

Språk och Datorer (729G49)

Språkteknologi 3:

Lingvistiskt uppmärkt text 1: Dependensanalys

14/4/2025



Innehåll

- Repetition: Lingvistiskt uppmärkt text
- Fraser och frasstrukturträd
- Dependensträd
- Utmaning: Syntaktisk Tvetydighet



Motivering

- Vi började i Labb 1 och 2 med att titta på **ord**.
- Ord kombineras till **fraser**,
 - som i sin tur kan kombineras till **större, nästlade fraser**.
- —> Språk har en hierarkisk struktur.
- Att bestämma den **syntaktiska strukturen** i en text är grundläggande för i princip all analys av naturligt språk.

Applikationer

- Lingvistik: Undersöka skillnader mellan språk, språkförändring, hitta vissa syntaktiska konstruktioner för att veta vilka är vanligare, etc.
- Informationsutvinning: Vem gör vad? Vilka relationer har olika koncept till varandra? Vem är det som jobbar, **betalar**, etc. i korpusen?
- Andraspråksinlärning: Analysera typiska fel, korrigera fel subjekt-verb-kongruens, generera grammatikövningar.
- Tidigare: Maskinöversättning, textgenerering.

Lingvistiskt uppmärkt text

1	Genom	genom	PP	3	AA
2	skattereformen	skattereform	NN	1	PA
3	införs	införa	VB	0	ROOT
4	individuell	individuell	JJ	5	AT
5	beskattning	beskattning	NN	3	SS
6	(	(	PAD	5	IR
7	särbeskattning	särbeskattning	NN	5	AN
8	)	)	PAD	5	JR
9	av	av	PP	5	ET
10	arbetsinkomster	arbetsinkomst	NN	9	PA
11	.	.	MAD	3	IP

Lingvistiska Annotationer

- Graford *skattereformen*
- Lemma *skattereform*
- Ordklass NN (substantiv)
- Huvudord Ord nummer 3 i meningen (*införs*)
- Grammatisk funktion SS (subjekt)
- Morfologiska egenskaper utrum, singular, definit, nominativ
- Betydelse skattereform..nn.1

1	Genom	genom	PP	3	AA
2	skattereformen	skattereform	NN	1	PA
3	införs	införa	VB	0	ROOT
4	individuell	individuell	JJ	5	AT
5	beskattning	beskattning	NN	3	SS
6	(	(	PAD	5	IR
7	särbeskattning	särbeskattning	NN	5	AN
8	)	)	PAD	5	JR
9	av	av	PP	5	ET
10	arbetsinkomster	arbetsinkomst	NN	9	PA
11	.	.	MAD	3	IP

Lingvistiska Annotationer

- **Graford** *skattereformen*
- ~~Lemma~~ **Labb 2** ~~skattereform~~
- **Ordklass** **NN (substantiv)**

- Huvudord Ord nummer 3 i meningen (*införs*)
- Grammatisk funktion SS (subjekt)
- Morfologiska egenskaper utrum, singular, definit, nominativ
- ~~Betydelse skattereform..nn.1~~

1	Genom	genom	PP	3	AA
2	skattereformen	skattereform	NN	1	PA
3	införs	införa	VB	0	ROOT
4	individuell	individuell	JJ	5	AT
5	beskattning	beskattning	NN	3	SS
6	(	(	PAD	5	IR
7	särbeskattning	särbeskattning	NN	5	AN
8	)	)	PAD	5	JR
9	av	av	PP	5	ET
10	arbetsinkomster	arbetsinkomst	NN	9	PA
11	.	.	MAD	3	IP

Lingvistiska Annotationer

1	Genom	genom	PP	3	AA
2	skattereformen	skattereform	NN	1	PA
3	införs	införa	VB	0	ROOT
4	individuell	individuell	JJ	5	AT
5	beskattning	beskattning	NN	3	SS
6	(	(	PAD	5	IR
7	särbeskattning	särbeskattning	NN	5	AN
8	)	)	PAD	5	JR
9	av	av	PP	5	ET
10	arbetsinkomster	arbetsinkomst	NN	9	PA
11	.	.	MAD	3	IP

- **Graford** *skattereformen*
- ~~Lemma~~ **Labb 2** ~~skattereform~~
- **Ordklass** **NN (substantiv)**

- Huvudord Ord nummer 3 i meningen (*införs*)
- Grammatisk funktion **SS (subjekt)**

Labb 3

- Morfologiska egenskaper utrum, singular, definit, nominativ
- ~~Betydelse~~ ~~skattereform..nn.1~~

Fraser och Frasstrukturträd

Fraser

- **Syntax** studerar de regler och villkor som kontrollerar hur ord kan forma större enheter, såsom fraser och satser, inom en mening.
- **Fraser** är sekvenser av sammanhängande ord. Det viktigaste ordet i en fras kallas **huvudord** och ger frasen dess namn.
- [Den galne mannen med den vilda blicken] [bet] [den stackars hunden].

