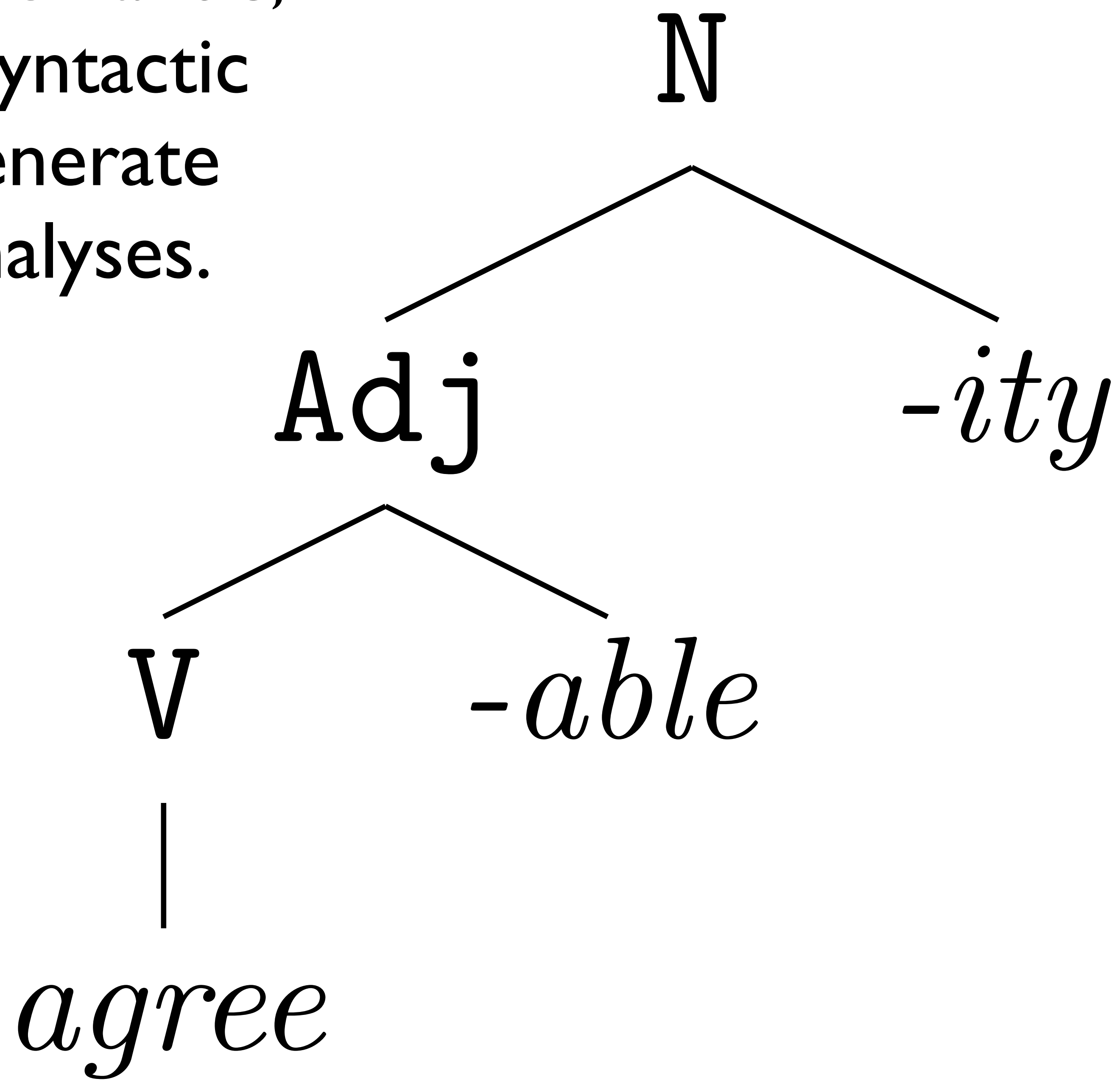
Unit-driven selection.

The Framework

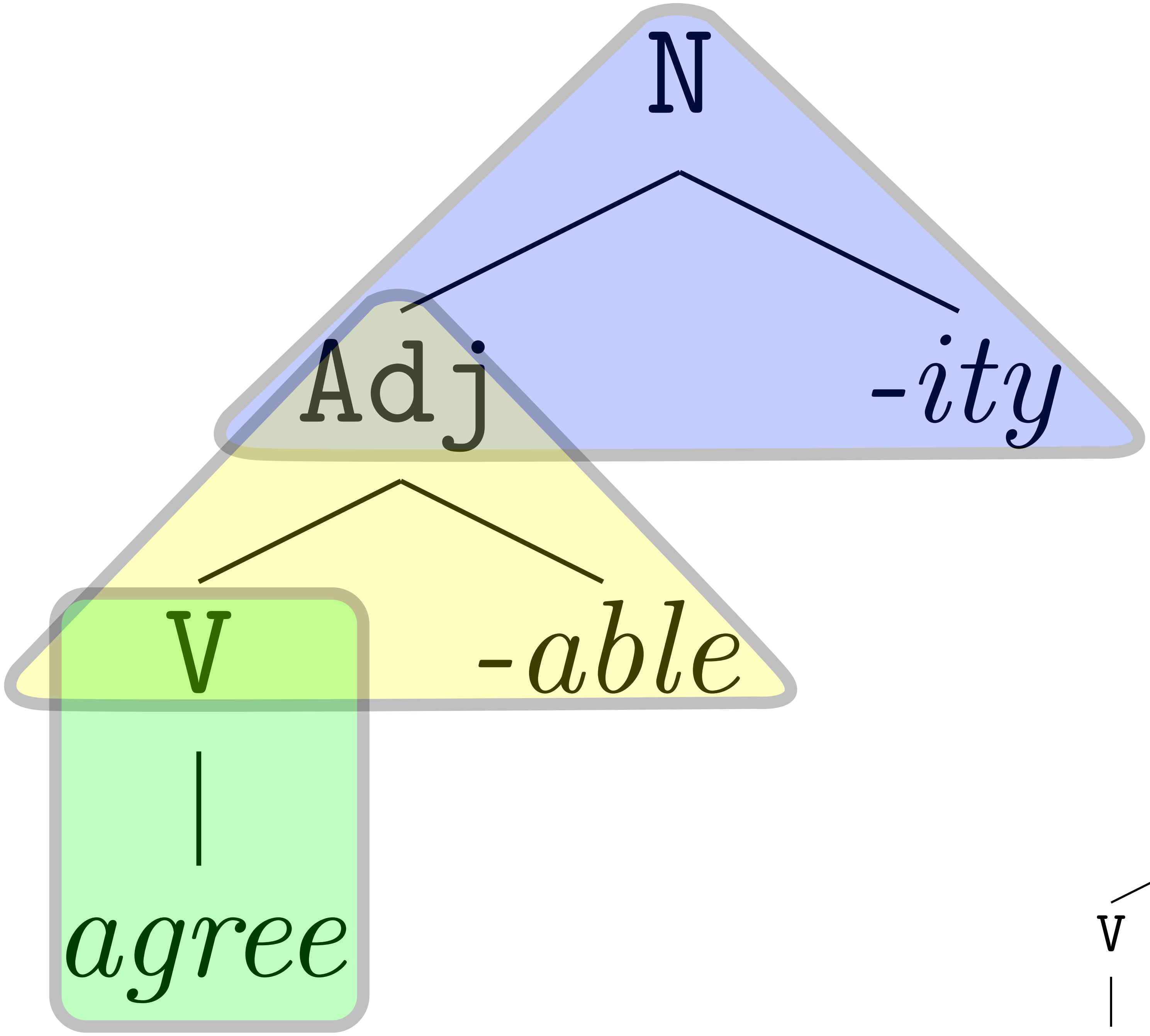
1. An underlying computational system that defines space of possible structures.
2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic model implementing a **tradeoff** between storage and computation (prior and likelihood).
3. Use probabilistic inference to derive language-specific predictions.

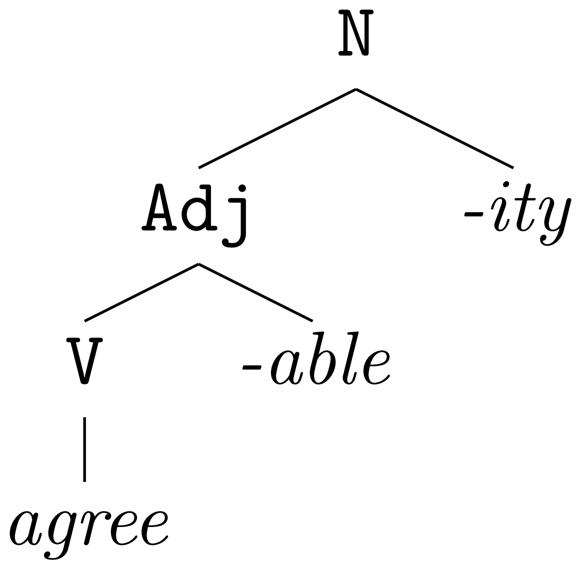
Uncertainty

Phonological, semantic, and morphosyntactic processes generate candidate analyses.

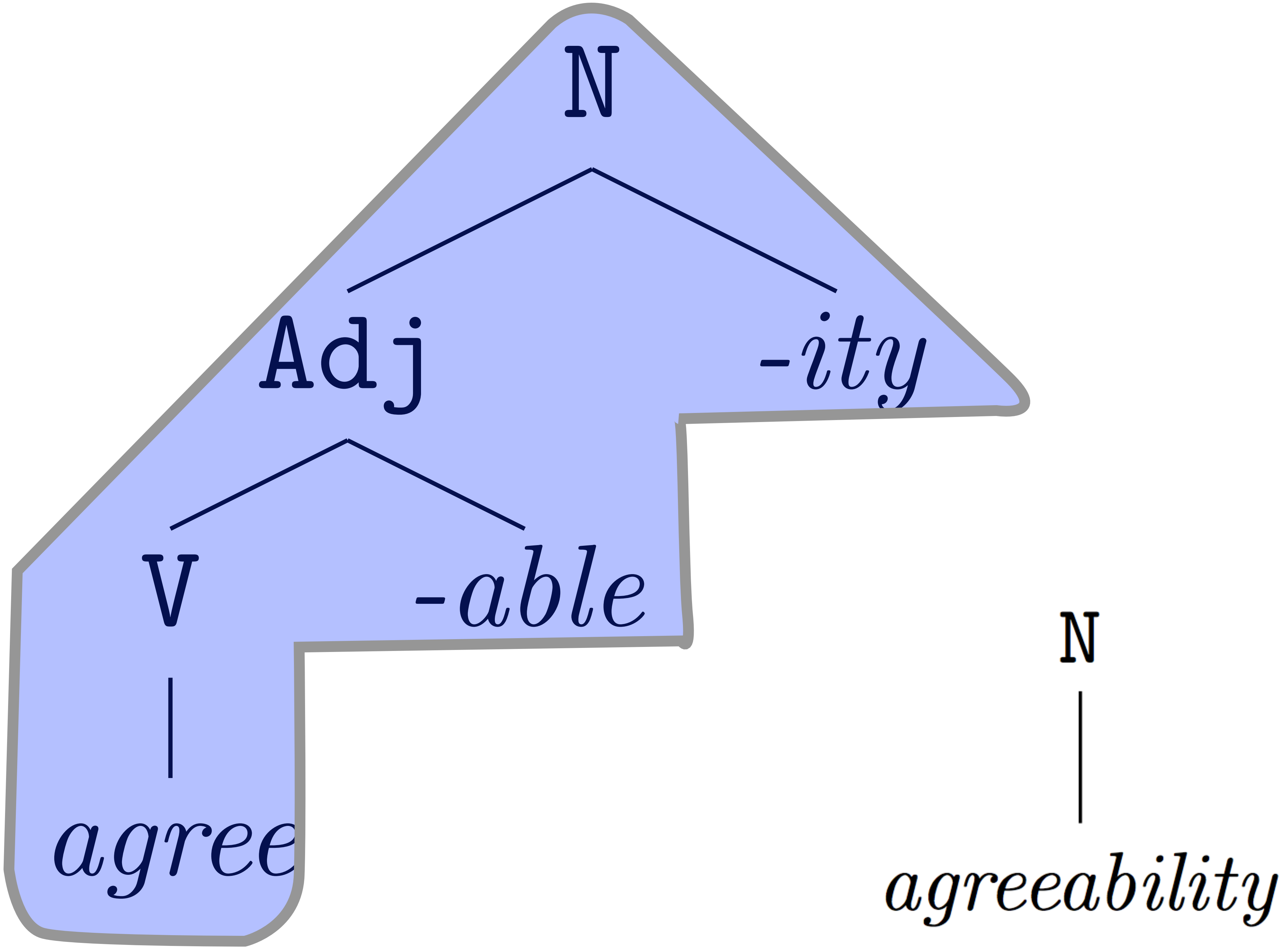


Uncertainty

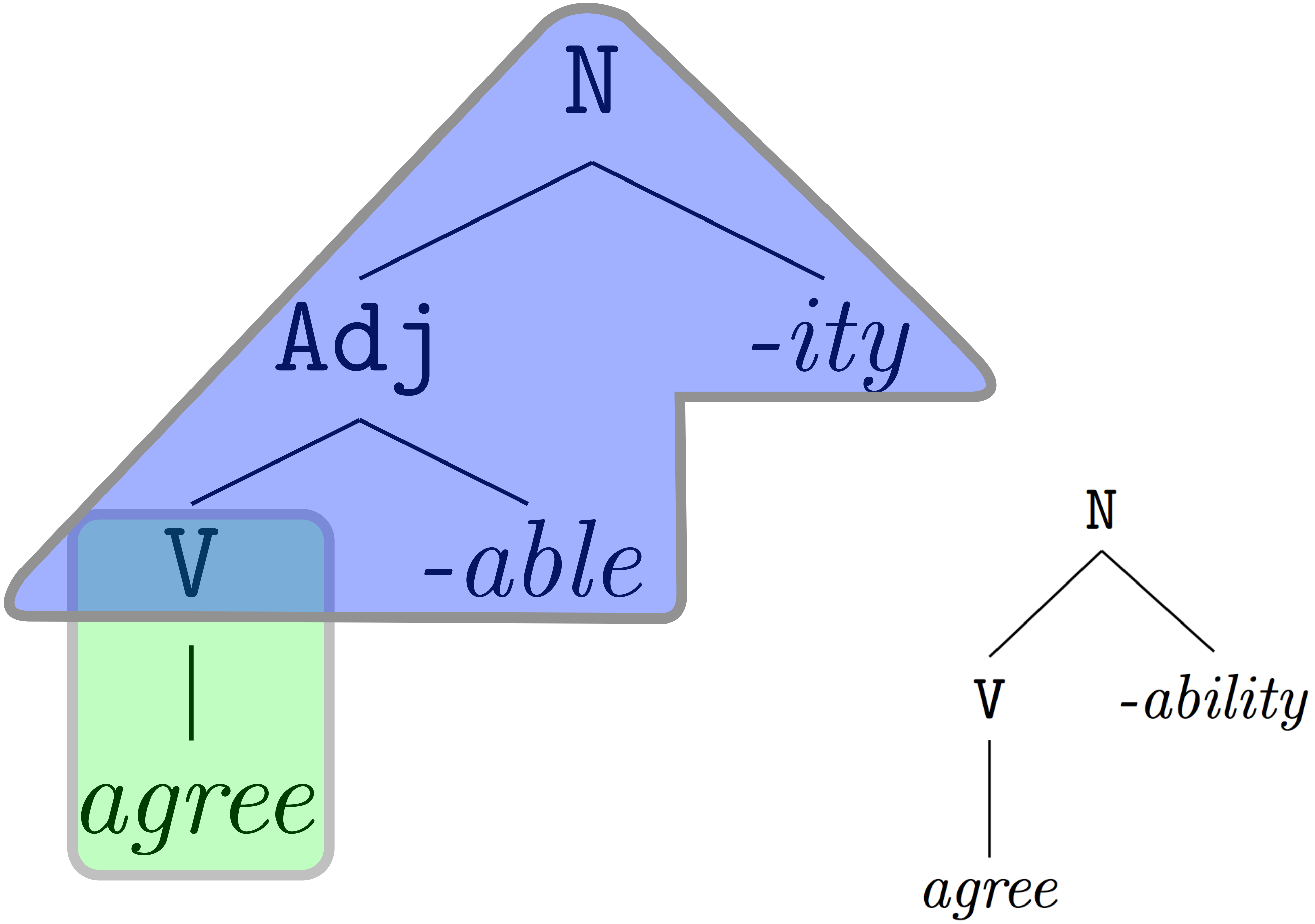




Uncertainty



Uncertainty



Probabilistic Model

- **Goal**: Define a probability distribution that quantifies uncertainty about likely and unlikely lexical units.
- Prior and likelihood lead to a **tradeoff** between storage and computation.

1. Prior over lexicon: P(L).
2. Likelihood over derivations of individual forms: P(D | L).

Prior on Lexicon

- Prefer a small number of highly reusable (i.e., generalizable) units.
- Dirichlet Process (Ferguson, 1973; Aldous, 1983; Sethuraman, 1994; Pitman 1995).
    - Allows unbounded number of stored units (non-parametric).
    - Prefers fewer units.
    - Prefers more reusable units.

$$P(\text{unit}) \propto \text{frequency of use}$$

Likelihood of Derived Forms

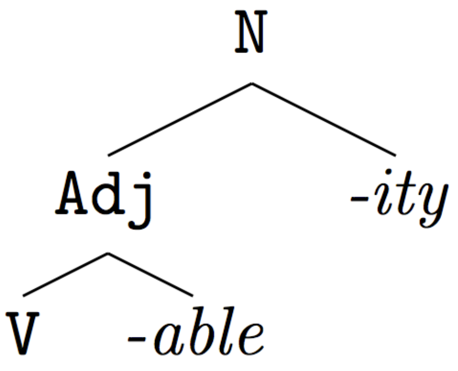
- Simple observed forms with few parts.
- Probability of derivation: Product of probability of individual units.
  - Probabilities between 0 and 1.
    - Geometric decrease in probability.
  - Prefers fewer lexical items per derivation.
  - Prefers to store more complex units.

The Framework

1. An underlying computational system that defines space of possible structures.
2. Manage uncertainty over the ways in which forms can be analyzed with a probabilistic implementing a **tradeoff** between storage and computation.
3. Use probabilistic inference to derive language-specific predictions.

The Mathematical Model: *Fragment Grammars*

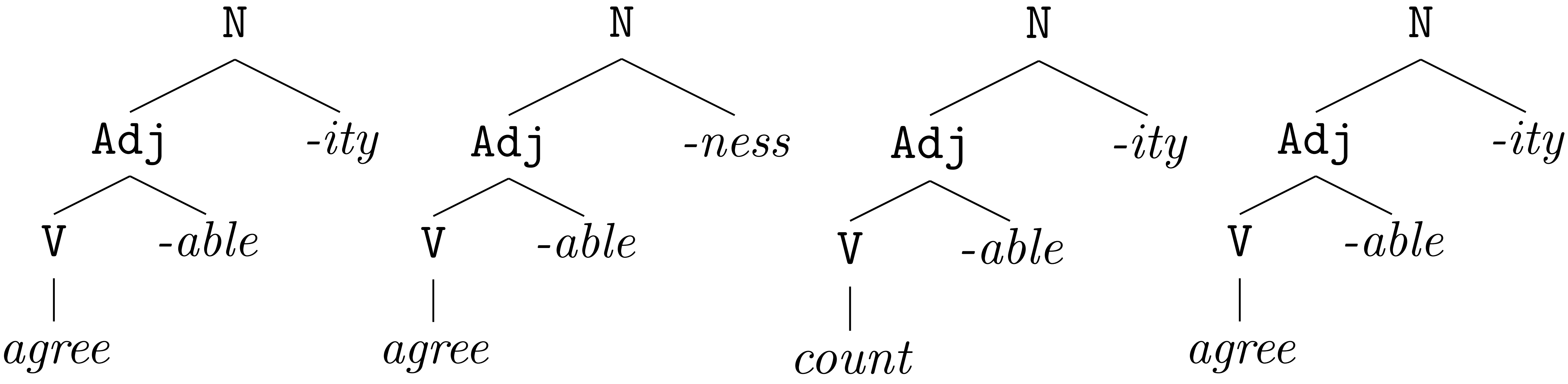
- Generalization of *Adaptor Grammars* (Johnson et al., 2007).
    - Allows storing of partial trees.
- Can be generalized with stochastic variants of *memoization* and *order of evaluation*.



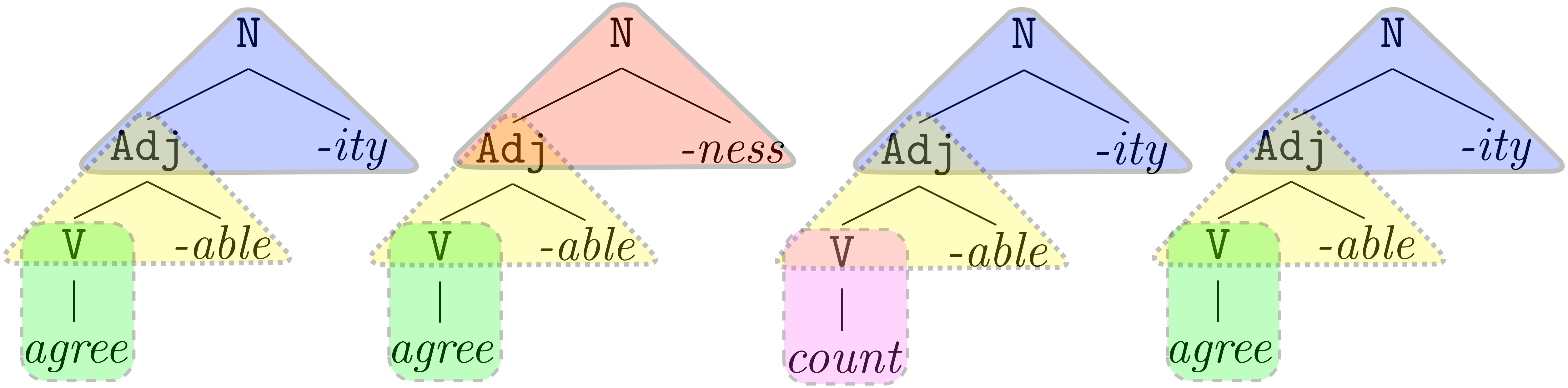
Inference Problem

Search for sets of stored units and derivations of individual forms which best explain the input taking into account preferences for a small inventory of highly reusable units and simple derivations of individual forms.

