Inferring unobserved co-occurrence events in Anchored Packed Trees

Thomas Kober
tkober@inf.ed.ac.uk

based on joint work with Julie Weeds, Jeremy Reffin and David Weir

13th February 2018 (1518526800)

Who am I?

Who am I?

- Started as a post-doc in January

Who am I?

- Started as a post-doc in January
- Finished my PhD at the University of Sussex (had my viva just over a week ago)

Who am I?

- Started as a post-doc in January
- Finished my PhD at the University of Sussex (had my viva just over a week ago)

Who am I?

- Started as a post-doc in January
- Finished my PhD at the University of Sussex (had my viva just over a week ago)

Who am I?

- Started as a post-doc in January
- Finished my PhD at the University of Sussex (had my viva just over a week ago)

Outline

Outline

- Introduction to Anchored Packed Trees

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

Outline

- **Introduction to Anchored Packed Trees**
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

What are APTs?

What are APTs?

- Compositional Distributional Semantic model

What are APTs?

- Compositional Distributional Semantic model

    - Semantics: The study of the meaning of words and phrases in a language

What are APTs?

- Compositional Distributional Semantic model

    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language

What are APTs?

- Compositional Distributional Semantic model
    - Compositional: Based on the product of combining elementary word representations
    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language

What are APTs?

- Compositional Distributional Semantic model
    - Compositional: Based on the product of combining elementary word representations
    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language
- Motivation

What are APTs?

- Compositional Distributional Semantic model
    - Compositional: Based on the product of combining elementary word representations
    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language
- Motivation
    - Open question what composition in distributional semantics means

What are APTs?

- Compositional Distributional Semantic model
    - Compositional: Based on the product of combining elementary word representations
    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language
- Motivation
    - Open question what composition in distributional semantics means
    - Existing models use linear algebraic operations in a vector space populated by words to mash together word representations

What are APTs?

- Compositional Distributional Semantic model
    - Compositional: Based on the product of combining elementary word representations
    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language

- Motivation
    - Open question what composition in distributional semantics means
    - Existing models use linear algebraic operations in a vector space populated by words to mash together word representations
    - Several shortcomings, e.g. commutativity for simple composition functions (e.g. point wise addition); or reliance on task specific training data for complex neural network based models

What are APTs?

- Compositional Distributional Semantic model
    - Compositional: Based on the product of combining elementary word representations
    - Distributional: Based on the co-occurrence statistics of words in a corpus
    - Semantics: The study of the meaning of words and phrases in a language
- Motivation
    - Open question what composition in distributional semantics means
    - Existing models use linear algebraic operations in a vector space populated by words to mash together word representations
    - Several shortcomings, e.g. commutativity for simple composition functions (e.g. point wise addition); or reliance on task specific training data for complex neural network based models
- APTs treating distributional composition as a process of contextualisation (Weir et al., 2016)

What are APTs?

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus
- Modelling forward dependencies:

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:

amod

white *clothes* **clothes:** `amod`:*white*

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:

  `amod`

  white *clothes* **clothes:** `amod`:*white*

- Inverse dependencies:

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:

  $\textit{white} \xleftarrow{\texttt{amod}} \textit{clothes}$

  clothes: `amod`:*white*

- Inverse dependencies:

  $\textit{white} \xleftarrow{\texttt{amod}} \textit{clothes}$

  white: $\overline{\texttt{amod}}$:*clothes*

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:

$\textit{white} \xleftarrow{\texttt{amod}} \textit{clothes}$

clothes: `amod`:*white*

- Inverse dependencies:

$\textit{white} \xleftarrow{\texttt{amod}} \textit{clothes}$

white: $\overline{\texttt{amod}}$:*clothes*

Inverse `amod`

What are APTs?

- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:

  amod

  white *clothes*

  clothes: amod:*white*

- Inverse dependencies:

  amod

  white *clothes*

  white: $\overline{\texttt{amod}}$:*clothes*

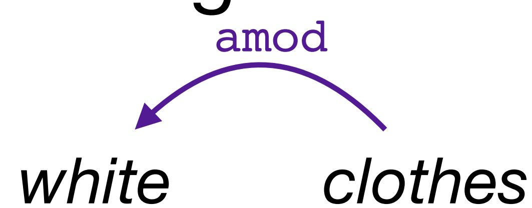
  Inverse amod

- Higher-order dependencies:

What are APTs?

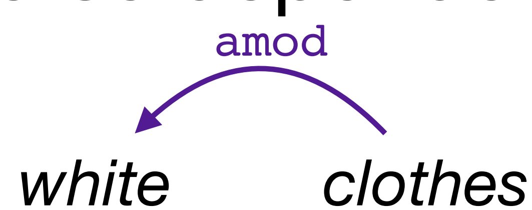
- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:



clothes: `amod`:*white*

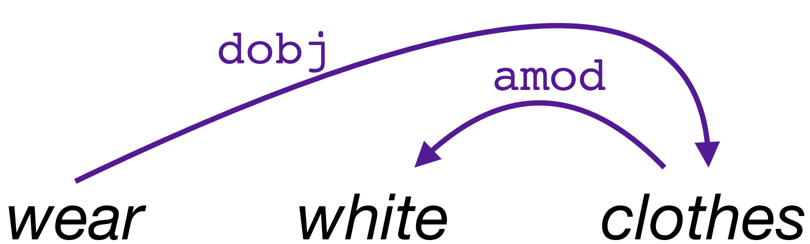
- Inverse dependencies:



Inverse `amod`

white: $\overline{\texttt{amod}}$:*clothes*

- Higher-order dependencies:

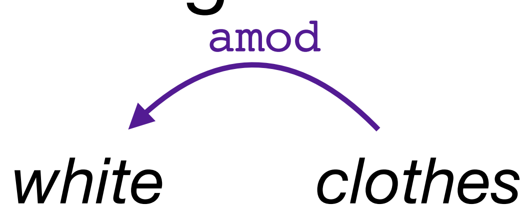


white: $\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:*wear*

What are APTs?

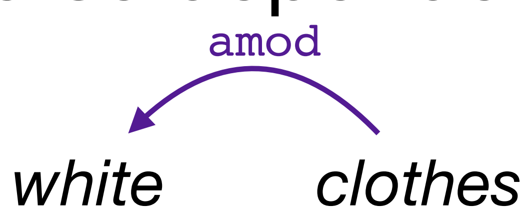
- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:



clothes: amod:*white*

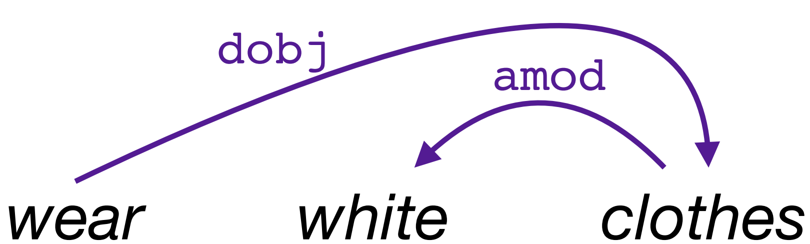
- Inverse dependencies:



Inverse amod

white: $\overline{\texttt{amod}}$:*clothes*

- Higher-order dependencies:



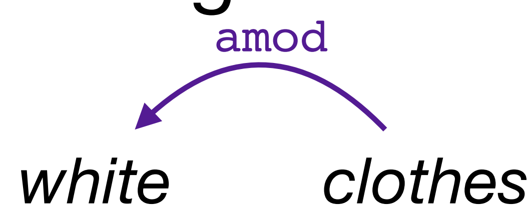
Higher order path

white: $\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:*wear*

What are APTs?

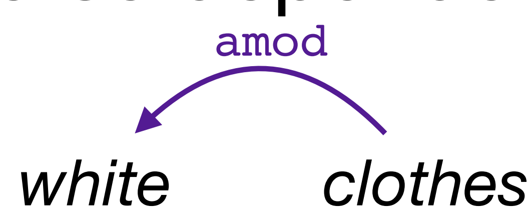
- Representations for individual lexemes are built from a dependency parsed corpus

- Modelling forward dependencies:



clothes: `amod`:*white*

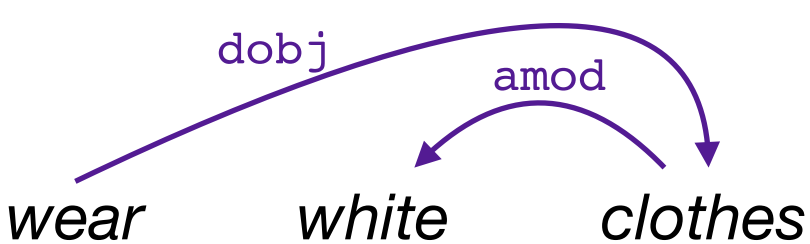
- Inverse dependencies:



Inverse `amod`

white: $\overline{\texttt{amod}}$:*clothes*

- Higher-order dependencies:



Higher order path

"**white** *things* can be *worn*"

white: $\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:*wear*

What are APTs?

What are APTs?

we folded the dry clean clothes

What are APTs?

we folded the dry clean clothes

we bought white shoes yesterday

What are APTs?

we folded the dry clean clothes

we bought white shoes yesterday

i like your clothes

What are APTs?

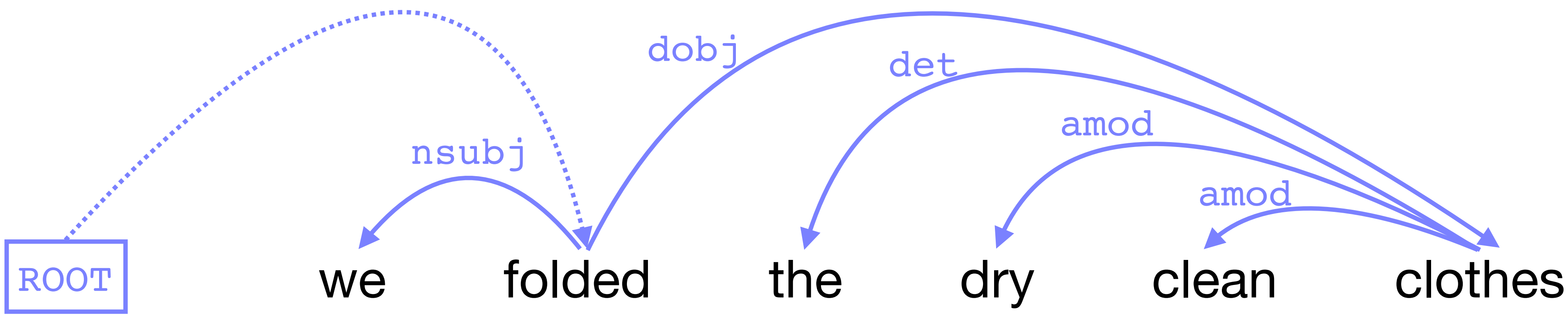
we folded the dry clean clothes

we bought white shoes yesterday

i like your clothes

he folded the white sheets

What are APTs?

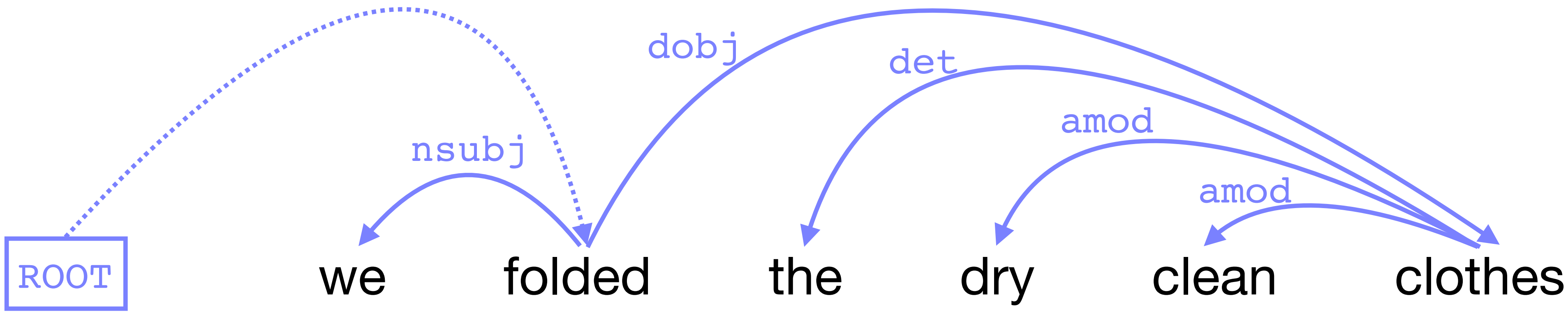


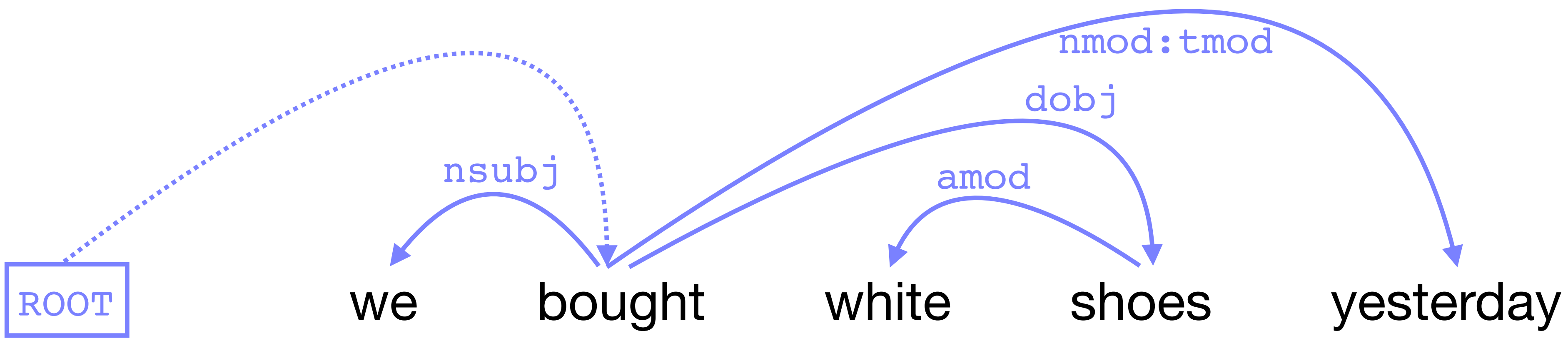
we bought white shoes yesterday

i like your clothes

he folded the white sheets

What are APTs?





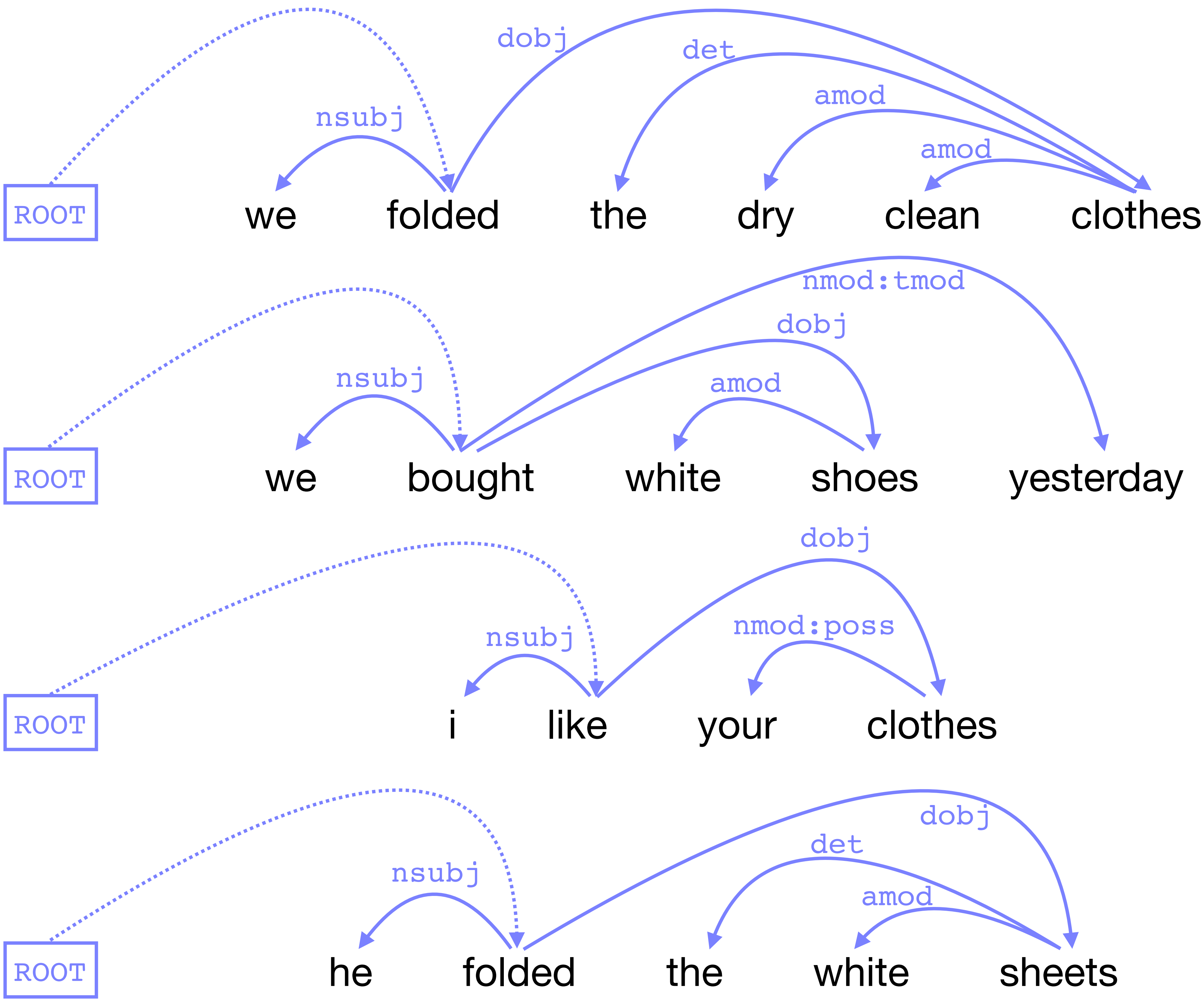
i like your clothes

he folded the white sheets

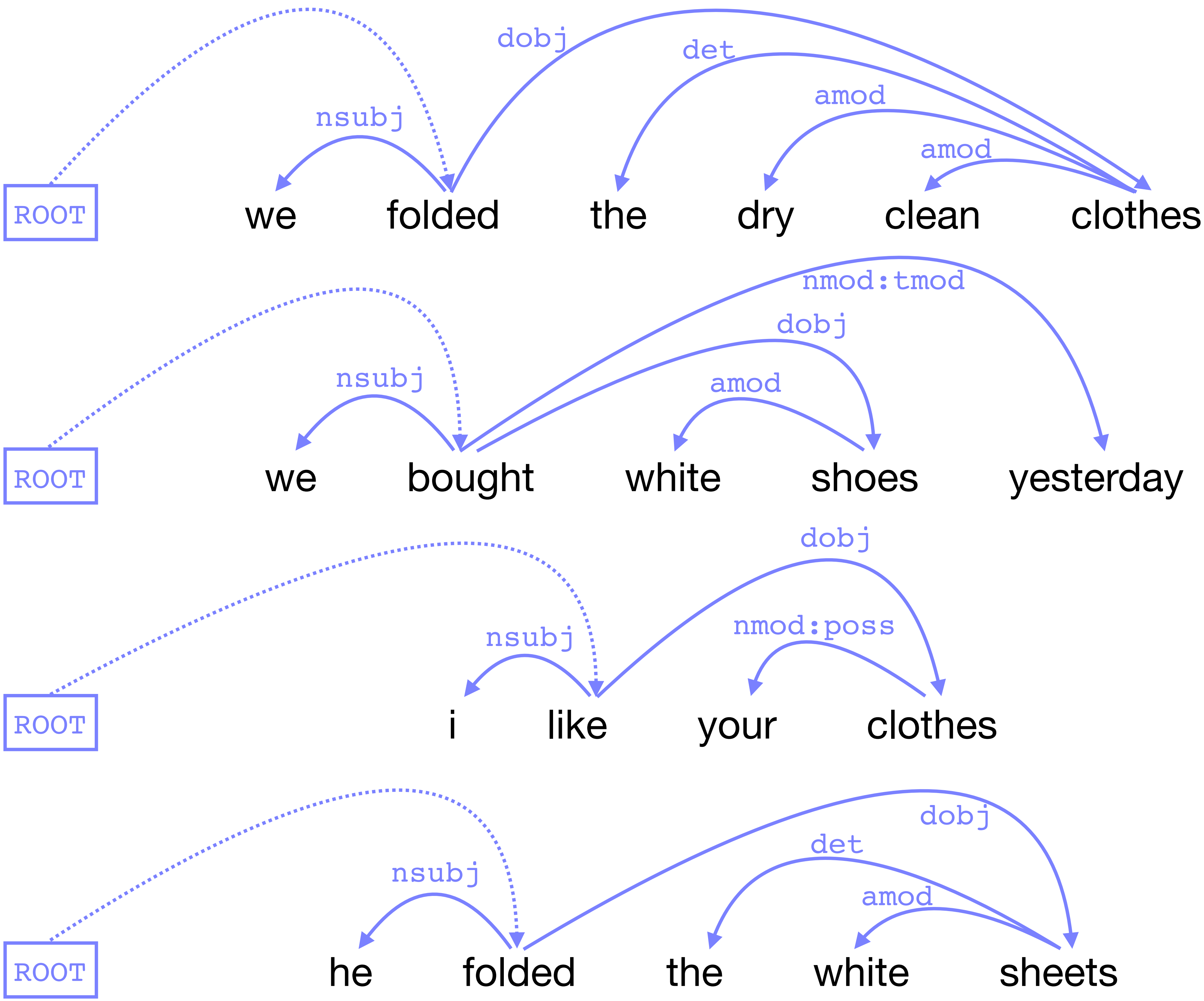
What are APTs?

he folded the white sheets

What are APTs?

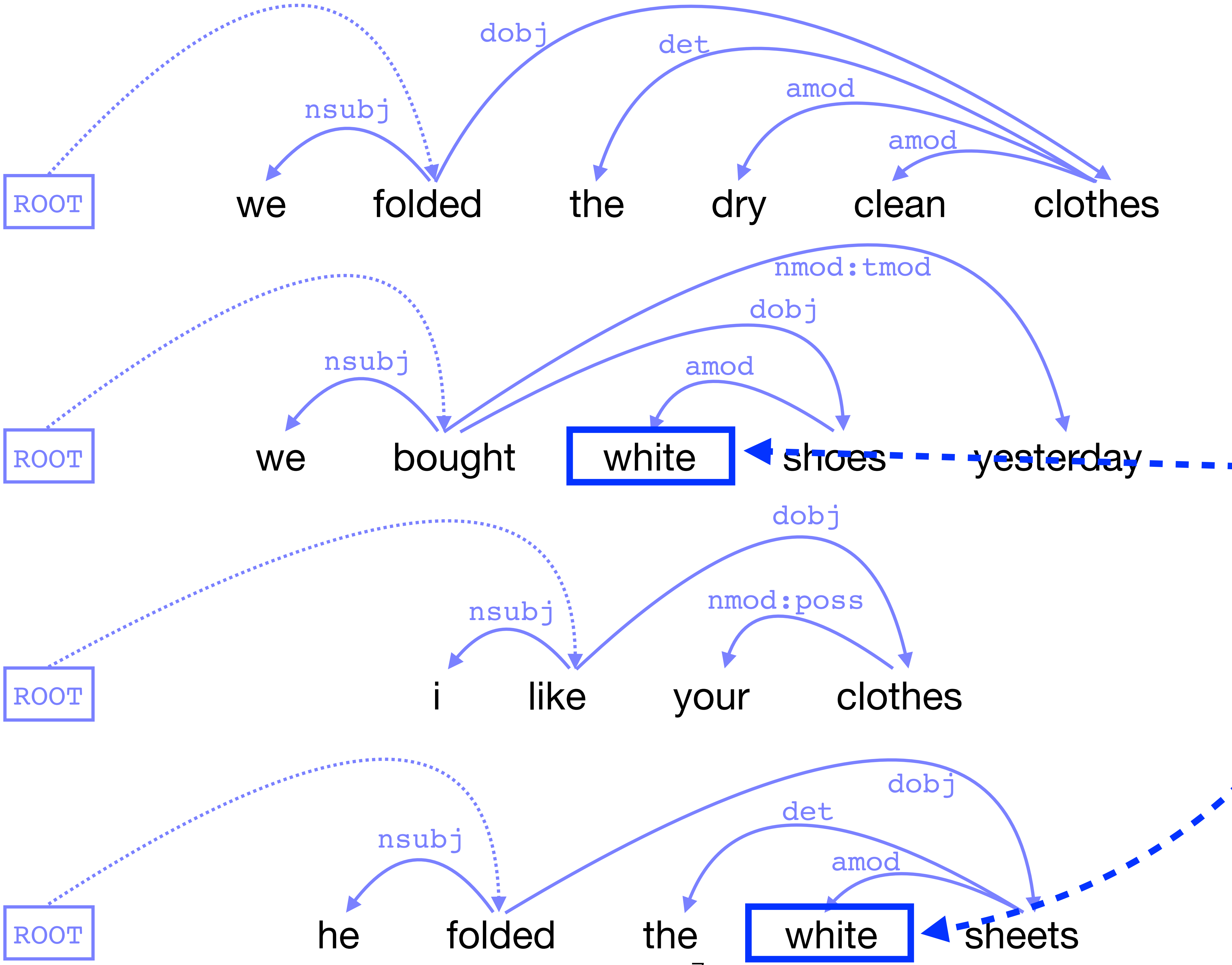


What are APTs?



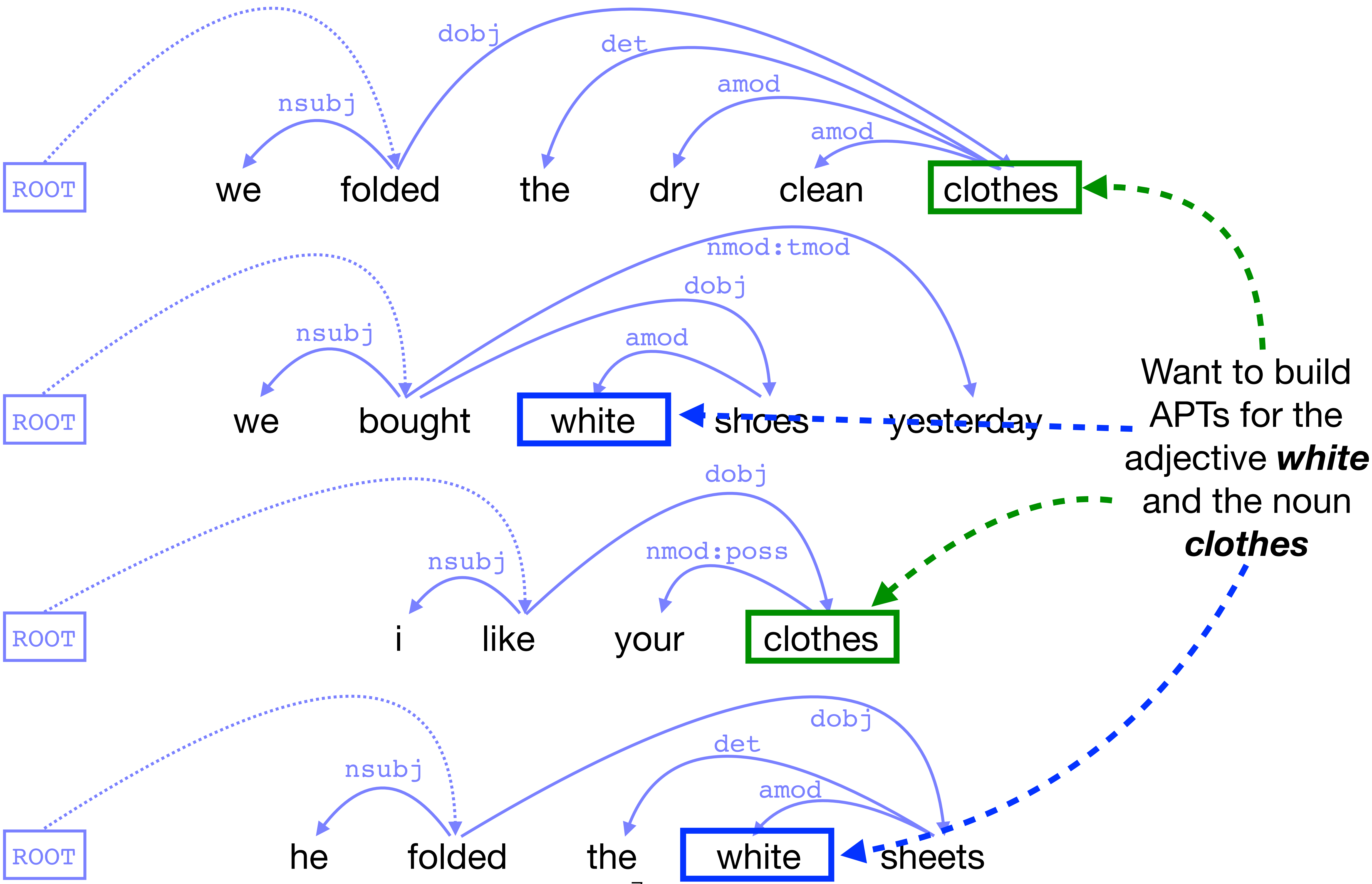
Want to build APTs for the adjective ***white*** and the noun ***clothes***

What are APTs?



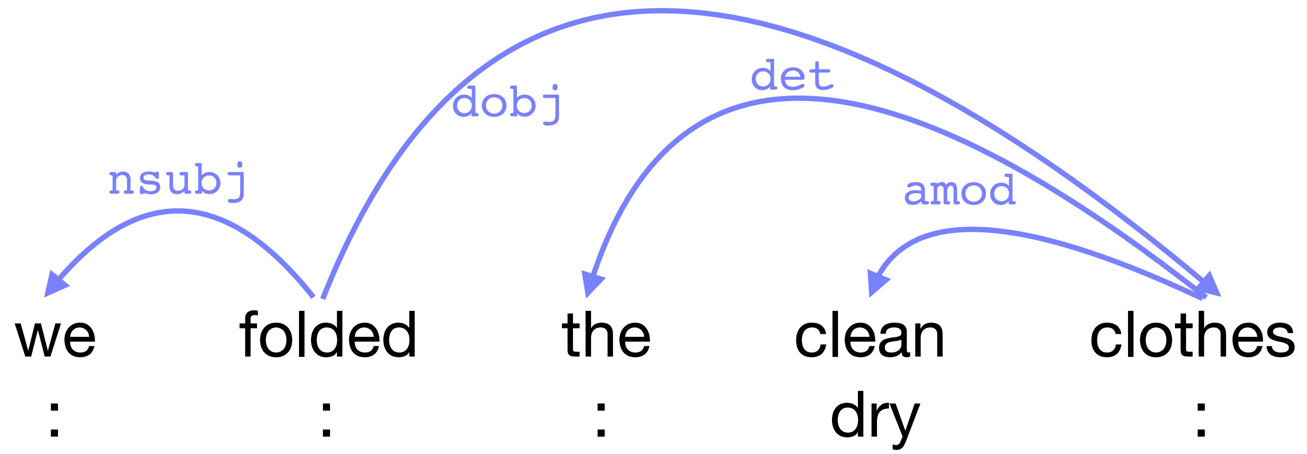
Want to build APTs for the adjective ***white*** and the noun ***clothes***

What are APTs?

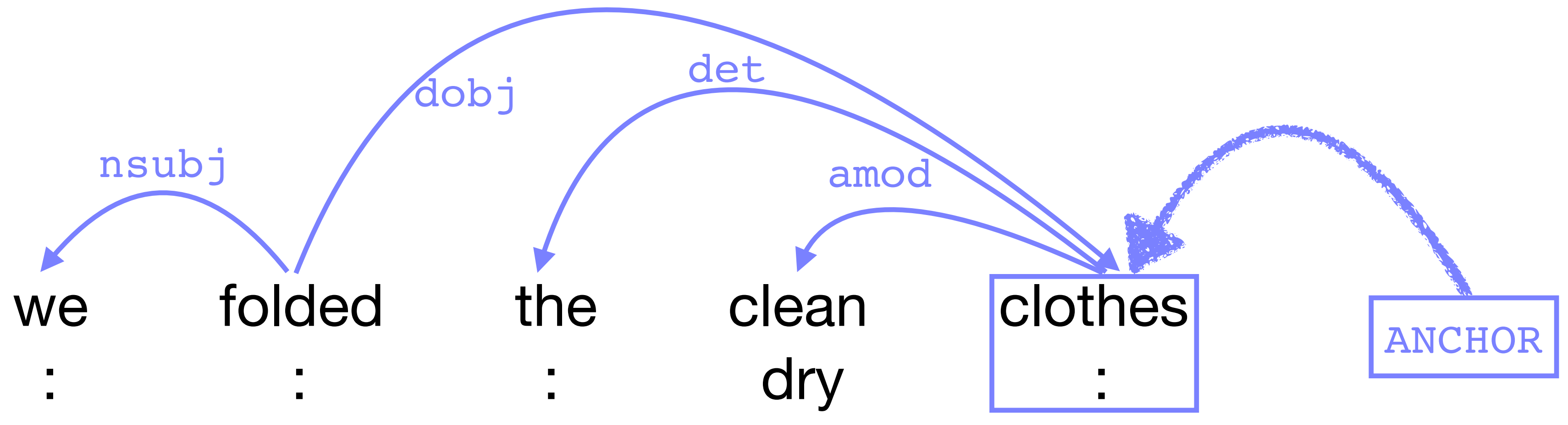


What are APTs?

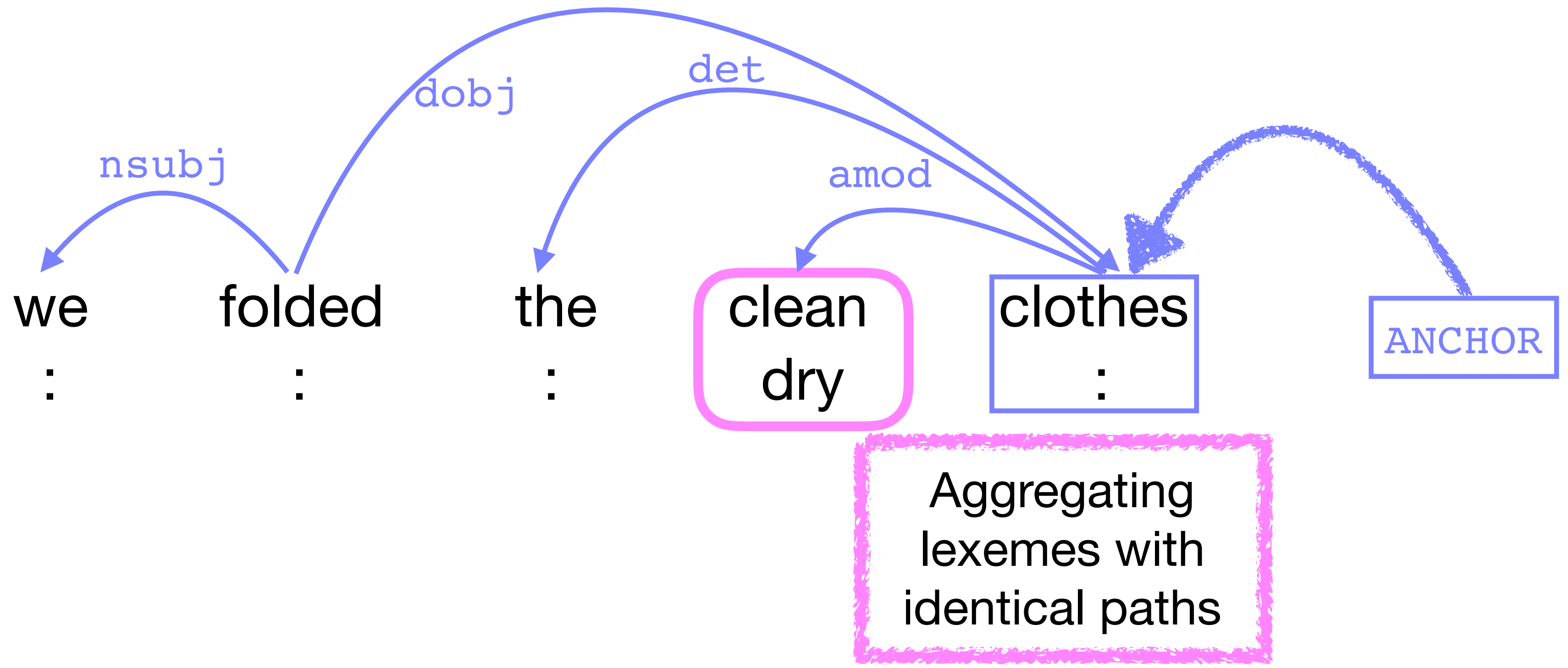
What are APTs?



