



Applied Natural Language Processing

Info 256

Lecture 22: Syntax (Nov 8, 2023)

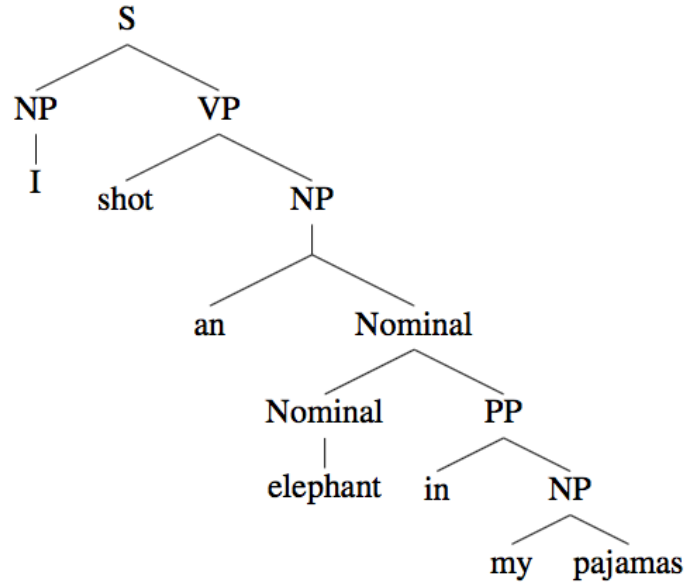
David Bamman, UC Berkeley

Syntax

- With syntax, we're moving from labels for discrete items — documents (sentiment analysis), tokens (POS tagging, NER) — to the **structure** between items.

PRP VBD DT NN IN PRP\$ NNS

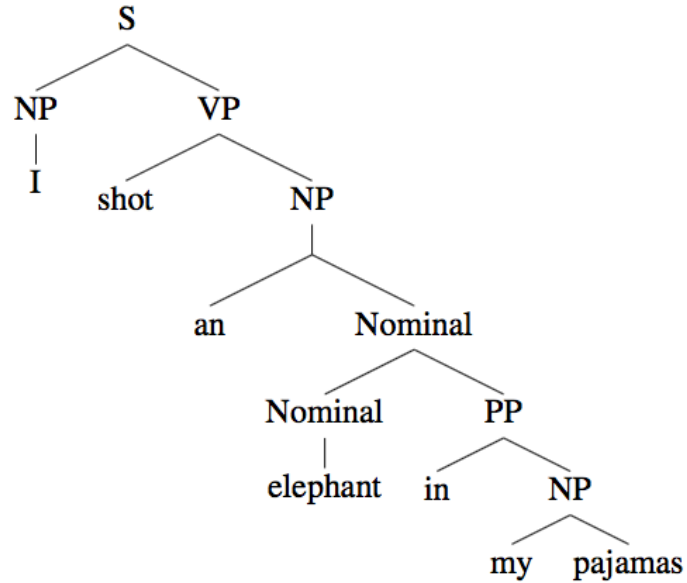
I shot an elephant in my pajamas



PRP VBD DT NN IN PRP\$ NNS

I shot an elephant in my pajamas

Why is syntax important?

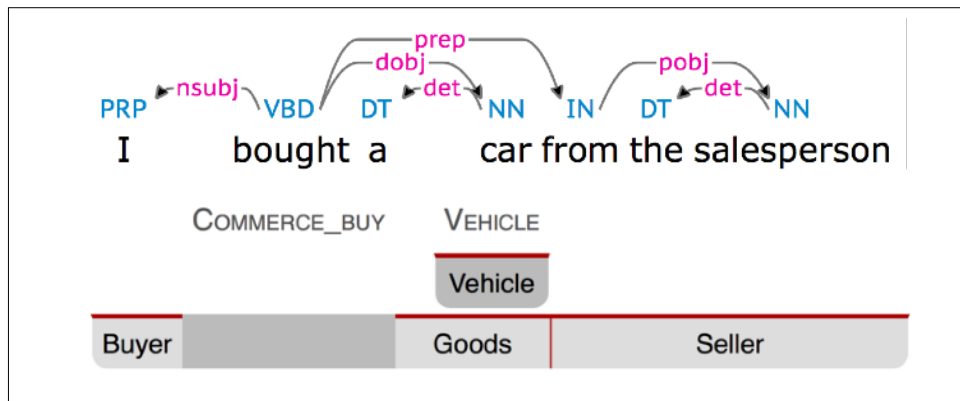


Why is POS important?