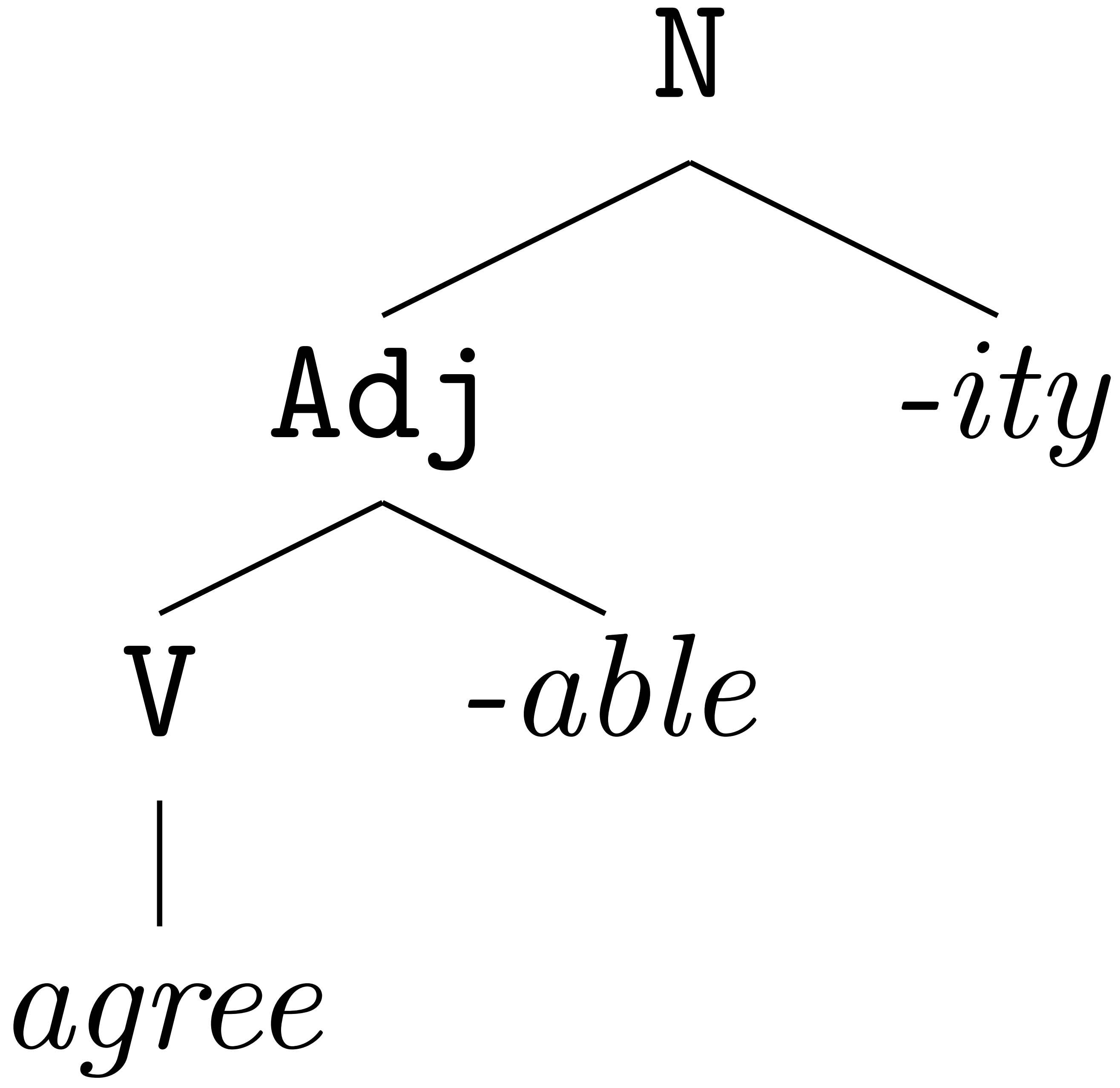
Example Input



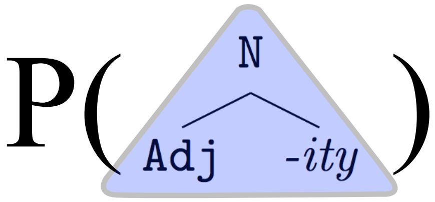
Maximal Decomposition

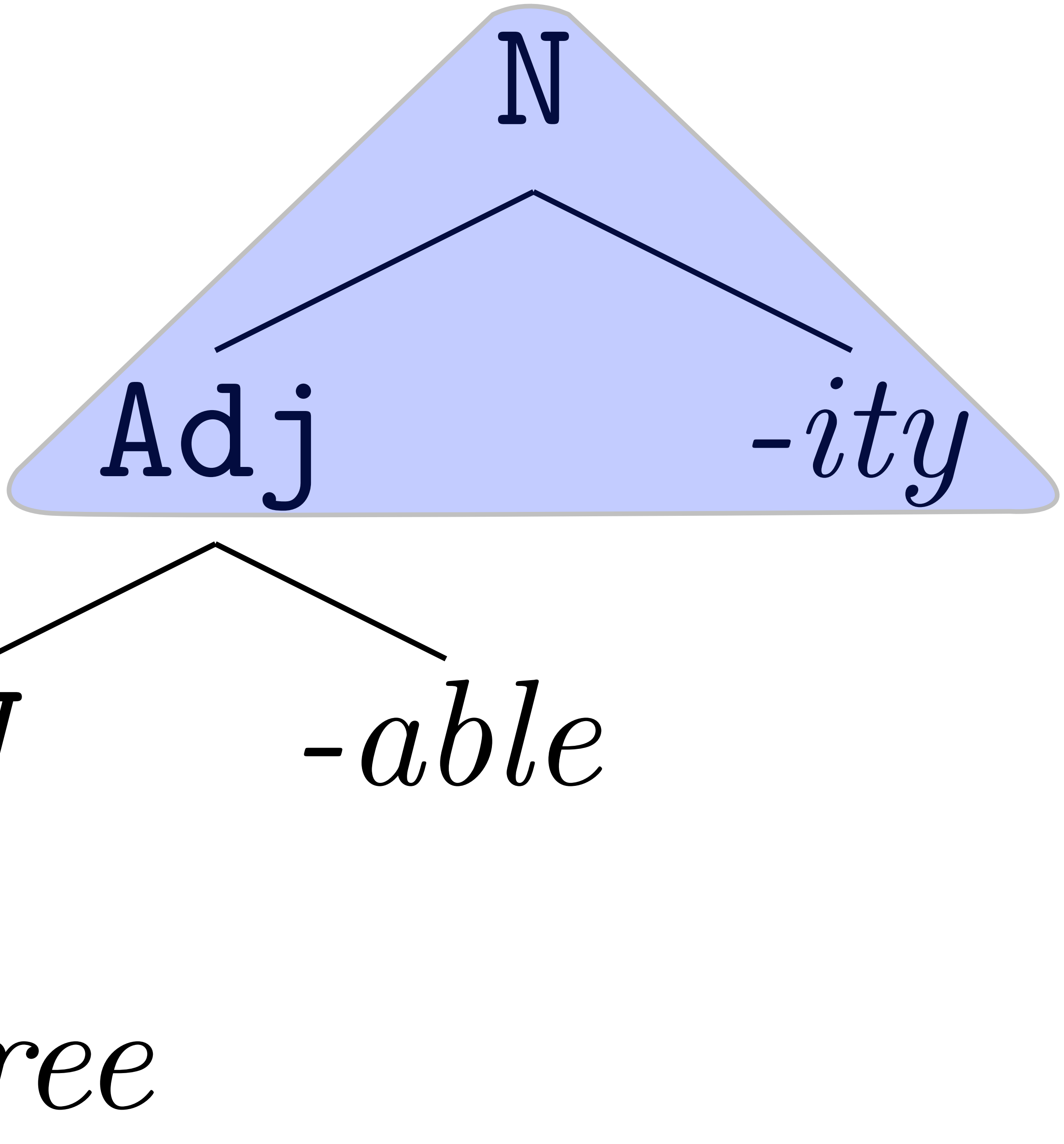


Derivation Complexity

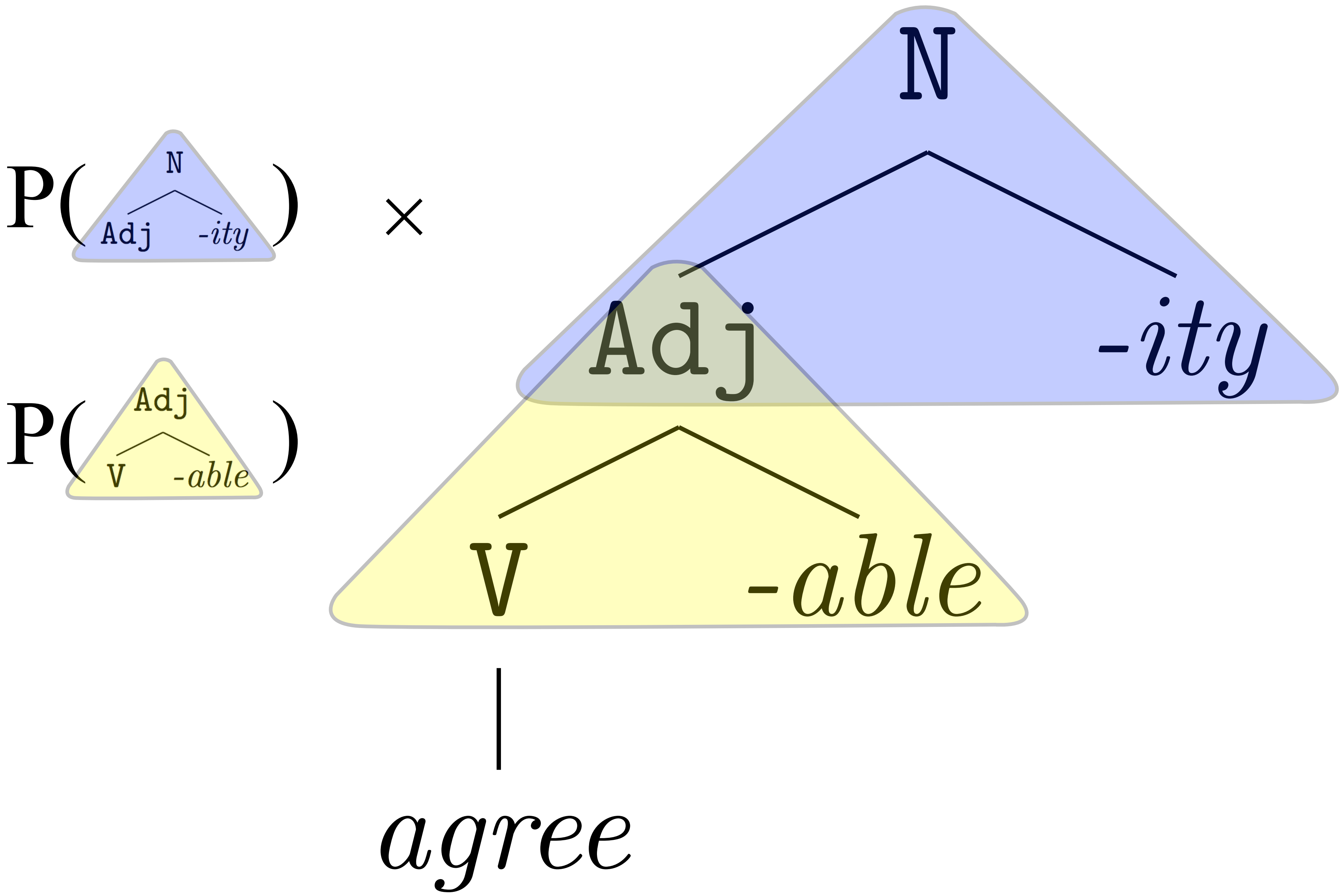


Derivation Complexity

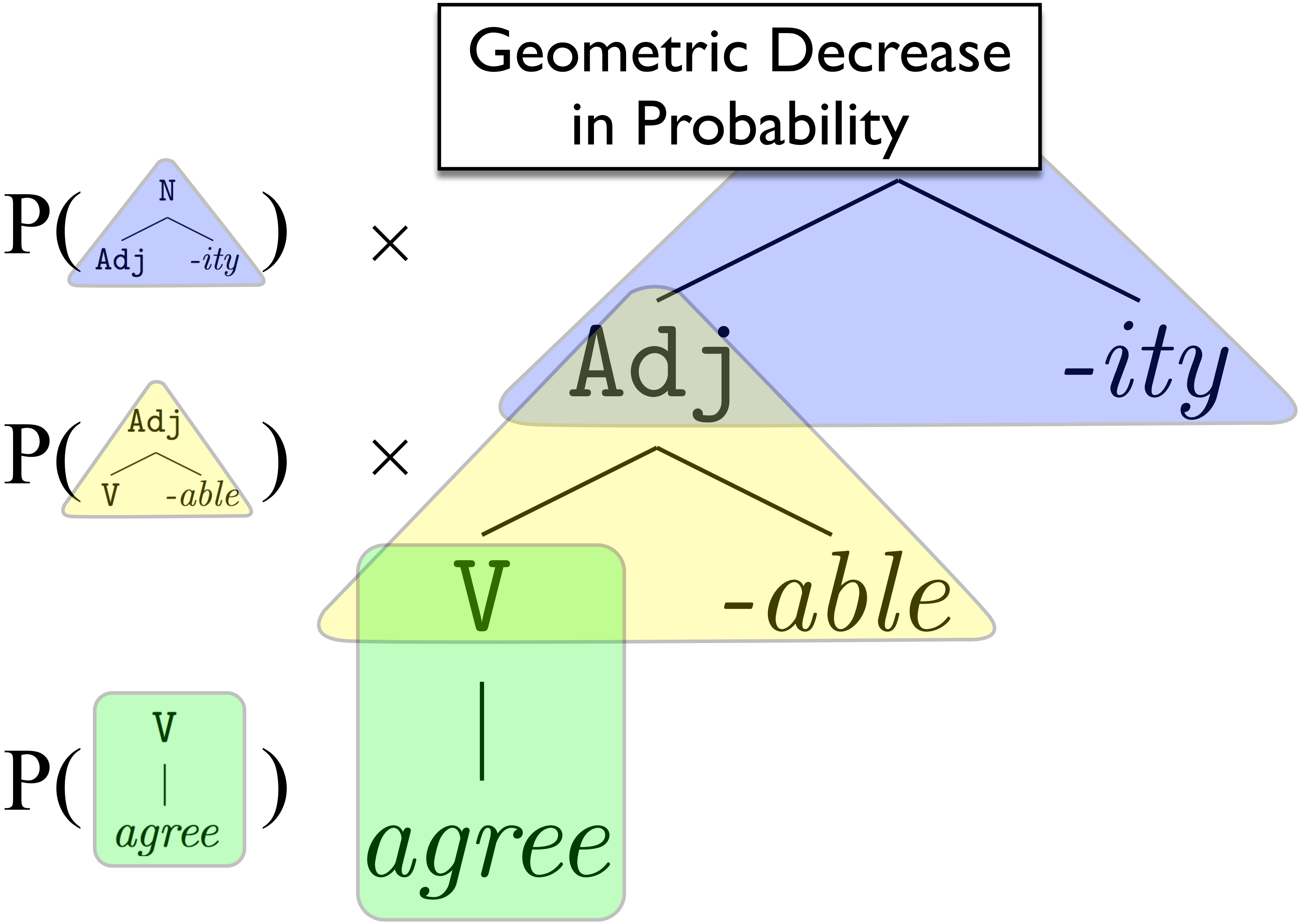




Derivation Complexity

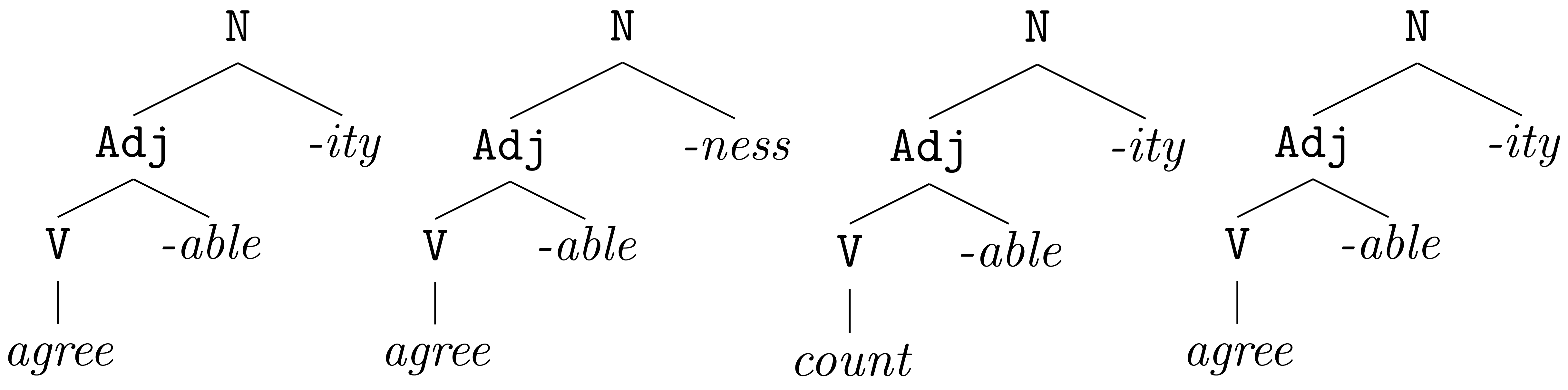


Derivation Complexity

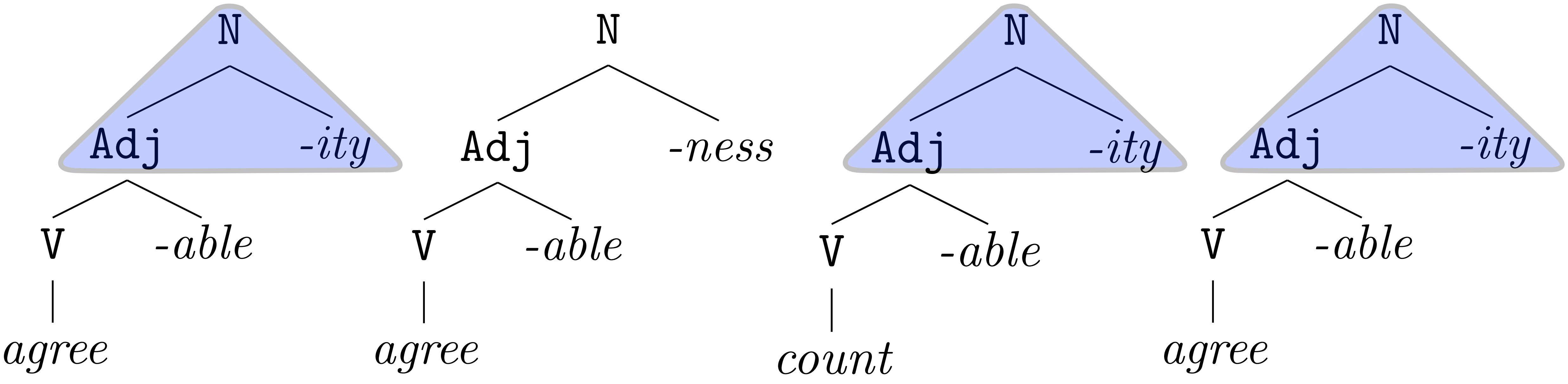


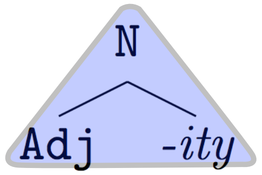
Geometric Decrease in Probability

Sharing Across Expressions

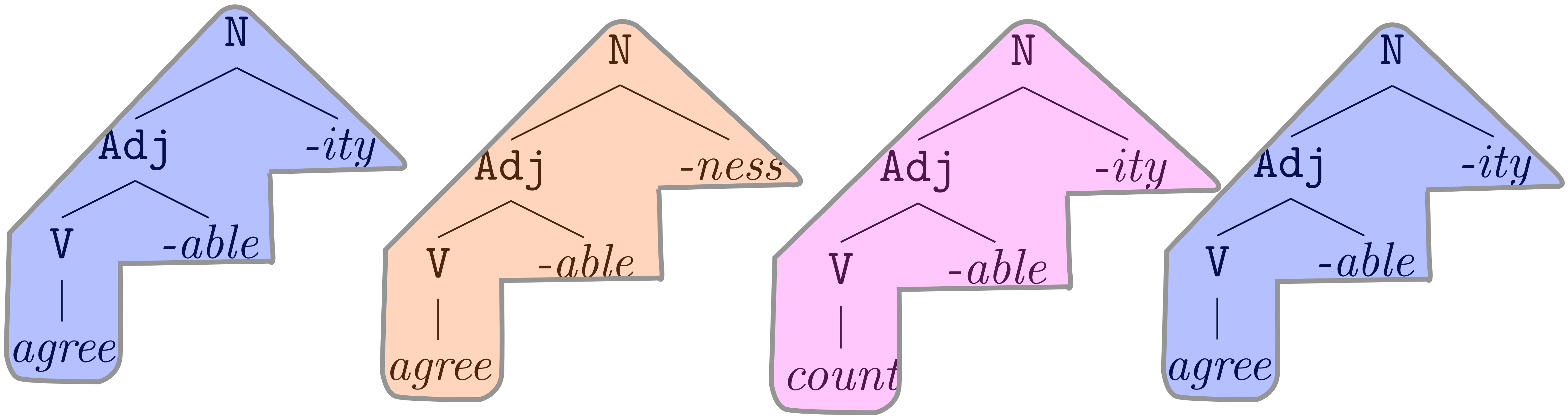


Sharing Across Expressions

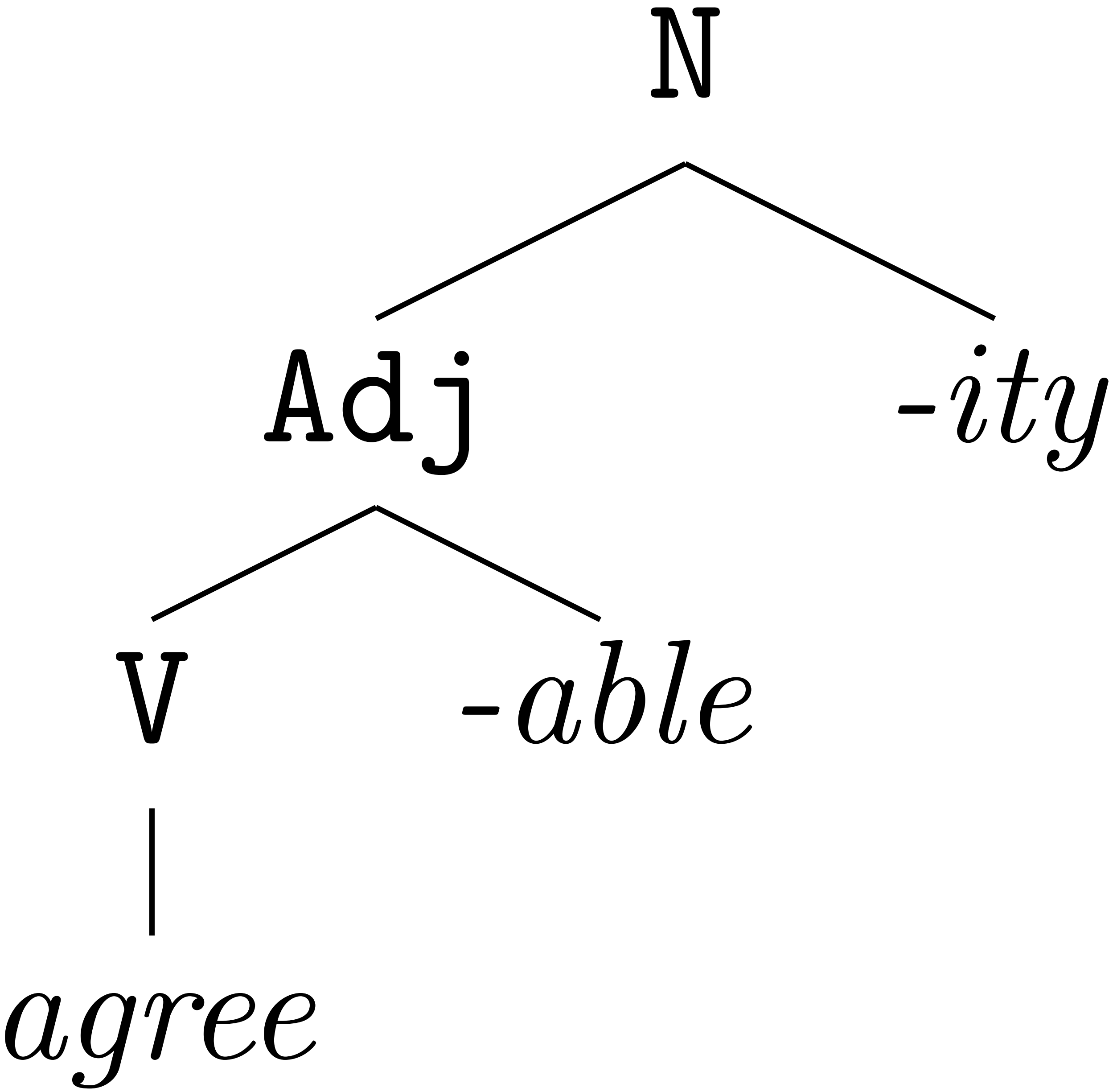


P() ∝ 3

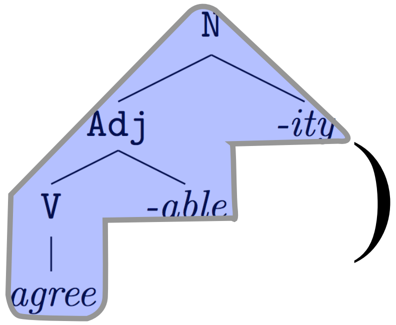
No Decomposition

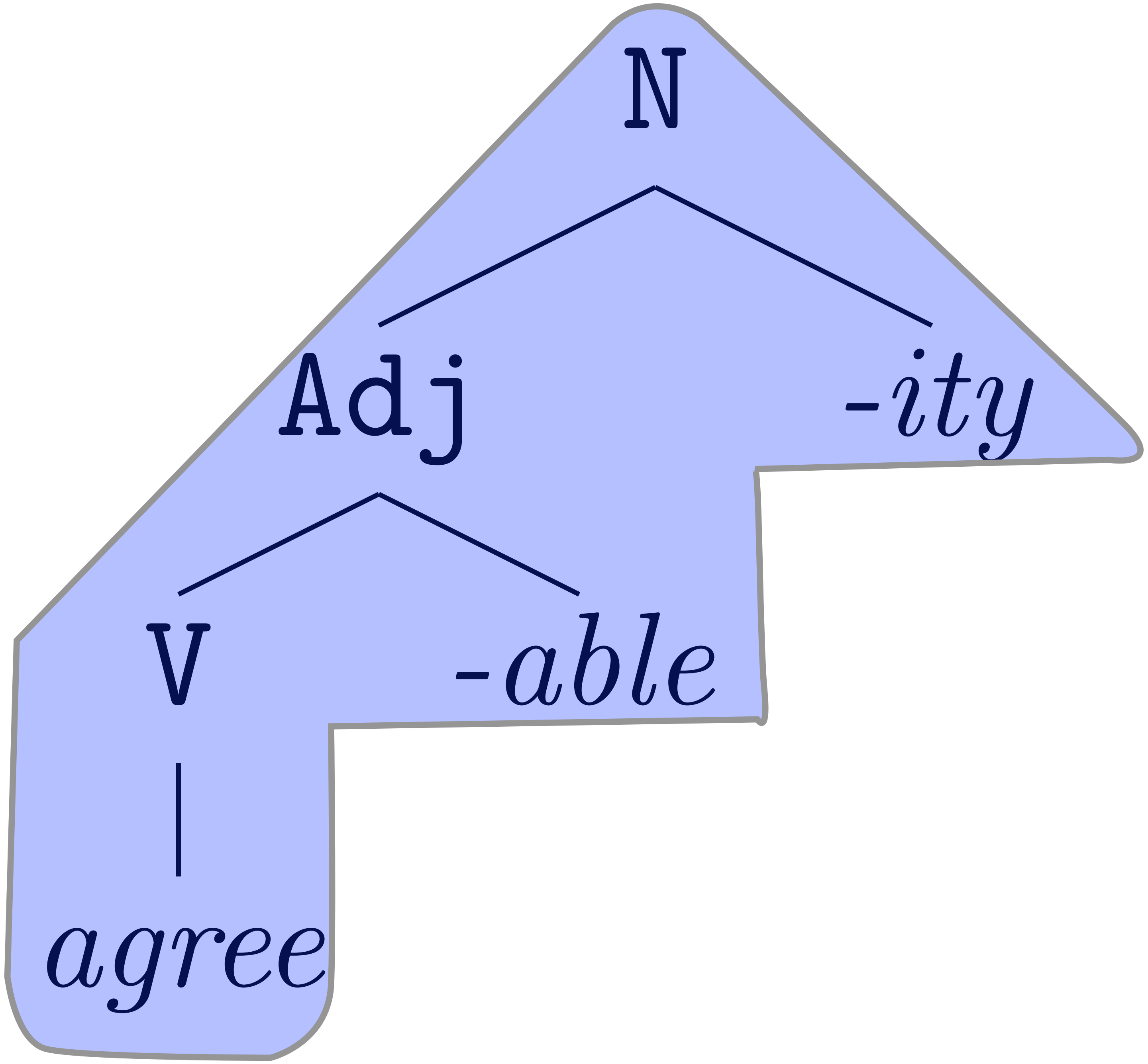


Derivation Complexity

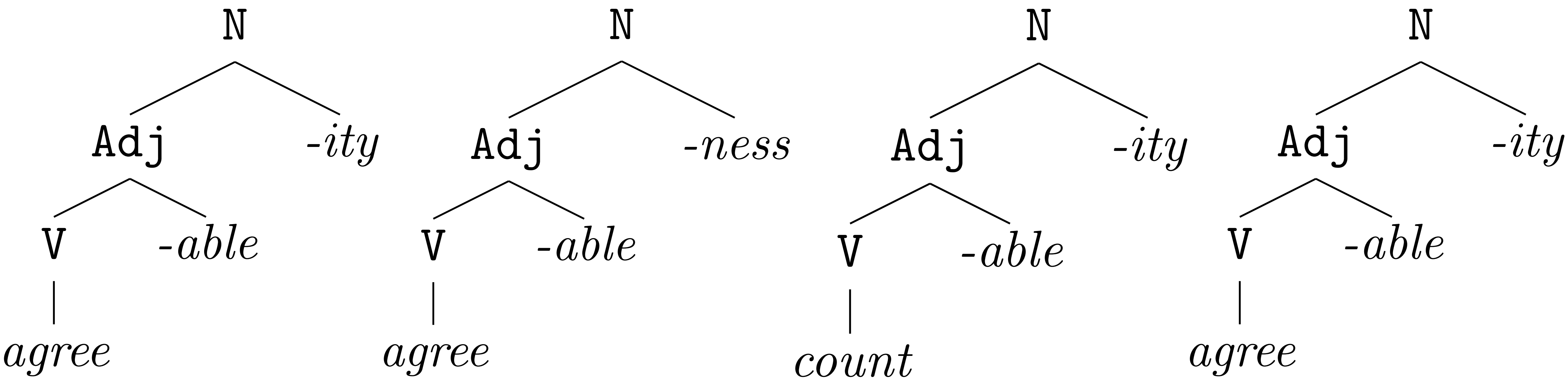


Derivation Complexity

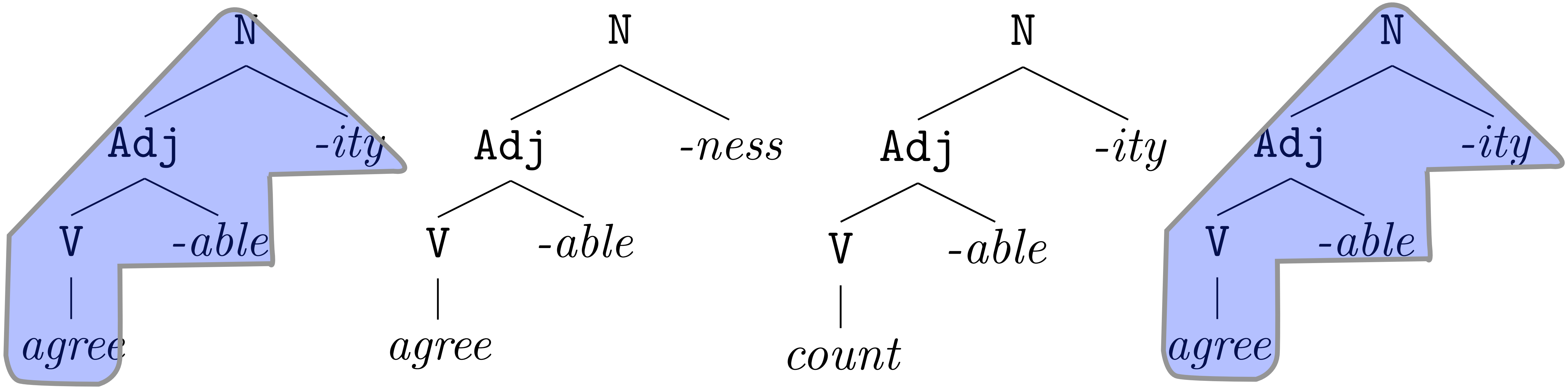
P()

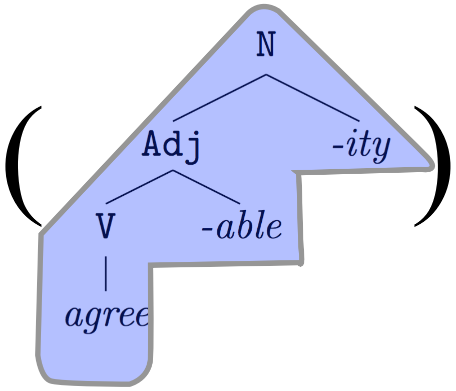


Sharing Across Expressions

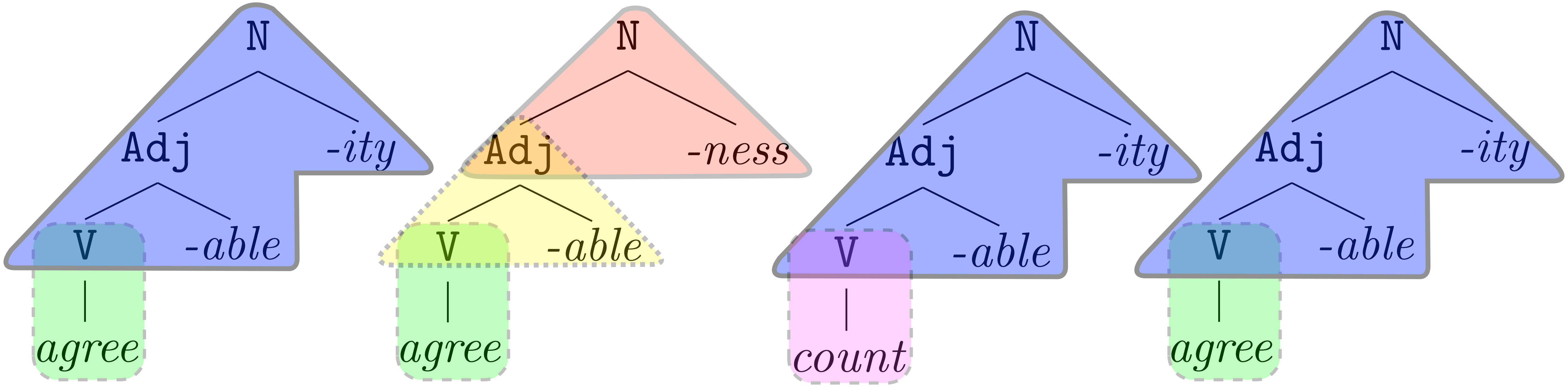


Sharing Across Expressions

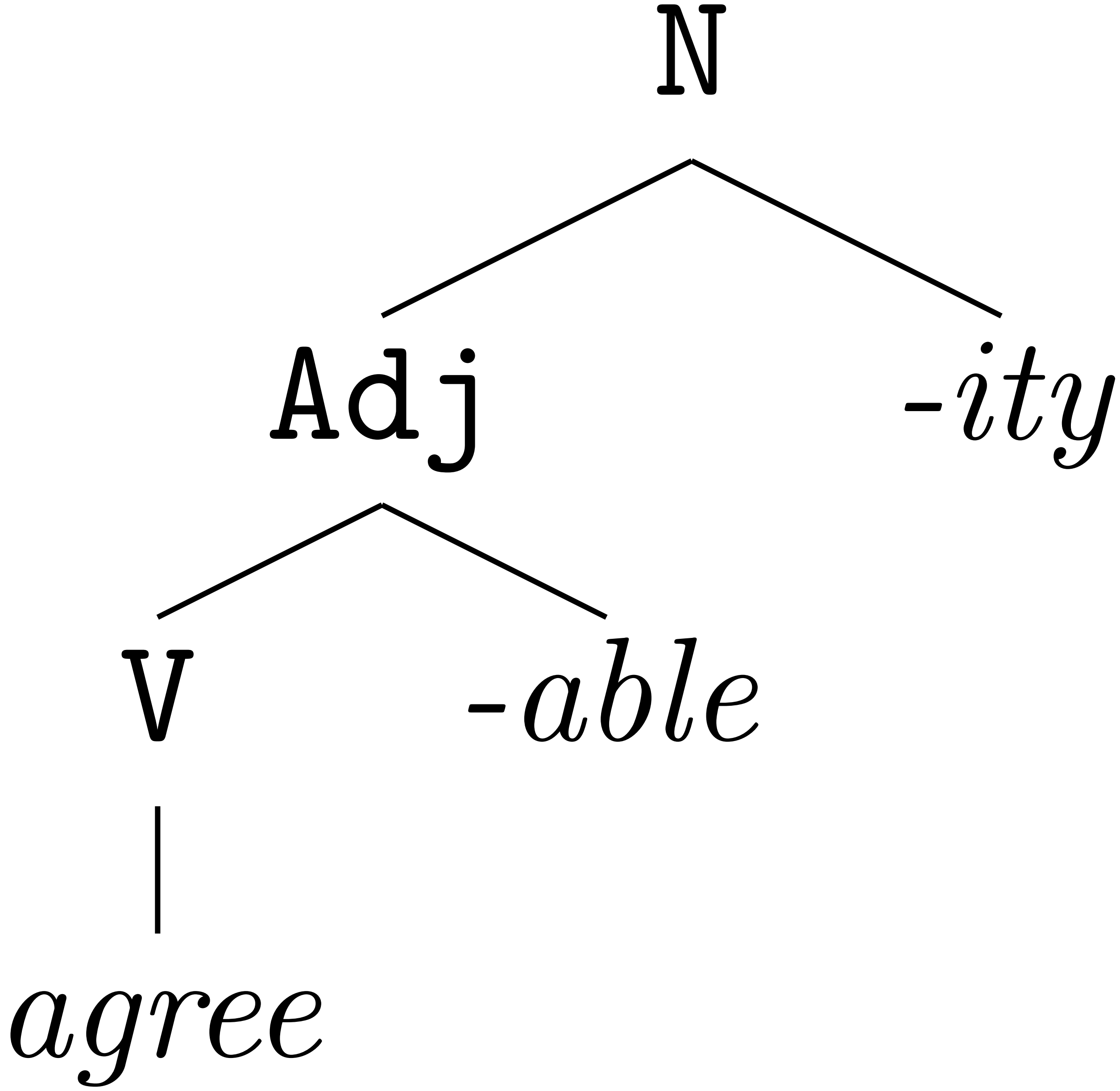


P() ∝ 2

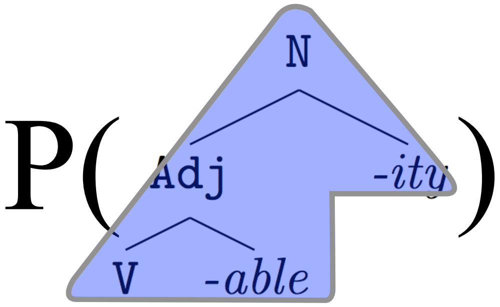
Intermediate Decomposition

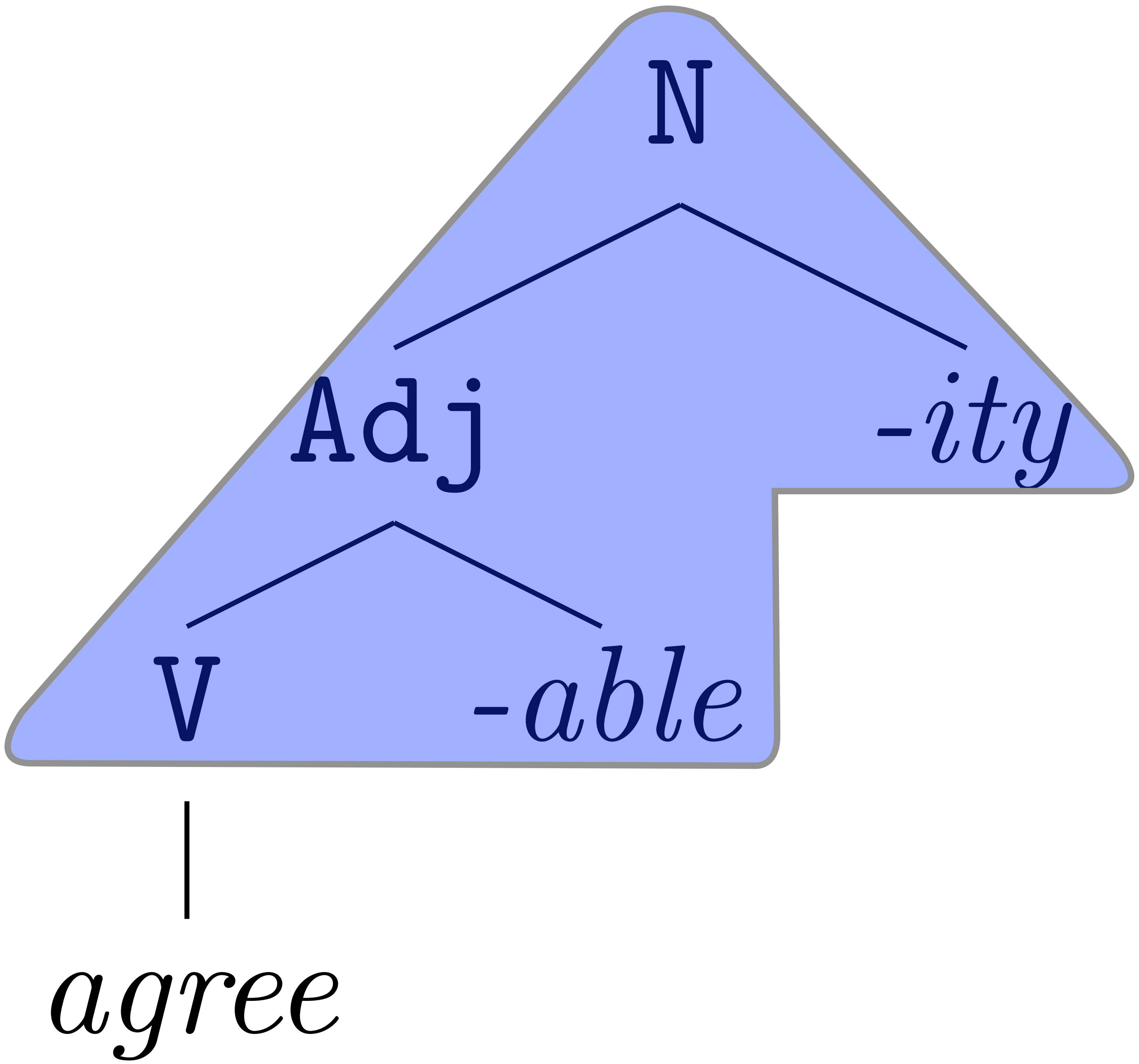


Derivation Complexity

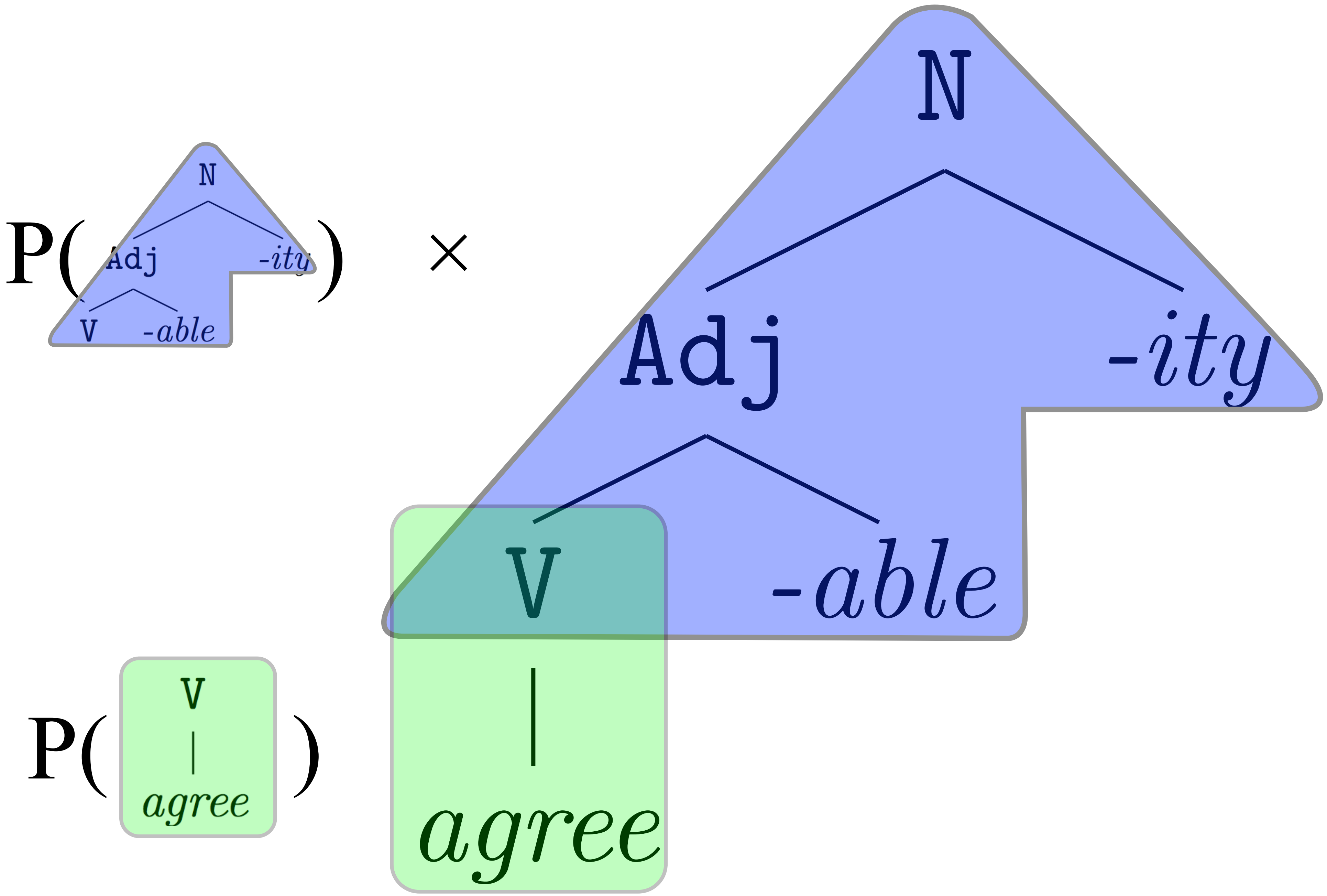


Derivation Complexity

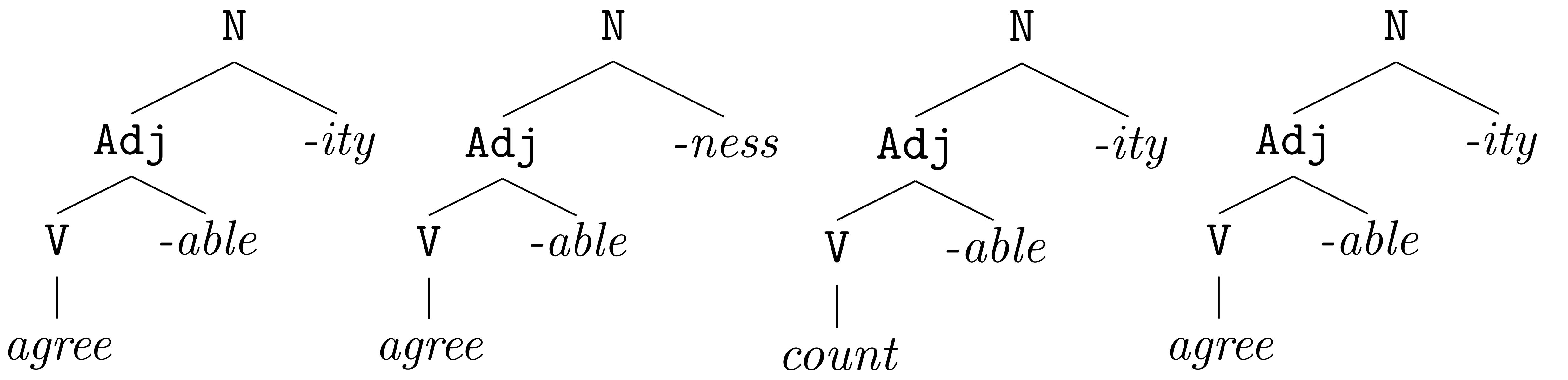
P()



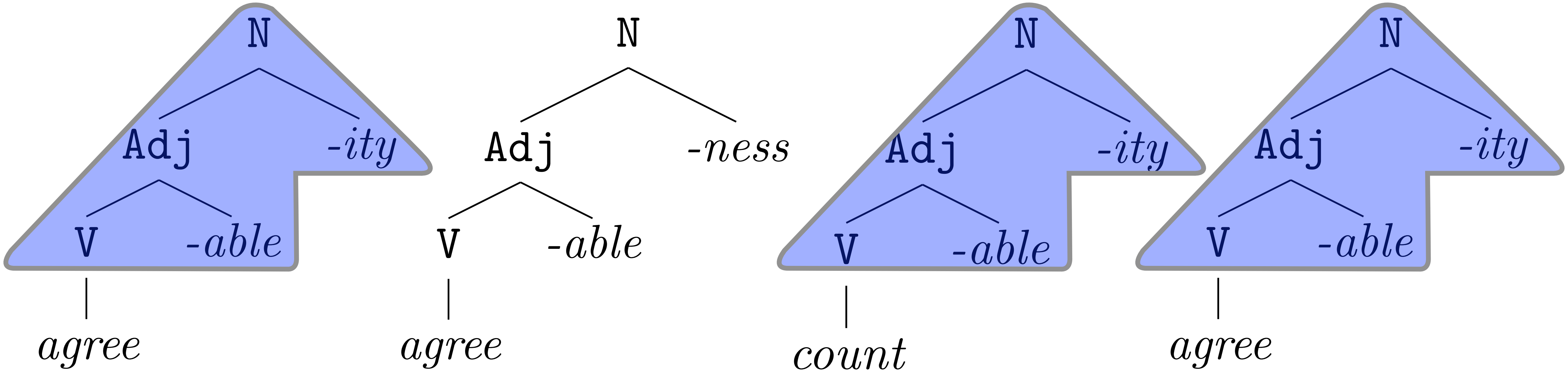
Derivation Complexity

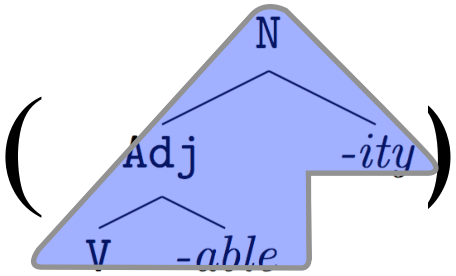


Sharing Across Expressions

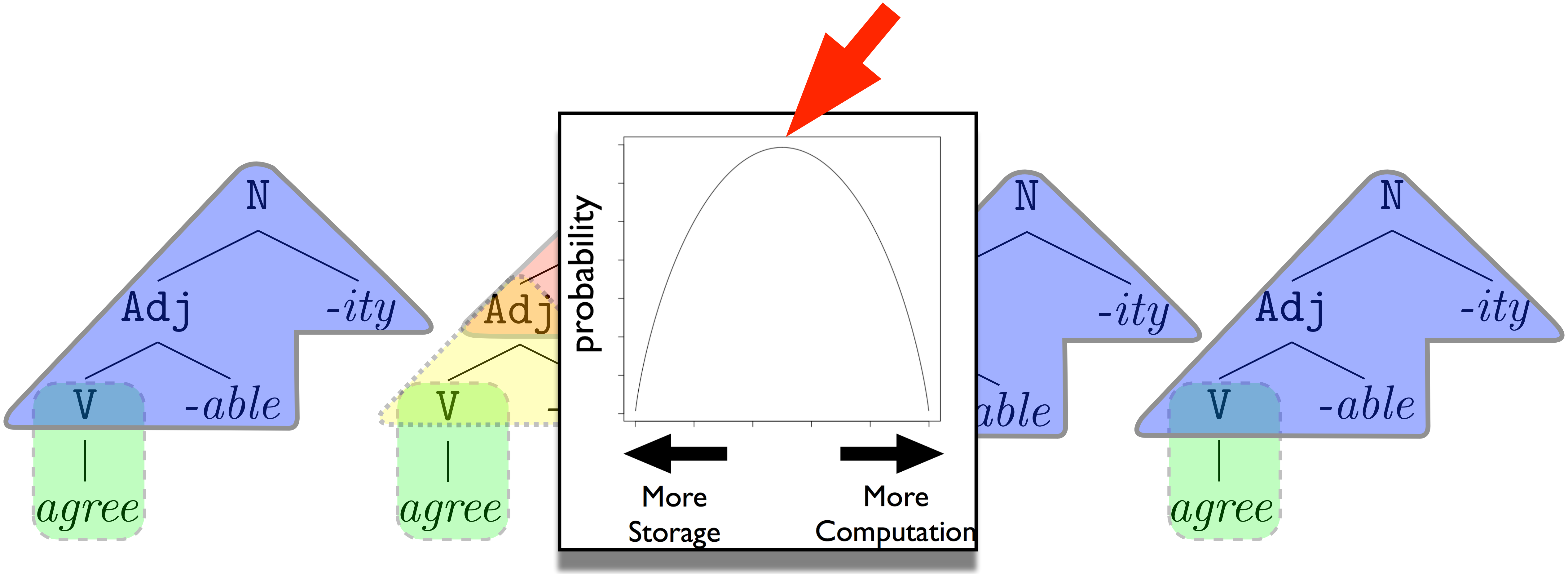


Sharing Across Expressions

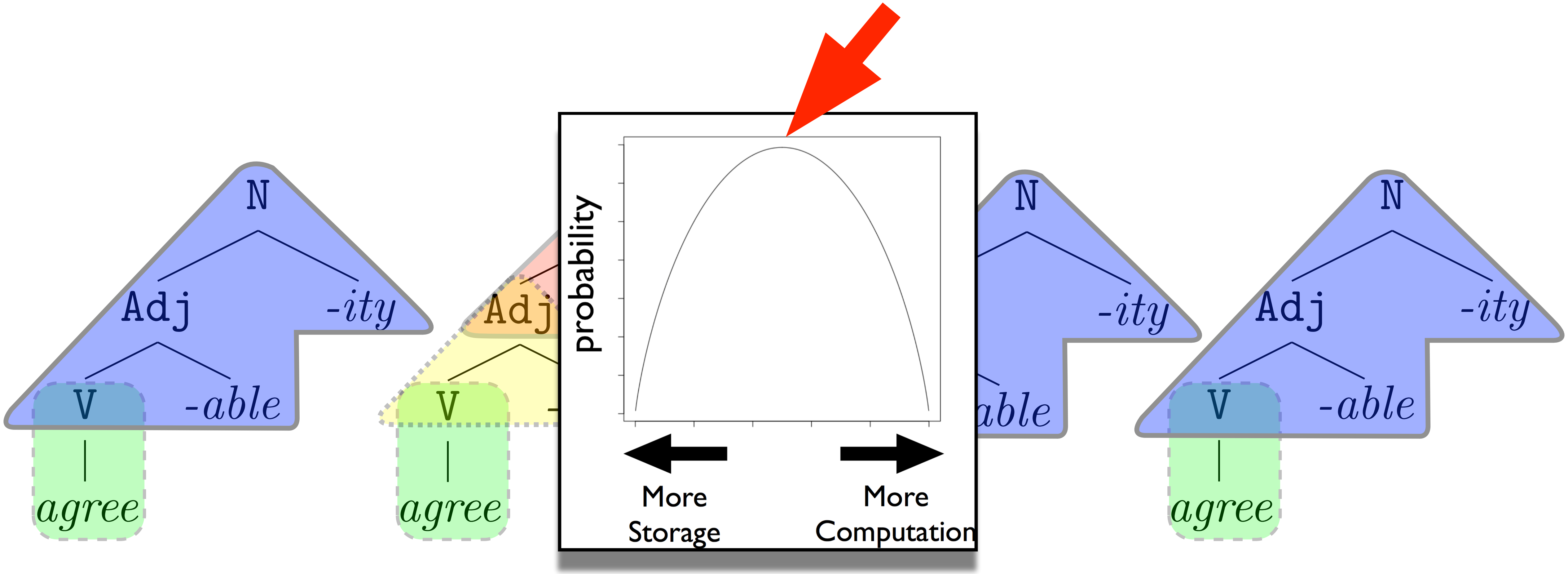


P() ∝ 3

Computation/Storage Tradeoff



Computation/Storage Tradeoff



Builds on both of classic and recent work in non-parametric Bayesian and minimum description length framework.s

Has connections to theory of programming languages, Kolmogorov complexity, and other areas, via probabilistic programming and program induction frameworks.

Overview

Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.

English Derivational Morphology

	Suffix
Productive	-ness
Unproductive	-th
Productive in Combination	-ity

Morphological Productivity

1. **Background**
2. **Productivity and frequency**
3. **Affix ordering**

Morphological Productivity

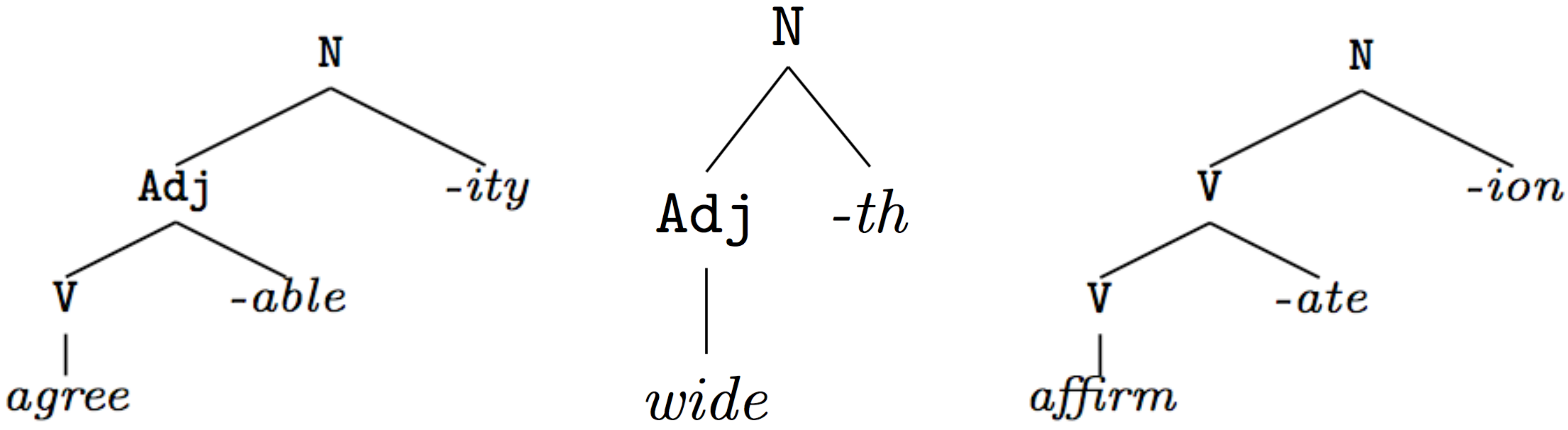
1. **Background**
2. **Productivity and frequency**
3. **Affix ordering**

Idealized Learner Study

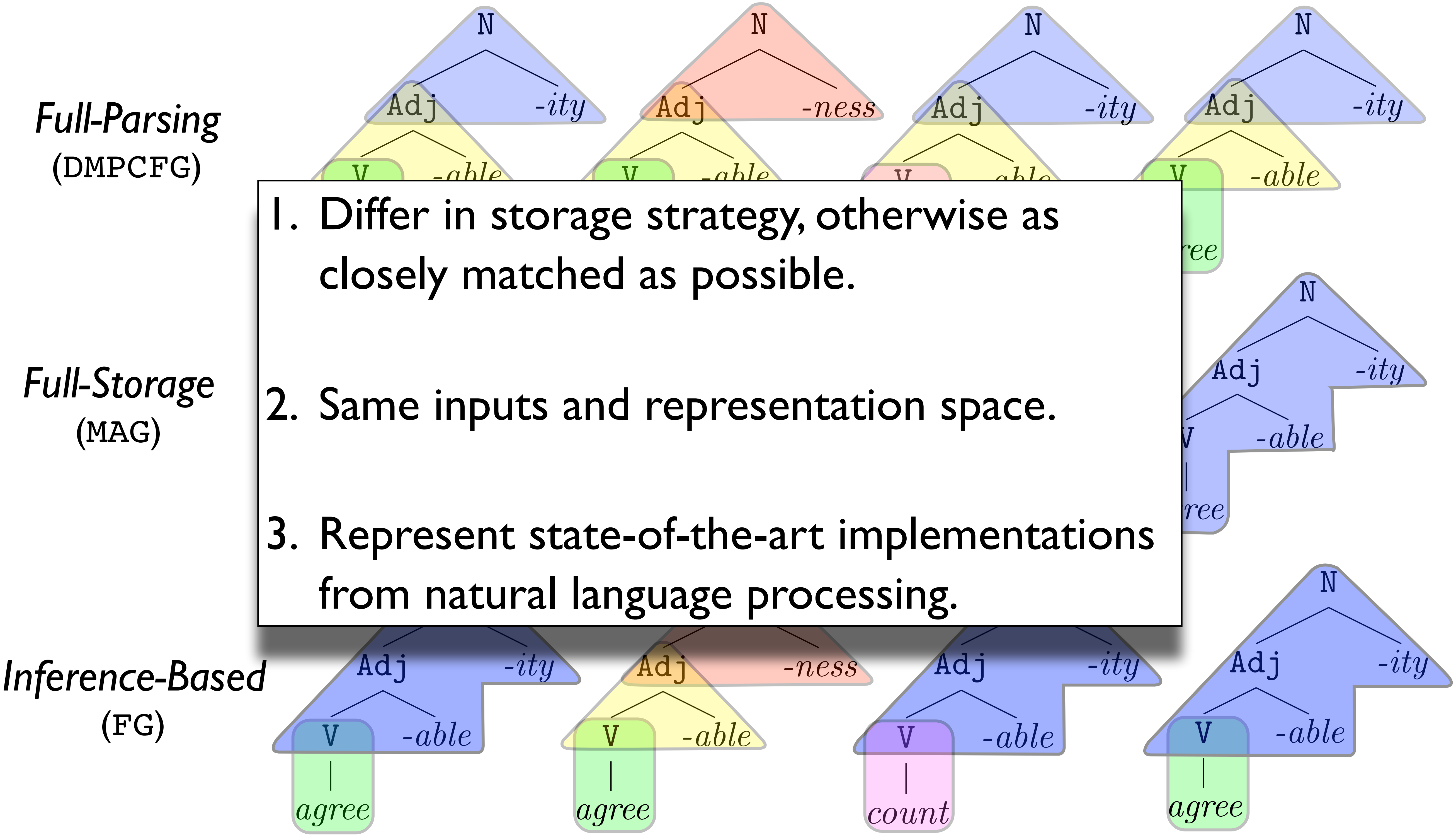
- Not intended to correspond to a particular stage of development or learning procedure used by children.
- **Question**: Does the computation/storage tradeoff provide an effective way of distinguishing between productive units like -*ness* and false generalizations like -*th*?

Input Representations

Phonologically, semantically, and morphosyntactically plausible candidate analyses, erring on the side of shareable structure (decomposition).