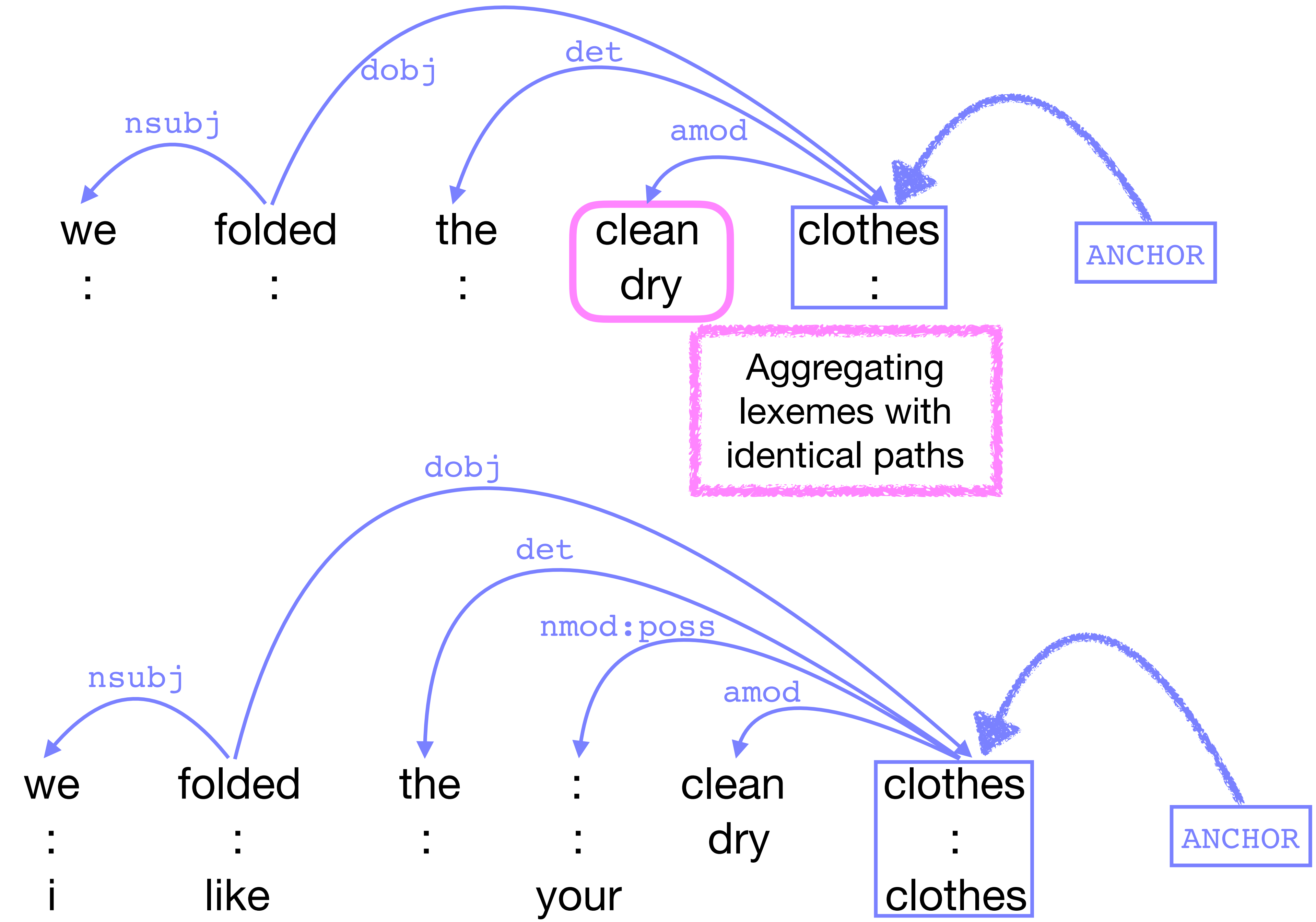
What are APTs?



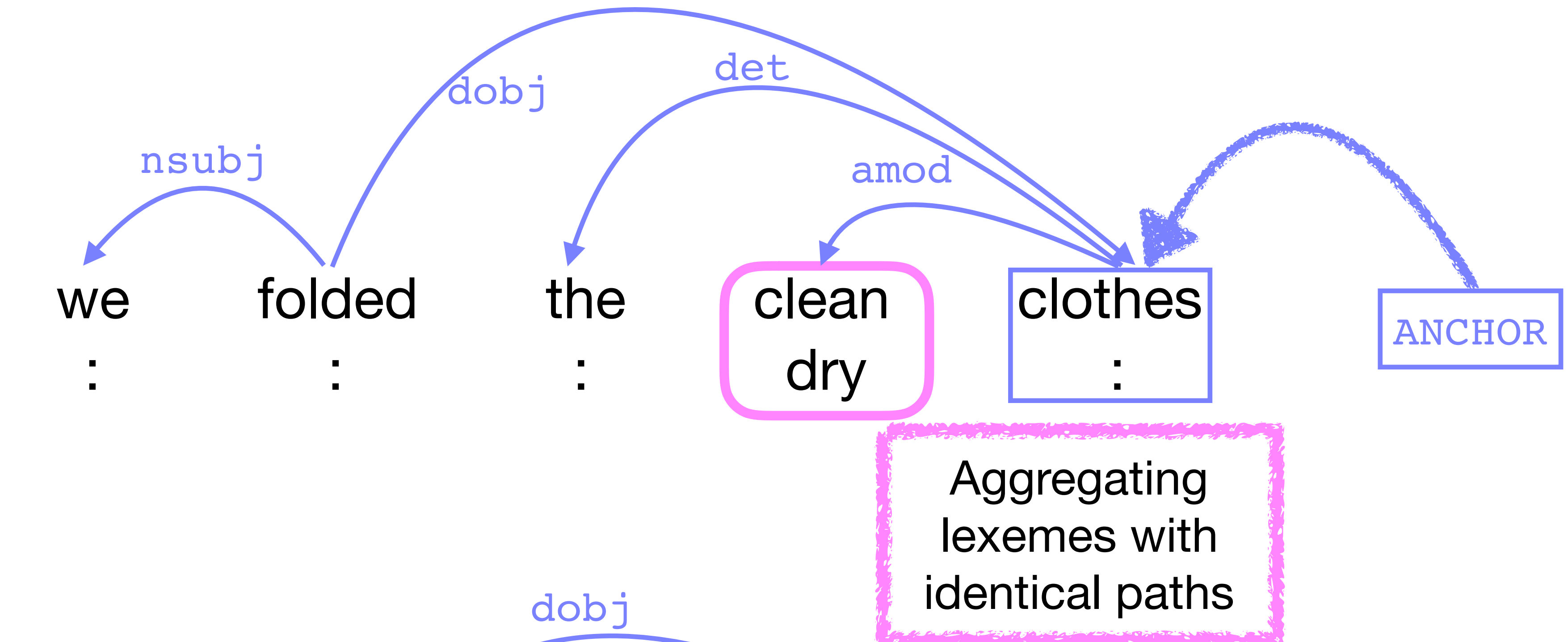
What are APTs?

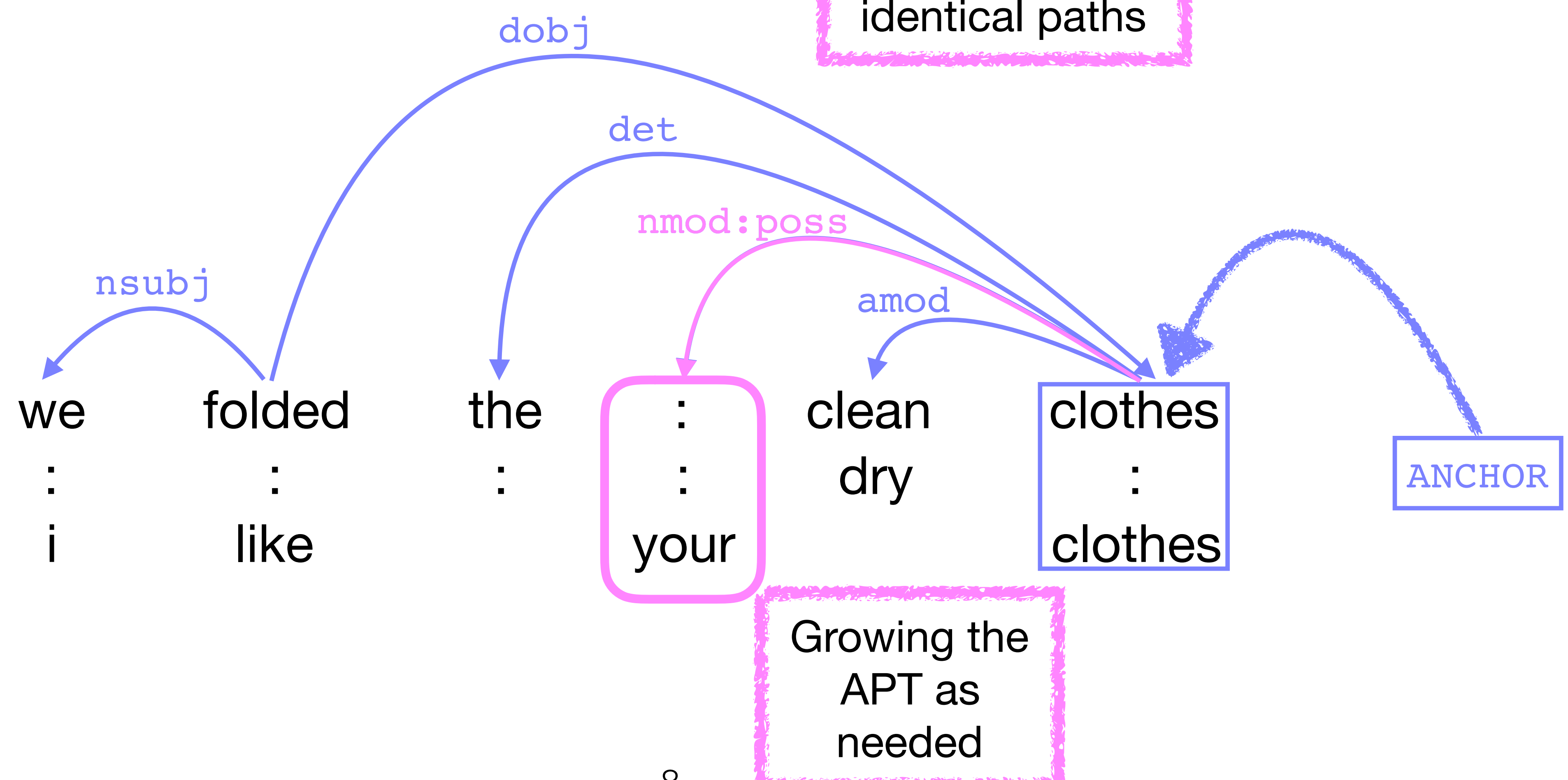


What are APTs?



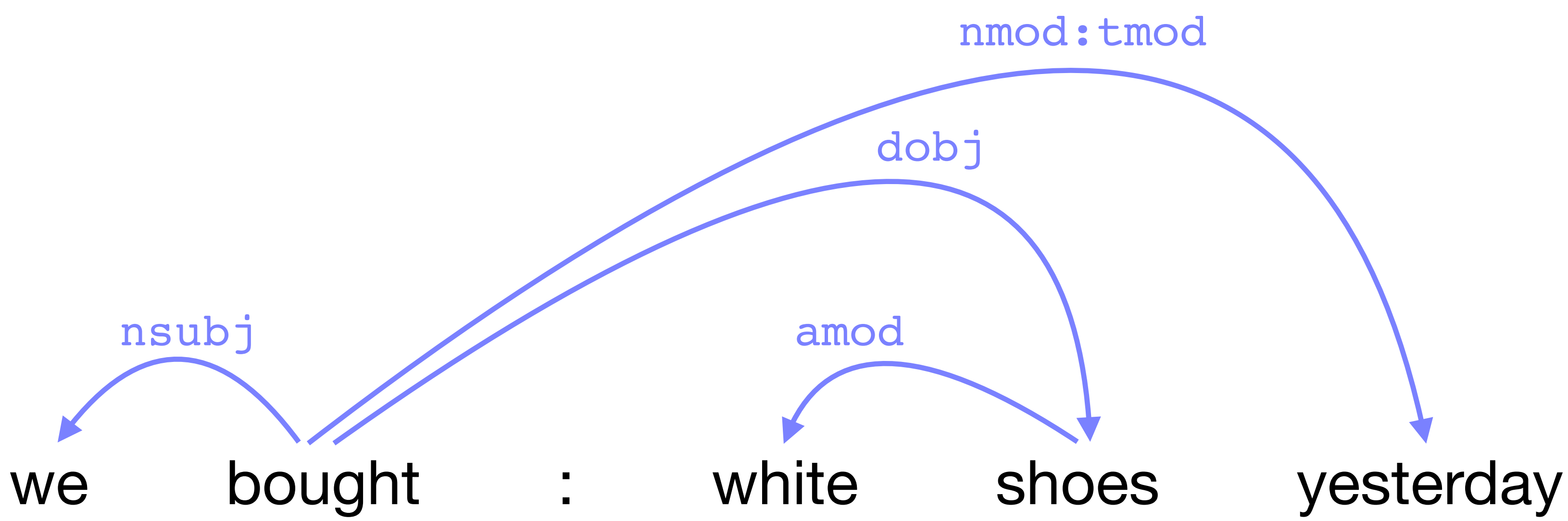
What are APTs?



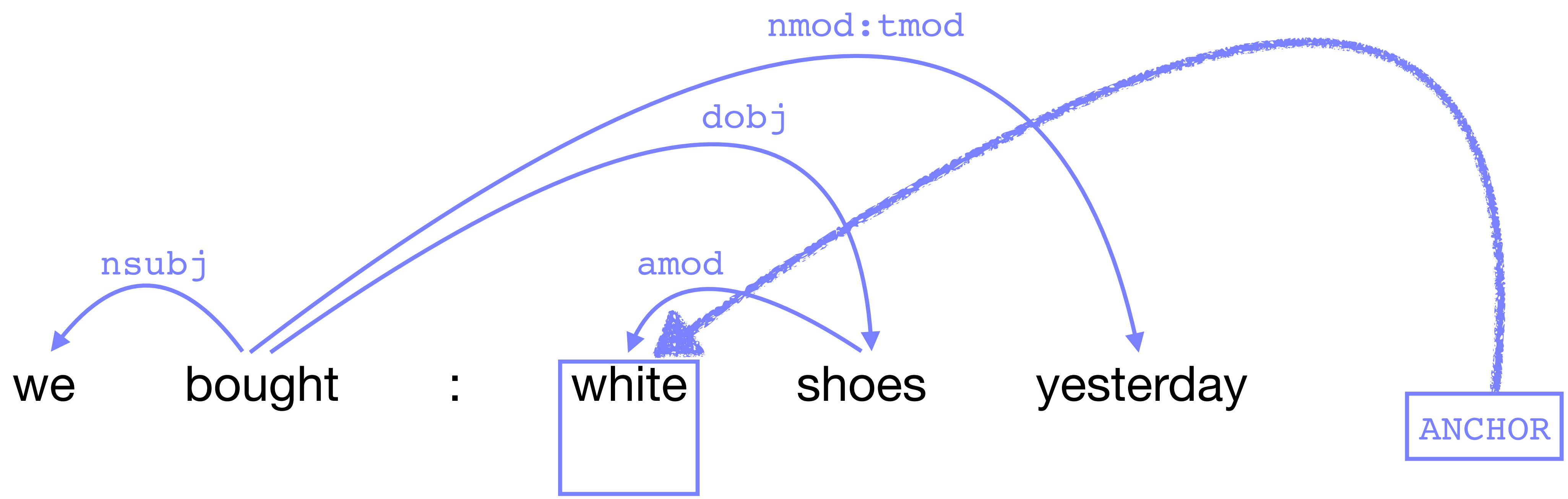


What are APTs?

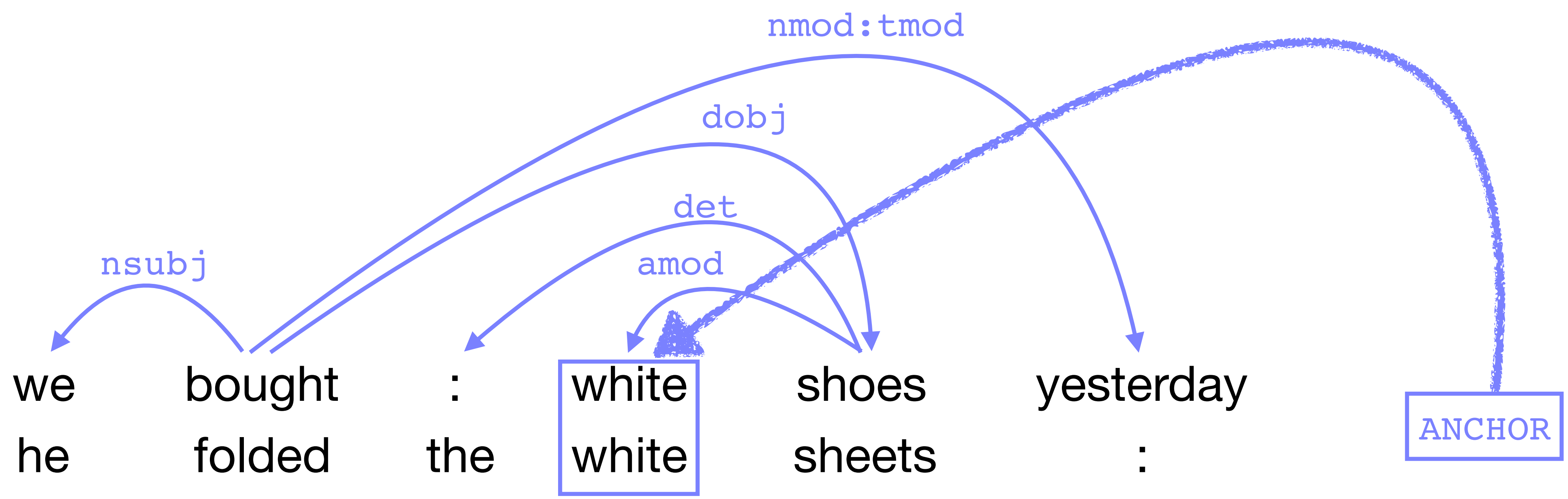
What are APTs?



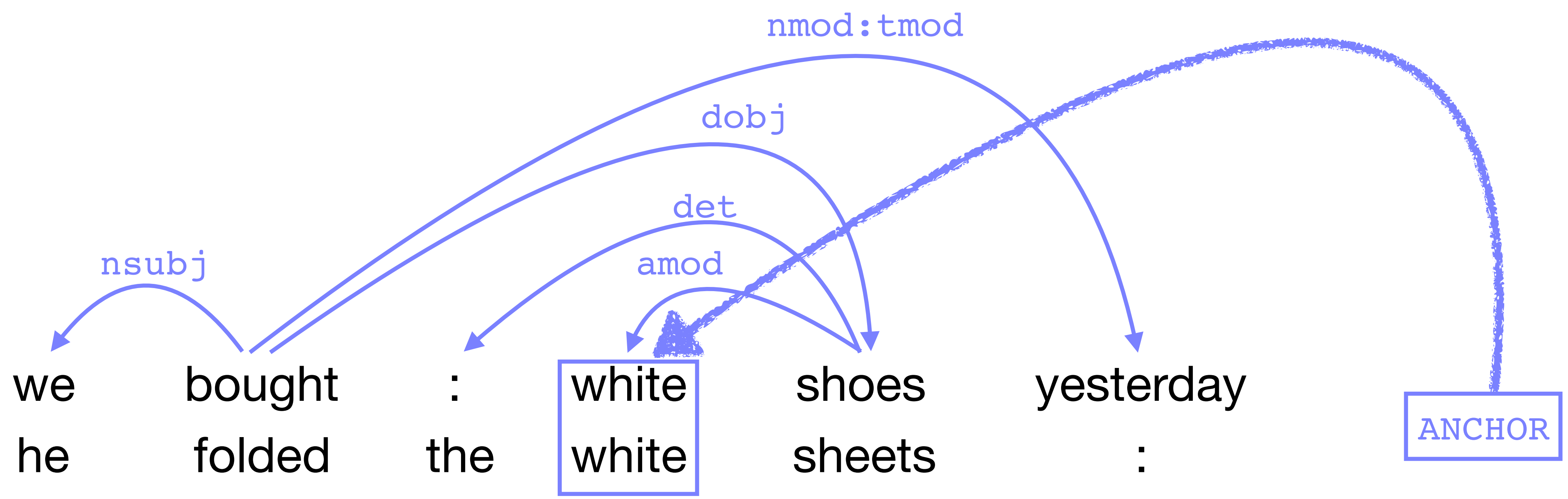
What are APTs?



What are APTs?

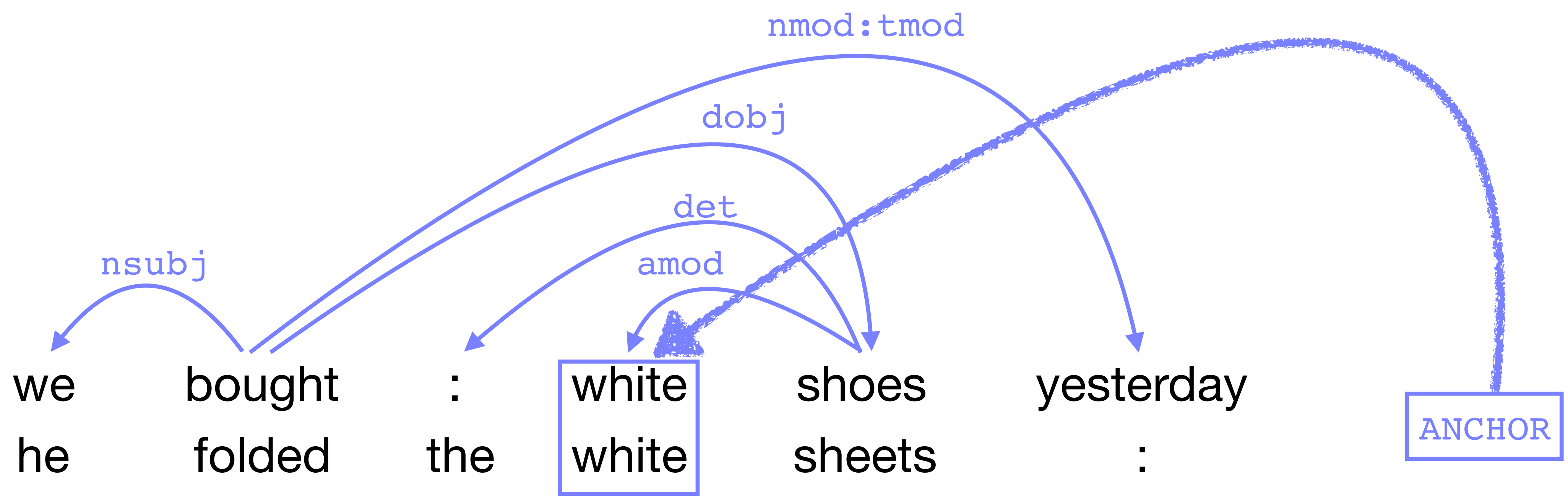


What are APTs?



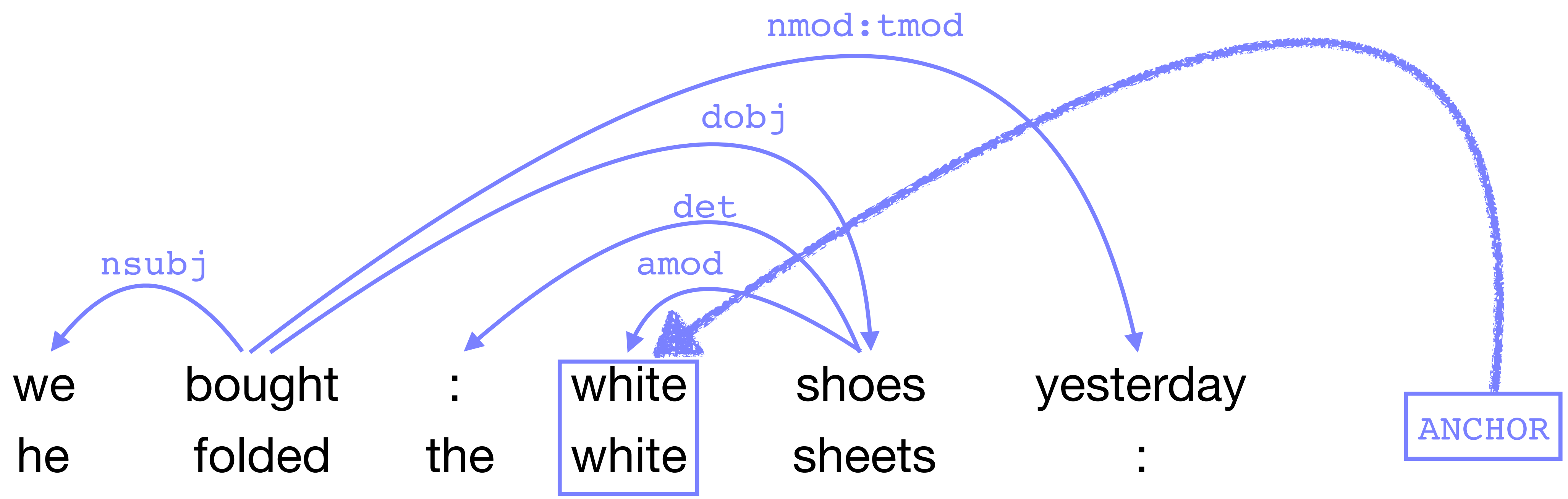
- Anchor is placed at every lexeme in a sentence during processing

What are APTs?



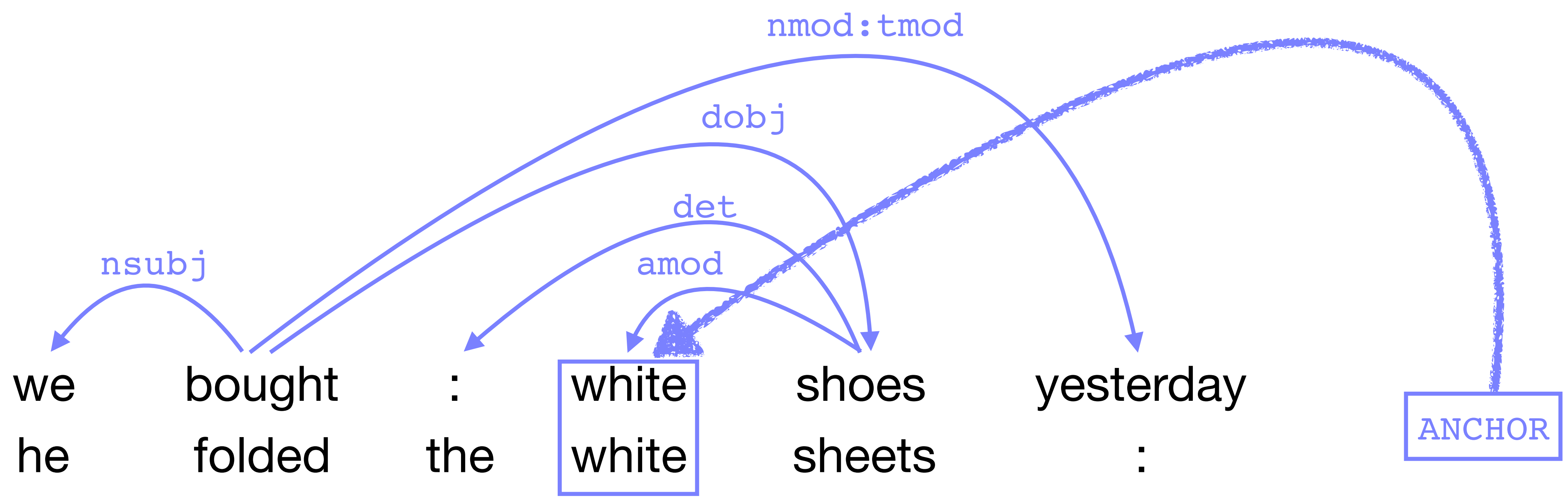
- Anchor is placed at every lexeme in a sentence during processing
- One APT per lexeme

What are APTs?



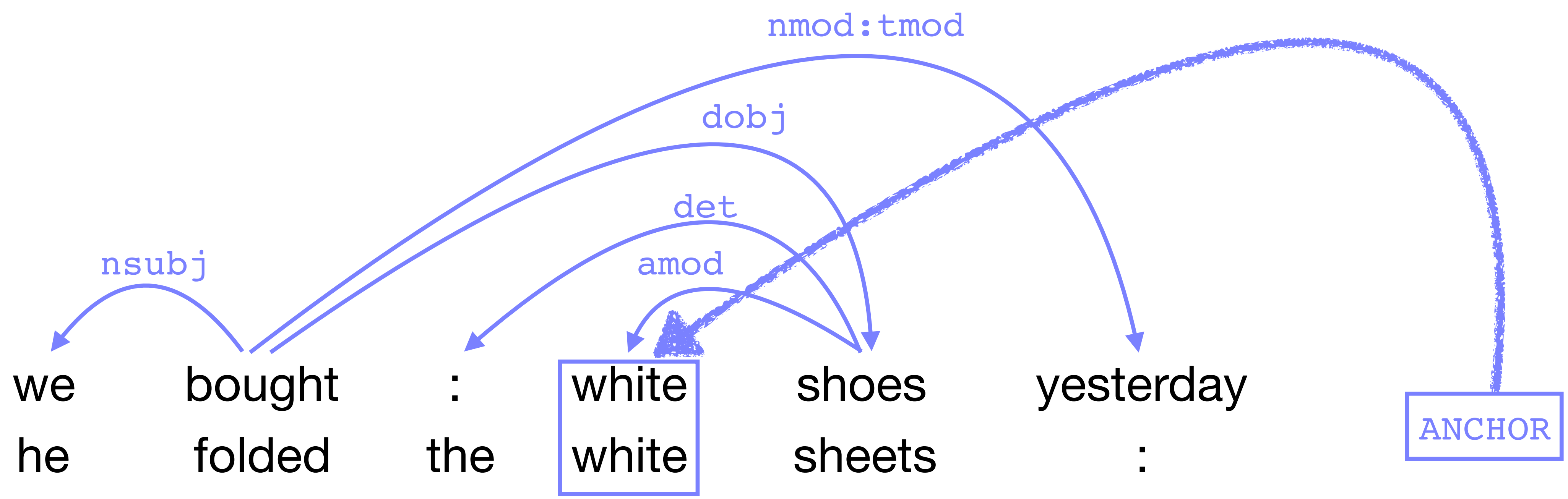
- Anchor is placed at every lexeme in a sentence during processing
- One APT per lexeme
- APTs are not a vector space *per se*, but define a graph

What are APTs?



- Anchor is placed at every lexeme in a sentence during processing
- One APT per lexeme
- APTs are not a vector space *per se*, but define a graph
  - Vertices contain lexemes

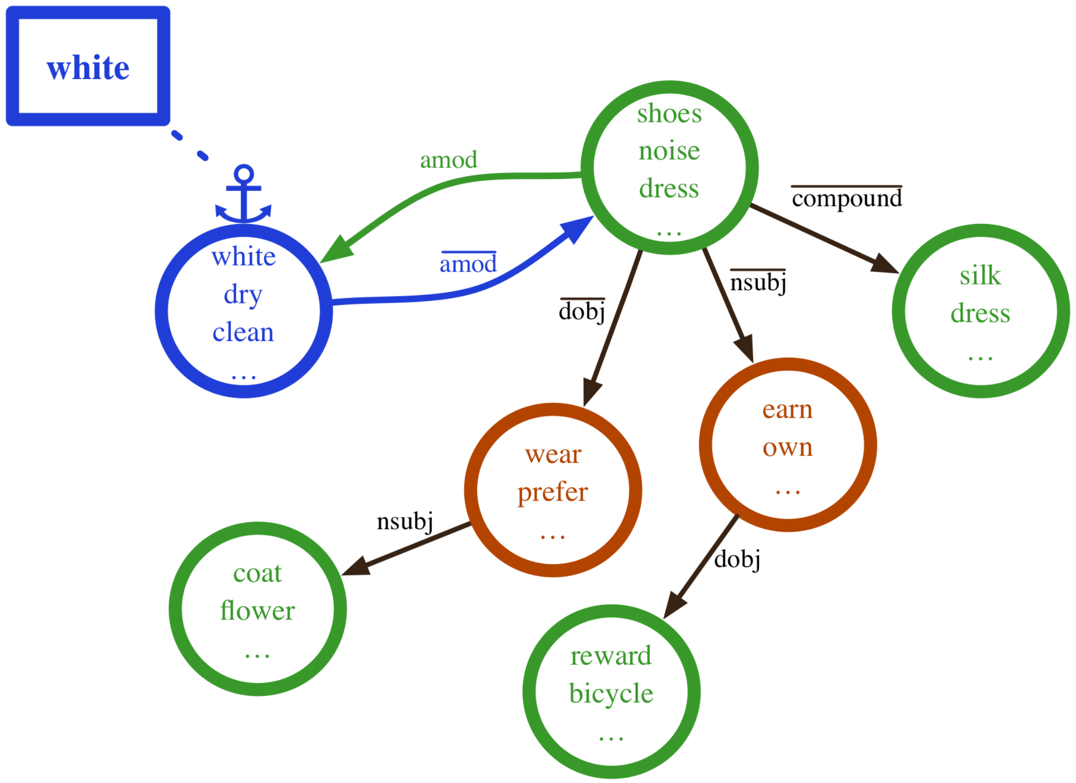
What are APTs?



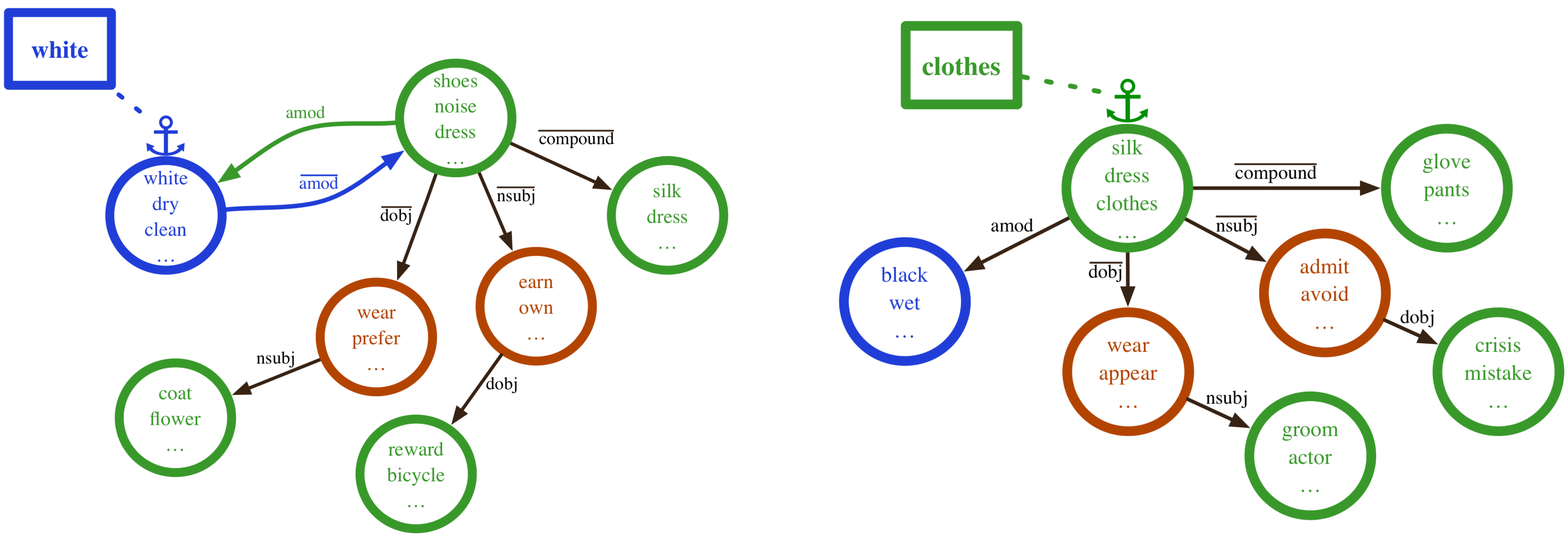
- Anchor is placed at every lexeme in a sentence during processing
- One APT per lexeme
- APTs are not a vector space *per se*, but define a graph
    - Vertices contain lexemes
    - Edges are dependency relations

What are APTs?

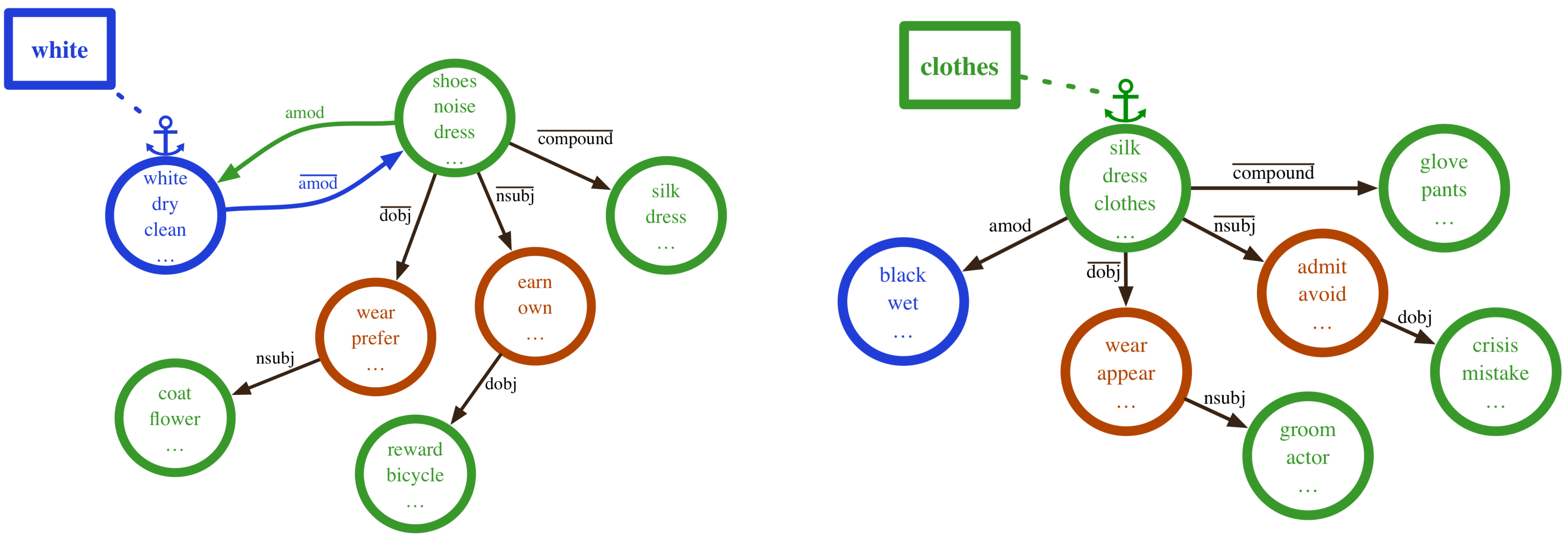
What are APTs?



What are APTs?