- POS tags are indicative of syntax
- POS = cheap multiword expressions [(JJ|NN)+ NN]
- POS tags are indicative of pronunciation (“I contest the ticket” vs “I won the contest”)

Why is syntax important?

- Foundation for **semantic analysis** (on many levels of representation: semantic roles, compositional semantics, frame semantics)



Why is syntax important?

- Strong representation for **discourse analysis** (e.g., coreference resolution)

Bill **VBD** Jon; he was having a good day.

- Many factors contribute to pronominal coreference (including the specific verb above), but syntactic subjects > objects > objects of prepositions are more likely to be antecedents

Why is syntax important?

Linguistic typology; relative positions of subjects (S), objects (O) and verbs (V)

SVO	English, Mandarin	I grabbed the chair
SOV	Latin, Japanese	I the chair grabbed
VSO	Arabic, Hawaiian	Grabbed I the chair
OSV	Yoda	Patience you must have
...

Sentiment analysis



"Unfortunately I already had this exact picture tattooed on my chest, but **this shirt** is very useful in colder weather."

[overlook1977]

Question answering

What did Barack Obama teach?

Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of Chicago Law School between 1992 and 2004.



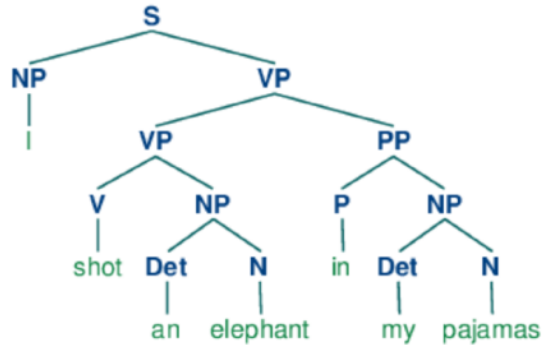
Syntax

- Syntax is fundamentally about the hierarchical structure of language and (in some theories) which sentences are **grammatical** in a language

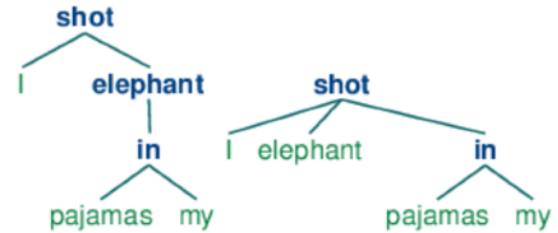
words → phrases → clauses → sentences

Formalisms

Phrase structure grammar
(Chomsky 1957)



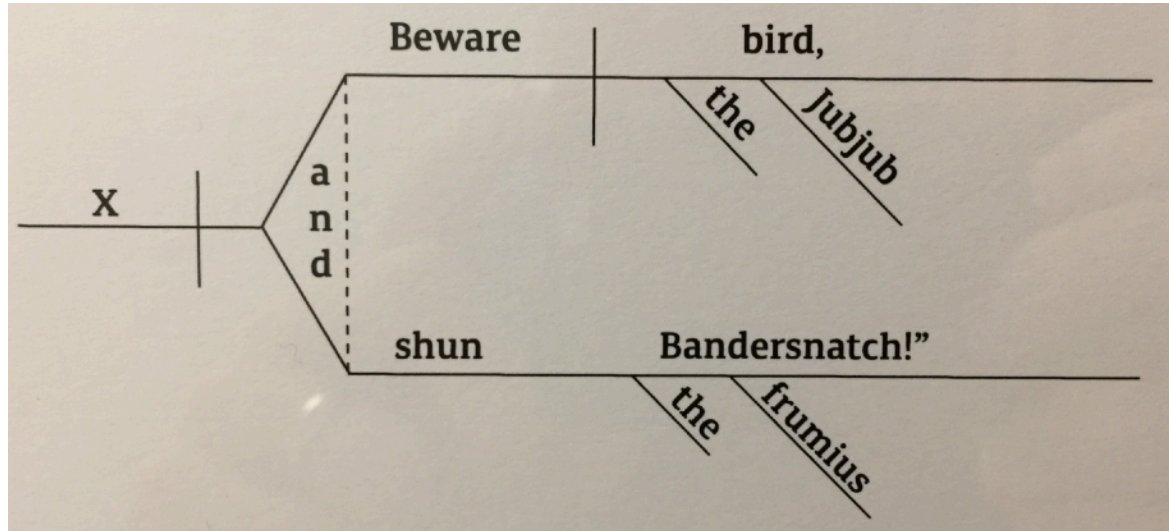
Dependency grammar
(Mel'čuk 1988; Tesnière 1959; Pāṇini)



