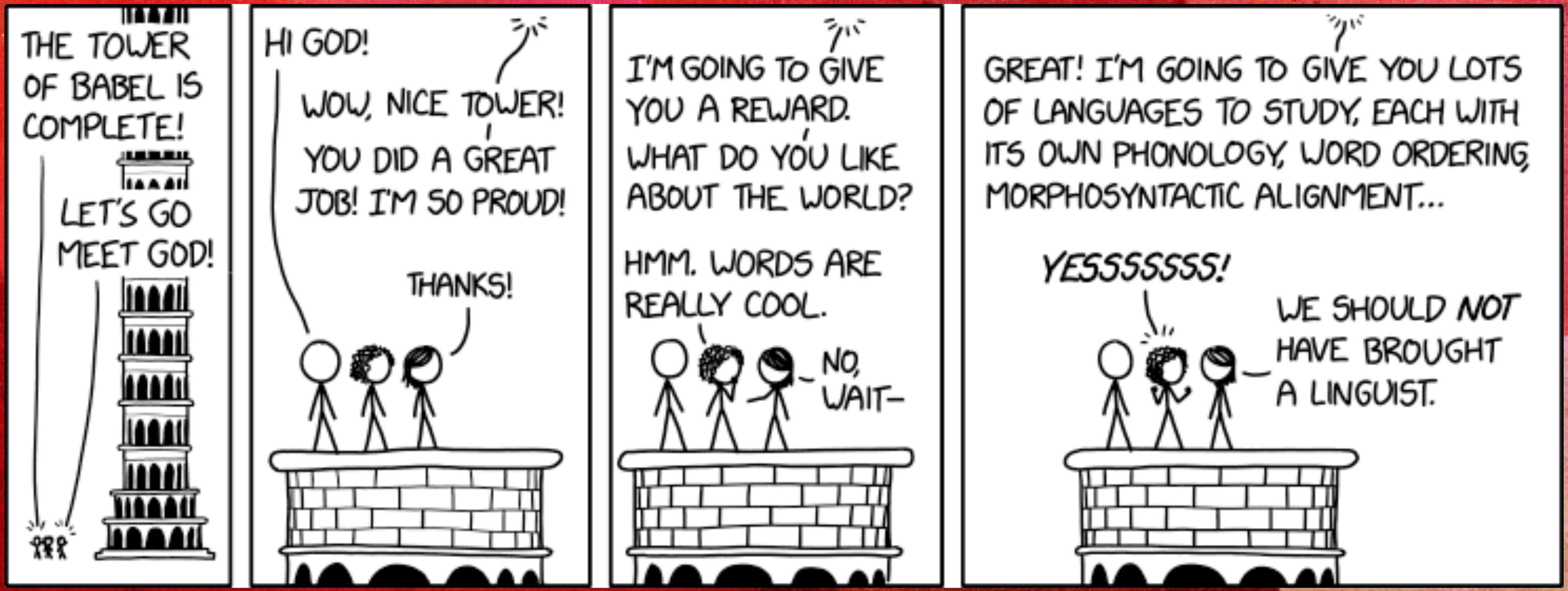


https://xkcd.com/2421/

TOWARDS MULTILINGUAL PARSING

Caio Corro

NATURAL LANGUAGE PROCESSING

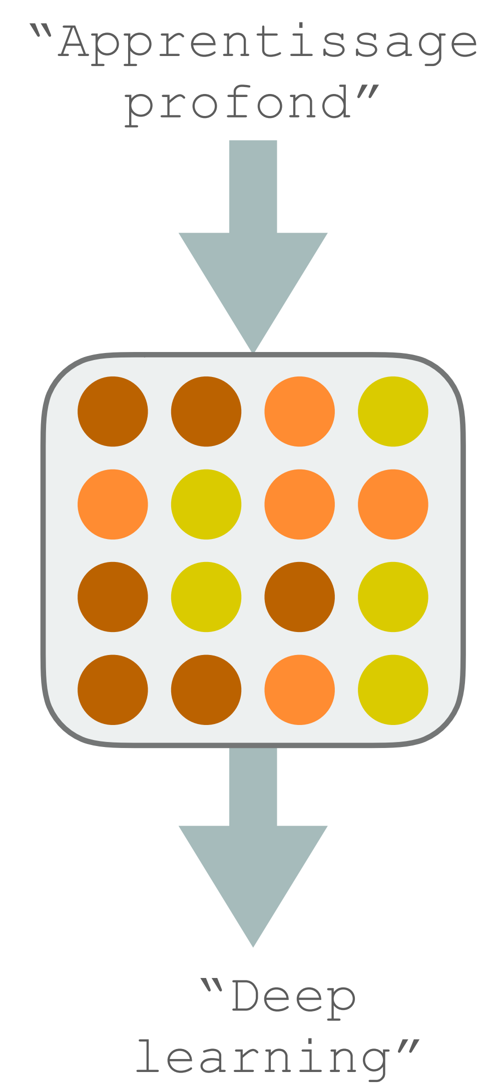
Applications

- Machine Translation
- Information Retrieval
- Sentiment Analysis
- Automatic Summarization
- Human-Computer Interaction
- …

Data type

- Structured: language is sequential
- Discrete: text processing
- Continuous: speech processing

How to build systems that works for many languages?



Difficulties

- **Many languages (> 6000)**
- Ambiguity
- Constantly evolving
- Noisy

TASK EXAMPLES 1/2

Named entity recognition

- Input: sentence
- Output: chunks (person, location, group, creative work, product, corporation)

```
          Person                              Location
[Neil  Armstrong] visited  the  [moon]
```

TASK EXAMPLES 1/2

Named entity recognition

- Input: sentence
- Output: chunks (person, location, group, creative work, product, corporation)

Person **Location**

Neil Armstrong visited the moon

Nested named entity recognition

- Input: sentence
- Output: nested chunks

Person **Location**

The president of the USA died

TASK EXAMPLES 2/2

Semantic parsing

- Input: sentence
- Output: formal representation
  (easier to manipulate by a software + (preferably) language agnostic)

```
The boy want to go outside.   ⇒   want(boy, go(boy, outside))
```

Would be better represented with lambda calculus, but that's not the point

SQL parsing

- Input: sentence
- Output: SQL query

```
I want to book a flight from Paris to Rome.
                    ⇓
SELECT * FROM flight WHERE from = "paris" AND to = "rome"
```

SUPERVISED LEARNING

The lazy solution for rich people

1. Annotate a lot of data for your task
2. Train a big neural network
3. That’s all!

Expensive and boring! $$$$

Limitation

You need to re-annotate data

- For each task and
- For each language

(without even mentioning out-of-domain data, language varieties…)

TOWARDS MULTILINGUAL PARSING (AND NLP)

Why should we care about syntax?

- Interesting field that relies on interesting formal systems (formal grammars, graph theory)
- Strong interaction between downstream tasks and syntax
  => can be used as a first step
- Expose structural differences between languages
  => could help us the build better neural architectures (explained later in the lecture)

Examples

If the syntactic structure of the sentence is known,
maybe we can just use a rule based system for the downstream task?

```
   Person                      Location
[Neil Armstrong] visited the [moon]
```

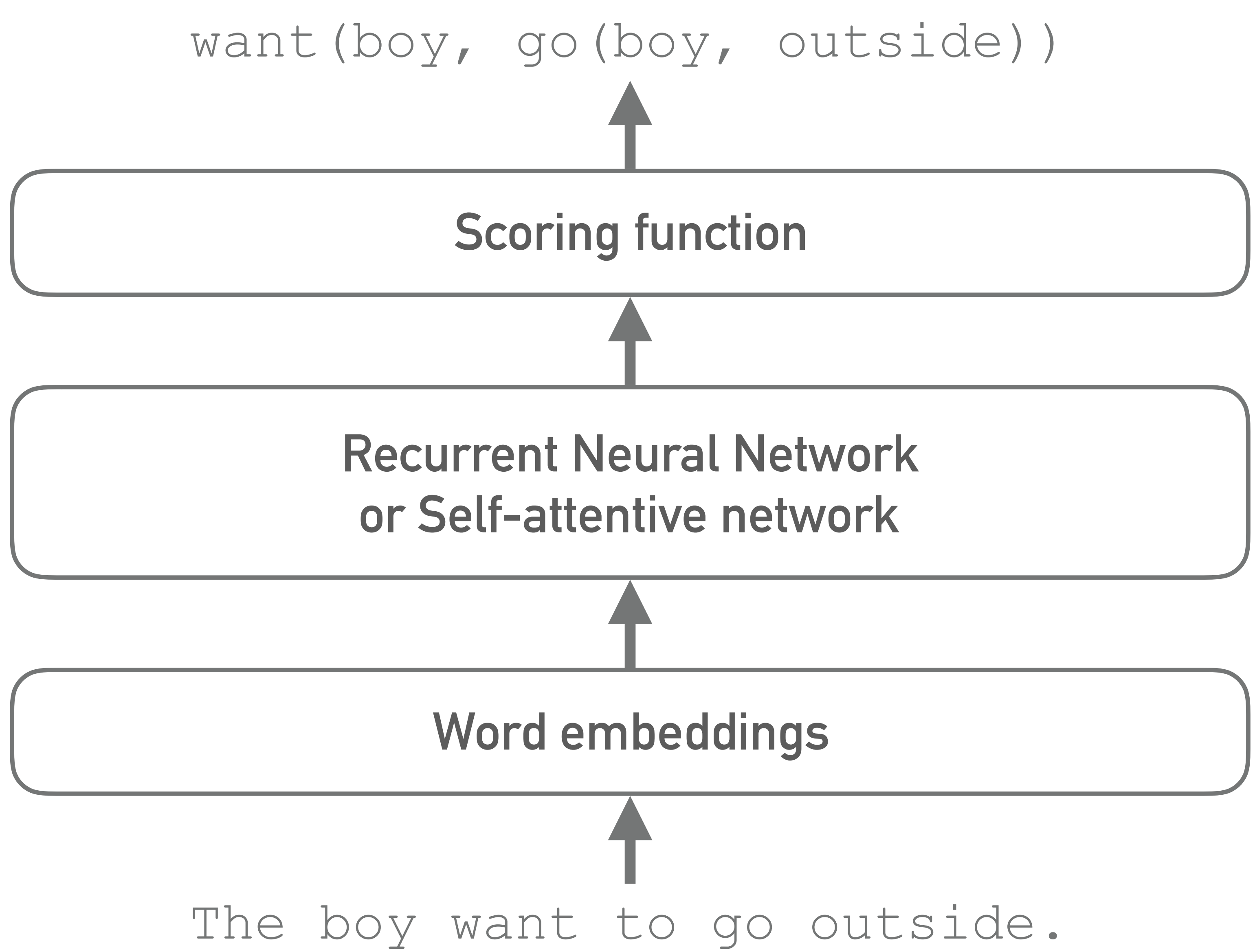
Entities are noun phrases

```
The boy want to go outside.  ⇒  want(boy, go(boy, outside))
```

« boy » is the subject of the sentence

DEEP LEARNING AND MULTILINGUAL NLP

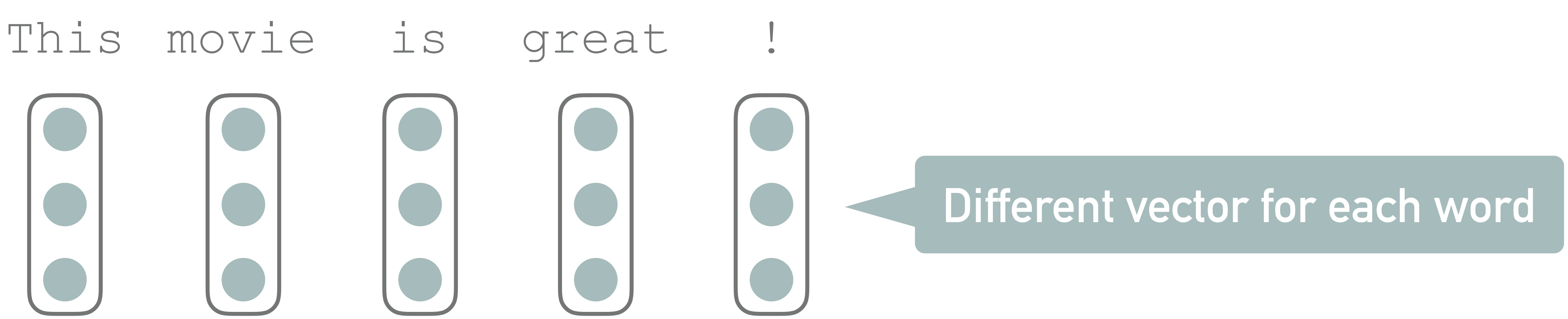
THE BIG PICTURE



WORD EMBEDDINGS

Intuition

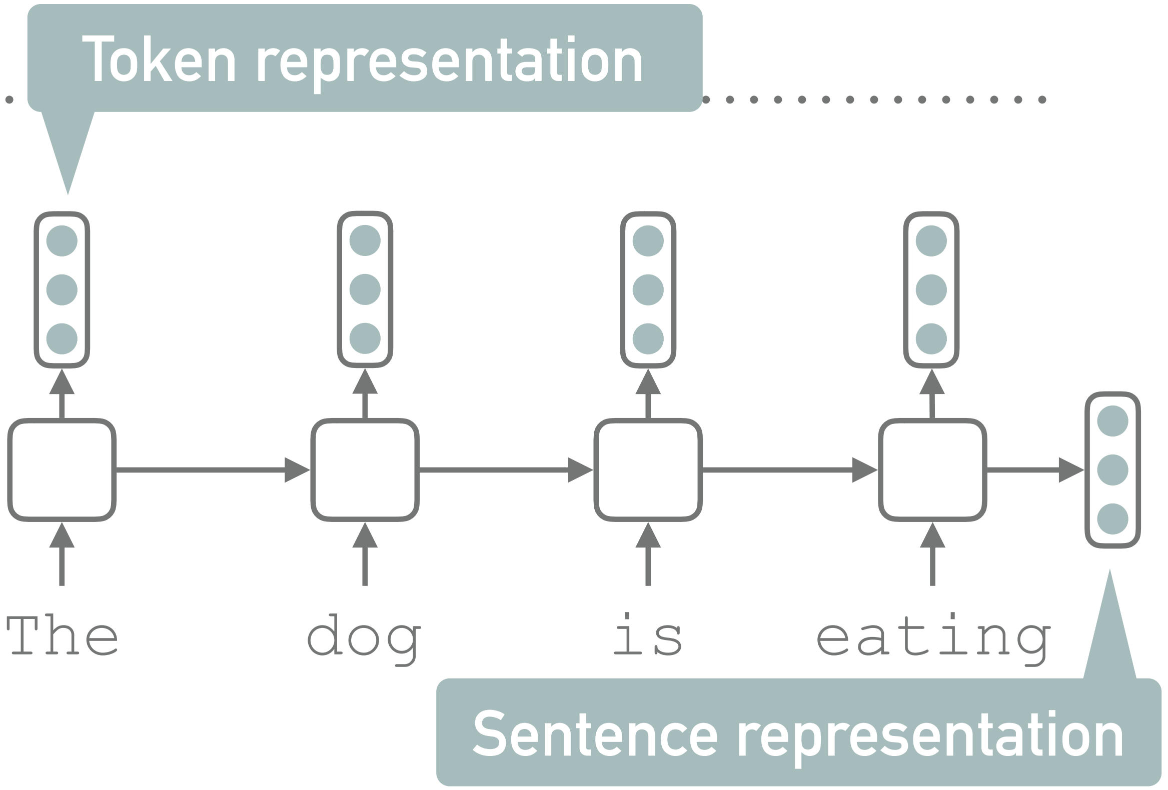
- Vector representation of words
- Given as input of a network
- Capture « features » about the word (Meaning? Syntactic features? Other? Who knows?!)



Pre-trained word embedding

- Unsupervised task
- Nowadays: context-sensitive pre-trained embeddings (i.e. ELMO, Bert)