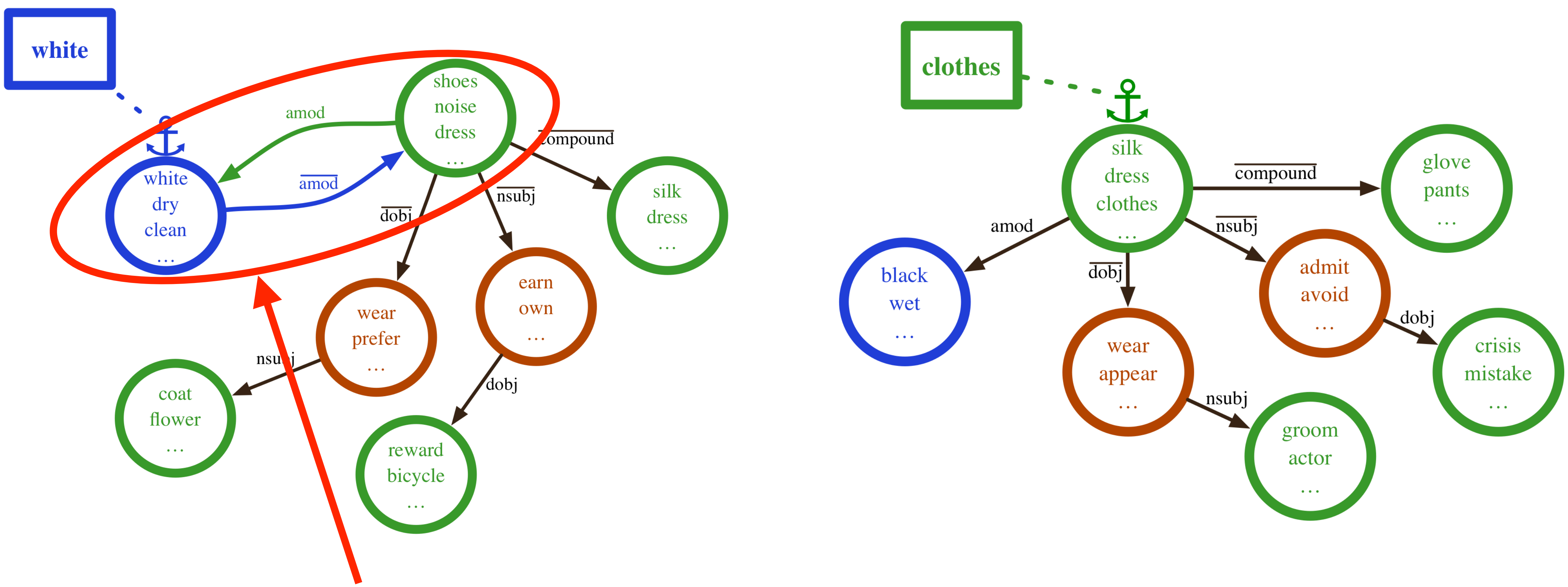


What are APTs?



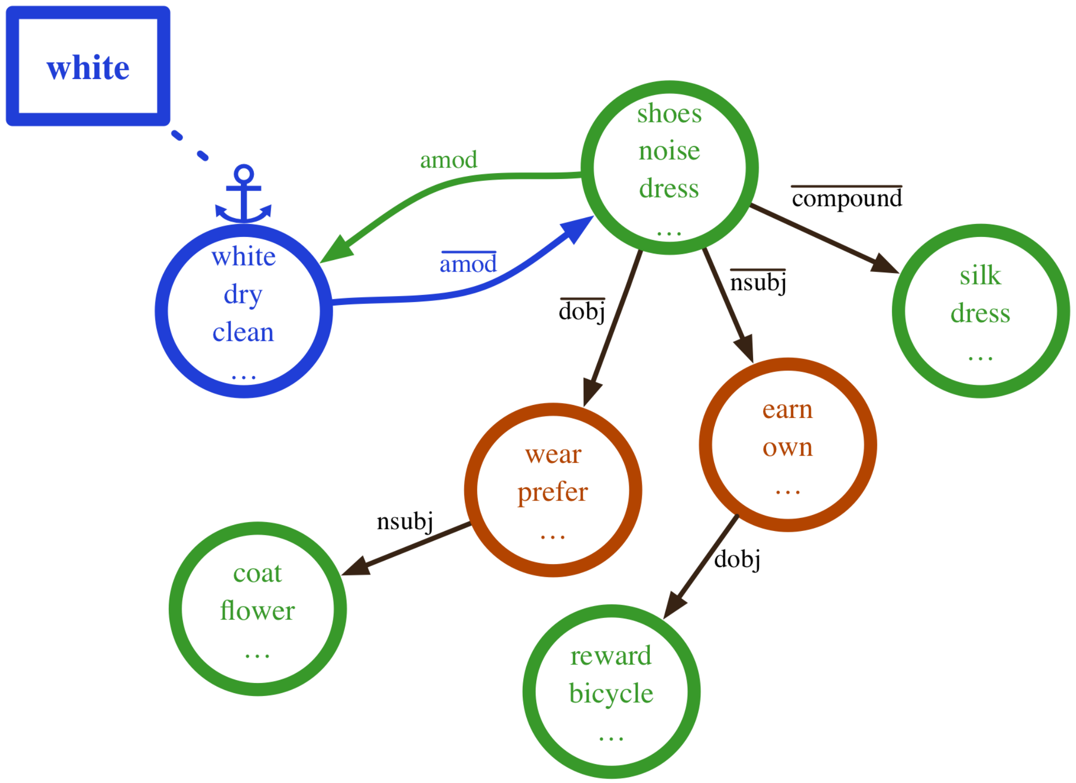
- All edges are bi-directional (see APT for white)

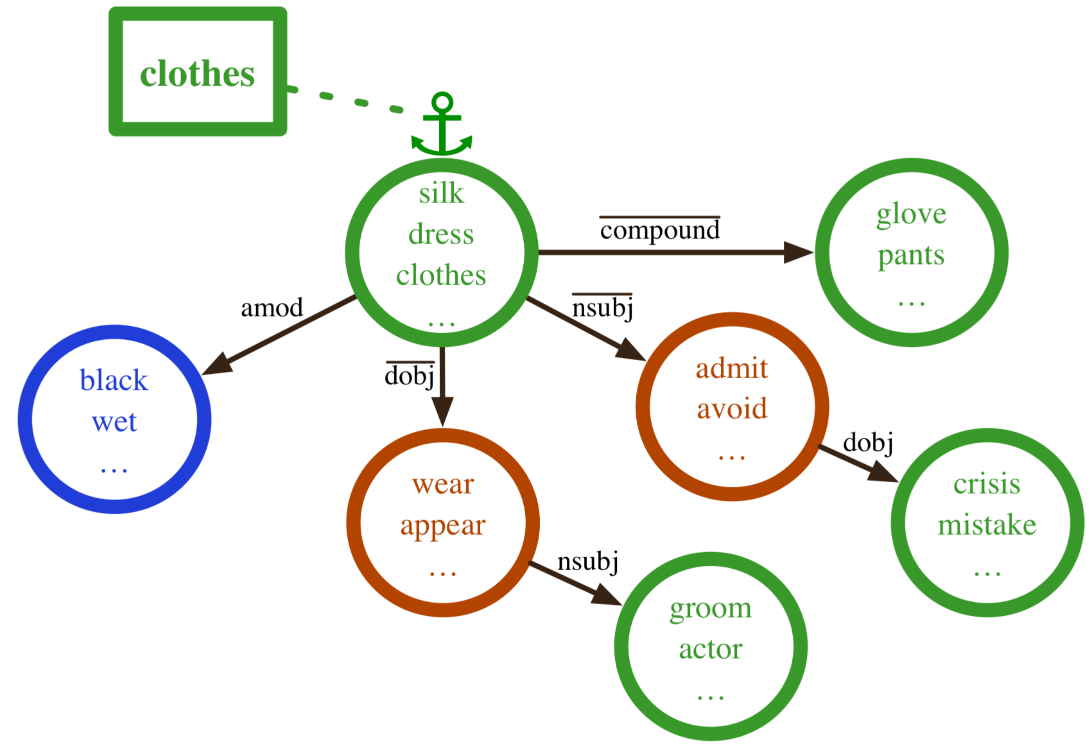
What are APTs?



- All edges are bi-directional (see APT for white)

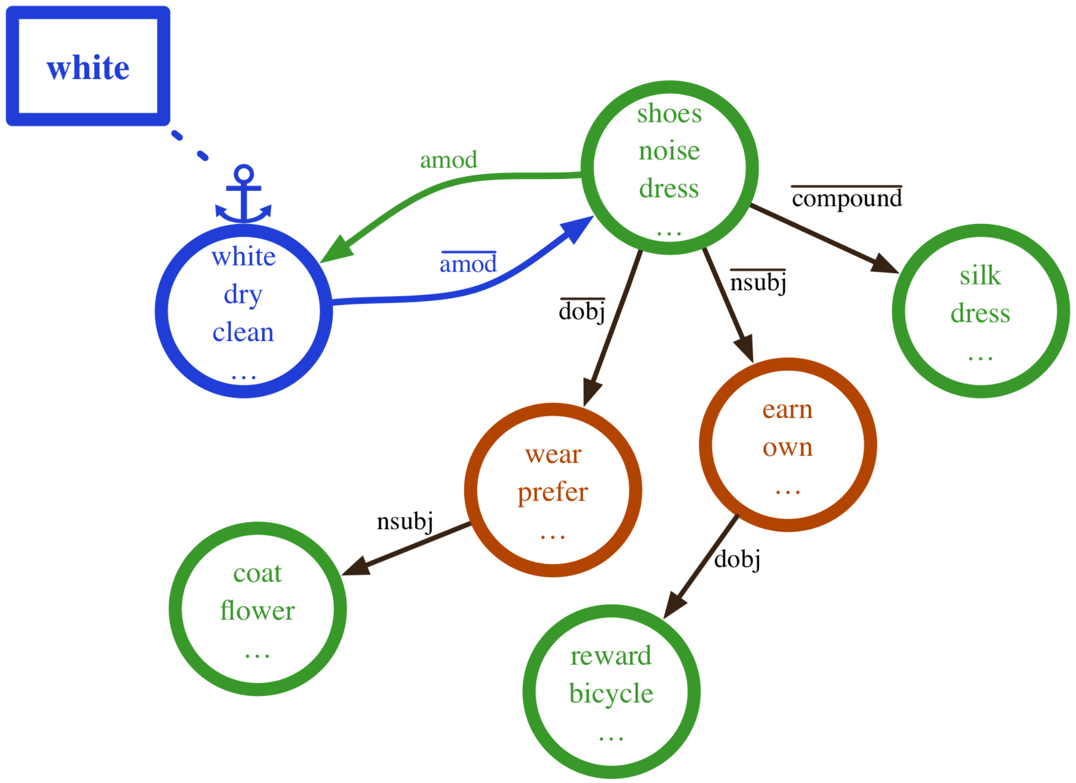
What are APTs?

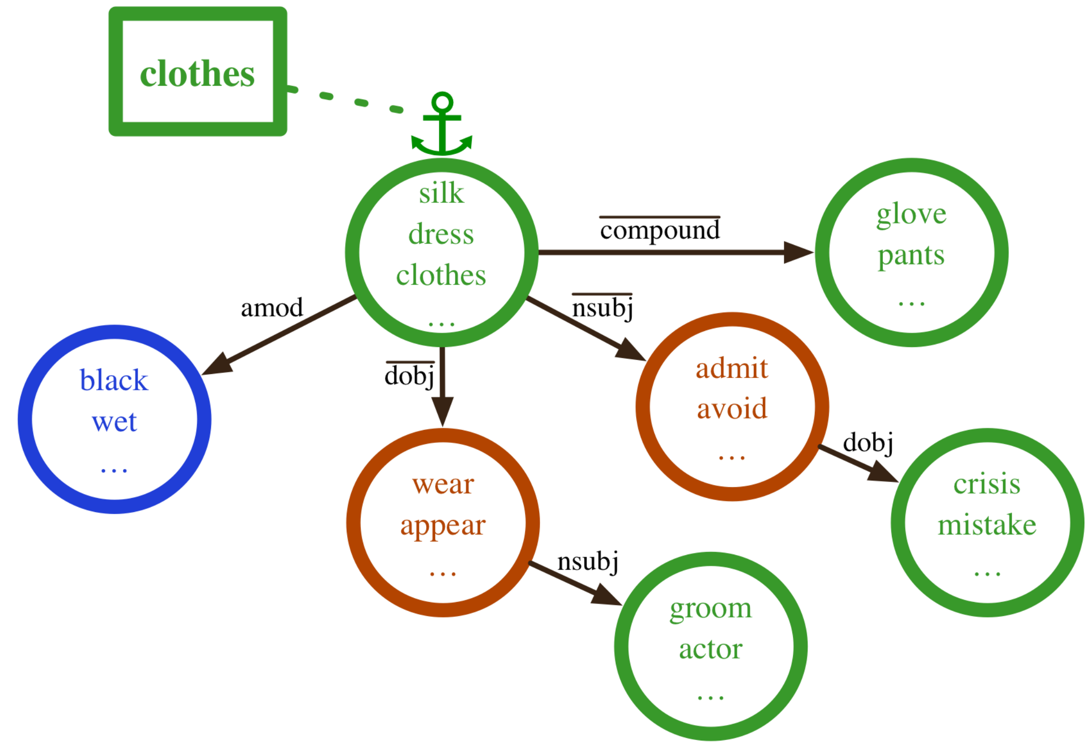




- All edges are bi-directional (see APT for white)

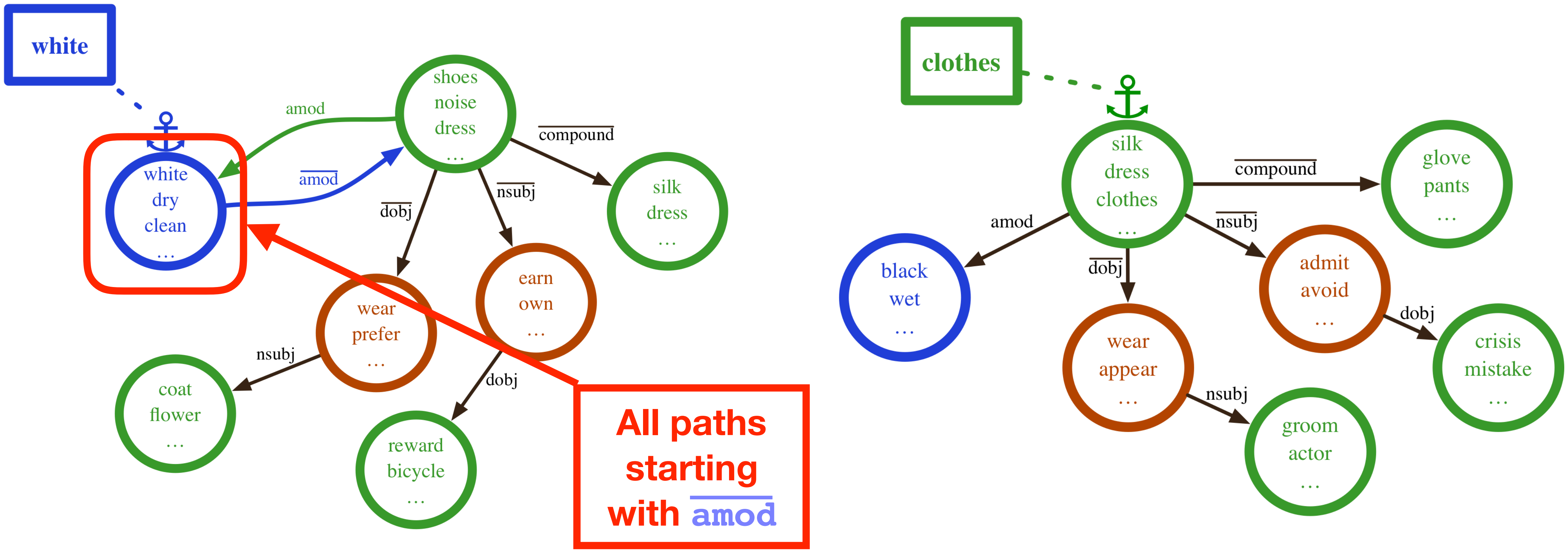
What are APTs?





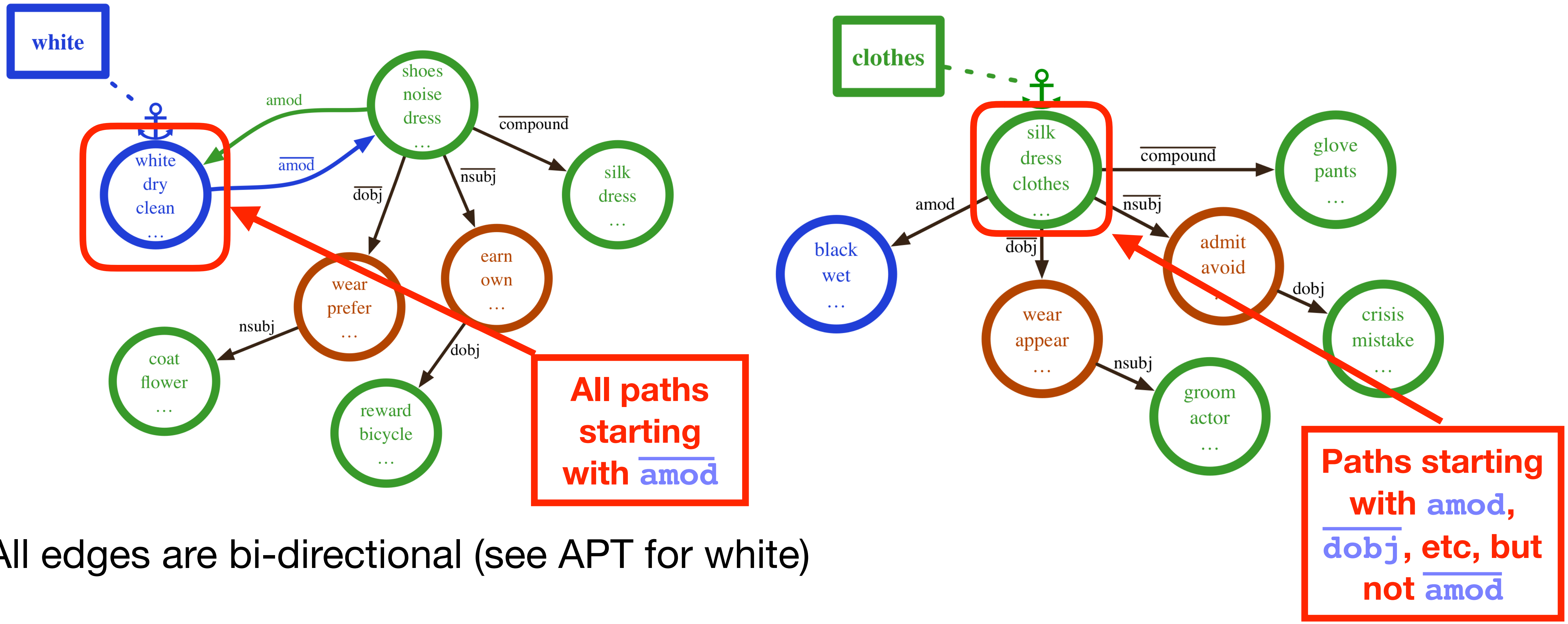
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different

What are APTs?



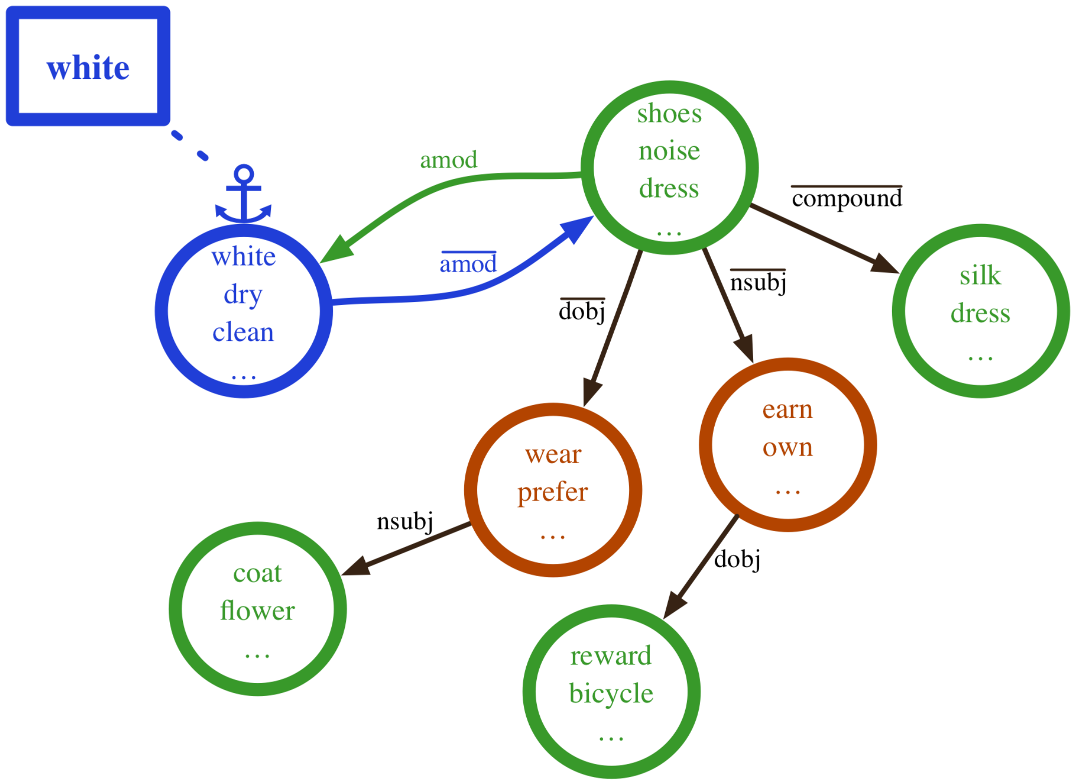
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different

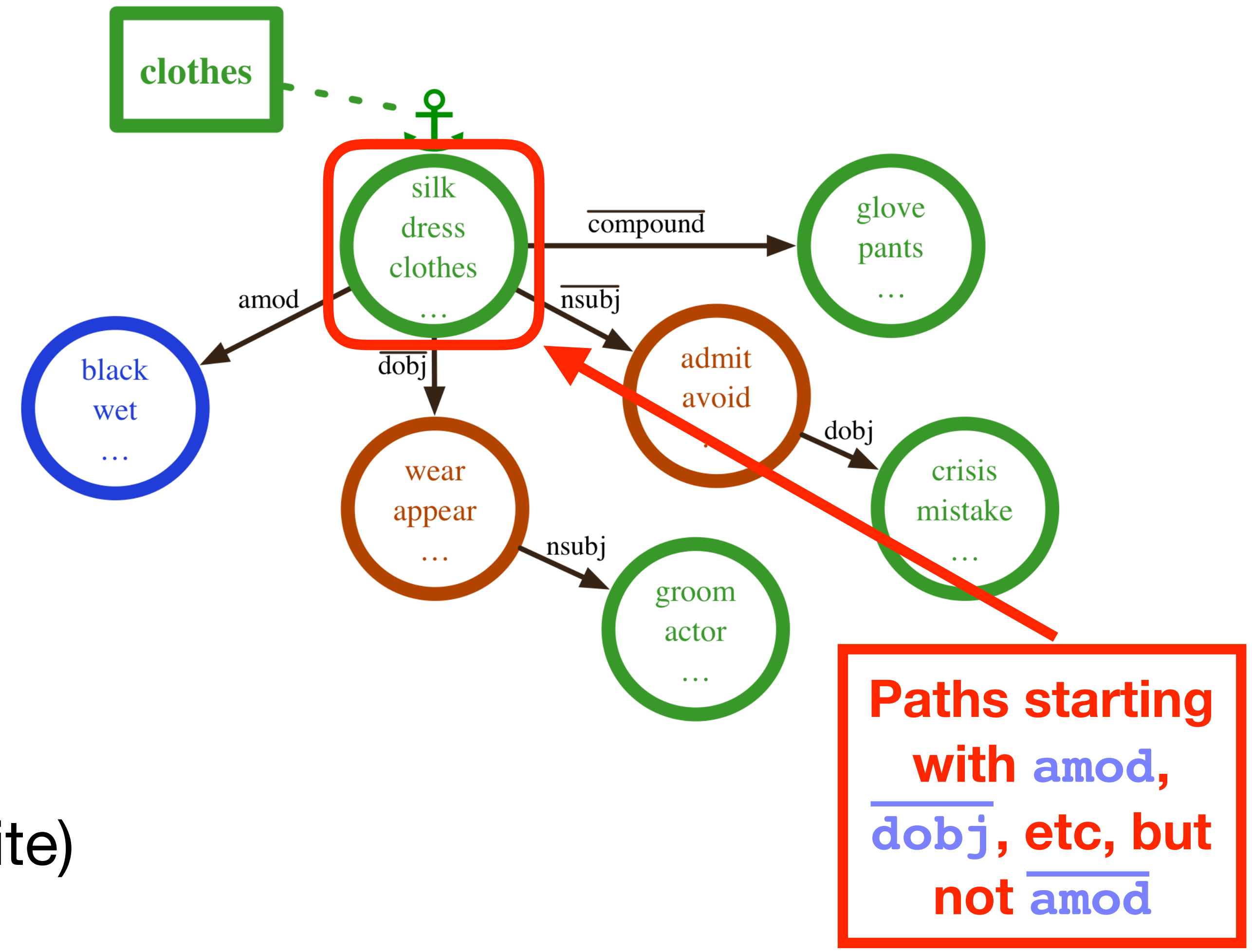
What are APTs?



- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different

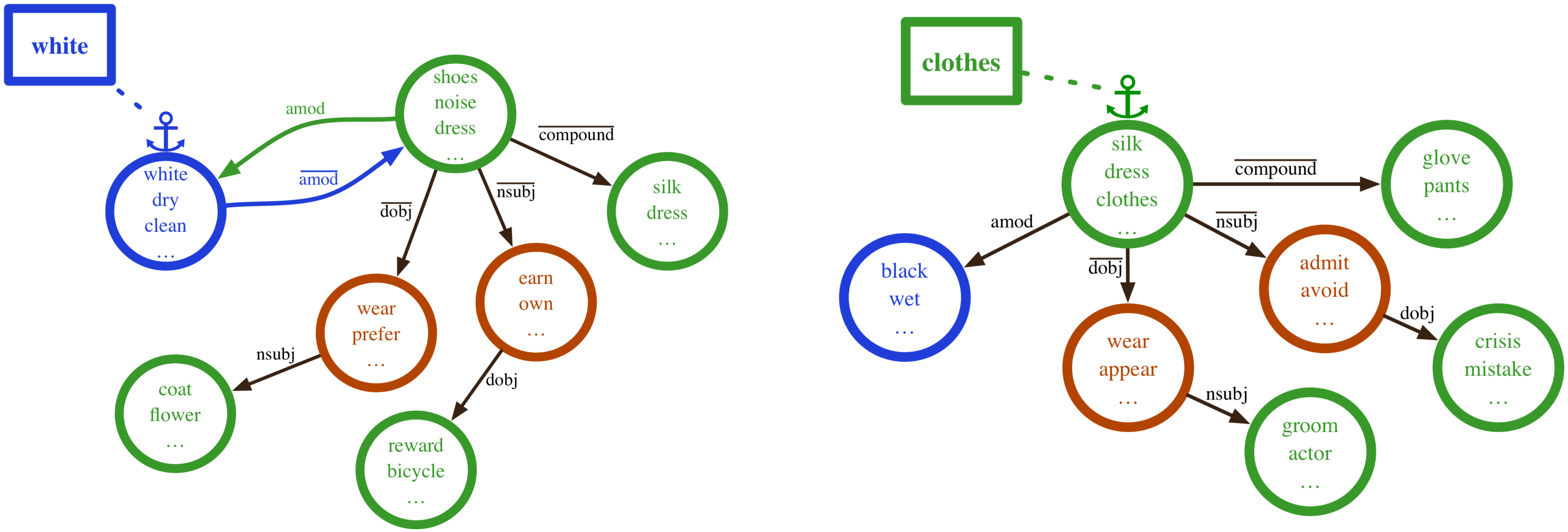
What are APTs?





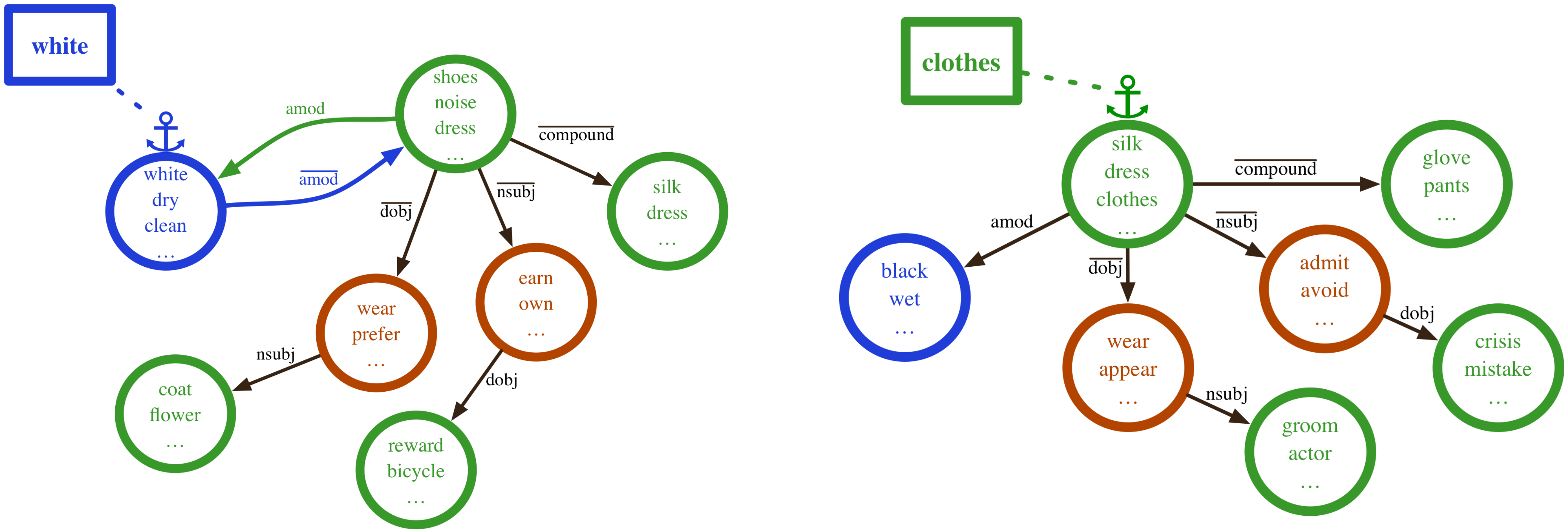
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different

What are APTs?



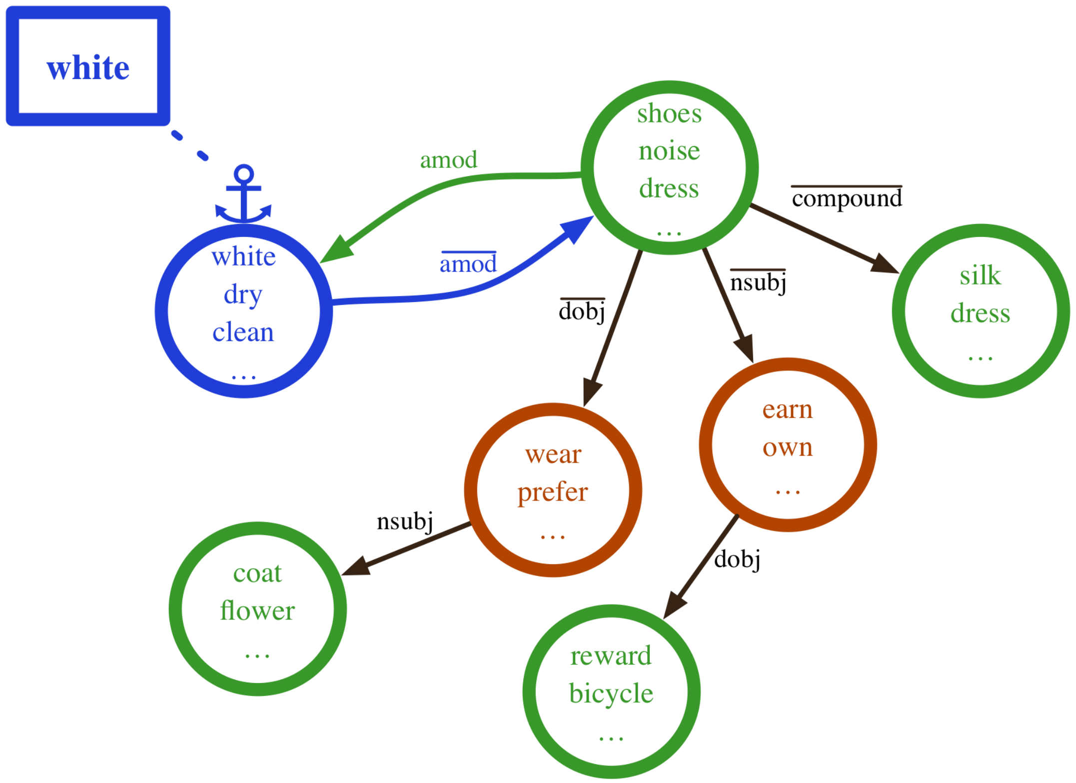
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different

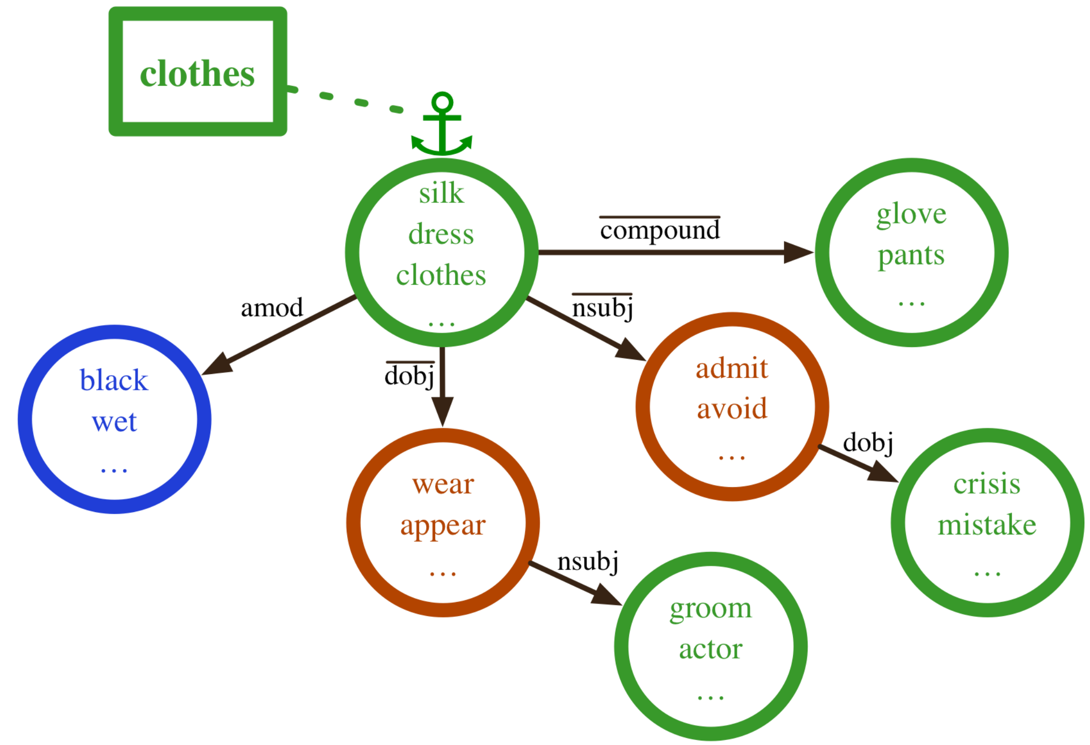
What are APTs?



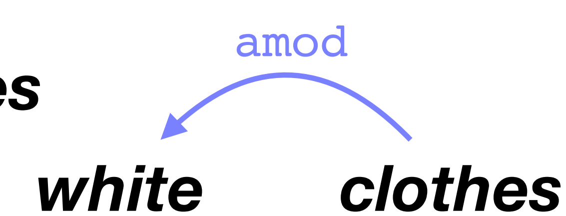
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different
- Suppose we want to compose the AN phrase ***white clothes***

What are APTs?

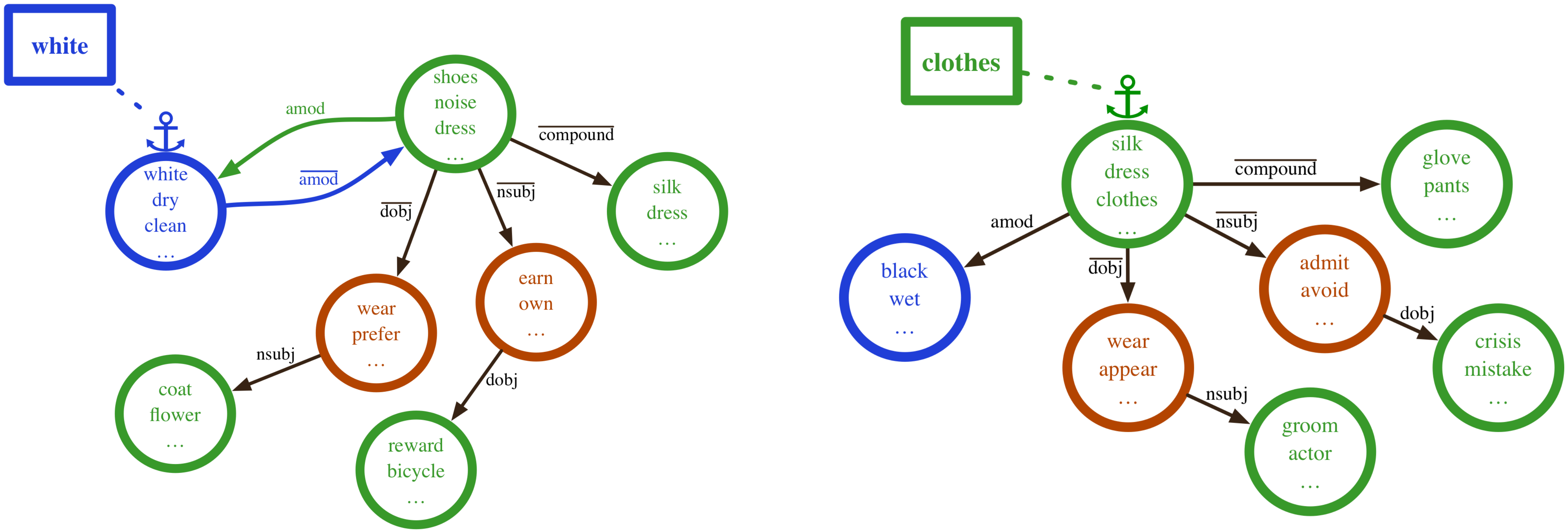




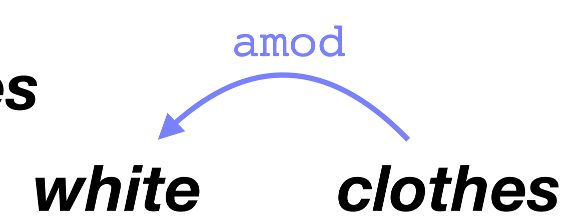
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different
- Suppose we want to compose the AN phrase ***white clothes***



What are APTs?