Models



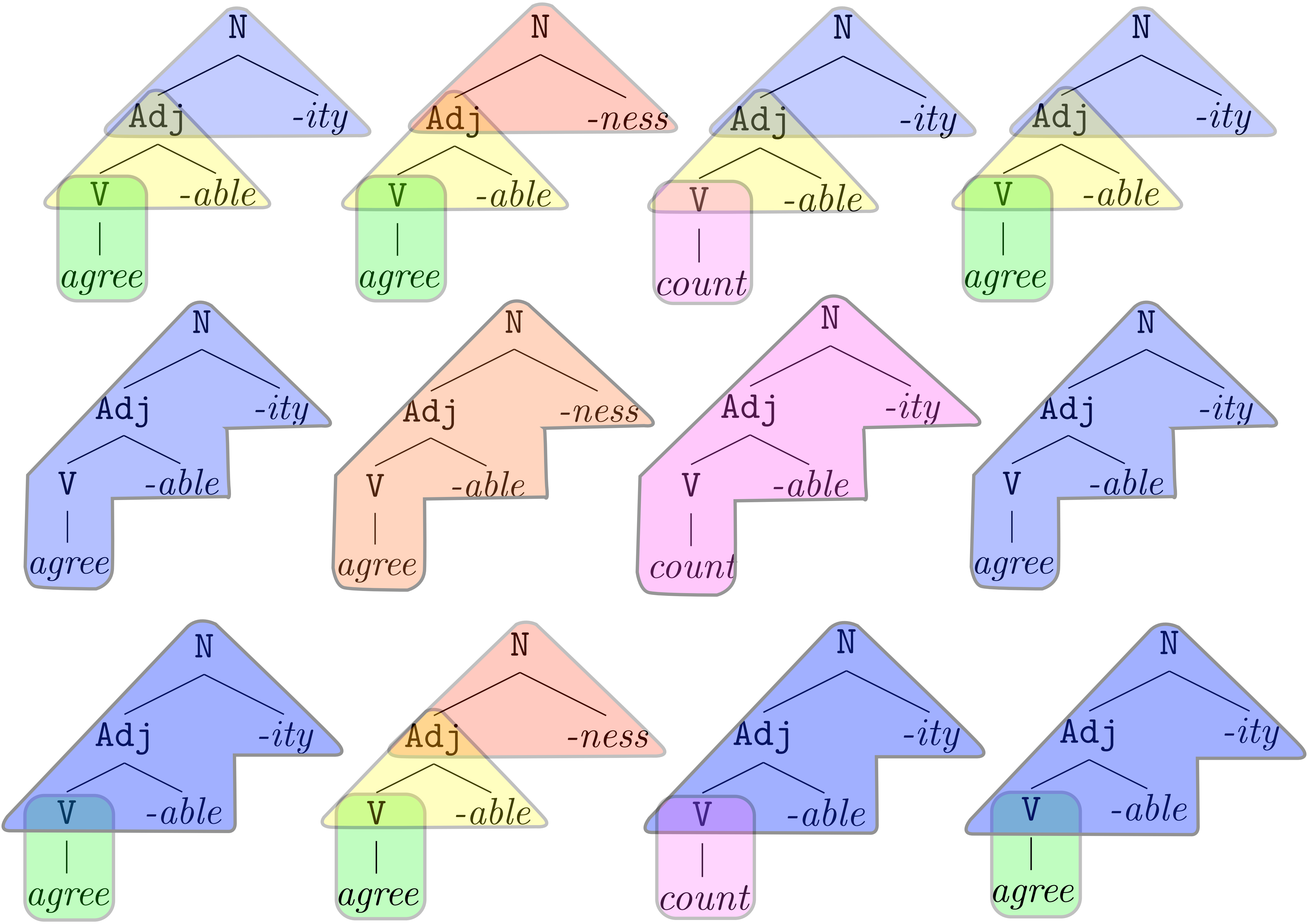
1. Differ in storage strategy, otherwise as closely matched as possible.
2. Same inputs and representation space.
3. Represent state-of-the-art implementations from natural language processing.

Models

Full-Parsing
(DMPCFG)

Full-Storage
(MAG)

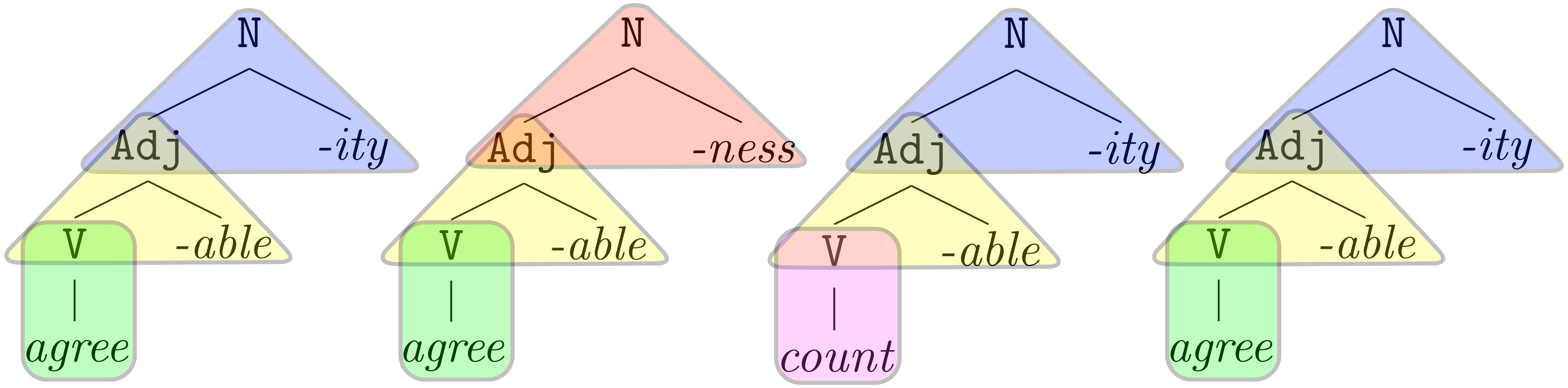
Inference-Based
(FG)



Full-Parsing

(MAP Dirichlet-Multinomial Context-Free Grammars: `DMPCFG`)

Johnson, Griffiths, & Goldwater (2007a)

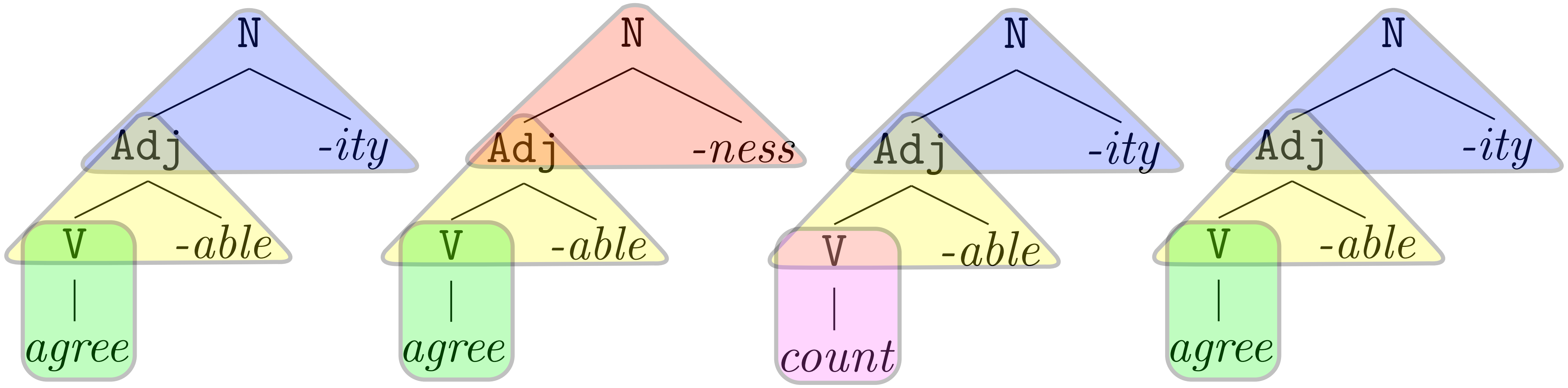


- <u>Storage Strategy</u>: Maximally Decompose.
- Implements radical version of decompositional theories from psychology (e.g., Taft, 1988) and linguistics (e.g., Halle & Marantz, 1983).

Full-Parsing

(MAP Dirichlet-Multinomial Context-Free Grammars: `DMPCFG`)

Johnson, Griffiths, & Goldwater (2007a)

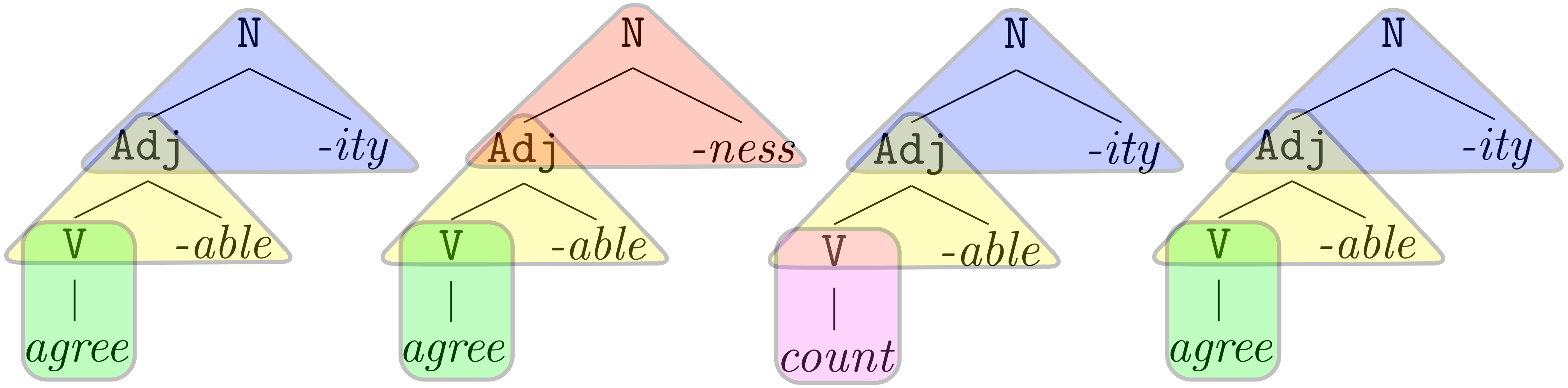


- <u>Storage Strategy</u>: Maximally Decompose.
- Best thought of as a baseline system for morphology.

Full-Parsing

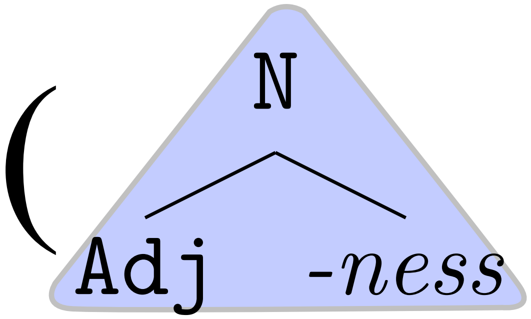
(MAP Dirichlet-Multinomial Context-Free Grammars: `DMPCFG`)

Johnson, Griffiths, & Goldwater (2007a)



- <u>Storage Strategy</u>: Maximally Decompose.

Productivity

P(

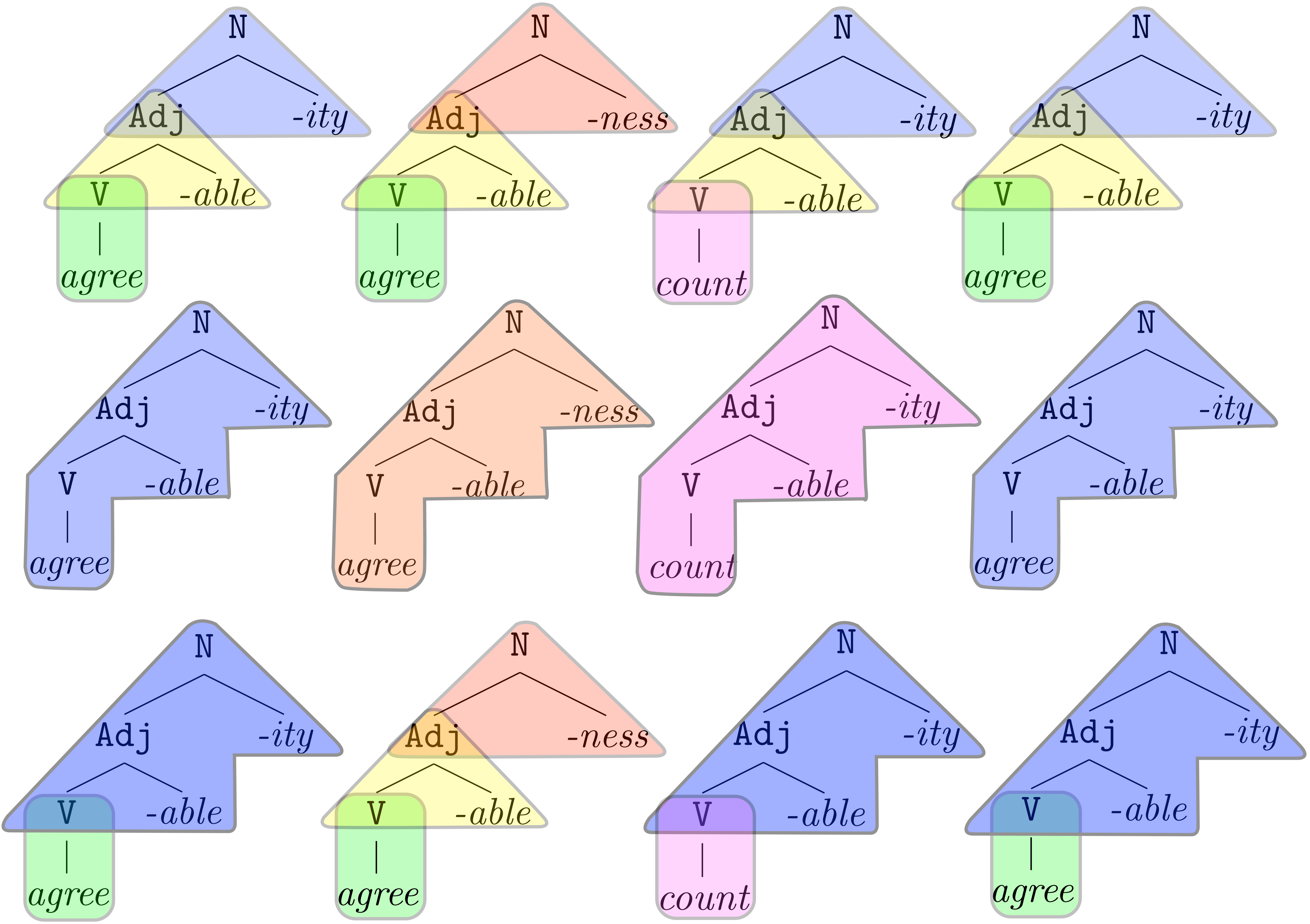
) ∝ Token Frequency

Models

Full-Parsing
(DMPCFG)

Full-Storage
(MAG)

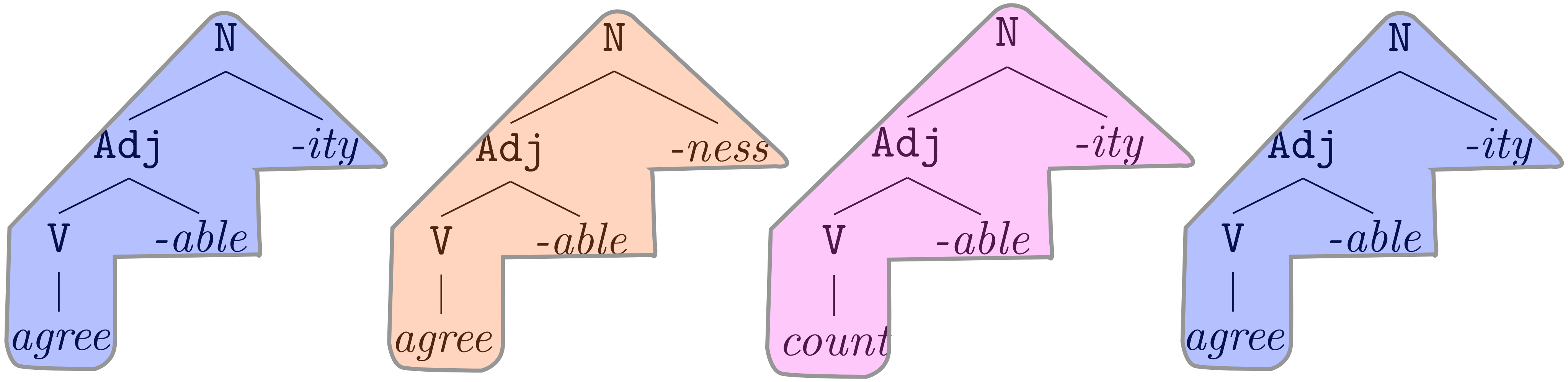
Inference-Based
(FG)



Full-Storage

(MAP All-Adapted Adaptor Grammars: `MAG`)

Johnson, Griffiths, & Goldwater (2007)

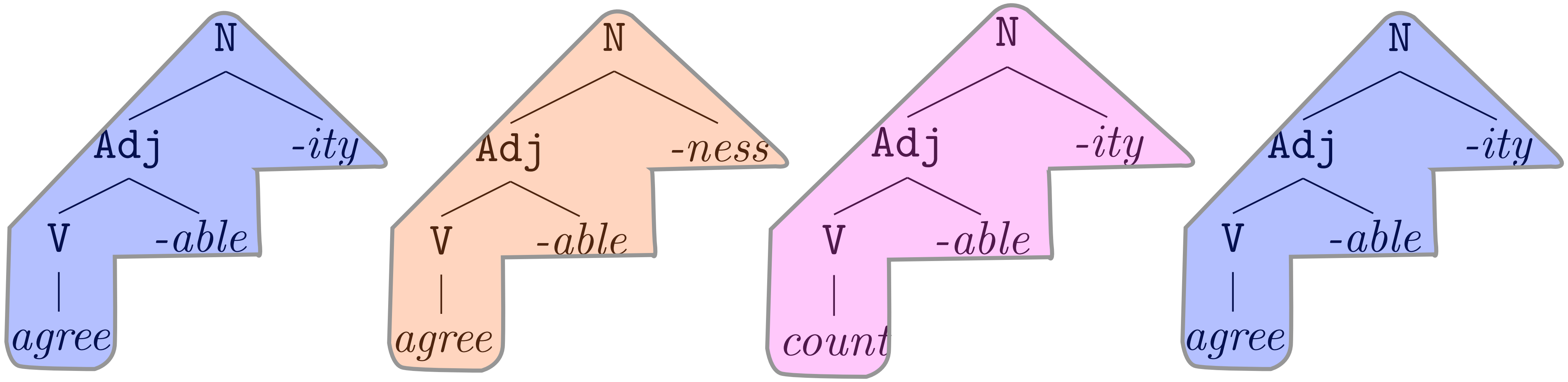


- <u>Storage Strategy</u>: Store structures in their entirety after first time generated.

Full-Storage

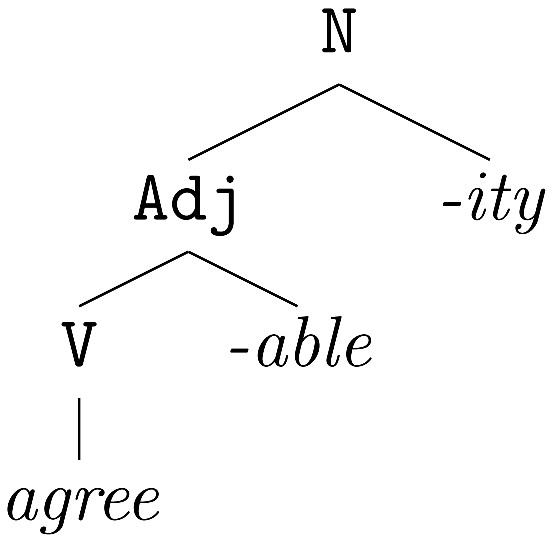
(MAP All-Adapted Adaptor Grammars: `MAG`)

Johnson, Griffiths, & Goldwater (2007)



- Storage Strategy: Store structures in their entirety after first time generated.

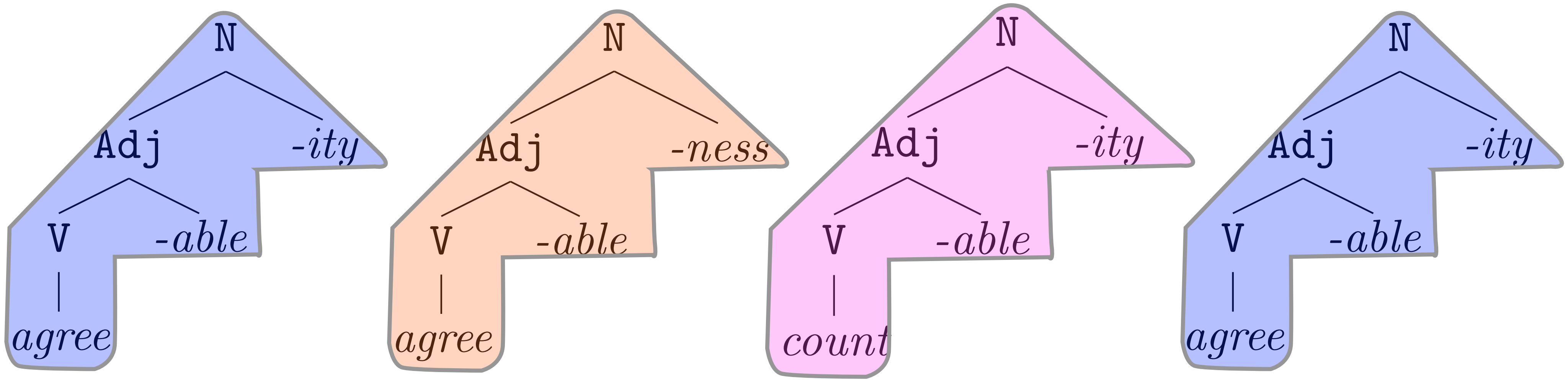
All full sub-trees are stored recursively.



Full-Storage

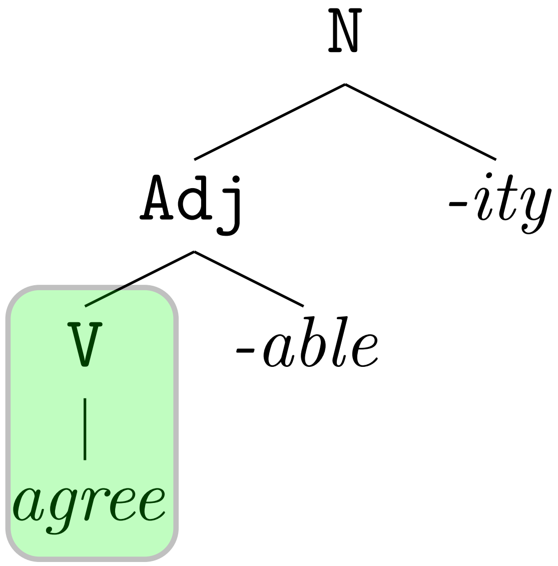
(MAP All-Adapted Adaptor Grammars: MAG)

Johnson, Griffiths, & Goldwater (2007)



- Storage Strategy: Store structures in their entirety after first time generated.

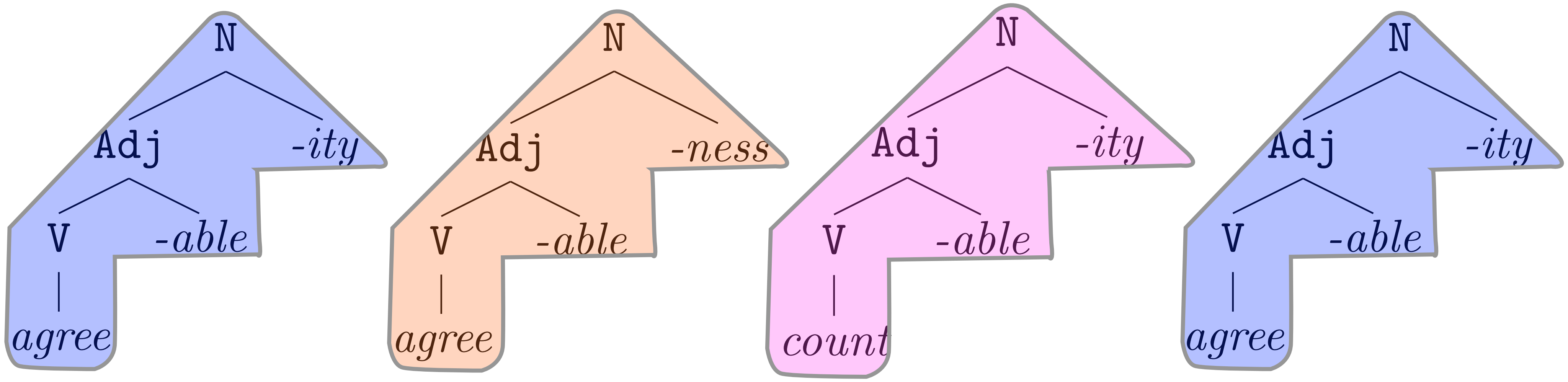
All full sub-trees are stored recursively.



Full-Storage

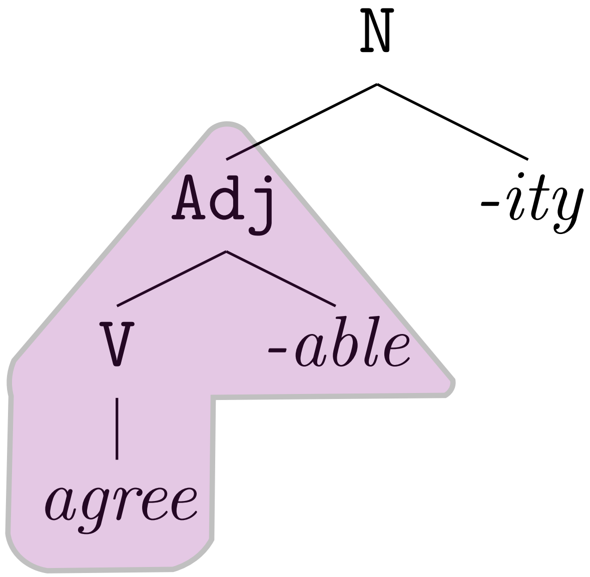
(MAP All-Adapted Adaptor Grammars: **MAG**)

Johnson, Griffiths, & Goldwater (2007)



- <u>Storage Strategy</u>: Store structures in their entirety after first time generated.

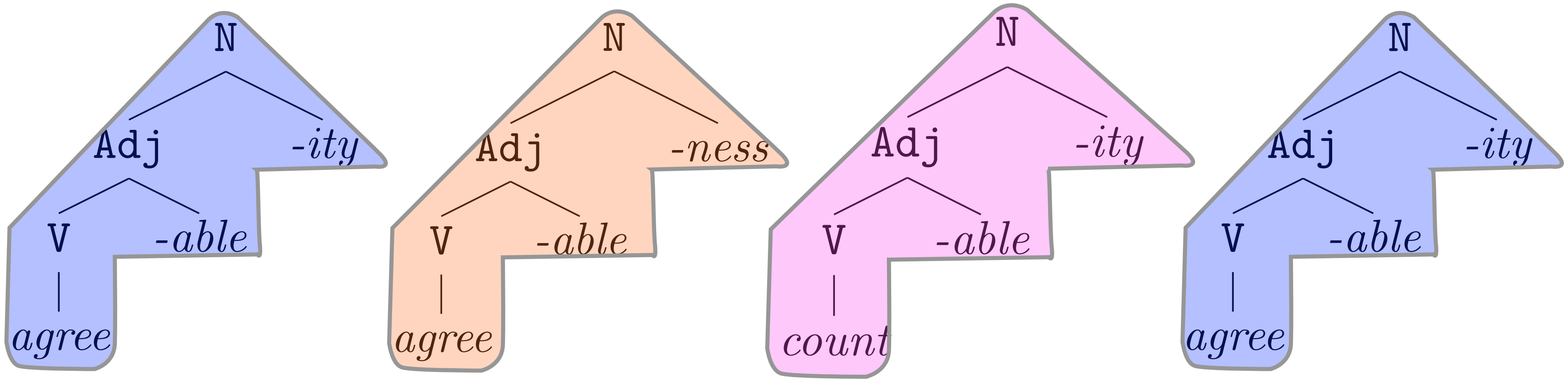
All full sub-trees are stored recursively.



Full-Storage

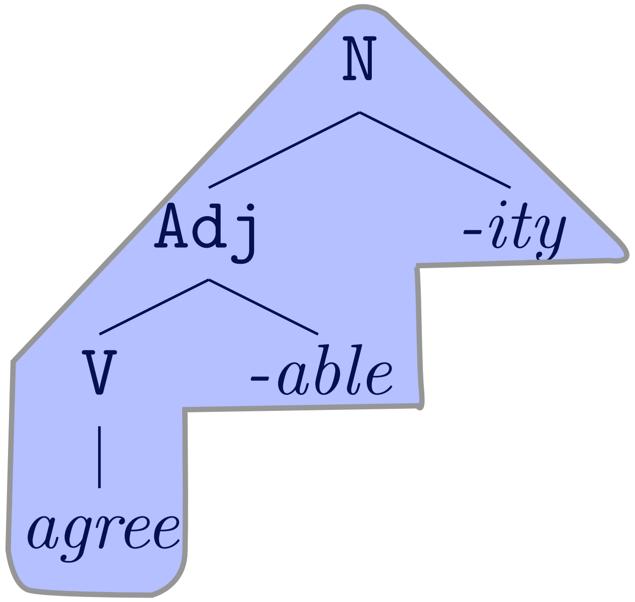
(MAP All-Adapted Adaptor Grammars: `MAG`)

Johnson, Griffiths, & Goldwater (2007)



- <u>Storage Strategy</u>: Store structures in their entirety after first time generated.

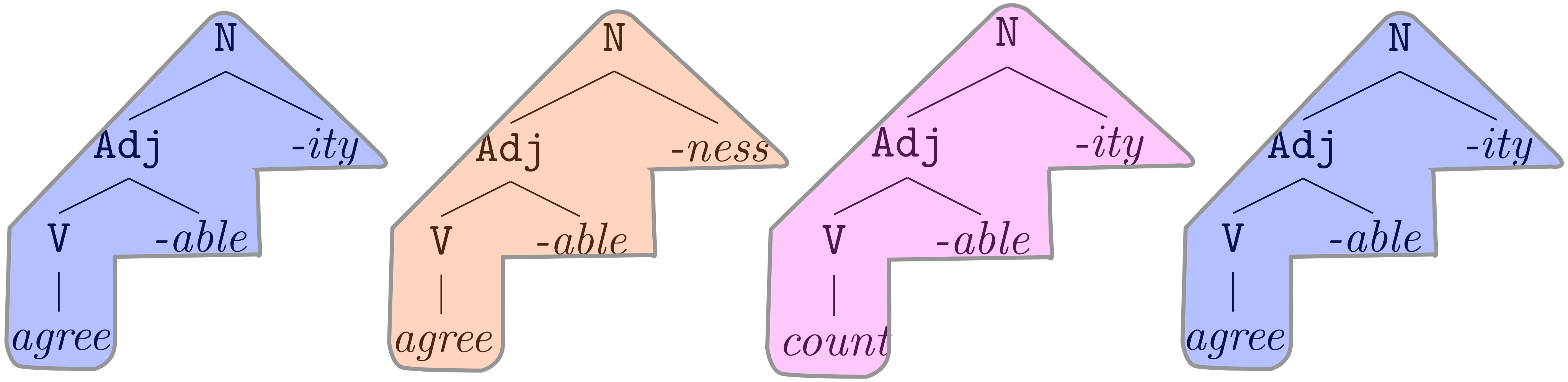
All full sub-trees are stored recursively.



Full-Storage

(MAP All-Adapted Adaptor Grammars: `MAG`)

Johnson, Griffiths, & Goldwater (2007)



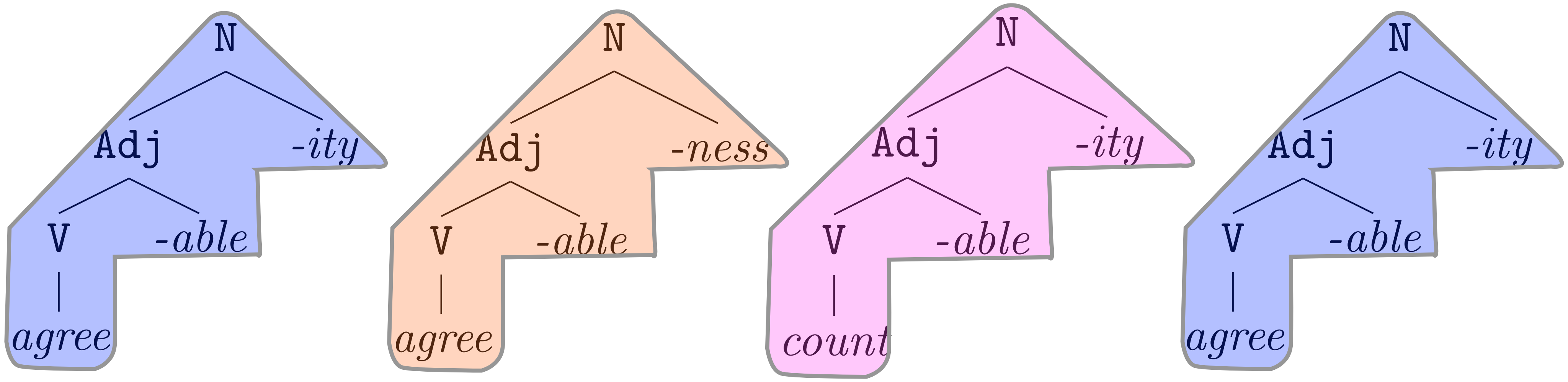
- <u>Storage Strategy</u>: Store structures in their entirety after first time generated.

- Modern probabilistic implementation of classical lexical redundancy rules (e.g., Jackendoff, 1975; Aronoff, 1976).

Full-Storage

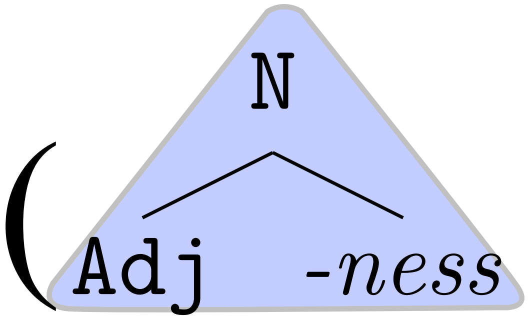
(MAP All-Adapted Adaptor Grammars: `MAG`)

Johnson, Griffiths, & Goldwater (2007)



- Storage Strategy: Store structures in their entirety after first time generated.

Productivity

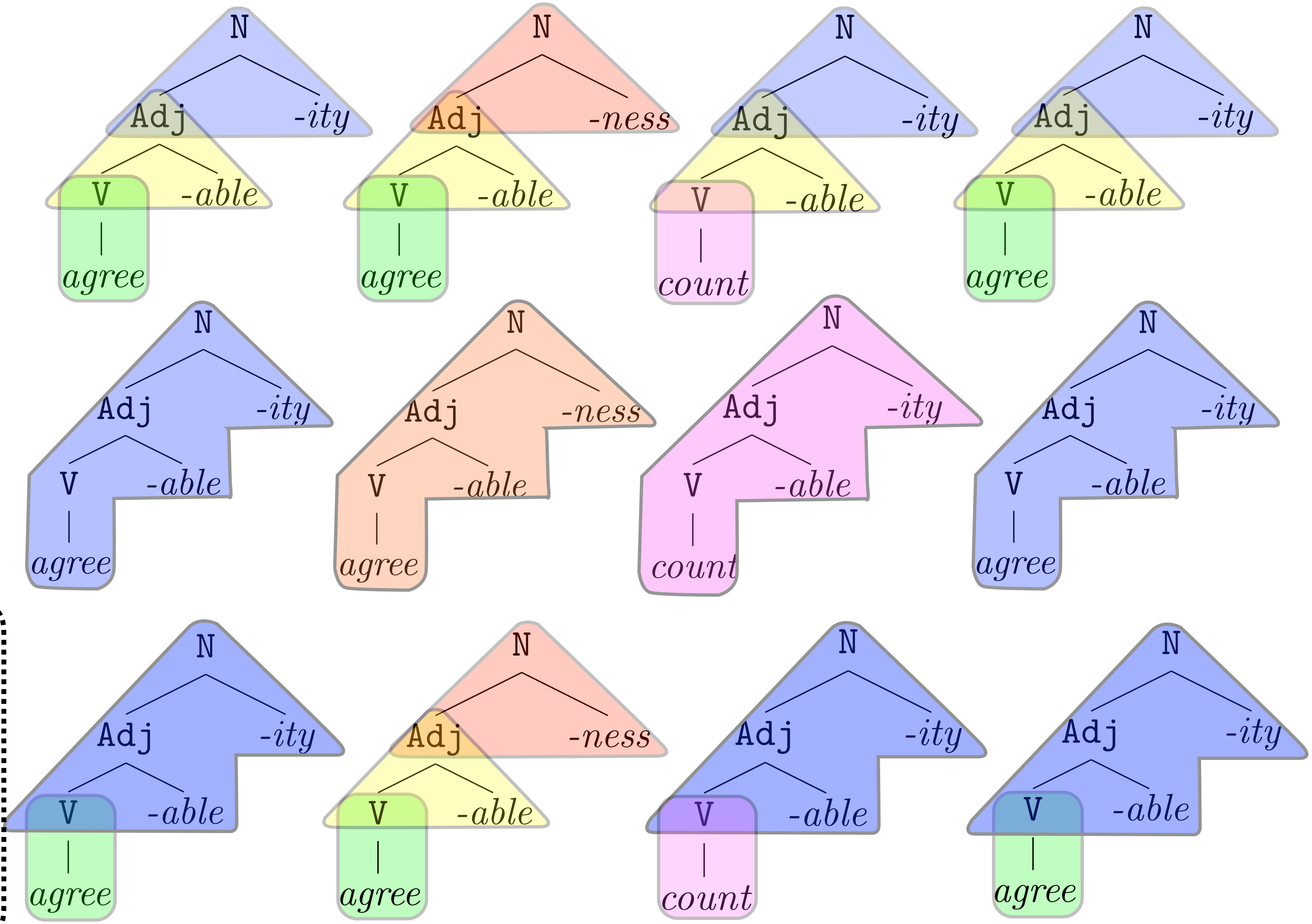
P() ∝ Type frequency

Models

Full-Parsing
(DMPCFG)

Full-Storage
(MAG)

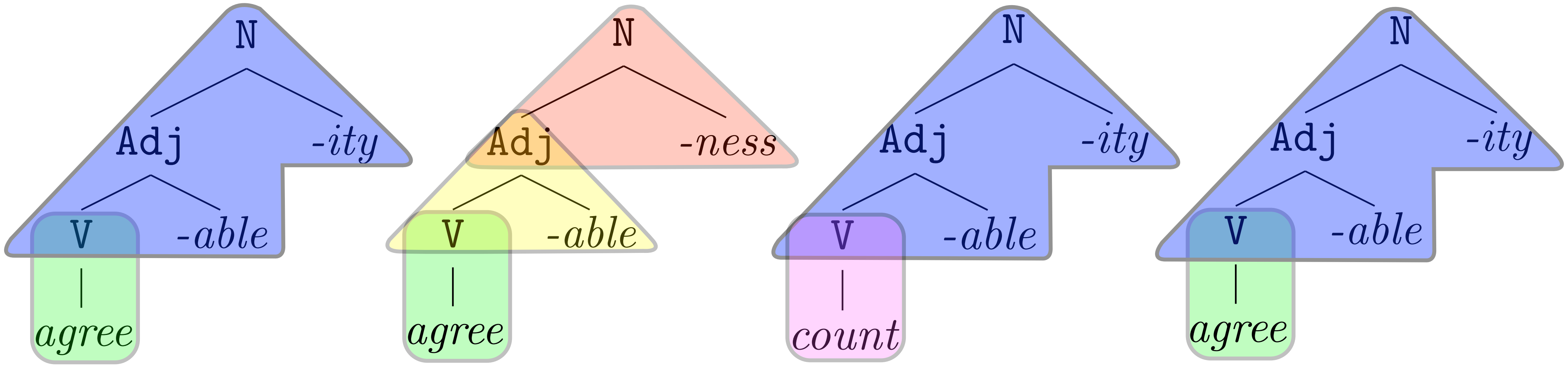
Inference-Based
(FG)



Inference-Based

(Fragment Grammars: **FG**)

O’Donnell et al. (2009); O’Donnell (2011, 2015)

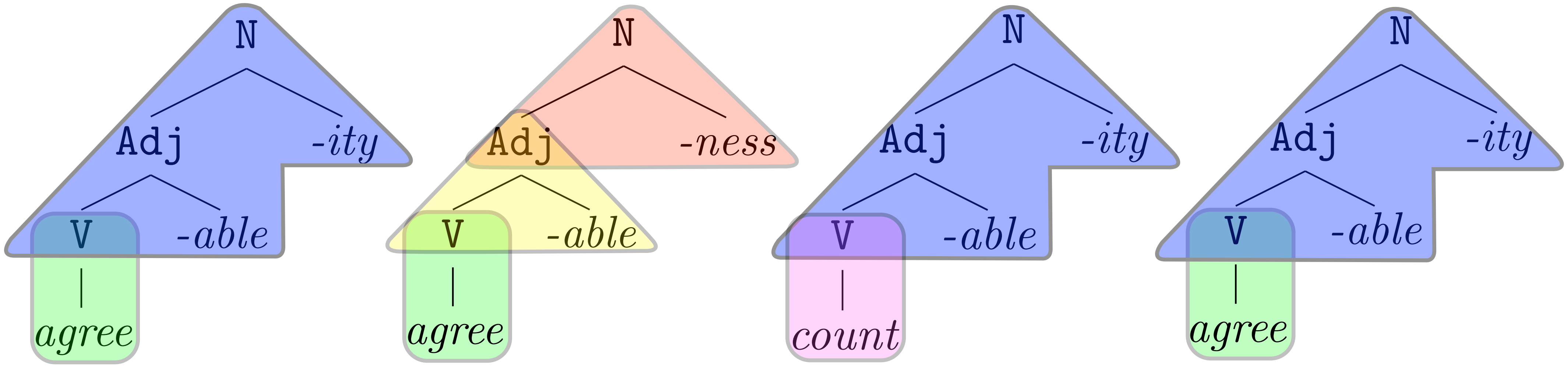


- Storage Strategy: Store set of units that best explains data (only inference-based storage proposal).

Inference-Based

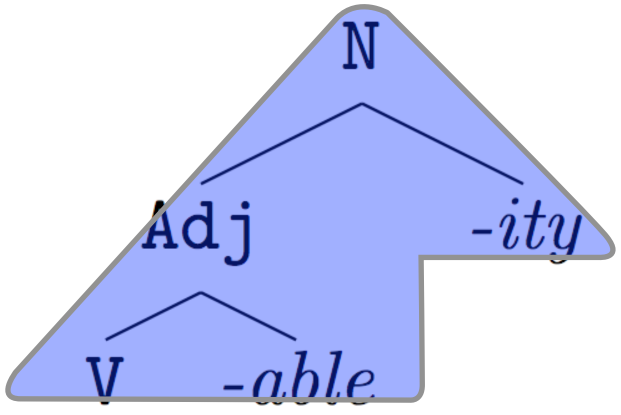
(Fragment Grammars: FG)

O'Donnell et al. (2009); O'Donnell (2011, 2015)



- <u>Storage Strategy</u>: Store set of units that best explains data (only inference-based storage proposal).