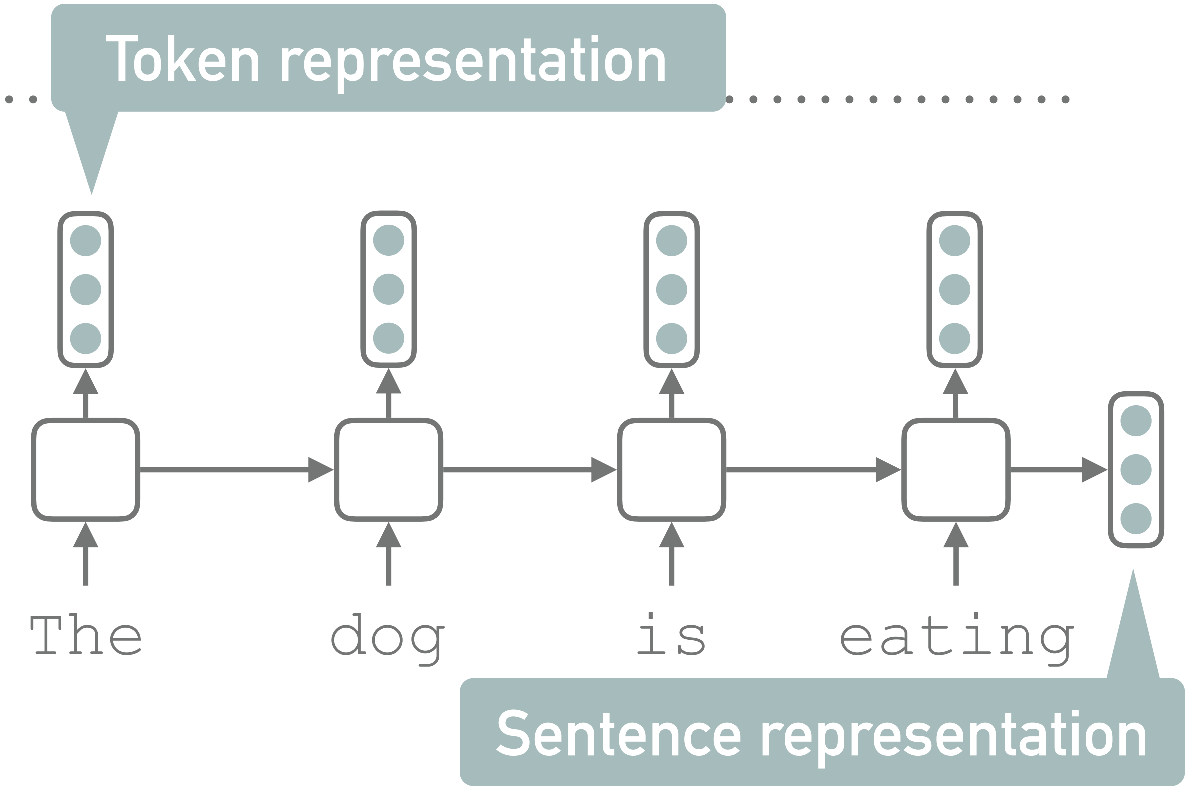
RECURRENT NEURAL NETWORKS



Recurrent neural networks

- Inputs are fed sequentially
- State representation updated at each input

RECURRENT NEURAL NETWORKS



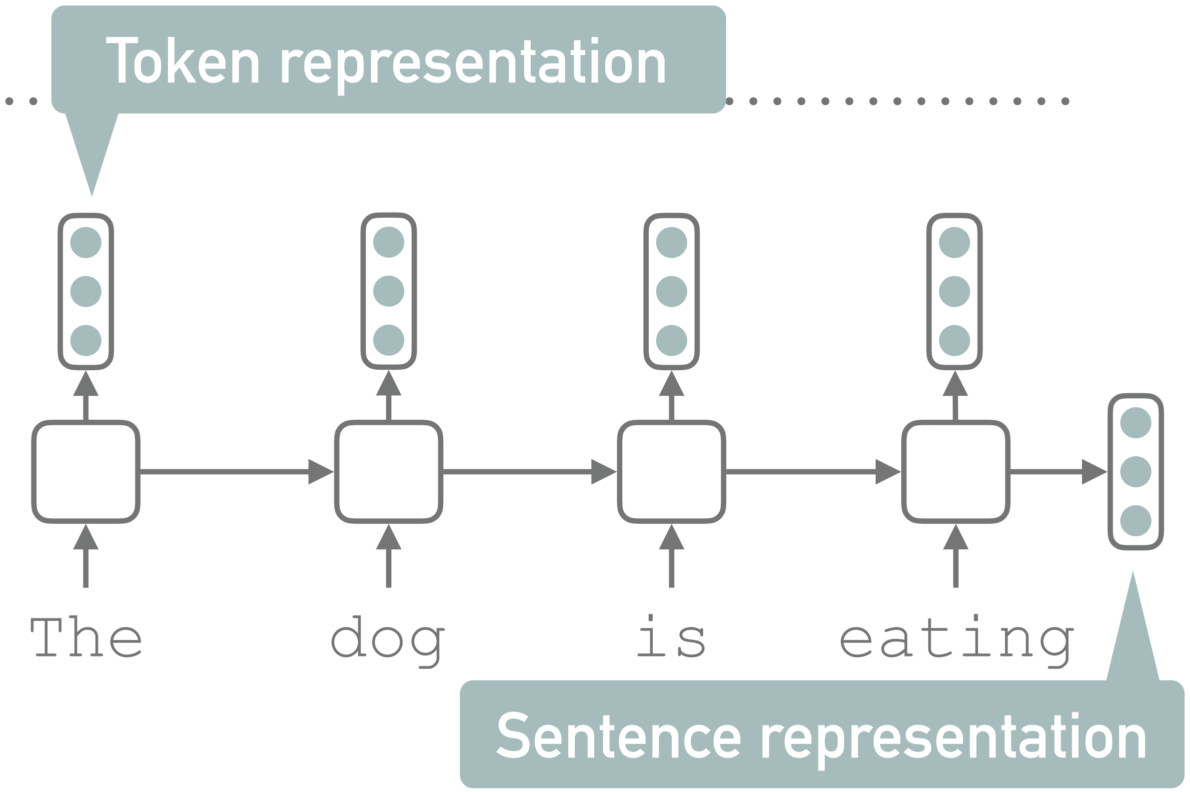
Recurrent neural networks

- Inputs are fed sequentially
- State representation updated at each input

Intuition

Use two RNNs with **different trainable parameters**

RECURRENT NEURAL NETWORKS

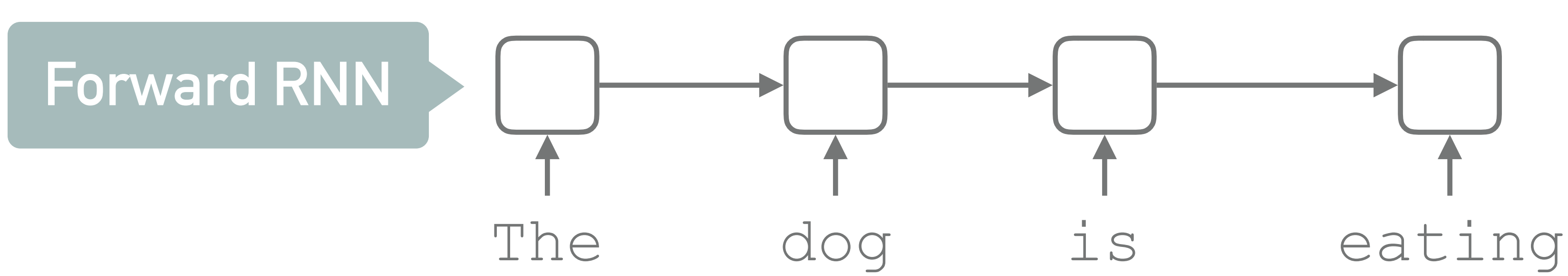


Recurrent neural networks

- Inputs are fed sequentially
- State representation updated at each input

Intuition

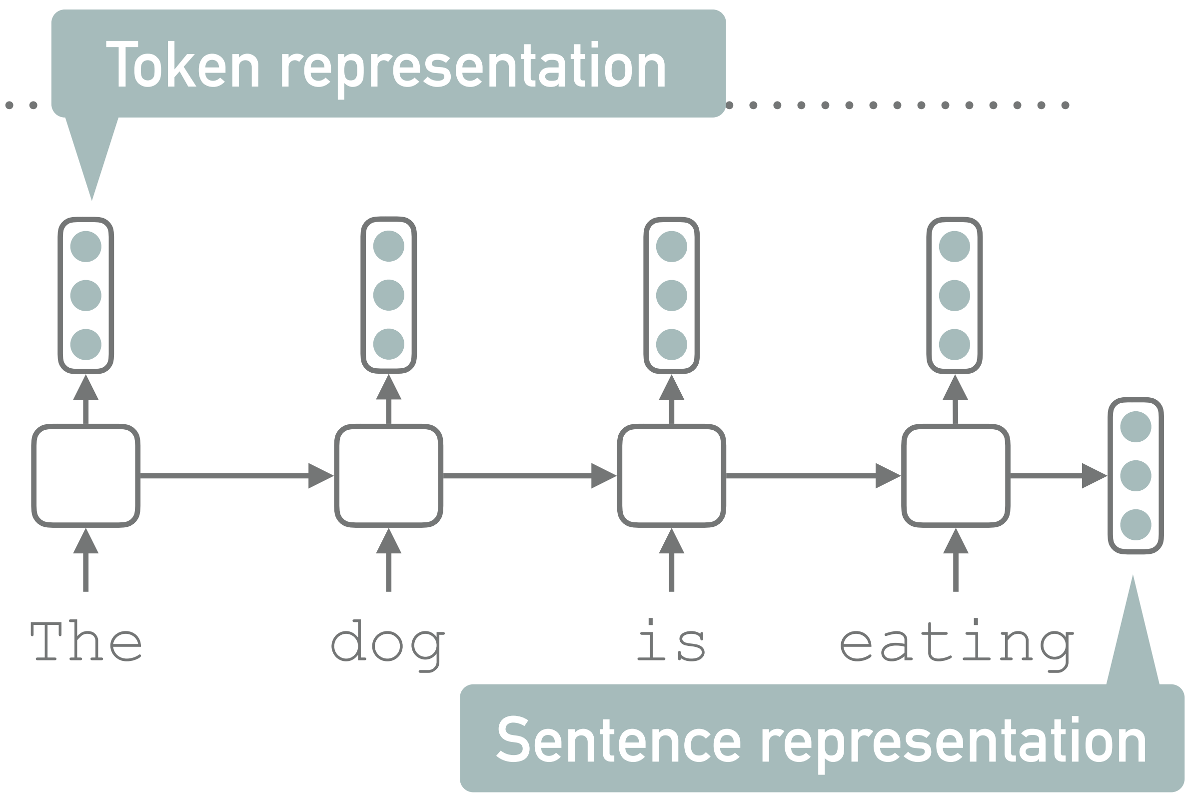
Use two RNNs with **different trainable parameters**



RECURRENT NEURAL NETWORKS

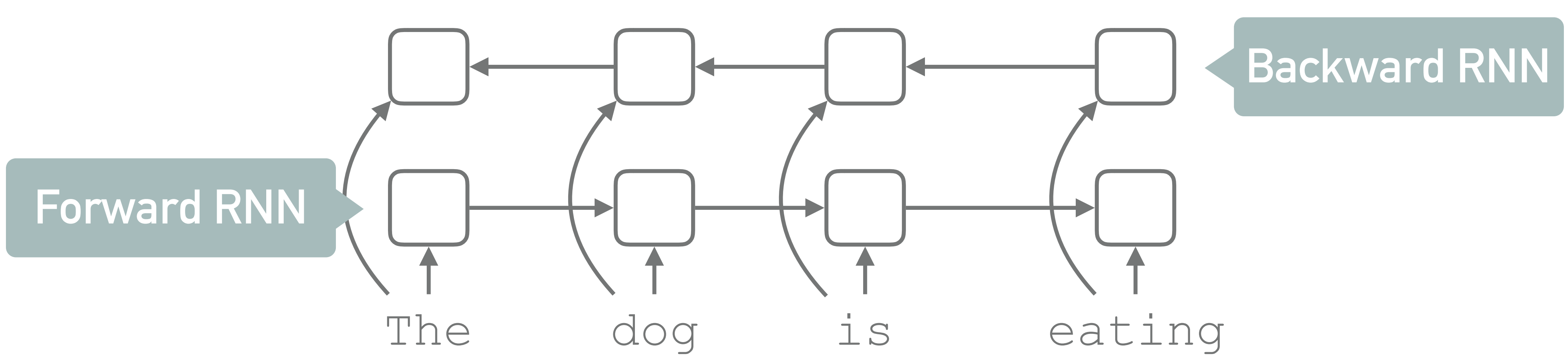
Recurrent neural networks

- Inputs are fed sequentially
- State representation updated at each input



Intuition

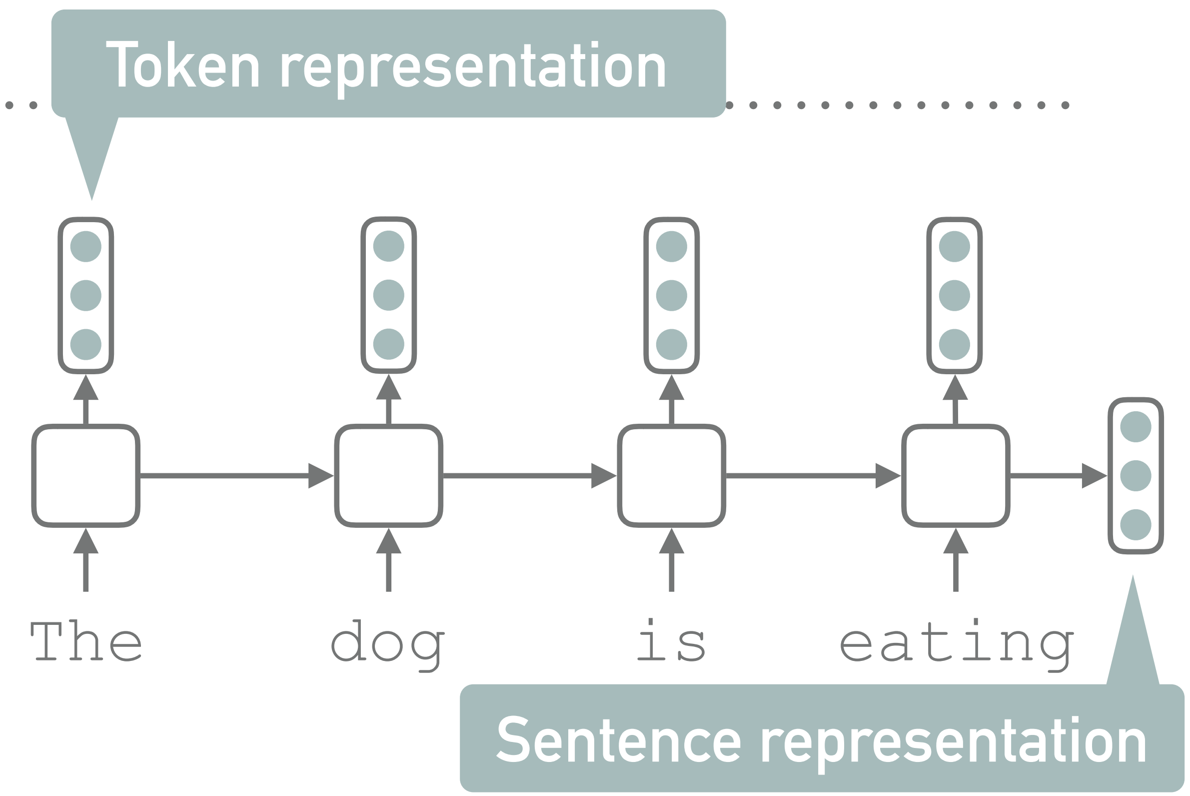
Use two RNNs with **different trainable parameters**



RECURRENT NEURAL NETWORKS

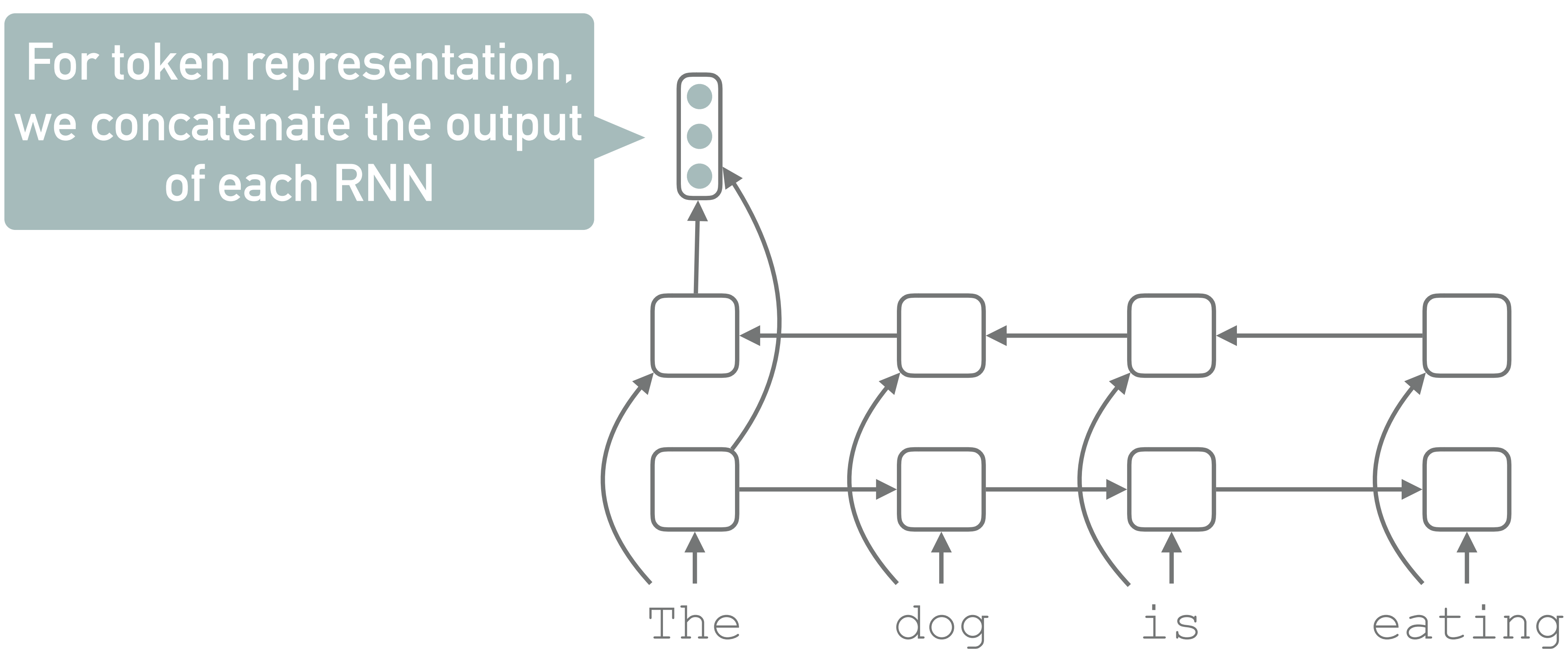
Recurrent neural networks

- Inputs are fed sequentially
- State representation updated at each input

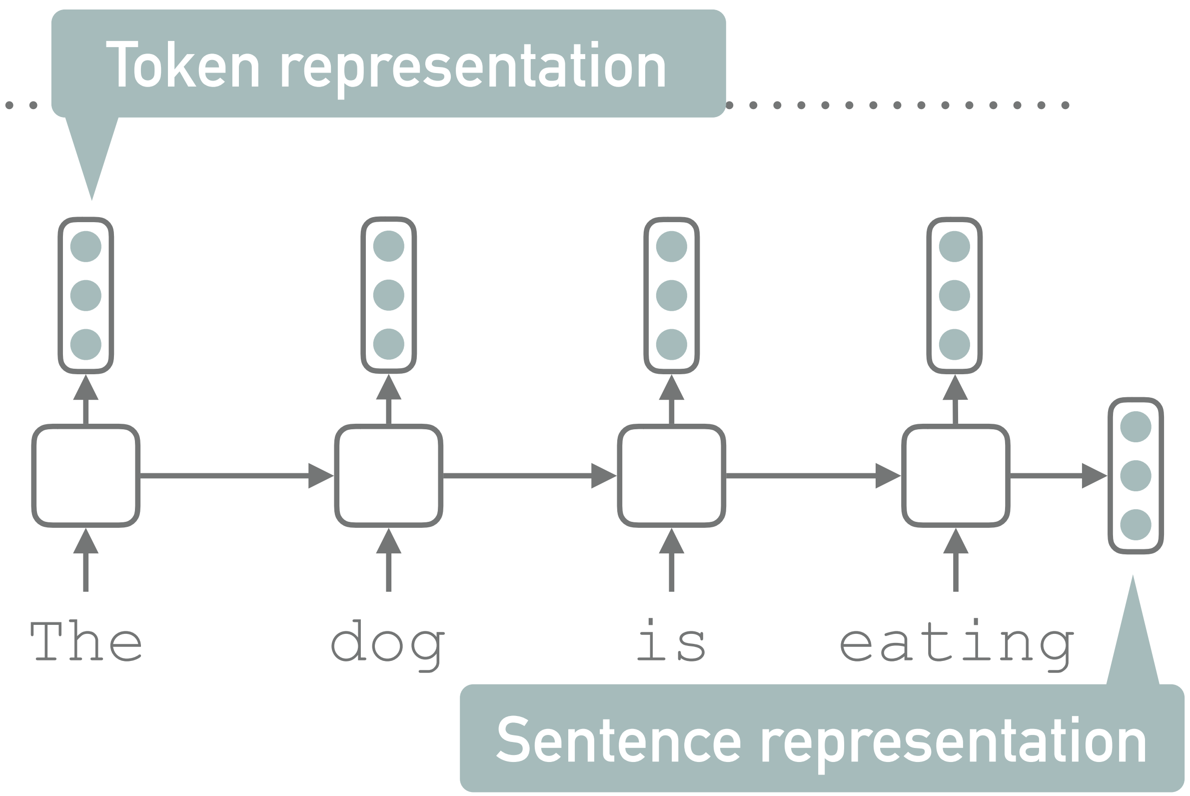


Intuition

Use two RNNs with **different trainable parameters**



RECURRENT NEURAL NETWORKS

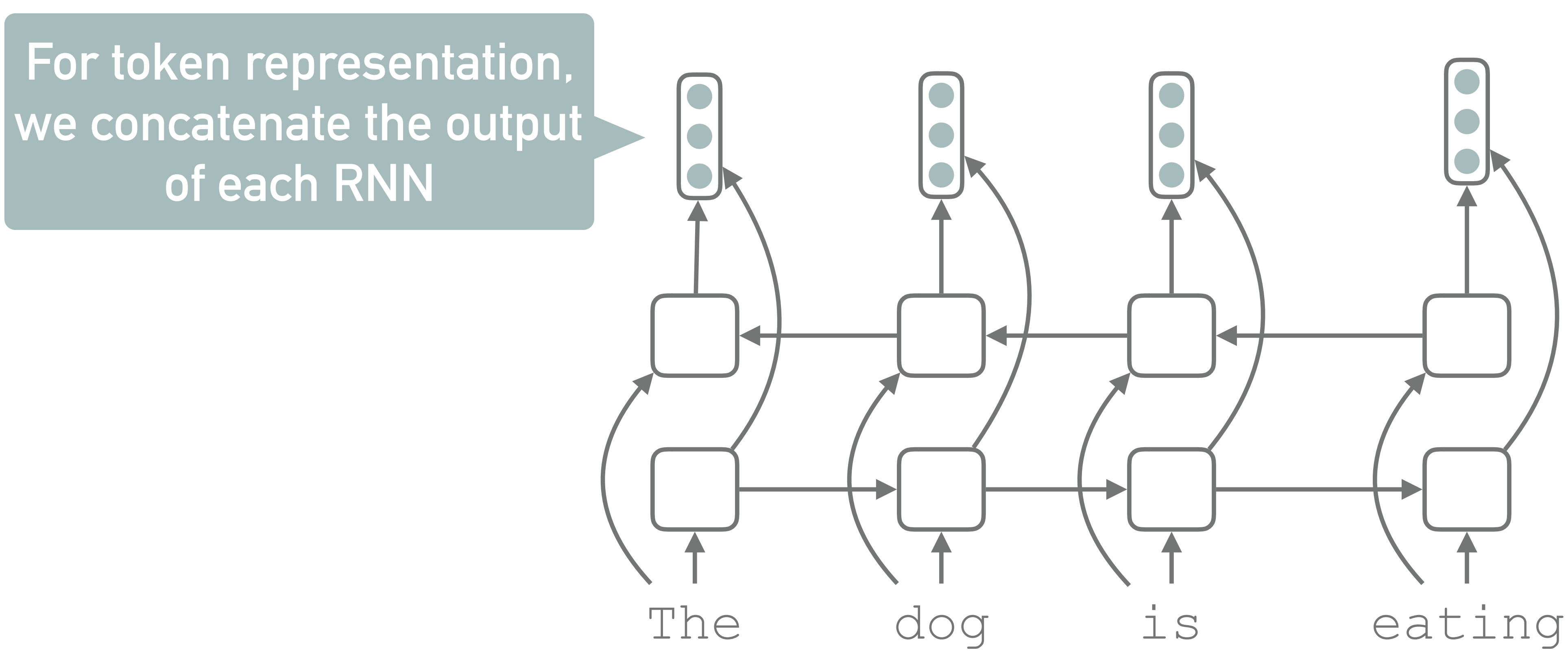


Recurrent neural networks

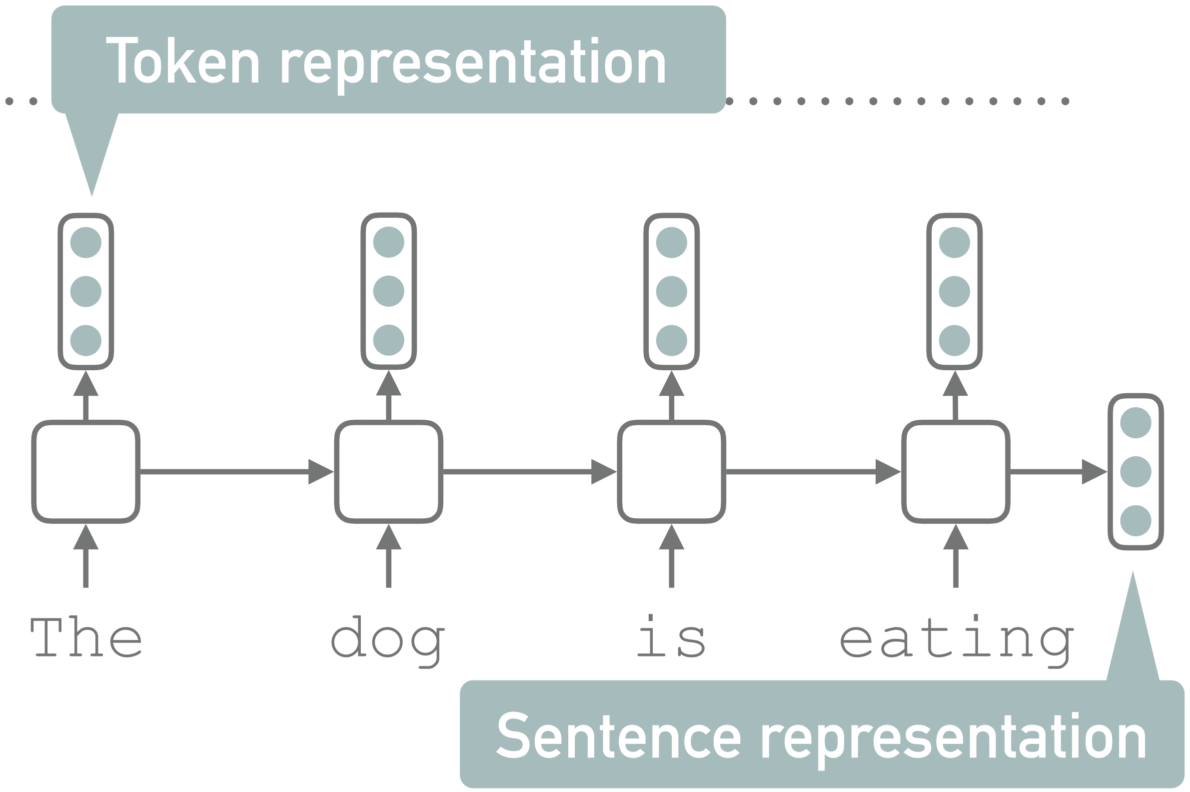
- Inputs are fed sequentially
- State representation updated at each input