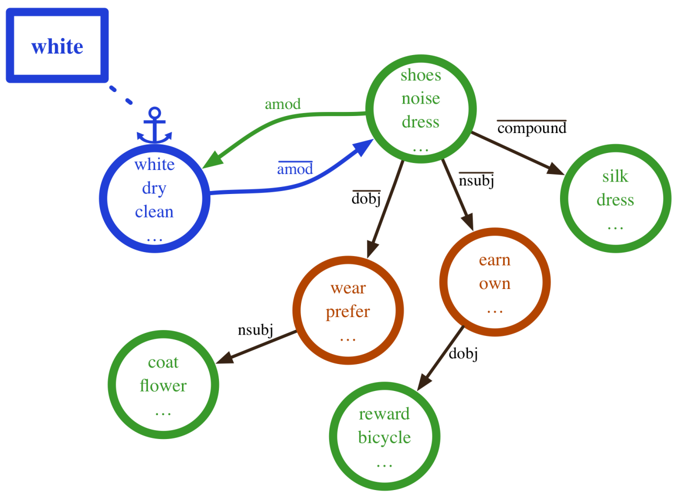
- All edges are bi-directional (see APT for white)
- Feature spaces of words with different grammatical roles are quite different
- Suppose we want to compose the AN phrase ***white clothes***



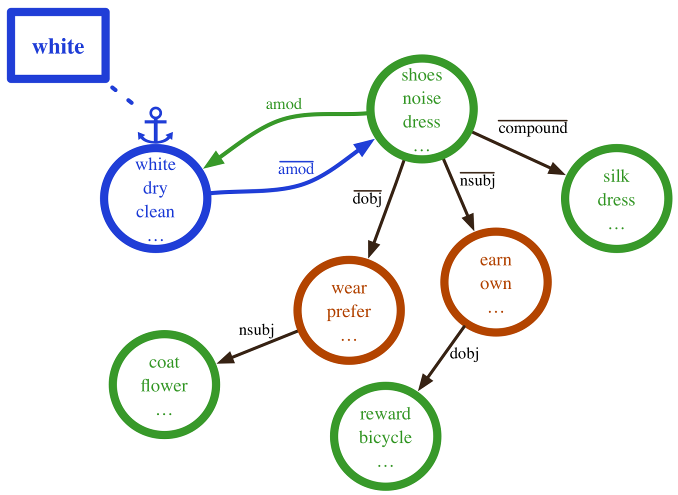
- Lets vectorise them…

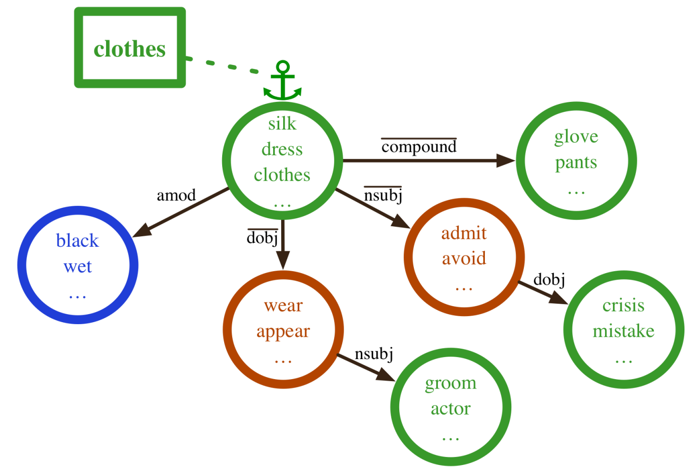
What are APTs?

What are APTs?

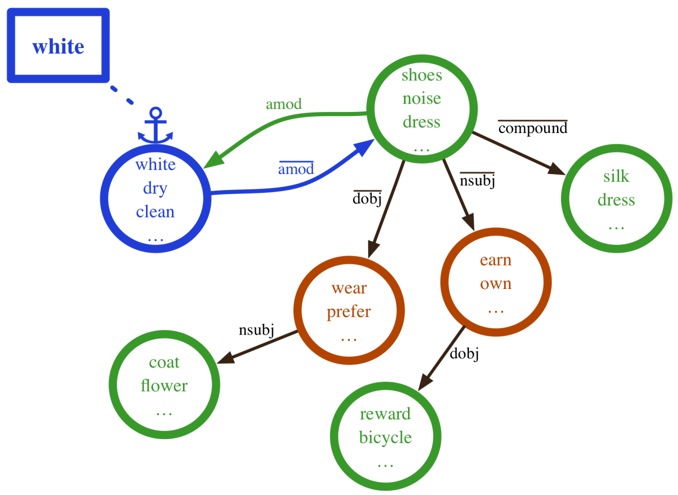


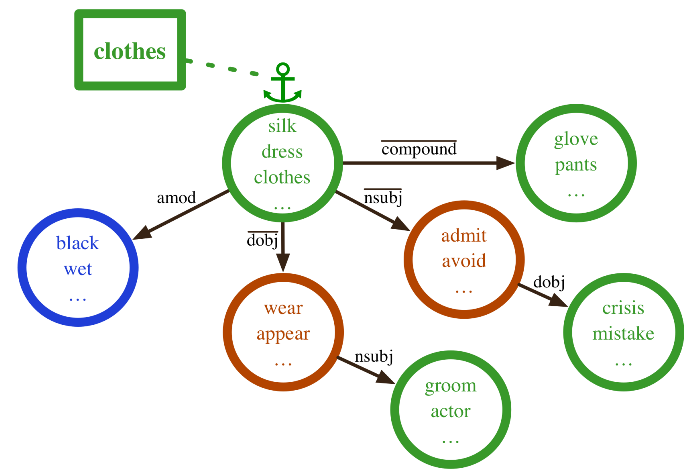
What are APTs?





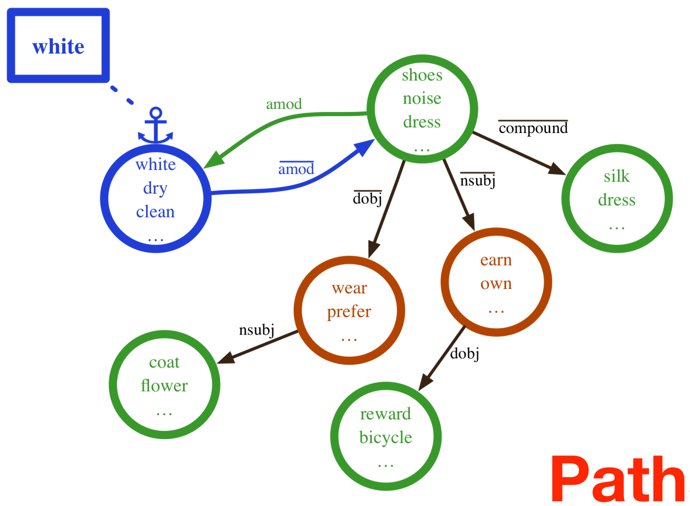
What are APTs?

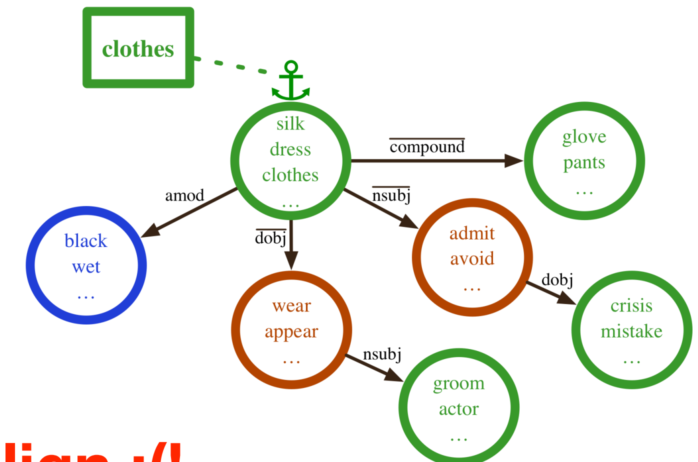




white	clothes
:clean	amod:wet
$\overline{\texttt{amod}}$:shoes	:dress
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}.\texttt{nsubj}$:coat	$\overline{\texttt{dobj}}.\texttt{nsubj}$:actor

What are APTs?





Paths don't align :(!

white	clothes
:clean	$\texttt{amod}$:wet
$\overline{\texttt{amod}}$:shoes	:dress
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}.\texttt{nsubj}$:coat	$\overline{\texttt{dobj}}.\texttt{nsubj}$:actor

What are APTs?

Paths don't align :(!

white	clothes
:clean	amod:wet
$\overline{\texttt{amod}}$:shoes	:dress
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}.\texttt{nsubj}$:coat	$\overline{\texttt{dobj}}.\texttt{nsubj}$:actor

- Can't leverage distributional commonalities between ***white*** and ***clothes***

What are APTs?

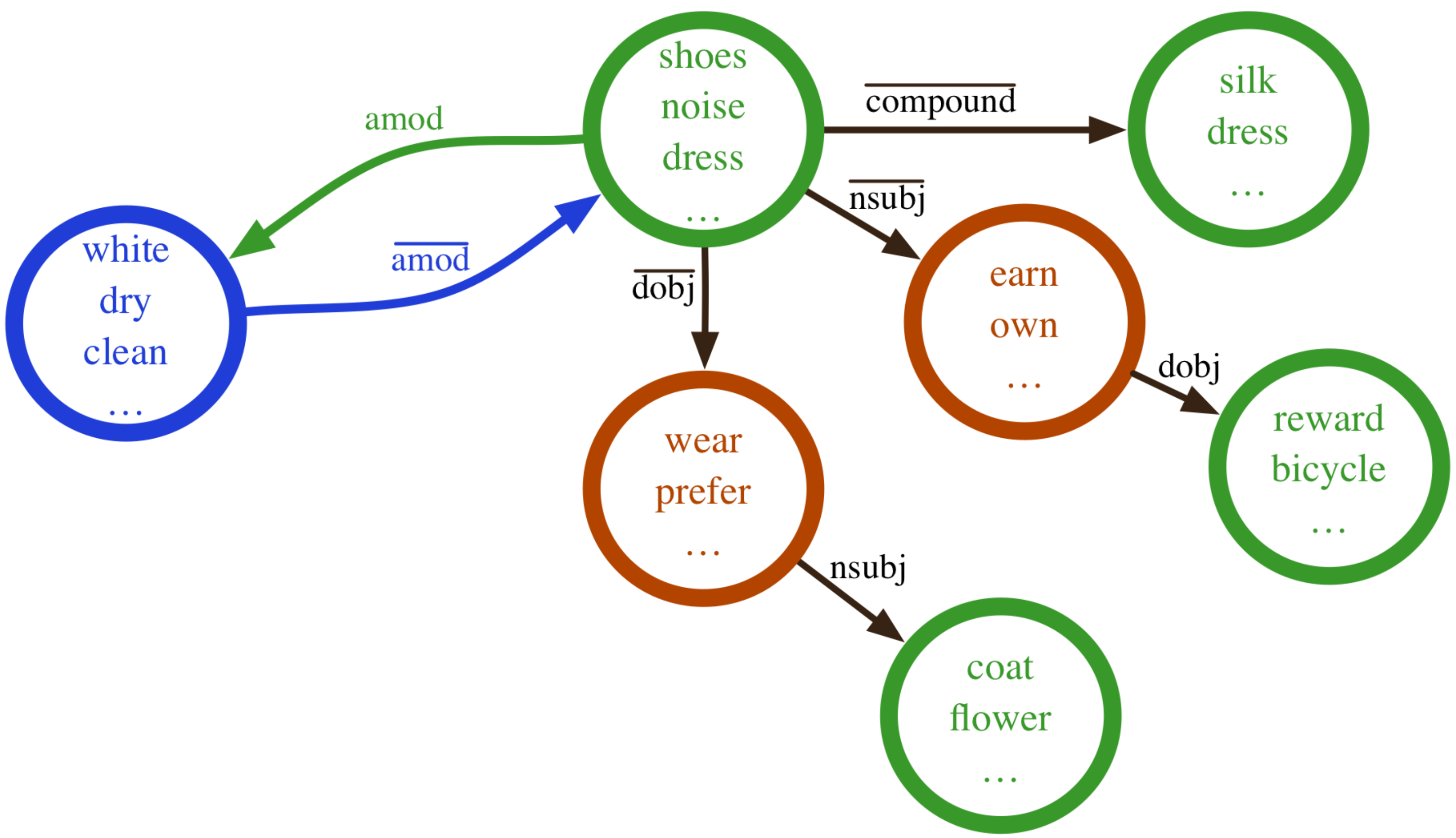
Paths don't align :(!

white	clothes
:clean	amod:wet
$\overline{\texttt{amod}}$:shoes	:dress
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{amod}}.\overline{\texttt{dobj}}.\texttt{nsubj}$:coat	$\overline{\texttt{dobj}}.\texttt{nsubj}$:actor

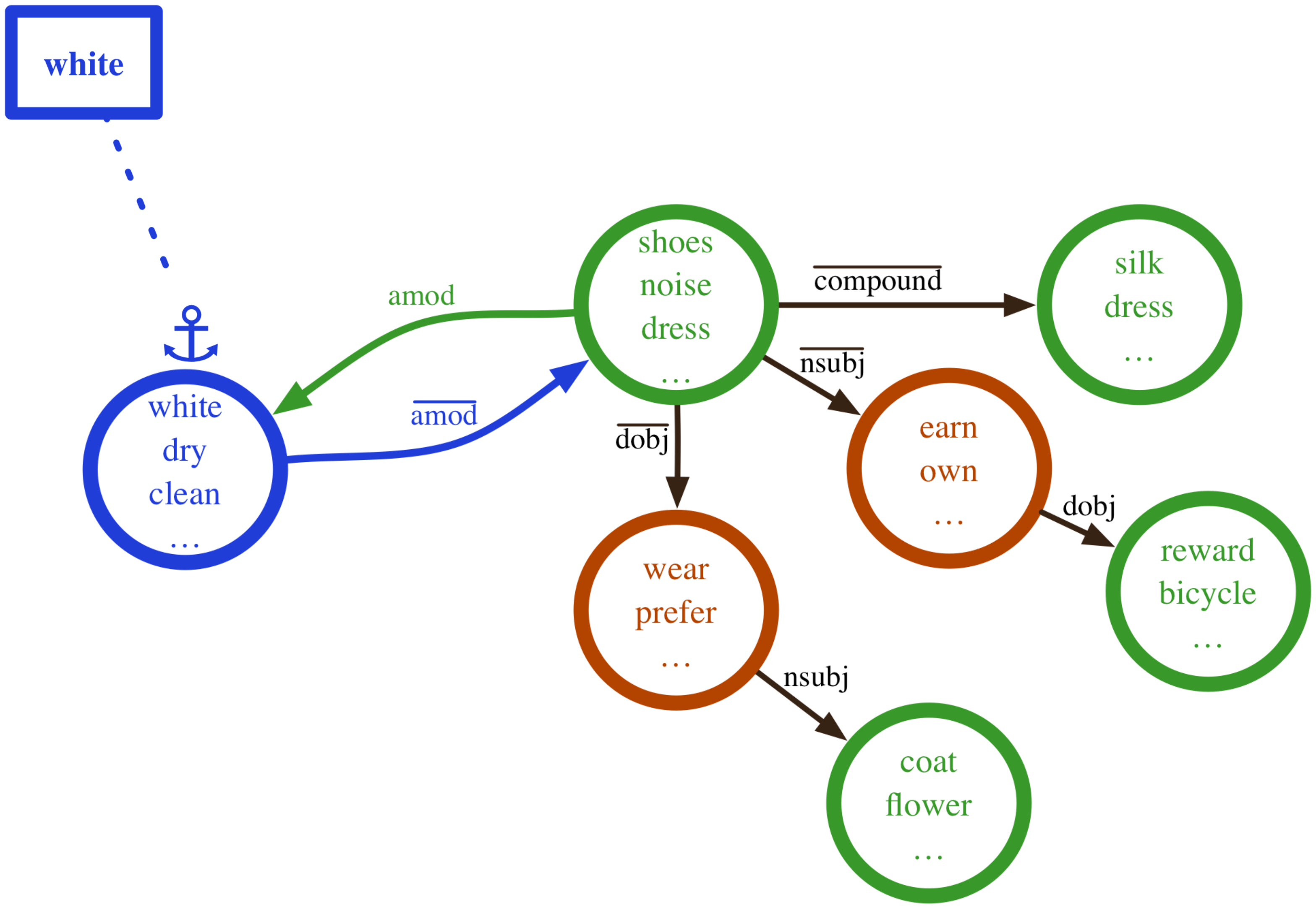
- Can't leverage distributional commonalities between ***white*** and ***clothes***
- Need a mechanism for aligning representations with different grammatical roles before composition

What are APTs?

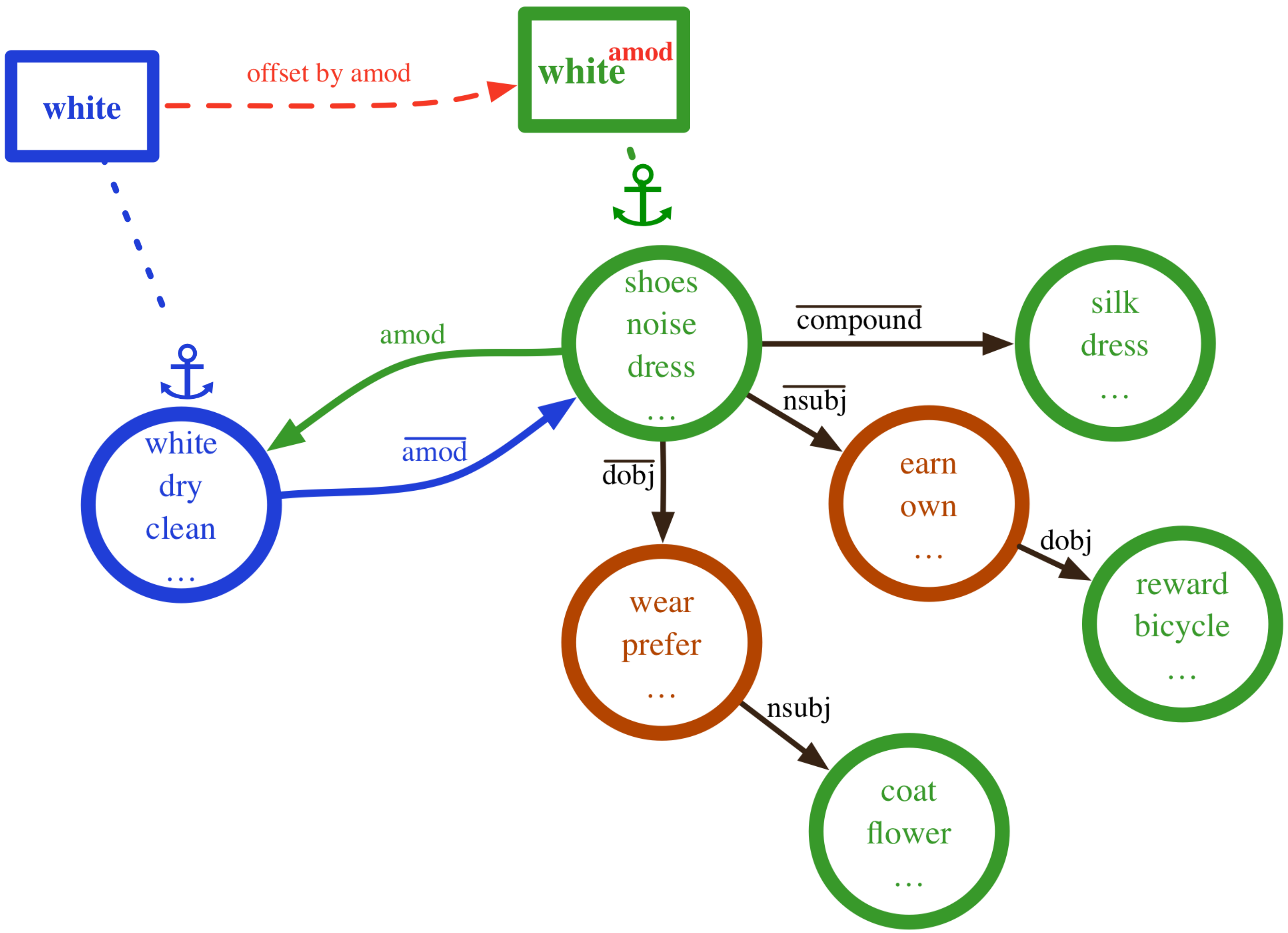
What are APTs?



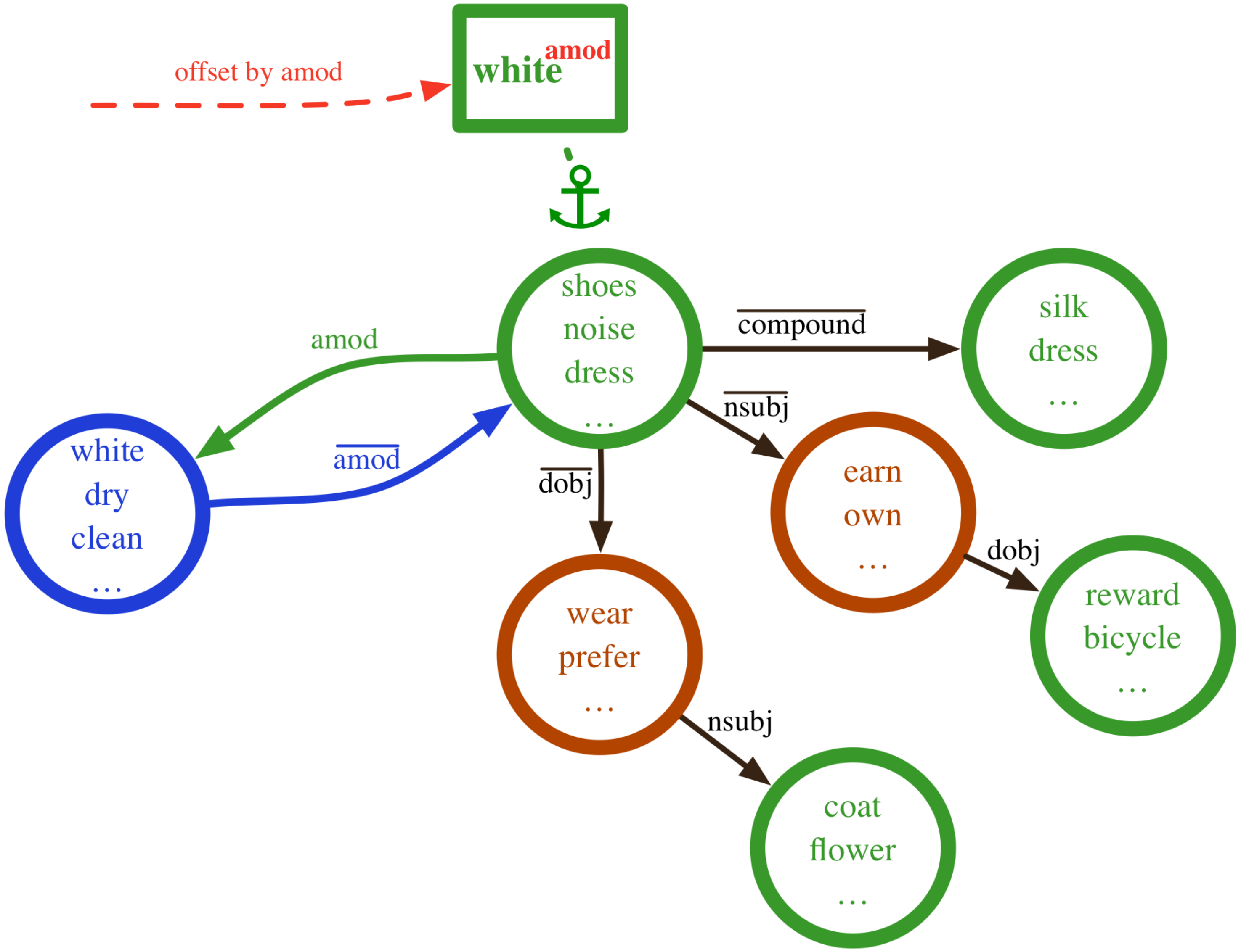
What are APTs?



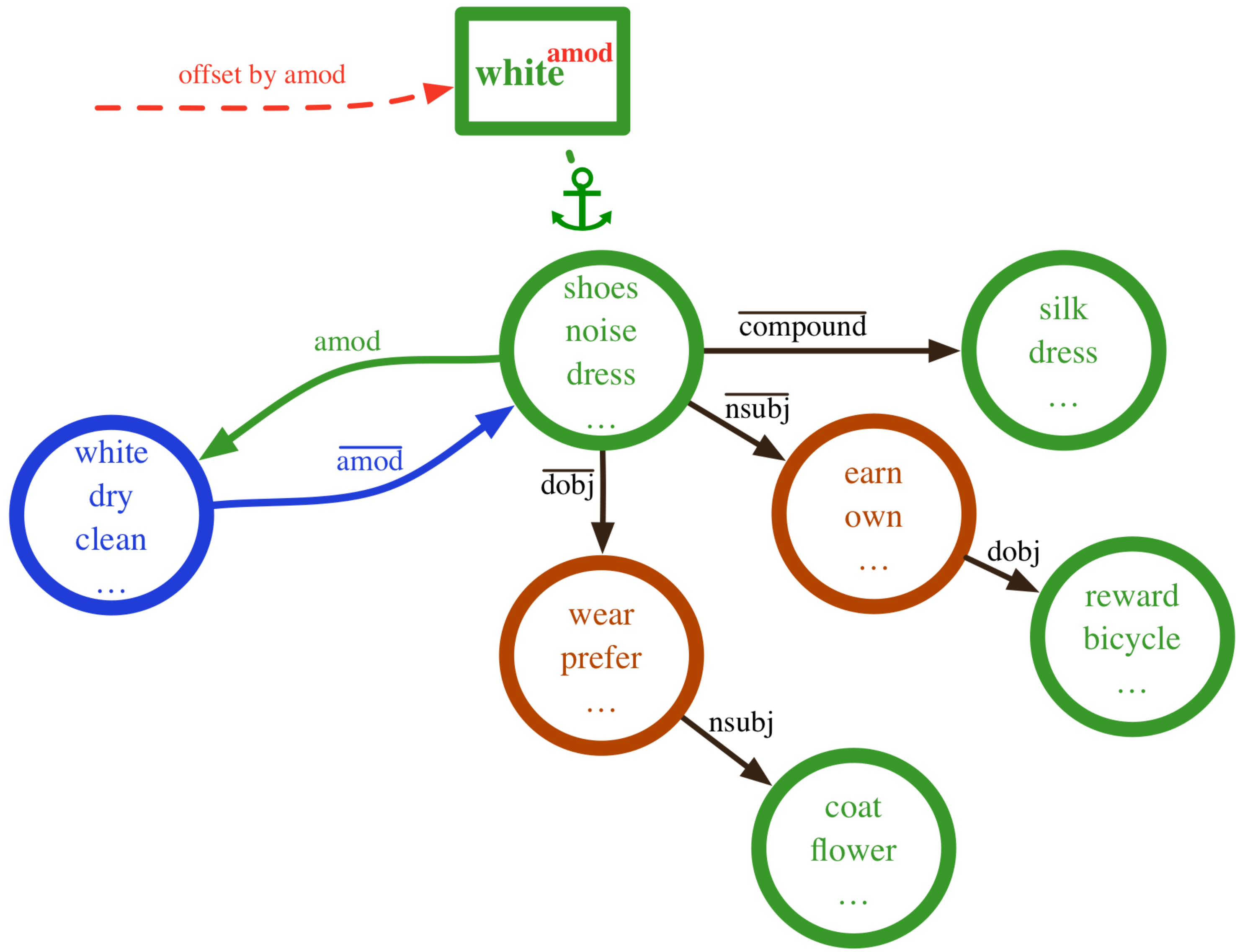
What are APTs?



What are APTs?

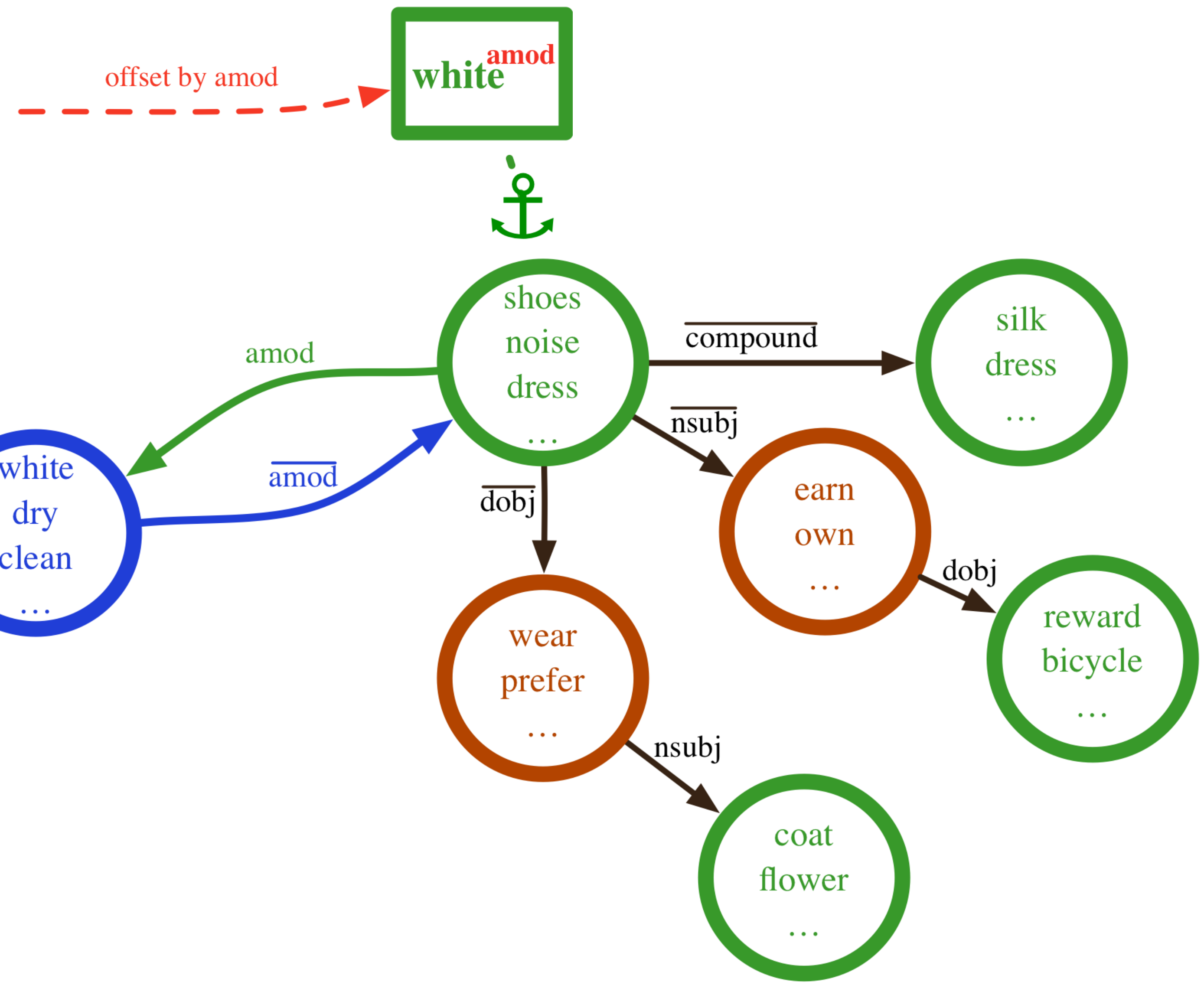


What are APTs?



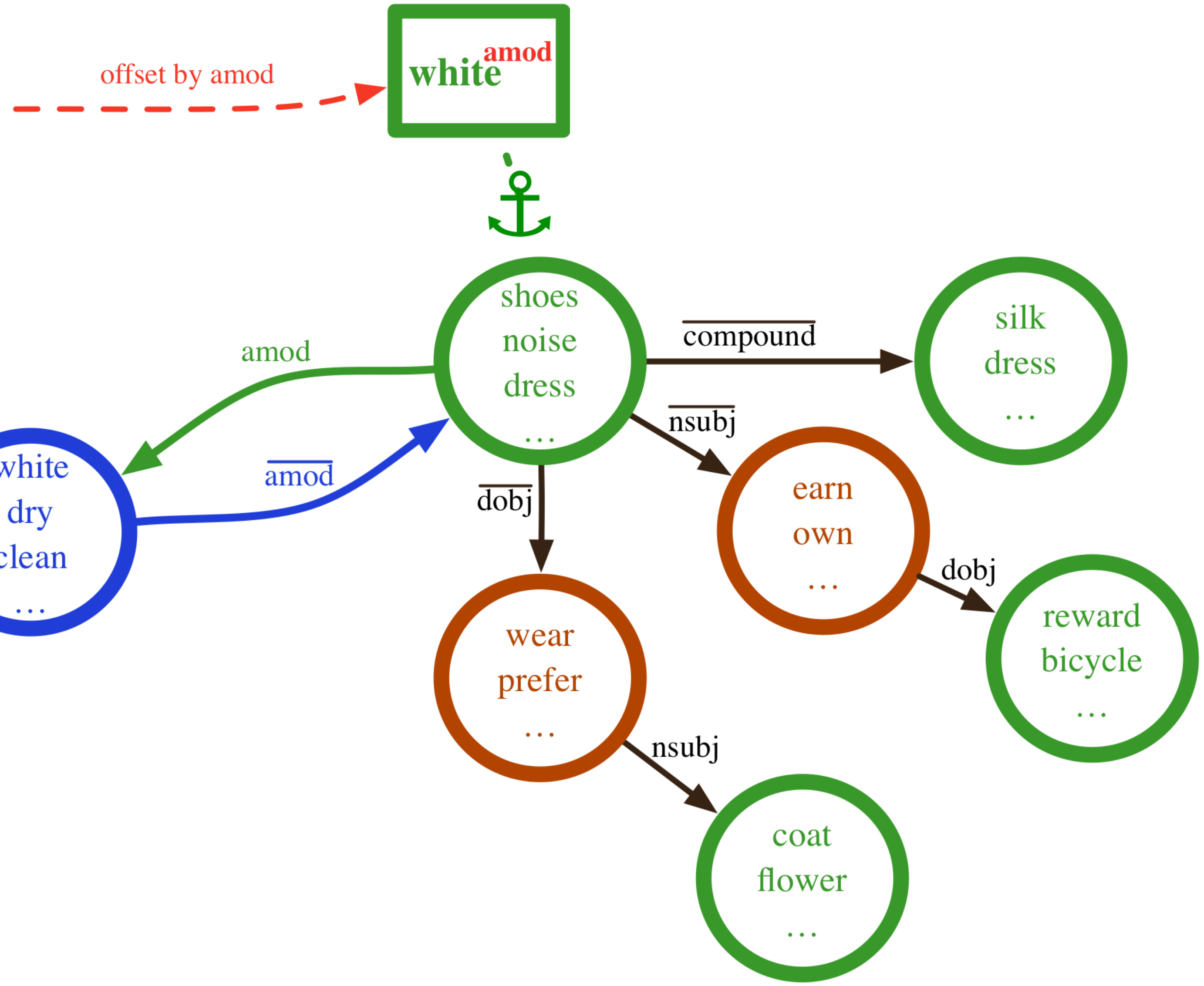
- Offset by `amod` to create a noun view for the adjective *white*

What are APTs?