Huvudord



- [Den galne mannen med den vilda blicken] [bet] [den stackars hunden].

Huvudord



- [Den galne **mannen** med den vilda blicken] **[bet]** [den stackars **hunden**].
- **Mannen:**
- **bet:**
- **hunden:**

Huvudord



- [Den galne **mannen** med den vilda blicken] **[bet]** [den stackars **hunden**].
- **Mannen:** Substantiv.
- **bet:** Verb.
- **hunden:** Substantiv.

Huvudord



- [Den galne **mannen** med den vilda blicken] **[bet]** [den stackars **hunden**].
- **Mannen:** Substantiv. Nominalfras.
- **bet:** Verb. Verbfras.
- **hunden:** Substantiv. Nominalfras.

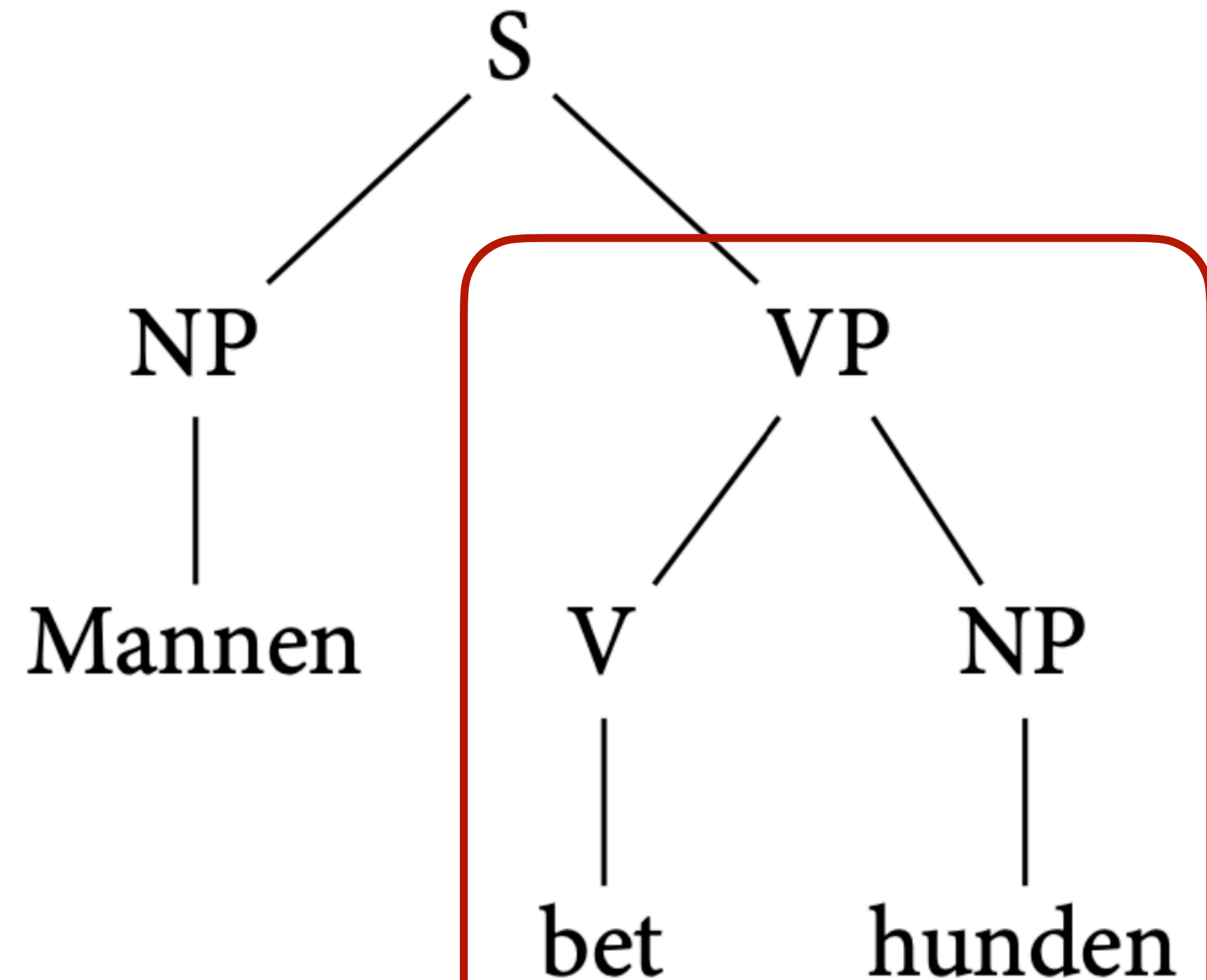
Huvudord



- [Den galne **mannen** med den vilda blicken] **[bet]** [den stackars **hunden**].
- **Mannen:** Substantiv. Nominalfras.
- **bet:** Verb. Verbfras.
- **hunden:** Substantiv. Nominalfras.
- Ordet fras används även om frasen består av **ett enda ord**.

Frasstrukturträd

- Fraser har en rekursiv struktur – de kan i sin tur bestå av mindre fraser.
- Strukturen kan beskrivas genom ett **frasstrukturträd**.