Intuition

Use two RNNs with **different trainable parameters**



RECURRENT NEURAL NETWORKS

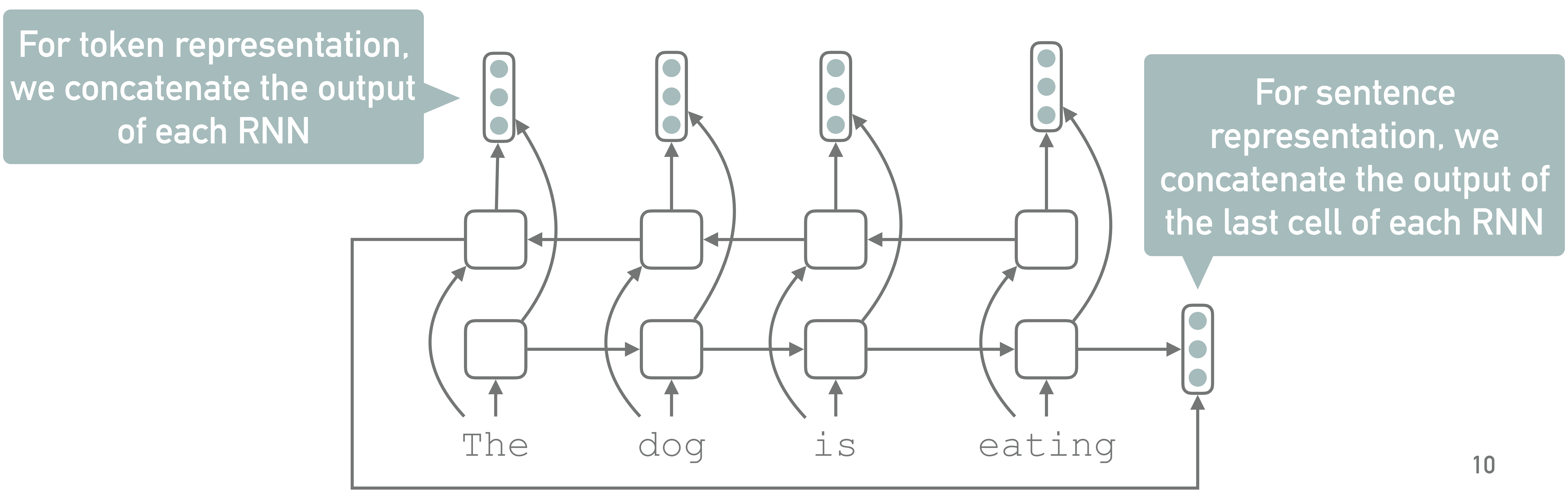


Recurrent neural networks

- Inputs are fed sequentially
- State representation updated at each input

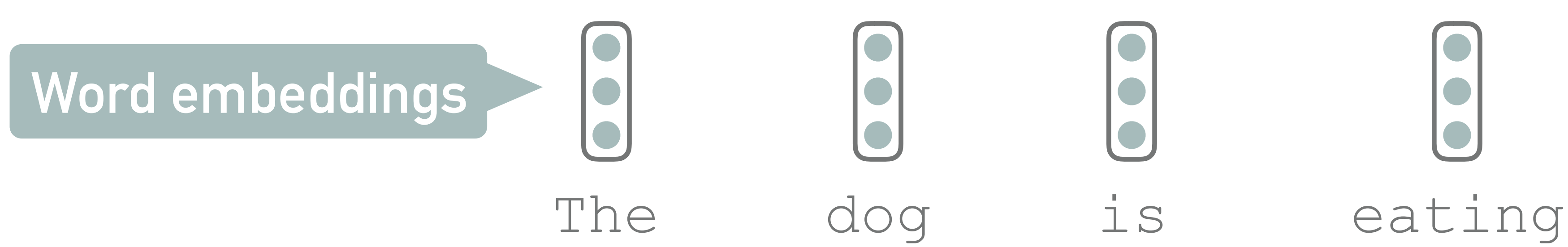
Intuition

Use two RNNs with **different trainable parameters**



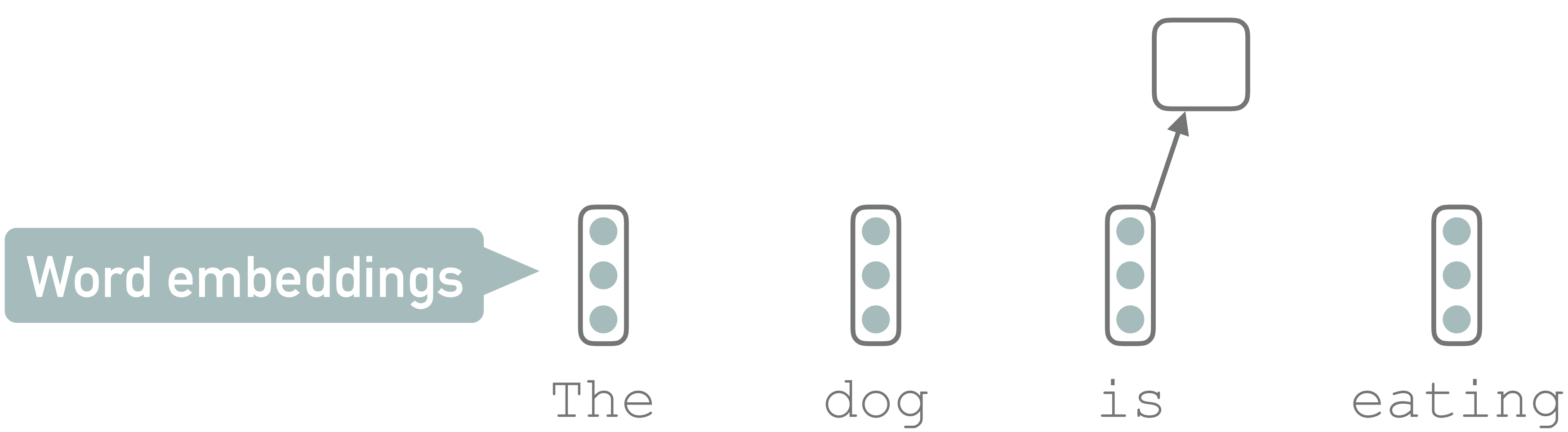
SELF-ATTENTIVE NETWORKS 1/2

- Look at other word using end-to-end trained attention instead of a given graph structure
  In practice we rely on « soft-selection » or attention instead of hard selection



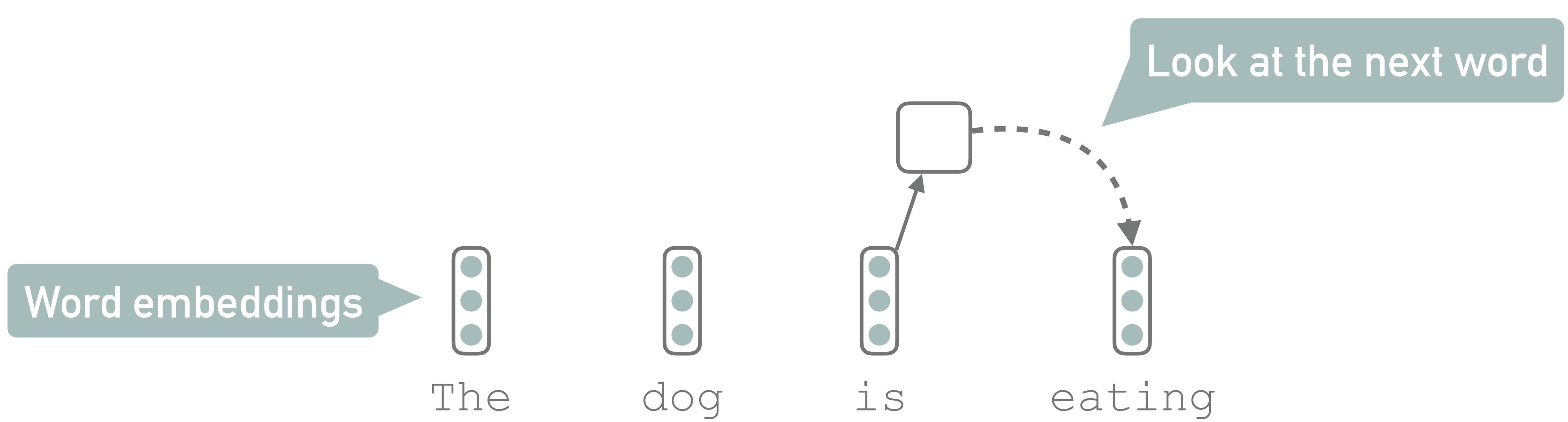
SELF-ATTENTIVE NETWORKS 1/2

- Look at other word using end-to-end trained attention instead of a given graph structure
  In practice we rely on « soft-selection » or attention instead of hard selection



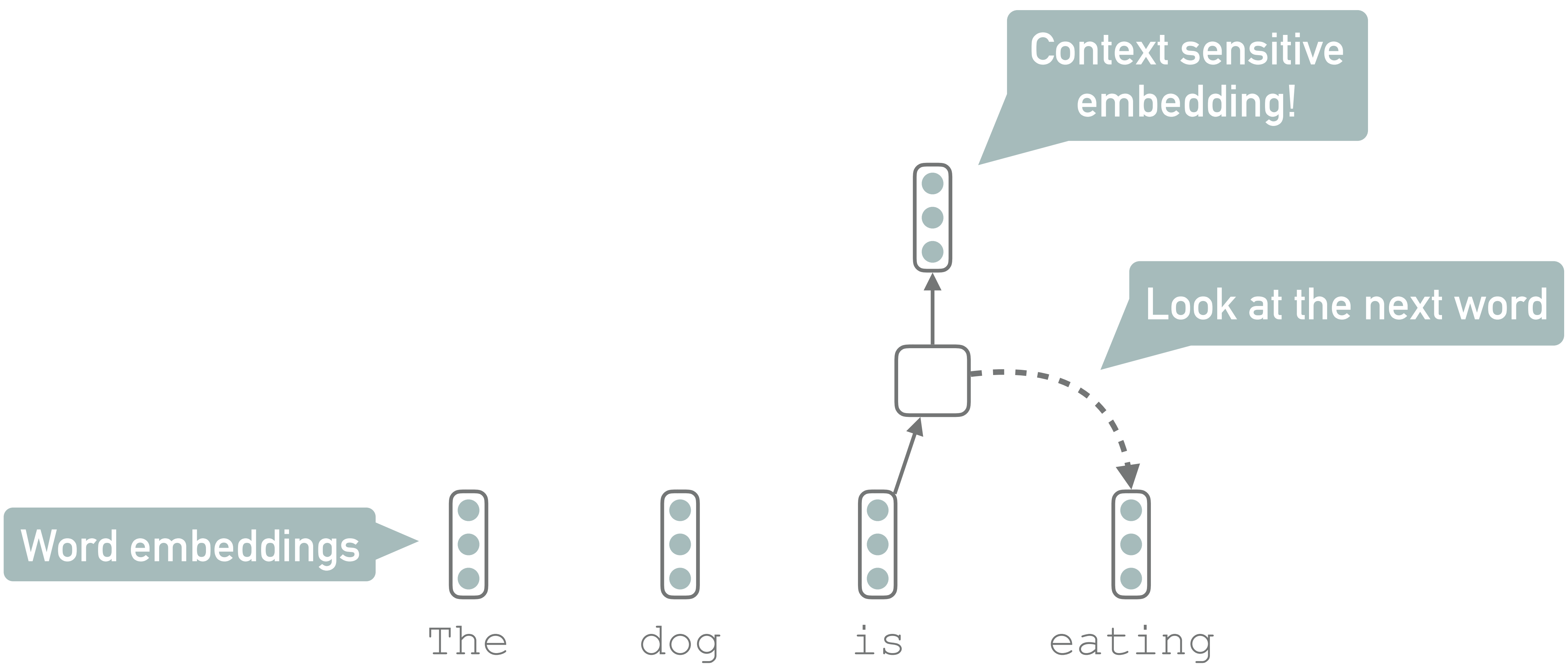
SELF-ATTENTIVE NETWORKS 1/2

- Look at other word using end-to-end trained attention instead of a given graph structure
  In practice we rely on « soft-selection » or attention instead of hard selection



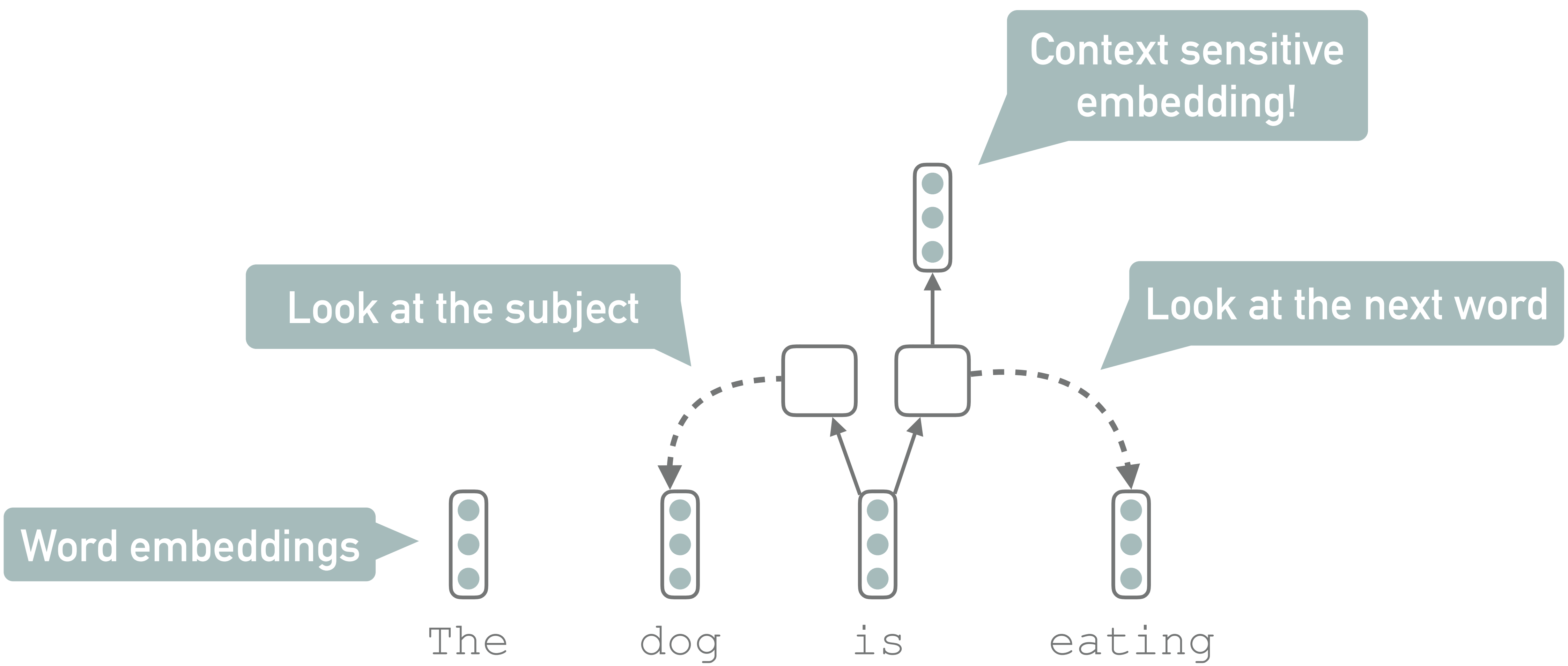
SELF-ATTENTIVE NETWORKS 1/2

- Look at other word using end-to-end trained attention instead of a given graph structure
  In practice we rely on « soft-selection » or attention instead of hard selection



SELF-ATTENTIVE NETWORKS 1/2

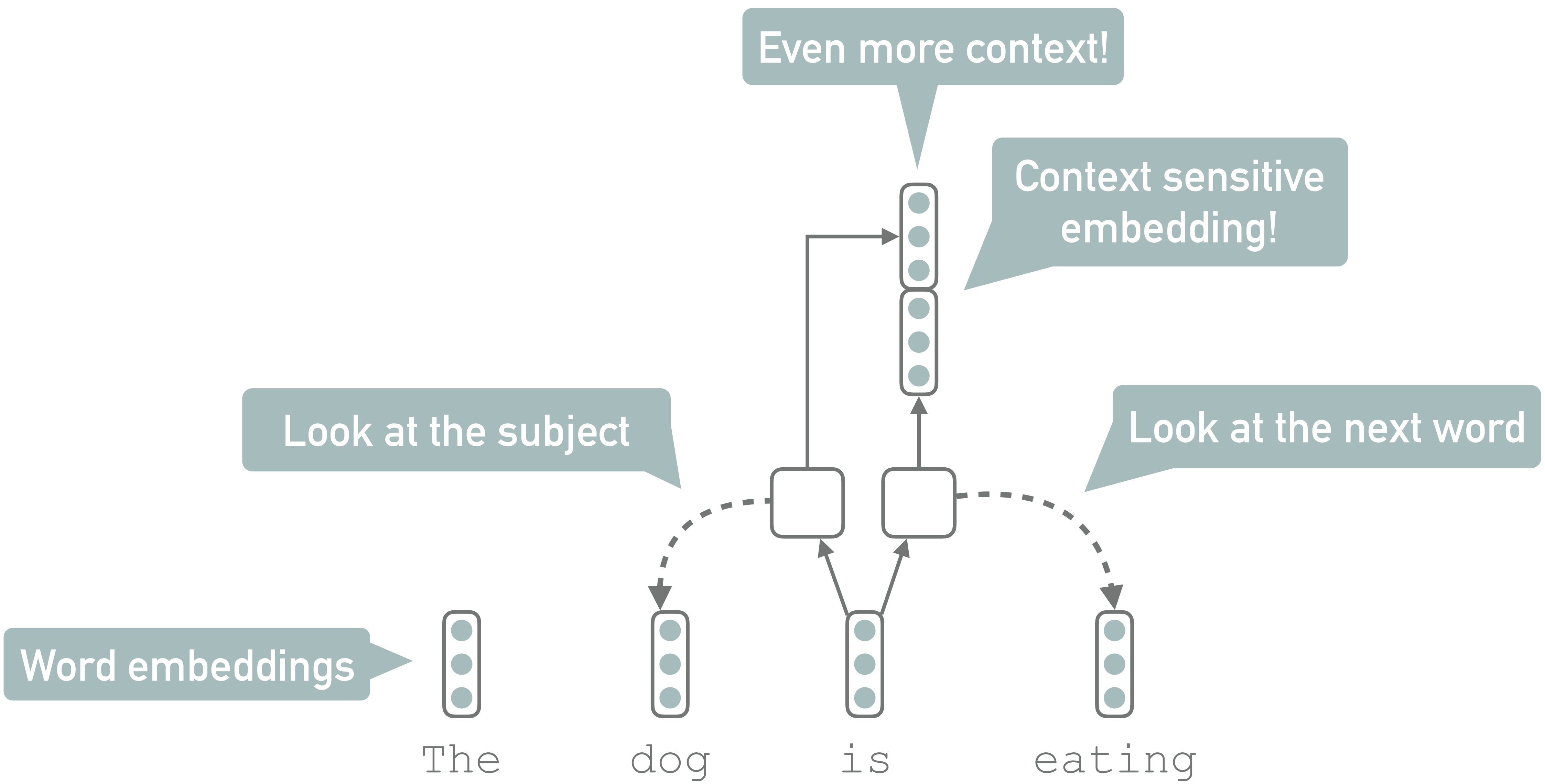
- Look at other word using end-to-end trained attention instead of a given graph structure
  In practice we rely on « soft-selection » or attention instead of hard selection
- Combine several attention modules to attend to several words



SELF-ATTENTIVE NETWORKS 1/2

- Look at other word using end-to-end trained attention instead of a given graph structure
  In practice we rely on « soft-selection » or attention instead of hard selection

- Combine several attention modules to attend to several words



SELF-ATTENTIVE NETWORKS 2/2

Word order information

As such, each head is based on a bag-of-word model! How can take into account word order?

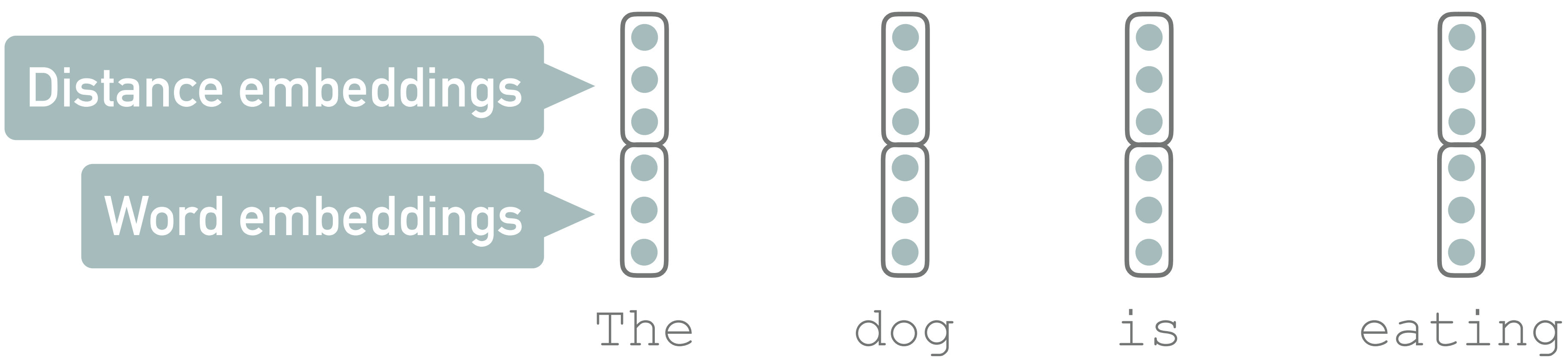
SELF-ATTENTIVE NETWORKS 2/2

Word order information

As such, each head is based on a bag-of-word model! How can take into account word order?

Distance embedding

Concatenate (or sum) a distance embedding to the input



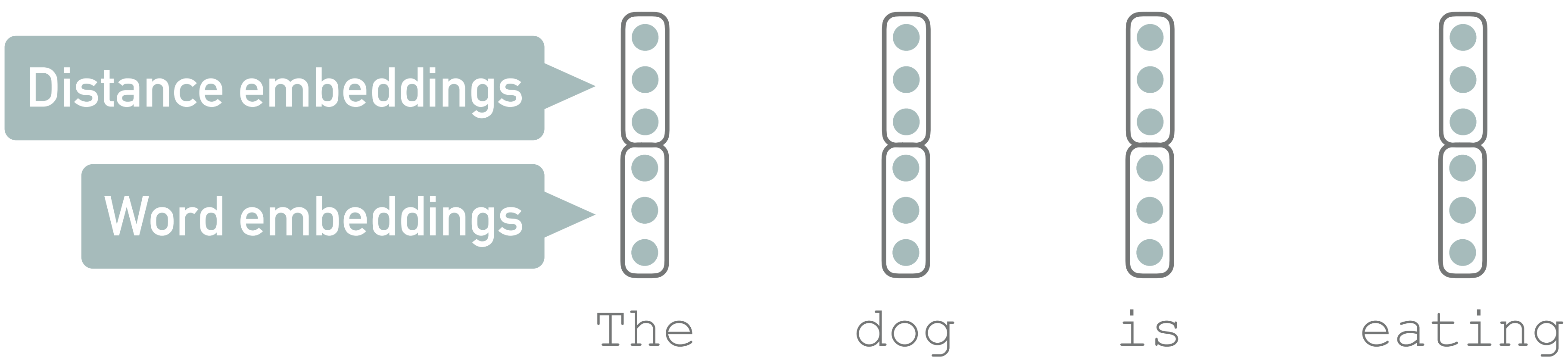
SELF-ATTENTIVE NETWORKS 2/2

Word order information

As such, each head is based on a bag-of-word model! How can take into account word order?

Distance embedding

Concatenate (or sum) a distance embedding to the input



Trainable embeddings

- One embedding for each position in the sentence (1, 2, 3, 4…)
- Randomly initialized and trained end-to-end

But what if at test time we have a sentence longer than any sentence in the training set?!