- Generalization of *Full-Storage* (AG) model (Johnson et al., 2007).

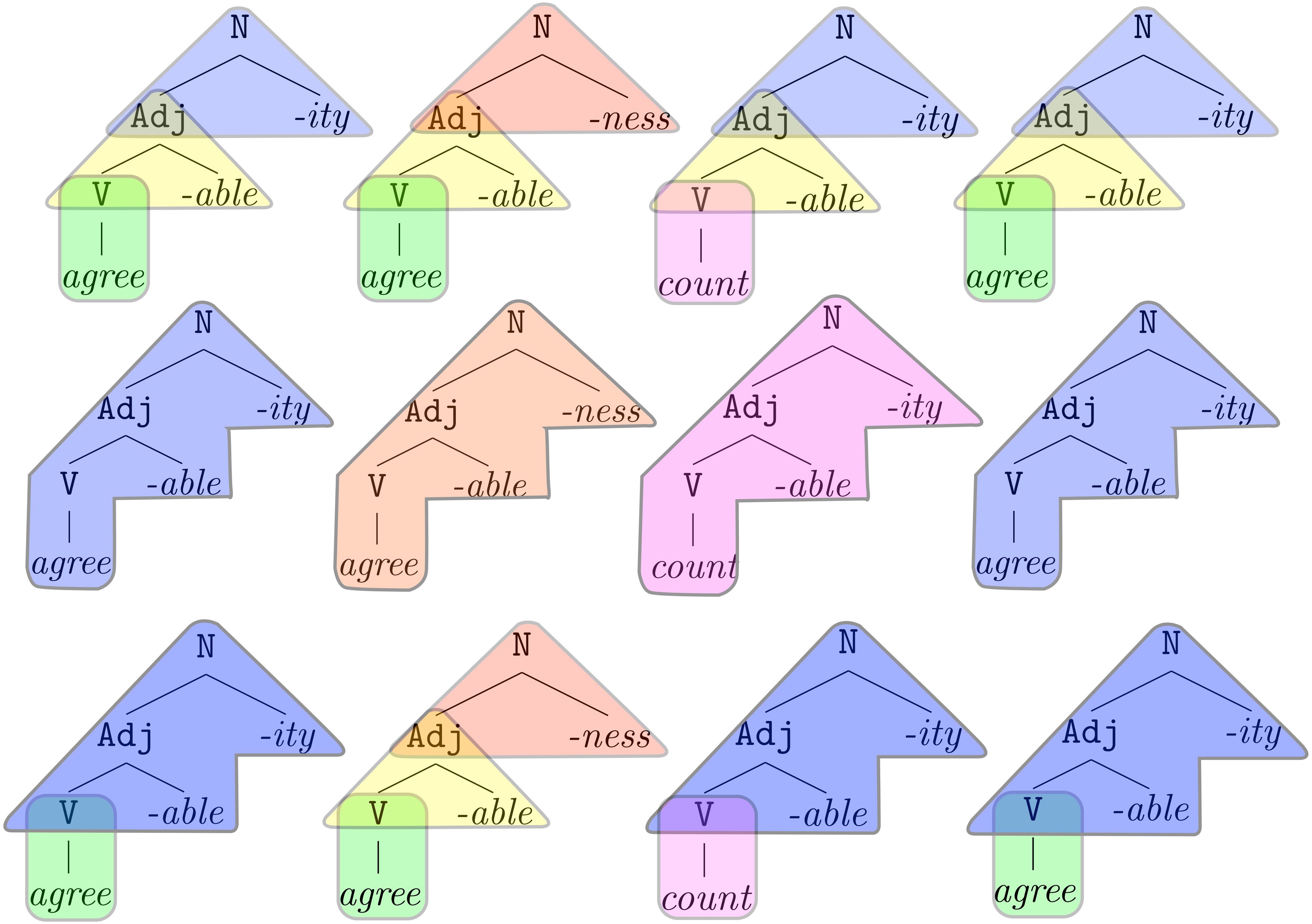


Models

Full-Parsing
(DMPCFG)

Full-Storage
(MAG)

Inference-Based
(FG)



Morphological Productivity

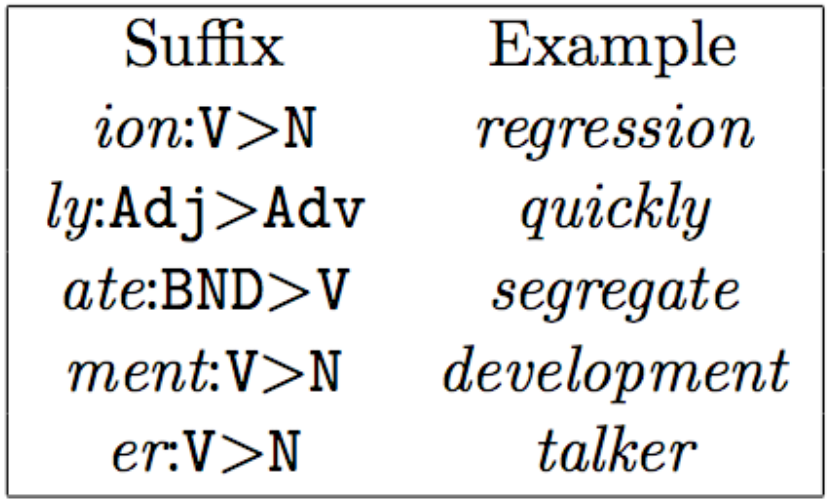
1. **Background**
2. **Productivity and frequency**
3. **Affix ordering**

Productivity and the Distribution of Forms

What aspects of the distribution of words and part-words signal that some structure is an independent, productive lexical unit?

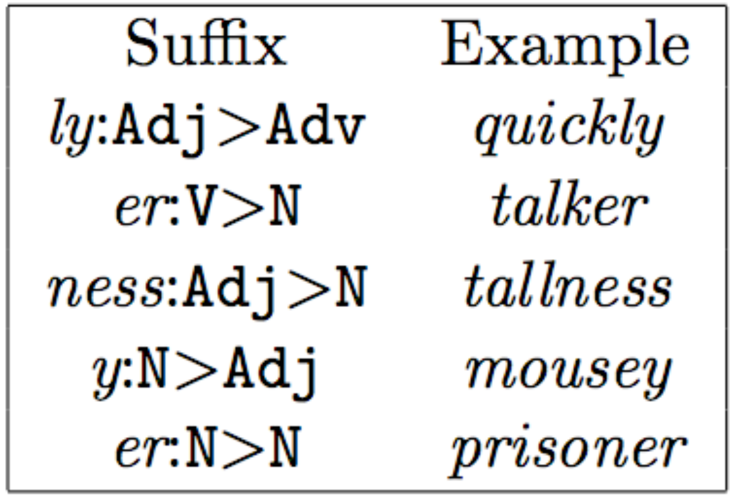
Five Most Productive Suffixes (Predicted)

Full-Parsing (MDPCFG)



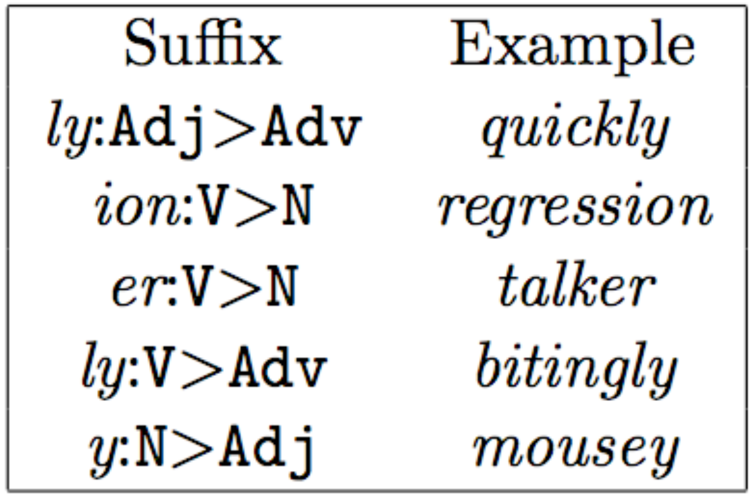
Suffix	Example
ion:V>N	*regression*
ly:Adj>Adv	*quickly*
ate:BND>V	*segregate*
ment:V>N	*development*
er:V>N	*talker*

Inference-Based (FG)



Suffix	Example
ly:Adj>Adv	*quickly*
er:V>N	*talker*
ness:Adj>N	*tallness*
y:N>Adj	*mousey*
er:N>N	*prisoner*

Full-Storage (MAG)

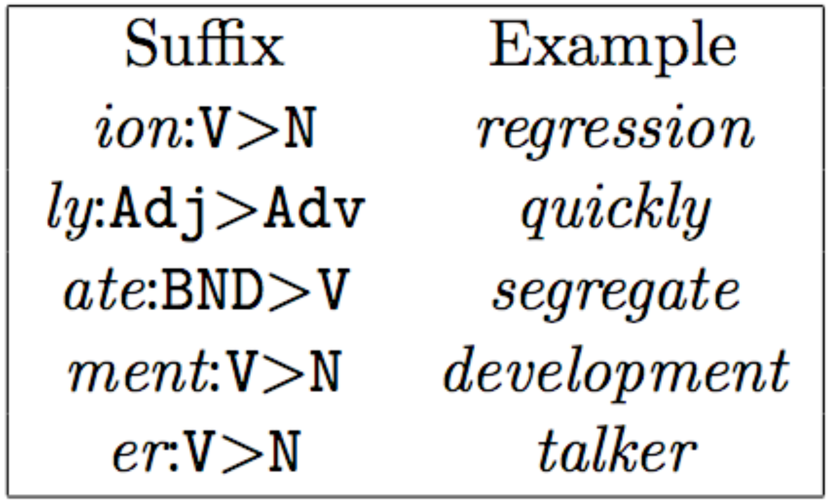


Suffix	Example
ly:Adj>Adv	*quickly*
ion:V>N	*regression*
er:V>N	*talker*
ly:V>Adv	*bitingly*
y:N>Adj	*mousey*

Five Most Productive Suffixes (Predicted)

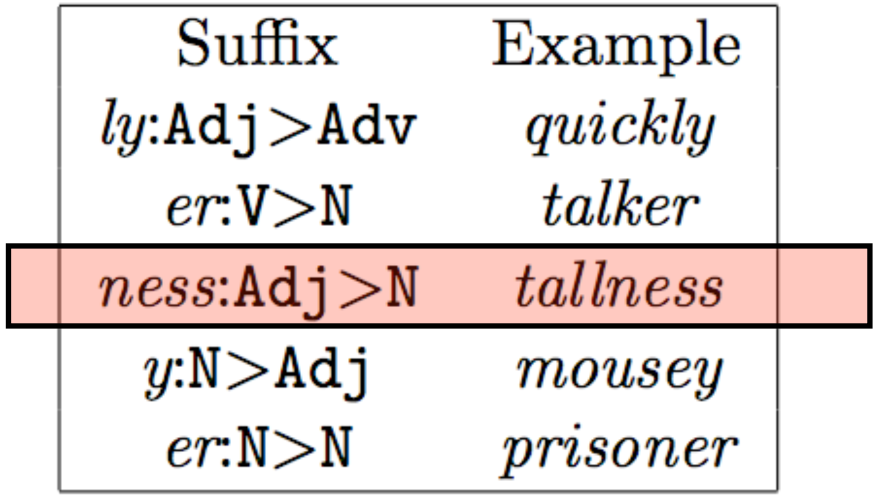
-ness

Full-Parsing (MDPCFG)



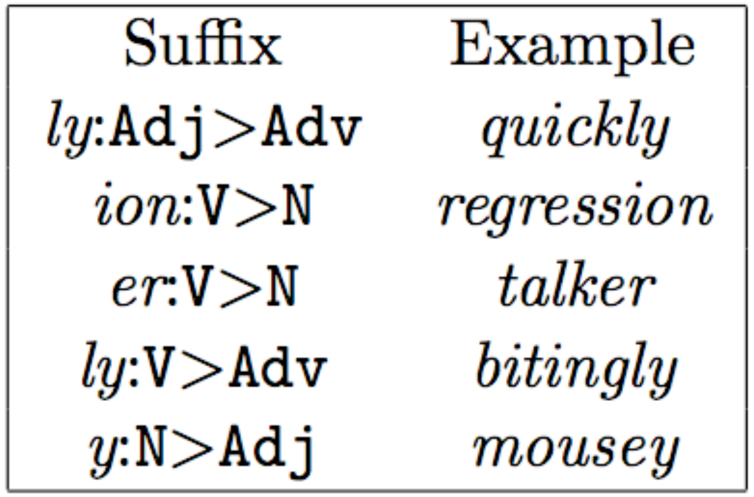
Suffix	Example
ion:V>N	*regression*
ly:Adj>Adv	*quickly*
ate:BND>V	*segregate*
ment:V>N	*development*
er:V>N	*talker*

Inference-Based (FG)



Suffix	Example
ly:Adj>Adv	*quickly*
er:V>N	*talker*
ness:Adj>N	*tallness*
y:N>Adj	*mousey*
er:N>N	*prisoner*

Full-Storage (MAG)

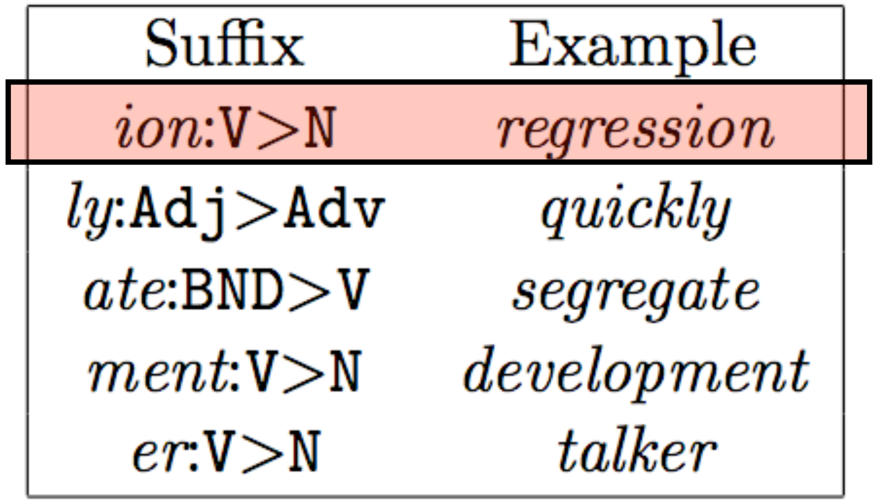


Suffix	Example
ly:Adj>Adv	*quickly*
ion:V>N	*regression*
er:V>N	*talker*
ly:V>Adv	*bitingly*
y:N>Adj	*mousey*

Five Most Productive Suffixes (Predicted)

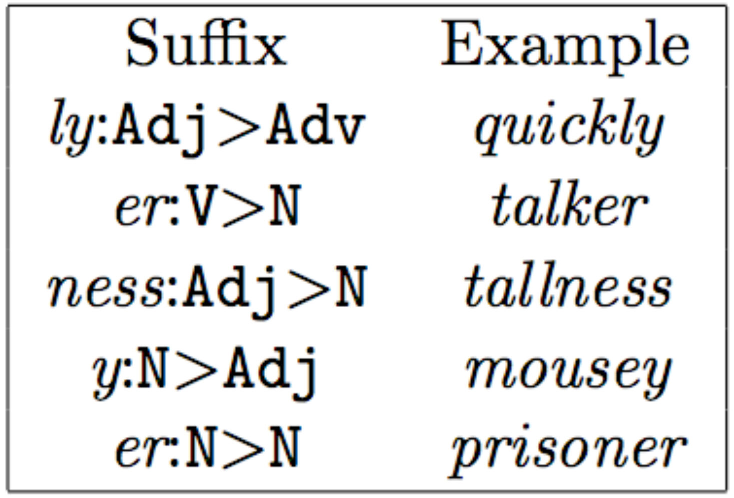
-ion

Full-Parsing (MDPCFG)



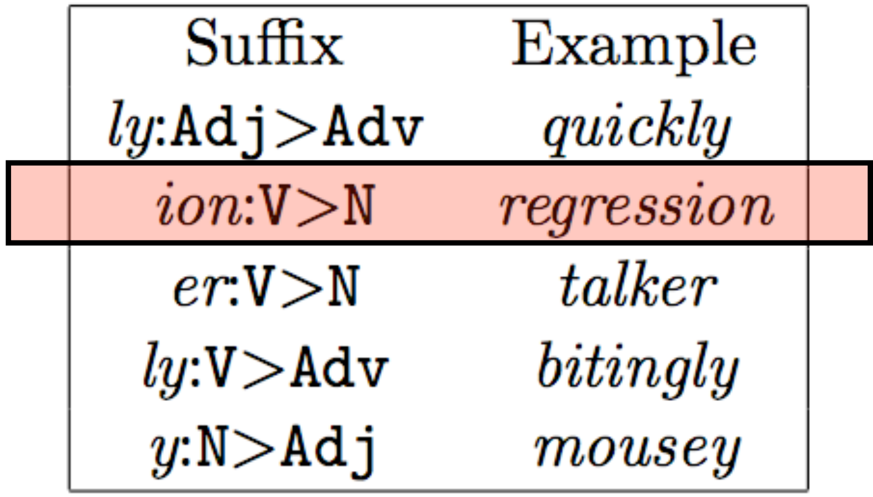
Suffix	Example
ion:V>N	*regression*
ly:Adj>Adv	*quickly*
ate:BND>V	*segregate*
ment:V>N	*development*
er:V>N	*talker*

Inference-Based (FG)



Suffix	Example
ly:Adj>Adv	*quickly*
er:V>N	*talker*
ness:Adj>N	*tallness*
y:N>Adj	*mousey*
er:N>N	*prisoner*

Full-Storage (MAG)



Suffix	Example
ly:Adj>Adv	*quickly*
ion:V>N	*regression*
er:V>N	*talker*
ly:V>Adv	*bitingly*
y:N>Adj	*mousey*

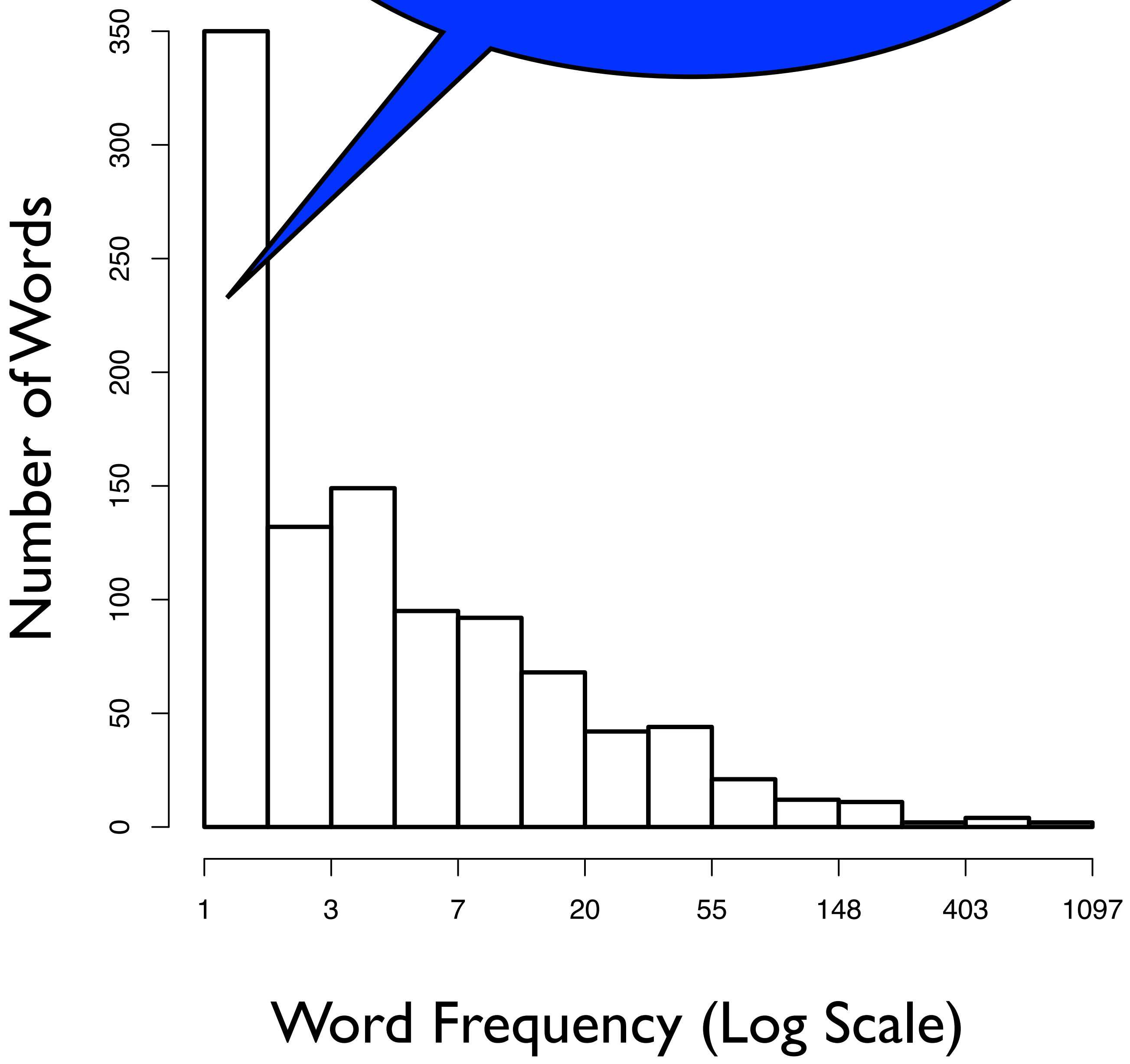
*eat : *eation*

Productivity and Frequency Distributions

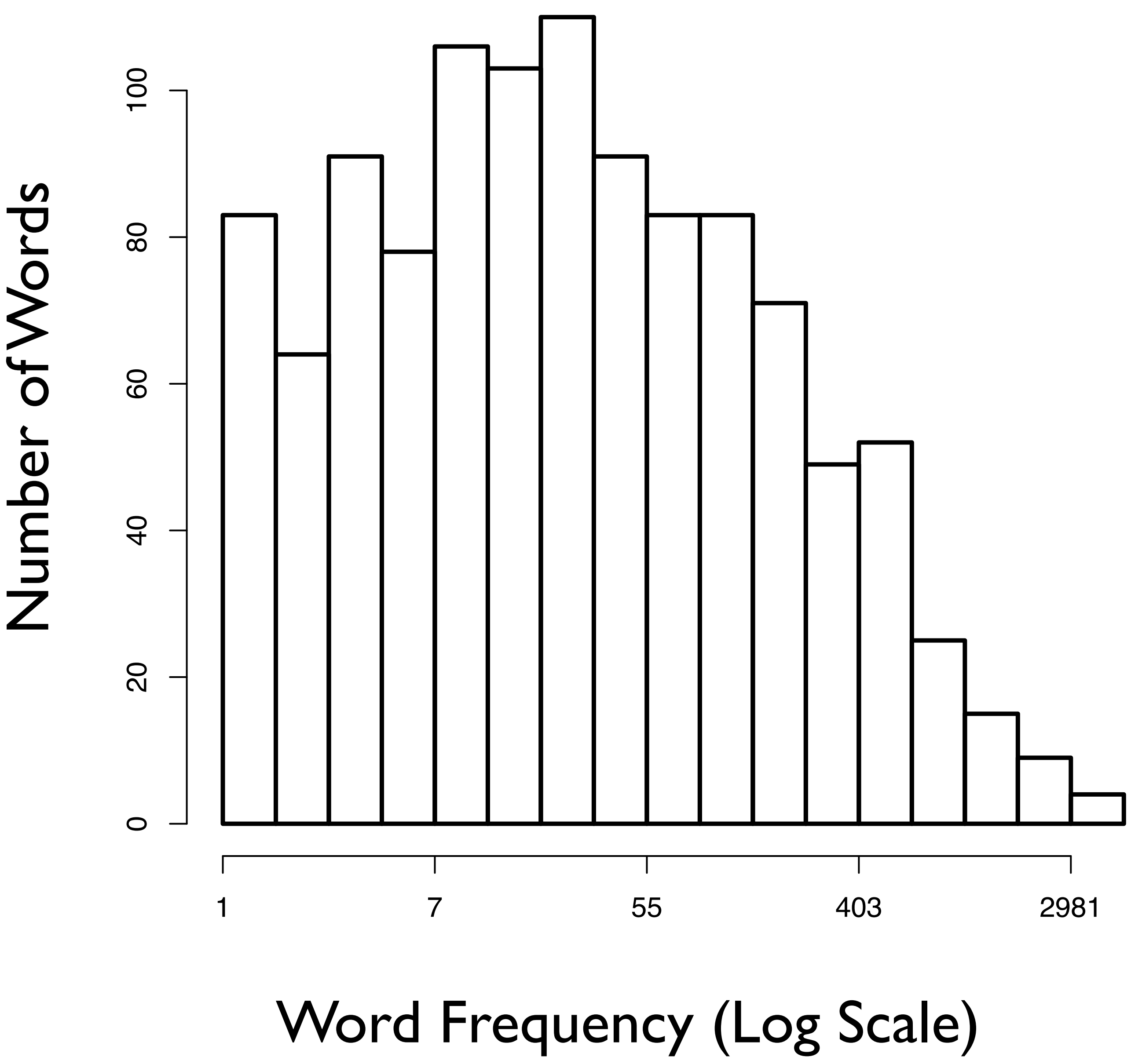
- Productive suffixes give rise to new forms.
- <u>Distributional consequence</u>: Large proportion of low-frequency forms.
- Large number of rare event distributions (LNRE) investigated mathematically (Khmaladze, 1987; Baayen, 2001).
- Used to develop useful statistical tools to quantify and study productivity in corpus samples (e.g., Baayen, 1992).
    - More below.

Pr d The Proportion of
ency Forms

High Proportion of Low-Frequency Forms

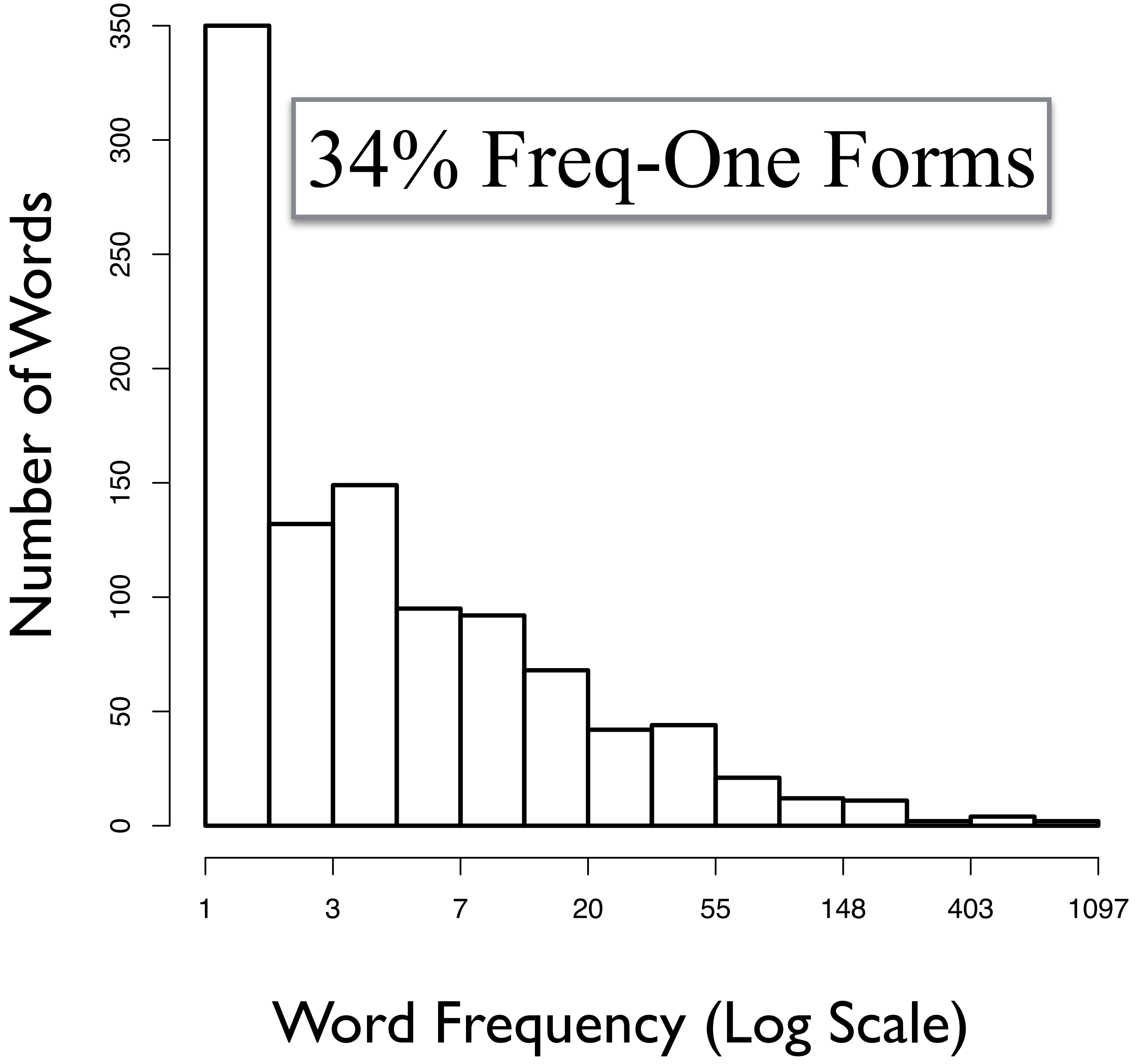


-ion

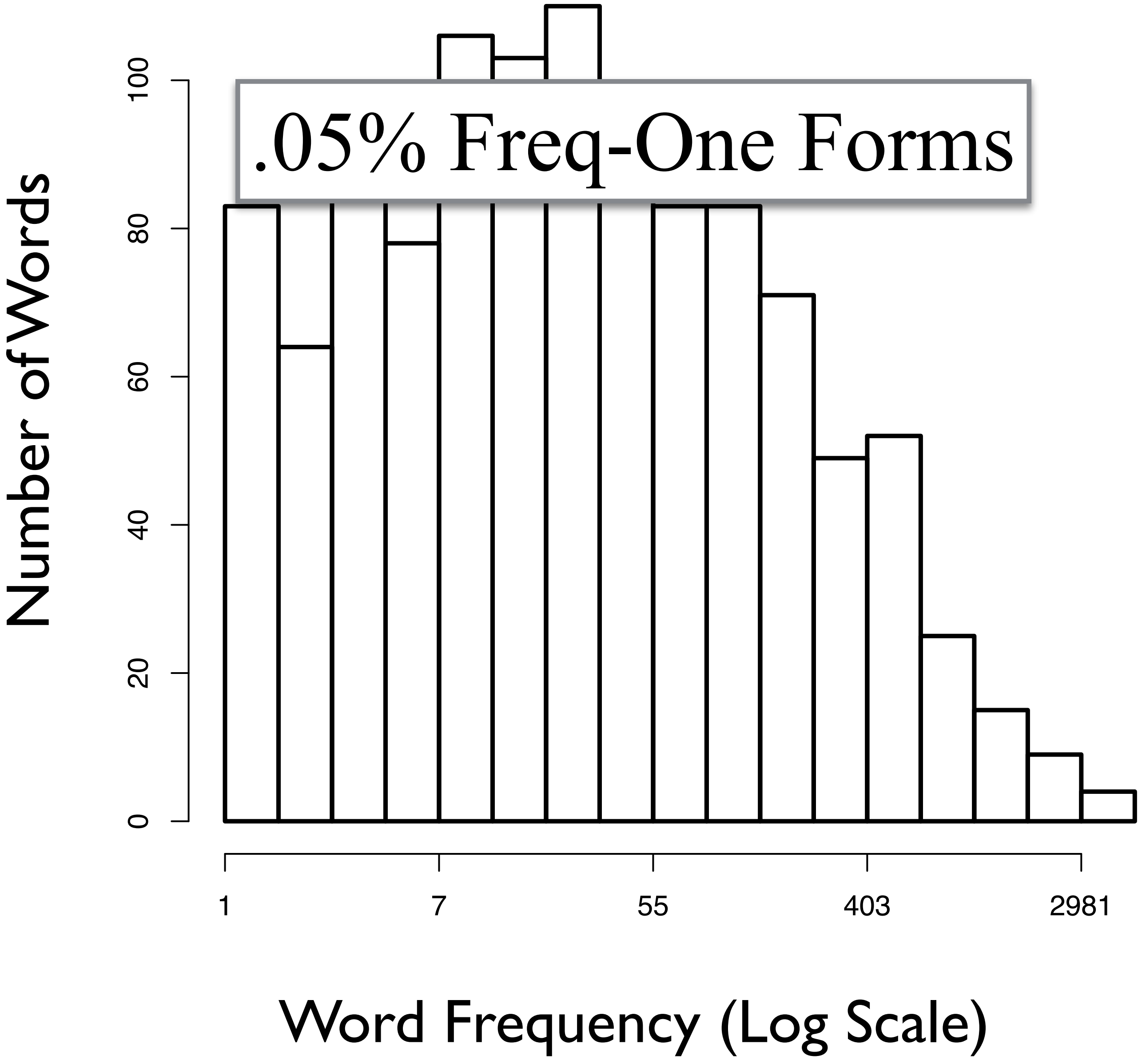


Productivity and The Proportion of Low-Frequency Forms

-ness

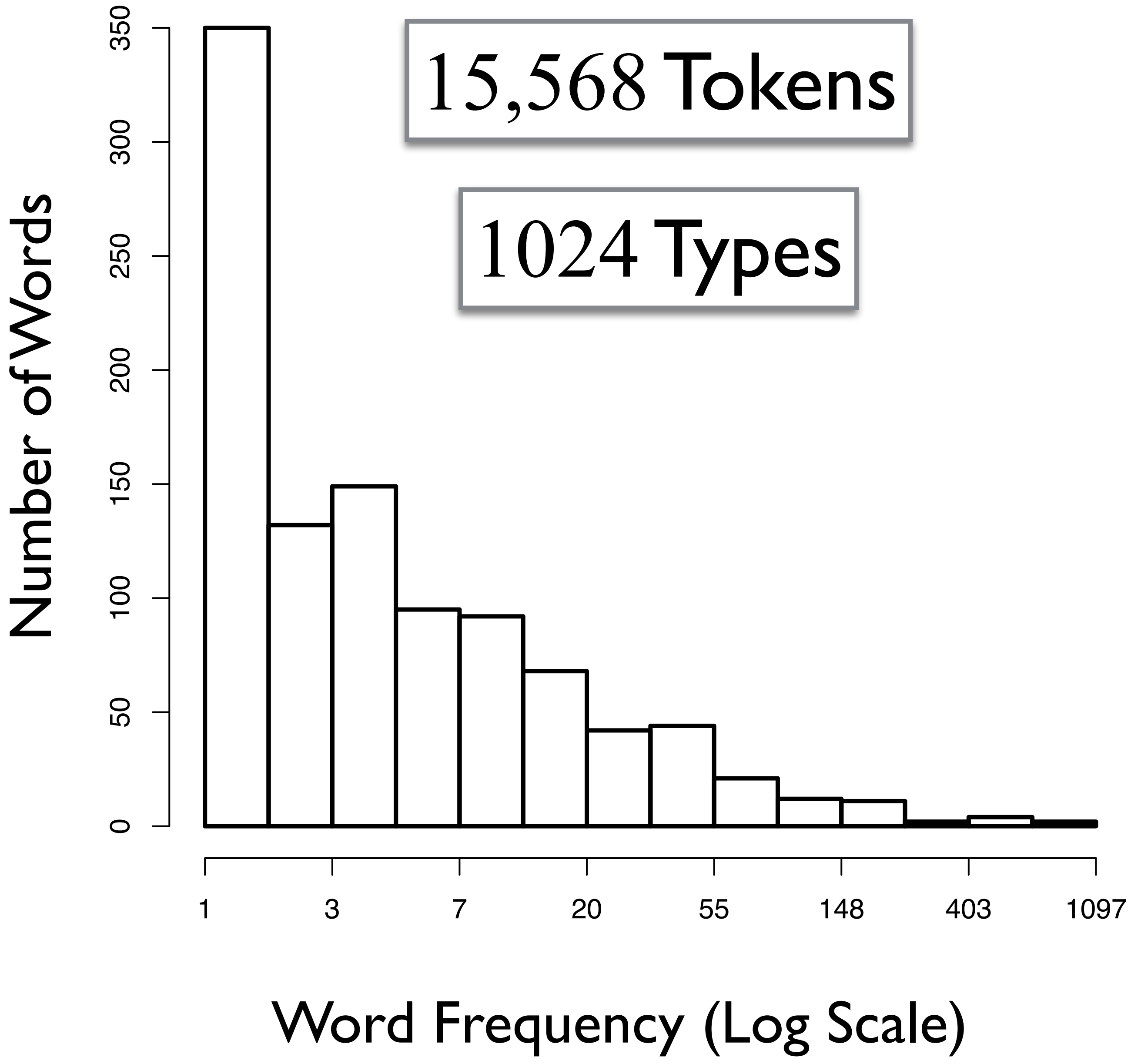


-ion



Type and Token Frequencies

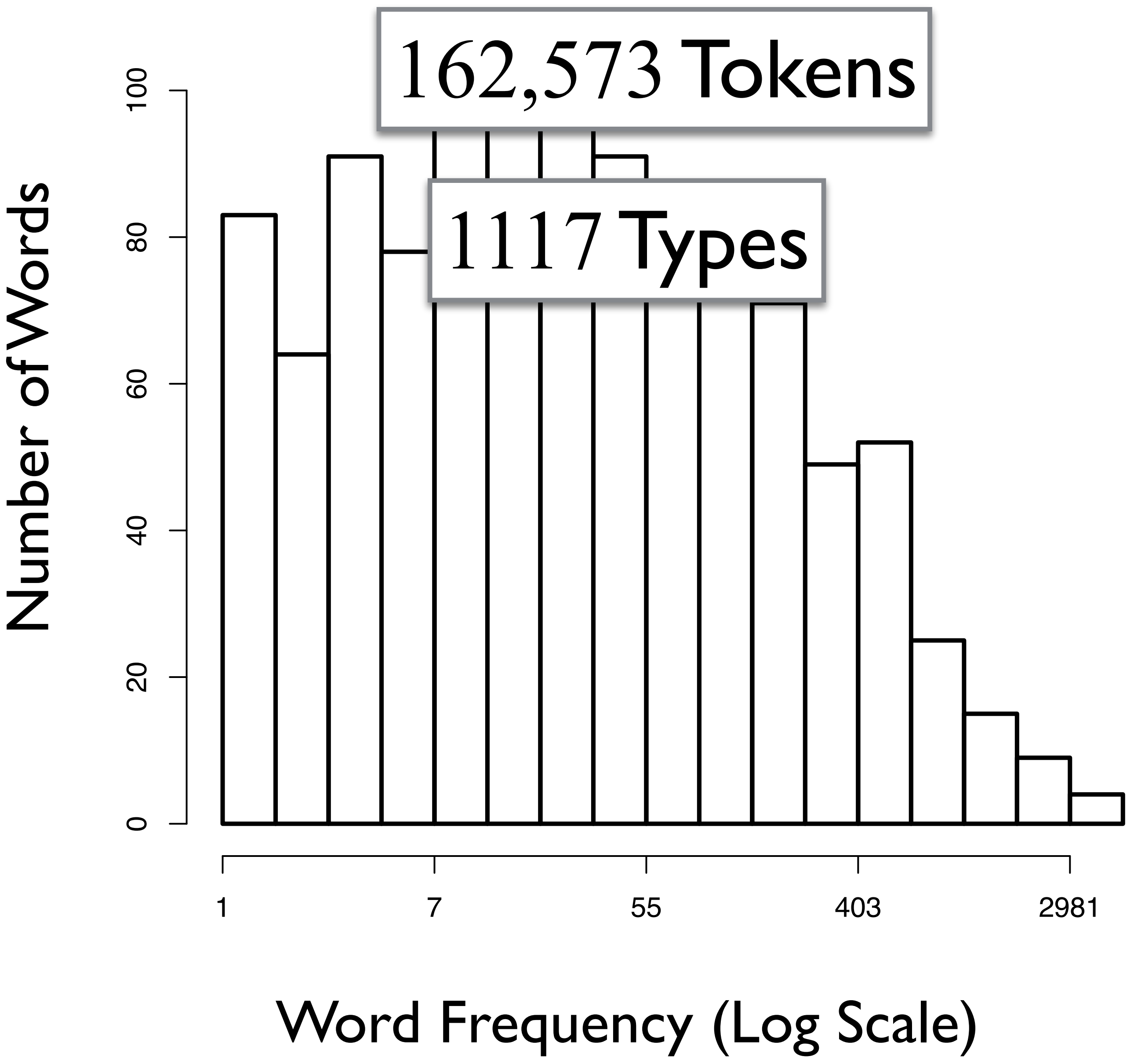
-ness



15,568 Tokens

1024 Types

-ion



162,573 Tokens

1117 Types

Five Most Productive Suffixes (Predicted)

-ne

High Proportion of Low-Frequency Forms

Full-Parsing (MDPCFG)

Suffix	Example
ion:V>N	*regression*
ly:Adj>Adv	*quickly*
ate:BND>V	*segregate*
ment:V>N	*development*
er:V>N	*talker*

Inference-Bas

Suffix	Exa ple
ly:Adj>Adv	*q ckly*
er:V>N	*talker*
ness:Adj>N	*tallness*
y:N>Adj	*mousey*
er:N>N	*prisoner*

ge (MAG)

Suffix	Example
ly:Adj>Adv	*quickly*
ion:V>N	*regression*
er:V>N	*talker*
ly:V>Adv	*bitingly*
y:N>Adj	*mousey*

Five Most Productive Suffixes
(Predict

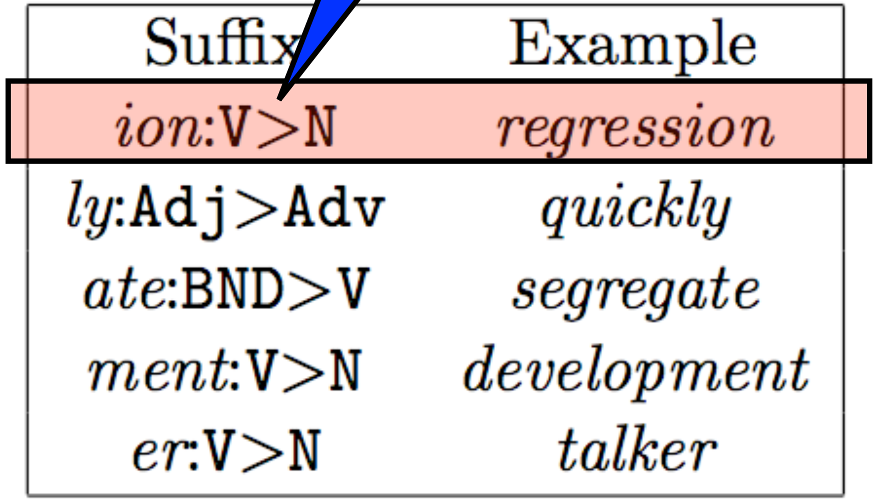
Frequently Proposed

-ion

High Token Frequency

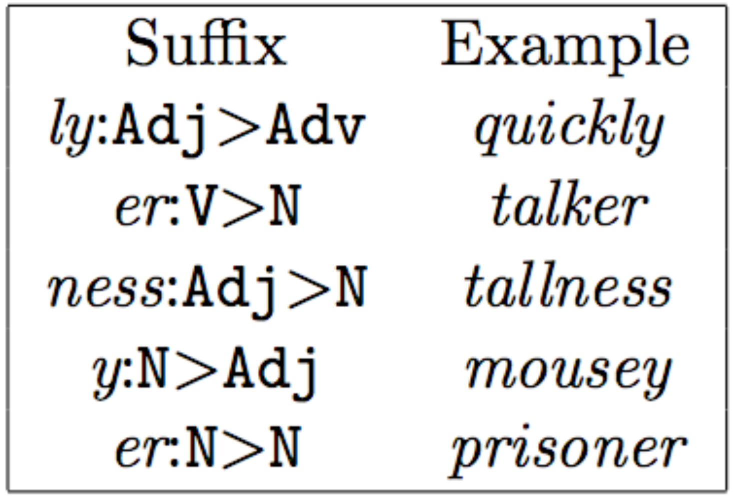
High Type Frequency

Full-Fu FG)



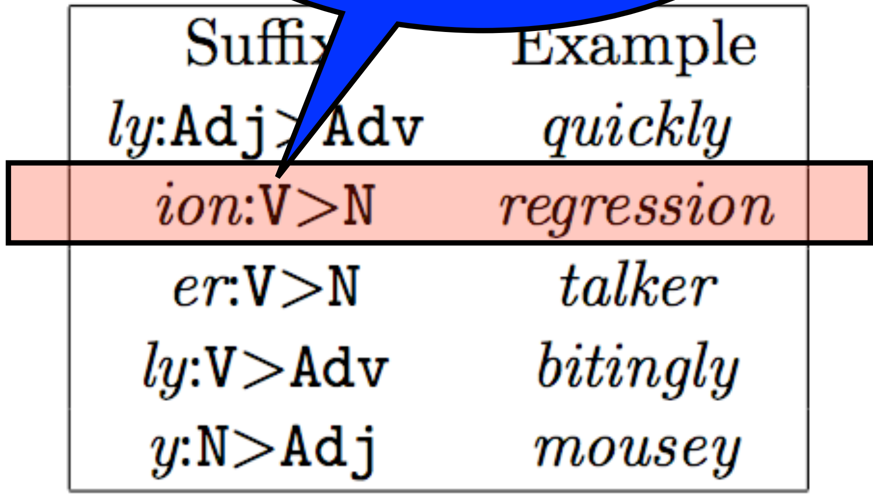
Suffix	Example
ion:V>N	*regression*
ly:Adj>Adv	*quickly*
ate:BND>V	*segregate*
ment:V>N	*development*
er:V>N	*talker*

Inference-Based (FG)



Suffix	Example
ly:Adj>Adv	*quickly*
er:V>N	*talker*
ness:Adj>N	*tallness*
y:N>Adj	*mousey*
er:N>N	*prisoner*

F



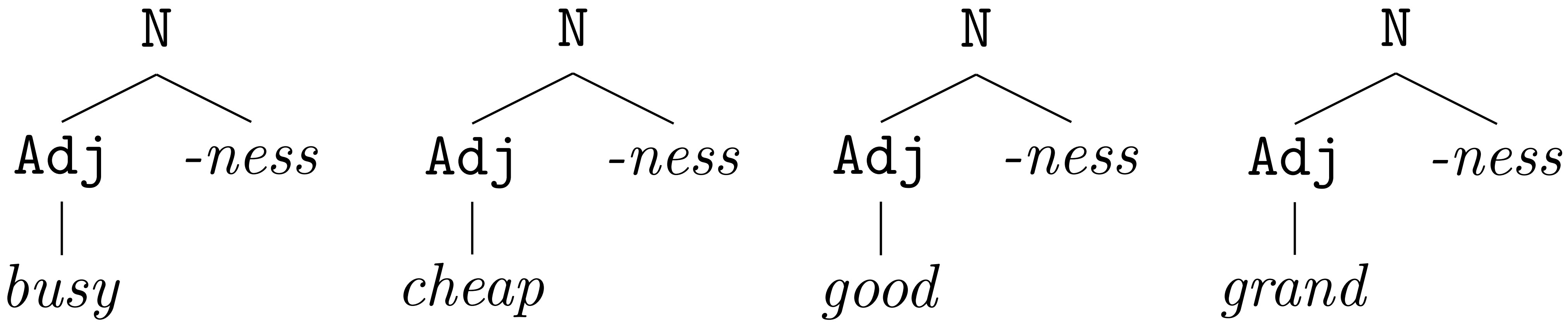
Suffix	Example
ly:Adj>Adv	*quickly*
ion:V>N	*regression*
er:V>N	*talker*
ly:V>Adv	*bitingly*
y:N>Adj	*mousey*

Productivity and The Proportion of Low-Frequency Forms

- The *Inference-Based* (`FG`) makes correct use of distributional facts about productivity.