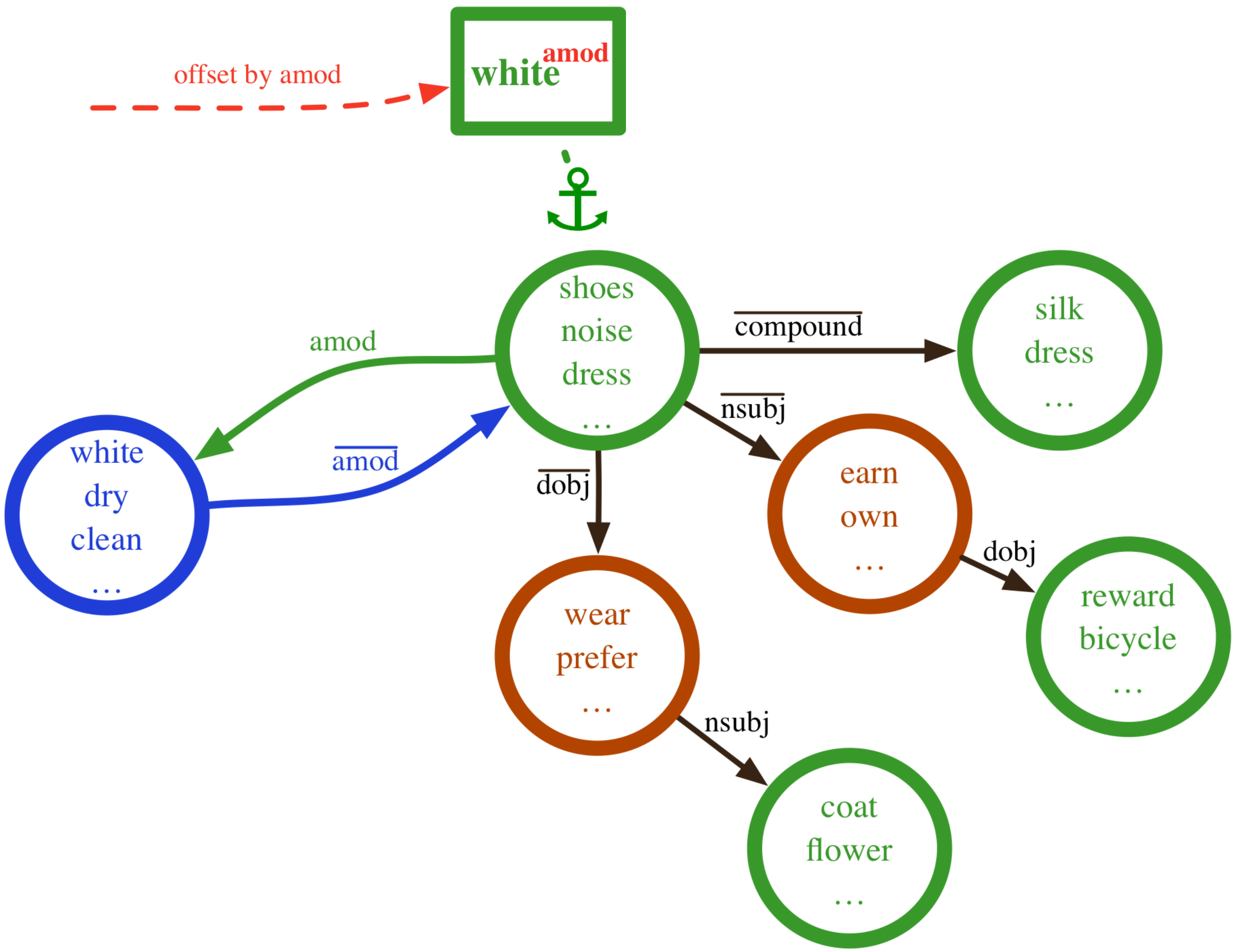
- Offset by `amod` to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"

What are APTs?



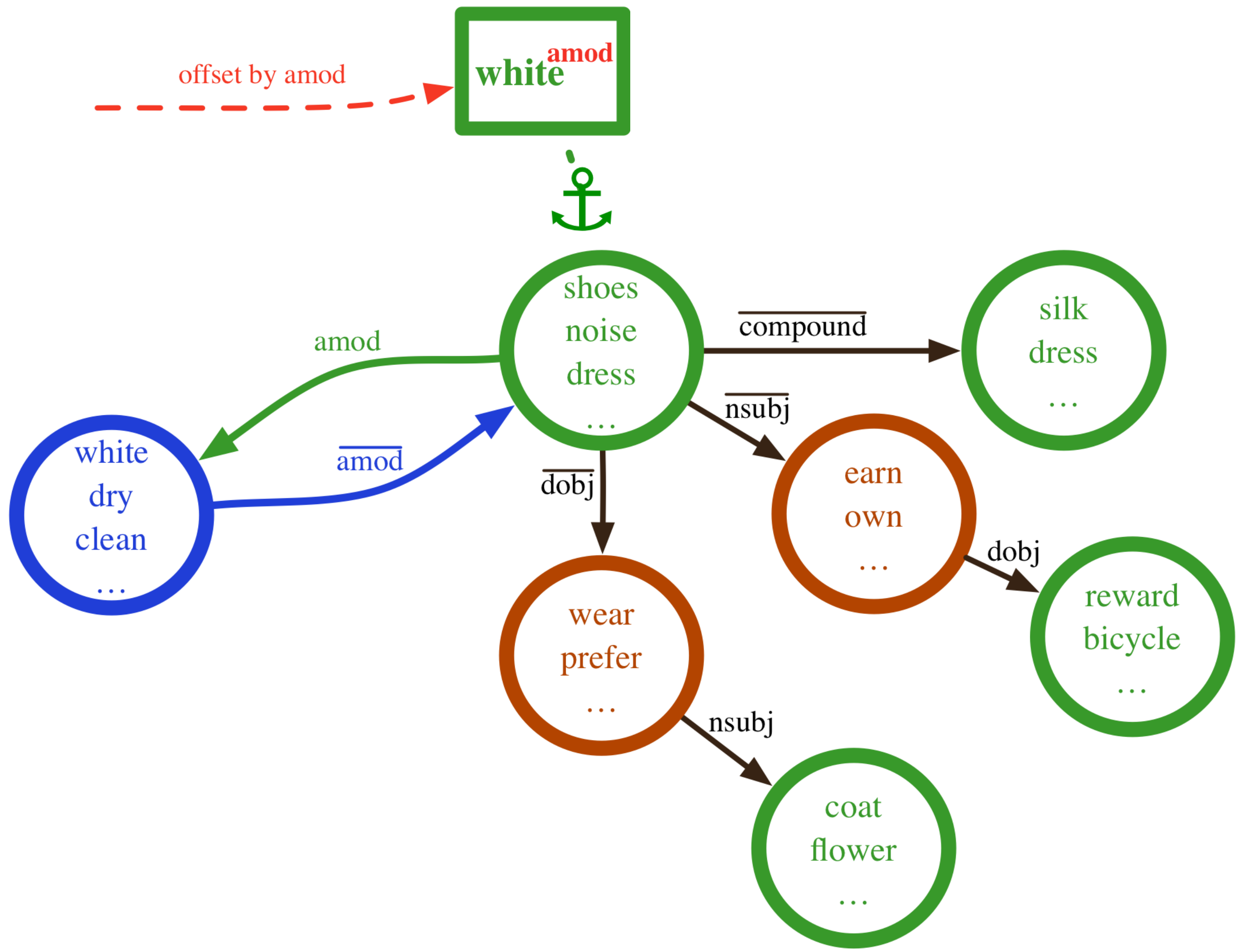
- Offset by `amod` to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

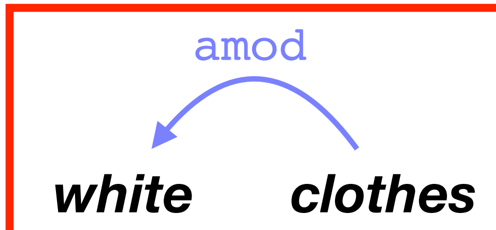
What are APTs?



- Offset by `amod` to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

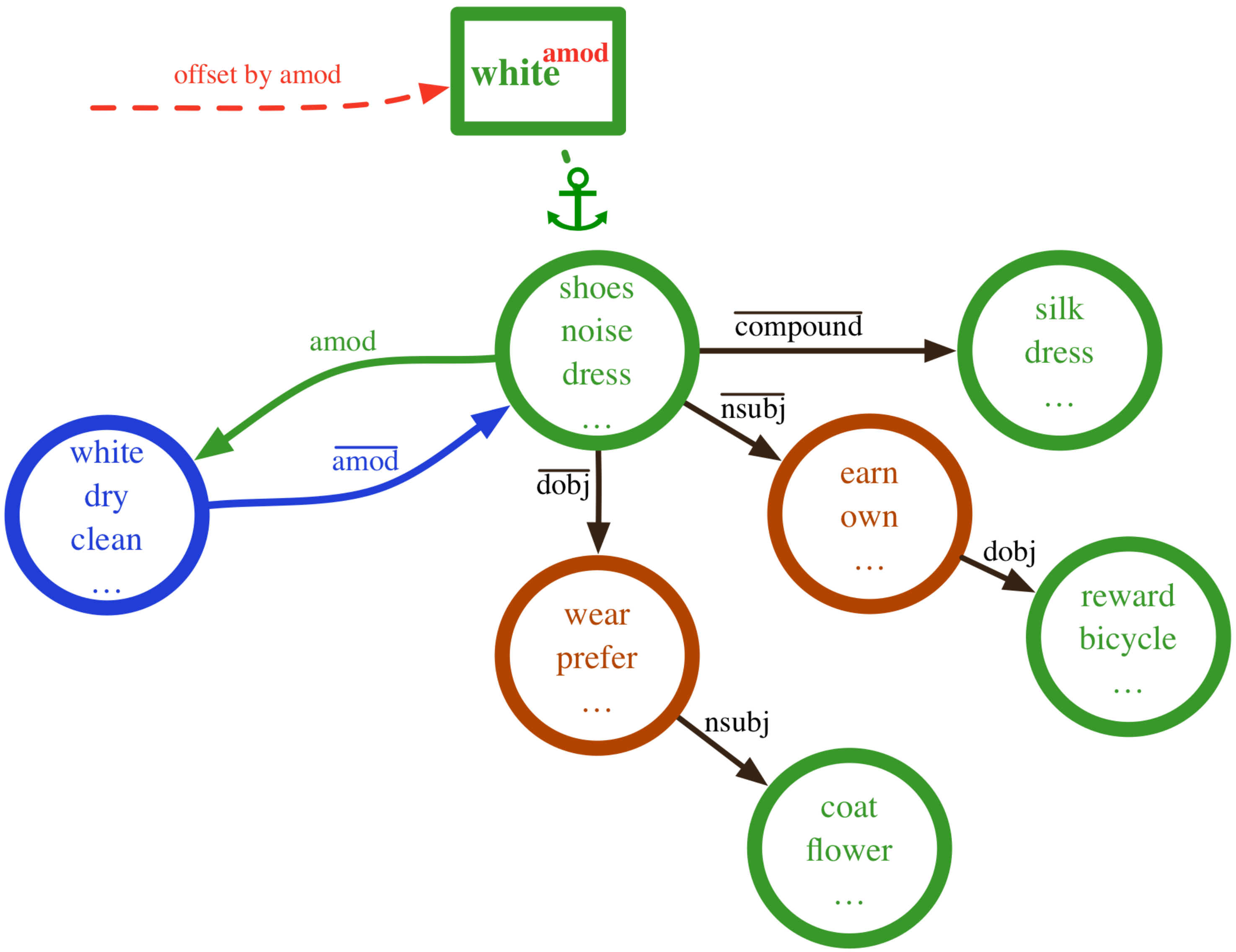
What are APTs?





- Offset by amod to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

What are APTs?



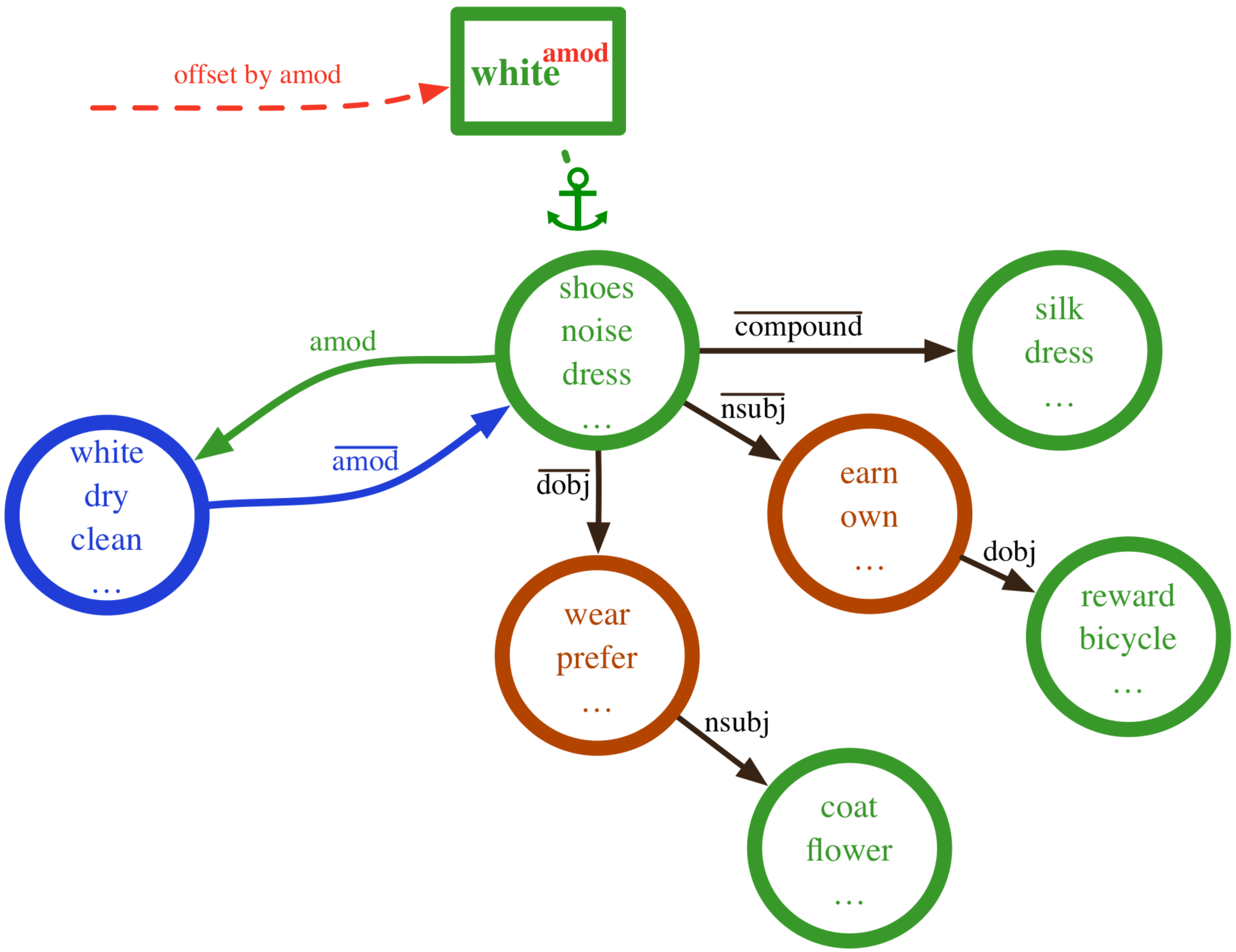
- Offset by `amod` to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

amod

white ***clothes***

- *white* connected to *clothes* via $\overline{\texttt{amod}}$

What are APTs?



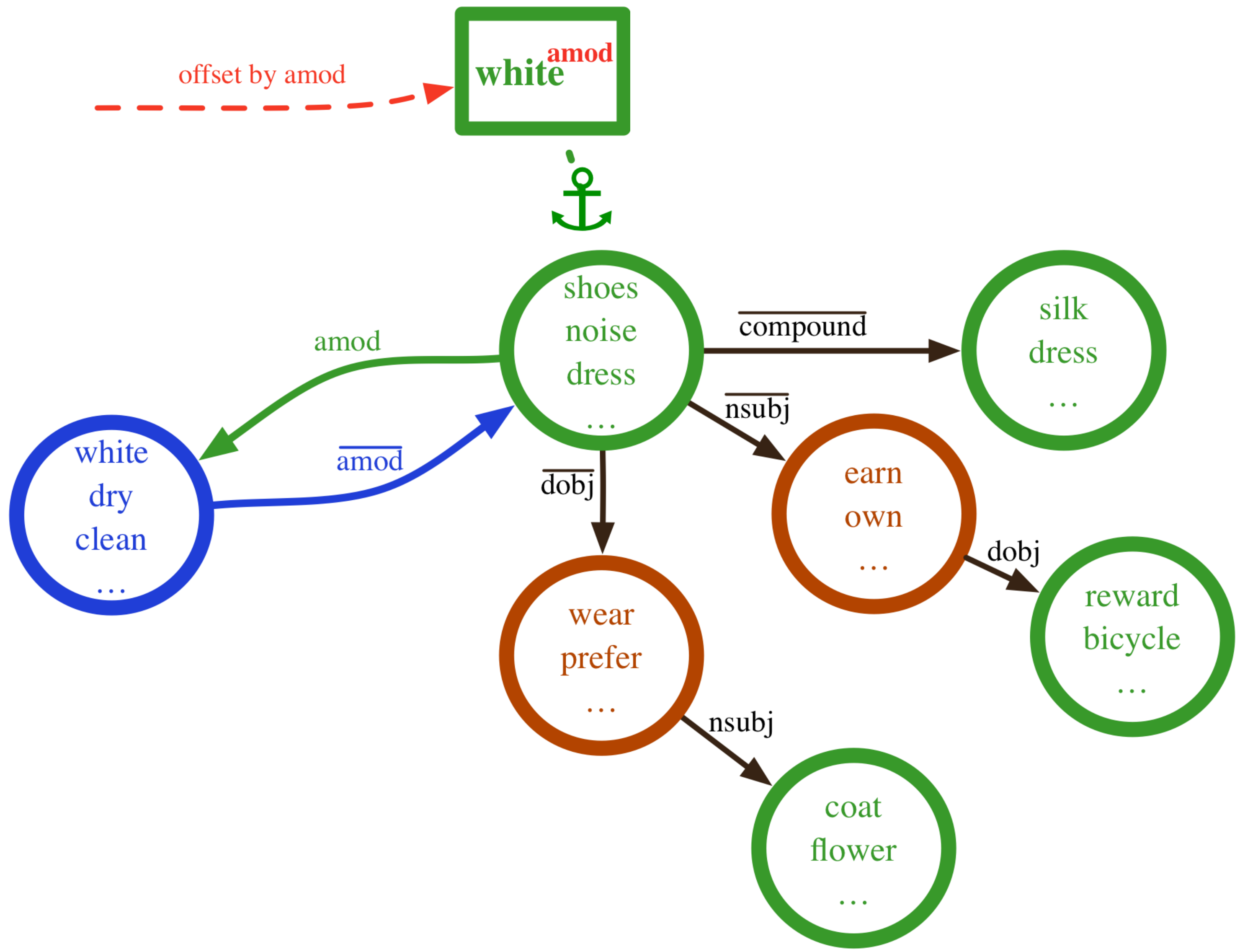
amod

white ***clothes***

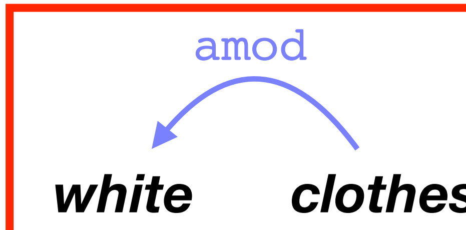
- *white* connected to *clothes* via amod
- for alignment, offset needs to happen in inverse direction to the head, so amod

- Offset by amod to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

What are APTs?

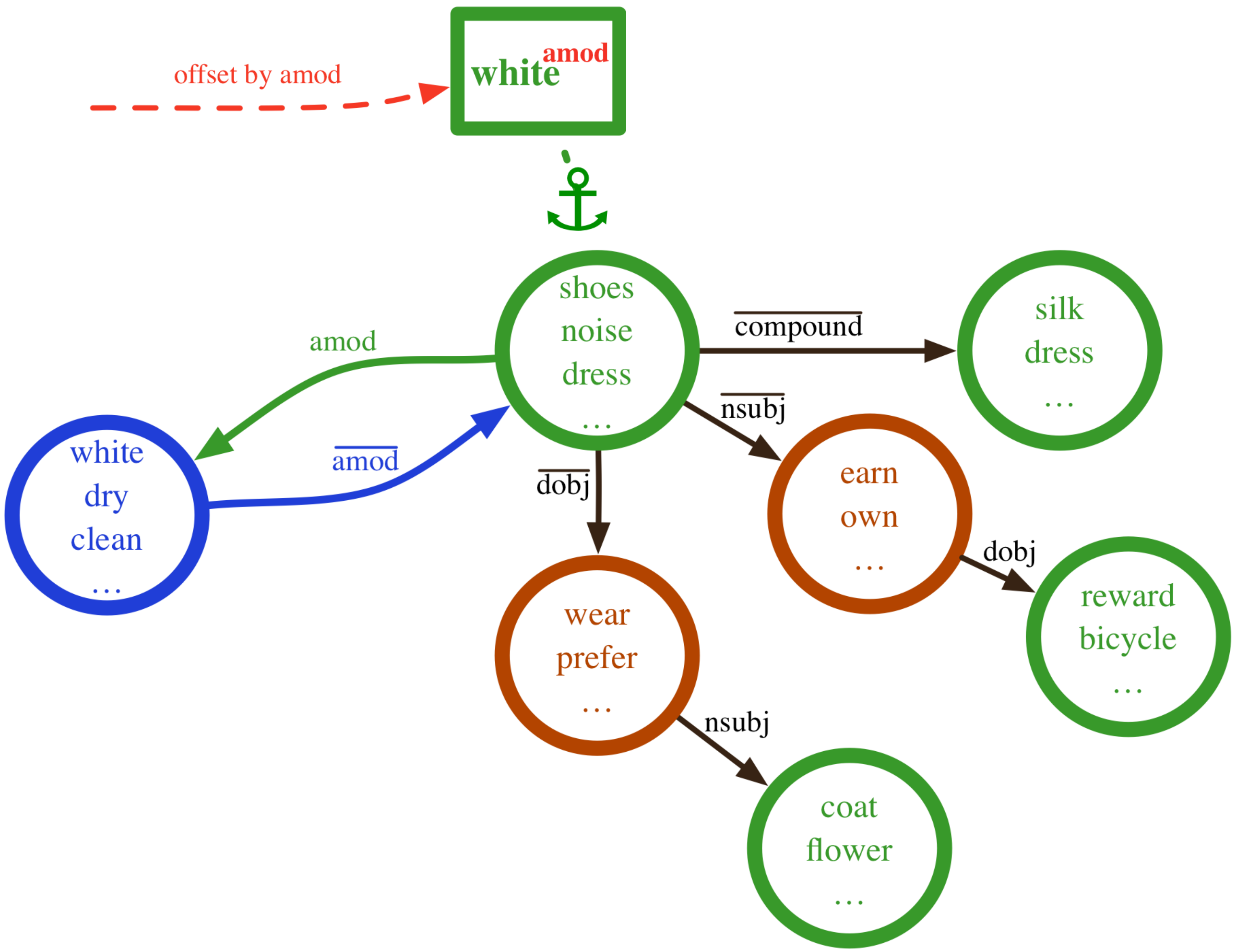


- Offset by amod to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted



- *white* connected to *clothes* via $\overline{\texttt{amod}}$
- for alignment, offset needs to happen in inverse direction to the head, so amod
- type reduction - $\texttt{amod}.\overline{\texttt{amod}}$ results in empty path ε

What are APTs?



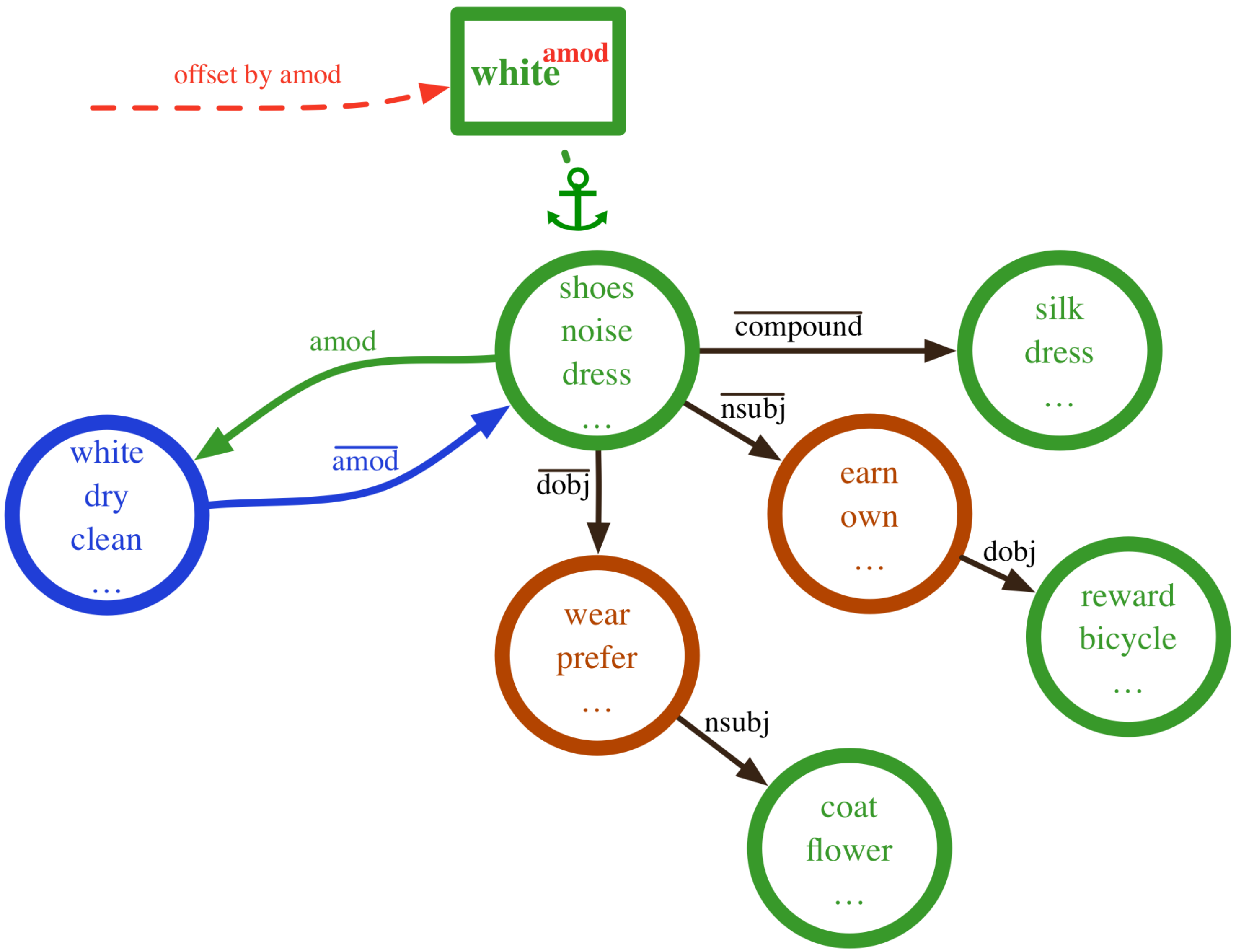
- Offset by amod to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

amod

white ***clothes***

- *white* connected to *clothes* via $\overline{\text{amod}}$
- for alignment, offset needs to happen in inverse direction to the head, so amod
- type reduction - amod.$\overline{\text{amod}}$ results in empty path **ε**
- Hence, travelling along the $\overline{\text{amod}}$ edge from *white* to *clothes* involves offsetting by amod

What are APTs?



- $\boxed{\text{Offset by amod}}$ to create a noun view for the adjective *white*
- Representing a "*thing that can be white*"
- Nothing structurally changes in the APT, only the position of the anchor is shifted

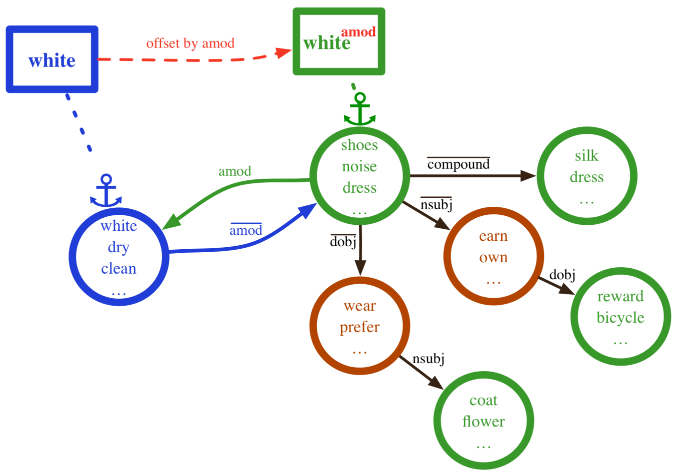
amod

white ***clothes***

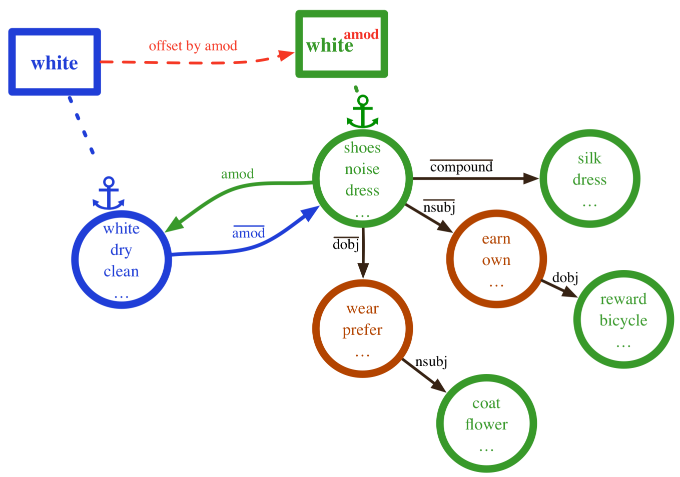
- *white* connected to *clothes* via $\overline{\text{amod}}$
- for alignment, offset needs to happen in inverse direction to the head, so amod
- type reduction - $\text{amod}.\overline{\text{amod}}$ results in empty path ε
- Hence, travelling along the $\overline{\text{amod}}$ edge from *white* to *clothes* involves offsetting by amod
- See Weir et al., (2016) for full details

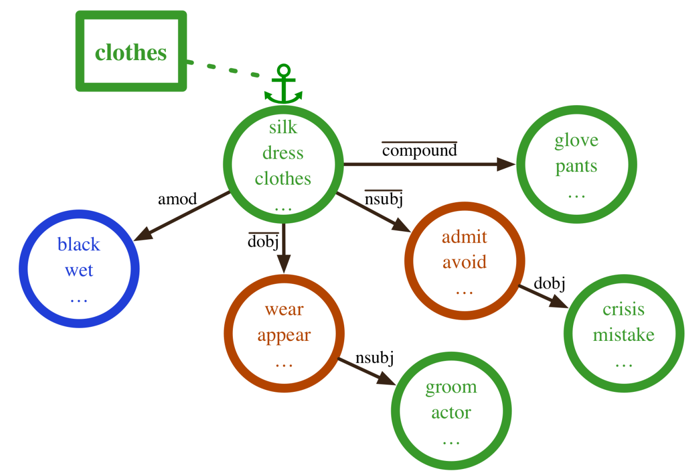
What are APTs?

What are APTs?

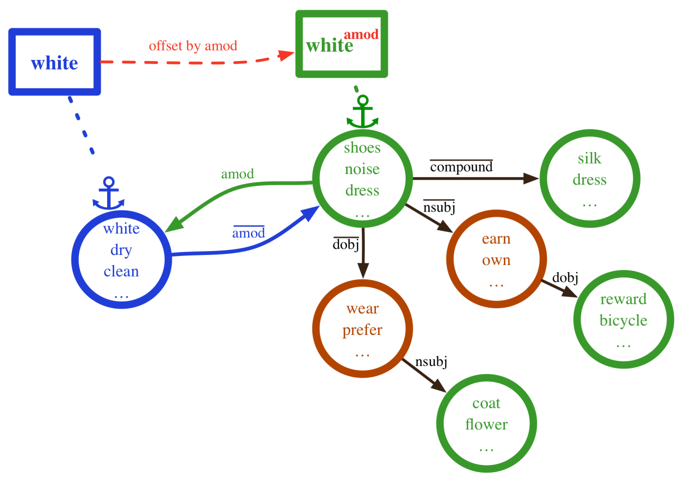


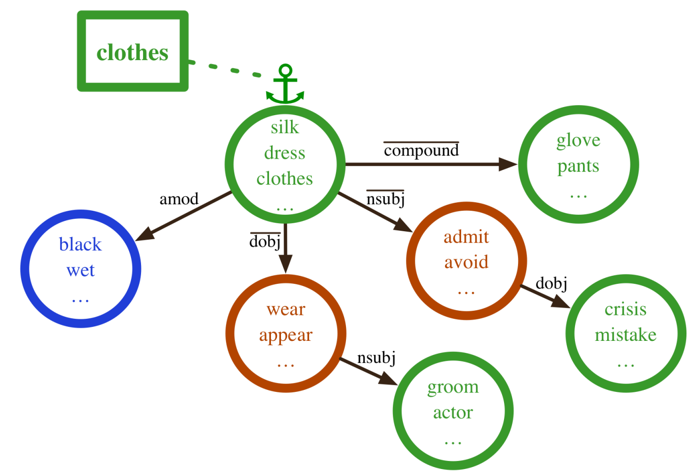
What are APTs?





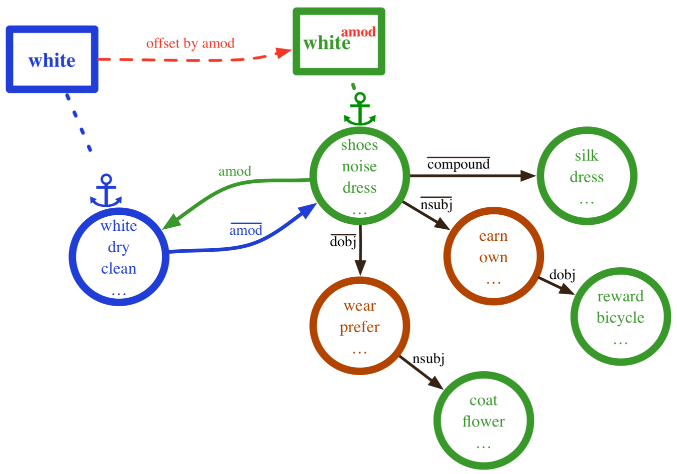
What are APTs?

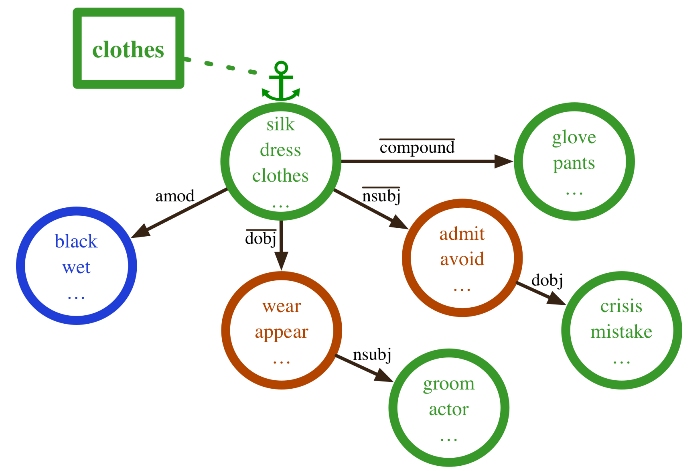




white	white$^{\text{amod}}$	clothes
:clean	amod:clean	amod:wet
$\overline{\text{amod}}$:shoes	:shoes	:dress
$\overline{\text{amod}}$.$\overline{\text{dobj}}$:wear	$\overline{\text{dobj}}$:wear	$\overline{\text{dobj}}$:wear
$\overline{\text{amod}}$.$\overline{\text{dobj}}$.nsubj:coat	$\overline{\text{dobj}}$.nsubj:coat	$\overline{\text{dobj}}$.nsubj:actor

What are APTs?

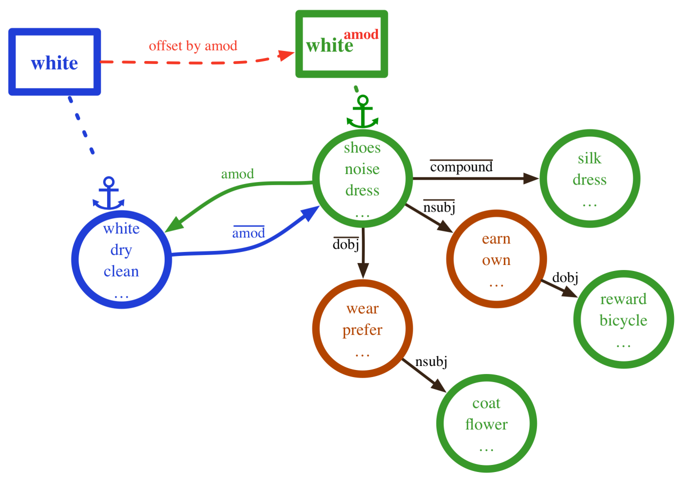


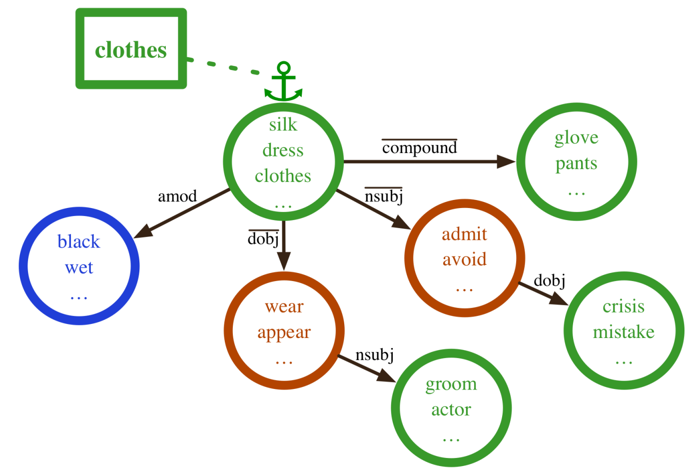


white	white$^{\text{amod}}$	clothes
:clean	amod:clean	amod:wet
$\overline{\text{amod}}$:shoes	:shoes	:dress
$\overline{\text{amod}}$.$\overline{\text{dobj}}$:wear	$\overline{\text{dobj}}$:wear	$\overline{\text{dobj}}$:wear
$\overline{\text{amod}}$.$\overline{\text{dobj}}$.nsubj:coat	$\overline{\text{dobj}}$.nsubj:coat	$\overline{\text{dobj}}$.nsubj:actor

Offset view - aligned with *clothes*

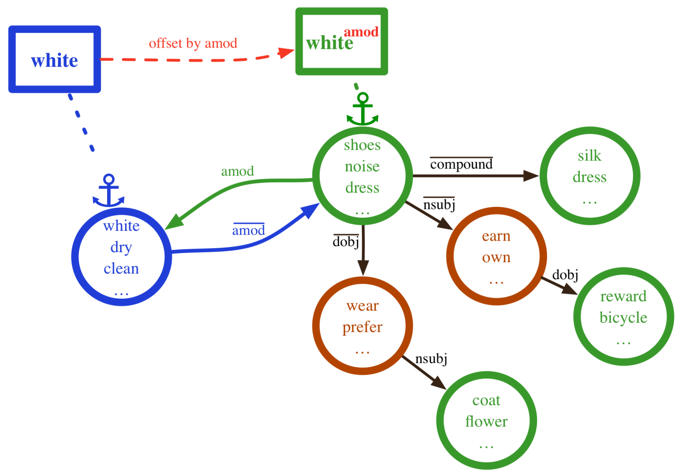
What are APTs?