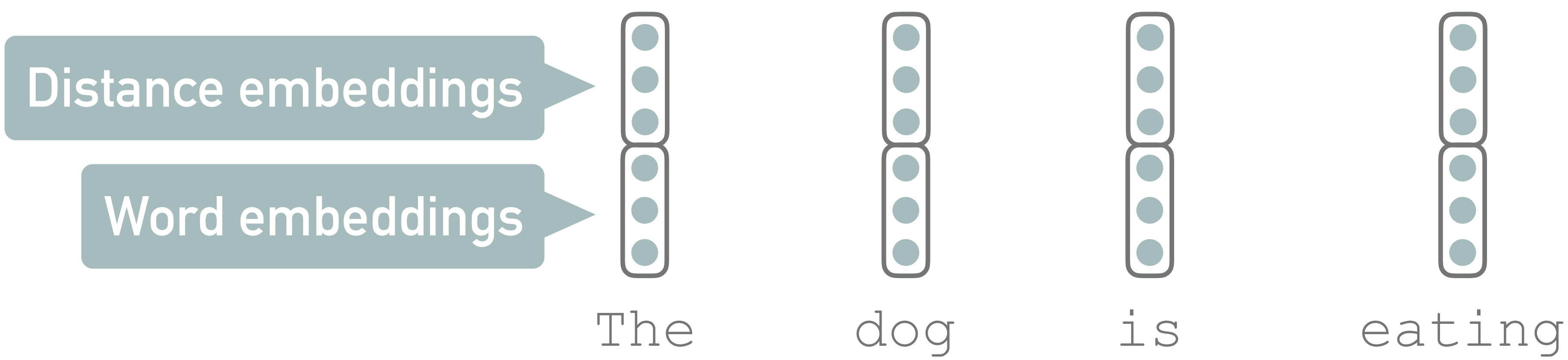
SELF-ATTENTIVE NETWORKS 2/2

Word order information

As such, each head is based on a bag-of-word model! How can take into account word order?

Distance embedding

Concatenate (or sum) a distance embedding to the input



Trainable embeddings

- One embedding for each position in the sentence (1, 2, 3, 4…)
- Randomly initialized and trained end-to-end

But what if at test time we have a sentence longer than any sentence in the training set?!

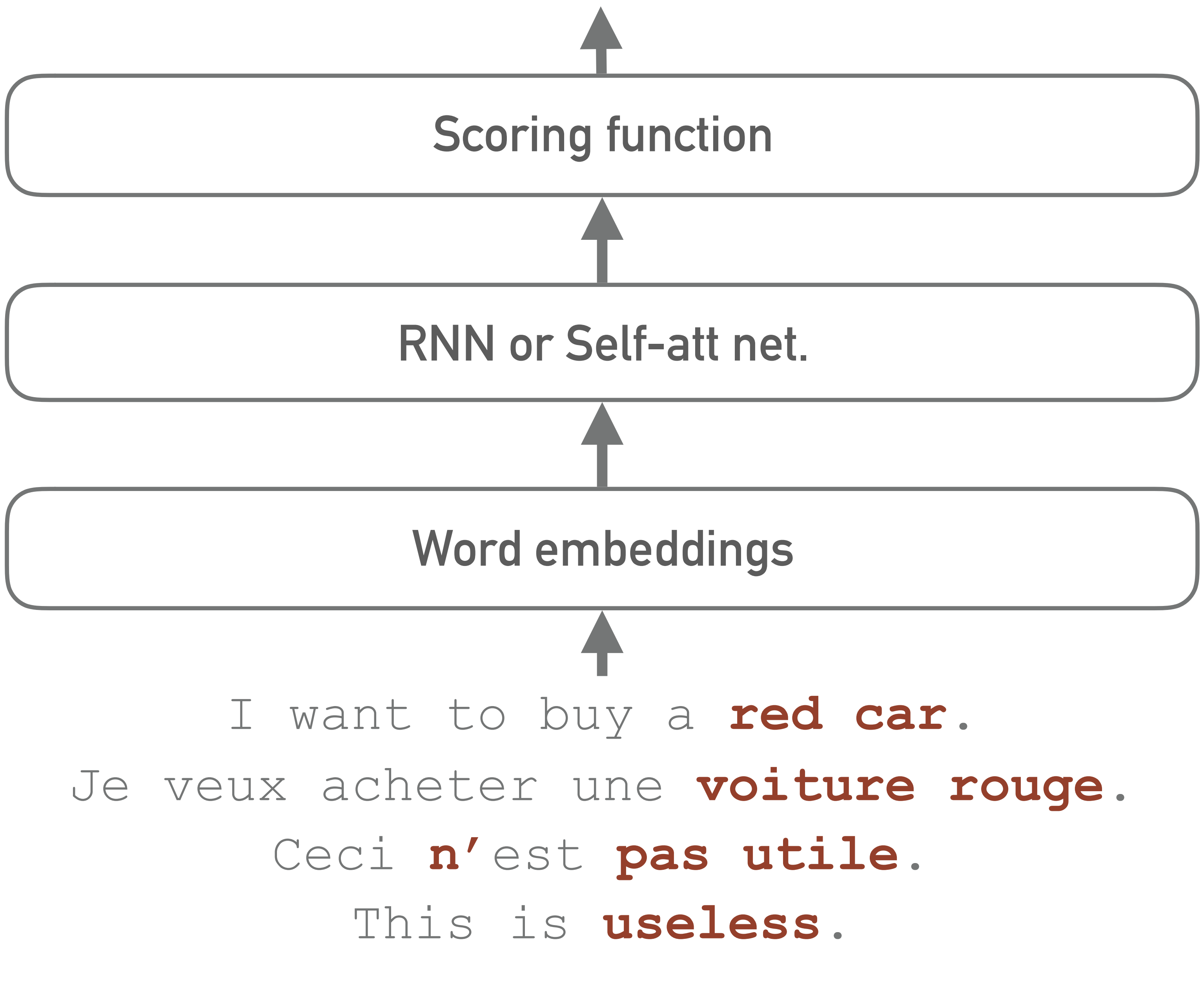
Fixed embeddings

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

TOWARD MULTILINGUAL ARCHITECTURES 2/2

Main idea

- Train one neural network and use it for several languages
- Hope it works: improve results in a multilingual settings + cross-lingual prediction



TOWARD MULTILINGUAL ARCHITECTURES 2/2

Limitations

- Word order is hard-coded in the architectures but differs in different languages
  - RNN: « by nature »
  - Self-attentive networks: via the positional embeddings
- Word embeddings are specific to a given language

MultiBERT

Just pre-train BERT on multilingual data and hope that it works

- Does it really work? And why?
- Does it work for languages where we have little data?

Why should we care about syntax?

- (Language agnostic?) intermediary step for a downstream task
- Tool to understand how to build better multi-lingual neural architectures

SYNTAX

SYNTAX 1/2

What is the syntactic analysis of a sentence?

- Exposes the (syntactic) role of each words
- Exposes of the relation between words in a sentence

The dog is eating.

Dependency: modifier of « dog »

The dog is eating.

Constituent: noun phrase

Concepts that you must not confuse! [Rambow, 2010]

- **Syntactic content:**
  the morphological and syntactic facts of the analyzed sentence
  (Syntactic dependencies? (Lexicalized) constituents?)
- **Representation type:**
  what type of mathematical object is used to represent syntactic facts?
  (Graphs? Phrase structure trees?)

SYNTAX 2/2

On syntactic theories

- There are several syntactic theories
- Different theories are useful in different settings

The simple question that does not make sense

What is the part of speech (or syntactic category) of the word « **their** » in the following sentence?

« `They focus on` **`their`** `own work.` »

Answer: it depends on the syntactic theory!

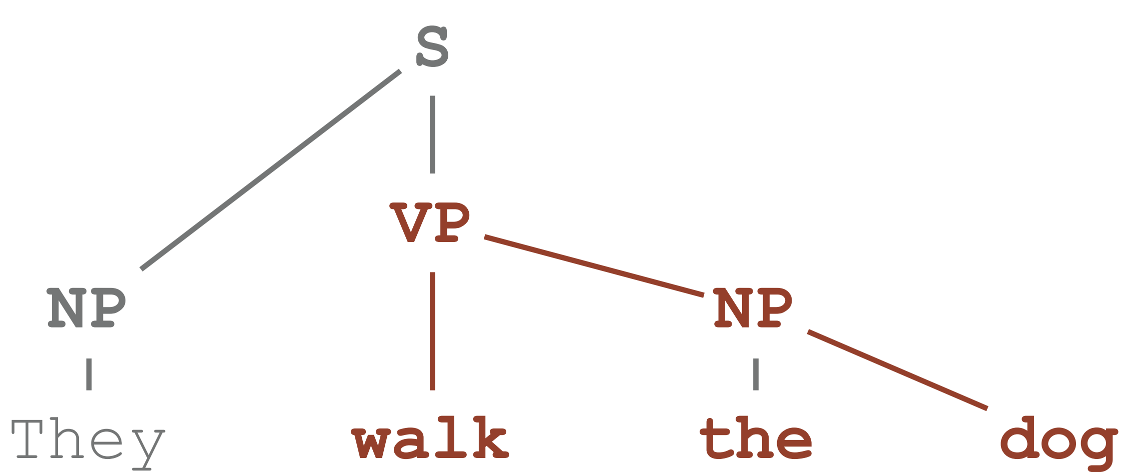
- Penn Treebank style: **Possessive pronoun**
- Universal Dependencies style: **Determiner**

CONSTITUENT ANALYSIS

Definition [Brinton, 2000]

A constituent is a set of words defining a syntactic unit in a hierarchical syntactic structure. As such, it is a part of a sentence that can be moved, modified or deleted alongside agreement adjustments.

Example in the Penn Treebank style



S	clause
NP	Noun Phrase
VP	Verbal Phrase

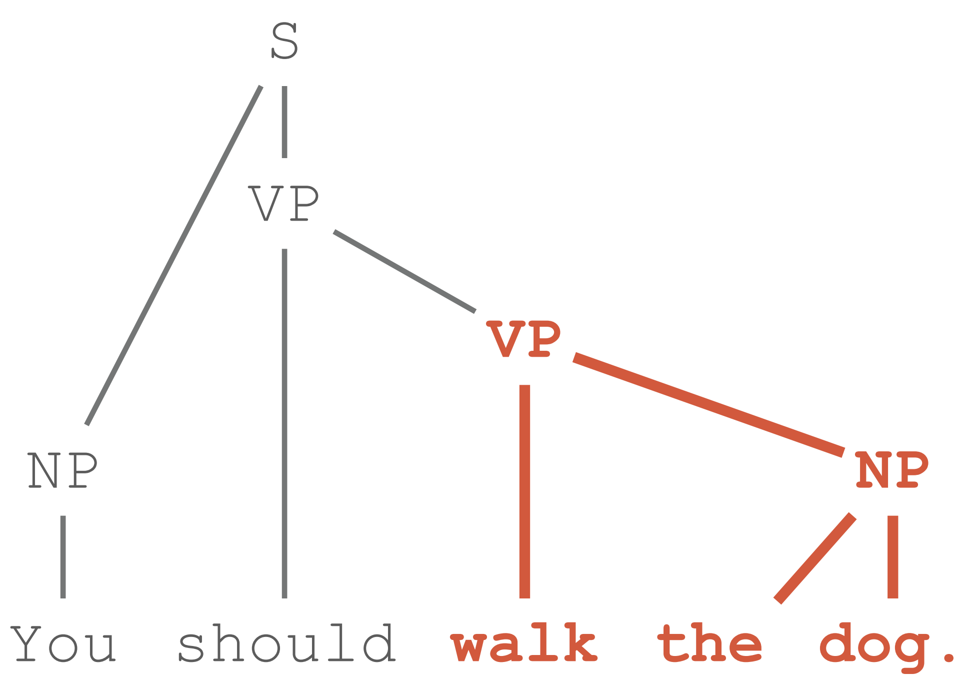
Yield

The **yield** of a constituent is a set of words it dominates
Example: The yield of the VP node is « walk the dog »

CONTINUOUS VS DISCONTINUOUS CONSTITUENT ANALYSIS 1/3

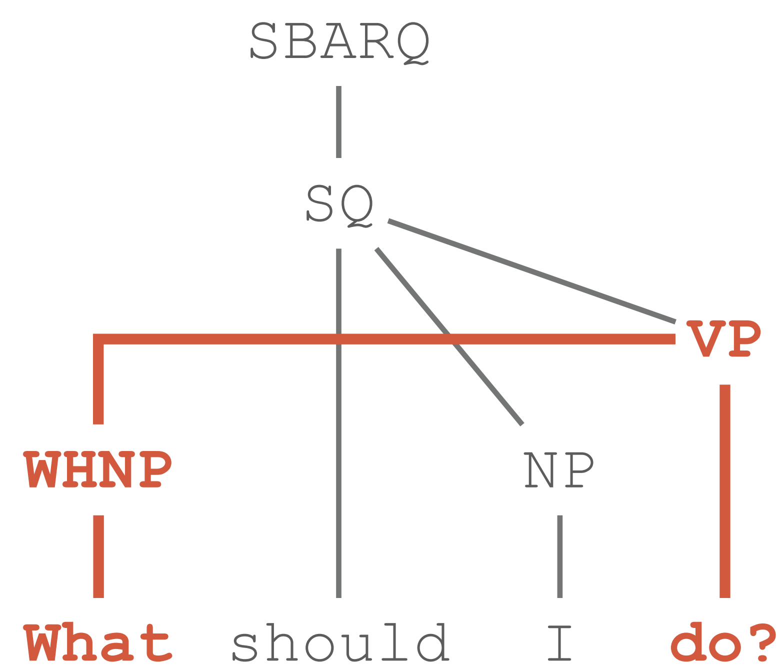
Continuous constituent tree

The yield of each constituent is a contiguous sequence of words



Discontinuous constituent tree

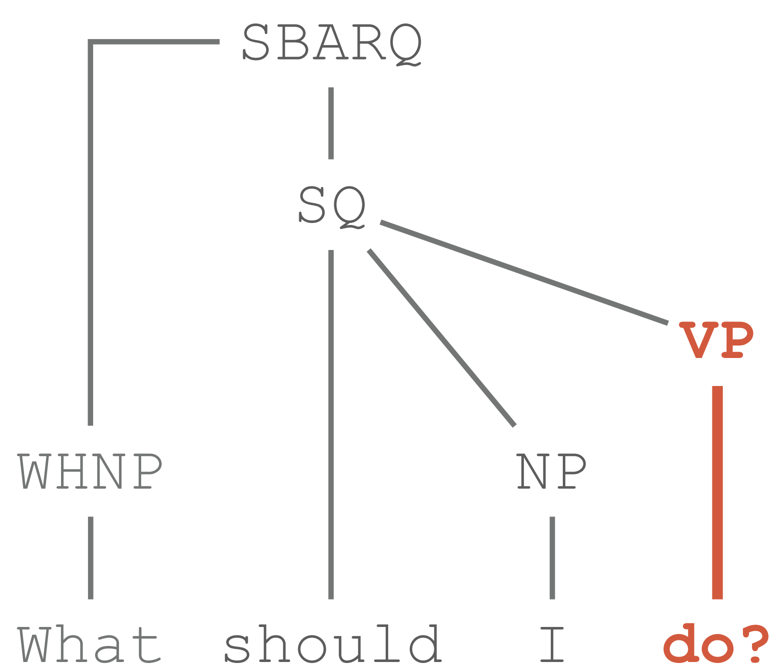
A constituent can yield a non-contiguous sequence of words



CONTINUOUS VS DISCONTINUOUS CONSTITUENT ANALYSIS 2/3

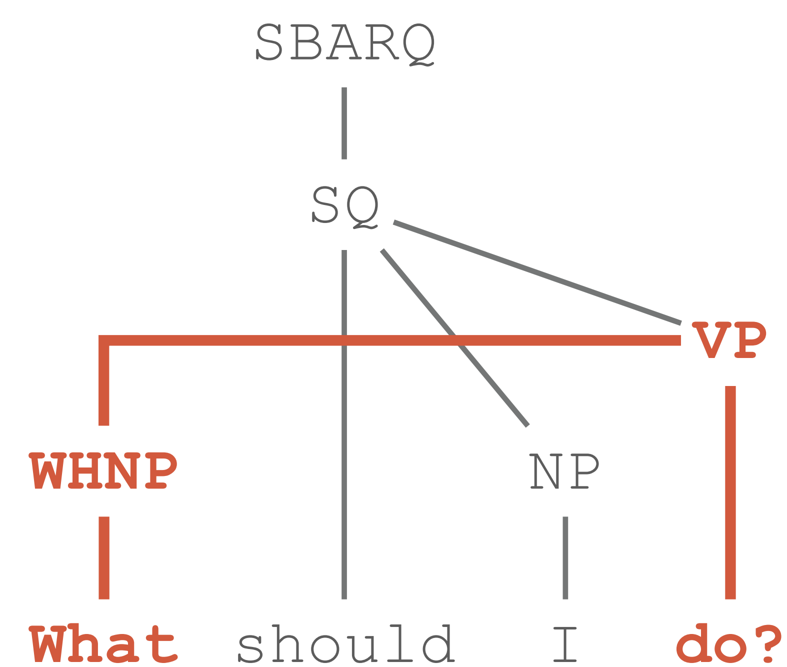
Original Penn Treebank annotation

+ traces to recover the movement origin



Discontinuous variant

Explicit representation of non-local dependencies in the constituency structure

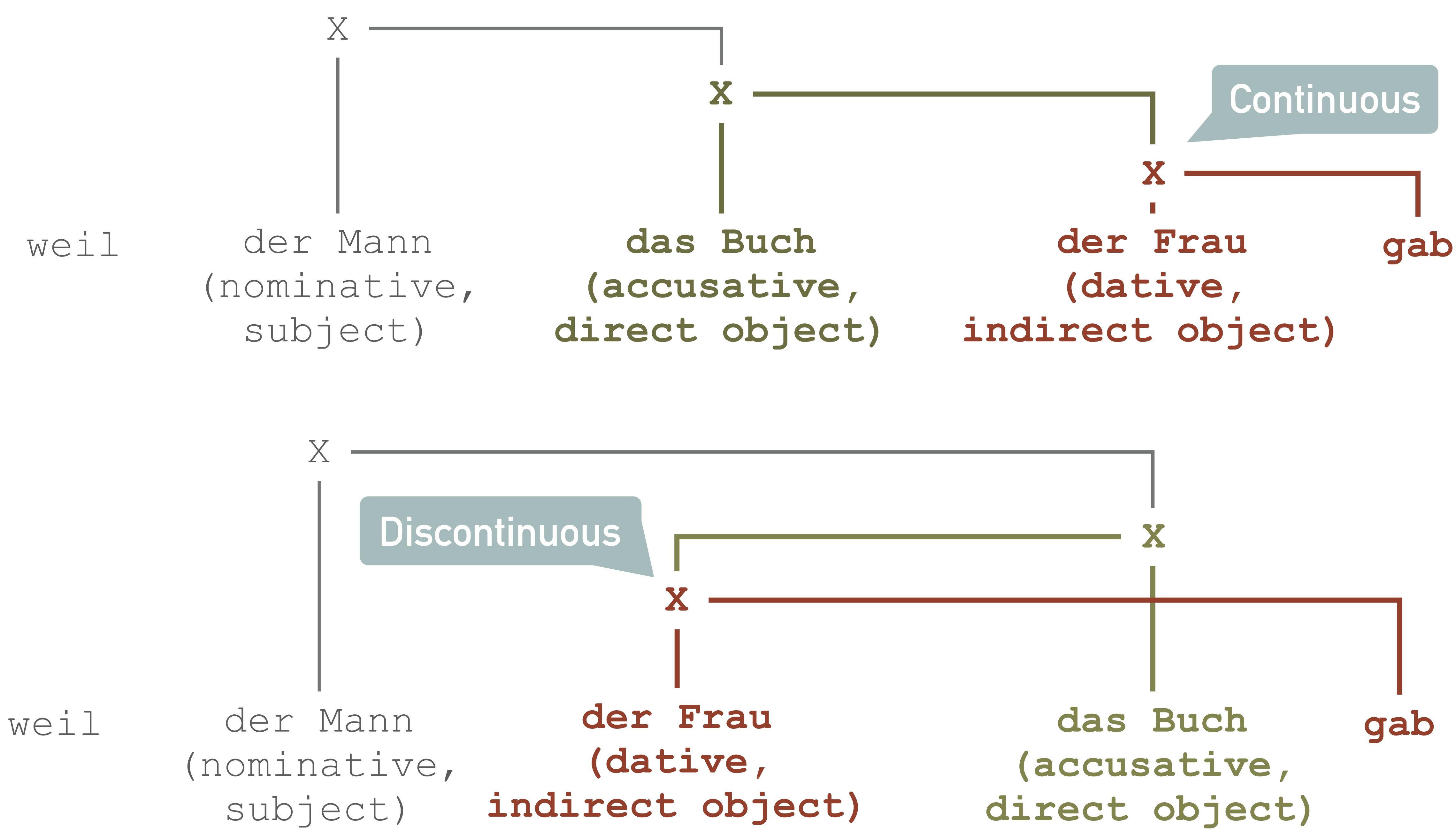


[Evang and Kallmeyer, 2001]

CONTINUOUS VS DISCONTINUOUS CONSTITUENT ANALYSIS 3/3

Do we need to model discontinuous constituents?