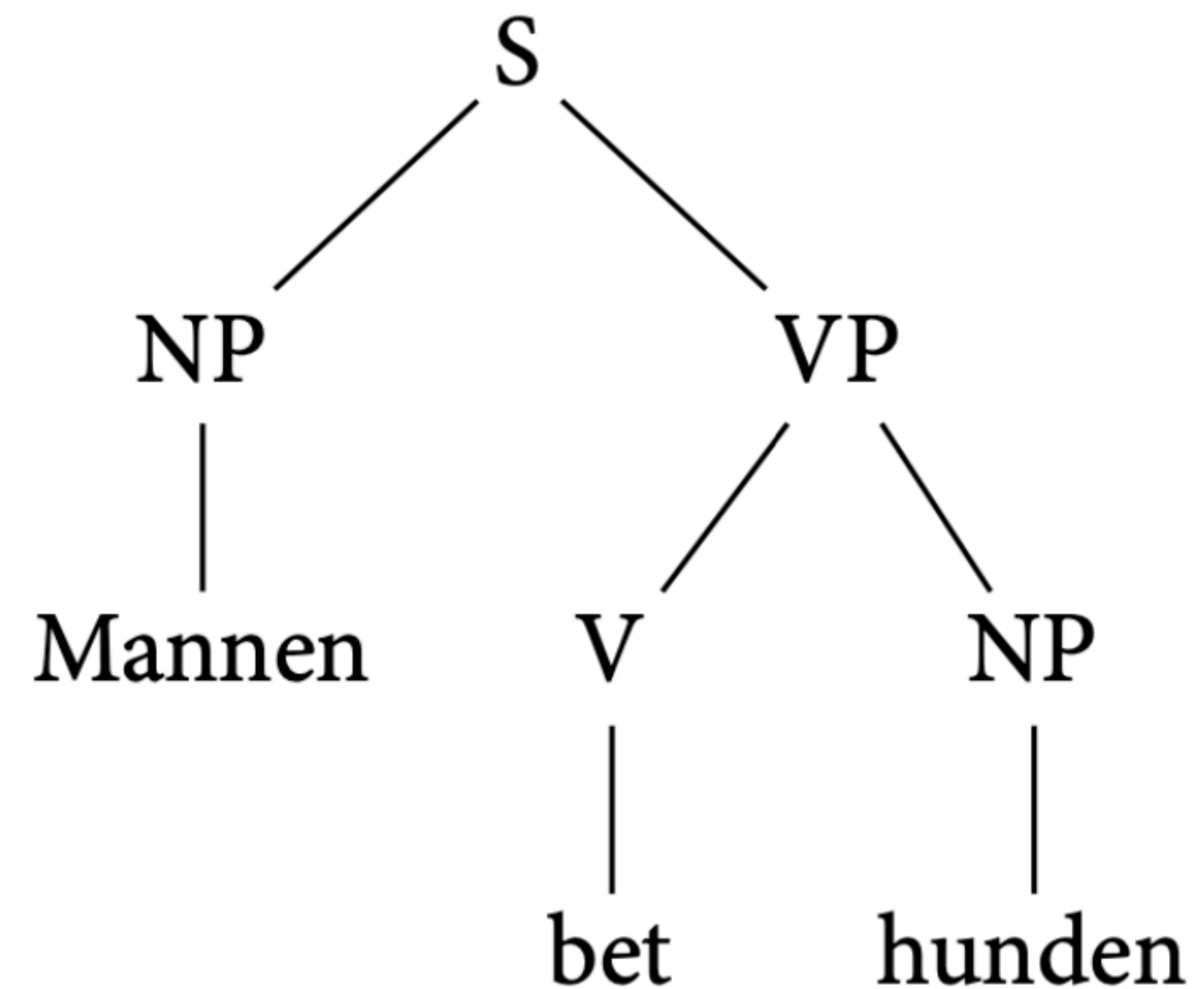