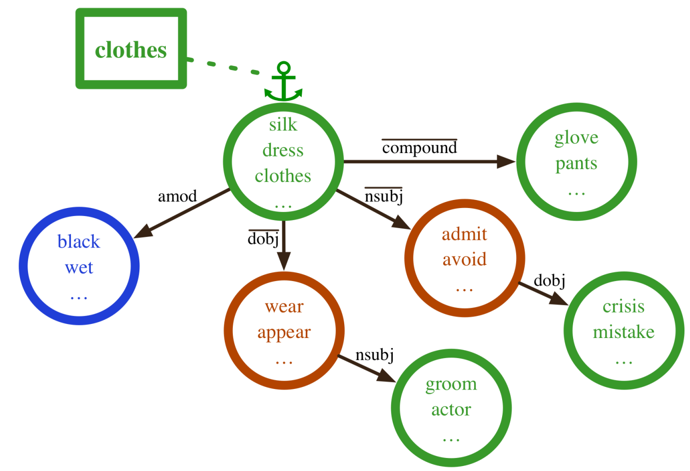




white	white$^{\text{amod}}$	clothes
:clean	amod:clean	amod:wet
$\overline{\text{amod}}$:shoes	:shoes	:dress
$\overline{\text{amod}}$.$\overline{\text{dobj}}$:wear	$\overline{\text{dobj}}$:wear	$\overline{\text{dobj}}$:wear
$\overline{\text{amod}}$.$\overline{\text{dobj}}$.nsubj:coat	$\overline{\text{dobj}}$.nsubj:coat	$\overline{\text{dobj}}$.nsubj:actor

What are APTs?





white	white $^{\texttt{amod}}$	clothes
:clean	amod:clean	amod:wet
$\overline{\texttt{amod}}$:shoes	:shoes	:dress
$\overline{\texttt{amod}}$.$\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{amod}}$.$\overline{\texttt{dobj}}$.nsubj:coat	$\overline{\texttt{dobj}}$.nsubj:coat	$\overline{\texttt{dobj}}$.nsubj:actor

Paths now aligned \o/!

What are APTs?

What are APTs?

- Can now compose the two aligned APTs

What are APTs?

- Can now compose the two aligned APTs
- Either by taking the intersection or the union of their aligned features

What are APTs?

- Can now compose the two aligned APTs
- Either by taking the intersection or the union of their aligned features
    - PPMI weights associated with distributional features can be combined in the usual ways (min, max, point wise addition/multiplication, etc)

What are APTs?

- Can now compose the two aligned APTs
- Either by taking the intersection or the union of their aligned features
    - PPMI weights associated with distributional features can be combined in the usual ways (min, max, point wise addition/multiplication, etc)
    - Composition is not commutative (due to offsetting and taking syntax into account)

What are APTs?

- Can now compose the two aligned APTs
- Either by taking the intersection or the union of their aligned features
    - PPMI weights associated with distributional features can be combined in the usual ways (min, max, point wise addition/multiplication, etc)
    - Composition is not commutative (due to offsetting and taking syntax into account)

white clothes	
Composition by union	*Composition by intersection*
$\overline{\texttt{amod}}$:clean	
$\overline{\texttt{amod}}$:wet	
:shoes	
:dress	
$\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{dobj}}\texttt{.nsubj:}$coat	
$\overline{\texttt{dobj}}\texttt{.nsubj}$:actor	

What are APTs?

- Can now compose the two aligned APTs
- Either by taking the intersection or the union of their aligned features
    - PPMI weights associated with distributional features can be combined in the usual ways (min, max, point wise addition/multiplication, etc)
    - Composition is not commutative (due to offsetting and taking syntax into account)

Composed APT treated as a noun

white clothes	
Composition by union	*Composition by intersection*
$\overline{\texttt{amod}}$:clean	
$\overline{\texttt{amod}}$:wet	
:shoes	
:dress	
$\overline{\texttt{dobj}}$:wear	$\overline{\texttt{dobj}}$:wear
$\overline{\texttt{dobj}}\texttt{.nsubj}$:coat	
$\overline{\texttt{dobj}}\texttt{.nsubj}$:actor	

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

Outline

- Introduction to Anchored Packed Trees
- **Evaluating APTs - A first attempt :((((**
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

Evaluation - A first attempt

Evaluation - A first attempt

- Using standard lexical and phrasal datasets

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
  - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
  - **MEN** (Bruni et al., 2012), containing 3000 word pairs
  - **SimLex-999** (Hill et al., 2015), containing 999 word pairs

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs
    - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
    - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs
    - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
    - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)
- Comparing human similarity ratings between words or phrases, to model similarity estimates by calculating Spearman's ρ

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs
    - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
    - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)
- Comparing human similarity ratings between words or phrases, to model similarity estimates by calculating Spearman's ρ
    - money - cash: 0.91

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs
    - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
    - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)
- Comparing human similarity ratings between words or phrases, to model similarity estimates by calculating Spearman's ρ
    - money - cash: 0.91
    - forest - graveyard: 0.19

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs
    - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
    - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)
- Comparing human similarity ratings between words or phrases, to model similarity estimates by calculating Spearman's ρ
    - money - cash: 0.91
    - forest - graveyard: 0.19
    - vast amount - large quantity: 0.96

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
  - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
  - **MEN** (Bruni et al., 2012), containing 3000 word pairs
  - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
  - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)
- Comparing human similarity ratings between words or phrases, to model similarity estimates by calculating Spearman's ρ
  - money - cash: 0.91
  - forest - graveyard: 0.19
  - vast amount - large quantity: 0.96
  - little room - similar result: 0.17

Evaluation - A first attempt

- Using standard lexical and phrasal datasets
    - **WS353** (Finkelstein et al., 2001), containing 353 word pairs; using the similarity/relatedness split of Agirre et al., (2009)
    - **MEN** (Bruni et al., 2012), containing 3000 word pairs
    - **SimLex-999** (Hill et al., 2015), containing 999 word pairs
    - **ML2010** (Mitchell & Lapata, 2010), containing 108 adjective-noun, 108 noun-noun, and 108 verb-object pairs (324 phrase pairs in total)
- Comparing human similarity ratings between words or phrases, to model similarity estimates by calculating Spearman's ρ
    - money - cash: 0.91
    - forest - graveyard: 0.19
    - vast amount - large quantity: 0.96
    - little room - similar result: 0.17
- Vectorised **order 2** APT space from the BNC, using **PPMI** as lexical association function

Evaluation - A first attempt

Evaluation - A first attempt

Dataset
WS353 (Sim)
WS353 (Rel)
MEN
SimLex-999
ML10 - AN
ML10 - NN
ML10 - VO

Evaluation - A first attempt

Dataset	word2vec*
WS353 (Sim)	0.64
WS353 (Rel)	0.42
MEN	0.63
SimLex-999	0.25
ML10 - AN	0.50
ML10 - NN	0.47
ML10 - VO	0.42

Evaluation - A first attempt

Dataset	**word2vec***
WS353 (Sim)	0.64
WS353 (Rel)	0.42
MEN	0.63
SimLex-999	0.25
ML10 - AN	0.50
ML10 - NN	0.47
ML10 - VO	0.42

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

Evaluation - A first attempt

Dataset	word2vec*	APTs
WS353 (Sim)	0.64	0.40
WS353 (Rel)	0.42	0.24
MEN	0.63	0.36
SimLex-999	0.25	0.22
ML10 - AN	0.50	0.39
ML10 - NN	0.47	0.41
ML10 - VO	0.42	0.35

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

Evaluation - A first attempt

Dataset	word2vec*	APTs	APTs tuned
WS353 (Sim)	0.64	0.40	0.52
WS353 (Rel)	0.42	0.24	0.35
MEN	0.63	0.36	0.43
SimLex-999	0.25	0.22	0.25
ML10 - AN	0.50	0.39	0.39
ML10 - NN	0.47	0.41	0.43
ML10 - VO	0.42	0.35	0.36

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

Evaluation - A first attempt

Primarily interested in the composition tasks

Dataset	word2vec*	APTs	APTs tuned
WS353 (Sim)	0.64	0.40	0.52
WS353 (Rel)	0.42	0.24	0.35
MEN	0.63	0.36	0.43
SimLex-999	0.25	0.22	0.25
ML10 - AN	0.50	0.39	0.39
ML10 - NN	0.47	0.41	0.43
ML10 - VO	0.42	0.35	0.36

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

Evaluation - A first attempt

Primarily interested in the composition tasks

Dataset	word2vec*	APTs	APTs tuned
WS353 (Sim)	0.64	0.40	0.52
WS353 (Rel)	0.42	0.24	0.35
MEN	0.63	0.36	0.43
SimLex-999	0.25	0.22	0.25
ML10 - AN	0.50	0.39	0.39
ML10 - NN	0.47	0.41	0.43
ML10 - VO	0.42	0.35	0.36

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

- The results are…well…*pretty underwhelming*

Evaluation - A first attempt

Primarily interested in the composition tasks

Dataset	word2vec*	APTs	APTs tuned
WS353 (Sim)	0.64	0.40	0.52
WS353 (Rel)	0.42	0.24	0.35
MEN	0.63	0.36	0.43
SimLex-999	0.25	0.22	0.25
ML10 - AN	0.50	0.39	0.39
ML10 - NN	0.47	0.41	0.43
ML10 - VO	0.42	0.35	0.36

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

- The results are…well…*pretty underwhelming*

Evaluation - A first attempt

Primarily interested in the composition tasks

Dataset	word2vec*	APTs	APTs tuned
WS353 (Sim)	0.64	0.40	0.52
WS353 (Rel)	0.42	0.24	0.35
MEN	0.63	0.36	0.43
SimLex-999	0.25	0.22	0.25
ML10 - AN	0.50	0.39	0.39
ML10 - NN	0.47	0.41	0.43
ML10 - VO	0.42	0.35	0.36

*) using 50dim pre-trained word vectors from the BNC (Hashimoto et al., 2014)

- The results are…well…*pretty underwhelming*

- Nice theory, but doesn't quite work out of the box - whats the problem?

So whats the problem?

So whats the problem?

- APTs are extremely sparse

So whats the problem?

- APTs are extremely sparse
    - Vectorised space of an APT model derived from the BNC has ~820k dimensions, the density of the co-occurrence matrix is 0.00058 ("half a per mill")

So whats the problem?

- APTs are extremely sparse
    - Vectorised space of an APT model derived from the BNC has ~820k dimensions, the density of the co-occurrence matrix is 0.00058 ("half a per mill")
- Due to modelling the dependency relation in a co-occurrence, the sparsity effect is amplified

So whats the problem?

- APTs are extremely sparse
    - Vectorised space of an APT model derived from the BNC has ~820k dimensions, the density of the co-occurrence matrix is 0.00058 ("half a per mill")
- Due to modelling the dependency relation in a co-occurrence, the sparsity effect is amplified
    - For example *fish* as object of *eat* and *fish* as subject of *eat* are modelled as two distinct contexts

So whats the problem?

- APTs are extremely sparse
    - Vectorised space of an APT model derived from the BNC has ~820k dimensions, the density of the co-occurrence matrix is 0.00058 ("half a per mill")
- Due to modelling the dependency relation in a co-occurrence, the sparsity effect is amplified
    - For example *fish* as object of *eat* and *fish* as subject of *eat* are modelled as two distinct contexts
- As a consequence, so is the “curse of dimensionality”, as there are fewer observations per dimension in the data

So whats the problem?

- APTs are extremely sparse
    - Vectorised space of an APT model derived from the BNC has ~820k dimensions, the density of the co-occurrence matrix is 0.00058 ("half a per mill")
- Due to modelling the dependency relation in a co-occurrence, the sparsity effect is amplified
    - For example *fish* as object of *eat* and *fish* as subject of *eat* are modelled as two distinct contexts
- As a consequence, so is the “curse of dimensionality”, as there are fewer observations per dimension in the data
- Are the representations too sparse to be useful?

So whats the problem?

So whats the problem?

- Evaluating semantic space using BLESS (Baroni & Lenci, 2011)

So whats the problem?

- Evaluating semantic space using BLESS (Baroni & Lenci, 2011)
- Compare 200 concrete nouns to a number of different relata, including hypernyms, co-hyponyms, meronyms, attributes (adjectives), events (verbs), and random lexemes for each PoS (NN, JJ, VB)

So whats the problem?

- Evaluating semantic space using BLESS (Baroni & Lenci, 2011)
- Compare 200 concrete nouns to a number of different relata, including hypernyms, co-hyponyms, meronyms, attributes (adjectives), events (verbs), and random lexemes for each PoS (NN, JJ, VB)
- Create box plot of the distribution of similarities per relation type - illustrates the bias towards any relation type in the distributional space

So whats the problem?

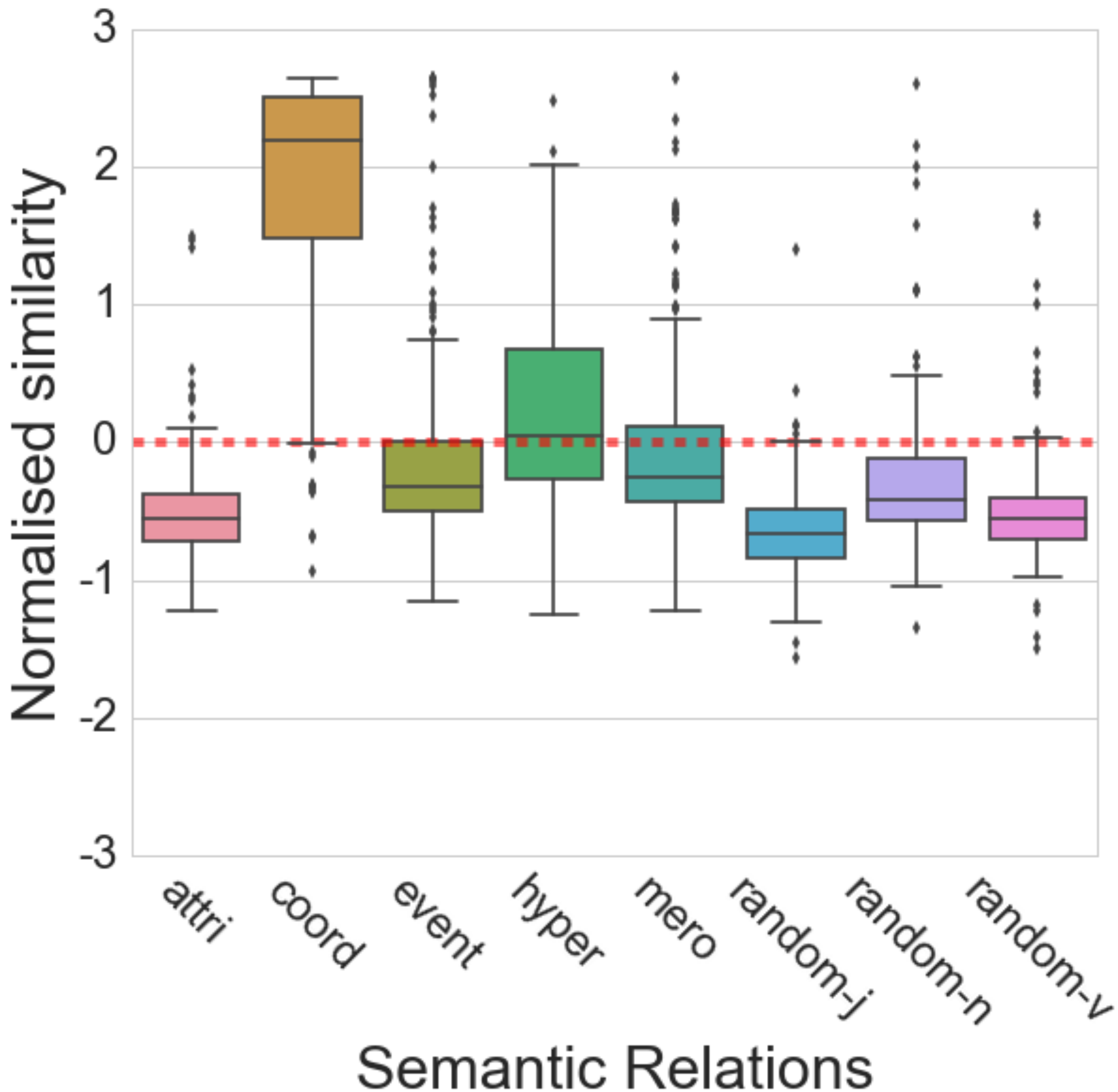
- Evaluating semantic space using BLESS (Baroni & Lenci, 2011)
- Compare 200 concrete nouns to a number of different relata, including hypernyms, co-hyponyms, meronyms, attributes (adjectives), events (verbs), and random lexemes for each PoS (NN, JJ, VB)
- Create box plot of the distribution of similarities per relation type - illustrates the bias towards any relation type in the distributional space
- Previous results found that typed DSMs have a bias towards co-hyponyms and hypernyms (Peirsman, 2008; Baroni & Lenci, 2011, Levy & Goldberg, 2014)

So whats the problem?

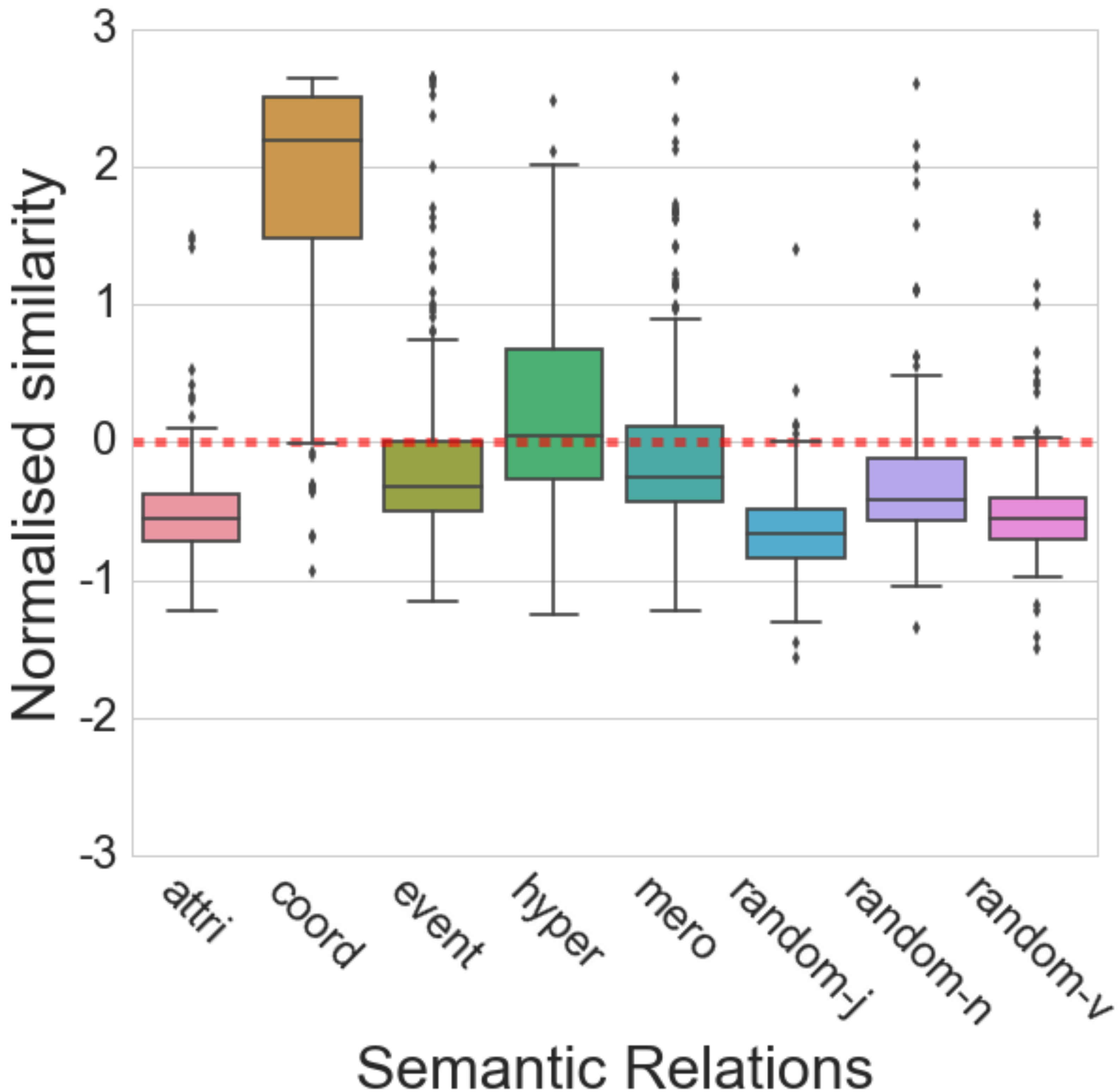
- Evaluating semantic space using BLESS (Baroni & Lenci, 2011)
- Compare 200 concrete nouns to a number of different relata, including hypernyms, co-hyponyms, meronyms, attributes (adjectives), events (verbs), and random lexemes for each PoS (NN, JJ, VB)
- Create box plot of the distribution of similarities per relation type - illustrates the bias towards any relation type in the distributional space
- Previous results found that typed DSMs have a bias towards co-hyponyms and hypernyms (Peirsman, 2008; Baroni & Lenci, 2011, Levy & Goldberg, 2014)
- If the APT space is too sparse to represent anything meaningful, we would expect to see (more or less) a uniform similarity distribution across all semantic relations

So whats the problem?

So whats the problem?

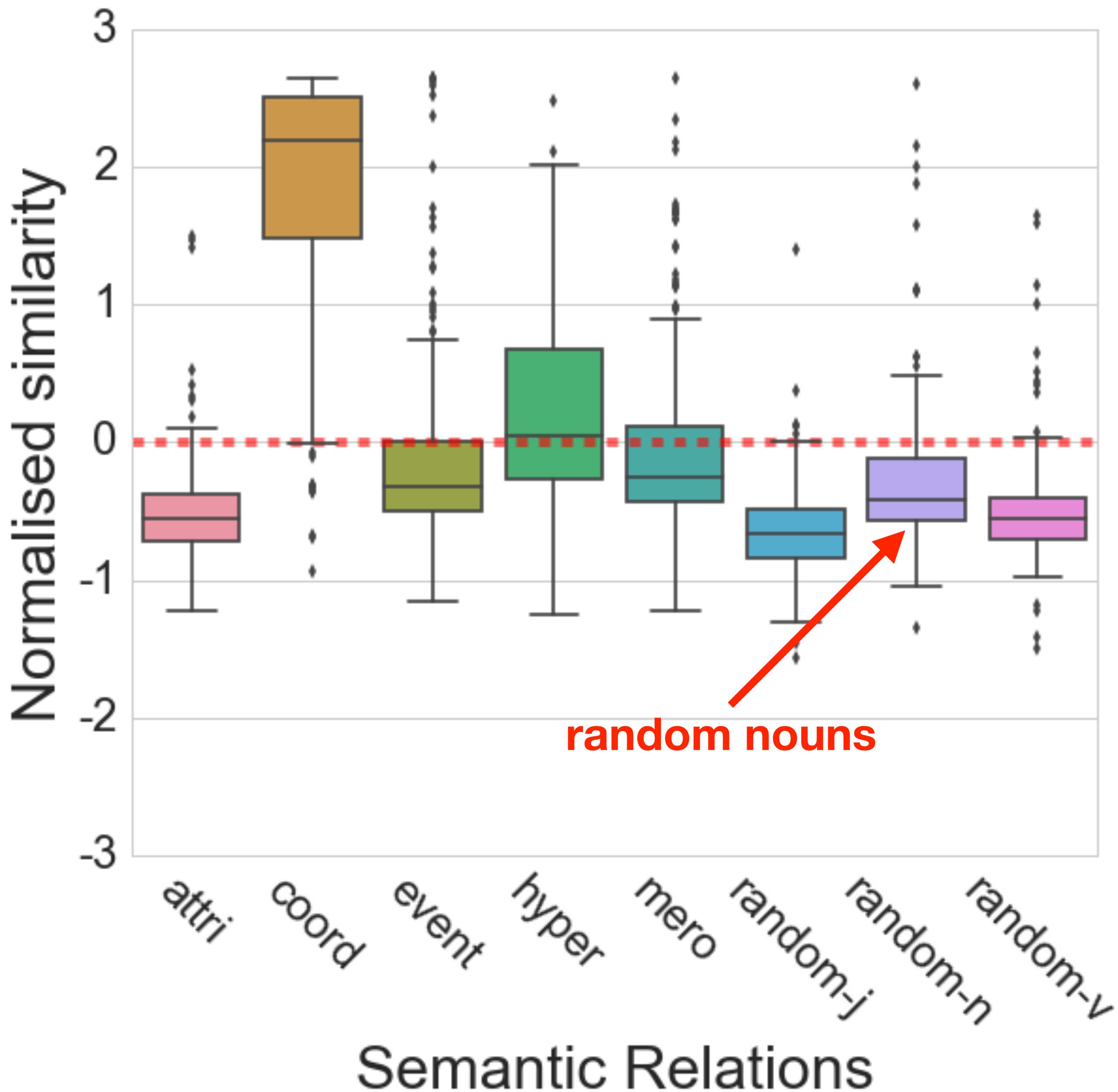


So whats the problem?



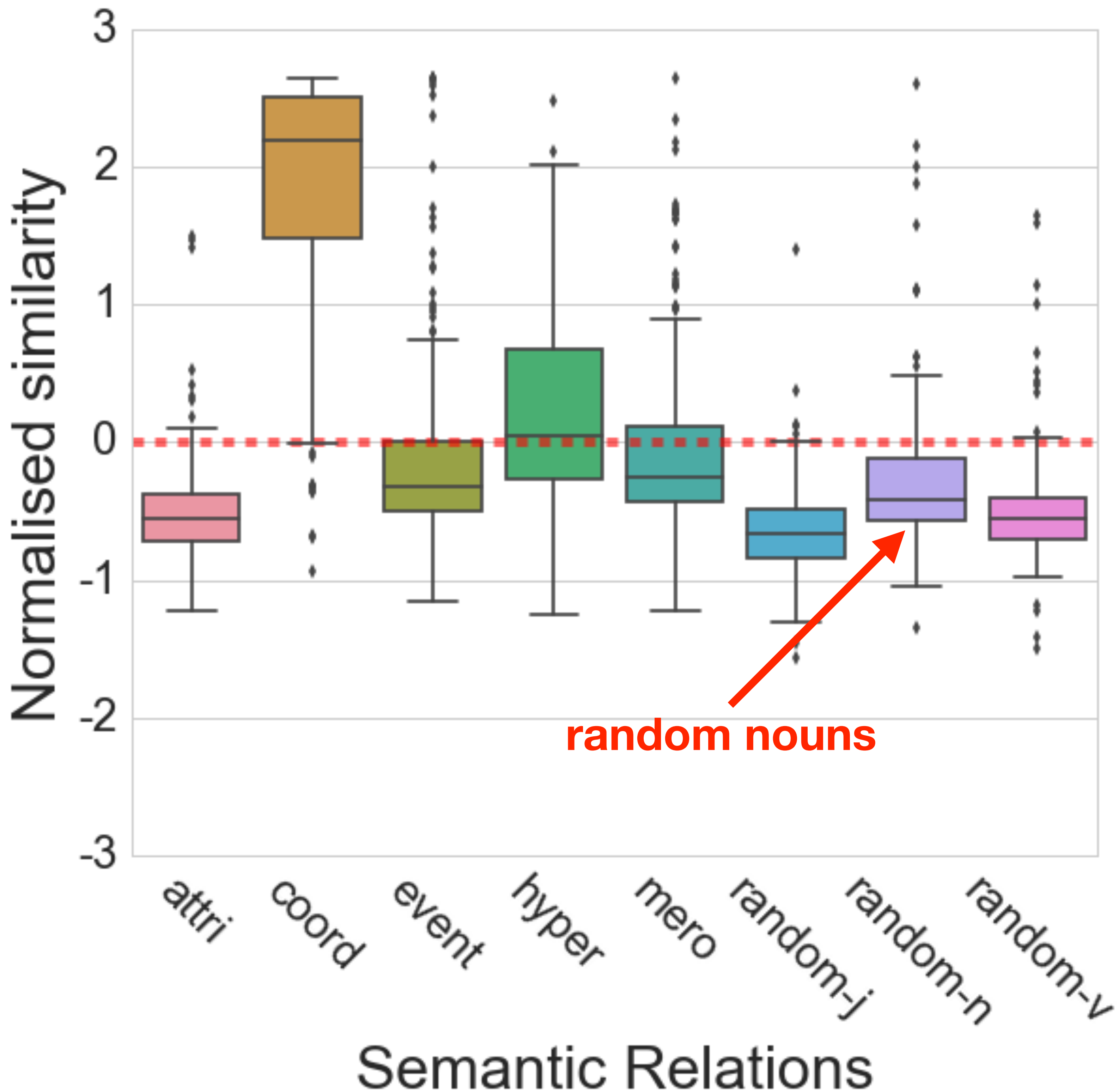
- Not so random really

So whats the problem?



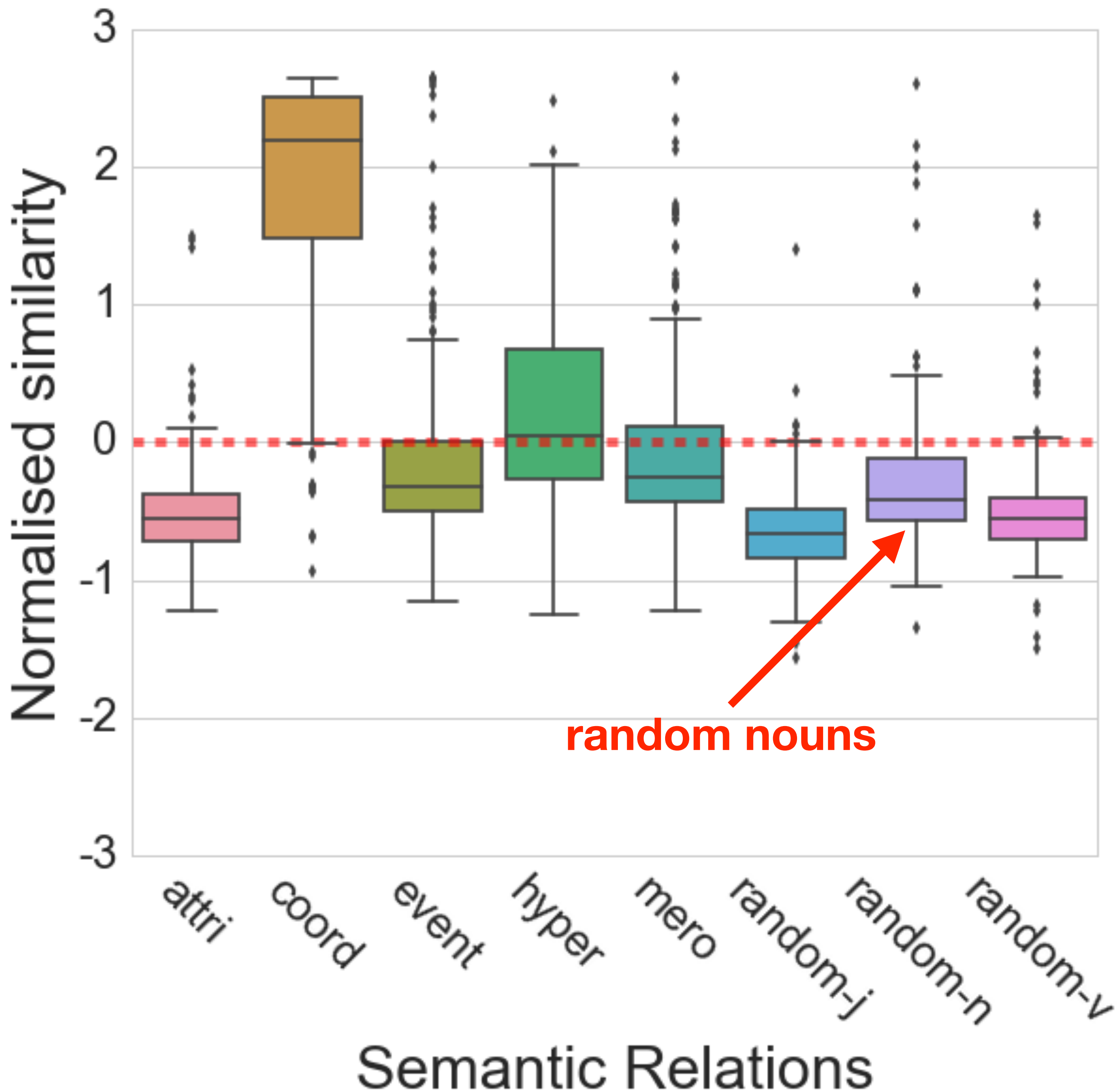
- Not so random really

So whats the problem?



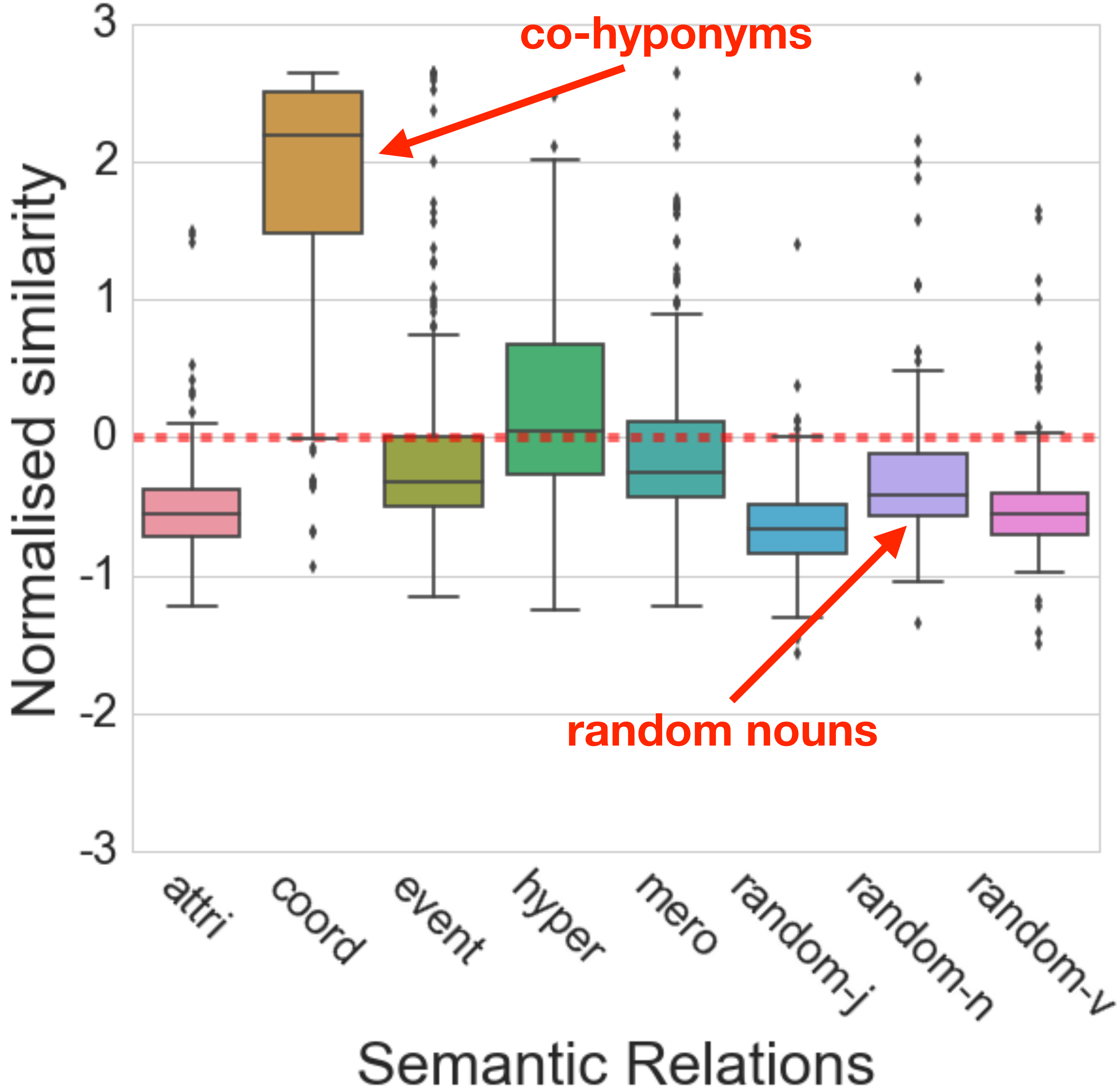
- Not so random really
- Results follow previous findings for typed DSMs

So whats the problem?