- It is arguably « better » in English [McCawley, 1982 ; Bunt et al., 1987]
- It is very import in « free word order » languages like German [Müller, 2004]



DEPENDENCY ANALYSIS

Definition [Brinton, 2000]

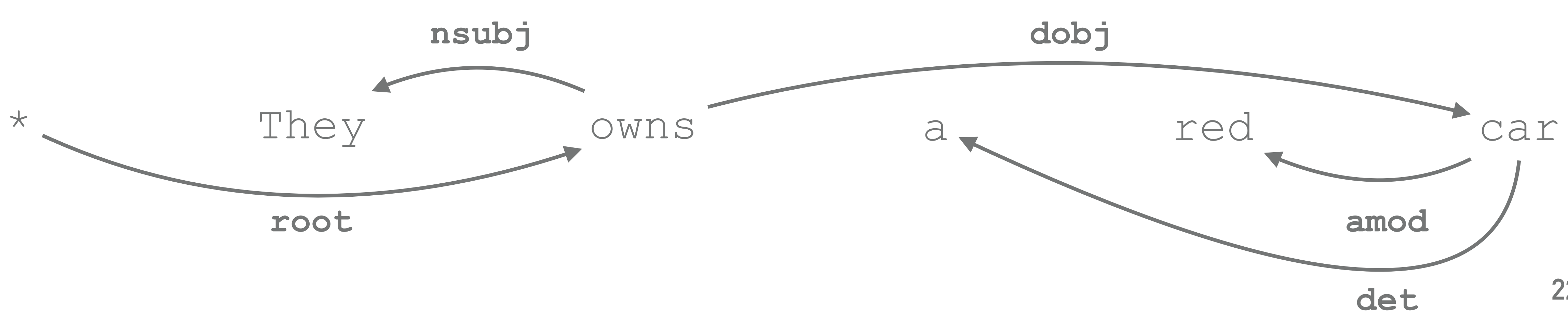
- **Head-modifier relationship:** given two words, their relationship is qualified as head-modifier if the modifier word is optional modifier of the head
- **Governor-complement relationship:** given two words, their relationship is qualified as governor-complement if both words are mutually dependent: one cannot occur without the other (e.g. predicate-subject relationship)

Example

- « red » is a modifier of its head « car »
- « they » is a complement of its governor « owns »

« They own a red car »

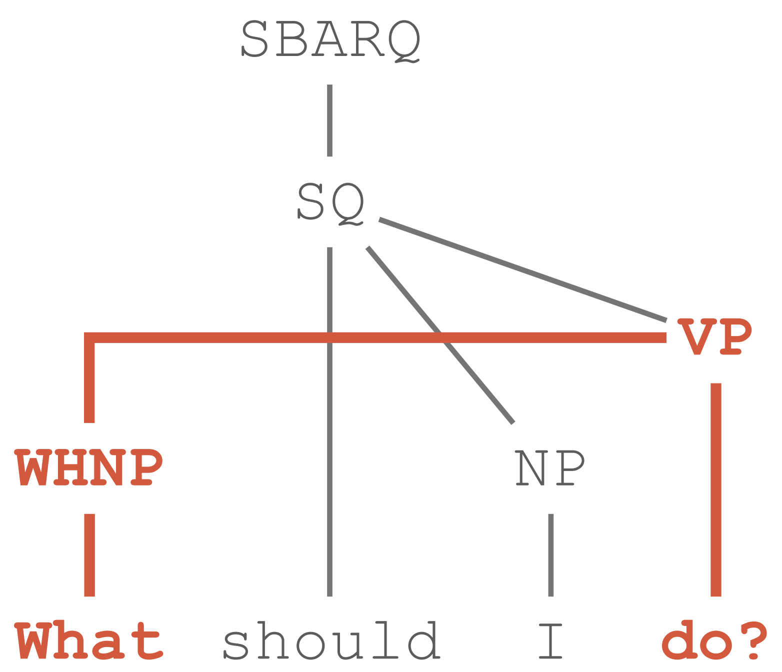
Full example using Stanford Dependencies



DISCONTINUOUS CONSTITUENTS VS NON-PROJECTIVE DEPS

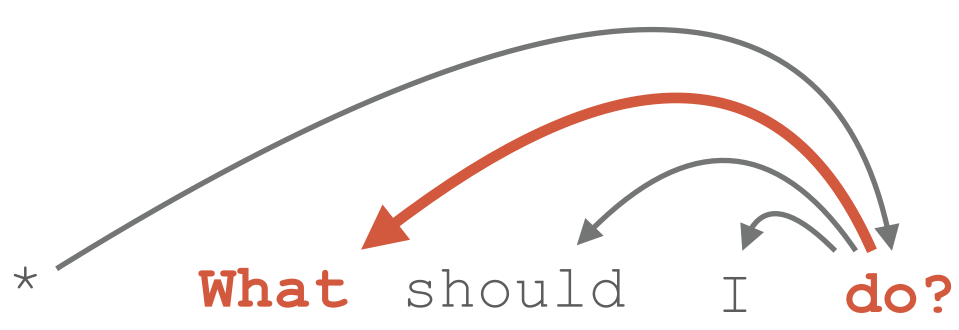
Projective dependency

- A dependency A->B is projective if all words between A and B are descendant of A
- Non-projective otherwise

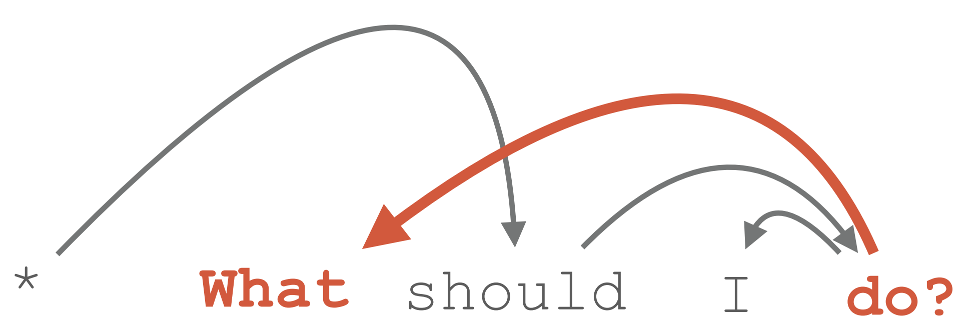


The dependency representation depends on the syntactic theory

- If the root of the sentence is « do »:



- If the root of the sentence is « should »:



THE PARSING PROBLEM

Constituency/dependency parsing

Given a sentence, predict its constituency/dependency structure

Intuition of the neural approach

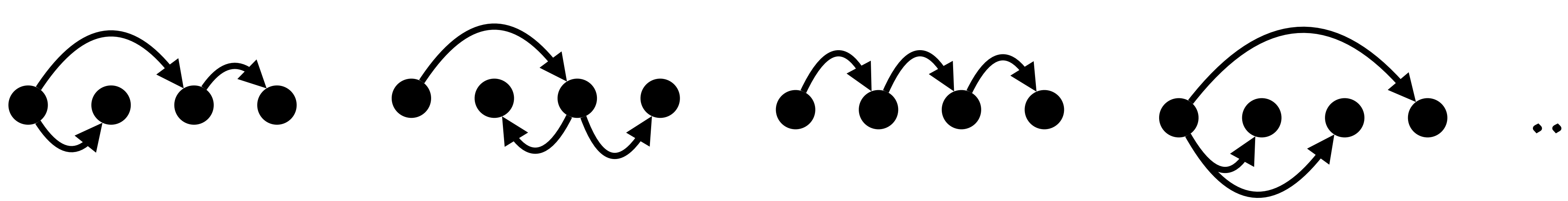
1. A neural network compute a score associated with each possible part
   - ➤ Constituency parsing: score every possible labeled constituent
   - ➤ Dependency parsing: score every possible labeled dependency
2. Compute the coherent structure of maximum score where

How? It depends…

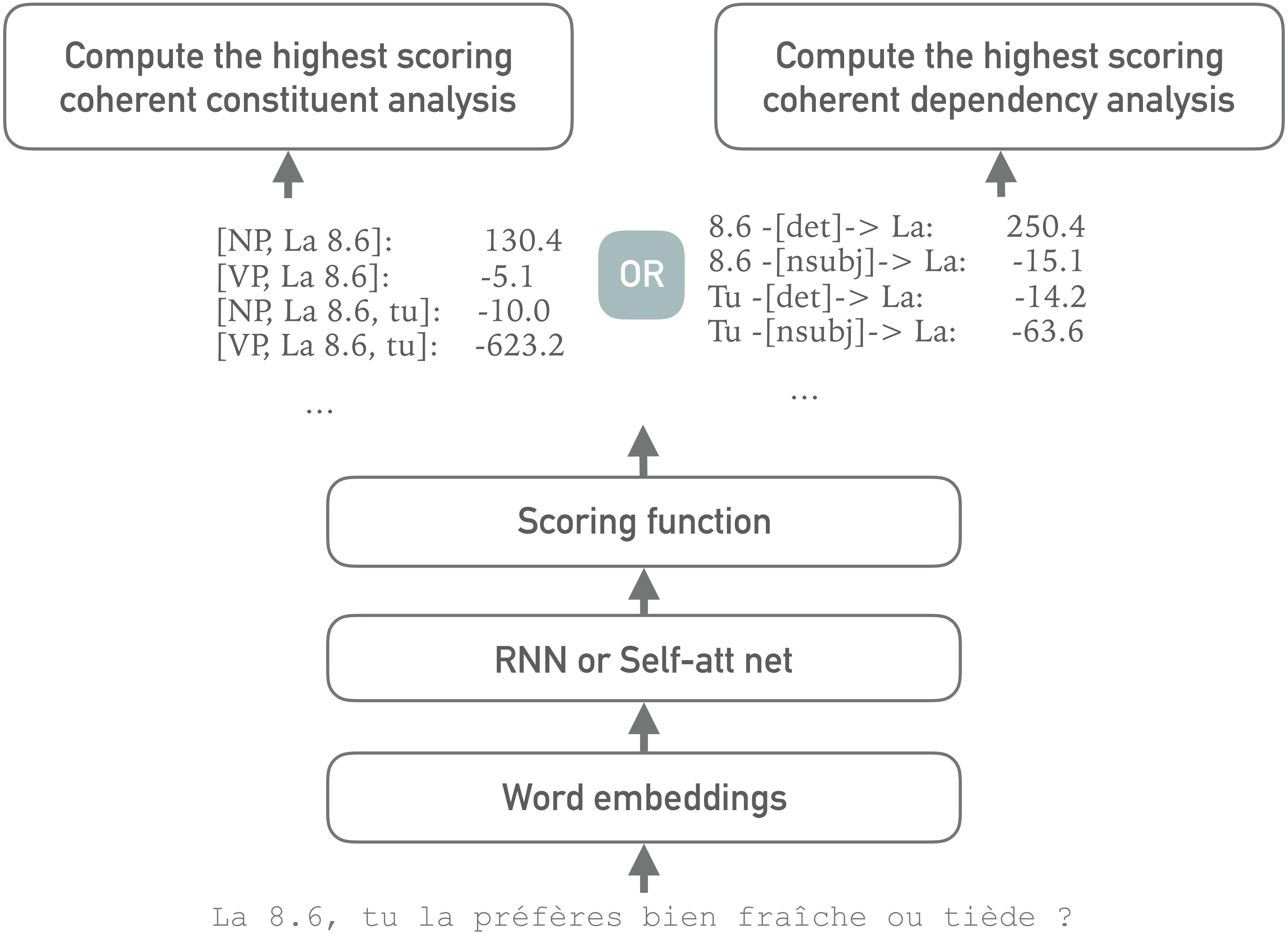
score of a structure = sum of the score of its parts

Example for dependency parsing

Compute the arborescence of maximum score given the score of each possible arc



ARCHITECTURE OF A NEURAL PARSER



LINGUISTIC TYPOLOGY

LINGUISTIC TYPOLOGY

How to classify languages with respect to their features?

- **Morphological features:** how do morphemes combine to build words?
  Example: « hopelessness » contains 3 morphemes « hope », « less » and « ness »
- **Syntactic features:** how do words combine to build sentences? i.e. word order

https://wals.info/

MORPHOLOGICAL DIFFERENCES 1/2

[Haspelmath and Sims, 2002]

Analytics or « morphologically poor » languages

Languages where morphology plays a (relatively) modest role:
Yoruba, Vietnamese, English, …

(1.2) Yoruba

Nwọn	*ó*	*maa*	*gbà*	*pọ́nùn*	*mẹ́wǎ*	*lọ́sọ̀ọ̀sẹ̀.*
they	FUT	PROG	get	pound	ten	weekly

'They will be getting £10 a week.'

(Rowlands 1969: 93)

(1.3) Vietnamese

Hai	*đú.a*	*bo*$^{?}$	*nhau*	*là*	*tại*	*gia-đình*	*thàng*	*chông.*
two	individual	leave	each.other	be	because.of	family	guy	husband

'They divorced because of his family.'

(Nguyen 1997: 223)

Synthetic of « morphologically rich » languages

Languages where morphology plays a (relatively) important role:
Sumerian, Swahili, Lezgian, …

(1.4) Swahili

Ndovu	*wa-wili*	*wa-ki-song-ana*	*zi-umia-zo*	*ni*	*nyika.*
elephants	PL-two	3PL-SUBORD-jostle-RECP	3SG-hurt-REL	is	grass

'When two elephants jostle, what is hurt is the grass.'

(Ashton 1947: 114)

(1.5) Lezgian

Marf-adi	*wiči-n*	*qalin*	*st'al-ra-ldi*	*qaw*	*gata-zwa-j.*
rain-ERG	self-GEN	dense	drop-PL-INS	roof	hit-IMPF-PST

'The rain was hitting the roof with its dense drops.'

(Haspelmath 1993: 140)

(1.6) West Greenlandic

Paasi-nngil-luinnar-para	*ilaa-juma-sutit.*
understand-not-completely-1SG.SBJ.3SG.OBJ.IND	come-want-2SG.PTCP

'I didn't understand at all that you wanted to come along.'

(Fortescue 1984: 36)

MORPHOLOGICAL DIFFERENCES 2/3

[Haspelmath and Sims, 2002]

Language	Ratio of morphemes per word
West Greenlandic	3.72
Sanskrit	2.59
Swahili	2.55
Old English	2.12
Lezgian	1.93
German	1.92
Modern English	1.68
Vietnamese	1.06

Table 1.1 The degree of synthesis of some languages

Source: based on Greenberg (1959), except for Lezgian

Morphology plays a relatively important role

Morphology plays a relatively modest role

MORPHOLOGICAL DIFFERENCES 2/3

« Tu **dormiras** demain. » « You **will** sleep tomorrow. »

Feature 67A: The Future Tense

This feature is described in the text of chapter 67 The Future Tense by Östen Dahl and Viveka Velupillai cite

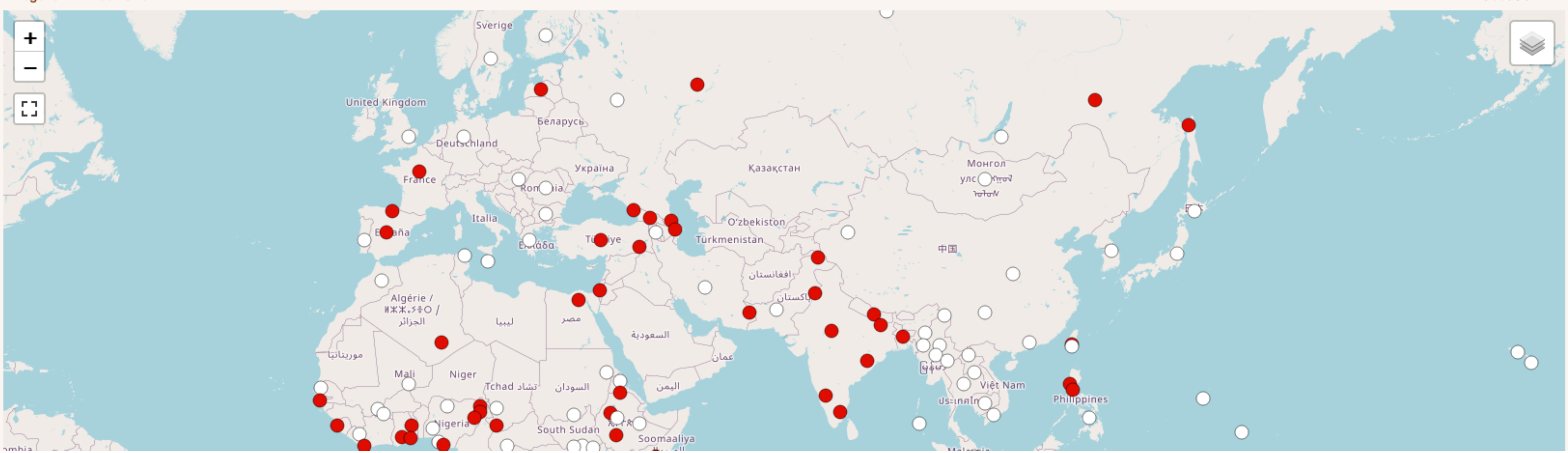
You may combine this feature with another one. Start typing the feature name or number in the field below.

× 67A: The Future Tense Submit

Values

●	Inflectional future exists	110
○	No inflectional future	112

Legend Icon size Show/hide Labels GeoJSON



MORPHOLOGICAL DIFFERENCES 2/3

Negative particle in French

« Je **n'** ai **pas** mangé de pomme »

Negative auxiliary verb in Finnish

« **En** syönyt omenaa »

Inflection for person and number

Indicative, conditional, and potential

Person	Singular	Plural
1.	en	emme
2.	et	ette
3.	ei	eivät

Imperative

Person	Singular	Plural
1.	-	älkäämme
2.	älä	älkää
3.	älköön	älkööt

https://en.wikipedia.org/wiki/Negative_verb#Finnish

Feature 112A: Negative Morphemes

This feature is described in the text of chapter 112 Negative Morphemes by Matthew S. Dryer cite

You may combine this feature with another one. Start typing the feature name or number in the field below.

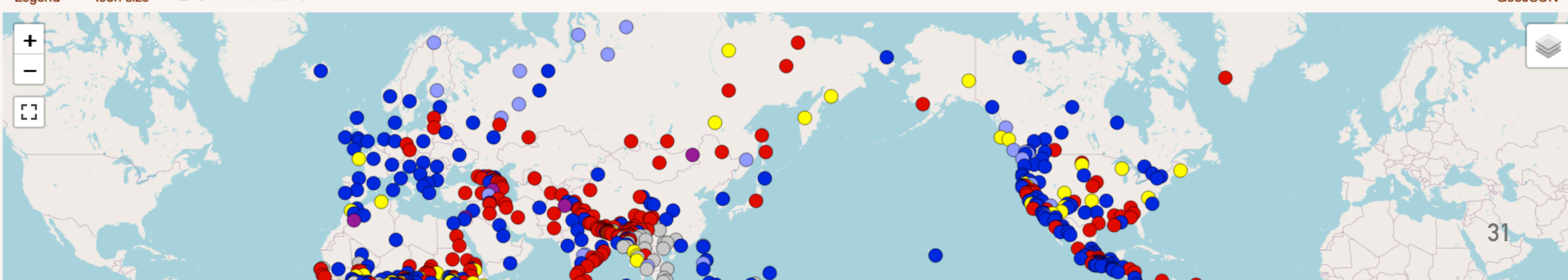
× 112A: Negative Morphemes

Submit

Values

Negative affix	395
Negative particle	502
Negative auxiliary verb	47
Negative word, unclear if verb or particle	73
Variation between negative word and affix	21
Double negation	119

Legend Icon size Show/hide Labels GeoJSON



SYNTACTIC DIFFERENCES 1/3

Feature 81A: Order of Subject, Object and Verb

This feature is described in the text of chapter 81 Order of Subject, Object and Verb by Matthew S. Dryer cite

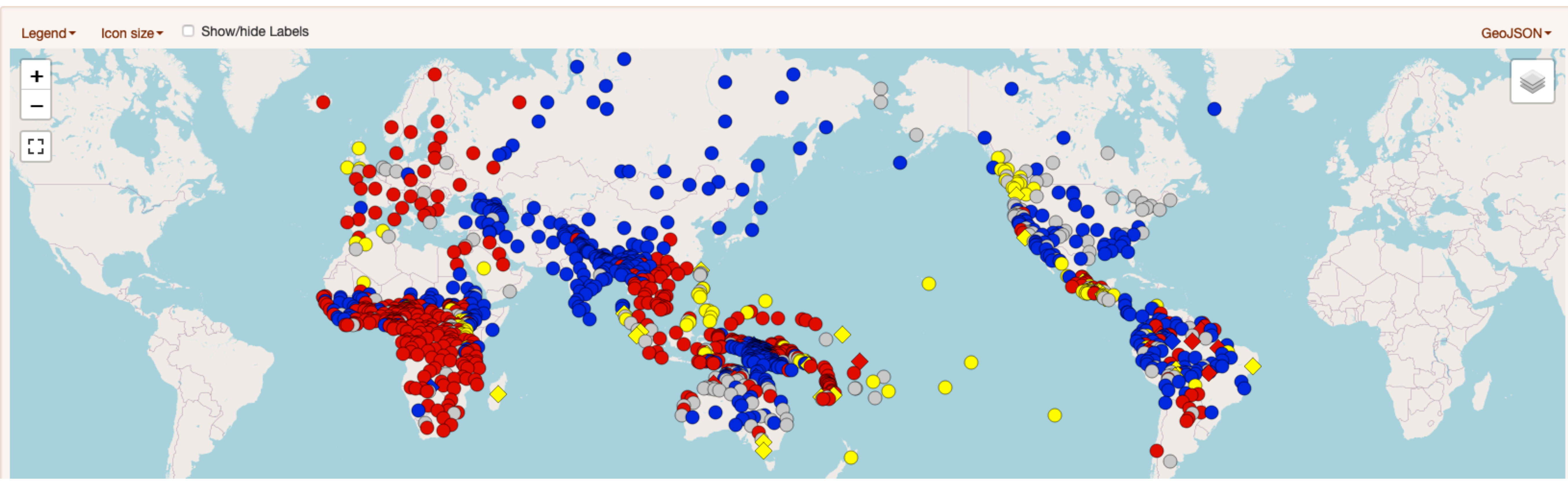
You may combine this feature with another one. Start typing the feature name or number in the field below.

× 81A: Order of Subject, Object and Verb

Submit

Values

	SOV	564
	SVO	488
	VSO	95
	VOS	25
	OVS	11
	OSV	4
	No dominant order	189



SYNTACTIC DIFFERENCES 2/3

Warning

WALS document the « dominant » word order, variations may exists!

« La **petite** maison **rouge** »

Feature 87A: Order of Adjective and Noun

This feature is described in the text of chapter 87 Order of Adjective and Noun by Matthew S. Dryer cite

You may combine this feature with another one. Start typing the feature name or number in the field below.

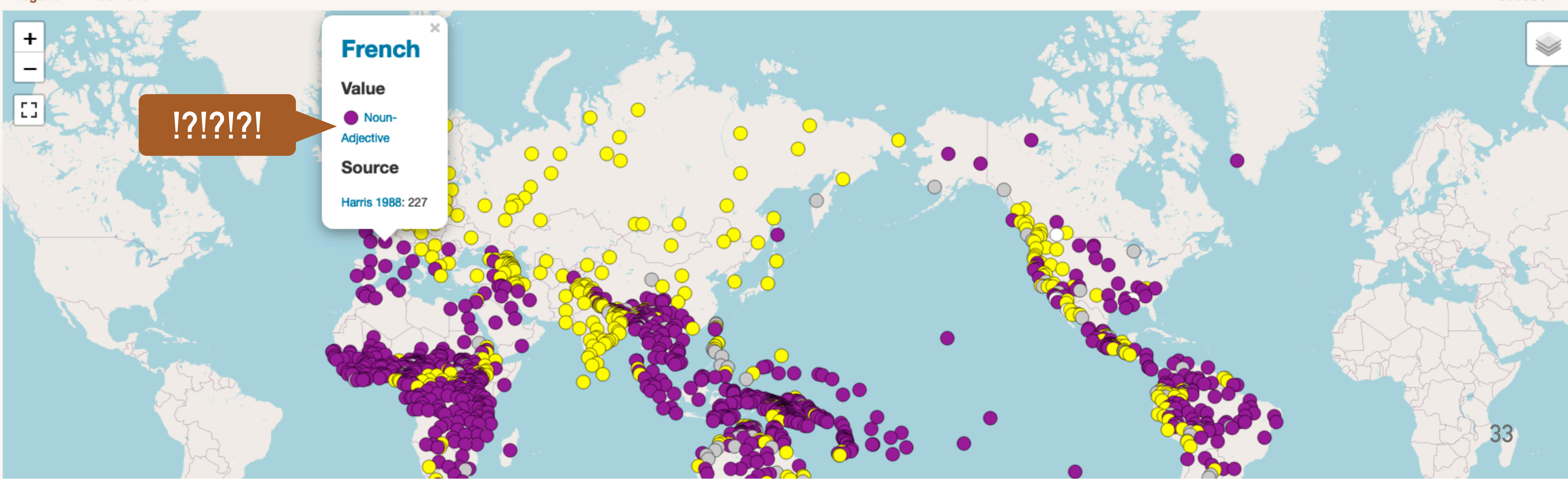
× 87A: Order of Adjective and Noun

Submit

Values

●	Adjective-Noun	373
●	Noun-Adjective	879
●	No dominant order	110
○	Only internally-headed relative clauses	5

Legend ▾ Icon size ▾ Show/hide Labels GeoJSON ▾



SYNTACTIC DIFFERENCES 3/3

Feature 81A: Order of Subject, Object and Verb

This feature is described in the text of chapter 81 Order of Subject, Object and Verb by Matthew S. Dryer cite

You may combine this feature with another one. Start typing the feature name or number in the field below.