Dependency syntax

- Sgall, Dependency-based formal description of language (1994)
- Mel'čuk, Dependency Syntax: Theory and Practice (1988)
- Tesnière, *Éléments de syntaxe structurale* (1959)
- Medieval theories of grammar (Covington 1984)
- Pānini grammar of Sanskrit (ca. 5th-century BCE)

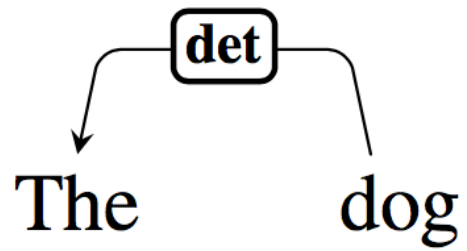
Dependency syntax

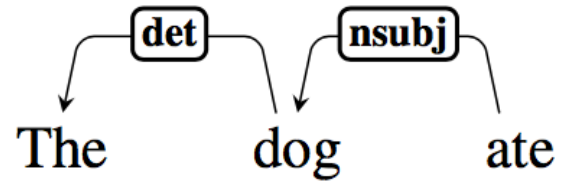


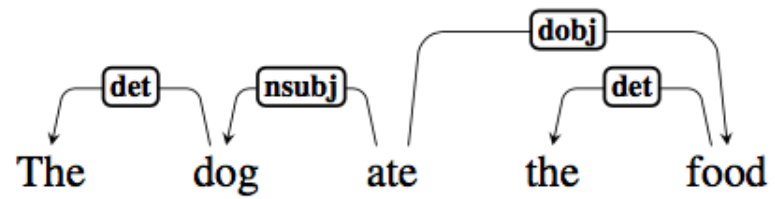
"Sentence diagramming"

Dependency syntax

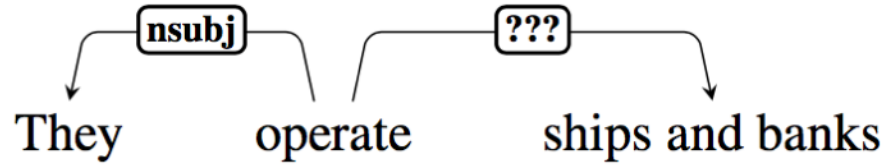
- Syntactic structure = **asymmetric**, **binary** relations between words.



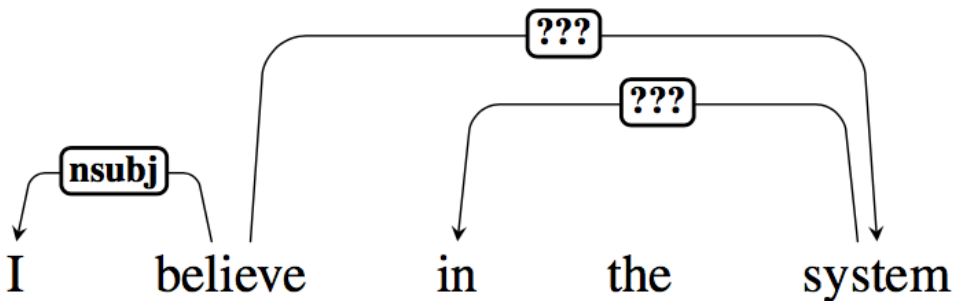
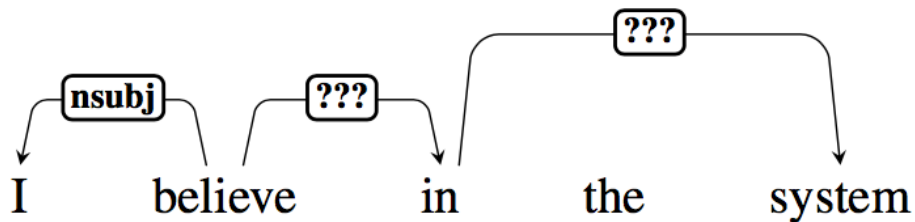