- Why is the model sensitive to the proportion of low-frequency forms?

Ideal Input for Productivity



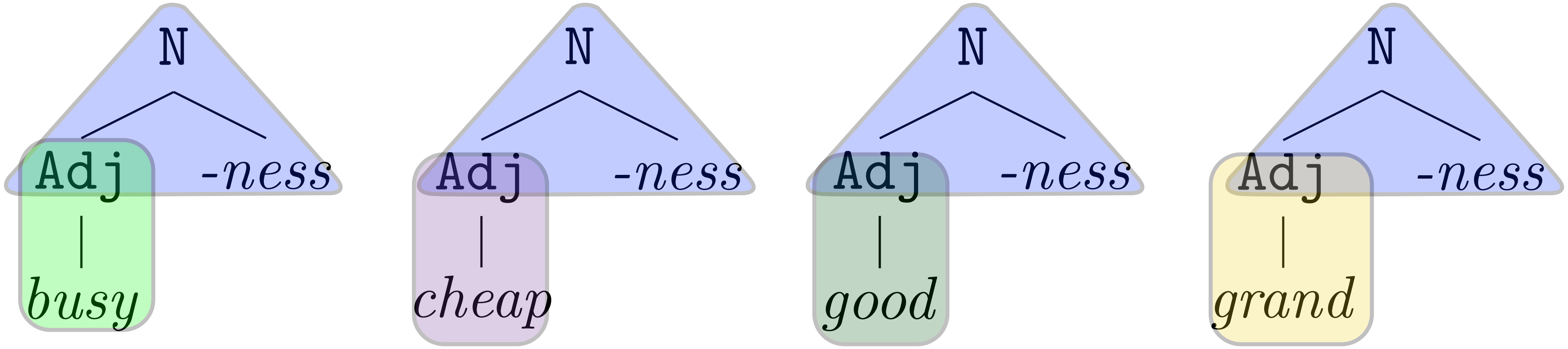
N
Adj -*ness*
busy

N
Adj -*ness*
cheap

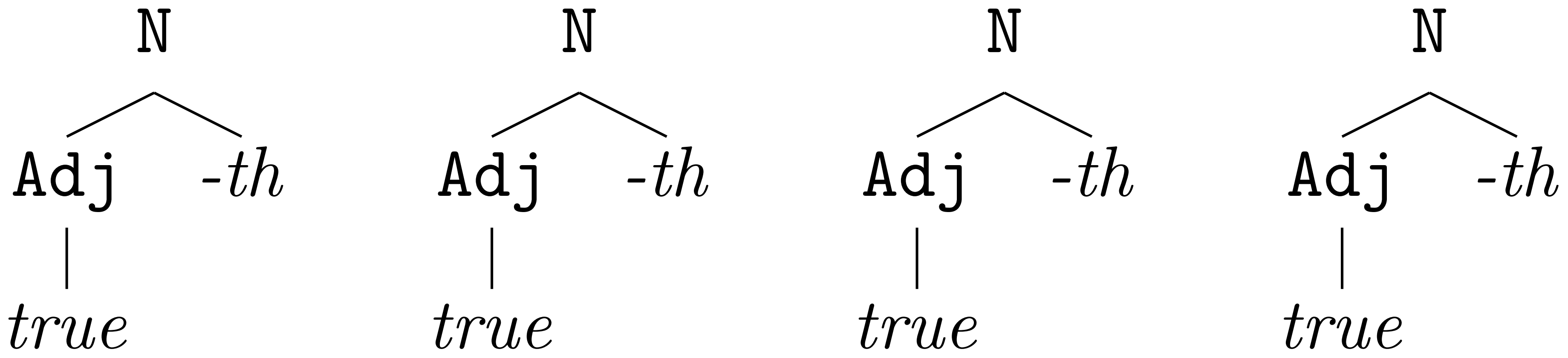
N
Adj -*ness*
good

N
Adj -*ness*
grand

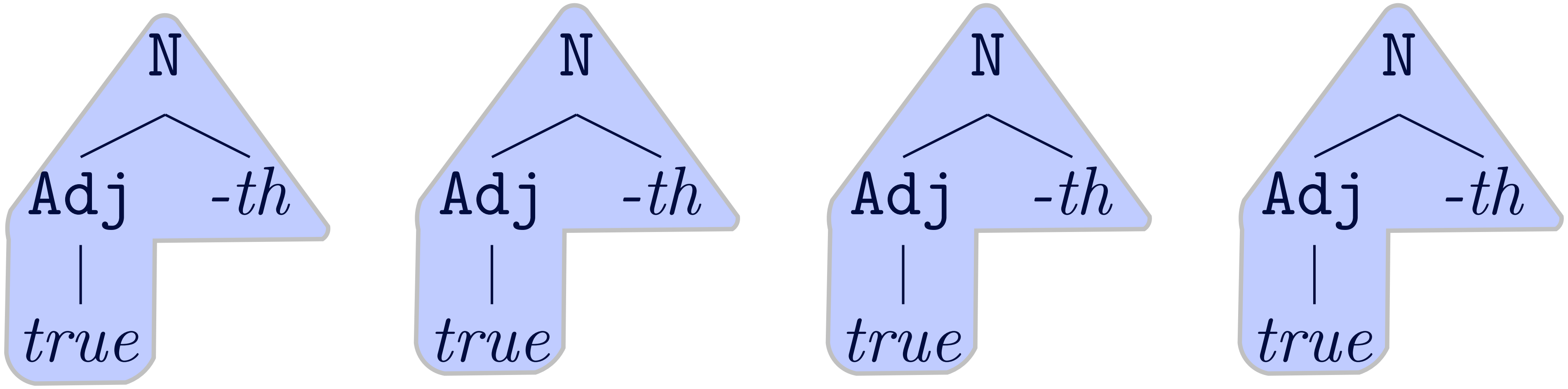
Ideal Input for Productivity



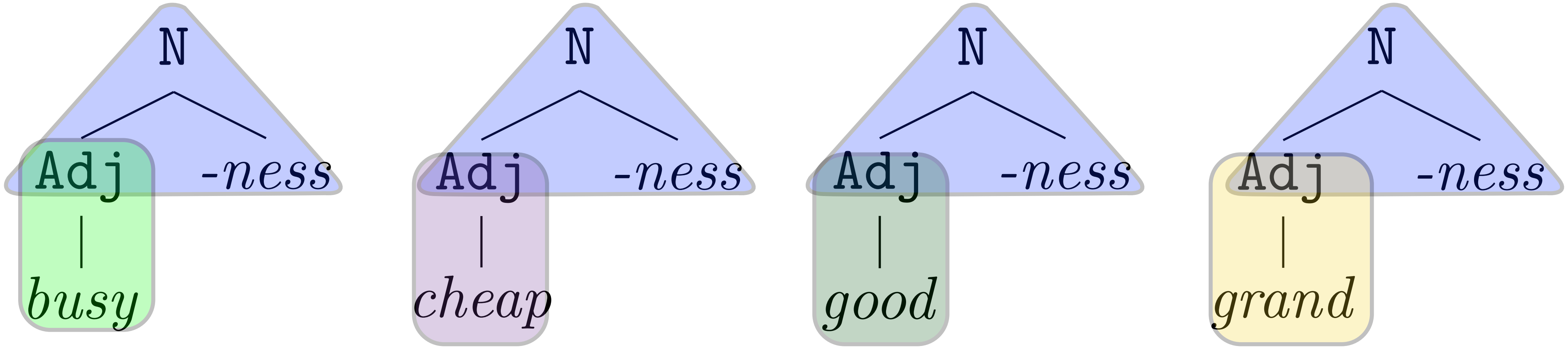
Ideal Input for Storage



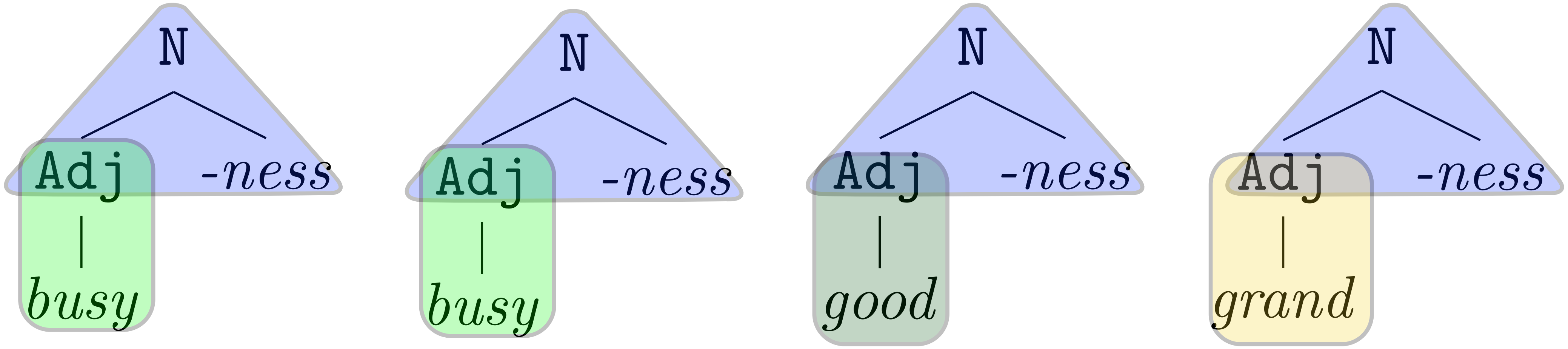
Ideal Input for Storage



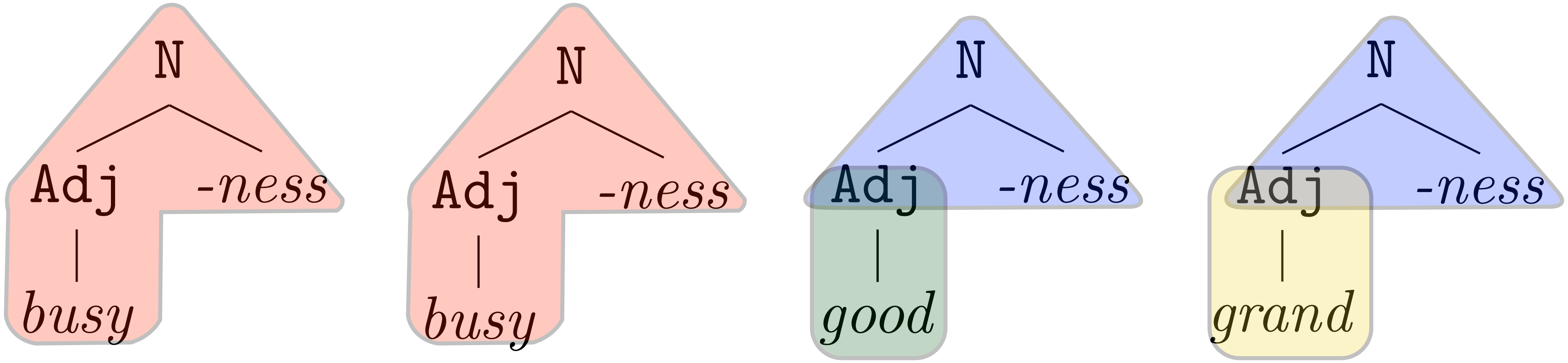
Productivity and the Proportion of Low-Frequency Forms



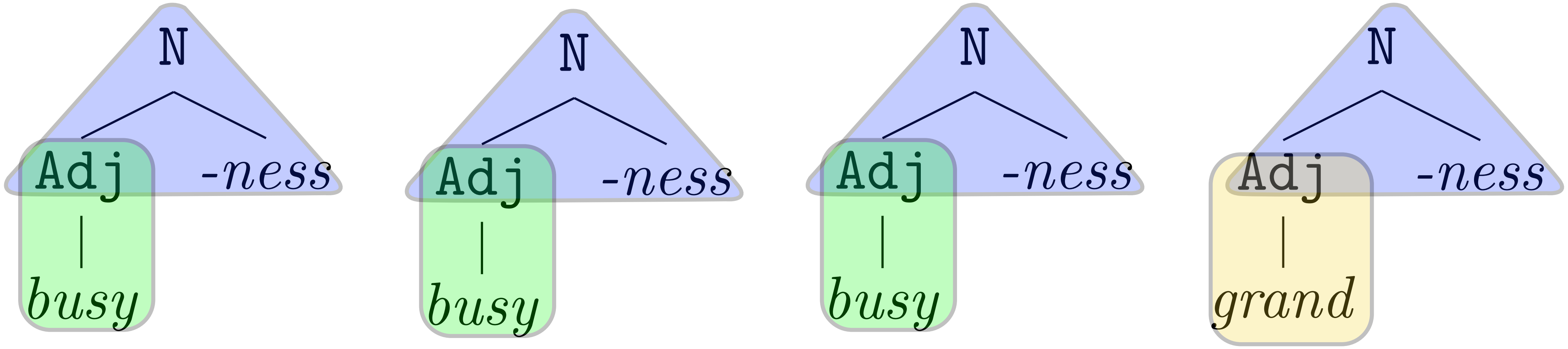
Productivity and the Proportion of Low-Frequency Forms



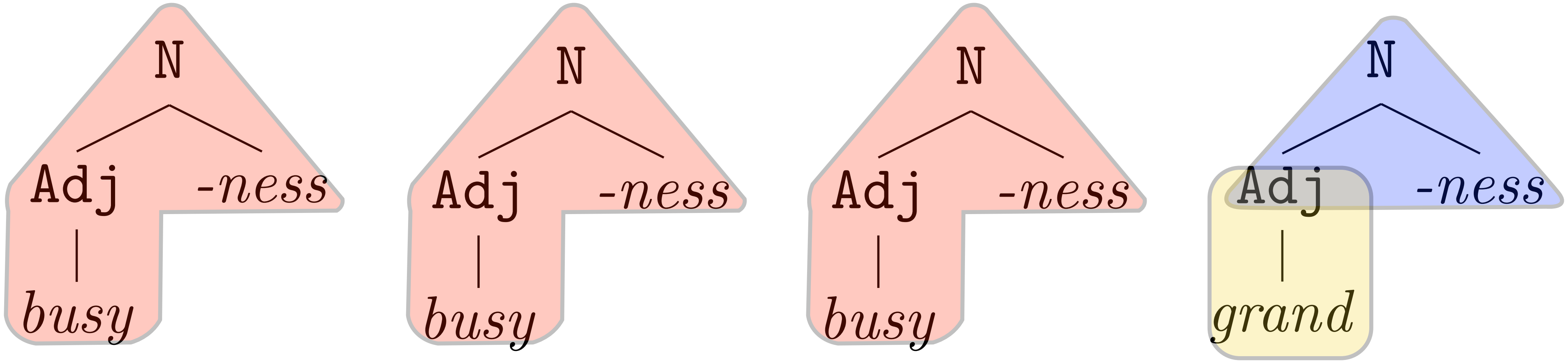
Productivity and the Proportion of Low-Frequency Forms



Productivity and the Proportion of Low-Frequency Forms



Productivity and the Proportion of Low-Frequency Forms

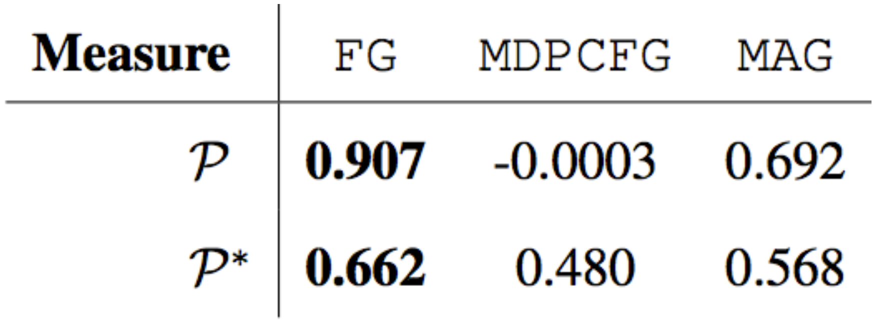


Baayen’s Hapax-Based Estimators

- Baayen’s $\mathcal{P}$/ $\mathcal{P}^*$ (e.g., Baayen, 1992)
- Most well-studied statistical corpus-oriented estimators of productivity.
- Estimators of *single-affix* productivity based on the proportion hapaxes.
- Various mathematical derivations.
    - Rate of vocabulary change.
    - Good-Turing.
- Non-psychological.

The *Inference-Based* Model and Baayen's Measures

- For individual suffixes predictions strongly correlate.



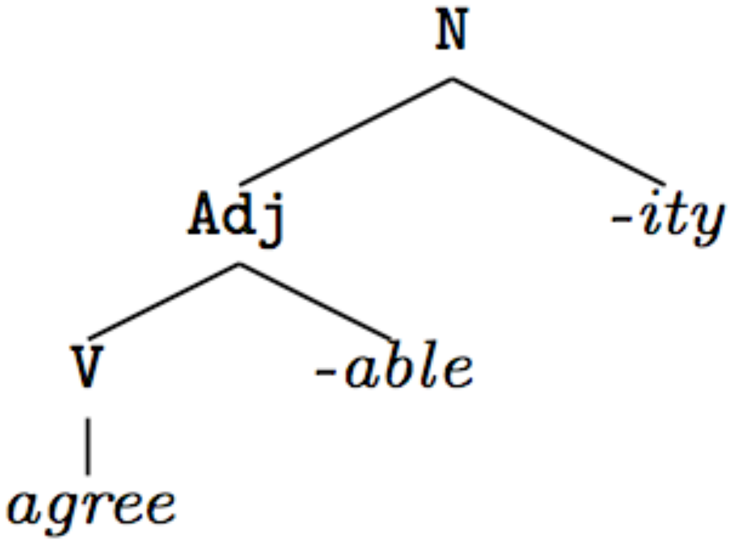
Measure	FG	MDPCFG	MAG
$\mathcal{P}$	**0.907**	-0.0003	0.692
$\mathcal{P}^*$	**0.662**	0.480	0.568

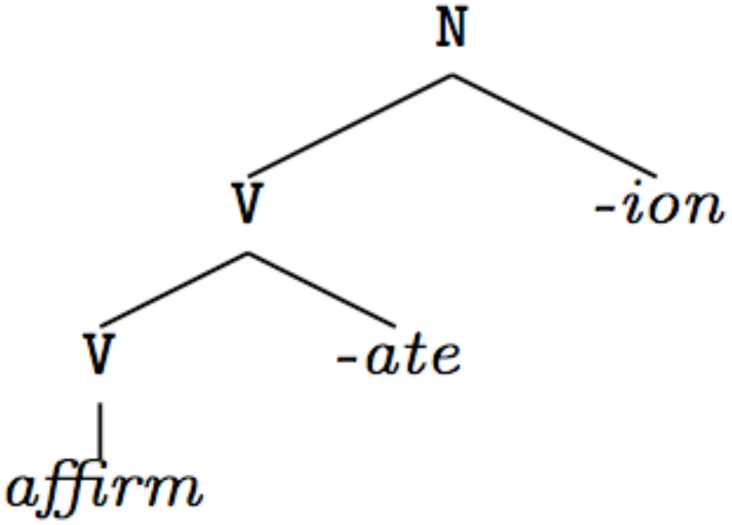
- *Inference-Based* (FG) model makes no explicit assumption about-low frequency forms.
- *Inference-Based* (FG) model derives relationship between productivity inferences and distributional facts from storage-computation tradeoff applied to the problem of lexicon learning.

Morphological Productivity

1. **Background**
2. **Productivity and frequency**
3. **Affix ordering**

The Affix-Ordering Problem





- Tiny fraction of logically possible affix combinations occur in practice.

- Fabb (1988)
  - 43 suffixes
  - 663 predicted possible.
  - ~50 attested.

The Affix-Ordering Problem

- Numerous accounts (see, O'Donnell, 2015 for review).
    - None entirely successful.
- <u>Main empirical generalization</u>: Ordering correlates with a number of other properties.

Correlated Properties

Earlier Suffixes

-ion, -ity, -y, -al, -ic,
-ate, -ous, -ive, …

Later Suffixes

-ness, -less, -hood, -ful,
-ly, -y, -like, -ist, …

1. Origin (Latin/Germanic)
2. Phonological regularity
3. Transparency of meaning
4. Productivity

Productivity and Ordering Generalization

On average more productive affixes appear outside of less productive affixes

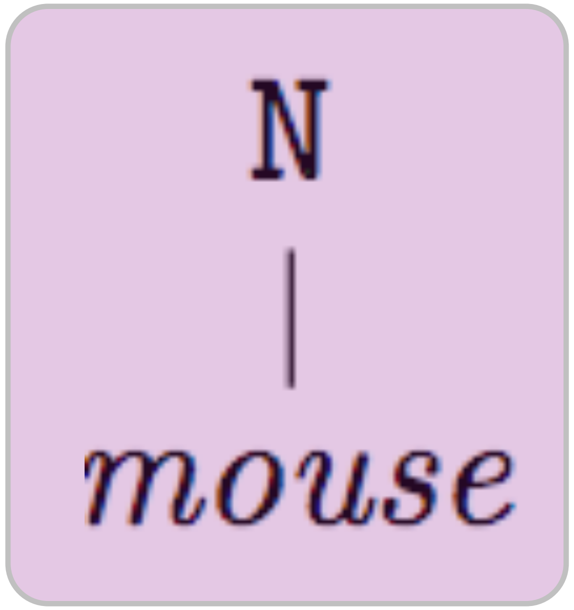
(Hay, 2002; Hay & Plag, 2004; Plag et al., 2009).

Productivity and Ordering Generalization

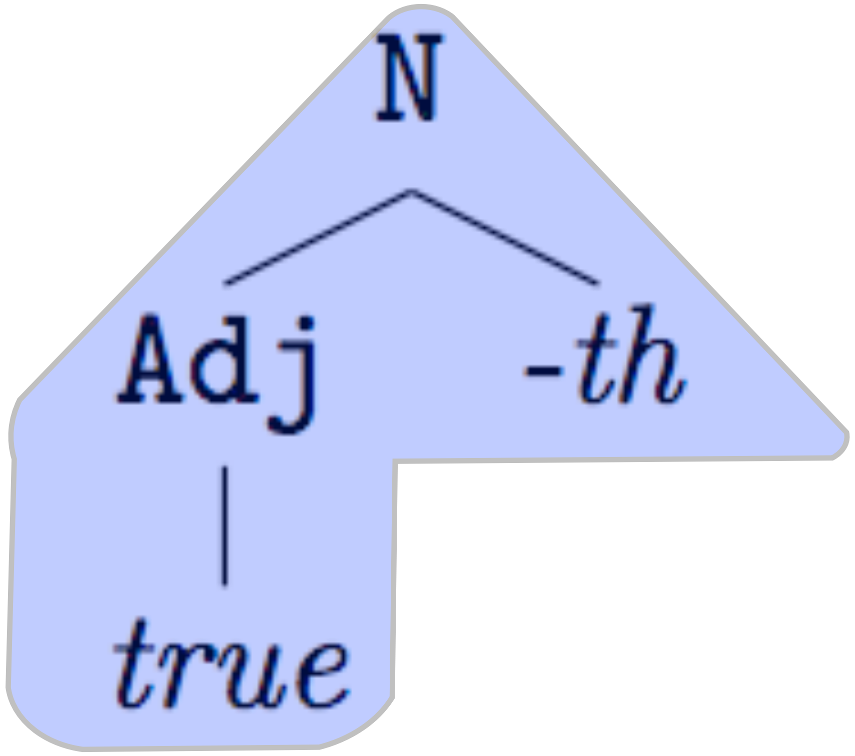
- Follows as a consequence of *Inference-Based* (FG) model's pattern of storage and computation.
- **Crucial point**: By definition, productive affixes are just independent units which can combine freely; unproductive affixes are subparts of other stored forms.

Productivity and Ordering

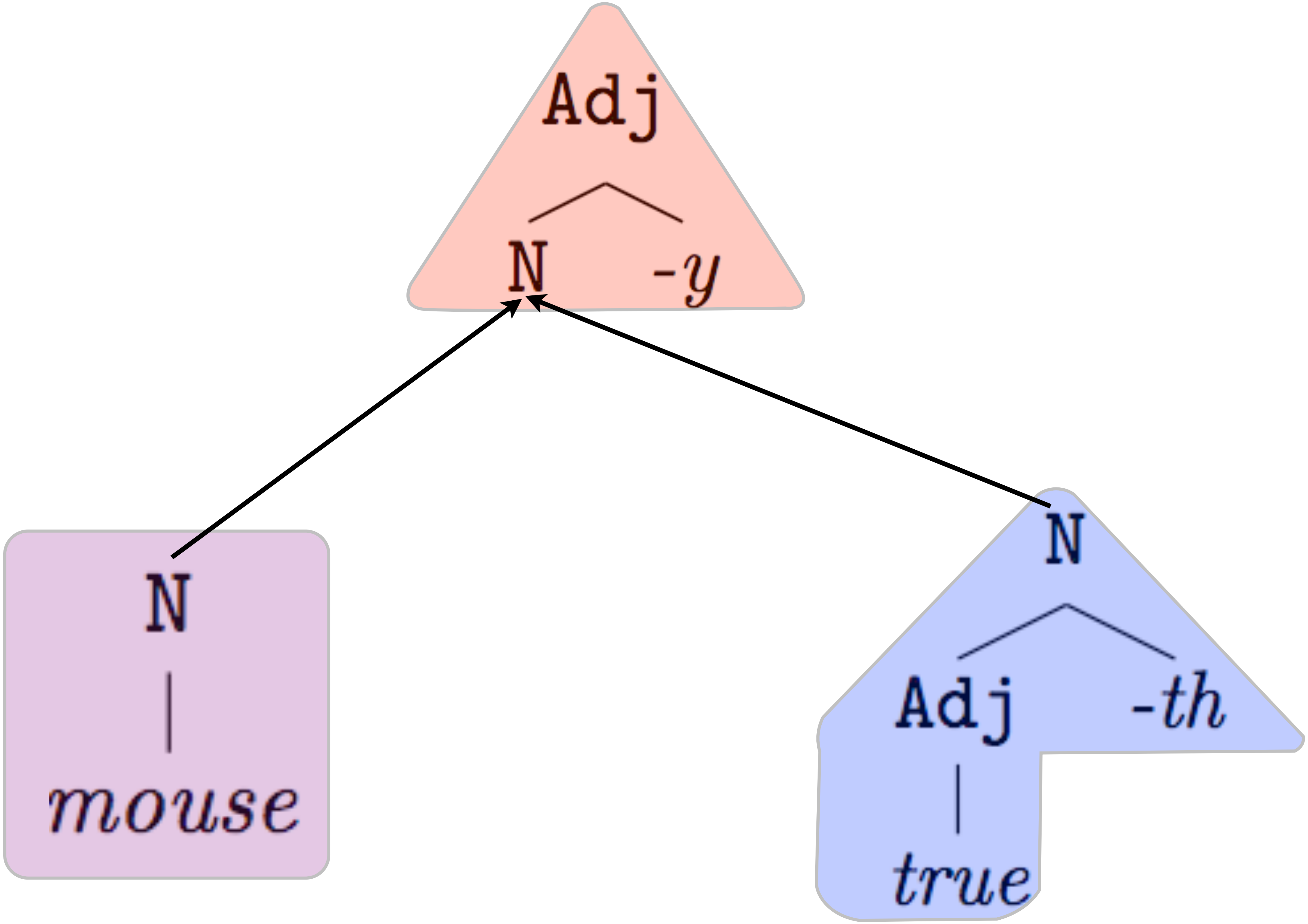
Simple words



Words with unproductive suffixes



Productivity and Ordering



Correlated Properties

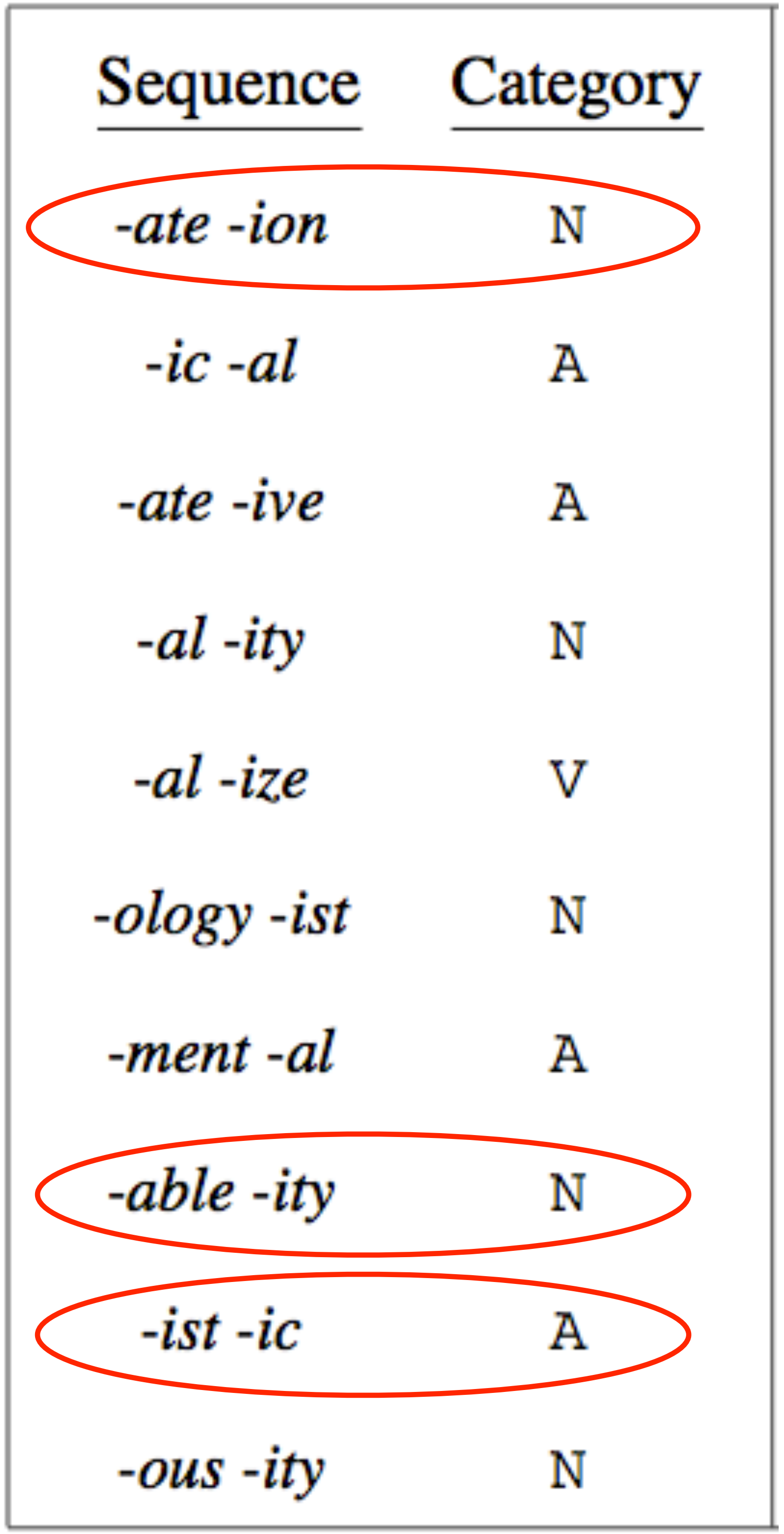
- Latinate v. Germanic reflects diachronic fact that Latinate affixes were borrowed as parts of loan words.

- Regularity and Transparency differences are accounted for if stored items can preserve or accrue idiosyncrasies.

Paradoxical Suffix Combinations

- Exceptions to the productivity and ordering generalization and all current theoretical accounts (Hay, 2003).
  - *-ability*, *-ation*, *-istic*
- **Idea**: *-ability*, *-ation*, and *-istic* are single, stored lexical units.

Predicted Generalizable Units



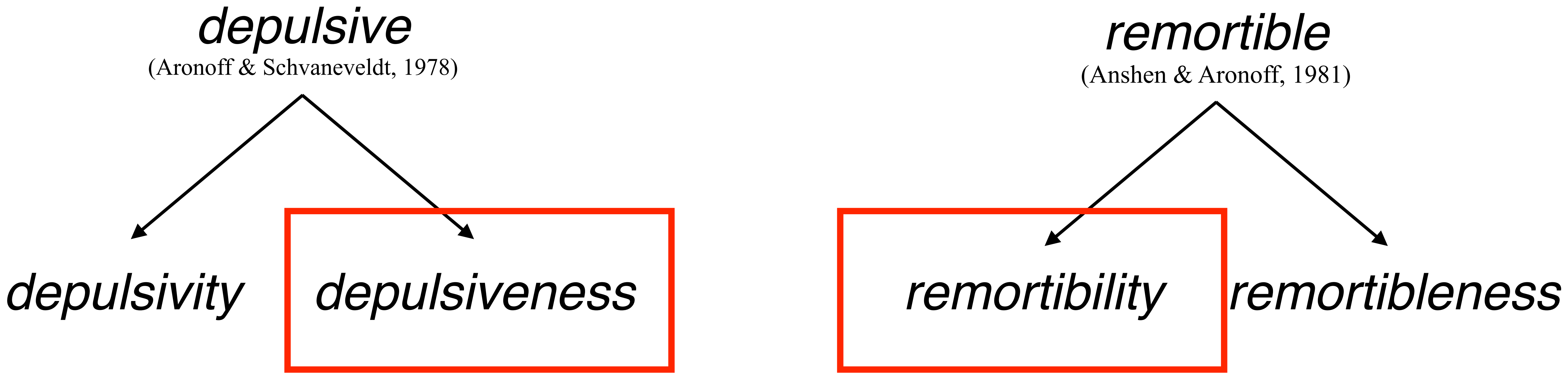
Sequence	Category
-ate -ion	N
-ic -al	A
-ate -ive	A
-al -ity	N
-al -ize	V
-ology -ist	N
-ment -al	A
-able -ity	N
-ist -ic	A
-ous -ity	N

Experimental Evidence

(Aronoff & Schvaneveldt,1978; Anshen and Aronoff, 1981)

Suffix I | Suffix 2

novel stem + *-ive v. -(a)ble* + *-ity v. -ness*



depulsive

(Aronoff & Schvaneveldt, 1978)