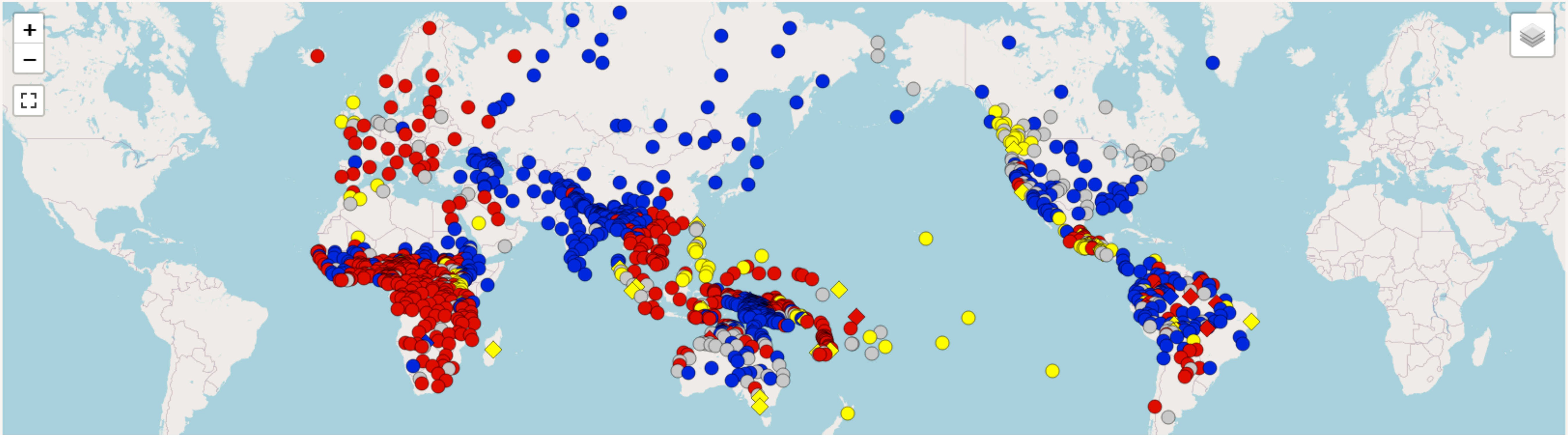
× 81A: Order of Subject, Object and Verb

Submit

Values

●	SOV	564
●	SVO	488
●	VSO	95
◆	VOS	25
◆	OVS	11
◆	OSV	4
●	No dominant order	189

Legend Icon size Show/hide Labels GeoJSON



MORPHOLOGY VS SYNTAX

« Tu vas manger. »

« **Est-ce que** tu vas manger ? »

MWE that indicates the question

« You **will** eat. »

« **Will** you eat? »

Word-order that indicates the

Feature 116A: Polar Questions

This feature is described in the text of chapter 116 Polar Questions by Matthew S. Dryer cite

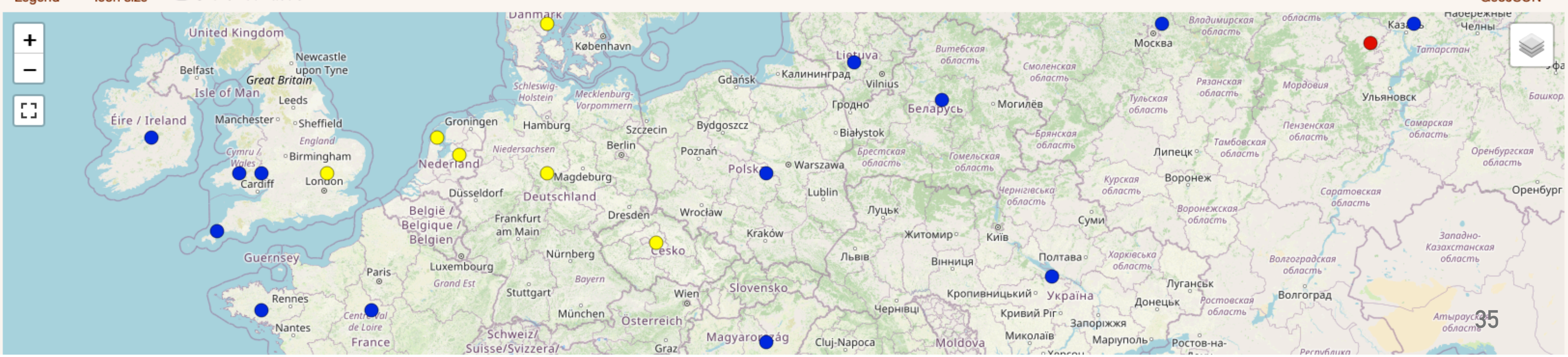
You may combine this feature with another one. Start typing the feature name or number in the field below.

× 116A: Polar Questions Submit

Values

●	Question particle	585
●	Interrogative verb morphology	164
●	Mixture of previous two types	15
●	Interrogative word order	13
◆	Absence of declarative morphemes	4
●	Interrogative intonation only	173
○	No interrogative-declarative distinction	1

Legend▾ Icon size▾ Show/hide Labels GeoJSON▾

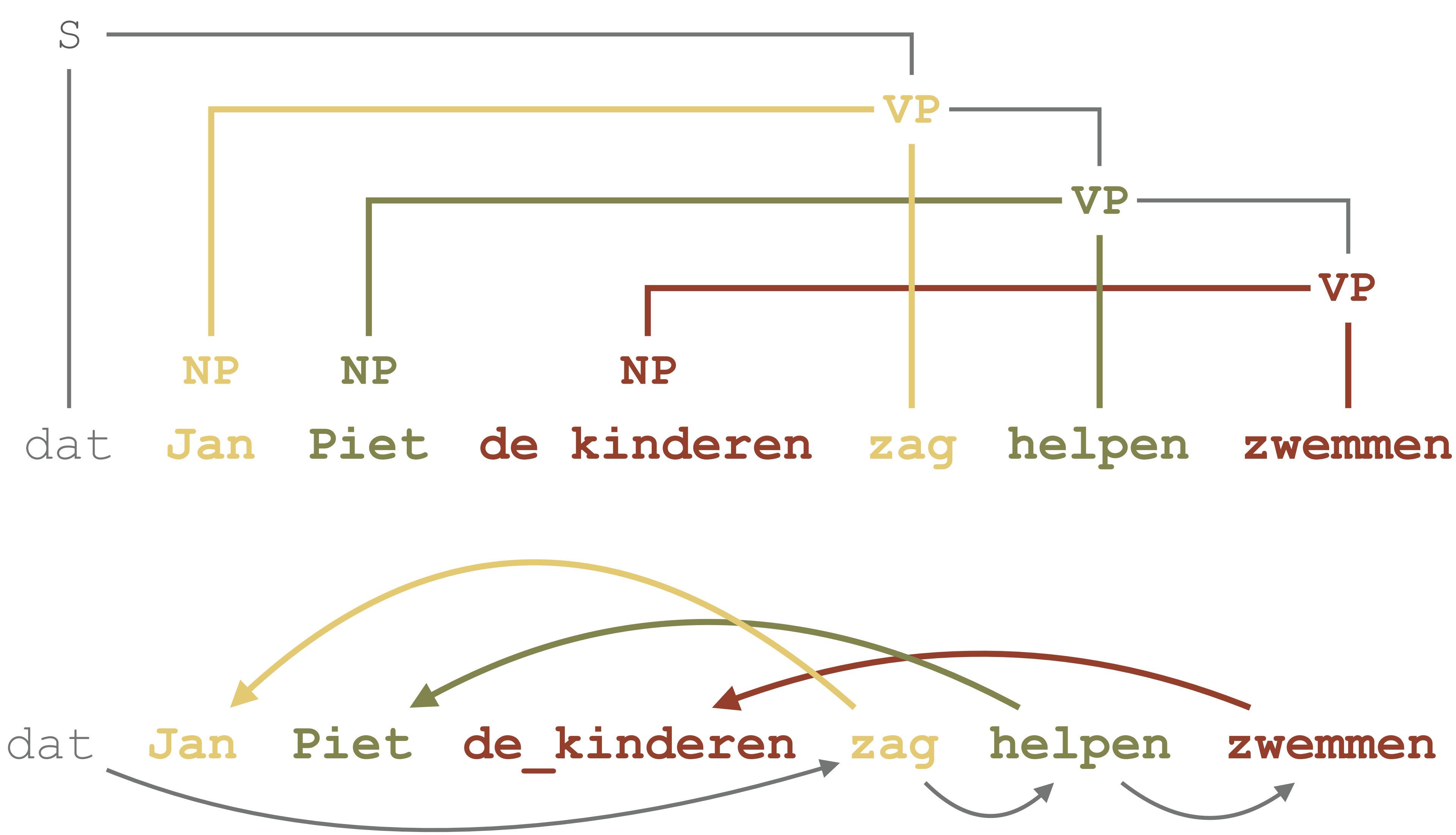


STRUCTURAL DIFFERENCES

Intuition

It's not only about word order, certain types of structures are not shared across languages

Example: cross-serial dependencies [Kallmeyer, 2010]



UNIVERSAL DEPENDENCIES

THE UNIVERSAL DEPENDENCY PROJECT

Universal Dependencies

- Annotation scheme to build cross-linguistically consistent treebank
- Over 100 languages and 200 treebanks
- Over 300 contributors
- Include code-switching treebanks (i.e. sentences that mix several languages)

Figure 5: Code-switching tweet showing grammatical fragments from Hindi and English.

[Bhat et al., 2018]

What can we do with that?

- Multilingual parsing
  (train and parse several languages with the same network)
- Cross-lingual parsing
  (train on one or several languages and evaluate on other languages)
- Research from a language typology perspective

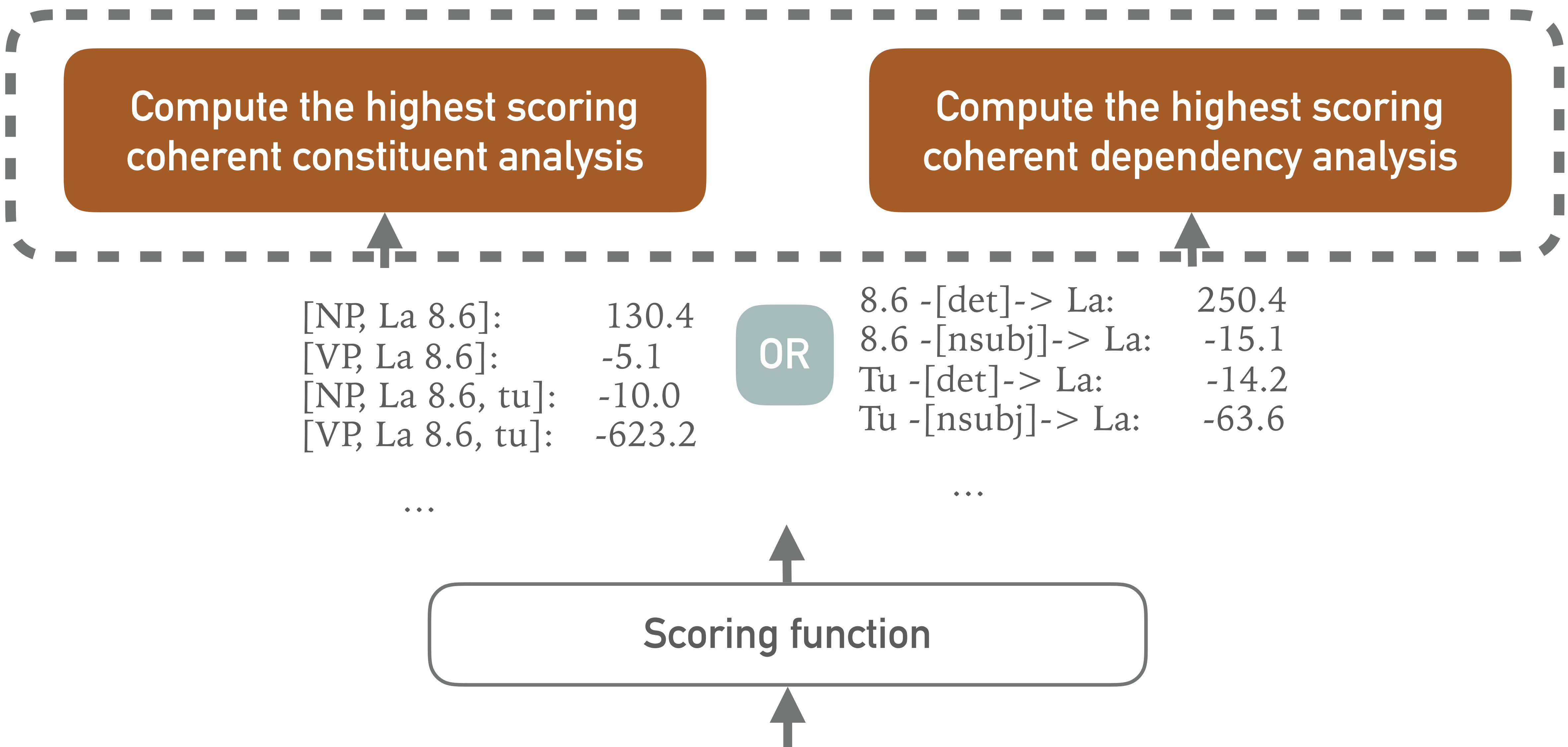
https://universaldependencies.org/

ONE BENEFIT OF DEPENDENCIES 1/2

Algorithmic complexity

What is the complexity of the prediction step?

- Depends on the structure: constituency or dependency parsing?
- Depends on the search space: continuous? discontinuous? What kind of discontinuity?



ONE BENEFIT OF DEPENDENCIES? 2/2

Intuition

We want the larger search space possible in order to be able to parse a maximum of linguistic phenomena (e.g. cross-serial dependencies)

Constituency parsing complexity (among many other algorithms!)

Increasing search space

Context-free:	$\mathcal{O}(n^3)$	[Sakai, 1961]
Well-nested + 2-bounded block degree: (Tree adjoining grammars, LCFRS)	$\mathcal{O}(n^6)$	[Joshi, 1987]
Joint tagging and parsing:	NP-hard	[Corro et al., 2017]

Dependency parsing complexity (among many other algorithms!)

Increasing search space

Projective:	$\mathcal{O}(n^3)$	[Eisner, 2000]
Well-nested + 2-bounded block degree:	$\mathcal{O}(n^7)$	[Gómez-Rodríguez et al. 2009]
Well-nested + k-bounded block degree, k > 2:	$\mathcal{O}(n^{5+2(k-1)})$	
k-bounded block degree, k > 2:	NP-complete	[Satta, 1992]
Non-projective:	$\mathcal{O}(n^2)$	[McDonald et al., 2005]

UNIVERSAL PART-OF-SPEECH TAGS [Petrov et al., 2012 ; Nivre et al. 2016]

17 categories that exist in most languages (whatever this means)

Open class words		Closed class words		Other	
ADJ	adjective	ADP	preposition/postposition	PUNCT	punctuation
ADV	adverb	**AUX**	auxiliary	**SYM**	symbol
INTJ	interjection	CONJ	coordinating conjunction	X	unspecified POS
NOUN	noun	DET	determiner		
PROPN	proper noun	NUM	numeral		
VERB	verb	PART	particle		
		PRON	pronoun		
		SCONJ	subordinating conjunction		

Basic unit is syntactic words

- Clitics are split off
  (Clitics = syntactically independent but phonologically dependent morphems)
- Contractions are undone

Example in Spanish (give me it) **Example in French**

« dámelo » => « dá me lo » « au » => « à le »

```
48      actes   acte    NOUN    _       Gender=Masc|Number=Plur 46      obj     _       _
49      inutiles        inutile ADJ     _       Gender=Masc|Number=Plur 48      amod    _       _
50      et      et      CCONJ   _       _       53      cc      _       _
51-52   au      _       _       _       _       _       _       _       _
51      à       à       ADP     _       _       53      case    _       _
52      le      le      DET     _       Definite=Def|Gender=Masc|Number=Sing|PronType=Art       53      det     _       _
53      dépassement     dépassement     NOUN    _       Gender=Masc|Number=Sing 44      conj    _       _
```

EXAMPLE OF TAGSET CONVERSION TABLE

Tagset en::penn, total 48 tags.

#	=>	SYM	_	*#*
$	=>	SYM	_	*$, C$, US$, A$, HK$*
"	=>	PUNCT	PunctSide=Fin|PunctType=Quot	*", '*
,	=>	PUNCT	PunctType=Comm	*,, 2, an*
-LRB-	=>	PUNCT	PunctSide=Ini|PunctType=Brck	
-RRB-	=>	PUNCT	PunctSide=Fin|PunctType=Brck	
.	=>	PUNCT	PunctType=Peri	*., ?, !*
:	=>	PUNCT	_	*--, :, ;, ..., -*
AFX	=>	ADJ	Hyph=Yes	
CC	=>	CCONJ	_	*and, or, but, &, nor*
CD	=>	NUM	NumType=Card	*million, billion, one, two, three*
DT	=>	DET	_	*the, a, an, this, some*
EX	=>	PRON	AdvType=Ex	*there*
FW	=>	X	Foreign=Yes	*de, perestroika, glasnost, vs., naczelnik*
HYPH	=>	PUNCT	PunctType=Dash	
IN	=>	ADP	_	*of, in, for, on, that*
JJ	=>	ADJ	Degree=Pos	*new, other, last, such, first*
JJR	=>	ADJ	Degree=Cmp	*more, higher, lower, less, better*
JJS	=>	ADJ	Degree=Sup	*most, least, largest, latest, best*
LS	=>	X	NumType=Ord	*3, 2, 1, 4, First*
MD	=>	VERB	VerbType=Mod	*will, would, could, can, may*

https://universaldependencies.org/tagset-conversion/

SYNTACTIC STRUCTURE CONVERSION

Definitions

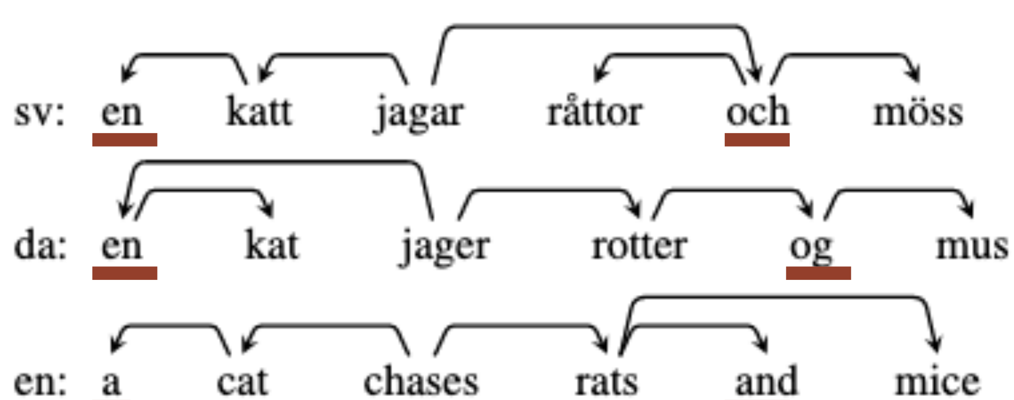
- **Content word:**
  Words that mainly contribute to the meaning
- **Function word:**
  Words that mainly express grammatical relationships (no or almost no meaning)

Intuition

- Function words in one language is « often » correspond to morphological inflection in another language
- To increase to probability of parallel structures, make content words simple « modifiers »

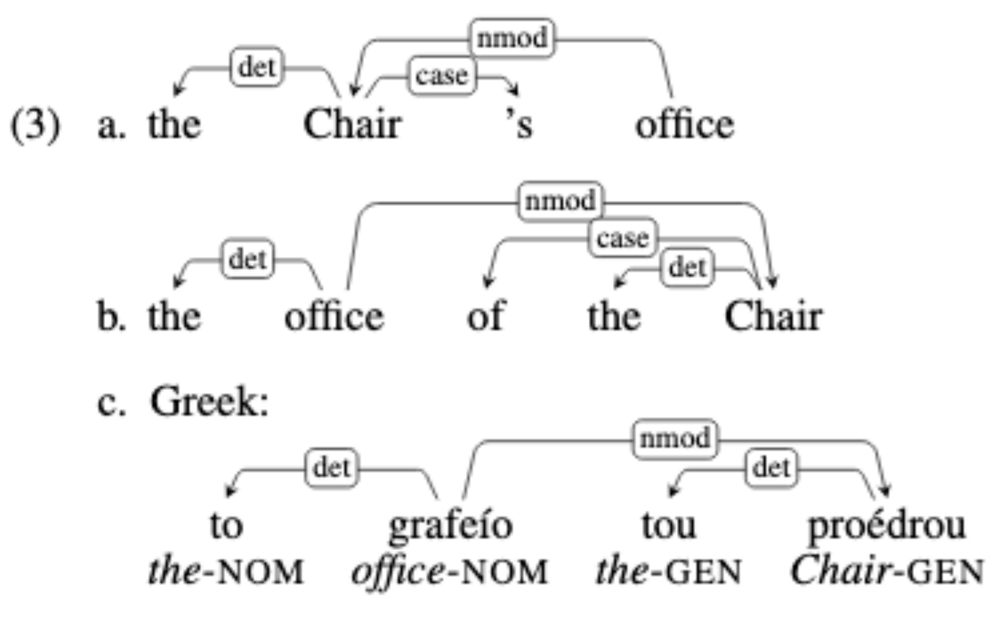
Example of function words annotation in different treebanks (not UD)

[Nivre at al., 2016]



CONTENT WORD AS HEAD WORD

- Using content word as the principle head/governors is **unusual** in syntactic theories
- This representation is very close to nucleus-based dependency structures where function words are merged with content words [Tesnière, 1959]



[Nivre at al., 2016]