Coordination



Case marking prepositions



Trees

- A dependency structure is a directed graph $G = (V,A)$ consisting of a set of vertices V and arcs A between them. Typically constrained to form a **tree**:
 - Single root vertex with no incoming arcs
 - Every vertex has exactly one incoming arc except root (**single head constraint**)
 - There is a unique path from the root to each vertex in V (**acyclic constraint**)

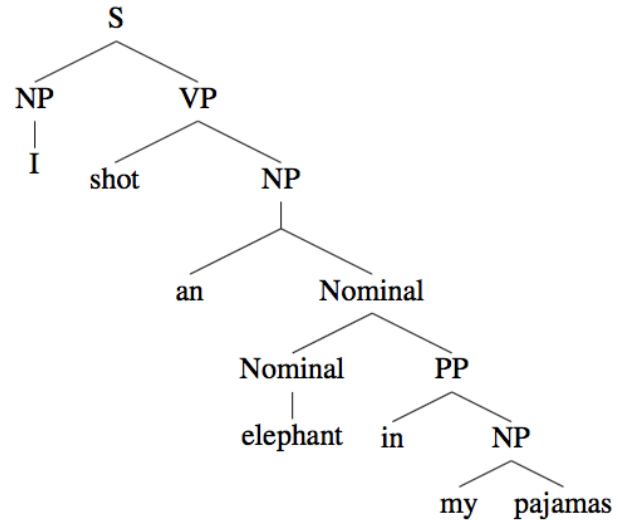
Trees

- Unlike phrase-structure trees, dependency trees aren't tied to the linear order of the words in a sentence.
- Adding a constraint derived from the linear order of words in a sentence allows for more efficient parsing algorithms.

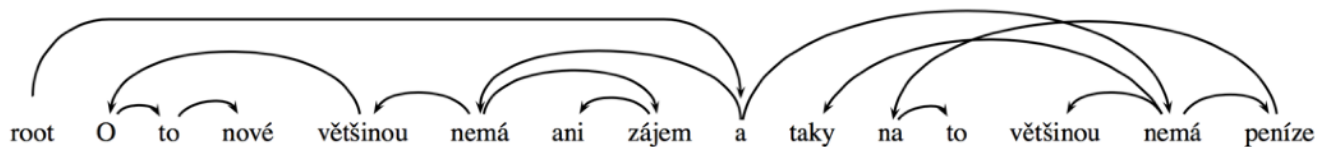
Word order

- Dependency relations belong to the structural order of a sentence, not the **linear order**.

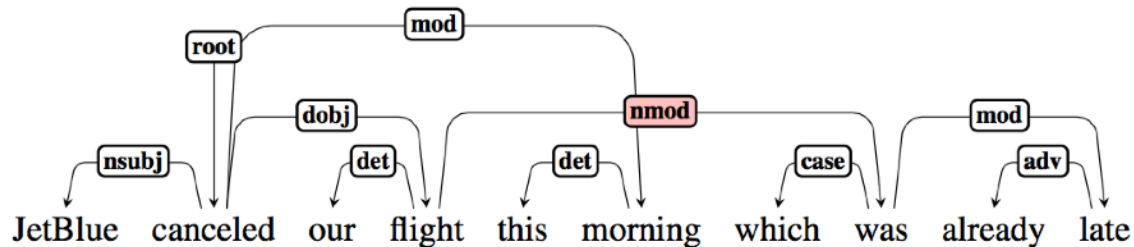
- This is different from a phrase-structure tree, where the syntax is constrained by the linear order of the sentence (a different linear order yields a different parse tree).