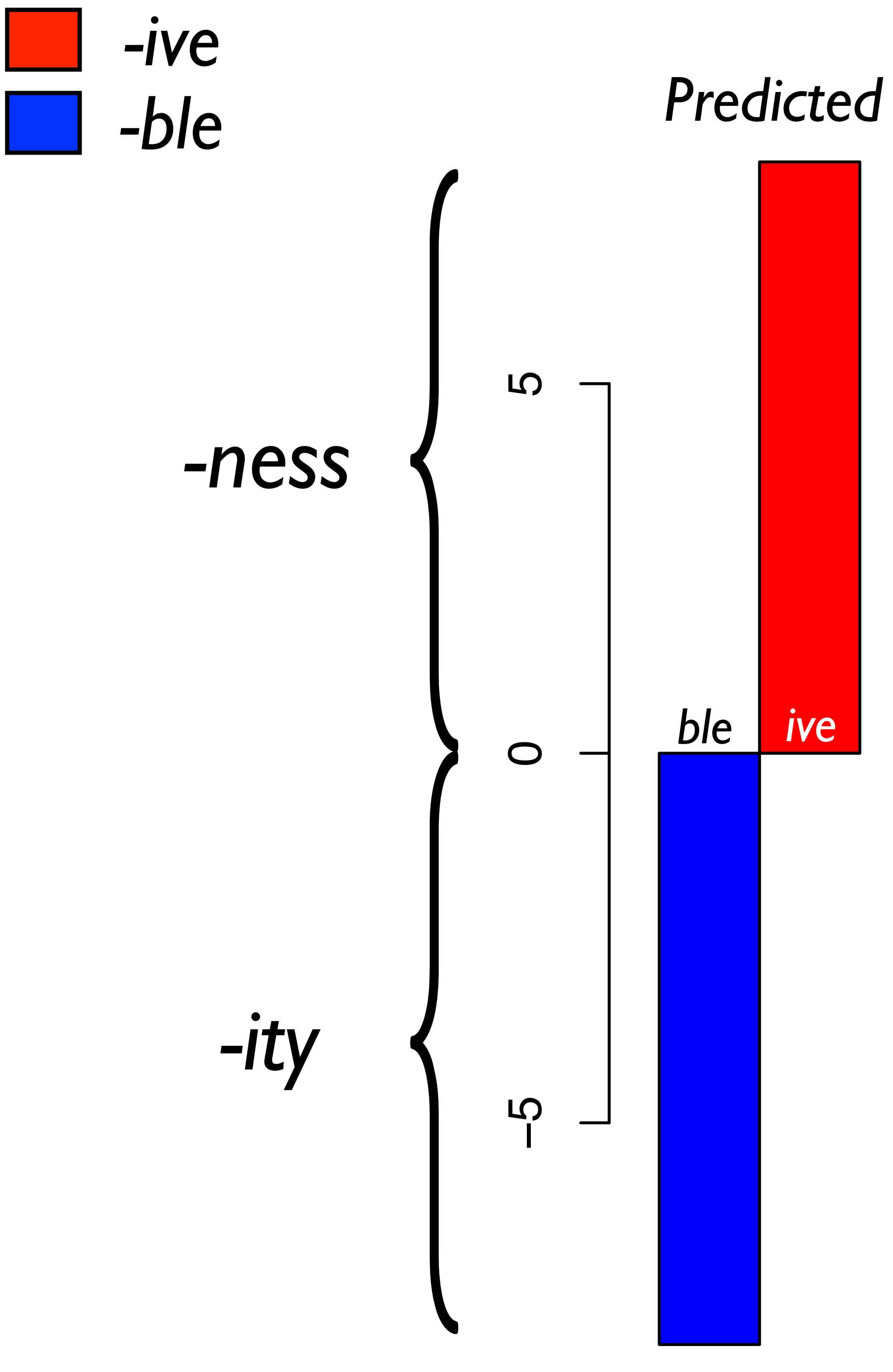
depulsivity *depulsiveness*

remortible

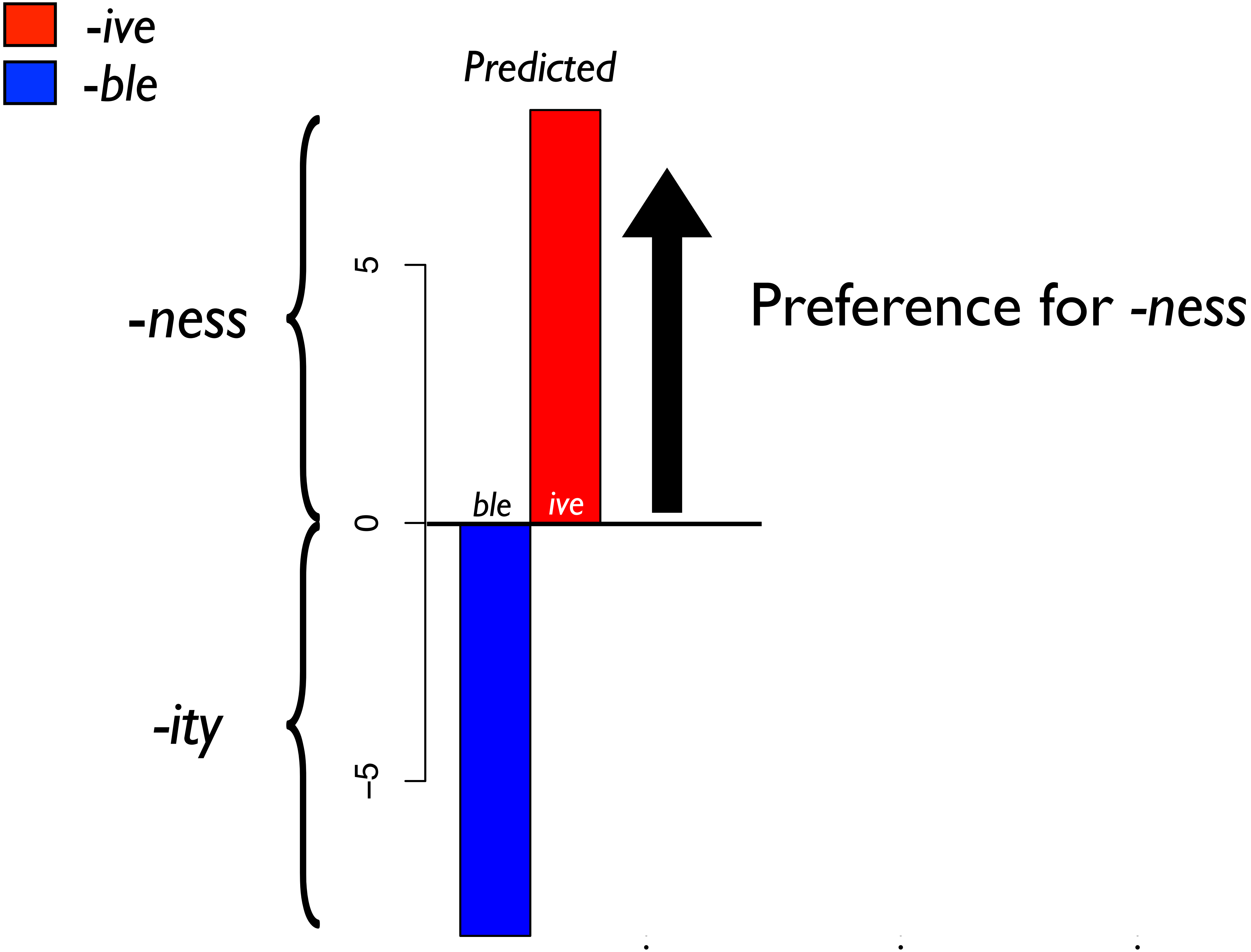
(Anshen & Aronoff, 1981)

remortibility *remortibleness*

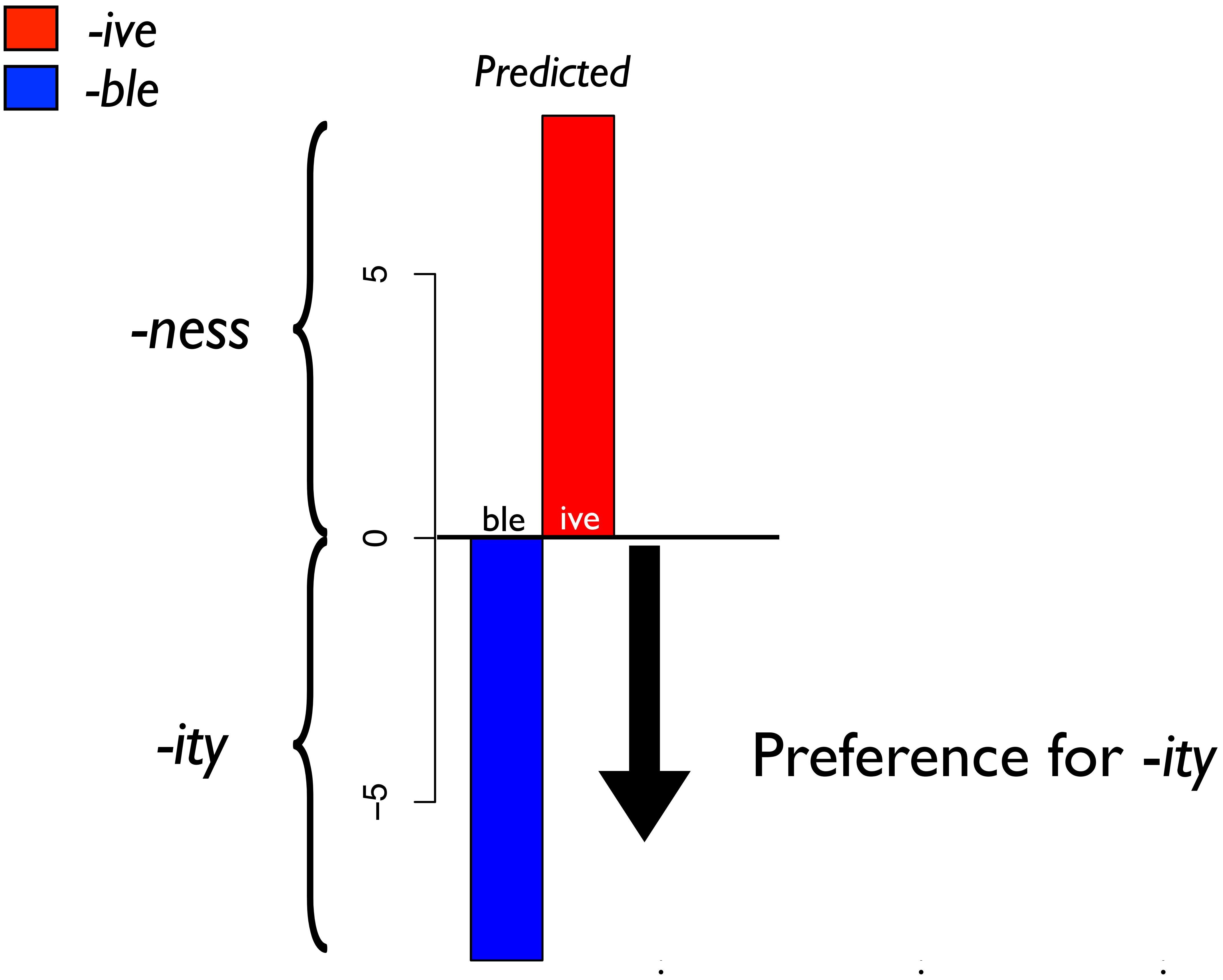
-ivity v. -bility



-ivity v. *-bility*

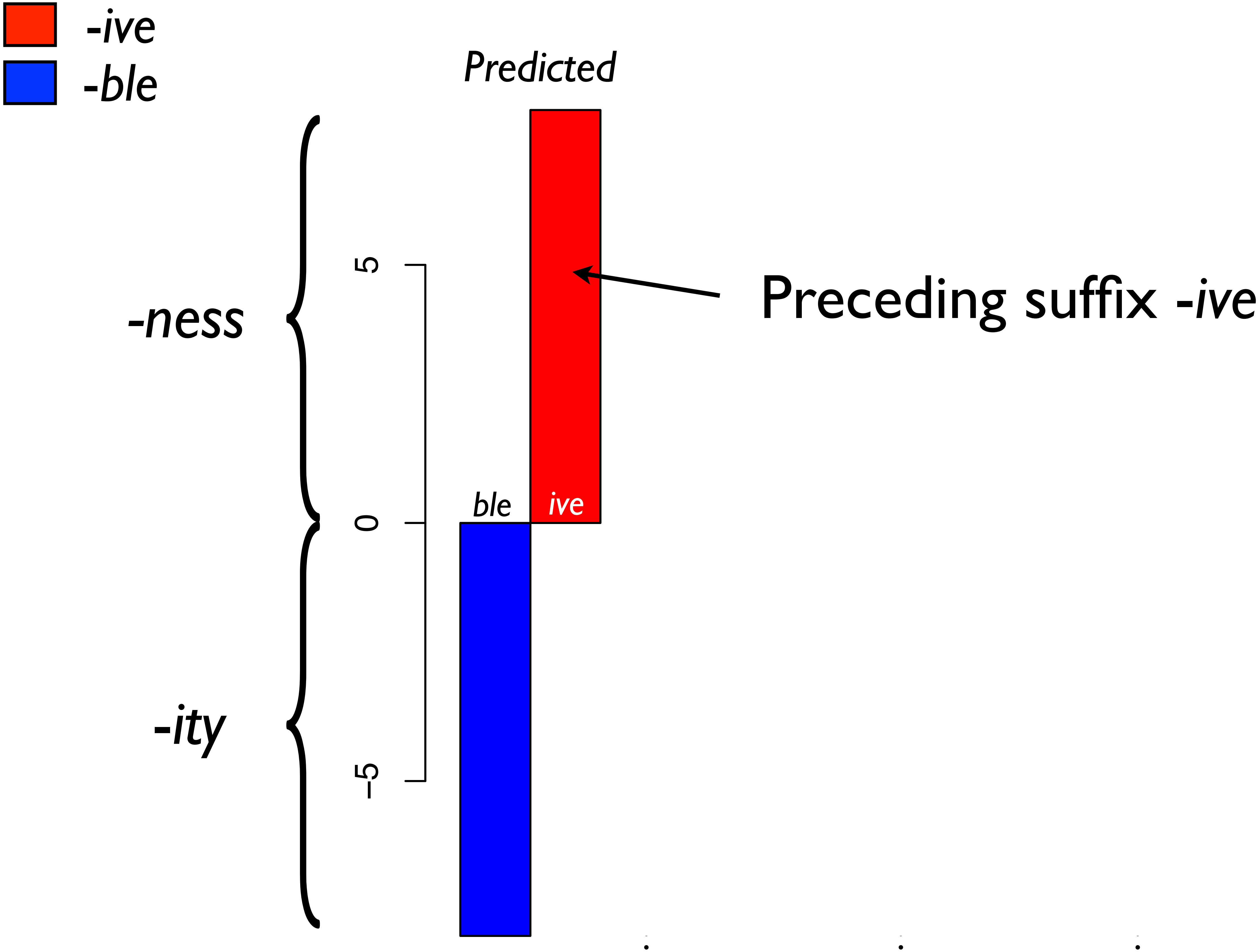


-ivity v. *-bility*



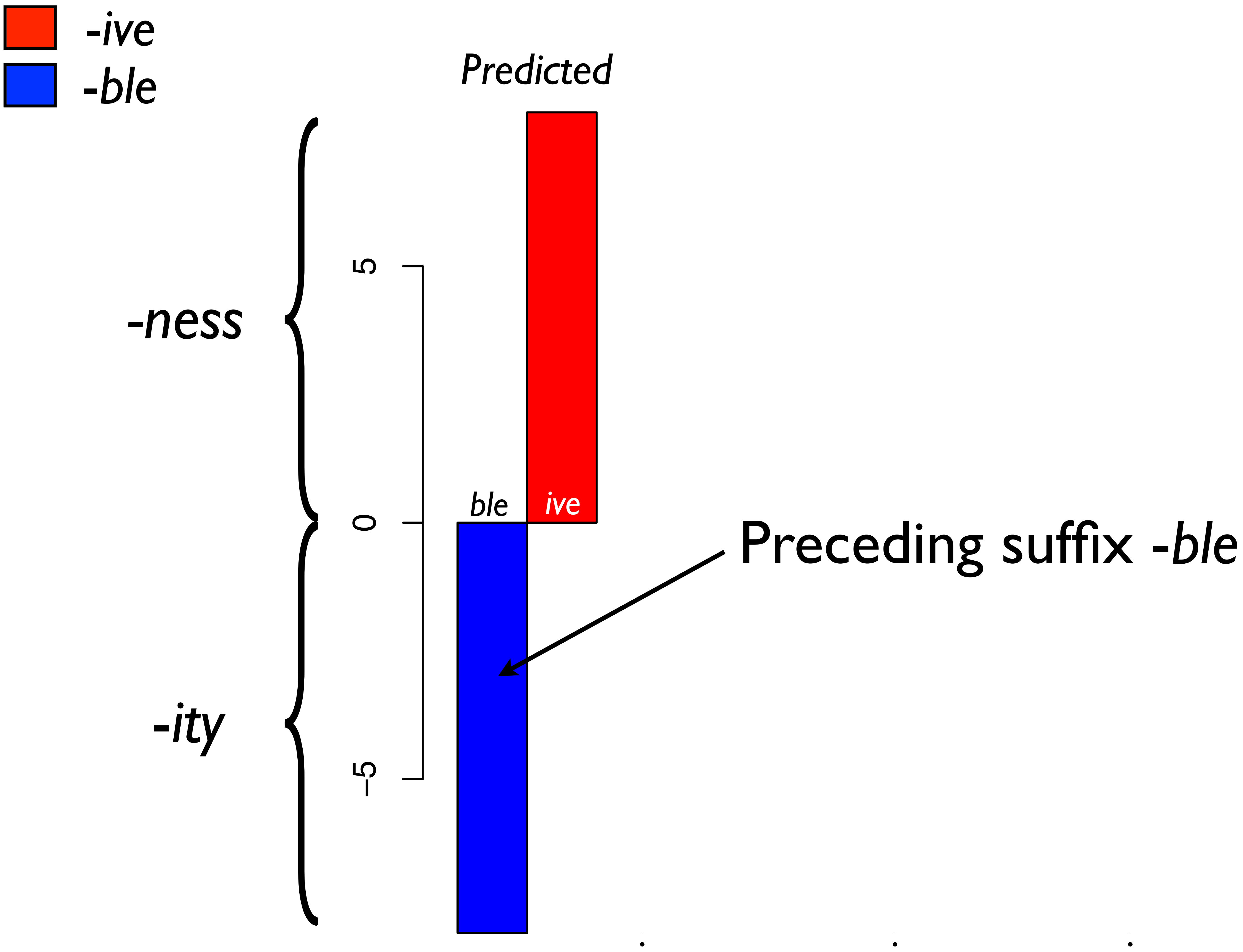
Preference for *-ity*

-ivity v. *-bility*



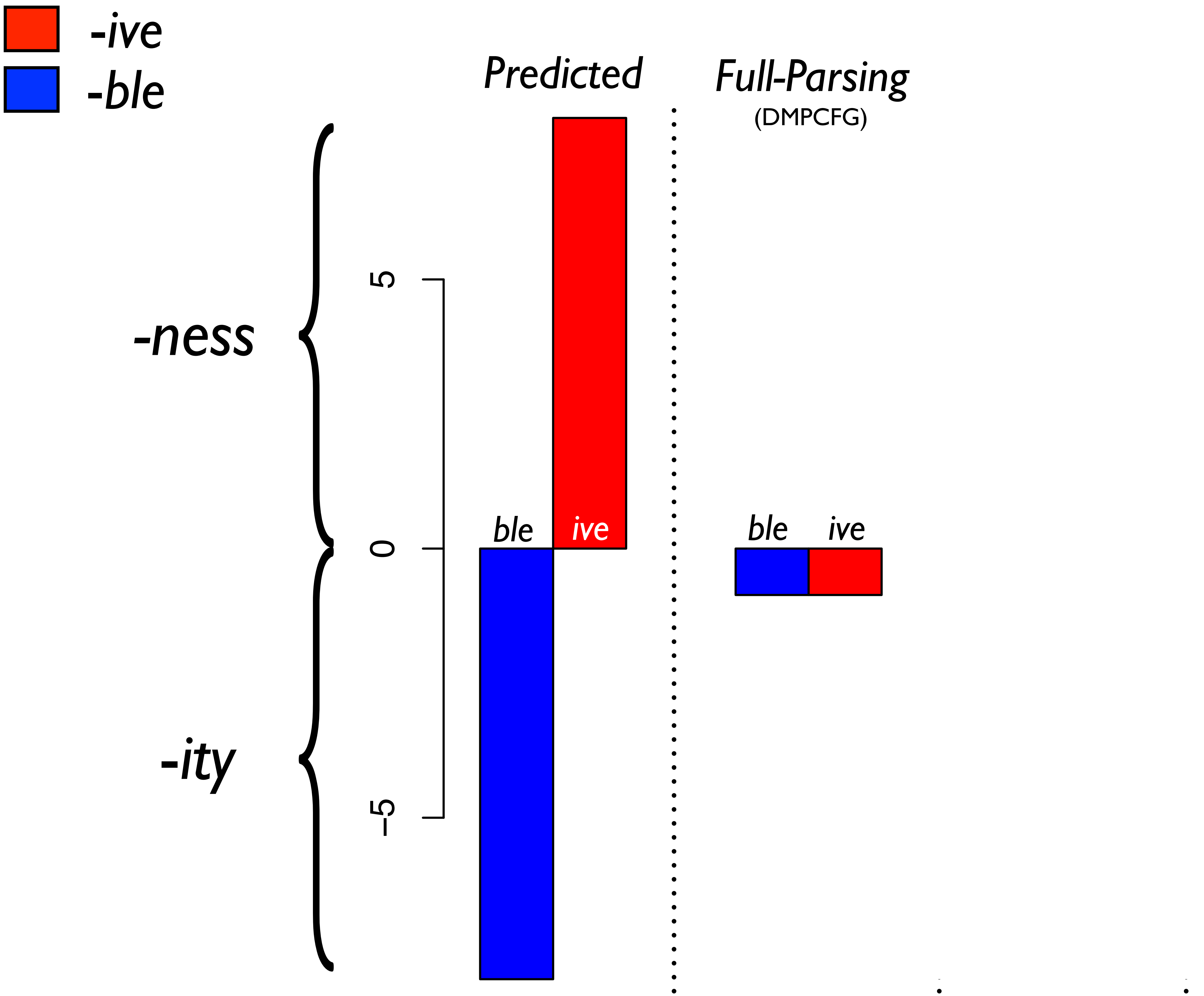
Preceding suffix *-ive*

-ivity v. *-bility*

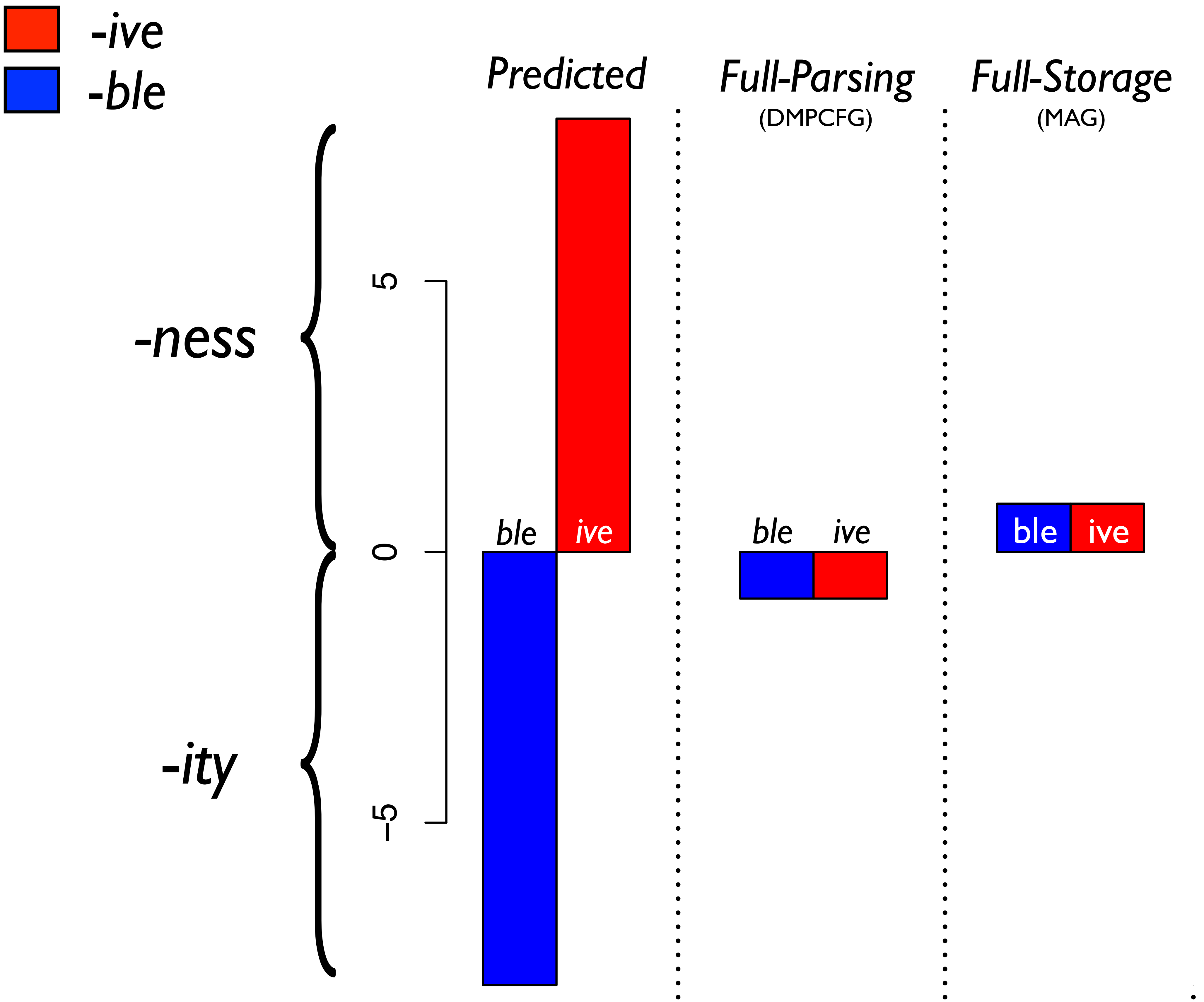


Preceding suffix *-ble*

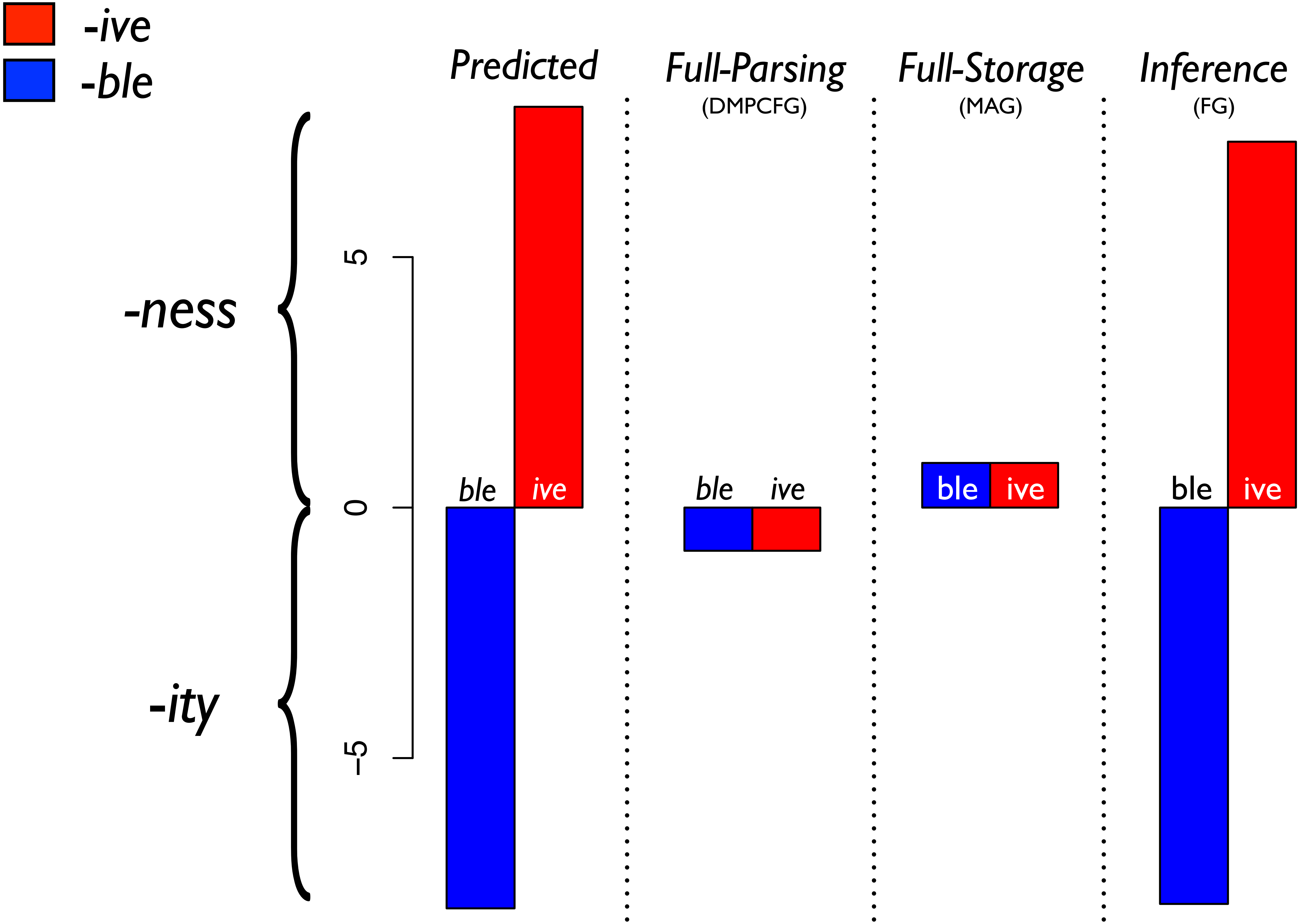
Full-Parsing (DMPCFG)



Full-Storage (MAG)



Inference-Based (FG)



Overview

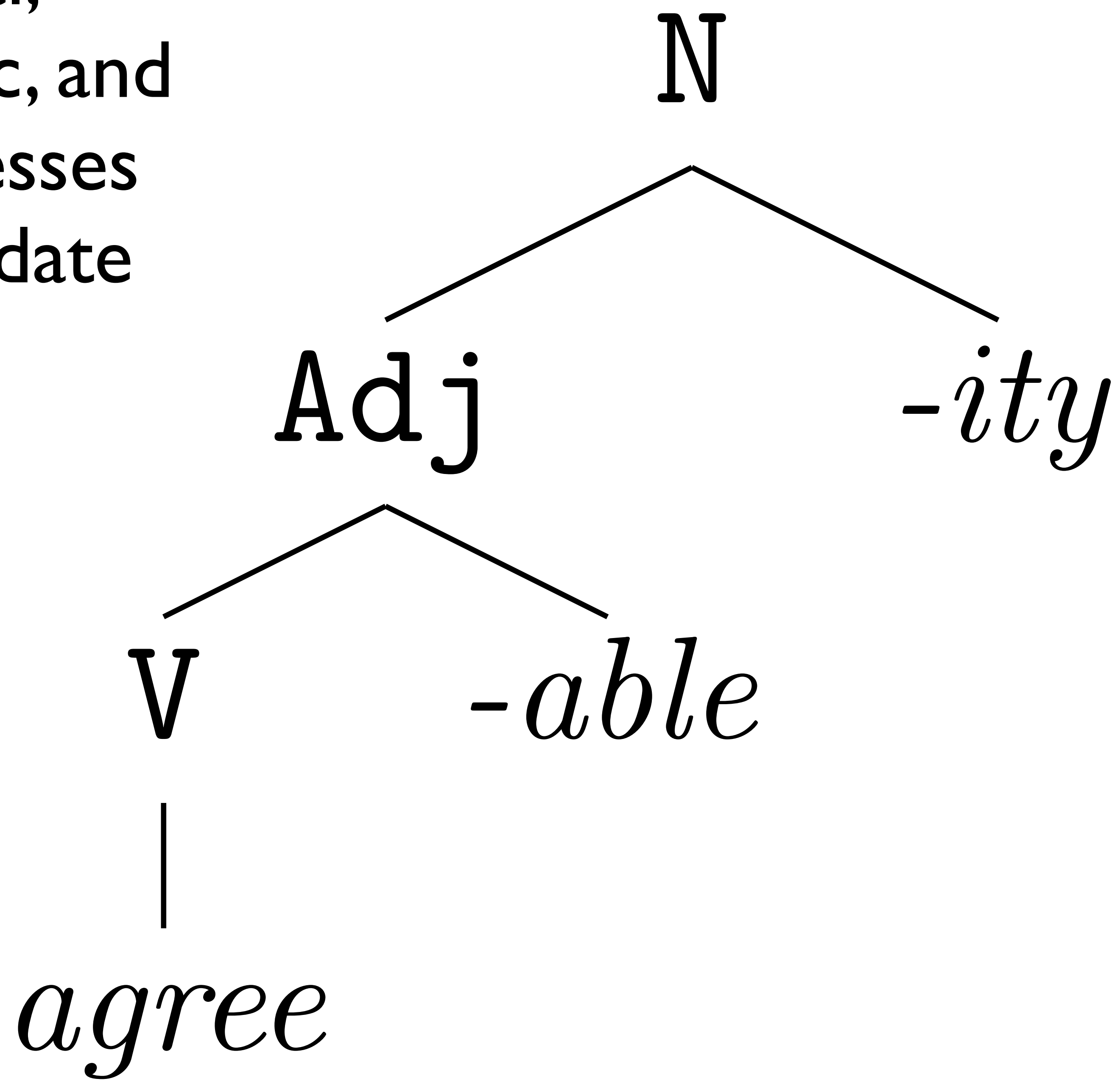
Part I: An approach to lexical uncertainty.

Part II: Detailed case study in morphology.

Part III: Unsupervised models of language learning.

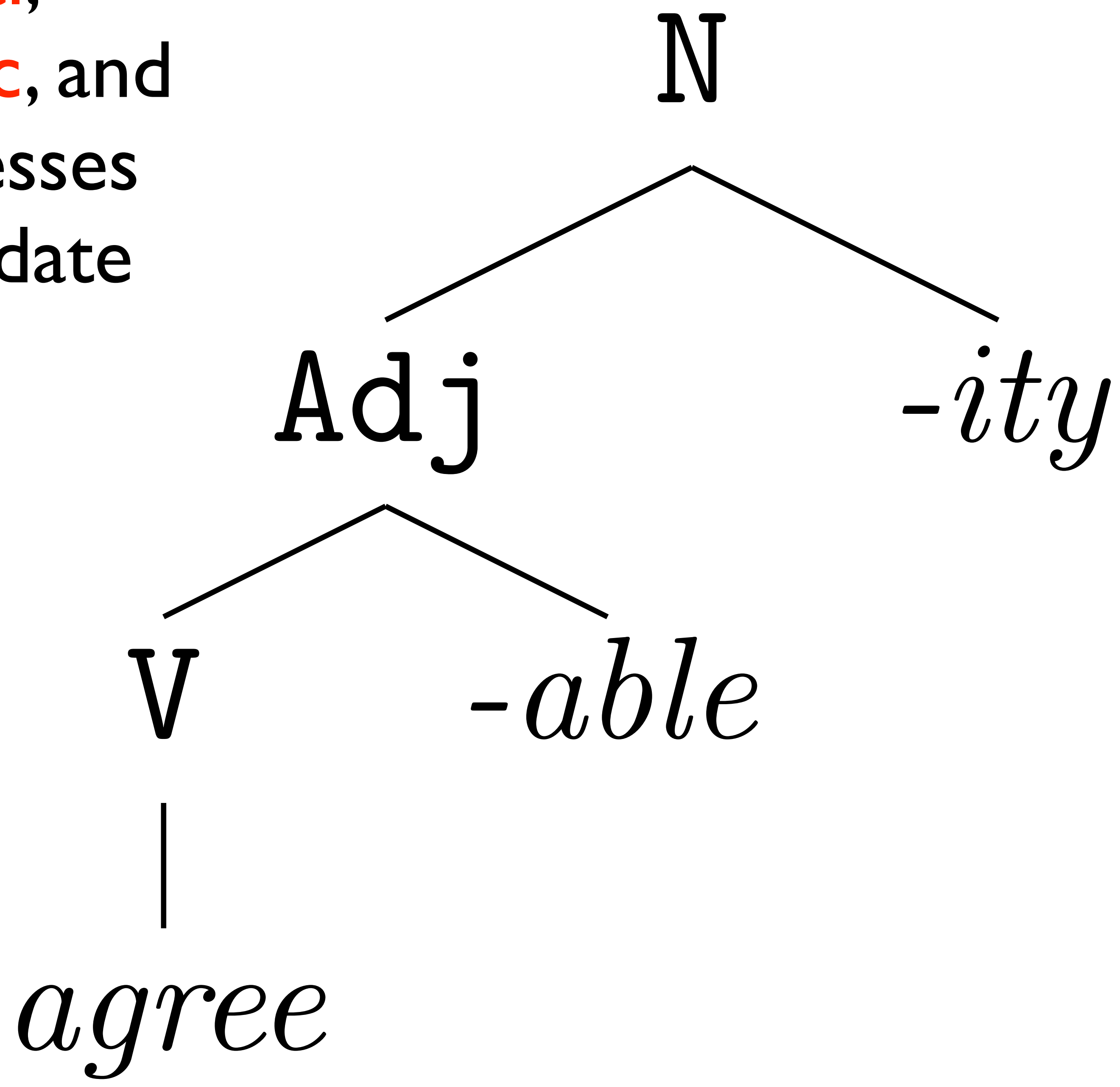
Morphology Studies

Phonological, morphosyntactic, and semantic processes generate candidate analyses.



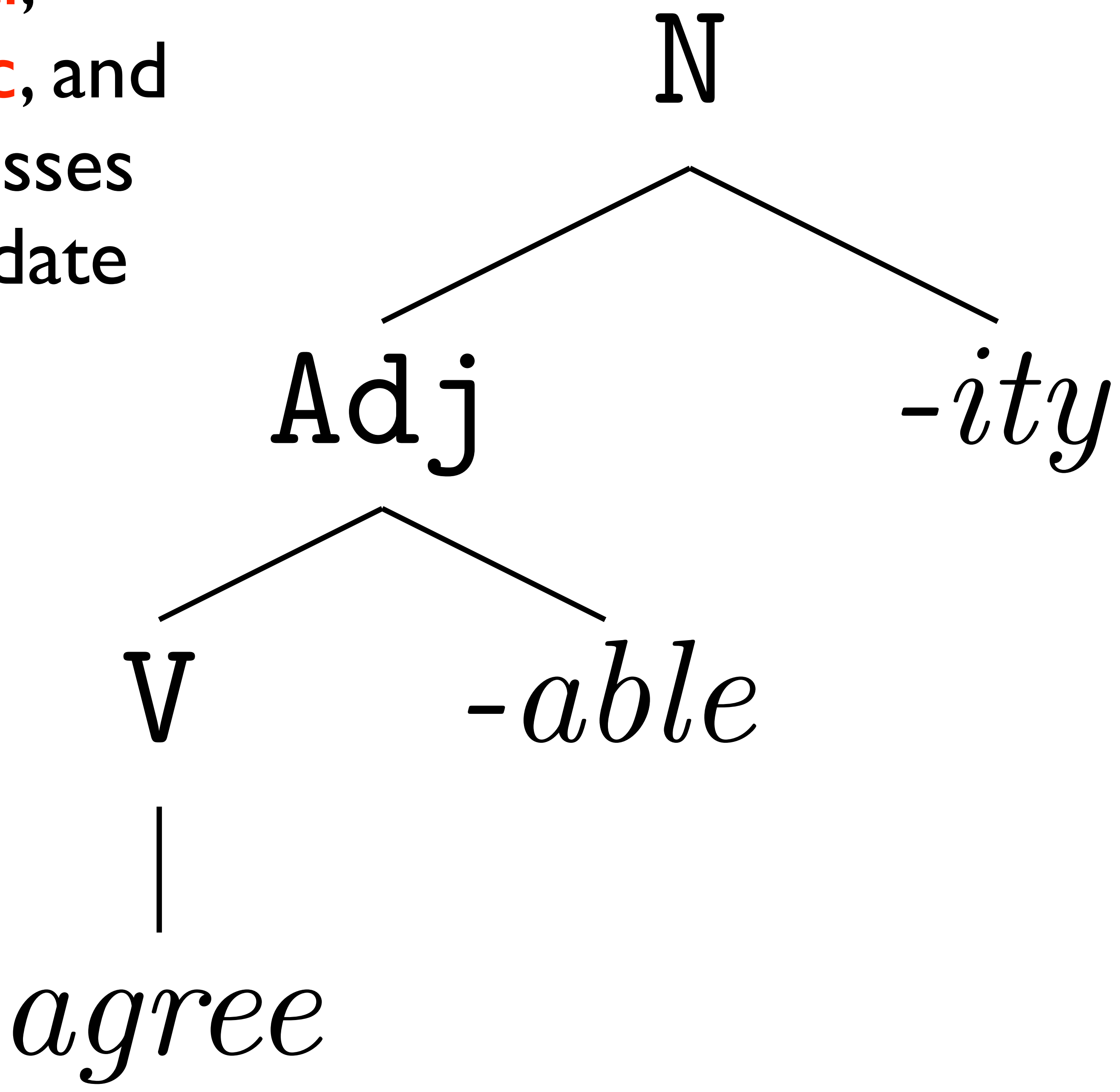
Morphology Studies

Phonological, morphosyntactic, and semantic processes generate candidate analyses.



Unsupervised Learning

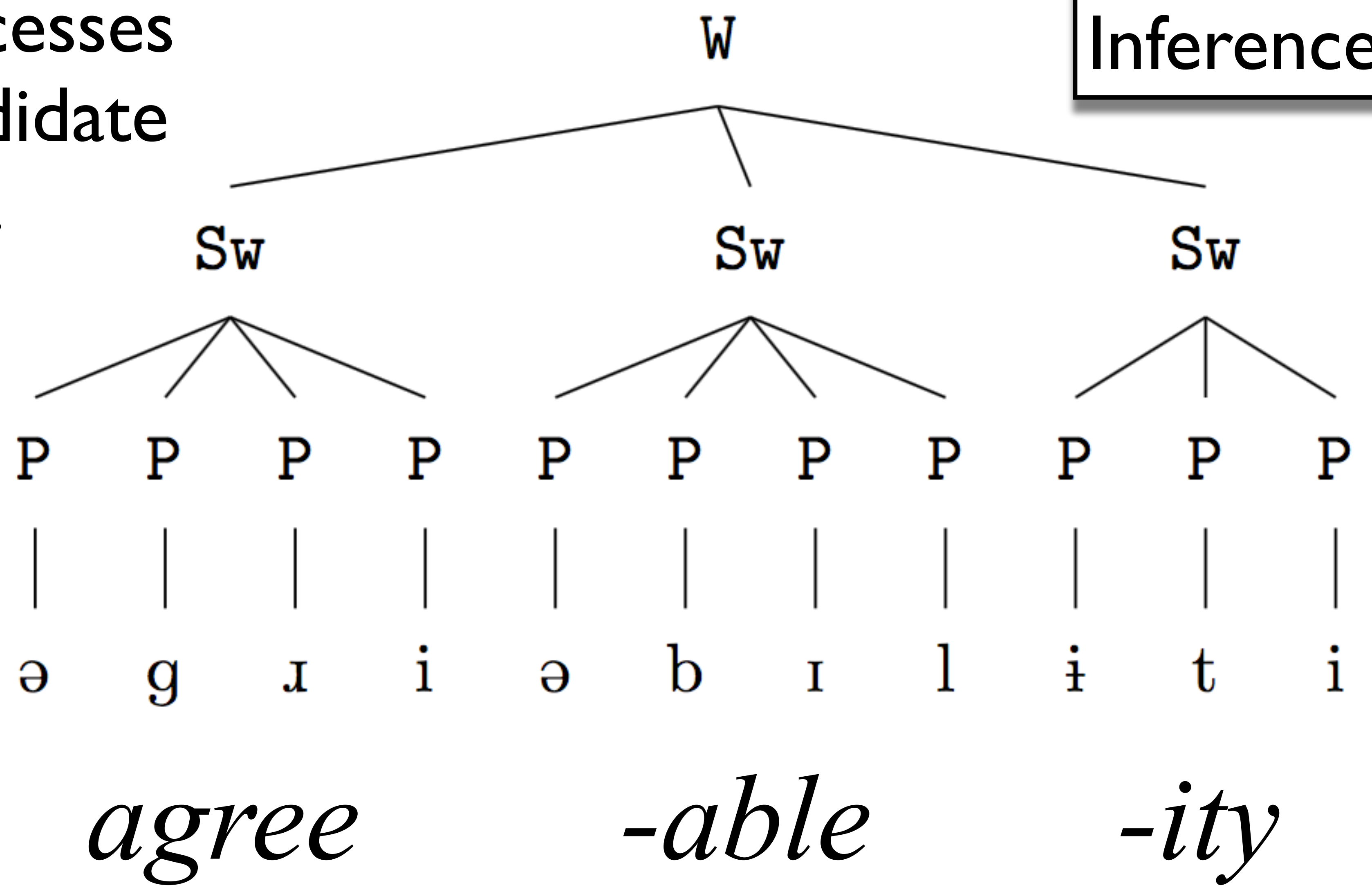
Phonological, morphosyntactic, and semantic processes generate candidate analyses.



Unsupervised Learning

Phonological, morphosyntactic, and semantic processes generate candidate analyses.

No Productivity Inference



Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)

- First step in integrating models of phonetic, phonological, and morphological structure learning.

- Completely unsupervised learning of words, and morphemes from speech (acoustic input).

- Uses similar storage-computation tradeoffs in each component.

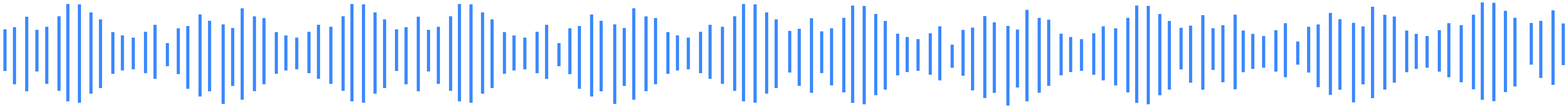
Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

1. Model of unsupervised learning of phonological units from acoustic data.
2. Model of mapping between underlying (lexical) to surface phonological units.
3. Model of morphological structure.
4. Model of lexical storage.