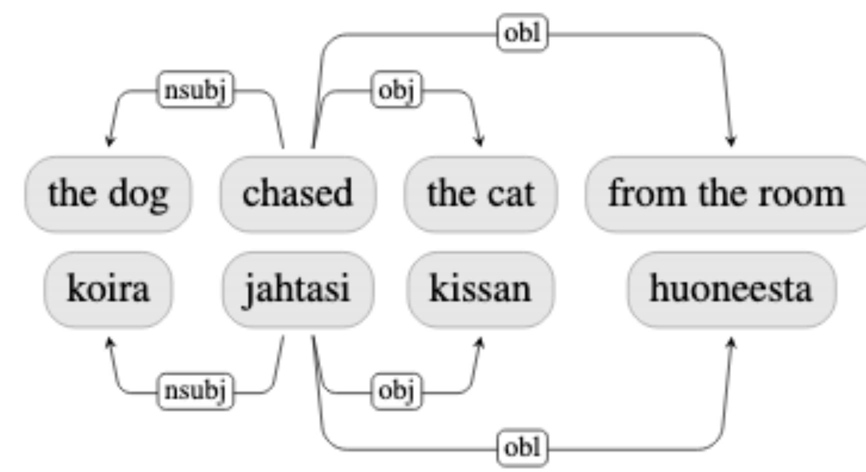
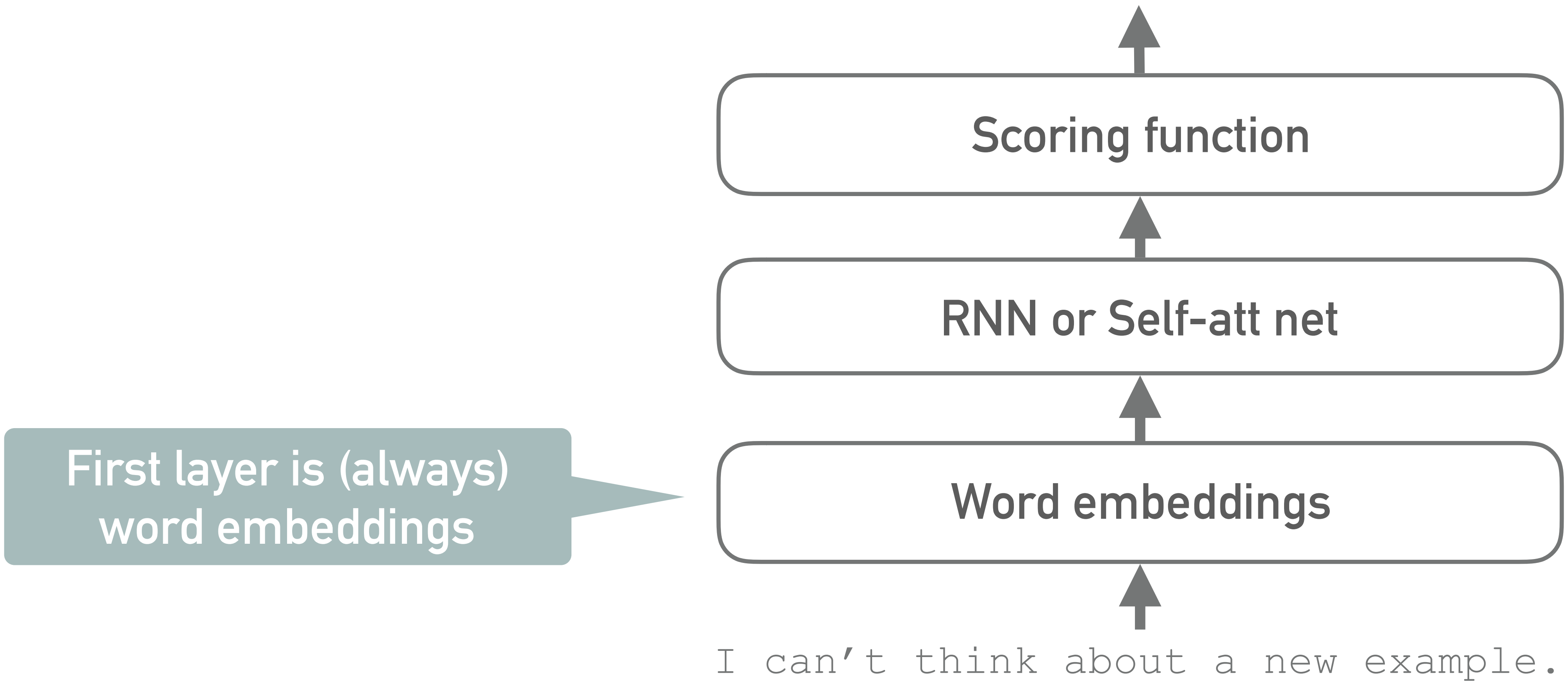


Figure 2: Nucleus-based dependency trees for equivalent sentences from English (top) and Finnish (bottom).

[Basirat and Nivre, 2021]

TOWARDS NEURAL ARCHITECTURE FOR MULTILINGUAL NLP

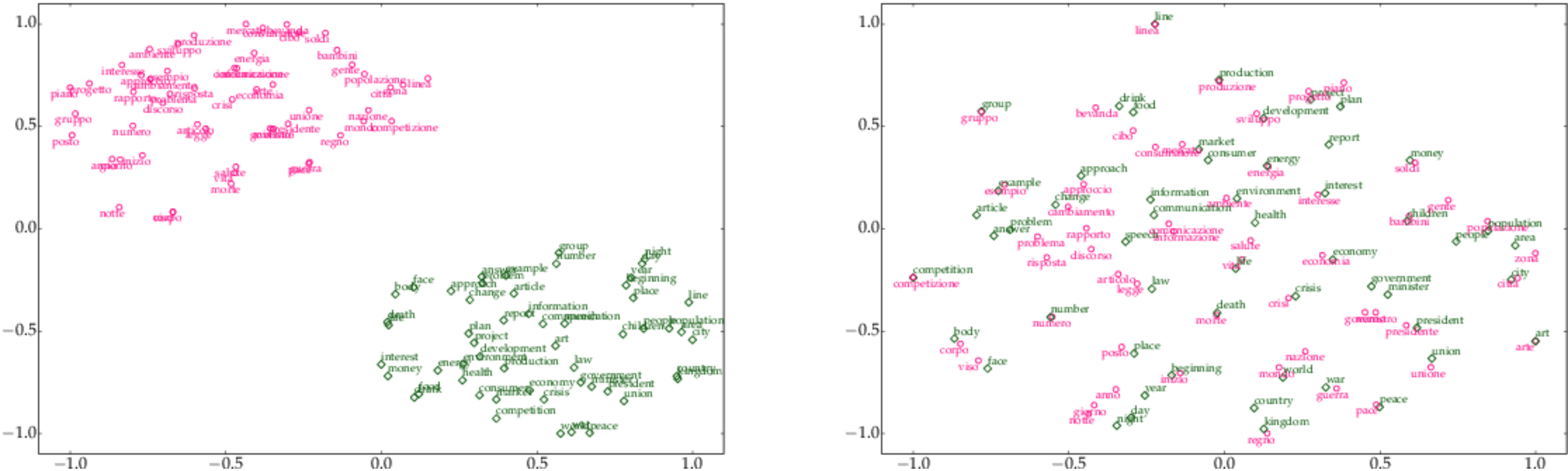
WORD EMBEDDINGS 1/2



Intuition

1. Pre-train language dependent word embeddings
2. Align the embeddings so all languages are in the same space

WORD EMBEDDINGS 1/2



[Ruder et al., 2019]

Does it make sense?

Can we really align embeddings of a morphologically rich language with the ones of a morphologically poor language?

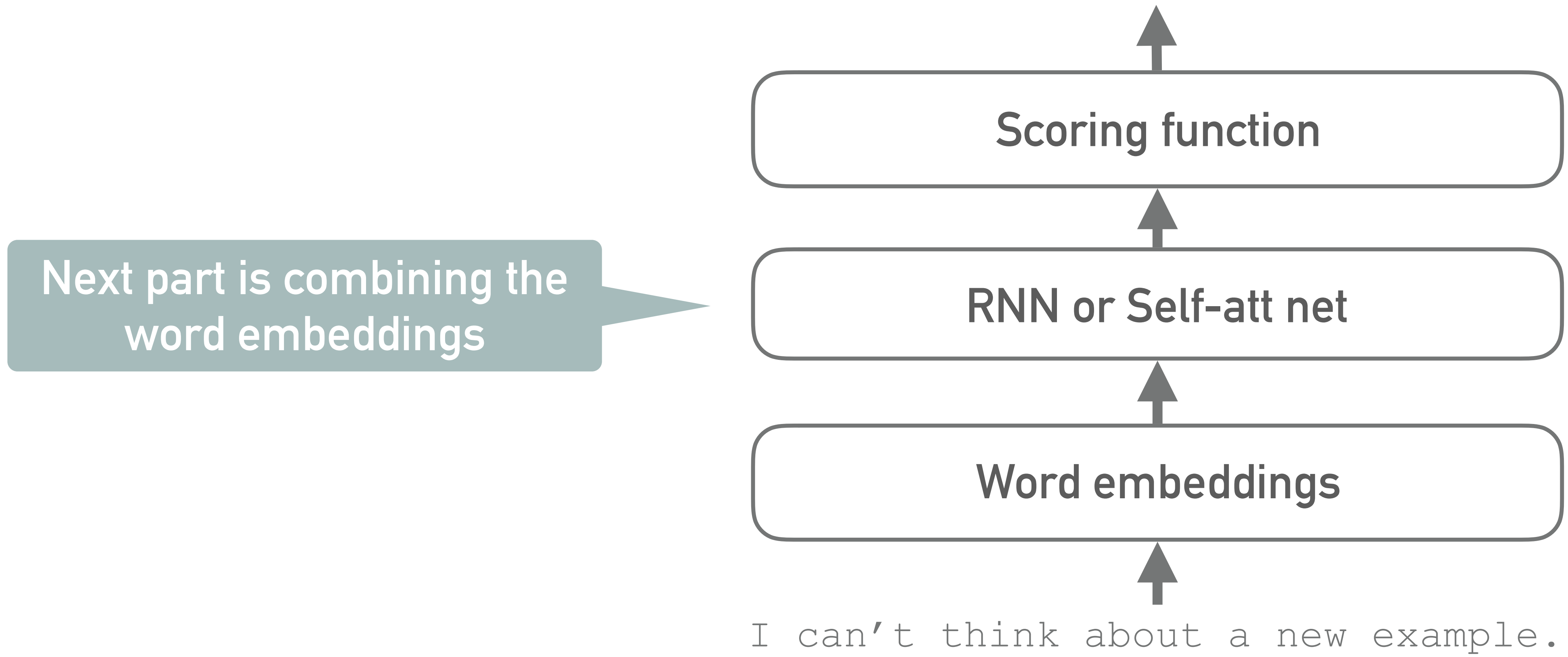
2 words

12 words

(1.6) West Greenlandic
Paasi-nngil-luinnar-para *ilaa-juma-sutit.*
understand-not-completely-1SG.SBJ.3SG.OBJ.IND come-want-2SG.PTCP
'I didn't understand at all that you wanted to come along.'
(Fortescue 1984: 36)

What should be aligned with what?!?!

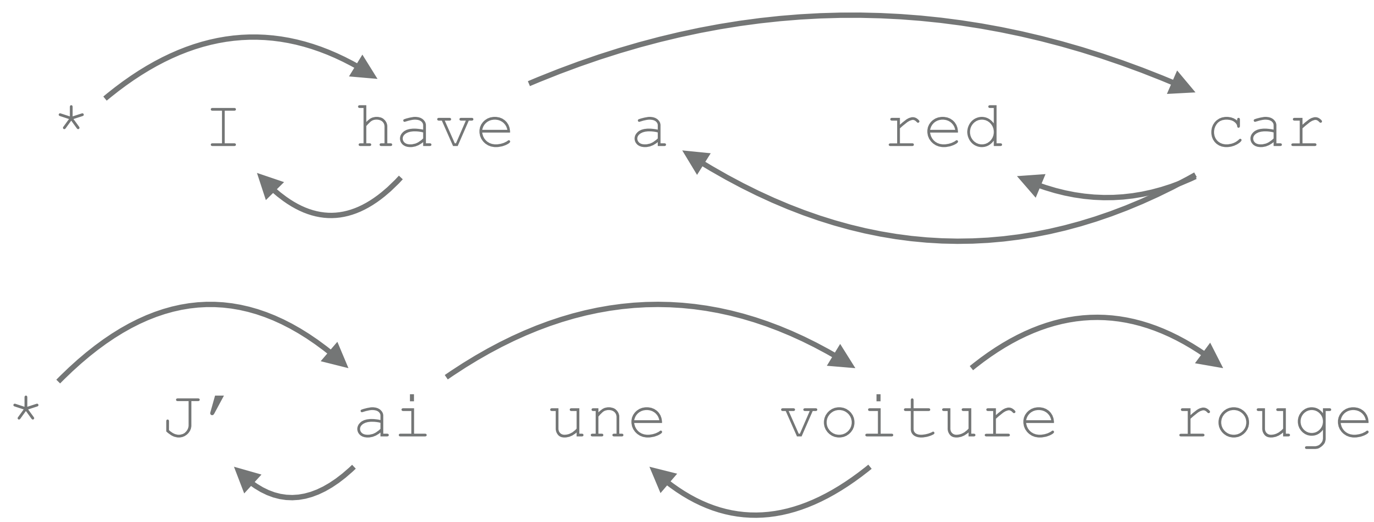
CONTEXTUAL REPRESENTATIONS



IDEA 1: COMPUTE REPRESENTATIONS OVER DEPENDENCIES

Intuition

- Let's assume aligned word embeddings make sense
- Instead of considering the sequential word order in the neural architecture, consider the universal dependency relations!



Different word order, same dependency graph

Problem

In a real scenario we don't have the dependency structure!

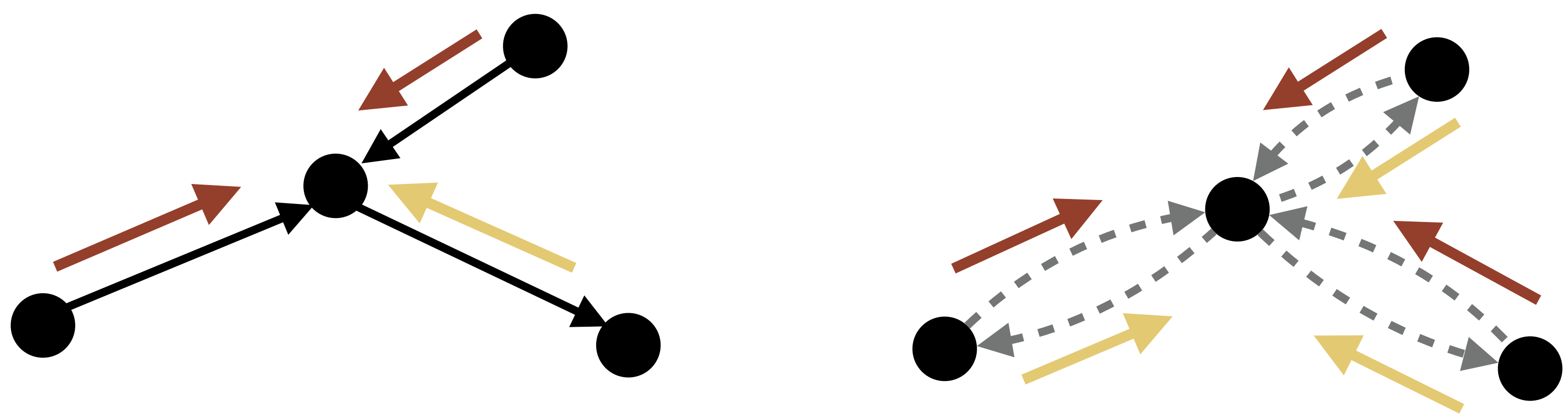
- We need to predict them
- To predict them we need to construct context-sensitive representations

=> so we go back to the same problem…

GRAPH NEURAL NETWORKS (GCN)

Intuition

- Vector representations of nodes (embeddings)
- (Parameterized) message passing on arcs



$$e^{n+1} = \tanh\left(\quad f(e^n) \quad + \quad g(e^n)T \quad + \quad h(e^n)T^\top \quad \right)$$

Self-loop

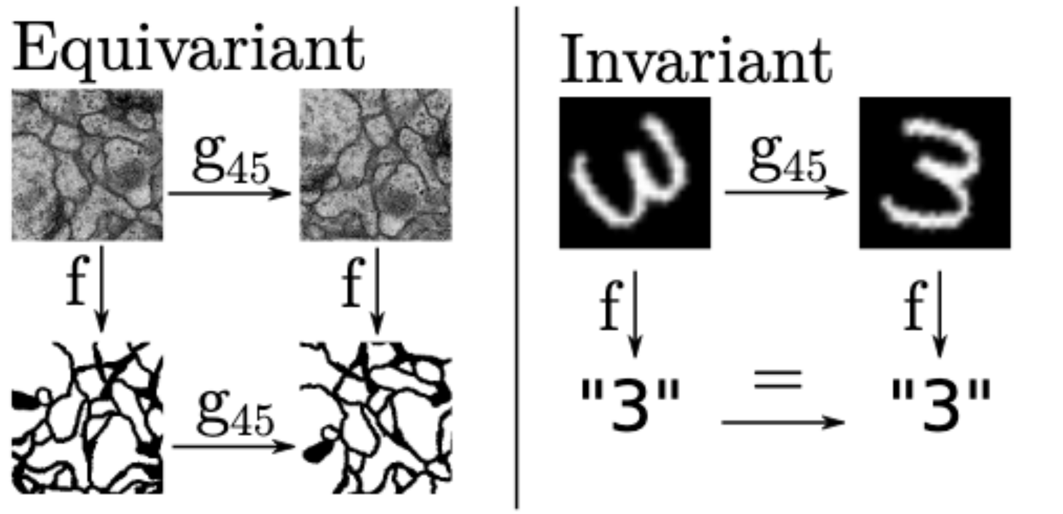
Messages on outgoing arcs

Messages on incoming arcs

[Kipf et Welling, 2016; Marcheggiani et Titov 2017]

IDEA 2: PERMUTATION EQUIVARIANT ARCHITECTURES?

Rotation equivariance vs. rotation invariance in computer vision



Exemple in NLP: bag-of-words model are invariant to word order permutation

A Neural Network is X equivariant if

- If the input is transformed by **X**
- Then the output is similarly transformed by **X**

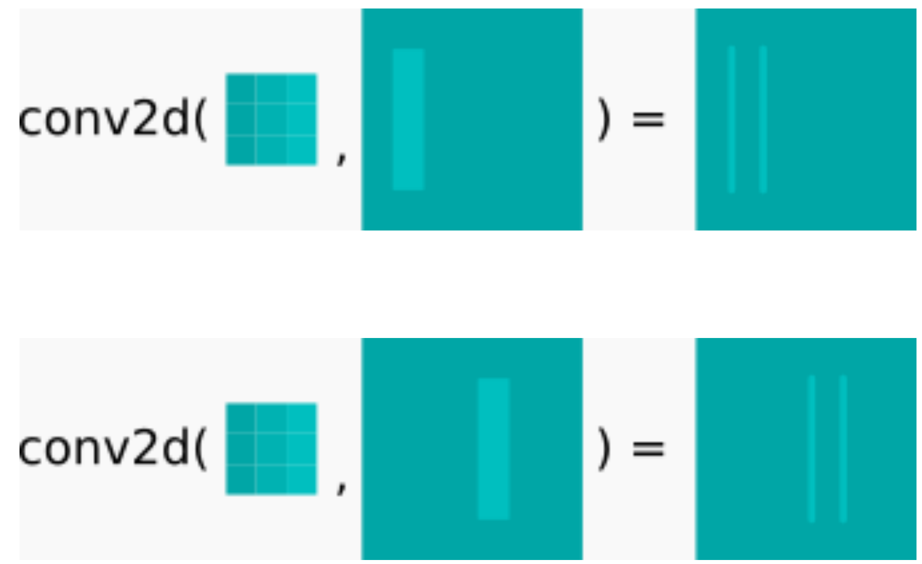
EQUIVARIANT CONVOLUTIONS IN COMPUTER VISION

Translation equivariant convolution

Preserves the « translation structure »

- If the input is transposed
- The output is also transposed

+ pooling will make the model invariant



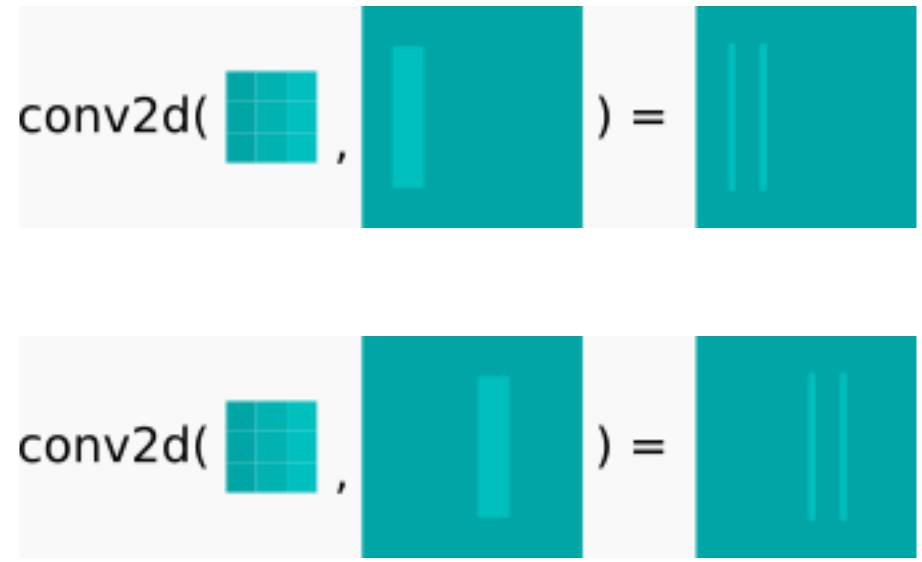
EQUIVARIANT CONVOLUTIONS IN COMPUTER VISION

Translation equivariant convolution

Preserves the « translation structure »

- ➤ If the input is transposed
- ➤ The output is also transposed

+ pooling will make the model invariant

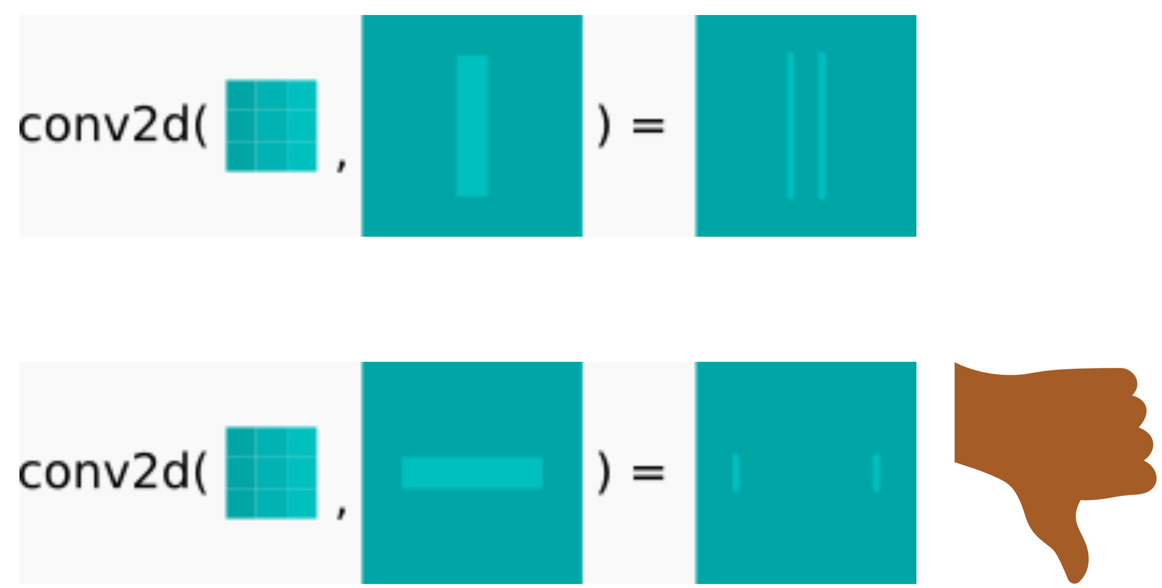


Rotation equivariant convolution

Preserves the « rotation structure »

- ➤ If the input is rotated
- ➤ The output is also rotated

Standard convolution **<u>is not</u>** rotation equivariant



GROUP CONVOLUTIONS

[Cohen and Weiling, 2016]

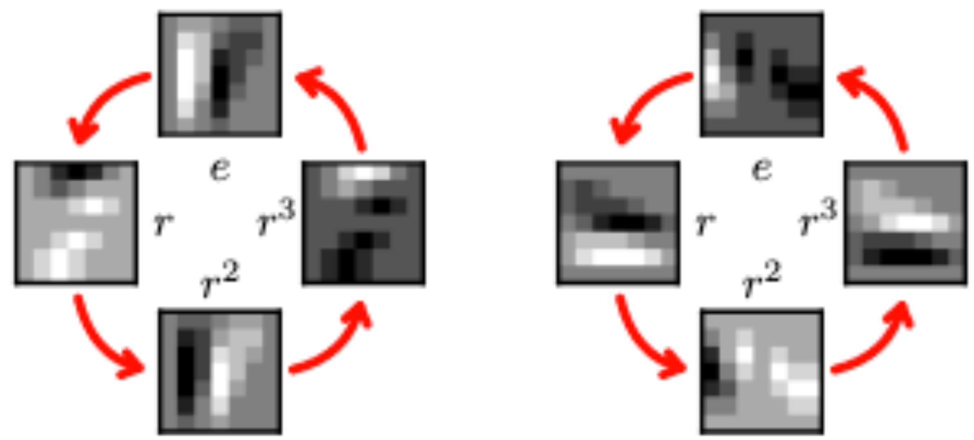


Figure 1. A p4 feature map and its rotation by r.