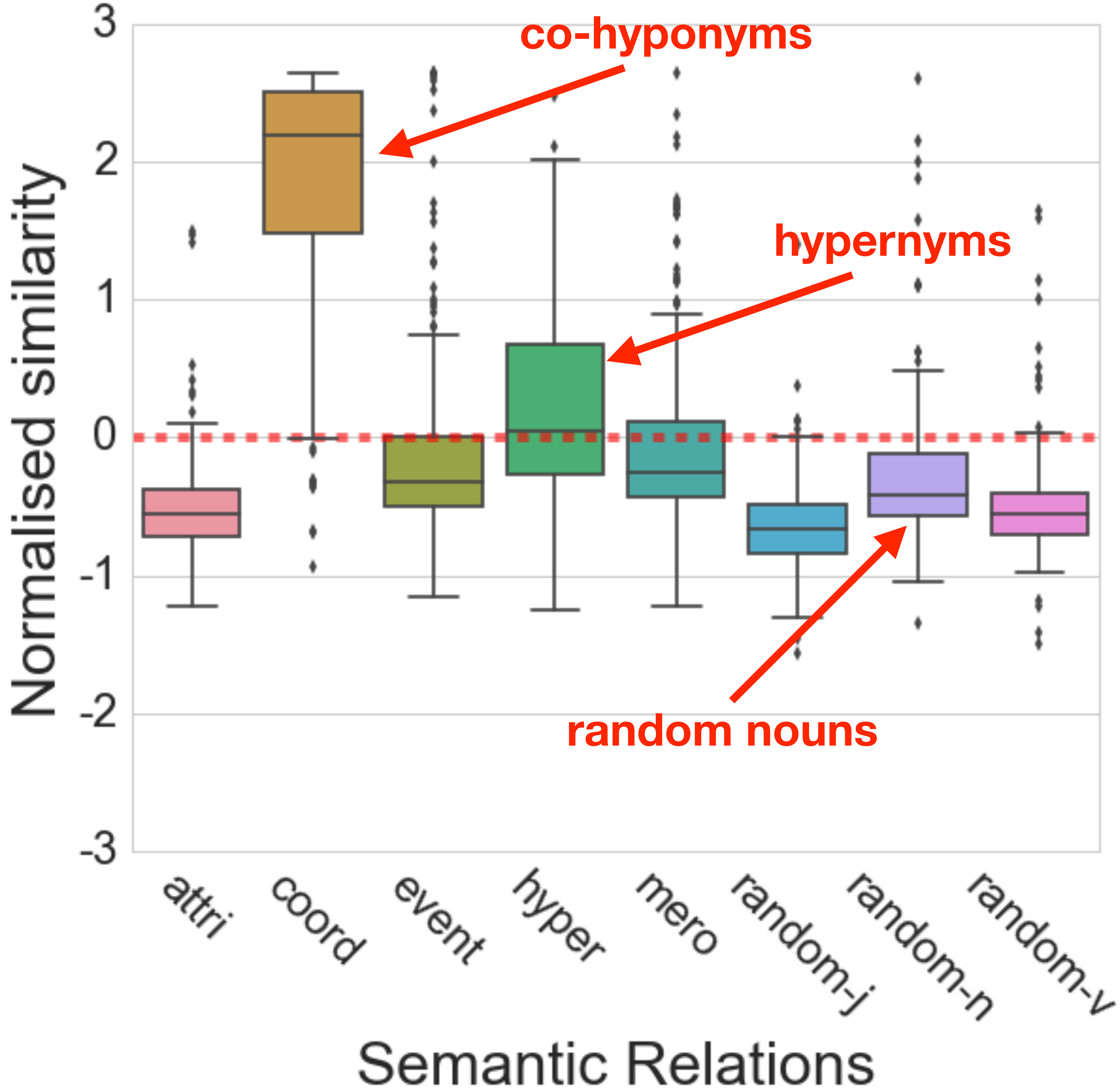
- Not so random really
- Results follow previous findings for typed DSMs
- Distributional space favours co-hyponymy and to a lesser extend hypernymy

So whats the problem?



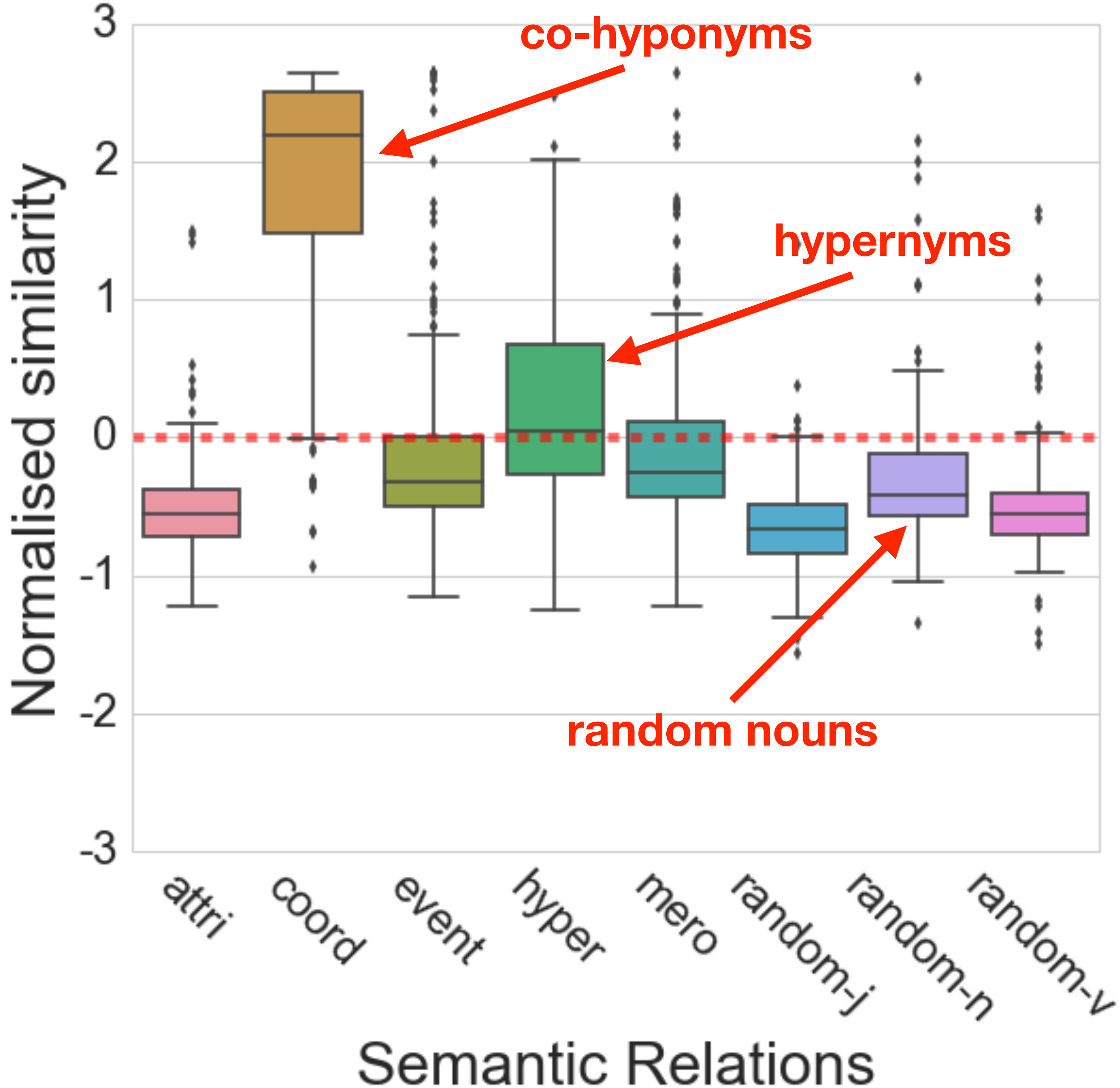
- Not so random really
- Results follow previous findings for typed DSMs
- Distributional space favours co-hyponymy and to a lesser extend hypernymy

So whats the problem?



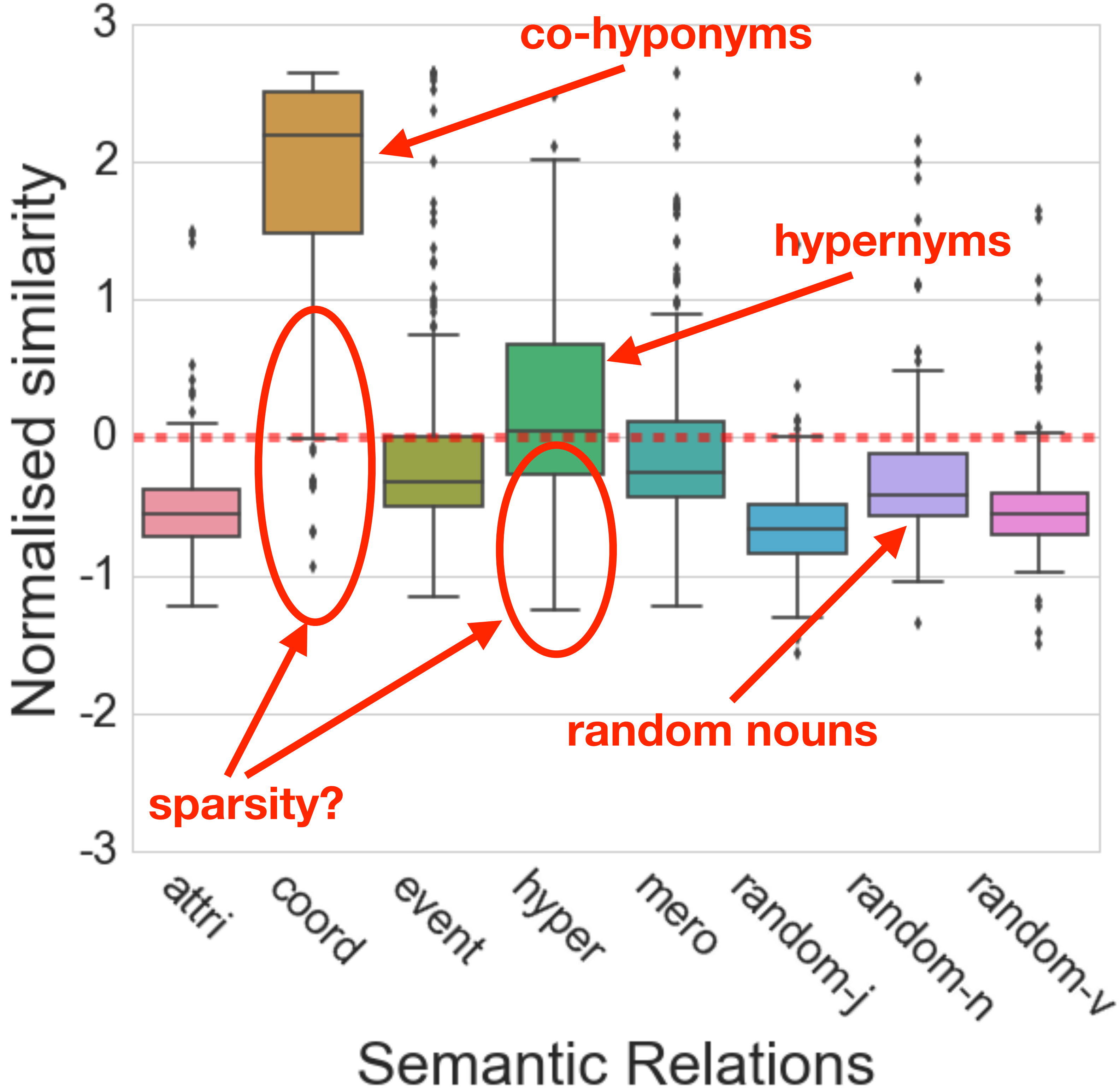
- Not so random really
- Results follow previous findings for typed DSMs
- Distributional space favours co-hyponymy and to a lesser extend hypernymy

So whats the problem?



- Not so random really
- Results follow previous findings for typed DSMs
- Distributional space favours co-hyponymy and to a lesser extend hypernymy
- While its very sparse, the distributional space is still intact

So whats the problem?



- Not so random really
- Results follow previous findings for typed DSMs
- Distributional space favours co-hyponymy and to a lesser extend hypernymy
- While its very sparse, the distributional space is still intact

So what do we do?

So what do we do?

- Cannot use standard dimensionality reduction techniques (e.g. SVD, NMF, …) because distributional composition relies on the explicit structure of the space

So what do we do?

- Cannot use standard dimensionality reduction techniques (e.g. SVD, NMF, …) because distributional composition relies on the explicit structure of the space

- Distributional composition is based on aligning the representations of words with different grammatical roles (e.g. adjectives and nouns) - not obvious/straightforward how to achieve that in a latent space

So what do we do?

- Cannot use standard dimensionality reduction techniques (e.g. SVD, NMF, …) because distributional composition relies on the explicit structure of the space

- Distributional composition is based on aligning the representations of words with different grammatical roles (e.g. adjectives and nouns) - not obvious/straightforward how to achieve that in a latent space

- Instead, leverage the distributional neighbourhood and explicitly infer co-occurrences from similar representations.

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- **Distributional Inference**
- Evaluating APTs - A better attempt :))))
- Conclusion

Distributional Inference

Distributional Inference

- Initial idea based on work by Essen & Steinbiss (1992) and Dagan et al., (1993) for smoothing language models

Distributional Inference

- Initial idea based on work by Essen & Steinbiss (1992) and Dagan et al., (1993) for smoothing language models

- For any lexeme ***w***, calculate its nearest neighbours and add features from the neighbours to ***w***

Distributional Inference

- Initial idea based on work by Essen & Steinbiss (1992) and Dagan et al., (1993) for smoothing language models

- For any lexeme ***w***, calculate its nearest neighbours and add features from the neighbours to ***w***

- ~Soft clustering of the distributional space, every lexeme is represented as the weighted average of its neighbourhood

Distributional Inference

- Initial idea based on work by Essen & Steinbiss (1992) and Dagan et al., (1993) for smoothing language models

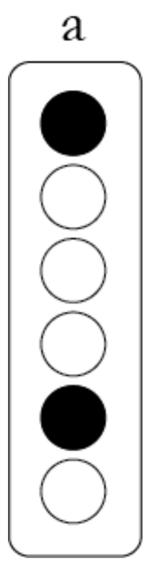
- For any lexeme ***w***, calculate its nearest neighbours and add features from the neighbours to ***w***

- ~Soft clustering of the distributional space, every lexeme is represented as the weighted average of its neighbourhood

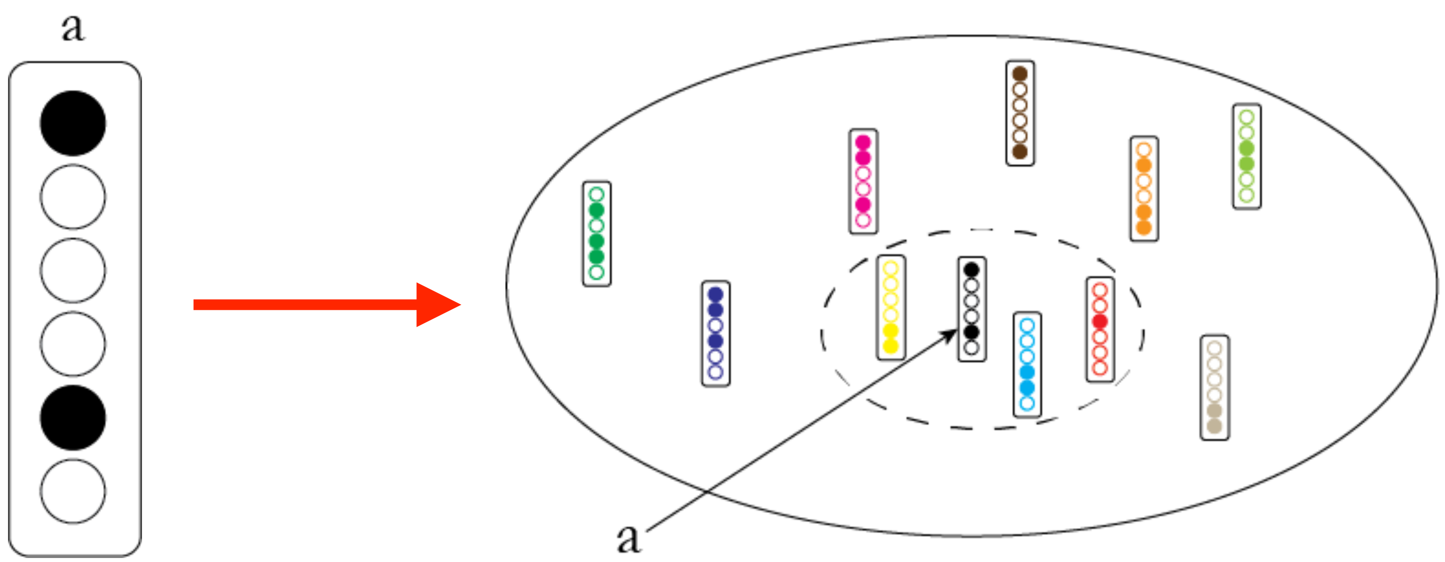
- The algorithm isn't just applicable to APTs but represents a general mechanism for enriching the representations in a sparse space (Kober et al., 2016)

Distributional Inference

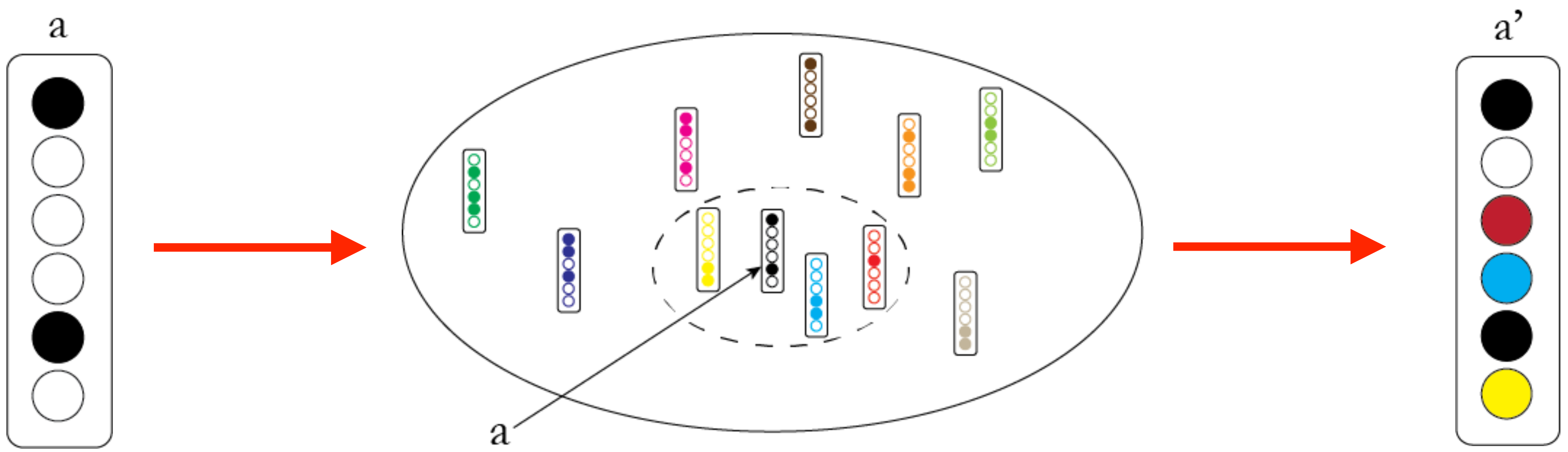
Distributional Inference



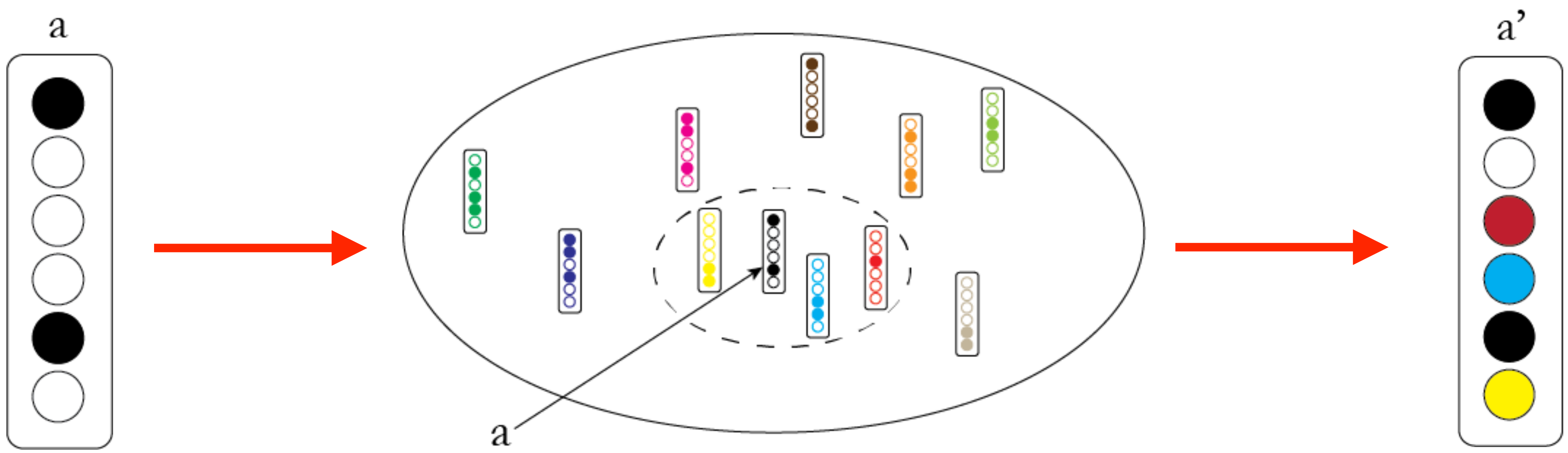
Distributional Inference



Distributional Inference

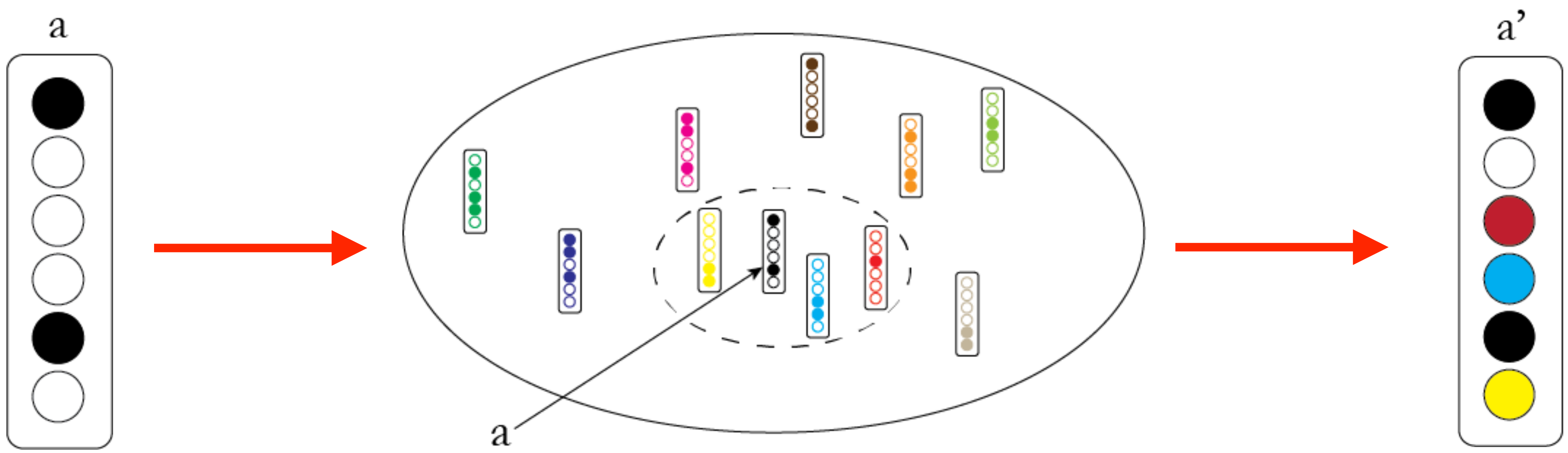


Distributional Inference



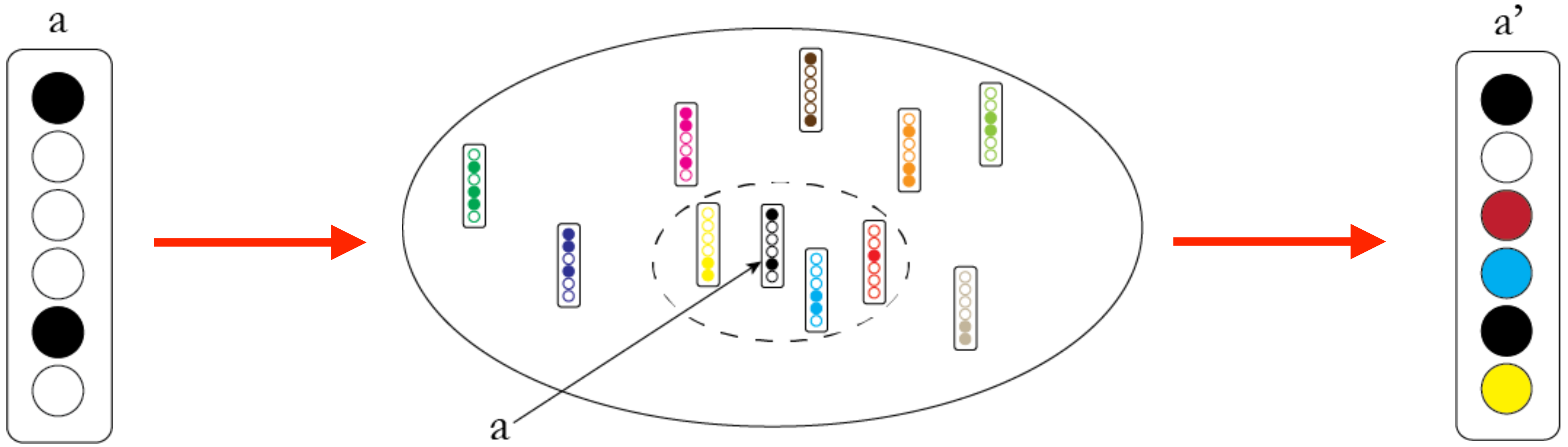
Lexeme

Distributional Inference



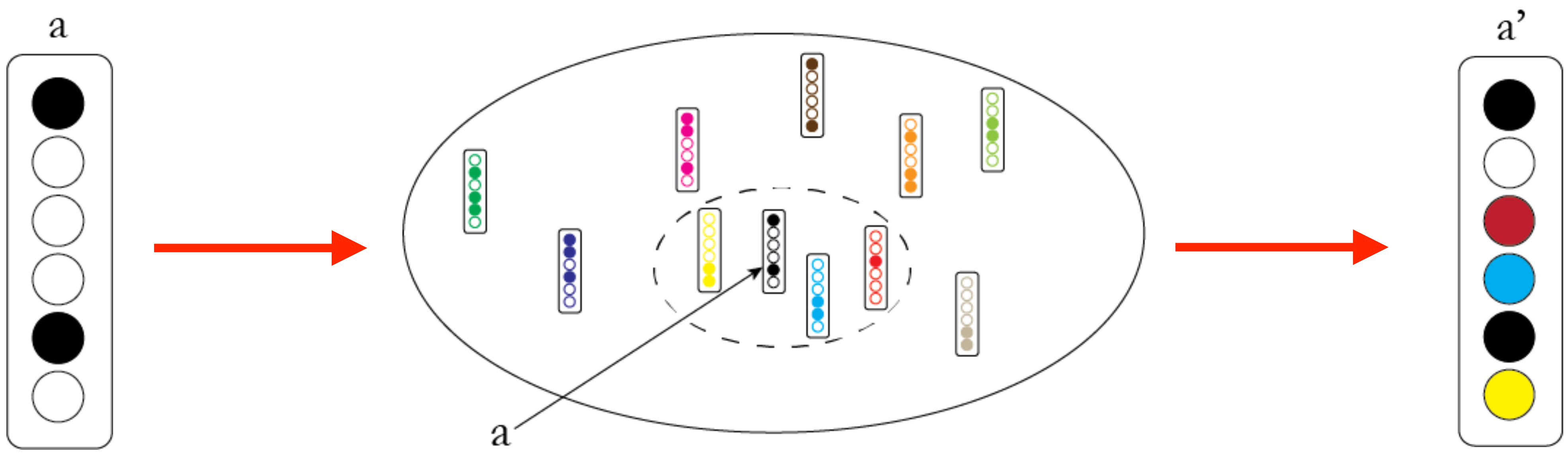
Lexeme	Neighbours

Distributional Inference



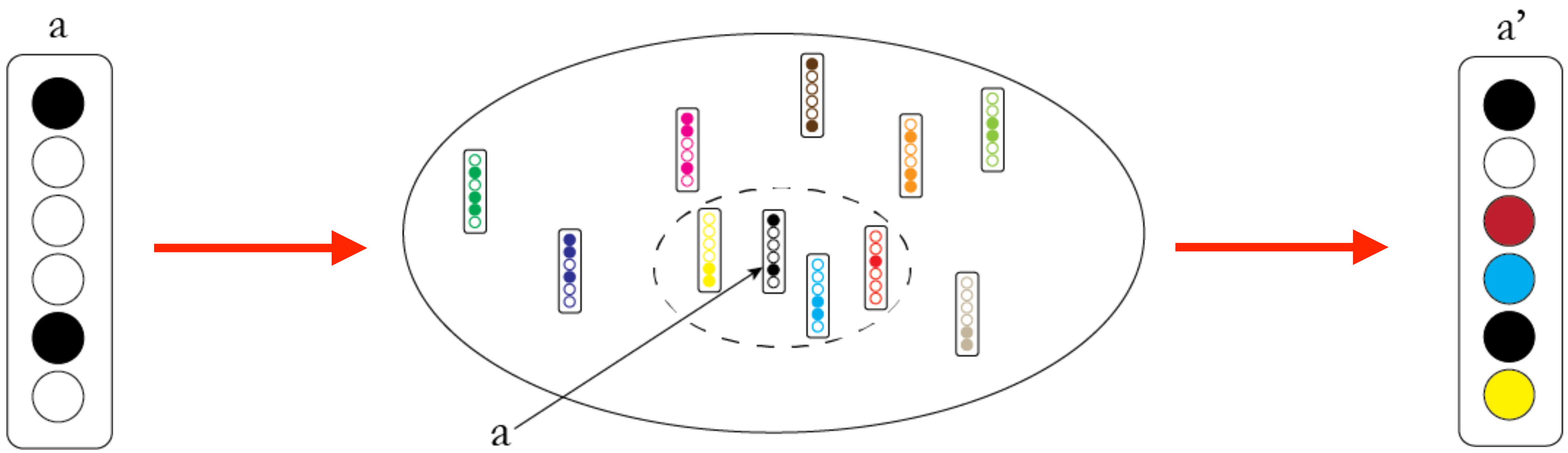
Lexeme	Neighbours	Inferred co-occurrences

Distributional Inference



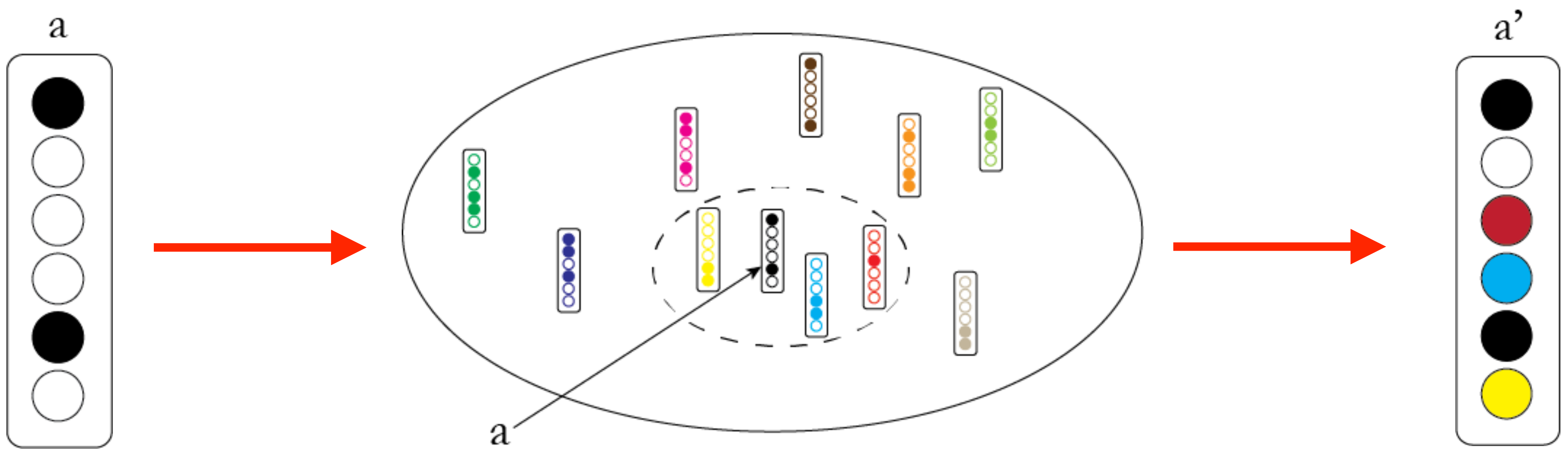
Lexeme	Neighbours	Inferred co-occurrences
magazine		

Distributional Inference



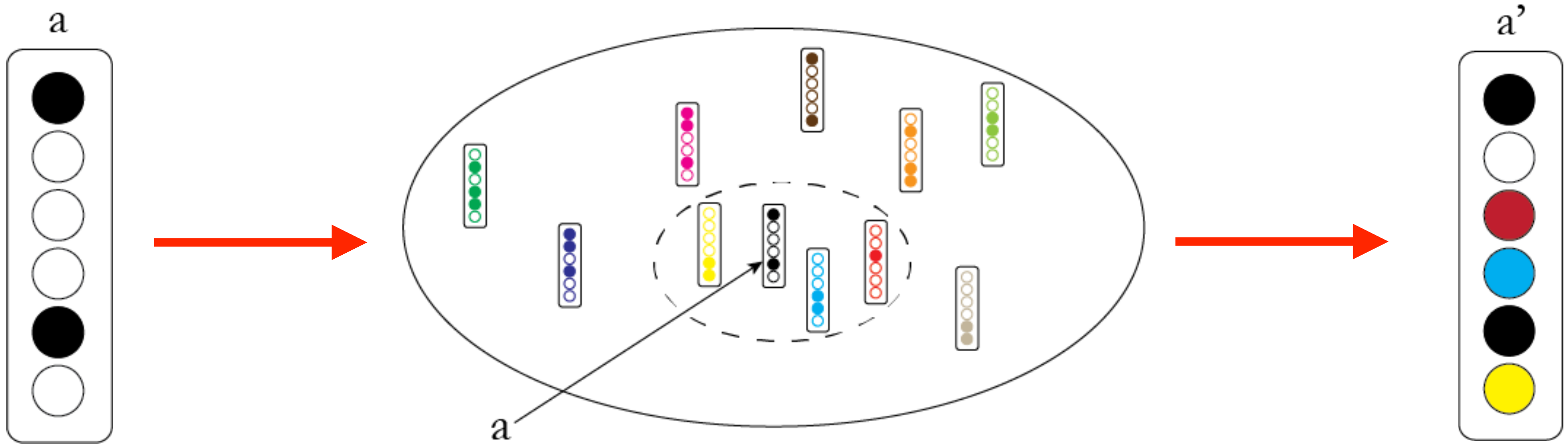
Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	

Distributional Inference



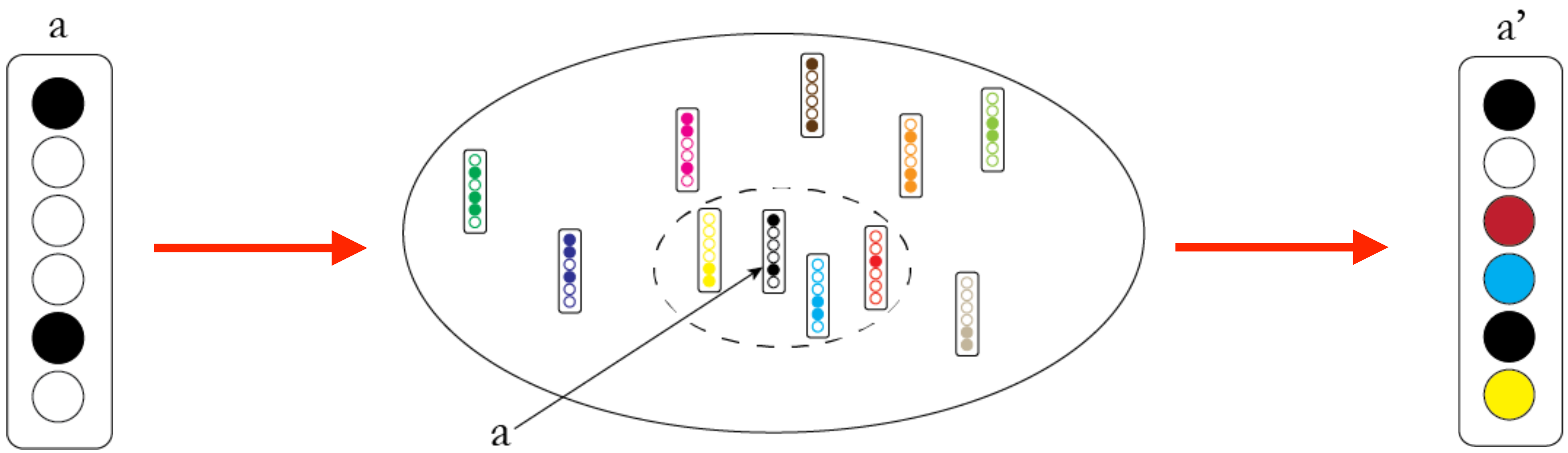
Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	——— ———

Distributional Inference



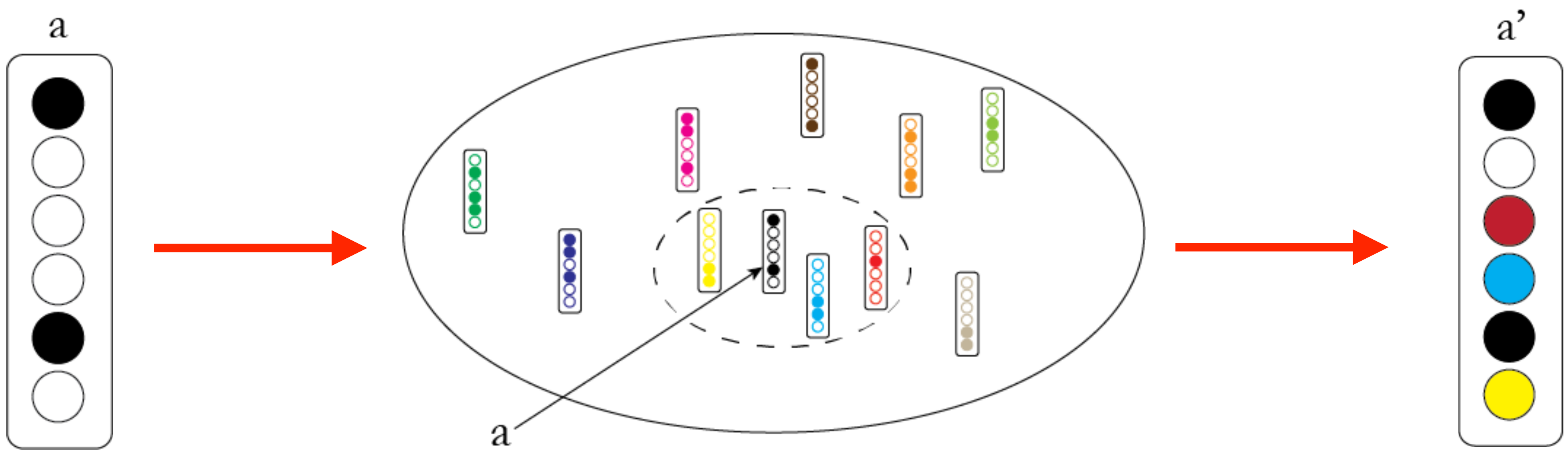
Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$:*sell*, $\overline{\texttt{nsubj}}$:*report*, `amod`:*daily*