Free word order

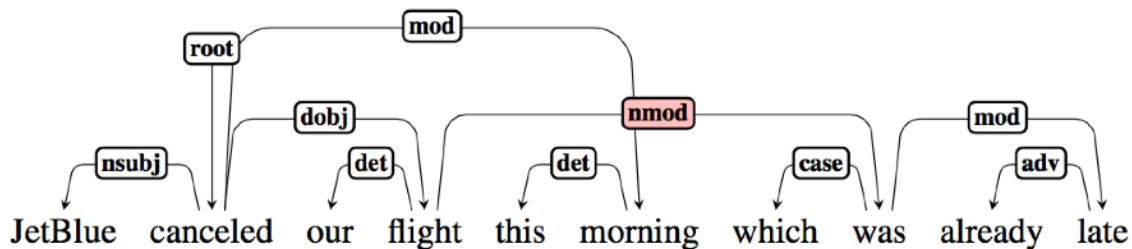


He is mostly not even interested in the new things and in most cases, he has no money for it either.



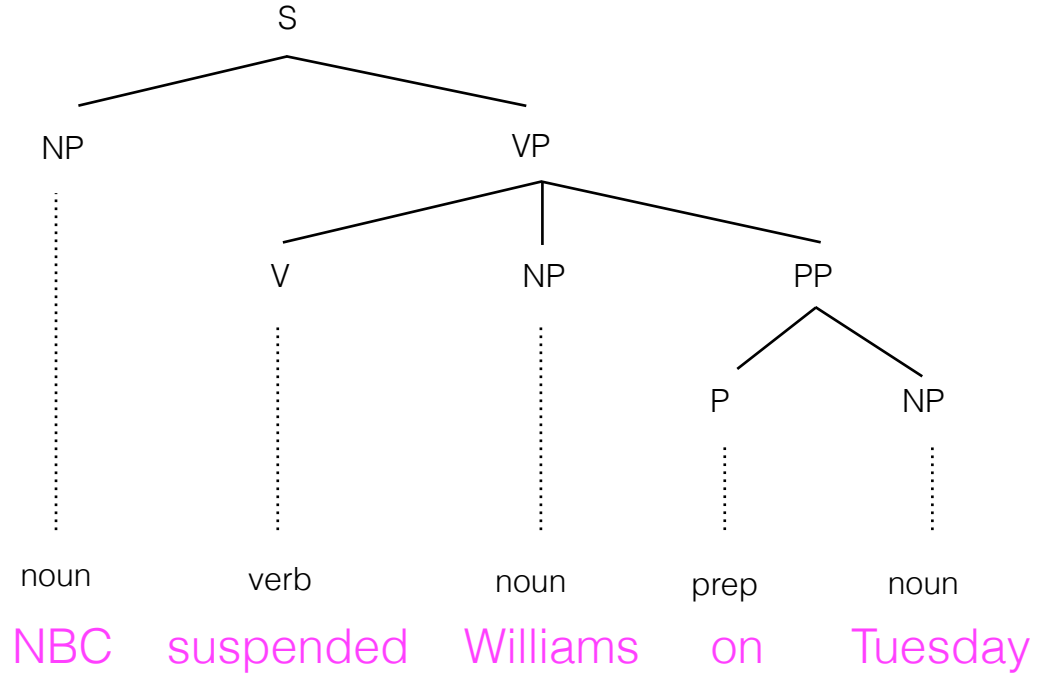
Projectivity

- An arc between a head and dependent is projective if there is a path from the head to every word between the head and dependent.