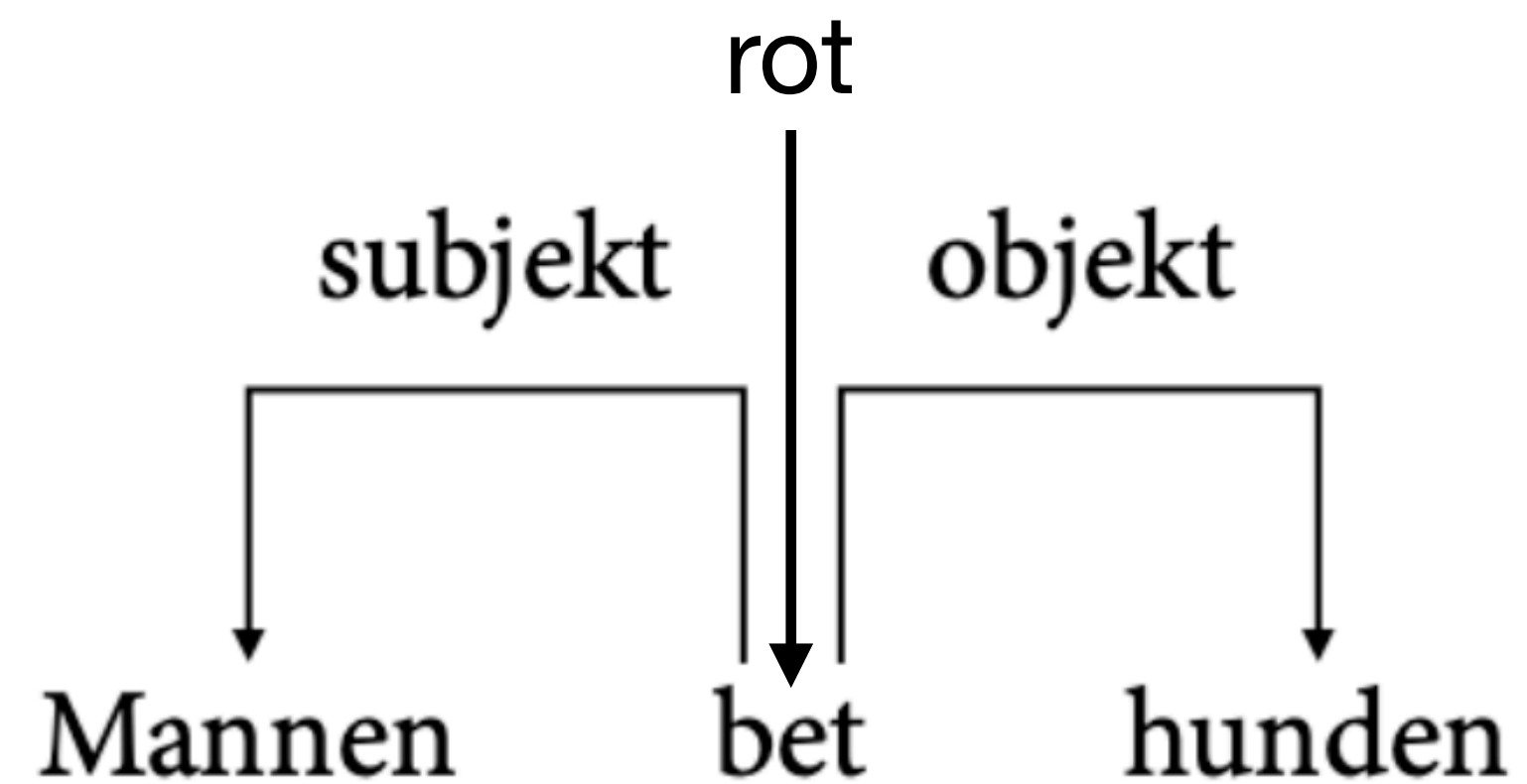
Dependensträd