Distributional Inference



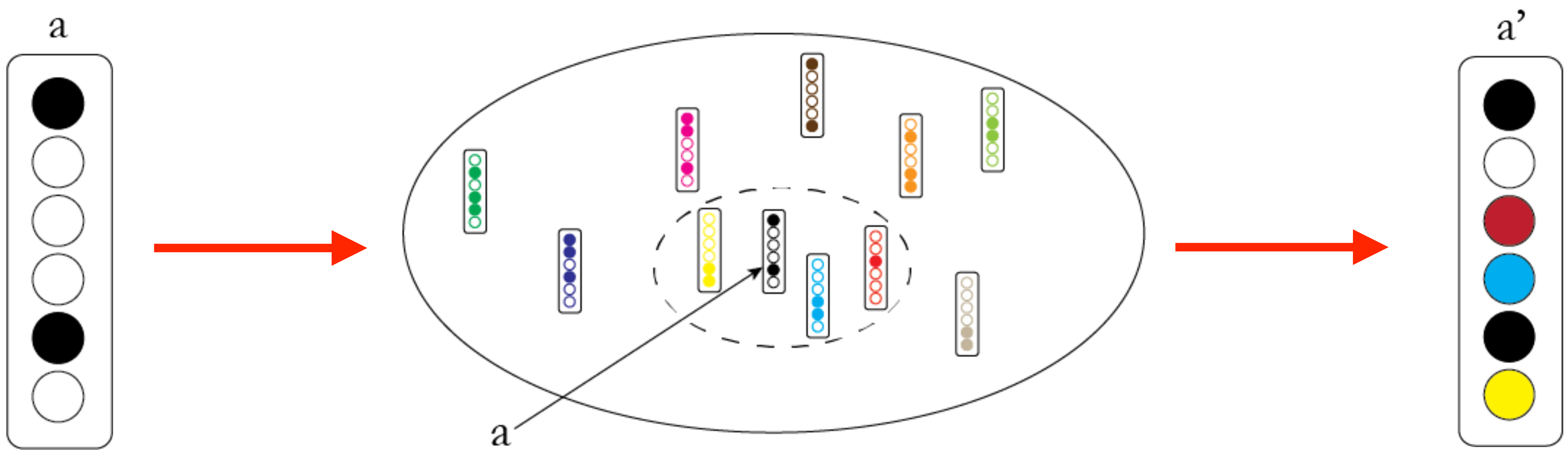
Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$:*sell*, $\overline{\texttt{nsubj}}$:*report*, `amod`:*daily*
cat		

Distributional Inference



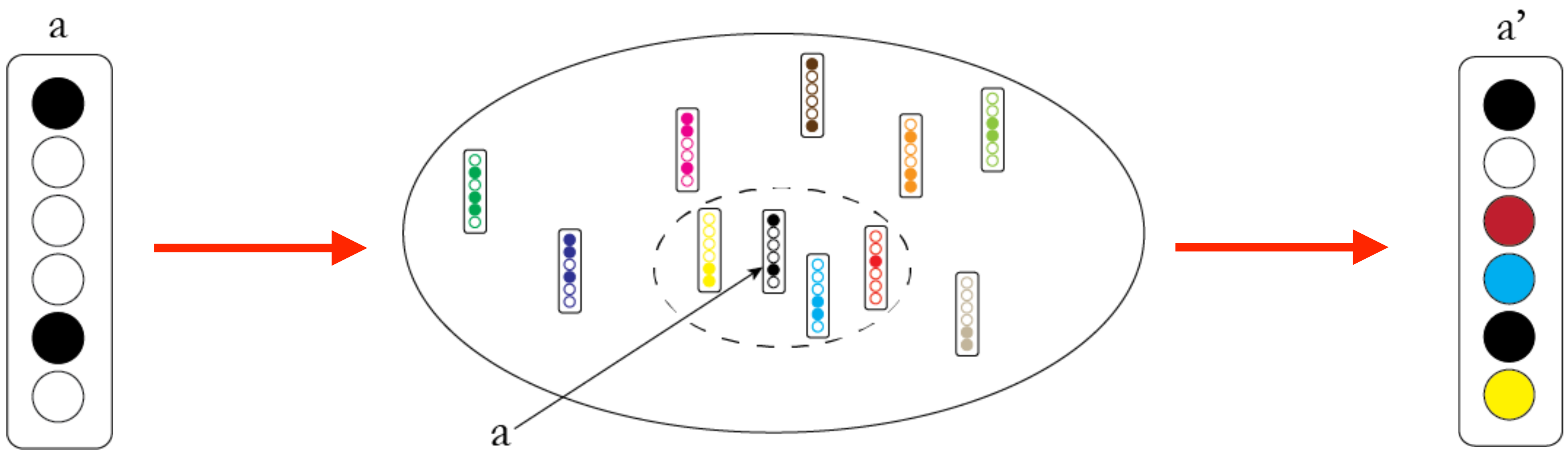
Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	

Distributional Inference



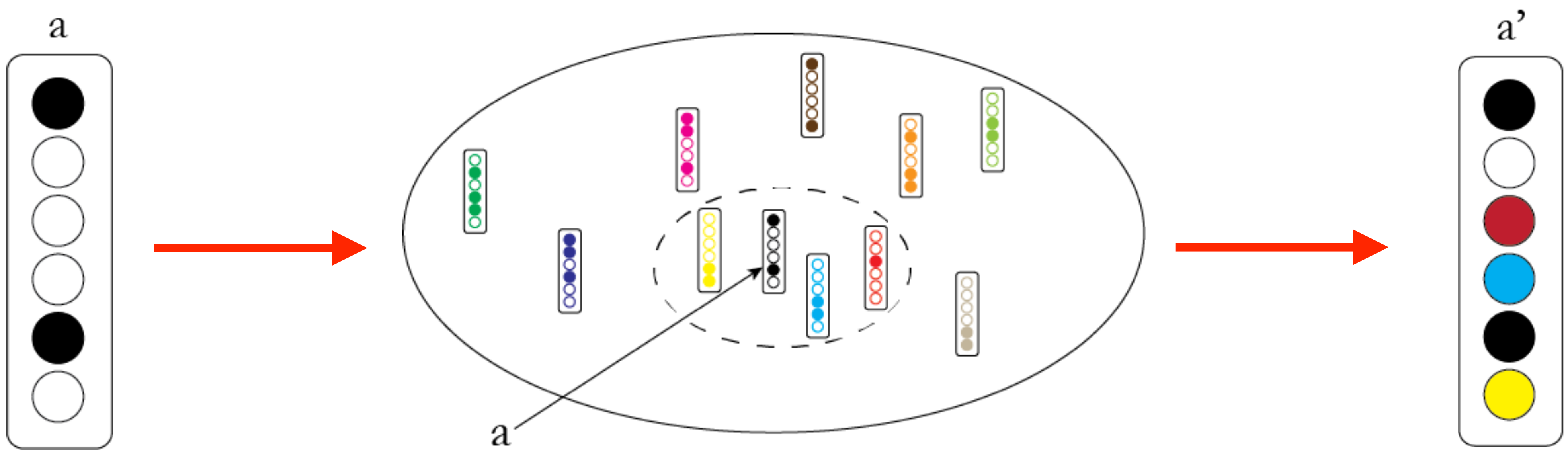
Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$:*sell*, $\overline{\texttt{nsubj}}$:*report*, `amod`:*daily*
cat	*dog*, rabbit, pet	

Distributional Inference



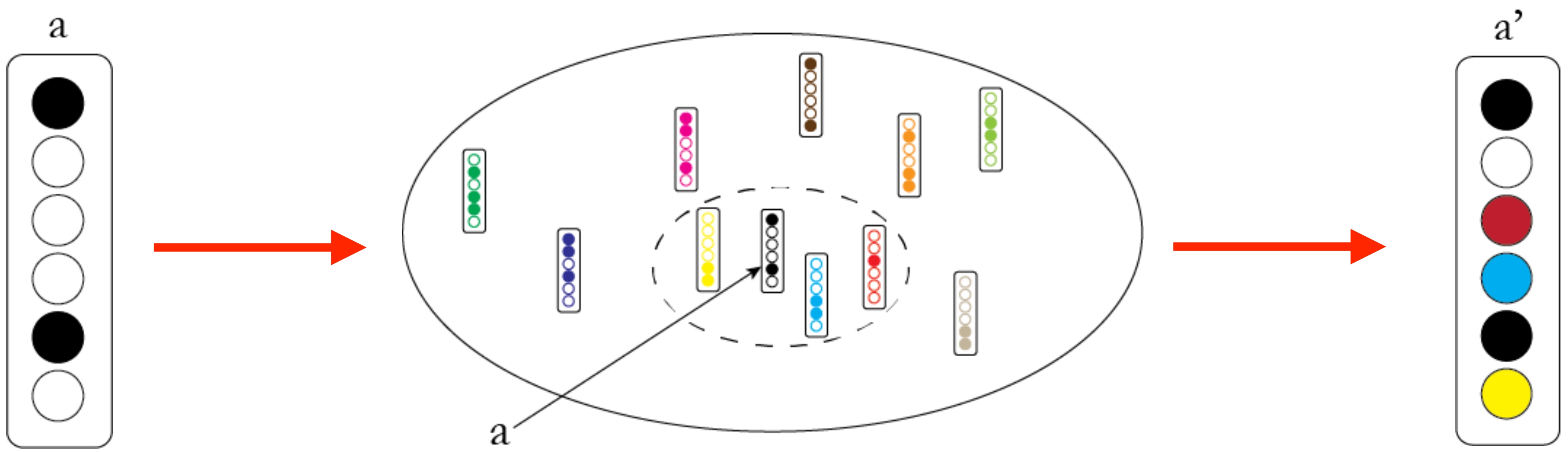
Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$`:`*walk*, $\overline{\texttt{nsubj}}$`:`*bark*, `amod:`*hot*

Distributional Inference



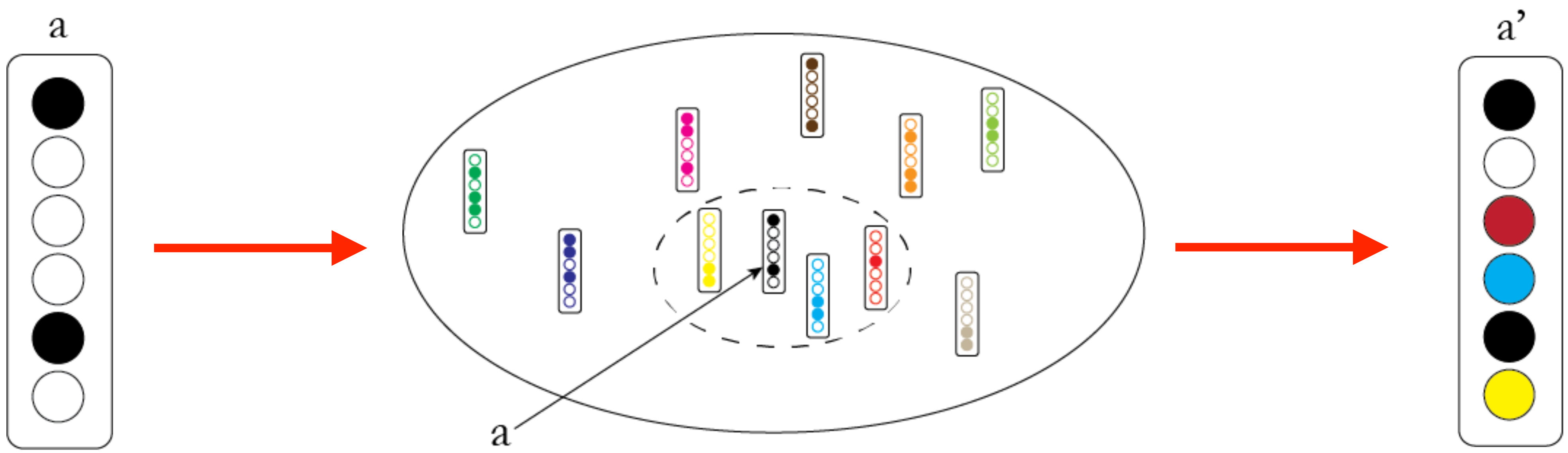
Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$:*sell*, $\overline{\texttt{nsubj}}$:*report*, `amod`:*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$:*walk*, $\overline{\texttt{nsubj}}$:*bark*, `amod`:*hot*
car		

Distributional Inference



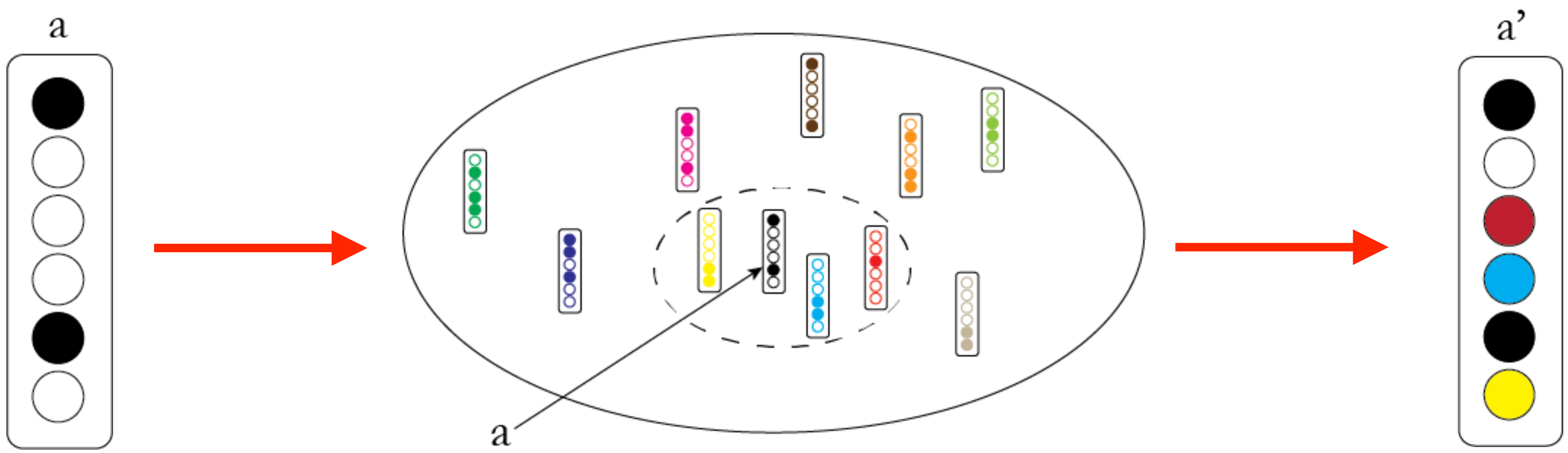
Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$:*sell*, $\overline{\texttt{nsubj}}$:*report*, `amod`:*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$:*walk*, $\overline{\texttt{nsubj}}$:*bark*, `amod`:*hot*
car	*vehicle*, lorry, bus	

Distributional Inference



Lexeme	Neighbours	Inferred co-occurrences
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$`:`*walk*, $\overline{\texttt{nsubj}}$`:`*bark*, `amod:`*hot*
car	*vehicle*, lorry, bus	`amod:`*four-wheel*, `amod:`*horse-drawn*, `amod:`*military*

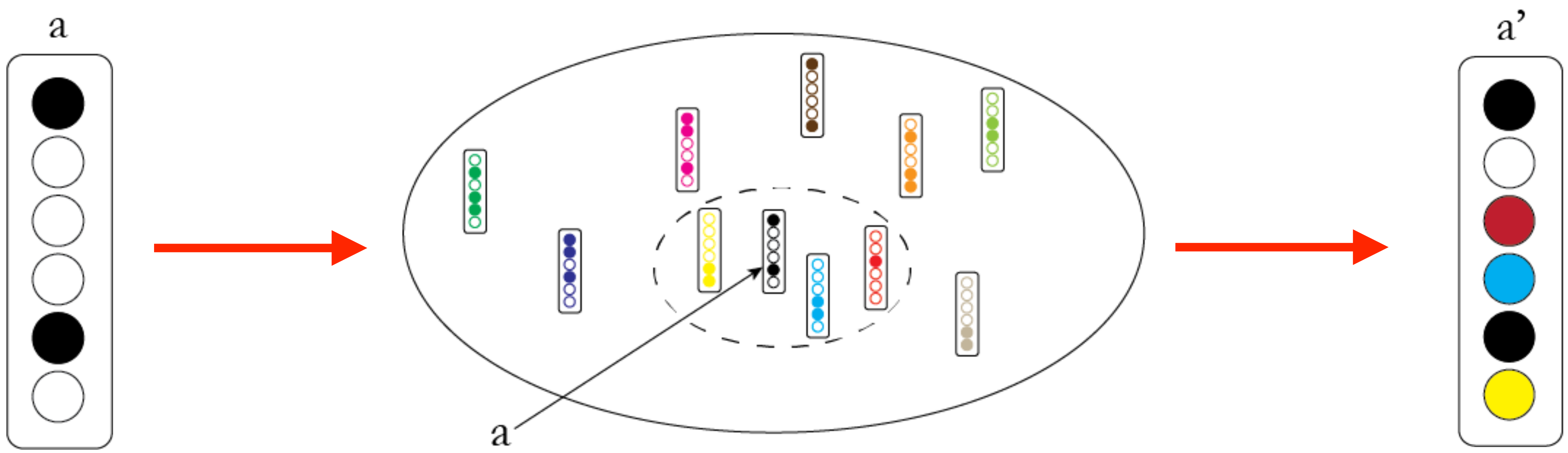
Distributional Inference



Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$`:`*walk*, $\overline{\texttt{nsubj}}$`:`*bark*, `amod:`*hot*
car	*vehicle*, lorry, bus	`amod:`*four-wheel*, `amod:`*horse-drawn*, `amod:`*military*

- Cats bark? Well…not so sure really…

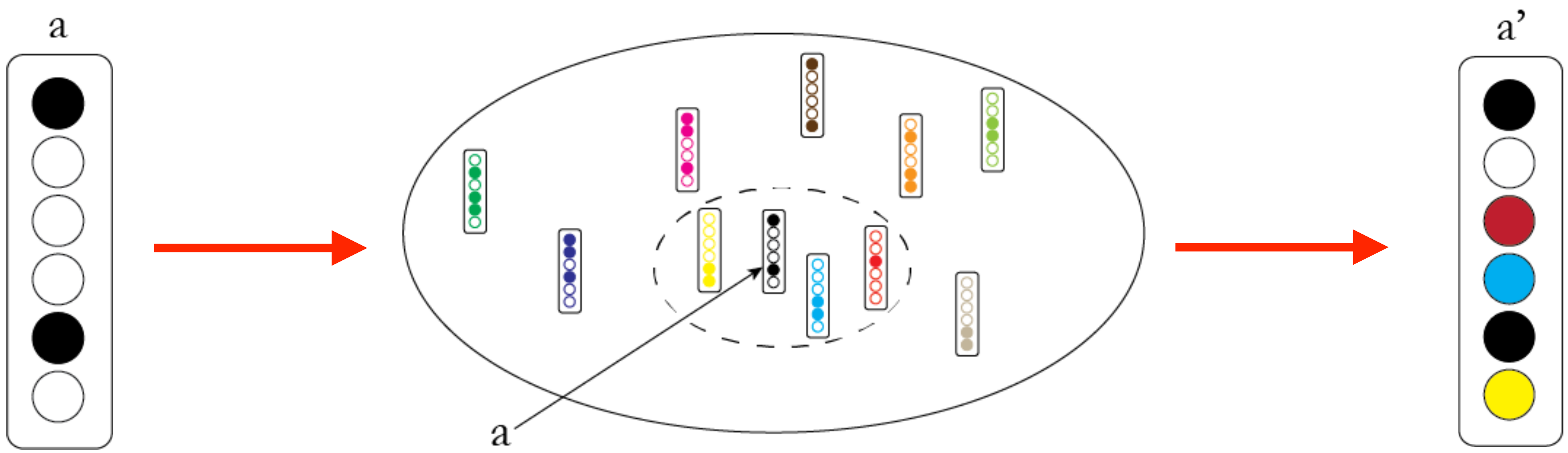
Distributional Inference



Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$`:`*walk*, $\overline{\texttt{nsubj}}$`:`*bark*, `amod:`*hot*
car	*vehicle*, lorry, bus	`amod:`*four-wheel*, `amod:`*horse-drawn*, `amod:`*military*

- Cats bark? Well…not so sure really…
- With too many neighbours might infer that there are *horse-drawn cats* or *military cats*

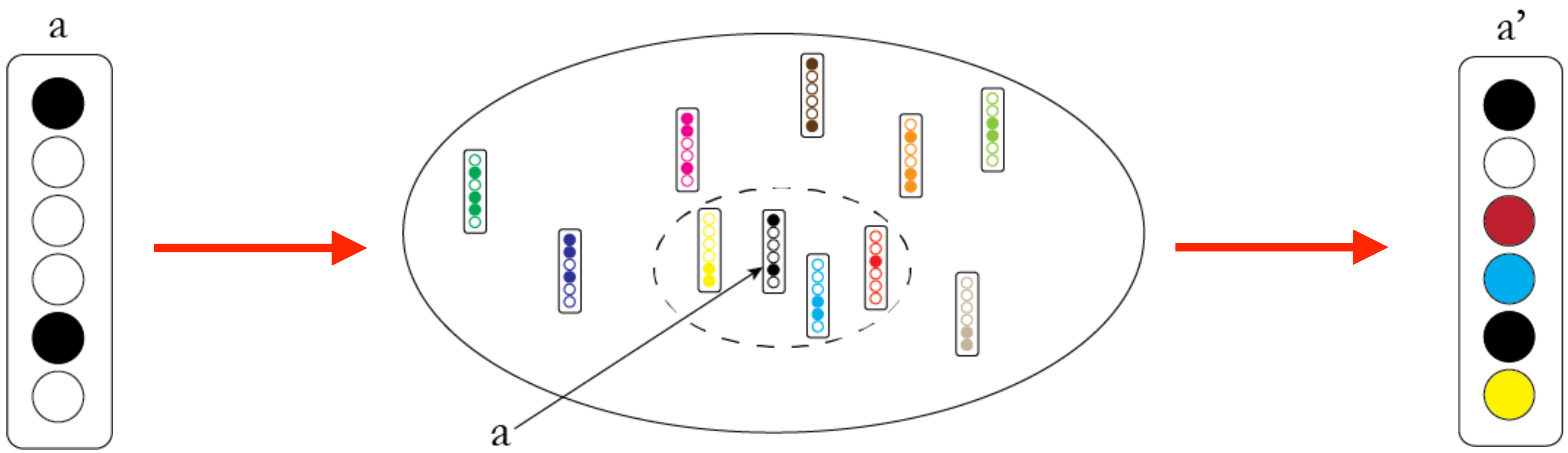
Distributional Inference



Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$`:`*walk*, $\overline{\texttt{nsubj}}$`:`*bark*, `amod:`*hot*
car	*vehicle*, lorry, bus	`amod:`*four-wheel*, `amod:`*horse-drawn*, `amod:`*military*

- Cats bark? Well…not so sure really…
- With too many neighbours might infer that there are *horse-drawn cats* or *military cats*
- The inference procedure does not assess the suitability of a feature

Distributional Inference



Lexeme	**Neighbours**	**Inferred co-occurrences**
magazine	*newspaper*, journal, paper	$\overline{\texttt{dobj}}$`:`*sell*, $\overline{\texttt{nsubj}}$`:`*report*, `amod:`*daily*
cat	*dog*, rabbit, pet	$\overline{\texttt{dobj}}$`:`*walk*, $\overline{\texttt{nsubj}}$`:`*bark*, `amod:`*hot*
car	*vehicle*, lorry, bus	`amod:`*four-wheel*, `amod:`*horse-drawn*, `amod:`*military*

- Cats bark? Well…not so sure really…
- With too many neighbours might infer that there are *horse-drawn cats* or *military cats*
- The inference procedure does not assess the suitability of a feature
- But would be useful to have some filtering mechanism (more on that later)

From Distributional Inference to Offset Inference

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)

From Distributional Inference to Offset Inference

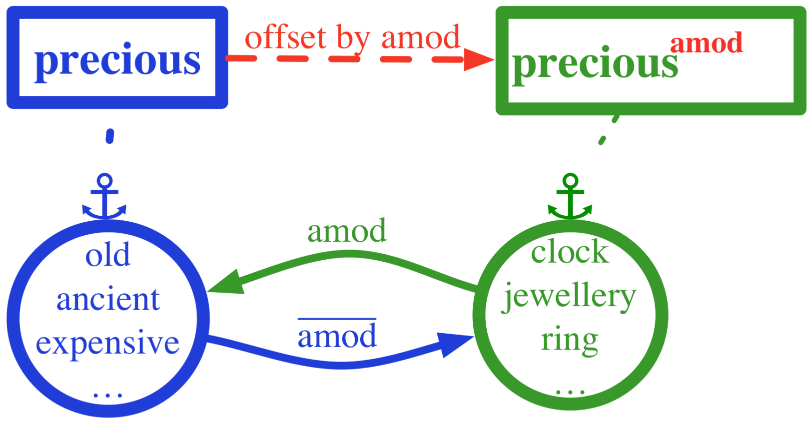
- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
    - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
  - Create a noun offset view (along its $\overline{\text{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")



From Distributional Inference to Offset Inference

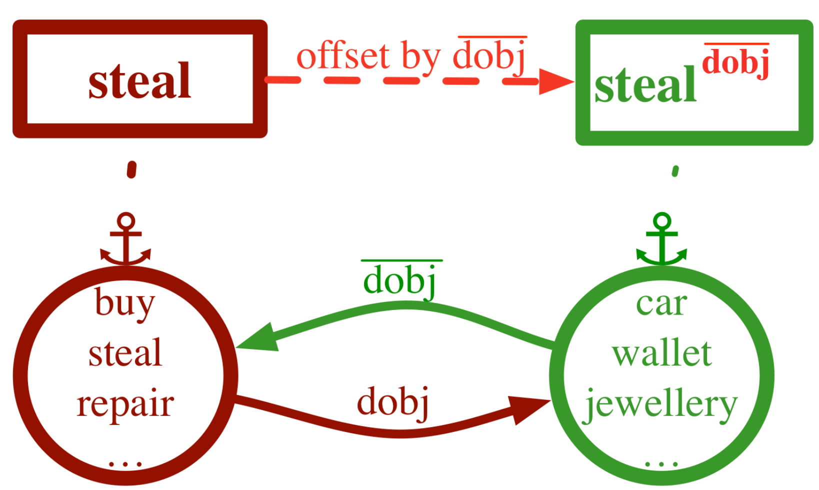
- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
  - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
    - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")
    - Create a noun offset view (along its $\overline{\texttt{dobj}}$ edge) for the verb ***stolen*** (representing "*a thing that can be stolen*")

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
    - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")
    - Create a noun offset view (along its $\overline{\texttt{dobj}}$ edge) for the verb ***stolen*** (representing "*a thing that can be stolen*")



From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
    - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")
    - Create a noun offset view (along its $\overline{\texttt{dobj}}$ edge) for the verb ***stolen*** (representing "*a thing that can be stolen*")

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
  - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")
  - Create a noun offset view (along its $\overline{\texttt{dobj}}$ edge) for the verb ***stolen*** (representing "*a thing that can be stolen*")
  - Realise that "*a thing that can be stolen*" is similar to "*a precious thing*" and add observed features from "*a precious thing*" to "*a thing that can be stolen*"

From Distributional Inference to Offset Inference

- Standard algorithm neglects the rich type structure of APTs
- Can leverage offset views to enrich elementary representations (Kober et al., 2017)
- Enables inferring knowledge on a more abstract level
    - Create a noun offset view (along its $\overline{\texttt{amod}}$ edge) for the adjective ***precious*** (representing "*a precious thing*")
    - Create a noun offset view (along its $\overline{\texttt{dobj}}$ edge) for the verb ***stolen*** (representing "*a thing that can be stolen*")
    - Realise that "*a thing that can be stolen*" is similar to "*a precious thing*" and add observed features from "*a precious thing*" to "*a thing that can be stolen*"
    - (In the given APT space from the BNC, the two offset views where 50% more similar to each other in terms of the cosine of their vector representations than the original representations)

From Distributional Inference to Offset Inference

From Distributional Inference to Offset Inference

- Furthermore uncovers relation between distributional inference and distributional composition in APTs

From Distributional Inference to Offset Inference

- Furthermore uncovers relation between distributional inference and distributional composition in APTs
    - Both mechanisms realised by the same operation (offset followed by a merge)

From Distributional Inference to Offset Inference

- Furthermore uncovers relation between distributional inference and distributional composition in APTs
    - Both mechanisms realised by the same operation (offset followed by a merge)
    - Can use in a complementary manner; distributional inference as a process of *co-occurrence embellishment*, distributional composition as a process of *co-occurrence filtering*

From Distributional Inference to Offset Inference

- Furthermore uncovers relation between distributional inference and distributional composition in APTs
    - Both mechanisms realised by the same operation (offset followed by a merge)
    - Can use in a complementary manner; distributional inference as a process of *co-occurrence embellishment*, distributional composition as a process of *co-occurrence filtering*
    - Using composition to filter noisy inferences that do not make sense in the given context (no more *barking cats*, *horse-drawn cats* or *military cats*)

From Distributional Inference to Offset Inference

- Furthermore uncovers relation between distributional inference and distributional composition in APTs
    - Both mechanisms realised by the same operation (offset followed by a merge)
    - Can use in a complementary manner; distributional inference as a process of *co-occurrence embellishment*, distributional composition as a process of *co-occurrence filtering*
    - Using composition to filter noisy inferences that do not make sense in the given context (no more *barking cats*, *horse-drawn cats* or *military cats*)
    - Inference mechanism falls out of the existing APT theory, no need to fiddle around with the formulation

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- **Evaluating APTs - A better attempt :))))**
- Conclusion

Evaluation - A second attempt

Evaluation - A second attempt

Dataset
WS353 (Sim)
WS353 (Rel)
MEN
SimLex-999
ML10 - AN
ML10 - NN
ML10 - VO

Evaluation - A second attempt

Dataset	word2vec
WS353 (Sim)	0.64
WS353 (Rel)	0.42
MEN	0.63
SimLex-999	0.25
ML10 - AN	0.50
ML10 - NN	0.47
ML10 - VO	0.42

Evaluation - A second attempt

Dataset	word2vec	APTs
WS353 (Sim)	0.64	0.40
WS353 (Rel)	0.42	0.24
MEN	0.63	0.36
SimLex-999	0.25	0.22
ML10 - AN	0.50	0.39
ML10 - NN	0.47	0.41
ML10 - VO	0.42	0.35

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs
WS353 (Sim)	0.64	0.40	0.52
WS353 (Rel)	0.42	0.24	0.35
MEN	0.63	0.36	0.43
SimLex-999	0.25	0.22	0.25
ML10 - AN	0.50	0.39	0.39
ML10 - NN	0.47	0.41	0.43
ML10 - VO	0.42	0.35	0.36

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

*) Can improve performance to up to 0.60 with a slightly different inference process; see Kober (2017)

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

*) Can improve performance to up to 0.60 with a slightly different inference process; see Kober (2017)

- Results substantially improved (especially for the composition task)

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

*) Can improve performance to up to 0.60 with a slightly different inference process; see Kober (2017)

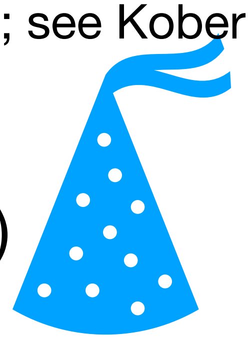
- Results substantially improved (especially for the composition task)

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

*) Can improve performance to up to 0.60 with a slightly different inference process; see Kober (2017)

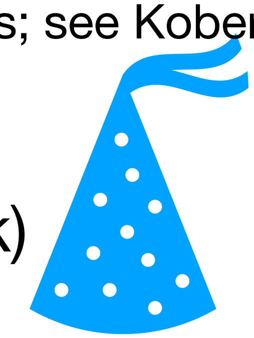
- Results substantially improved (especially for the composition task)



Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

*) Can improve performance to up to 0.60 with a slightly different inference process; see Kober (2017)

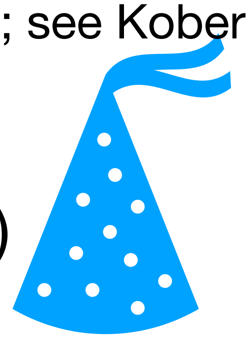


- Results substantially improved (especially for the composition task)
- Sparsity has a large impact, but distributional inference can successfully address it

Evaluation - A second attempt

Dataset	word2vec	APTs	Tuned APTs	APTs + DI
WS353 (Sim)	0.64	0.40	0.52	0.59
WS353 (Rel)	0.42	0.24	0.35	0.35
MEN	0.63	0.36	0.43	0.49
SimLex-999	0.25	0.22	0.25	0.30*
ML10 - AN	0.50	0.39	0.39	0.52
ML10 - NN	0.47	0.41	0.43	0.51
ML10 - VO	0.42	0.35	0.36	0.45

*) Can improve performance to up to 0.60 with a slightly different inference process; see Kober (2017)



- Results substantially improved (especially for the composition task)
- Sparsity has a large impact, but distributional inference can successfully address it
- Even with more data, distributional inference is helpful (see Kober 2017)

Works quite well, but…

Works quite well, but…

- Can address the issue of data sparsity up to some point

Works quite well, but…

- Can address the issue of data sparsity up to some point
- Distributional Inference suffers from the “cold start” problem

Works quite well, but…

- Can address the issue of data sparsity up to some point
- Distributional Inference suffers from the “cold start” problem
    - Trying to improve a distributional space on the basis of the same space that we know is slightly dodgy

Works quite well, but…

- Can address the issue of data sparsity up to some point
- Distributional Inference suffers from the “cold start” problem
    - Trying to improve a distributional space on the basis of the same space that we know is slightly dodgy
- Scalability issues

Works quite well, but…

- Can address the issue of data sparsity up to some point
- Distributional Inference suffers from the “cold start” problem
    - Trying to improve a distributional space on the basis of the same space that we know is slightly dodgy
- Scalability issues
    - Difficult to scale beyond 3-4 word phrases, because the distributional space is still mostly made up of unigrams, so its hard to find “good neighbours” for longer phrases from which to infer useful features from

Works quite well, but…

- Can address the issue of data sparsity up to some point
- Distributional Inference suffers from the “cold start” problem
    - Trying to improve a distributional space on the basis of the same space that we know is slightly dodgy
- Scalability issues
    - Difficult to scale beyond 3-4 word phrases, because the distributional space is still mostly made up of unigrams, so its hard to find “good neighbours” for longer phrases from which to infer useful features from
    - Could compose all high-frequency *n-1* grams and add them to the space to build better representations for *n* grams, but that has severe scalability issues.

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- Conclusion

Outline

- Introduction to Anchored Packed Trees
- Evaluating APTs - A first attempt :((((
- Distributional Inference
- Evaluating APTs - A better attempt :))))
- **Conclusion**

Conclusion

Conclusion

- APTs as a compositional distributional semantic model

Conclusion

- APTs as a compositional distributional semantic model
- Semantic APT space is very sparse, resulting in low performance on standard lexical and phrasal tasks

Conclusion

- APTs as a compositional distributional semantic model
- Semantic APT space is very sparse, resulting in low performance on standard lexical and phrasal tasks
- Proposed distributional inference (and subsequently generalised to offset inference) to address the sparsity issue

Conclusion

- APTs as a compositional distributional semantic model
- Semantic APT space is very sparse, resulting in low performance on standard lexical and phrasal tasks
- Proposed distributional inference (and subsequently generalised to offset inference) to address the sparsity issue
- Highlighted relation between distributional composition and distributional inference in APTs

Conclusion

- APTs as a compositional distributional semantic model
- Semantic APT space is very sparse, resulting in low performance on standard lexical and phrasal tasks
- Proposed distributional inference (and subsequently generalised to offset inference) to address the sparsity issue
- Highlighted relation between distributional composition and distributional inference in APTs
- Performance - especially on phrasal composition tasks - substantially improved

Thats it, I'm done!

Thats it, I’m done!

Thats it, I’m done!

Q & (maybe) A

tkober@inf.ed.ac.uk

References

- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In Proceedings of ACL, 19-27
- Marco Baroni and Alessandro Lenci. 2011. How we BLESSed Semantic Evaluation. In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 1-10
- Elia Bruni, Nam Khanh Tran and Marco Baroni. 2014. Multimodal Distributional Semantics. In JAIR (49), 1-47
- Ido Dagan, Shaul Marcus and Shaul Markovitch. 1993. Contextual Word Similarity and Estimation from Sparse Data. In Proceedings of ACL, 164-171
- Ute Essen and Volker Steinbiss. 1992. Co-occurrence Smoothing for Stochastic Language Modeling. In Proceedings of ICASSP, 161-164
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eytan Ruppin. 2001. Placing Search in Context: The Context Revisited. In Proceedings of WWW, 406-414
- Felix Hill, Roi Reichart and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. In CL, 41(4), 665-695
- Thomas Kober, Julie Weeds, Jeremy Reffin and David Weir. 2016. Improving Sparse Word Representations with Distributional Inference for Semantic Composition. In Proceedings of EMNLP, 1691-1702
- Thomas Kober, Julie Weeds, Jeremy Reffin and David Weir. 2017. Improving Semantic Composition with Offset Inference. In Proceedings of ACL, 433-440
- Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. Cognitive Science, 34(8):1388–1429
- Yves Peirsman. 2008. Word space models of semantic similarity and relatedness. In Proceedings of ESSLLI, 143-152
- David Weir, Julie Weeds, Jeremy Reffin and Thomas Kober. 2016. Aligning Packed Dependency Trees: A theory of composition for distributional semantics. Computational Linguistics 42(4), 727-761