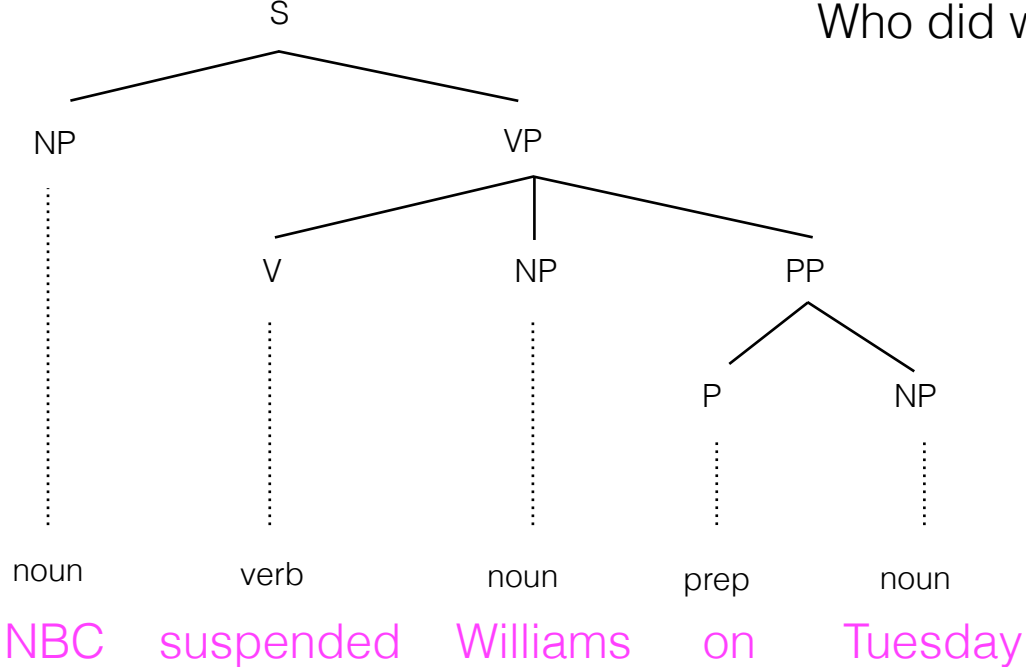


Dependencies vs constituents

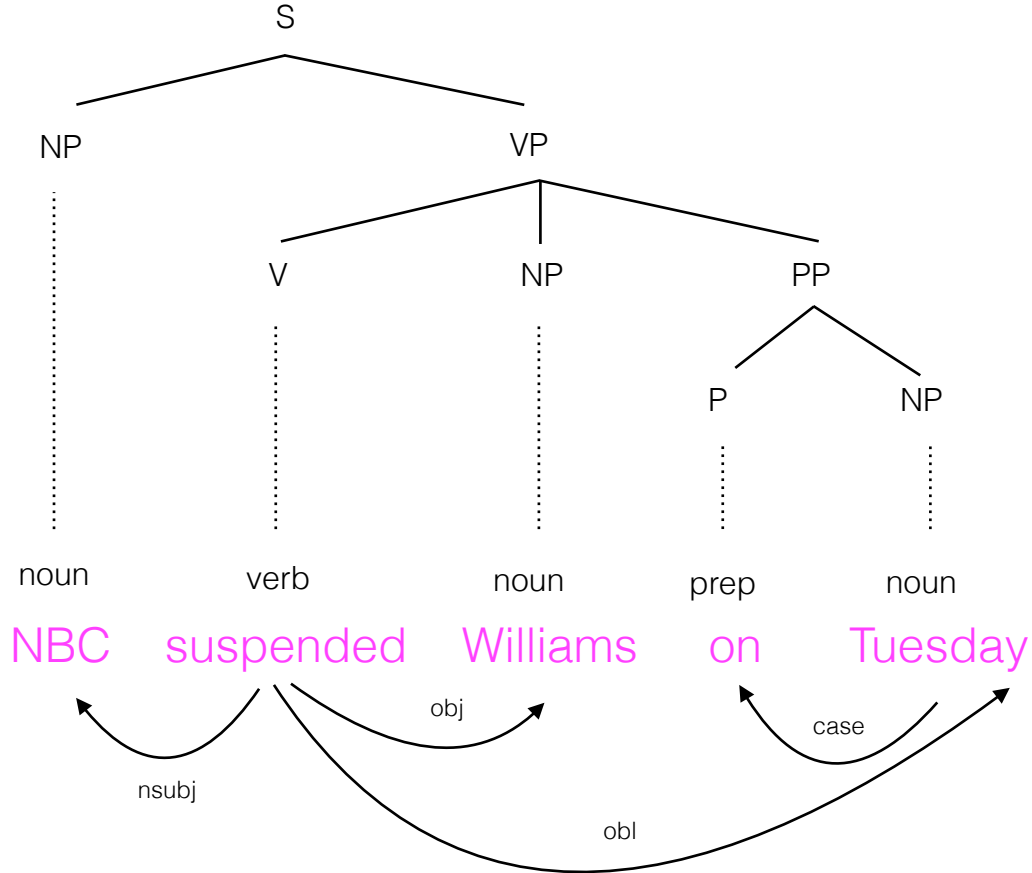
- Dependency links are closer to semantic relationships; no need to infer the relationships from the structure of a tree
- A dependency tree contains one edge for each word, no intermediate hidden structures that also need to be learned for parsing.
- Easier to represent languages with free word order.