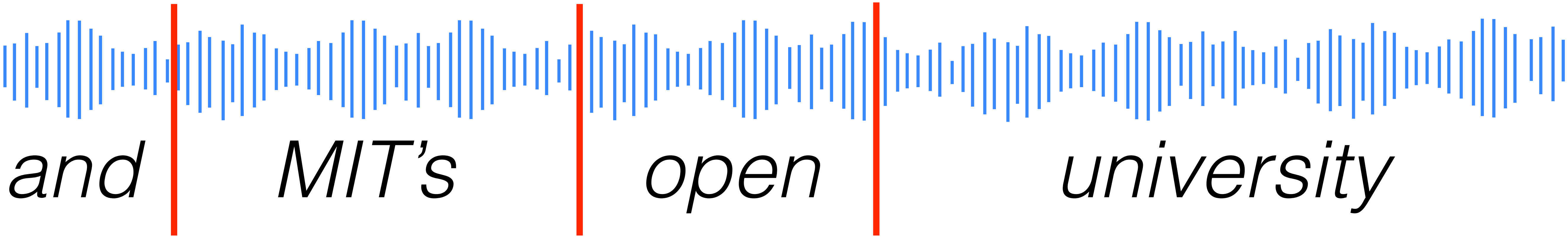
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)



Unsupervised Lexicon Discovery from Speech

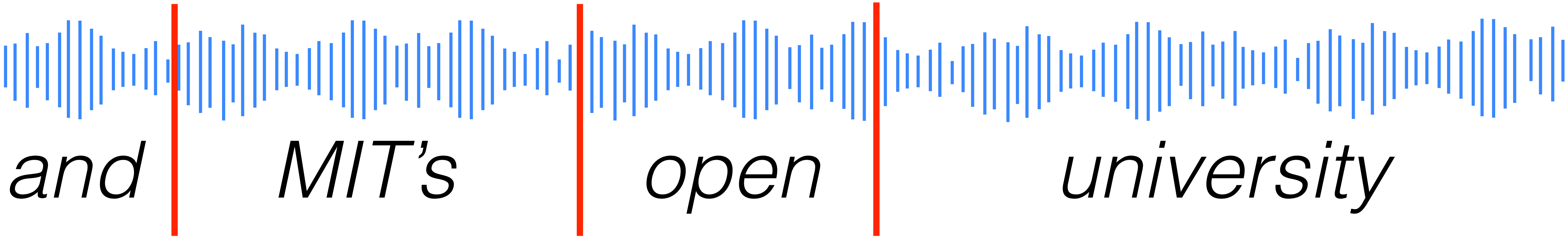
(Lee, O'Donnell, & Glass, 2015)



Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

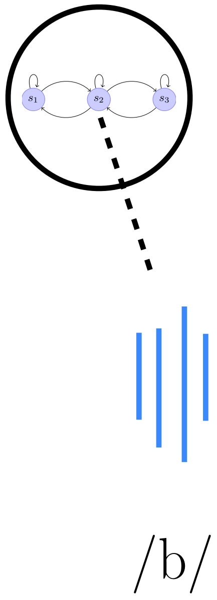
1. **Learning Phonemes** (Lee & Glass, 2012)
   - Phones are hidden Markov models generating acoustics.
   - Unbounded number of phones.
   - Dirichlet process prior over phone inventory.
     - Fewer more reusable phones.

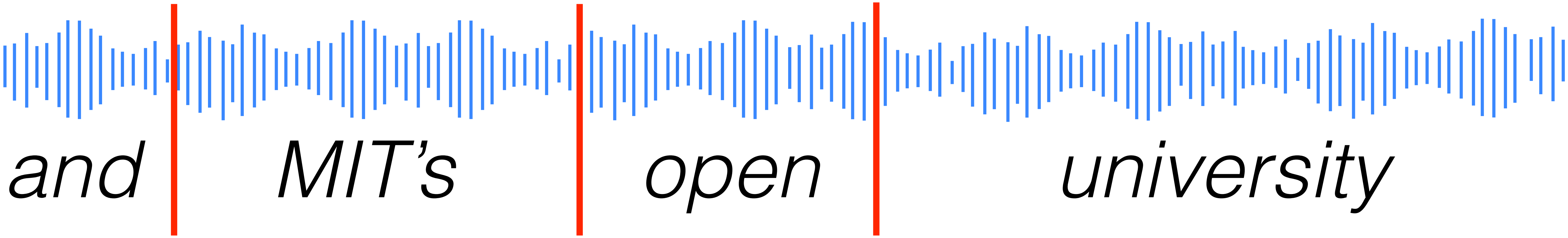


Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

1.Learning Phonemes (Lee & Glass, 2012)

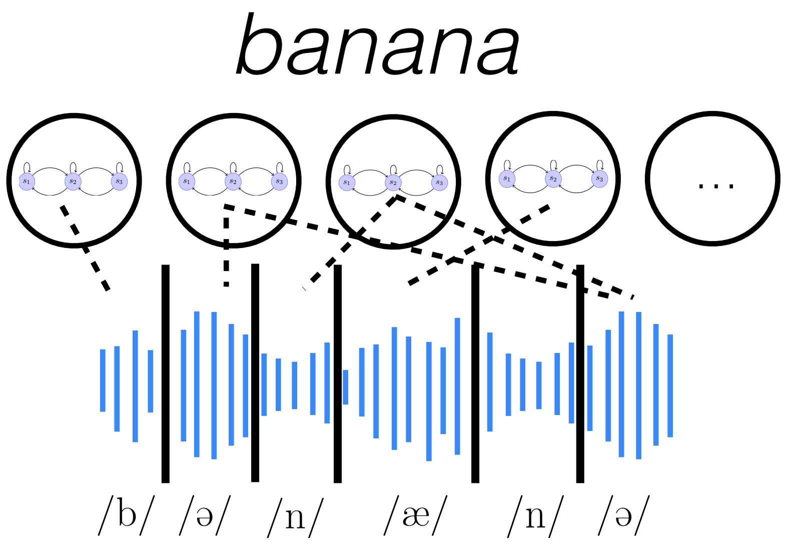




Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

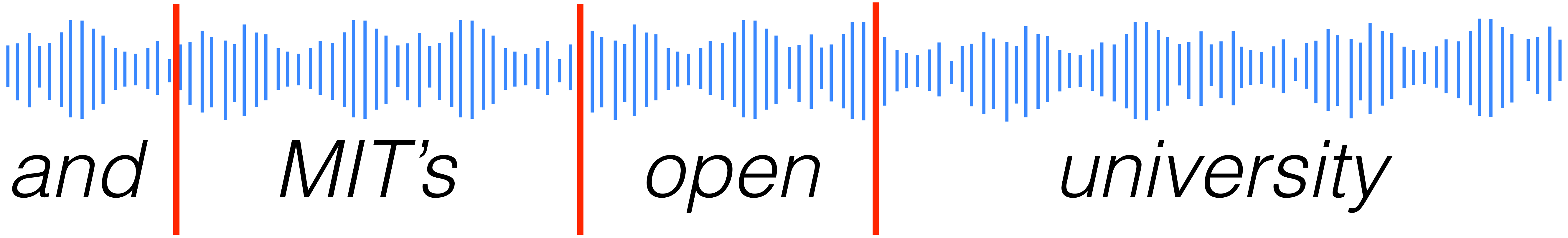
1. Learning Phonemes (Lee & Glass, 2012)



banana

…

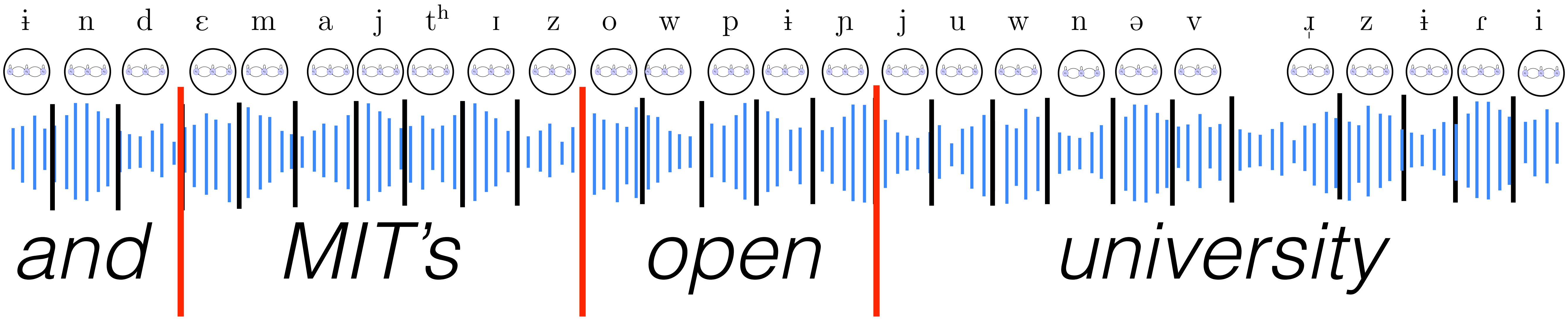
/b/ /ə/ /n/ /æ/ /n/ /ə/



and *MIT's* *open* *university*

Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)

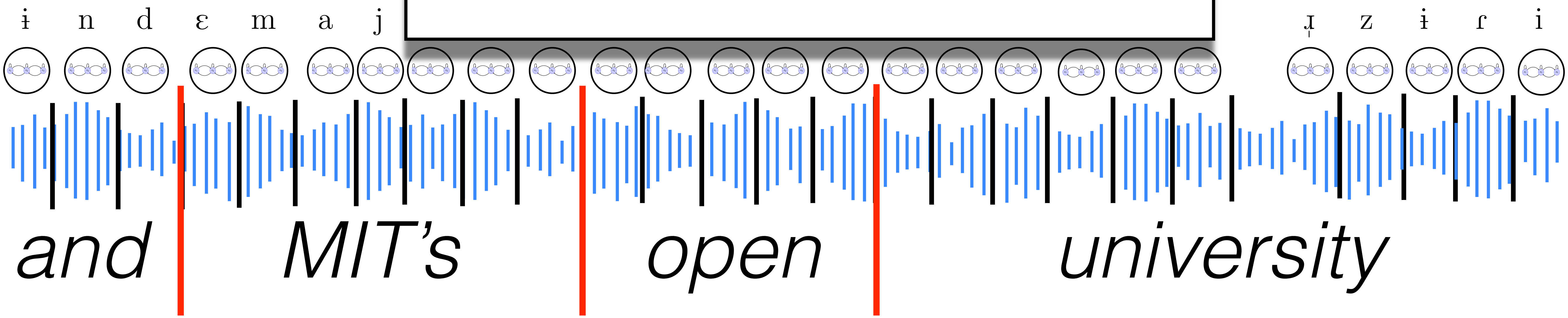


Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

2. Underlying and Surface Phones

- Highly simplified phonetics/phonology.
- Allophony, phonetic variation, coarticulation, noise, etc.

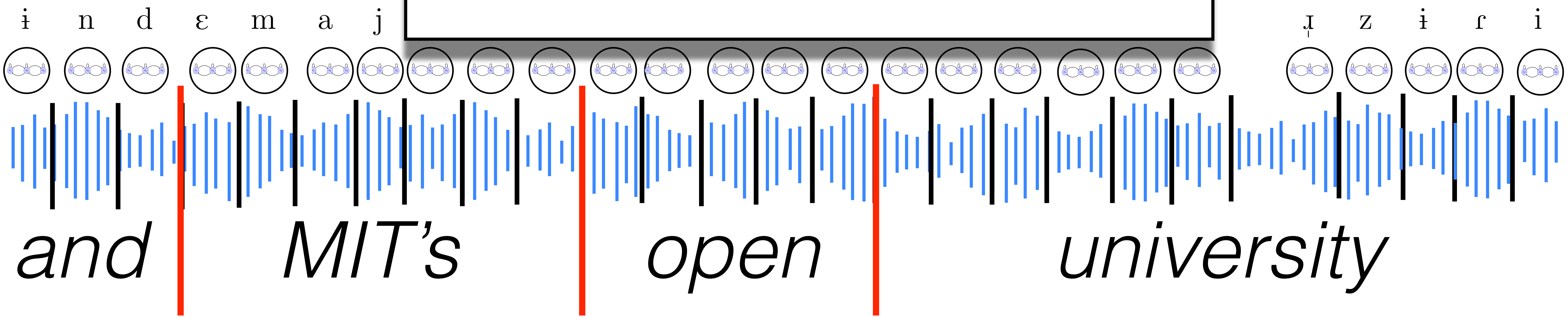


Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

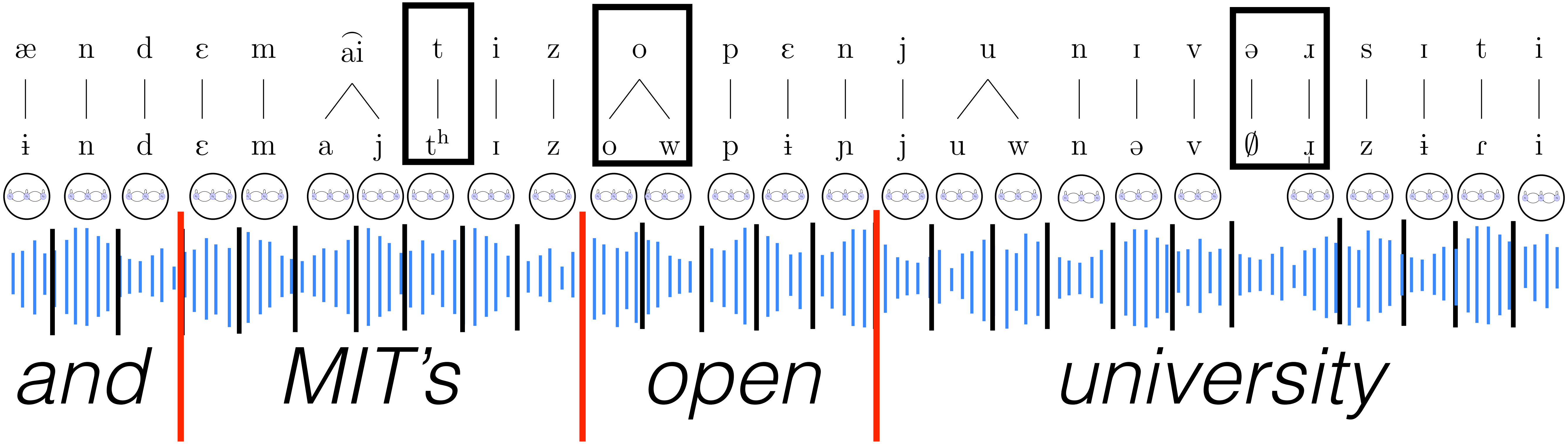
2. Underlying and Surface Phones

- Highly simplified phonetics/phonology.
  - Substitutions.
  - Splits.
  - Deletions.



Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

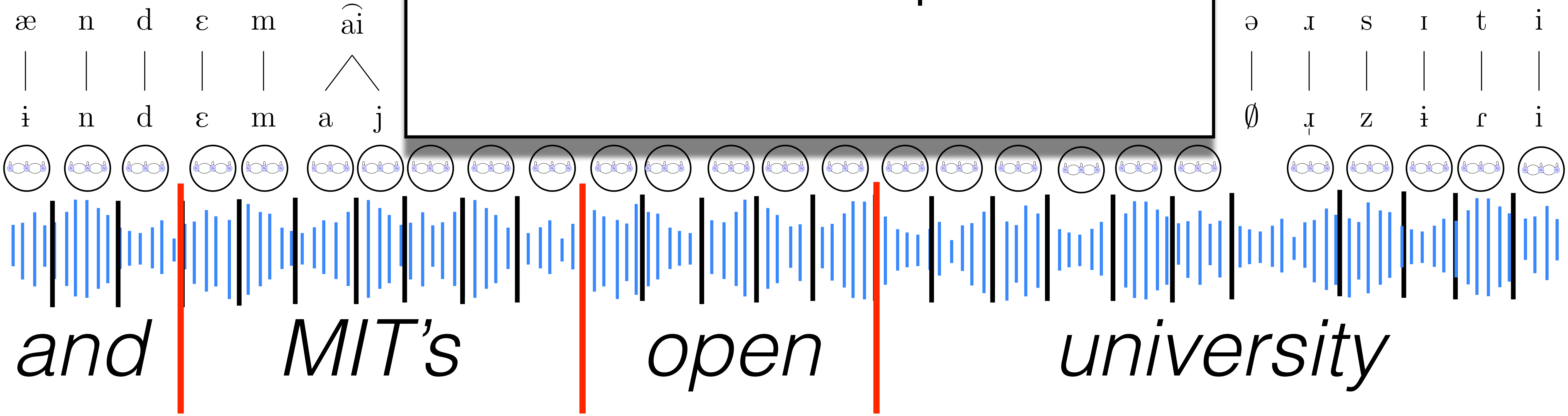

æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ z ɨ ɾ i
and
MIT's
open
university

Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)

3. Model of Morphological Structure

- Utterances consist of words.
- Words consist of sub-words.
- Sub-words consist of phonemes.

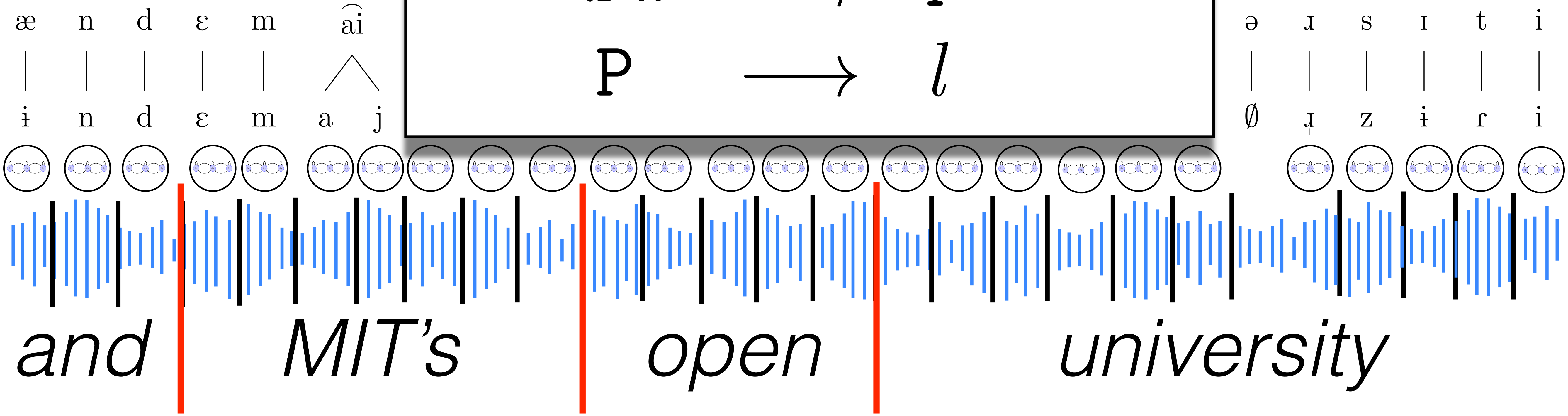


Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)

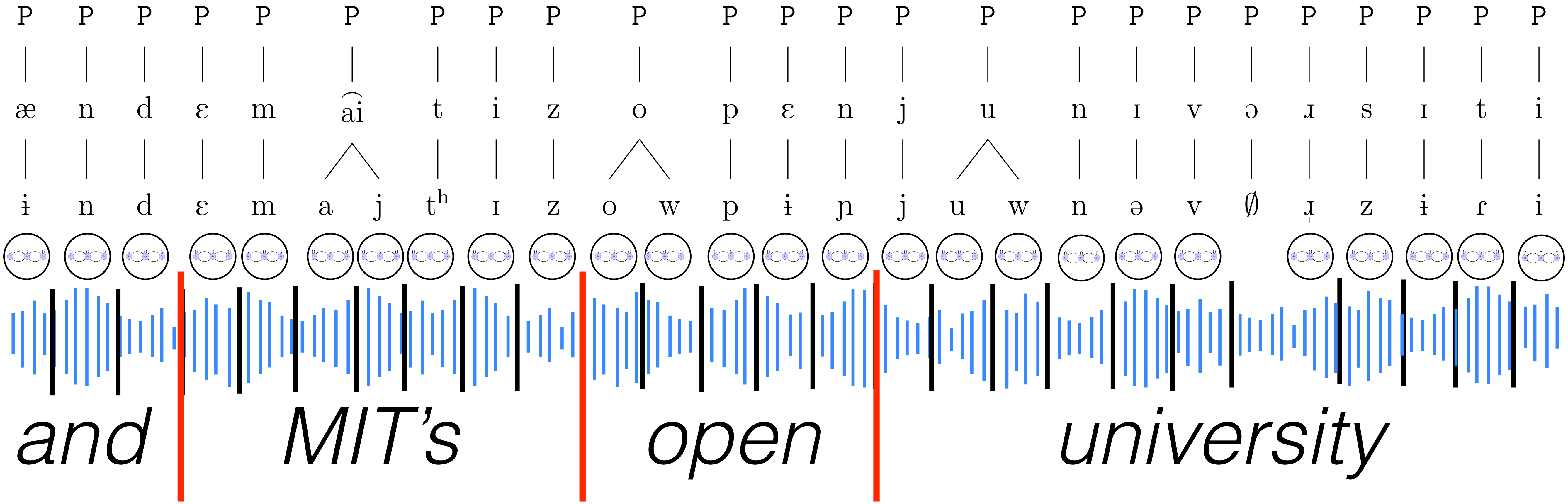
3. Model of Morphological Structure

$$
\begin{aligned}
\mathtt{U} &\longrightarrow \mathtt{W}^* \\
\mathtt{W} &\longrightarrow \mathtt{Sw}^* \\
\mathtt{Sw} &\longrightarrow \mathtt{P}^* \\
\mathtt{P} &\longrightarrow l
\end{aligned}
$$



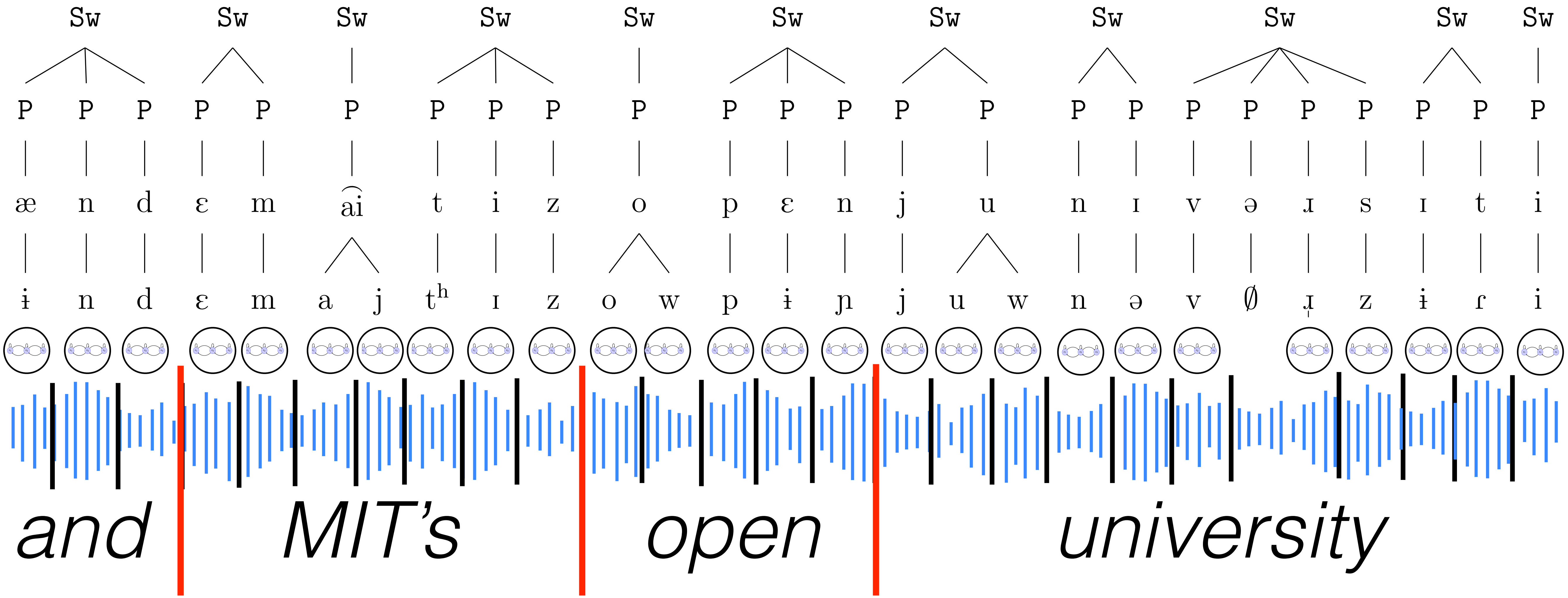
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)


P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i
and
MIT’s
open
university

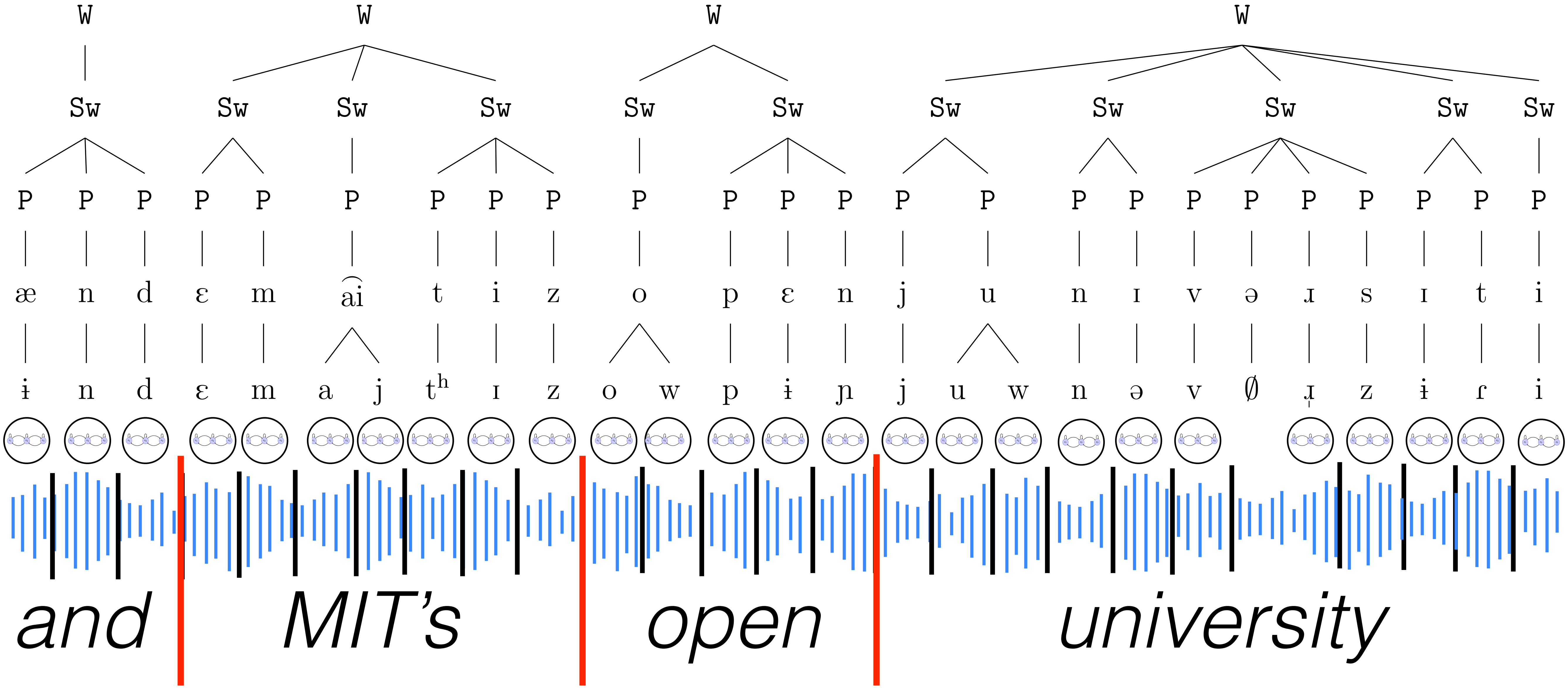
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)


Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i
and
MIT’s
open
university

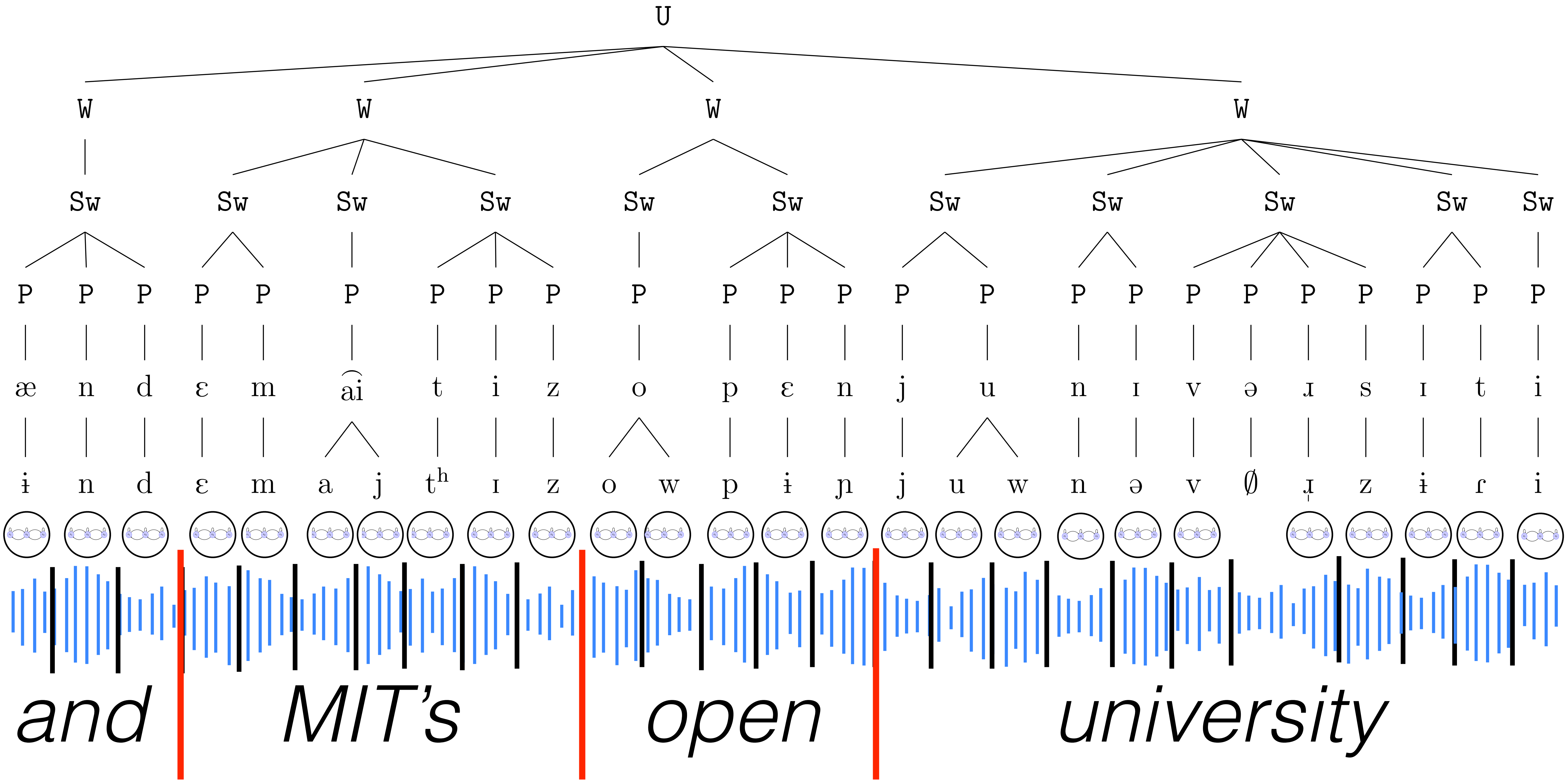
Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)


W W W W
Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i
and
MIT's
open
university

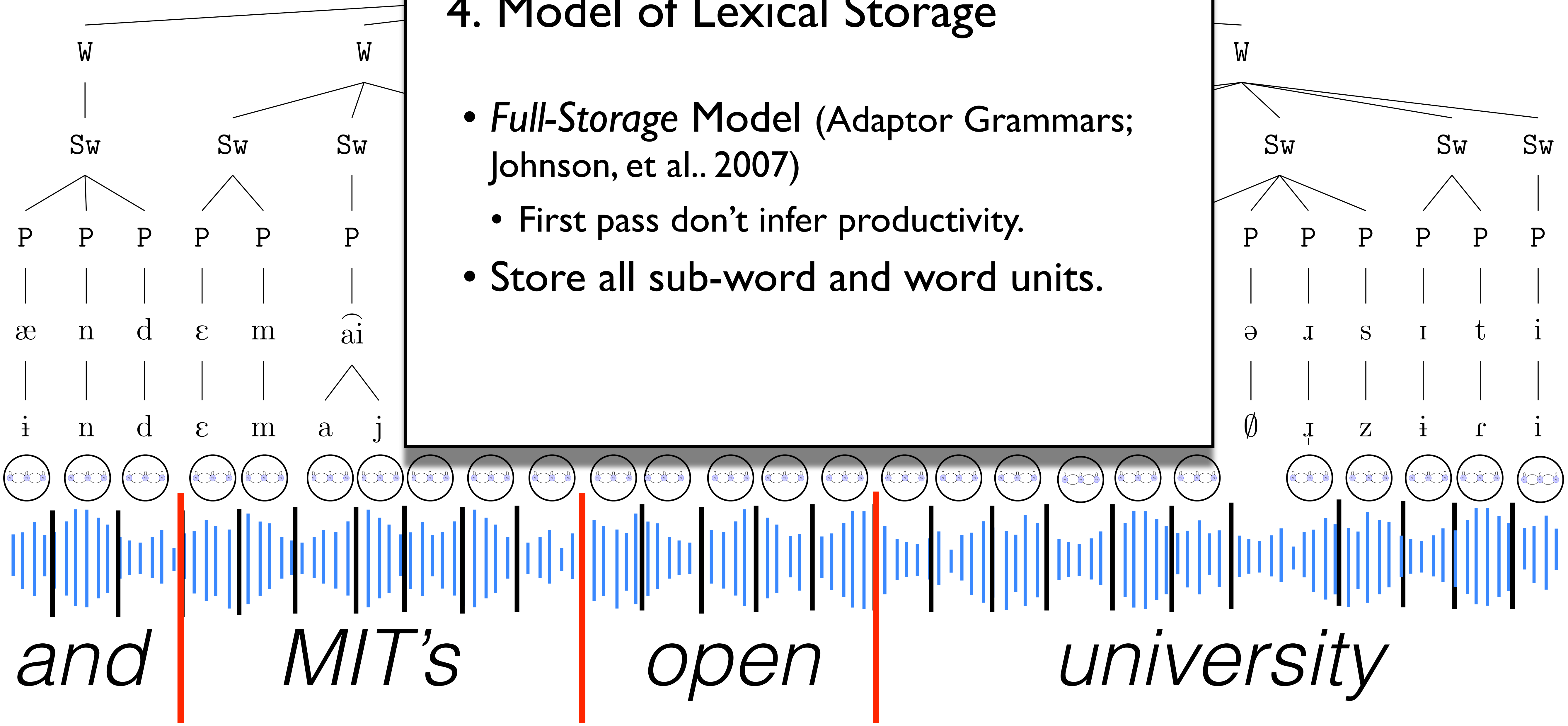
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)


U
W W W W
Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i
and
MIT’s
open
university

Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)

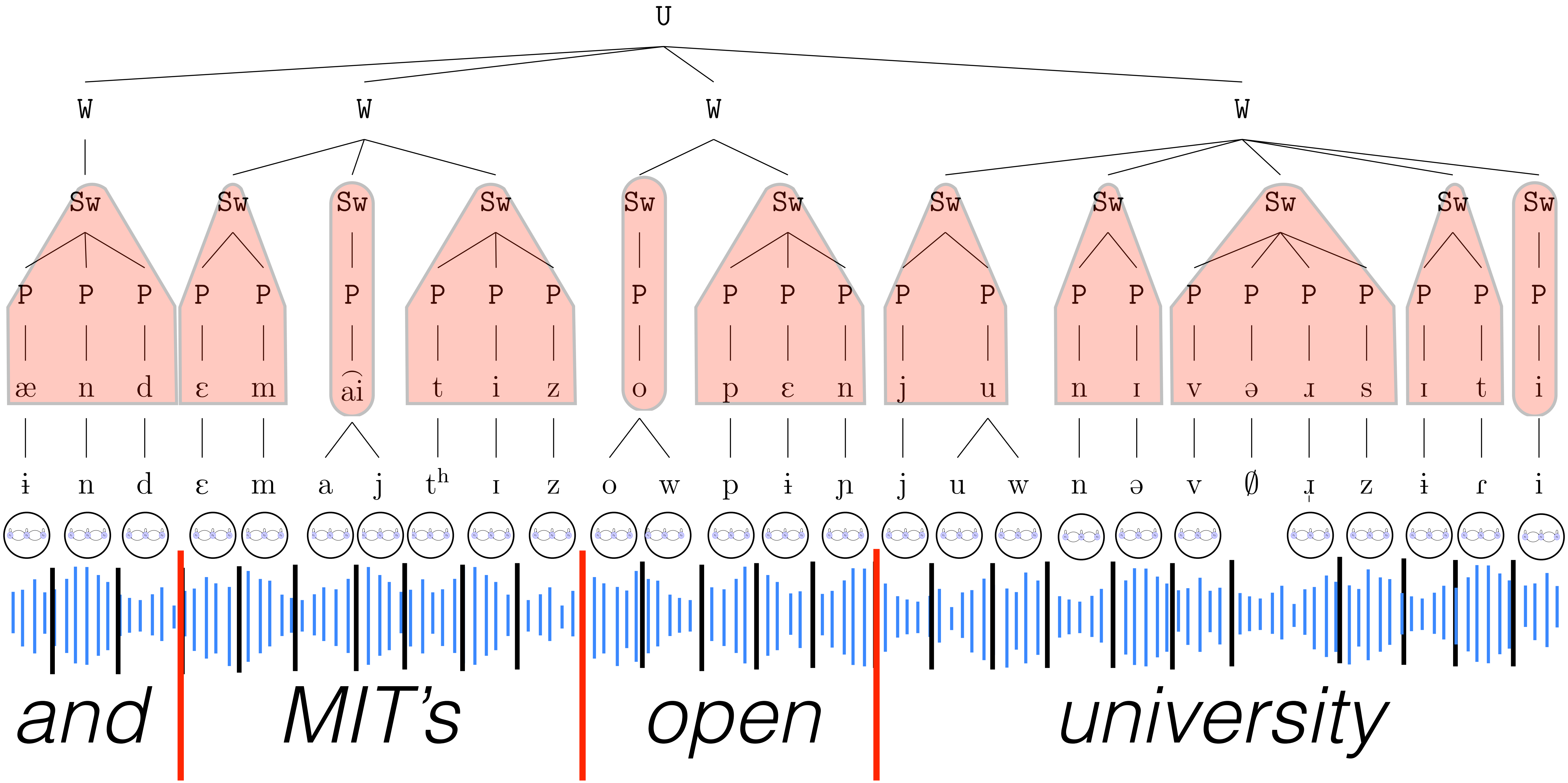


4. Model of Lexical Storage

- *Full-Storage* Model (Adaptor Grammars; Johnson, et al.. 2007)
  - First pass don’t infer productivity.
- Store all sub-word and word units.