Dependensanalys



- En dependensanalys är ett annat sätt att beskriva den syntaktiska strukturen hos en fras eller sats.
 - Vilka ord är deponenter till vilka andra ord?
- Precis som i en frasstrukturanalys avbildas dependensstrukturen med hjälp av ett träd, ett så kallat dependensträd.
- Dependensträd är mer minimalistiska och plana:
 - I motsats till ett frasstrukturträd har ett dependensträd inga frasnoder; **alla noder svarar mot ord.**

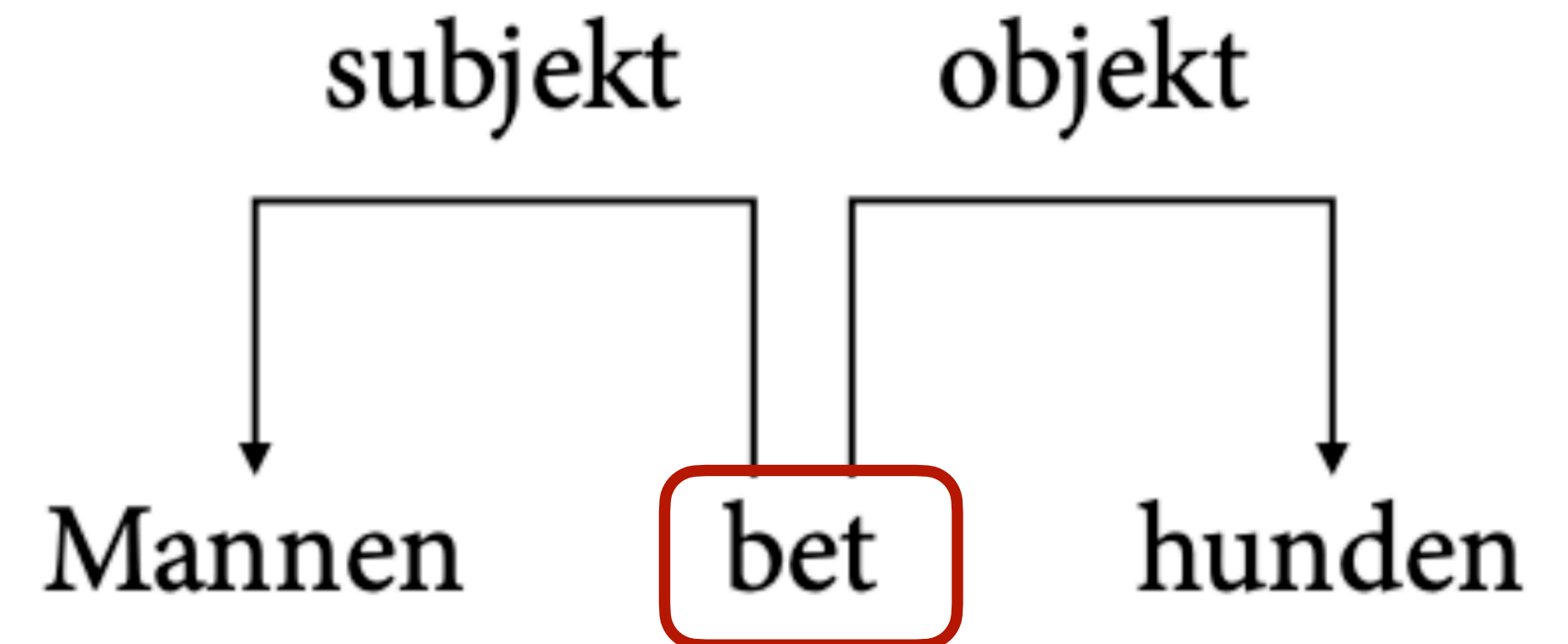
Dependensträd och Frasstrukturträd



Huvudord och Dependenter



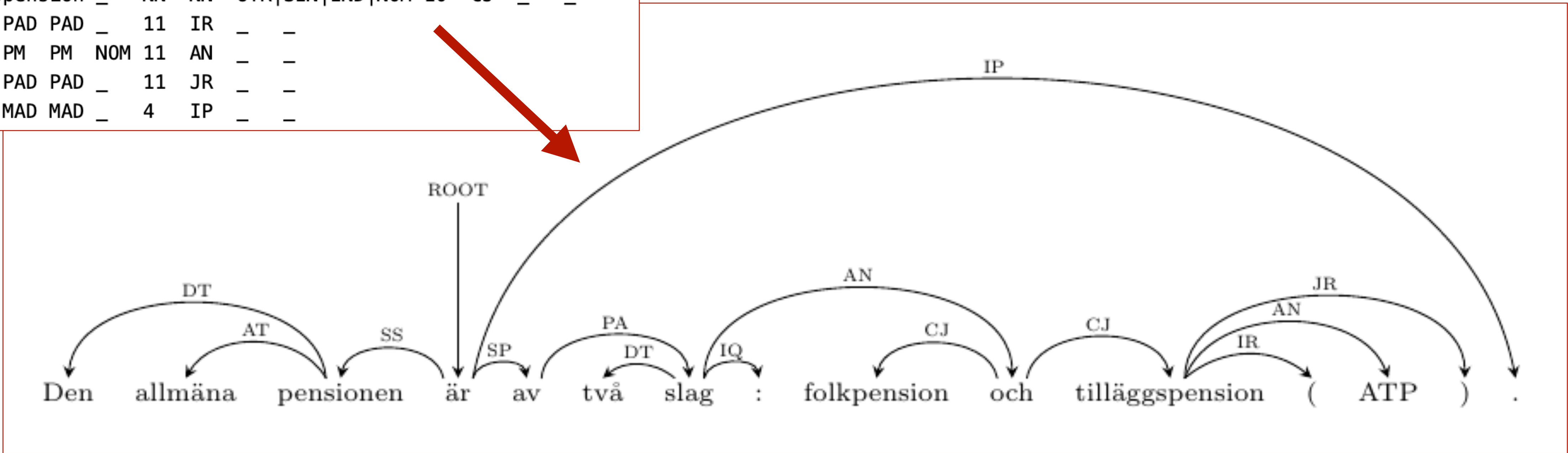
- Varje båge i ett dependensträd går från ett **huvudord** till ett underordnat ord som kallas **dependent**.
- Bågarna är etiketterade med **grammatiska relationer** såsom subjekt, objekt eller determinerare.
- Ett dependensträd formaliserar således den information som uttrycks vid traditionell satslösning (**funktionell analys**).



Från Annotation till Träd