Who did what to whom?



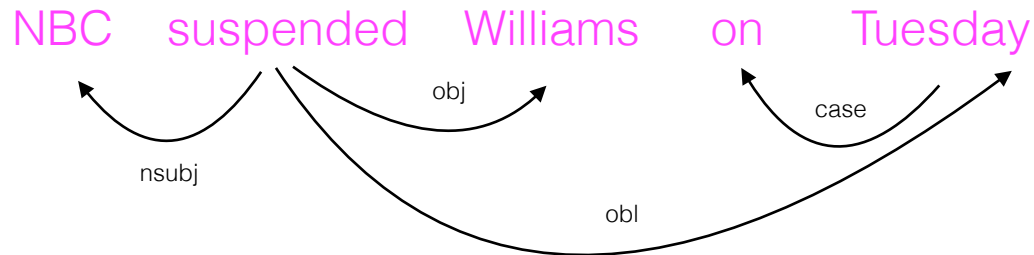
subject: S → NP VP
direct object: S → NP (VP → ... NP ...)



Dependency grammar

Captures binary relations between words

- nsubj(NBC, suspended)
- obj(Williams, suspended)



Data

- NELL SVO triples (604 million nsubj+dobj relations from 230B words on the web)

police	found	five .030 bullets	1
police	found	seven dead rebels	3
police	found	two hidden cameras	2
police	found	wanders lover	1
police	found	211 pounds	4
police	found	Marcia	3
police	found	bank draft	1
police	found	diskette	2
police	found	five marijuana plants	3
police	found	items used	1
police	found	judge	5

Dependency- Based Word Embeddings




























































Levy & Goldberg,
ACL 2014

<http://irsrv2.cs.biu.ac.il:9998/>

Target Word	BoW5	BoW2	DEPS
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	pauling hotelling heting lessing hamming
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina
object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl	event-driven domain-specific rule-based data-driven human-centered
dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning	singing rapping breakdancing miming busking