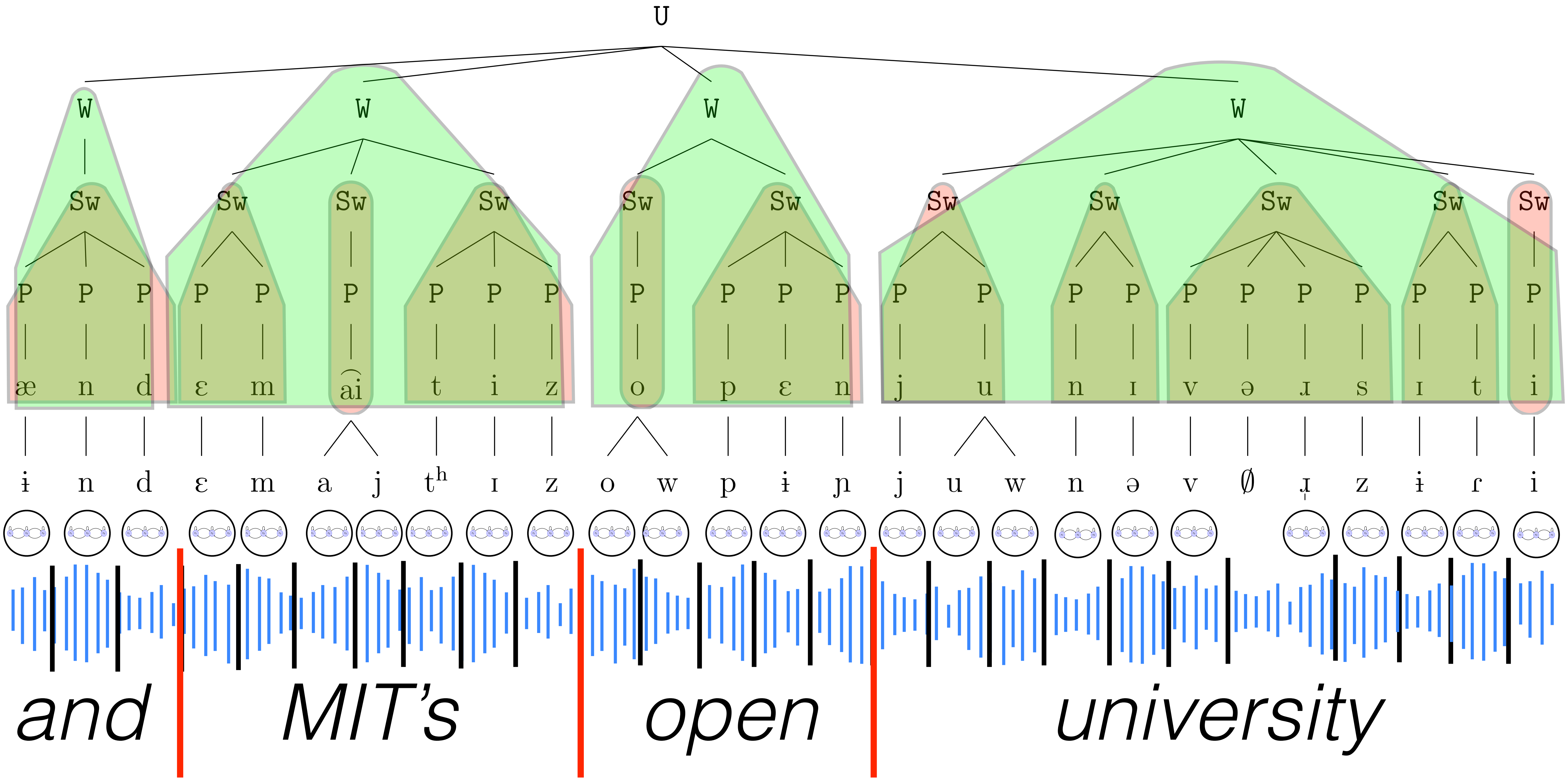
Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)


U
W W W W
Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i
and
MIT's
open
university

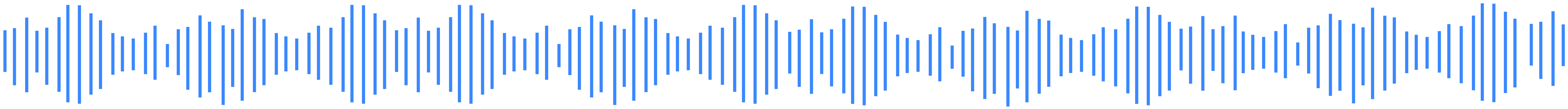
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)


U
W W W W
Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ z ɨ ɾ i
and
MIT’s
open
university

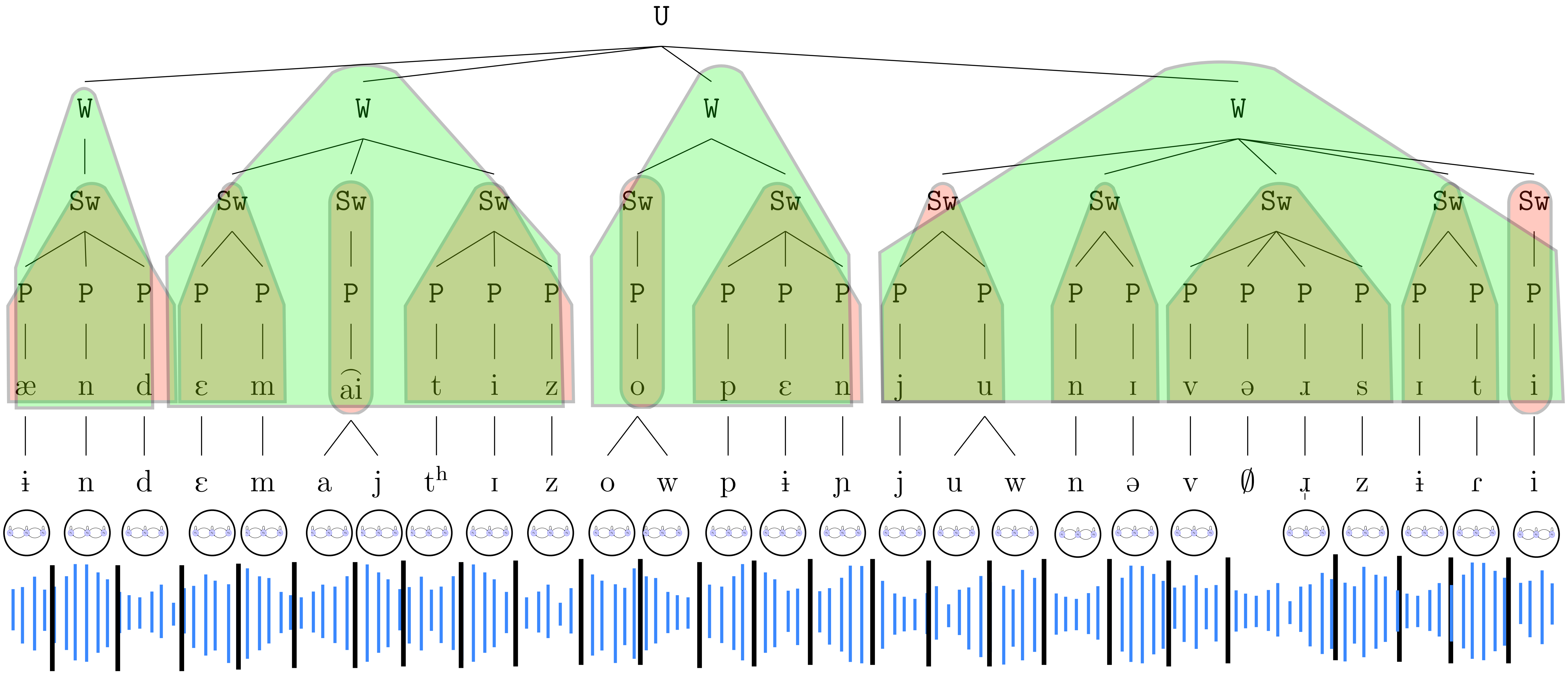
Unsupervised Lexicon Discovery from Speech

(Lee, O’Donnell, & Glass, 2015)



Unsupervised Lexicon Discovery from Speech

(Lee, O'Donnell, & Glass, 2015)


U
W W W W
Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m ai t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i

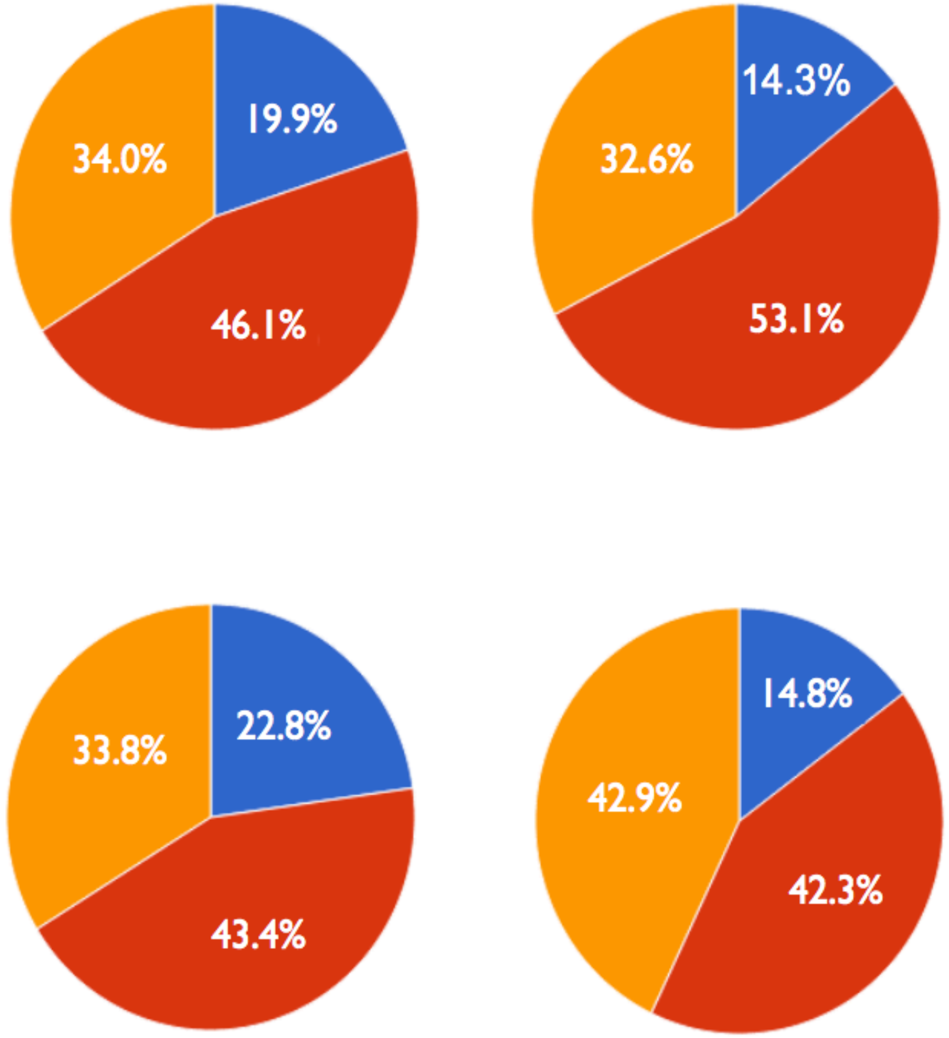
Key Scientific Results

- First unsupervised model of this scope.
- Outperforms state-of-art system for spoken term detection (Zhang, et al. 2013).
- Phone segmentation ~75%, word segmentation less great <20%.
- Lexical units.
- Synergistic Interactions.

Key Scientific Results

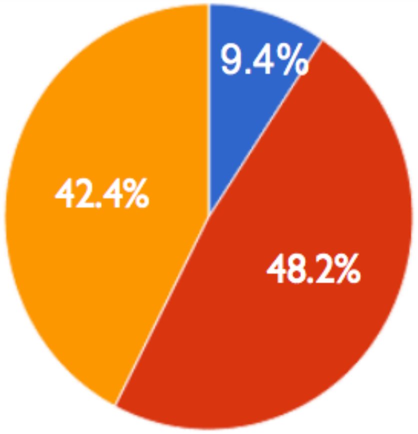
- Lexical units.
- Synergistic Interactions.

Lexical Unit Study

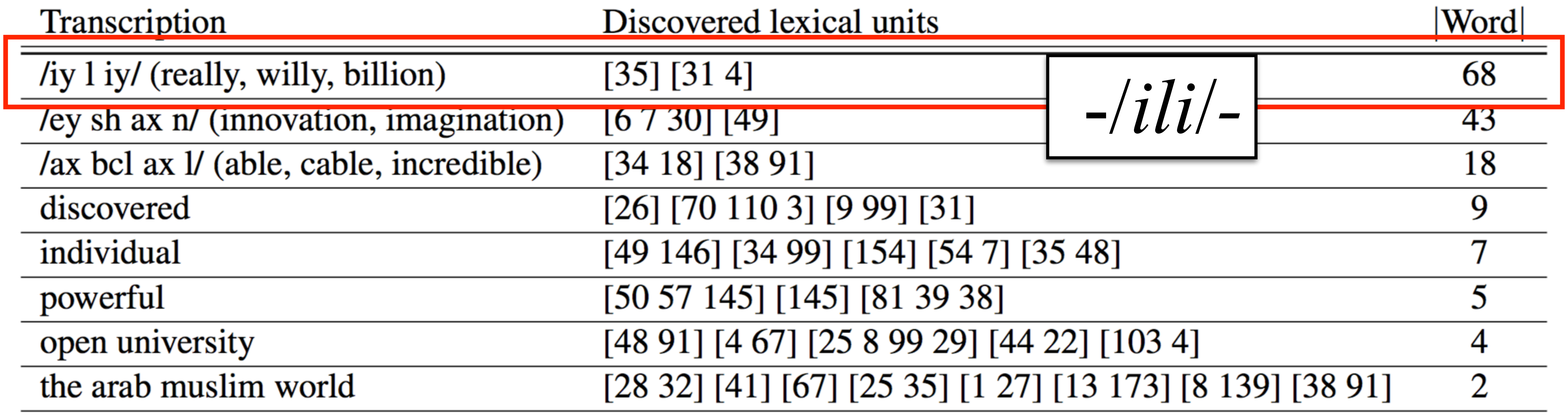


<1.5 Hours of Speech

$$\begin{array}{lcl} \mathtt{U} & \longrightarrow & \mathtt{W}^* \\ \mathtt{W} & \longrightarrow & \mathtt{Sw}^* \\ \mathtt{Sw} & \longrightarrow & \mathtt{P}^* \\ \mathtt{P} & \longrightarrow & \mathtt{l} \end{array}$$



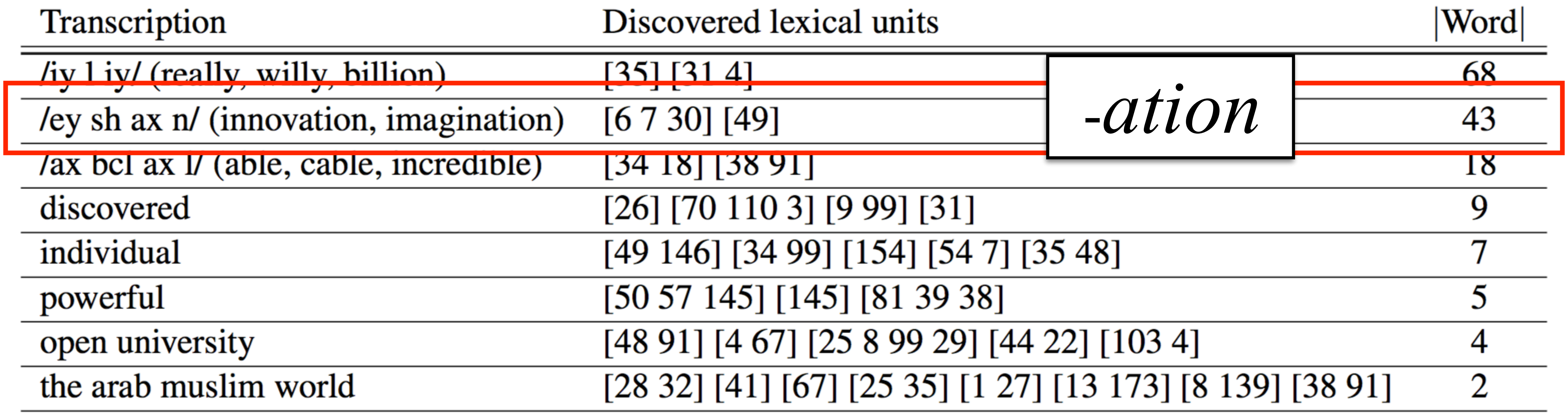
Lexical Units



Transcription	Discovered lexical units	|Word|
/iy l iy/ (really, willy, billion)	[35] [31 4]	68
/ey sh ax n/ (innovation, imagination)	[6 7 30] [49]	43
/ax bcl ax l/ (able, cable, incredible)	[34 18] [38 91]	18
discovered	[26] [70 110 3] [9 99] [31]	9
individual	[49 146] [34 99] [154] [54 7] [35 48]	7
powerful	[50 57 145] [145] [81 39 38]	5
open university	[48 91] [4 67] [25 8 99 29] [44 22] [103 4]	4
the arab muslim world	[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]	2

-/ili/-

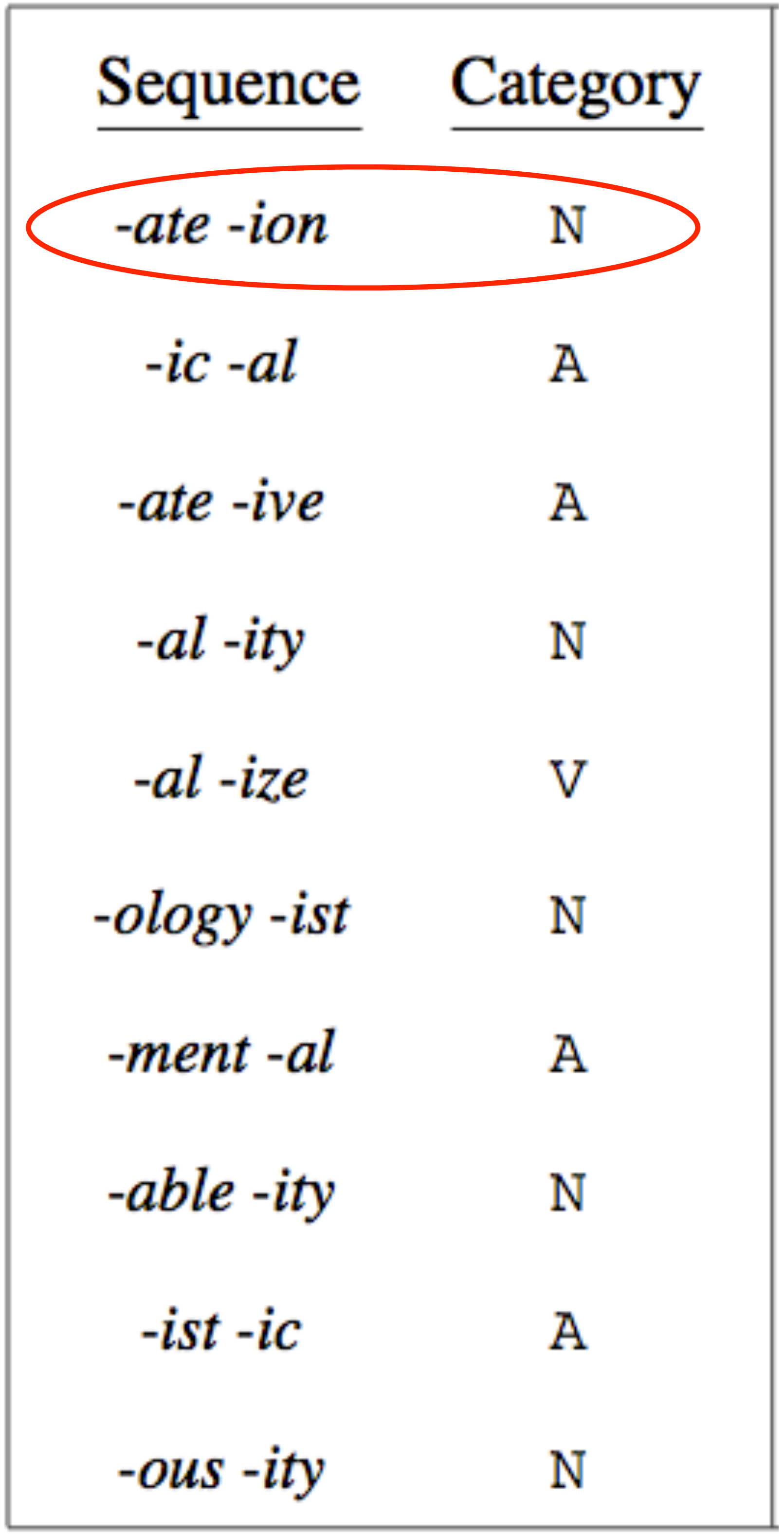
Lexical Units



Transcription	Discovered lexical units	|Word|
/iy l iy/ (really, willy, billion)	[35] [31 4]	68
/ey sh ax n/ (innovation, imagination)	[6 7 30] [49]	43
/ax bcl ax l/ (able, cable, incredible)	[34 18] [38 91]	18
discovered	[26] [70 110 3] [9 99] [31]	9
individual	[49 146] [34 99] [154] [54 7] [35 48]	7
powerful	[50 57 145] [145] [81 39 38]	5
open university	[48 91] [4 67] [25 8 99 29] [44 22] [103 4]	4
the arab muslim world	[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]	2

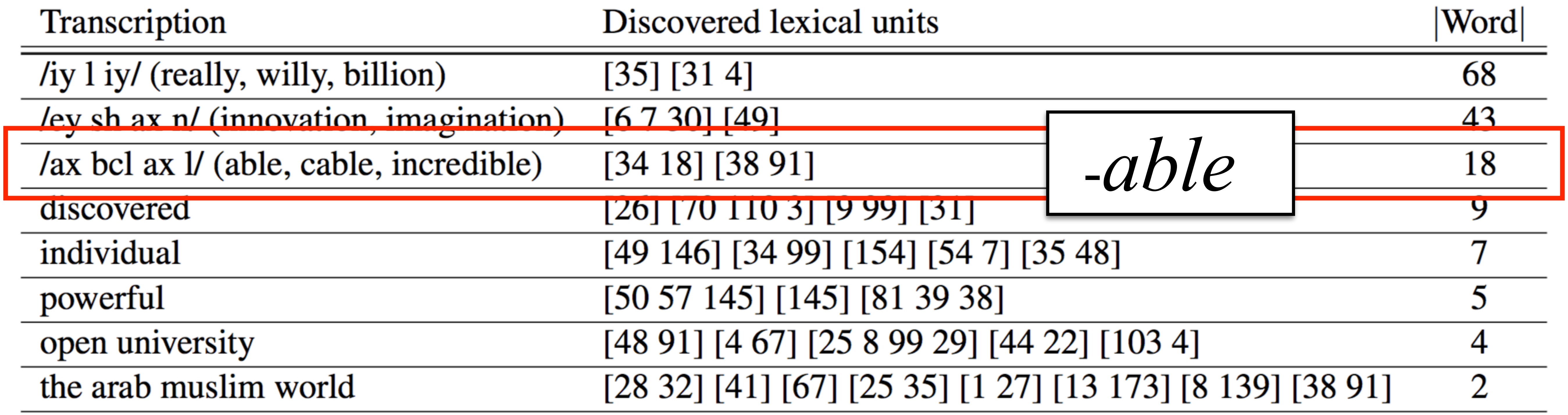
-ation

Predicted Generalizable Combinations



Sequence	Category
-ate -ion	N
-ic -al	A
-ate -ive	A
-al -ity	N
-al -ize	V
-ology -ist	N
-ment -al	A
-able -ity	N
-ist -ic	A
-ous -ity	N

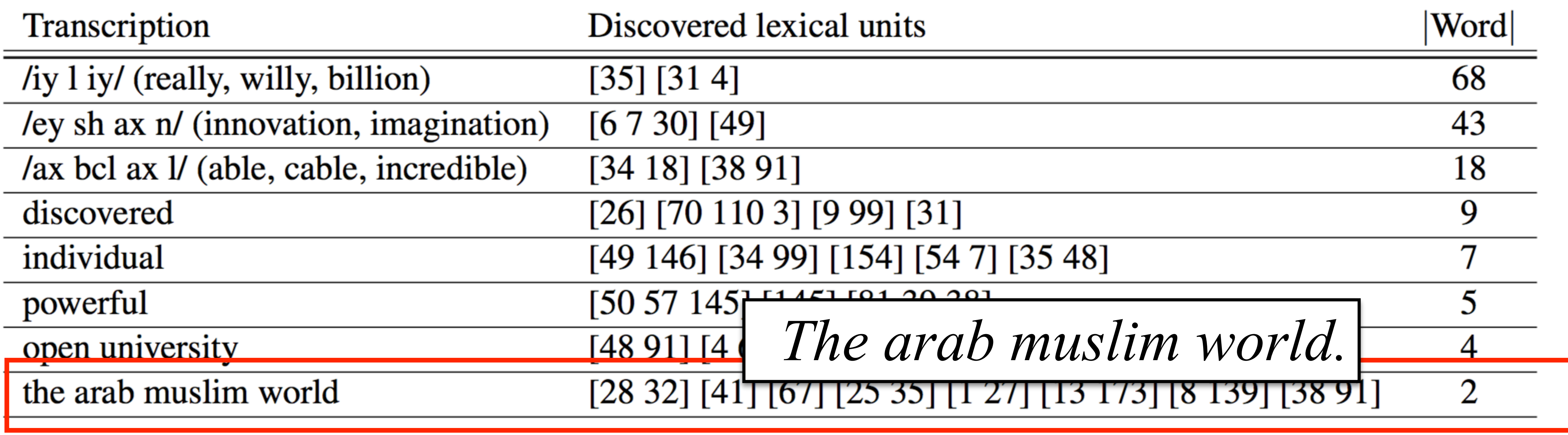
Lexical Units



Transcription	Discovered lexical units	|Word|
/iy l iy/ (really, willy, billion)	[35] [31 4]	68
/ey sh ax n/ (innovation, imagination)	[6 7 30] [49]	43
/ax bcl ax l/ (able, cable, incredible)	[34 18] [38 91]	18
discovered	[26] [70 110 3] [9 99] [31]	9
individual	[49 146] [34 99] [154] [54 7] [35 48]	7
powerful	[50 57 145] [145] [81 39 38]	5
open university	[48 91] [4 67] [25 8 99 29] [44 22] [103 4]	4
the arab muslim world	[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]	2

-able

Lexical Units



Transcription	Discovered lexical units	|Word|
/iy l iy/ (really, willy, billion)	[35] [31 4]	68
/ey sh ax n/ (innovation, imagination)	[6 7 30] [49]	43
/ax bcl ax l/ (able, cable, incredible)	[34 18] [38 91]	18
discovered	[26] [70 110 3] [9 99] [31]	9
individual	[49 146] [34 99] [154] [54 7] [35 48]	7
powerful	[50 57 145] [145] [81 20 28]	5
open university	[48 91] [4	4
the arab muslim world	[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]	2

The arab muslim world.

Key Scientific Results

- Lexical units.
- Synergistic Interactions.

Synergistic Interactions

Integrating multiple components often makes learning better.

of system mutually constrain one linguistic structures they learn are more accurate than if they were learned independently.

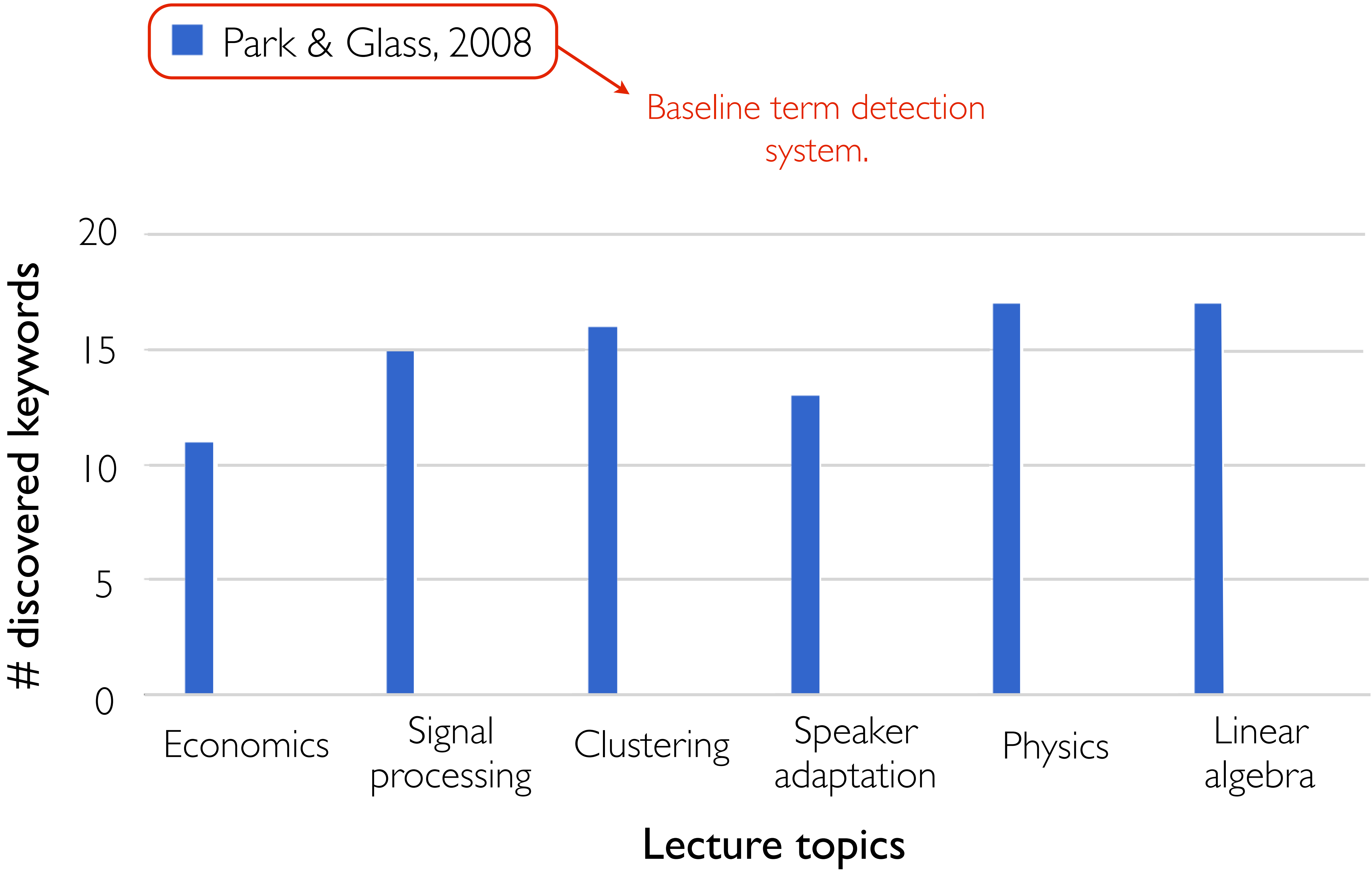
- Lexical units and Phones: Top-down influence of lexical unit learning improves phone unit learning.

- Underlying-to-Surface Mapping and Words: Modeling underlying-to-surface mapping improves ability to find correct words and sub-words.

Lesion Study

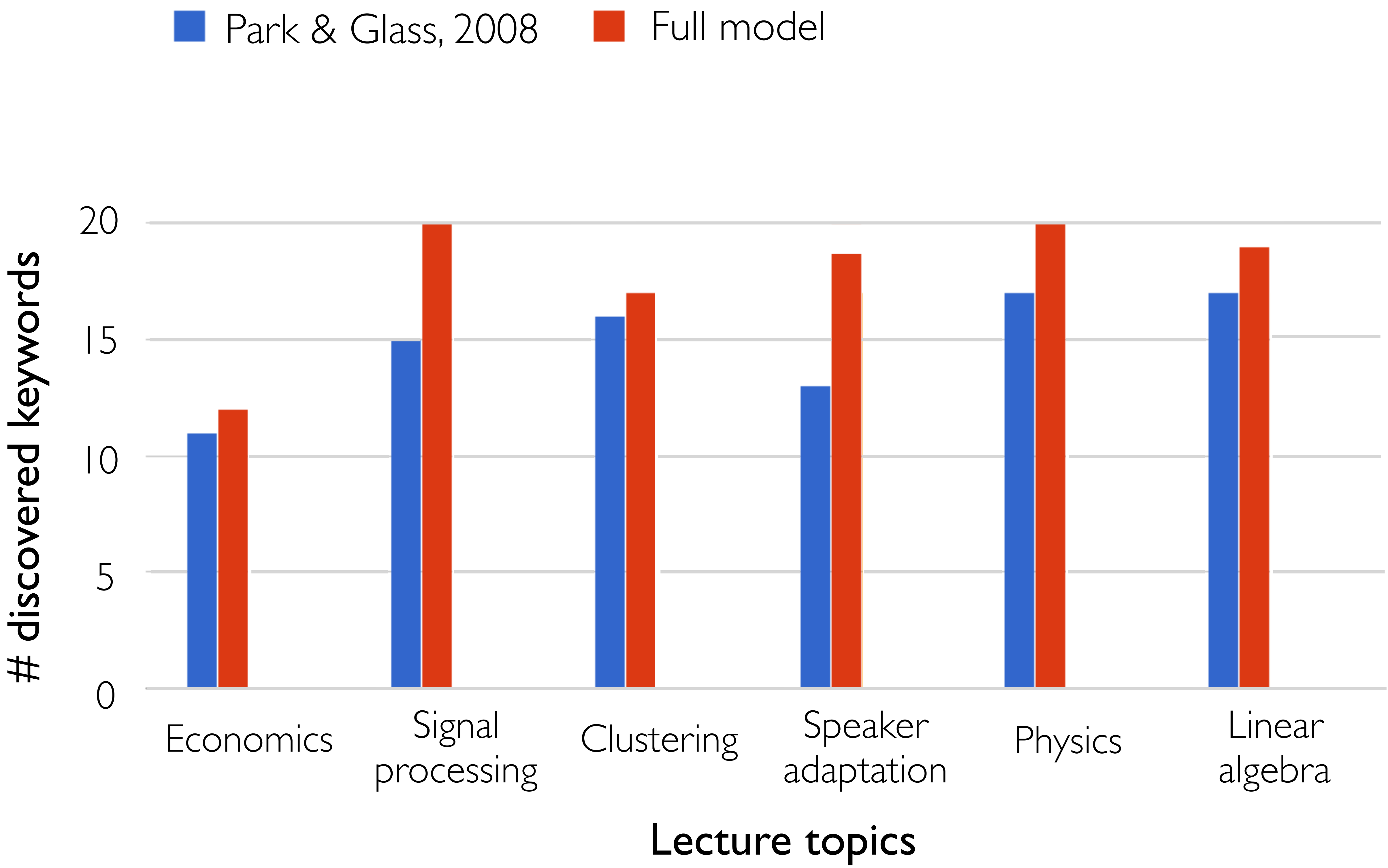
- **Lesioned** model components to study importance of difference components.
  - Examined top 20 most important words in each lecture (term frequency inverse-document frequency)

Baseline

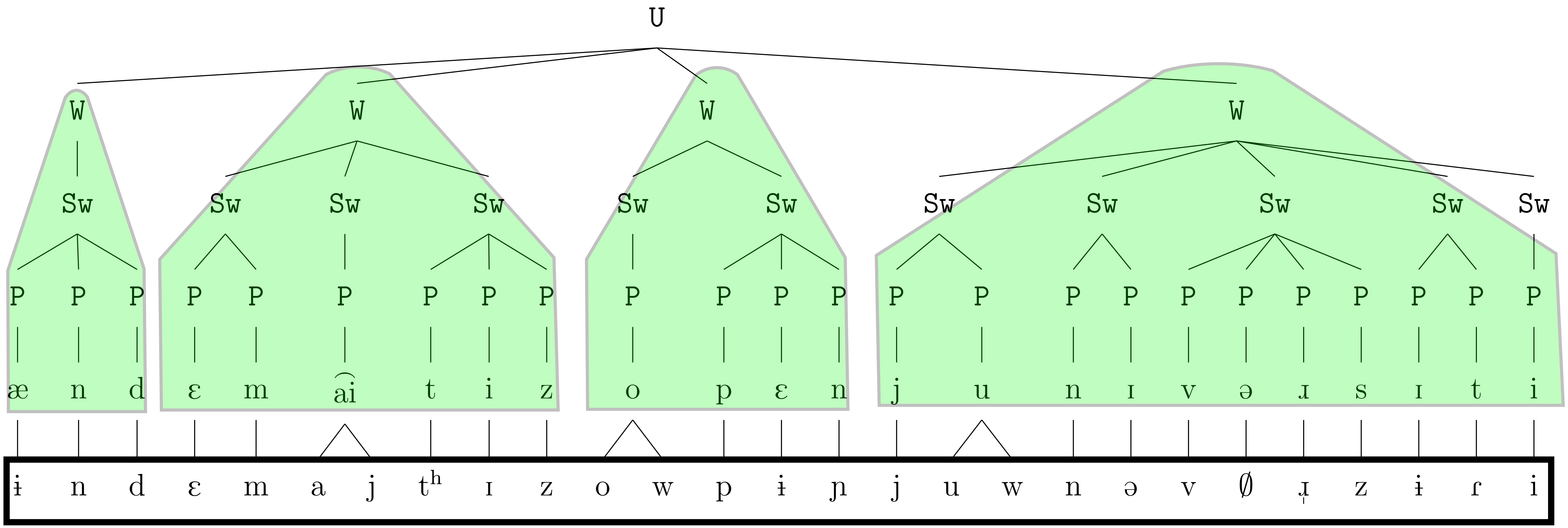


Baseline term detection system.

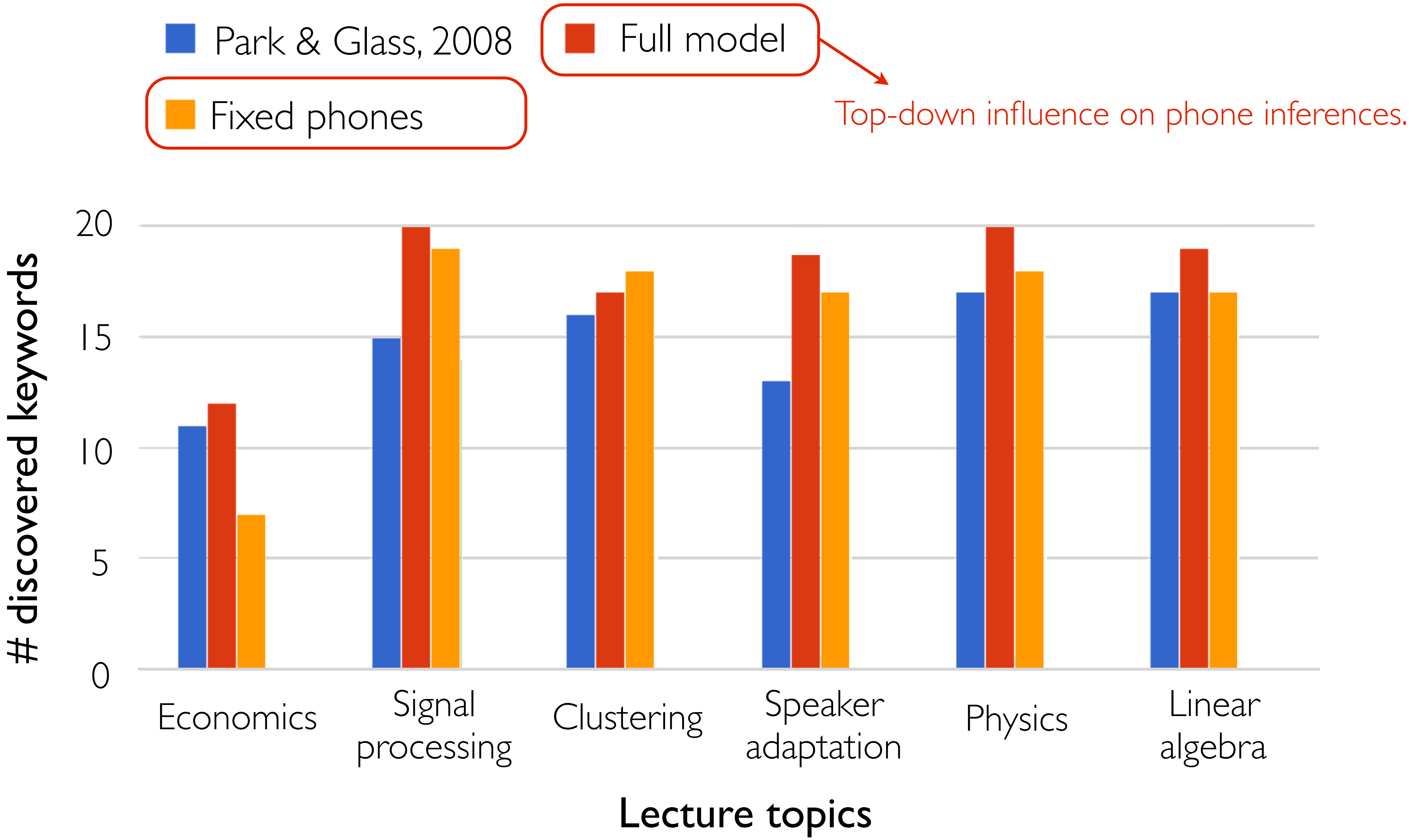
Full Model


Park & Glass, 2008
Full model
20
15
10
5
0
# discovered keywords
Economics
Signal processing
Clustering
Speaker adaptation
Physics
Linear algebra
Lecture topics

Fixed Phones

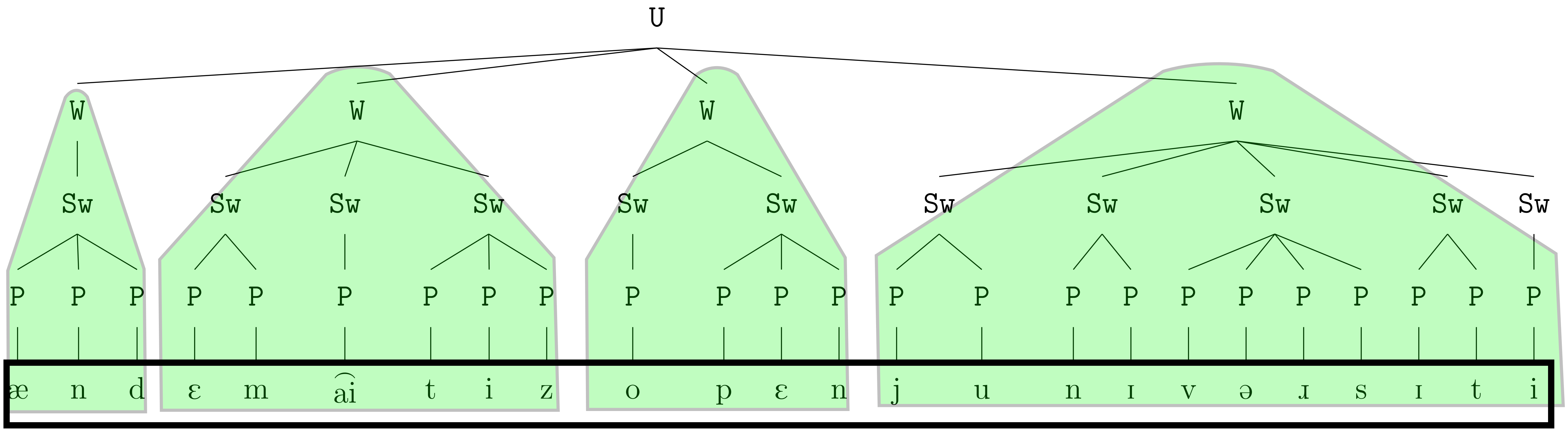

U
W W W W
Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw Sw
P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i
ɨ n d ɛ m a j tʰ ɪ z o w p ɨ ɲ j u w n ə v ∅ ɹ̩ z ɨ ɾ i

Joint Model

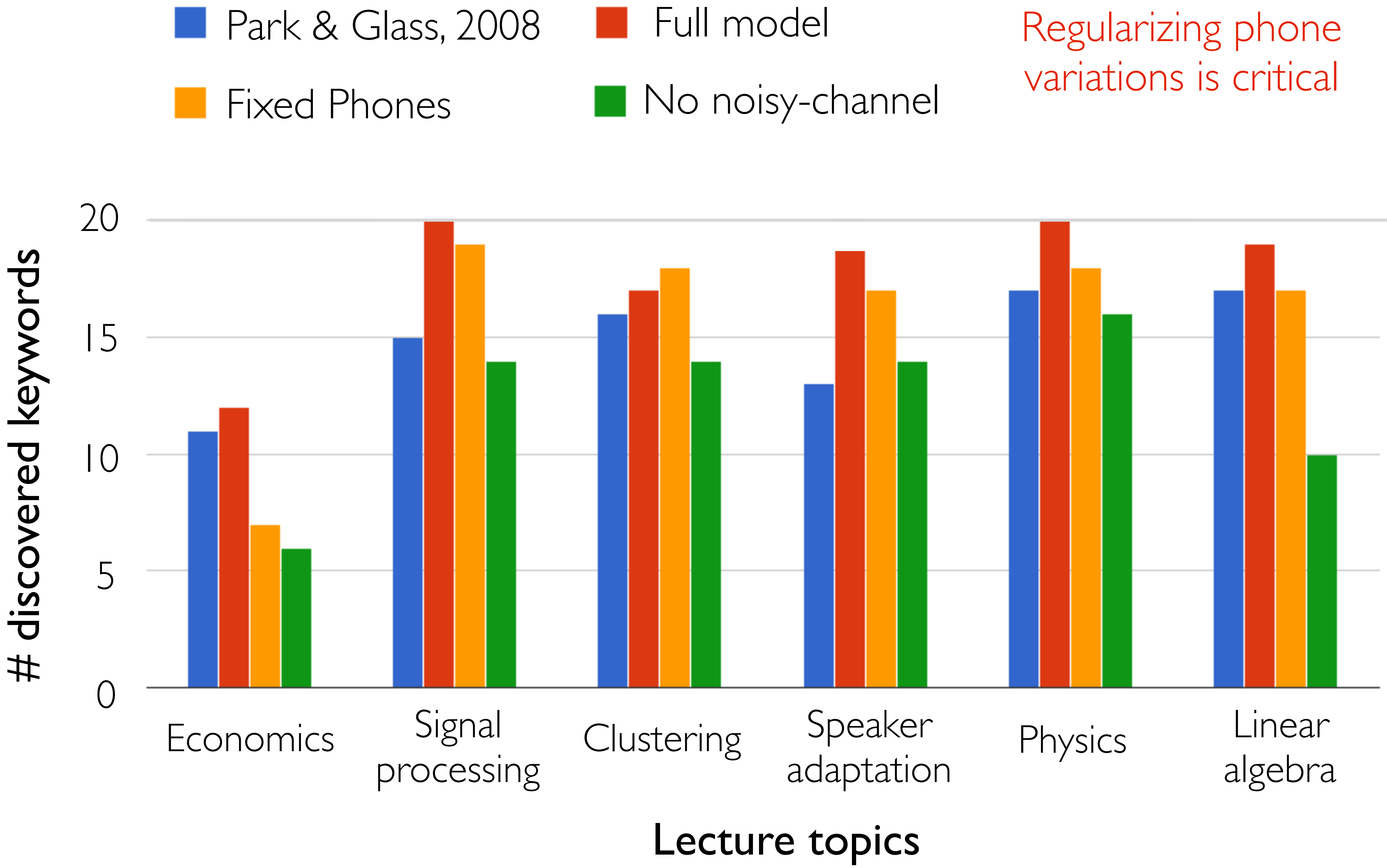


Top-down influence on phone inferences.

No Noisy Channel


U
W
W
W
W
Sw
Sw
Sw
Sw
Sw
Sw
Sw
Sw
Sw
Sw
Sw
P P P P P P P P P P P P P P P P P P P P P P P P P
æ n d ɛ m a͡i t i z o p ɛ n j u n ɪ v ə ɹ s ɪ t i

No Noisy-Channel



Regularizing phone variations is critical

Overall Conclusion

- Can use tradeoff-based approach to learn lexical units.
- Makes fine-grained predictions for linguistics and psychology.
- Can scale up to more unsupervised settings.

Thanks