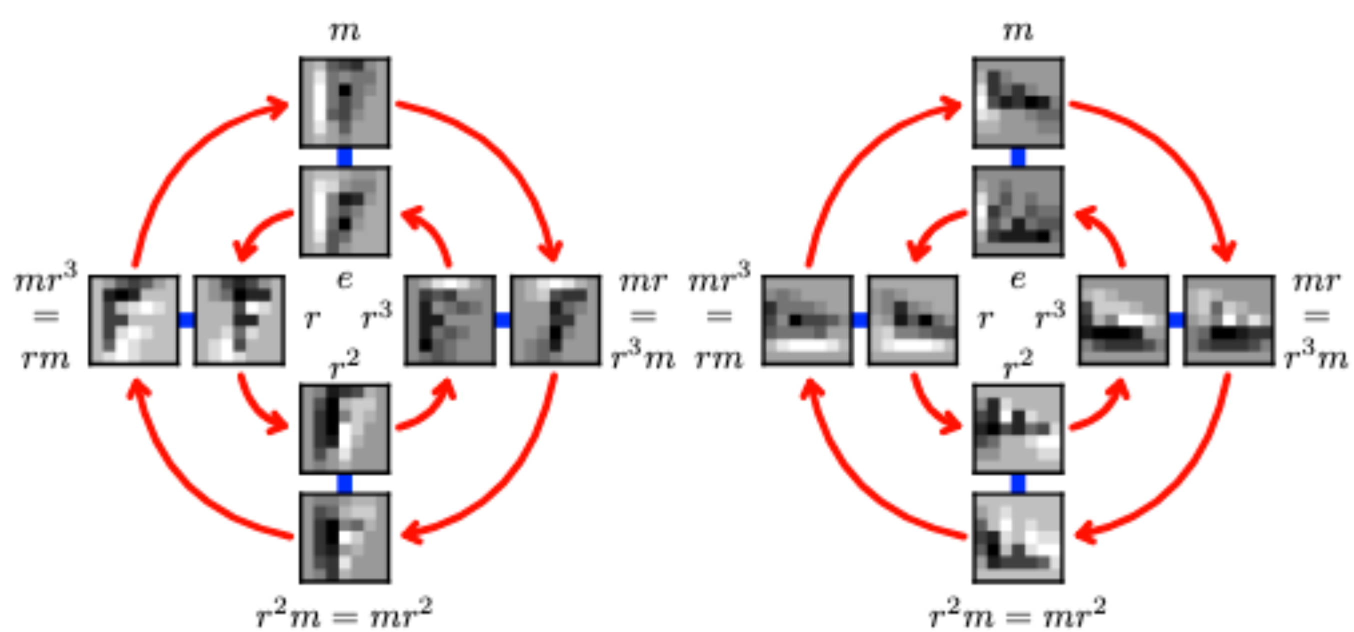
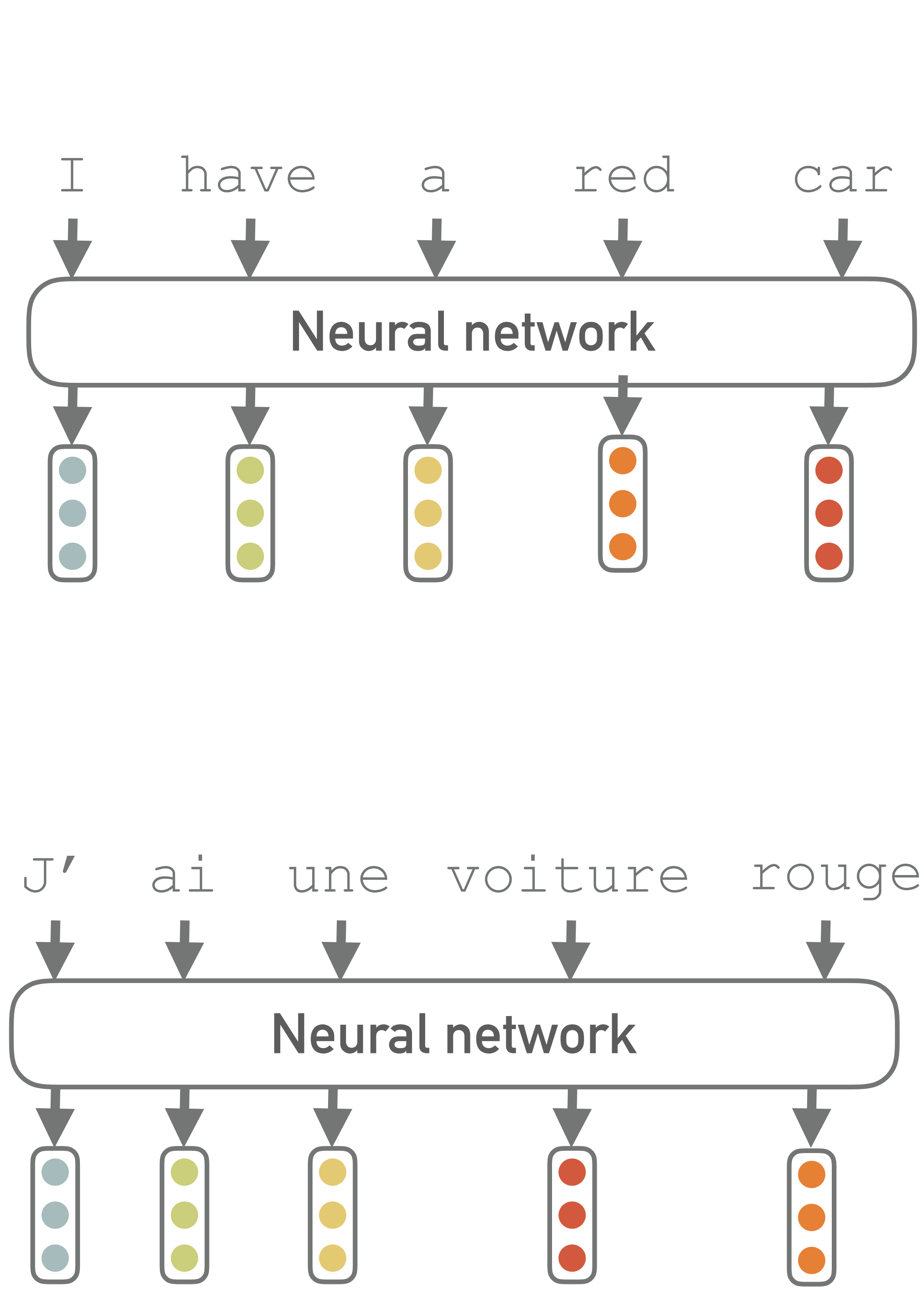


Figure 2. A p4m feature map and its rotation by r.

IDEA 2: PERMUTATION EQUIVARIANT ARCHITECTURES?



Amazing!

If the network is permutation equivariant, then we could score the dependencies and predict the same dependency structure for these three inputs! :)

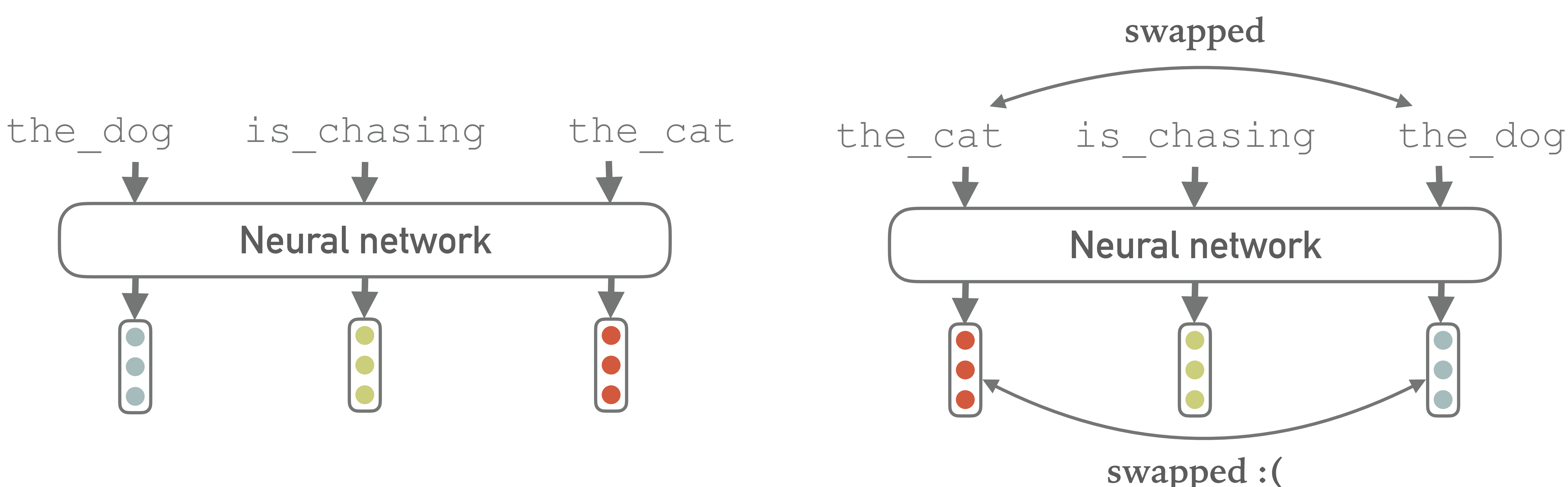
IDEA 2: PERMUTATION EQUIVARIANT ARCHITECTURES?

This is bad idea in general! :(

- We want to be permutation equivariant only for specific syntactic relations
- So we need to know the dependency structure to do that…

Example

By swapping « the_dog » and the « the_cat », we changed their grammatical role (subject vs. object), so **we also want to change their output representations!** :(



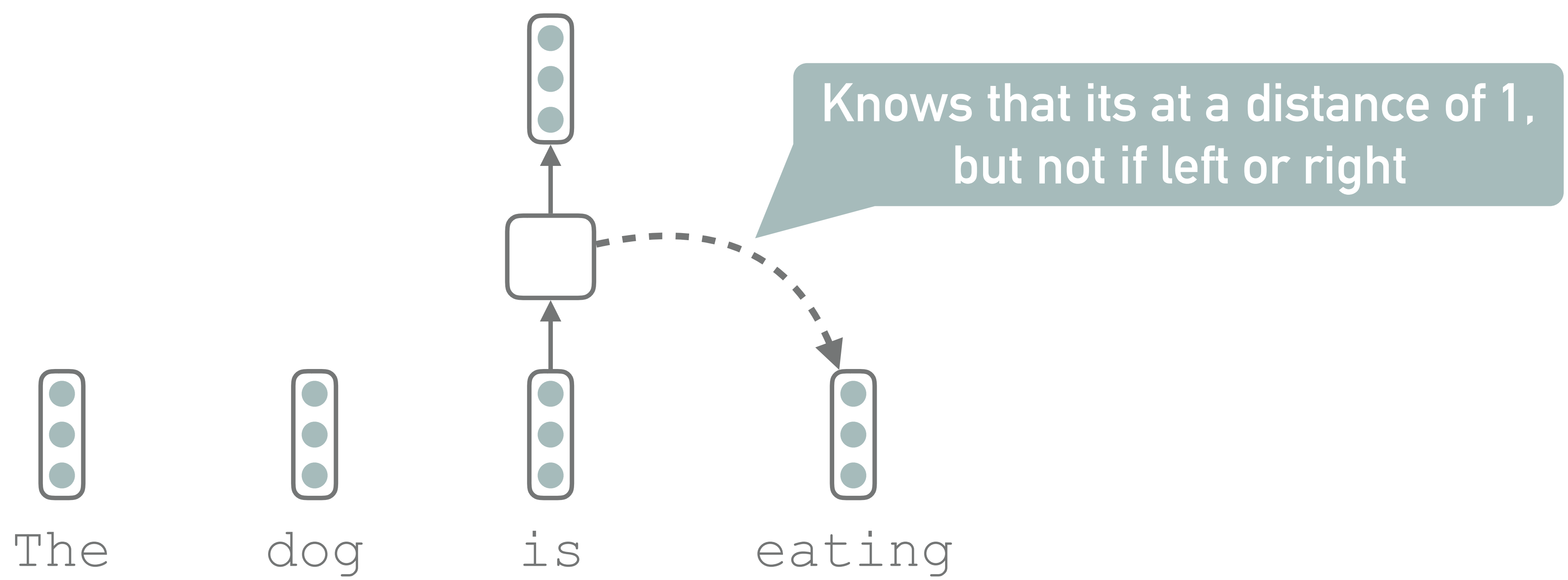
EXPERIMENTAL WORK IN THIS DIRECTION 1/2 [ahmad et al., 2019]

Transformer with relative position representations [Shaw et al., 2018]

- No distance embeddings

If only this => bag-of-word model

- Add direction agnostic distance information in heads



EXPERIMENTAL WORK IN THIS DIRECTION 2/2 [ahmad et al., 2019]

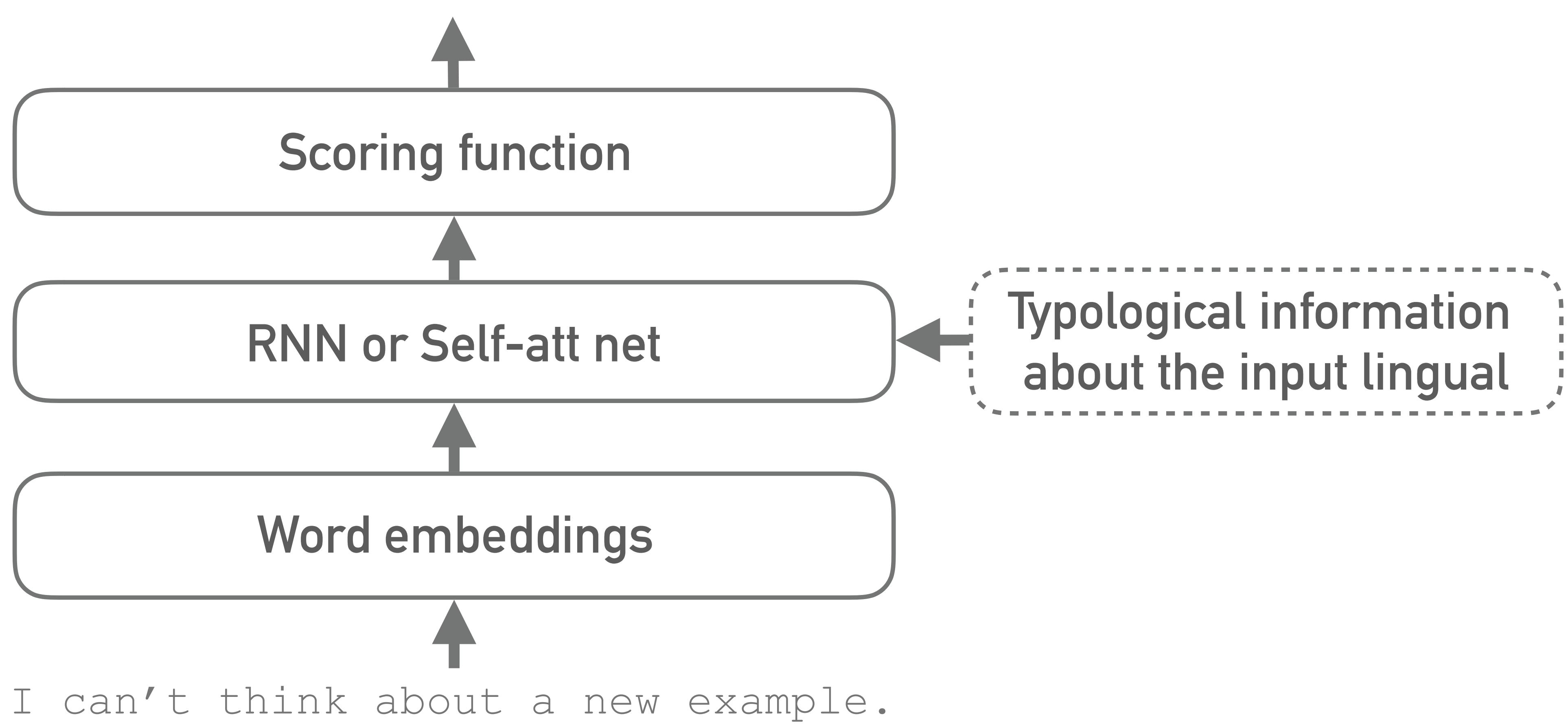
Lang	Dist. to English	SelfAtt-Graph (OF-OF)	RNN-Graph (OS-OF)
en	0.00	90.35/88.40	90.44/88.31
no	0.06	80.80/72.81	80.67/72.83
sv	0.07	80.98/73.17	81.23/73.49
fr	0.09	77.87/72.78	**78.35**†/**73.46**†
pt	0.09	**76.61**†/67.75	76.46/**67.98**
da	0.10	76.64/67.87	77.36/68.81
es	0.12	74.49/66.44	**74.92**†/**66.91**†
it	0.12	80.80/75.82	**81.10/76.23**†
hr	0.13	**61.91**†/**52.86**†	60.09/50.67
ca	0.13	73.83/65.13	**74.24**†/**65.57**†
pl	0.13	**74.56**†/**62.23**†	71.89/58.59
uk	0.13	**60.05/52.28**†	58.49/51.14
sl	0.13	**68.21**†/**56.54**†	66.27/54.57
nl	0.14	68.55/60.26	67.88/60.11
bg	0.14	**79.40**†/**68.21**†	78.05/66.68
ru	0.14	60.63/51.63	59.99/50.81
de	0.14	**71.34**†/**61.62**†	69.49/59.31
he	0.14	**55.29/48.00**†	54.55/46.93
cs	0.14	**63.10**†/**53.80**†	61.88/52.80
ro	0.15	**65.05**†/**54.10**†	63.23/52.11
sk	0.17	**66.65/58.15**†	65.41/56.98

IDEA 3: TYPOLOGICAL INFORMATION AS FEATURES 1/2

Intuition

Instead of build a « smart architecture », we give to the network features representing the typological information of the input language

- Train on many languages and hope it learns to be « equivariant in the good way »
- At test time, we could give input in an unknown language as long as it exists in WALS :)



IDEA 3: TYPOLOGICAL INFORMATION AS FEATURES 2/2

[Scholivet et al., 2019]

- Multilingual settings
- No cross-lingual experiments :(

L: Monolingual corpus.

Σ: Multilingual corpus.

Σ ID: Multilingual corpus + language ID.

ΣW_N, ΣW_{80}: Multilingual corpus + WALS.

L	Σ	Σ ID	ΣW_N	ΣW_{80}	Lang.
65.89	60.59	62.97	63.15	64.38	ar
78.59	74.32	76.43	76.26	77.47	bg S
77.18	72.76	74.11	73.03	76.27	ca R
68.92	68.01	68.91	68.72	69.61	cs S
73.62	67.38	70.25	70.19	70.25	da G
71.07	63.76	66.47	69.18	69.22	de G
77.11	71.26	72.72	73.29	75.84	el
70.05	66.02	69.3	69.91	70.19	en G
71.47	71.98	72.38	72.29	73.22	es R
66.98	63.76	66.89	65.75	67.79	et
63.26	55.76	59.54	60.22	59.39	eu
72.85	66.02	67.23	69.63	70	fa
60.97	56.29	58.86	57.37	59.28	fi
75.74	74.25	75.44	74.79	75.82	fr R
66.55	60.41	63.12	64.68	65.96	ga
70.21	63.03	65.69	66.09	67.45	he
78.91	73.86	75.81	75.77	74.45	hi
71.03	67.49	68.39	70	70.4	hr S
67.08	62.55	66.89	67.19	67.51	hu
68.64	58.38	63.99	62.61	64.57	id
81.44	76.45	78.03	76.97	79.83	it R

CONCLUSION

CONCLUSION

In brief

We have done **almost nothing** yet

- Aligned word embeddings don't even make sense
- Best we can do is pretrain BERT on big multilingual datasets

Exciting research coming in the future? Maybe by you? :-)

Most of the work is English-centric :/

> The annotators were then tasked with producing language-specific annotation guidelines with the expressed goal of keeping the label and construction set as close as possible to the original English set, only adding labels for phenomena that do not exist in English. Making fine-grained label dis-

[McDonald et al., 2013]

I didn't cover many topics!

- Transition-based parsers
- etc etc.