Universal Dependencies

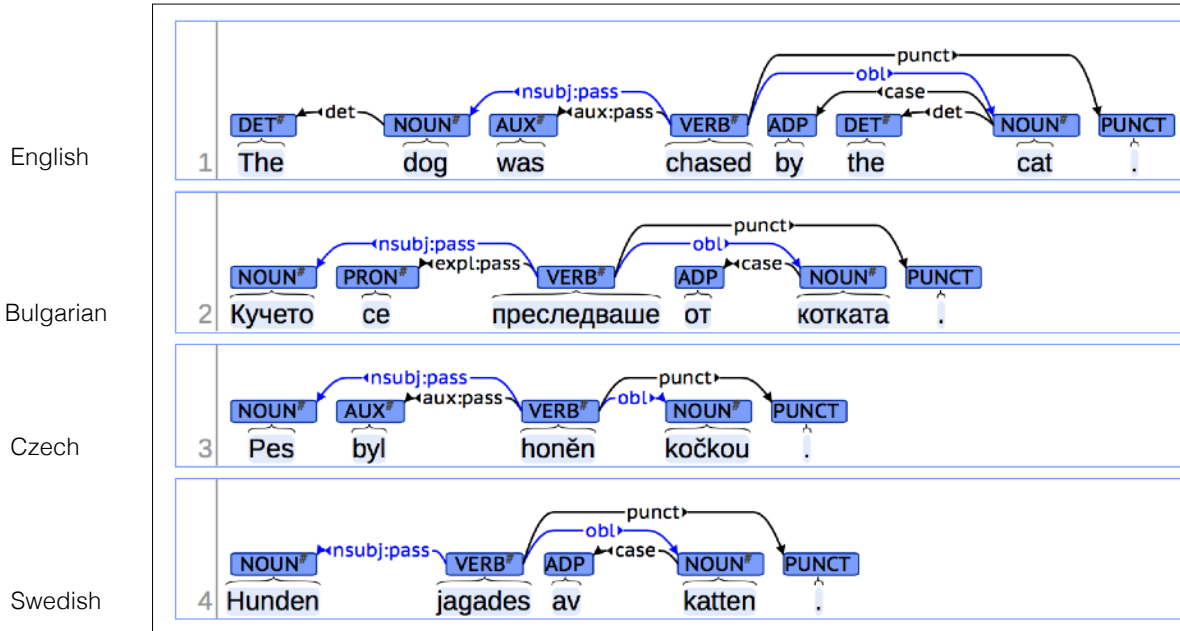
UD Treebanks

▶		Afrikaans	49K	LF	-	⚙️	⏸️		
▶		Ancient Greek	202K	LF		⚙️	✓		
▶		Ancient Greek-PROIEL	211K	LF	-	⚙️	✓		
▶		Arabic	242K	LF	-	⚙️	✓		
▶		Arabic-NYUAD	629K	LF	-	⚙️	✓		
▶		Arabic-PUD	20K	LF	-	👤	⏸️		
▶		Basque	121K	LF		⚙️	✓		
▶		Belarusian	8K	LF	-	👤	✓		
▶		Bulgarian	156K	LF		⚙️ ✓	✓		
▶		Buryat	10K	LF	-	👤	⏸️		
▶		Catalan	530K	LF		⚙️ ✓	✓		
▶		Chinese	123K	LF		⚙️ ✓	✓		
▶		Chinese-CFL	7K	L		👤	⏸️		
▶		Chinese-PUD	21K	F	-	👤	⏸️		
▶		Coptic	4K	LF		👤	✓		
▶		Croatian	197K	LF	-	⚙️ ✓	✓		
▶		Czech	1,503K	LF		⚙️ ✓	✓		

Universal Dependencies

- Developing cross-linguistically consistent treebank annotation for many languages
- Goals:
 - Facilitating multilingual parser development
 - Cross-lingual learning
 - Parsing research from a language typology perspective.

Universal Dependencies



Spacy

- Spacy uses the ClearNLP dependency labels (derived from phrase-structure trees) that are quite similar to the well-document [Stanford typed dependencies](#).

Stanford typed dependencies

nsubj nominal subject the **dog** ran

dobj direct object the dog chased the **cat**

amod adjectival modifier the **big** dog ran

det determiner **the** big dog ran

prep preposition the dog ran **into** the house

pobj object of preposition the dog ran into the **house**

I saw the man with the telescope

nsubj nominal subject the **dog** ran

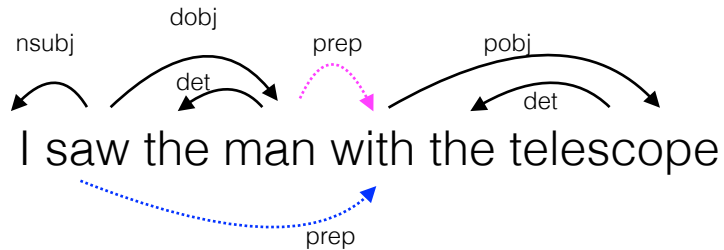
dobj direct object the dog chased the **cat**

amod adjectival modifier the **big** dog ran

det determiner **the** big dog ran

prep preposition the dog ran **into** the house

pobj object of preposition the dog ran into the **house**



I	saw	nsubj	I	saw	nsubj
the	man	det	the	man	det
man	saw	dobj	man	saw	dobj
the	telescope	det	the	telescope	det
telescope	with	pobj	telescope	with	pobj
with	saw	prep	with	man	prep

ikr smh he asked fir yo last name so he can
add u on fb lololol

Training a parser for a new language

1. Annotate texts for syntax (phrase structure or dependency)
2. Train a parser

English	Penn Treebank	Phrase s
English	CCGbank	Combina categori gramma
English	Prague English Dependency Treebank	Depende
English	BLLIP WSJ corpus	Phrase s
English	British Component of the International Corpus of English (ICE-GB)	Phrase s
English	Diachronic Corpus of Present-Day Spoken English (DCPSE)	Phrase s
English	Lancaster Parsed Corpus	Phrase s
English	Susanne Corpus	Phrase s
English	Christine Corpus	Phrase s
English	Lucy Corpus	Phrase s
English	Tübingen Treebank of English / Spontaneous Speech (TüBa-E/S)	HPSG
English	LinGO Redwoods	HPSG
English	Multi-Treebank	Phrase s
English	The PARC 700 Dependency Bank	Depende