1	Den	_	DT	DT	UTR SIN DEF	3	DT	_	_
2	allmänna	_	JJ	JJ	POS UTR/NEU SIN DEF NOM	3	AT	_	_
3	pensionen	_	NN	NN	UTR SIN DEF NOM	4	SS	_	_
4	är	_	VB	VB	PRS AKT	Ø	ROOT	_	_
5	av	_	PP	PP	_	4	SP	_	_
6	två	_	RG	RG	NOM	7	DT	_	_
7	slag	_	NN	NN	NEU PLU IND NOM	5	PA	_	_
8	:	_	MID	MID	_	7	IQ	_	_
9	folkpension	_	NN	NN	UTR SIN IND NOM	10	CJ	_	_
10	och	_	KN	KN	_	7	AN	_	_
11	tilläggspension	_	NN	NN	UTR SIN IND NOM	10	CJ	_	_
12	(	_	PAD	PAD	_	11	IR	_	_
13	ATP	_	PM	PM	NOM	11	AN	_	_
14	)	_	PAD	PAD	_	11	JR	_	_
15	.	_	MAD	MAD	_	4	IP	_	_



Hitta huvudet



- Huvudordet kan ofta ersätta frasen:
 - [Den galne mannen med den vilda blicken] bet [den stackars hunden].
 - Mannen bet hunden.
- Huvudordet är obligatoriskt; dependenten behöver inte vara det
 - Genom skattereformen införs [individuell beskattning].
- Dependents form beror på huvudordets form:
 - Sanne älskar [röda bilar].

Exempel



- En student läste en bok.
- Lektorn slog studenten med linjalen.

Fraser svarar mot delträd (1)



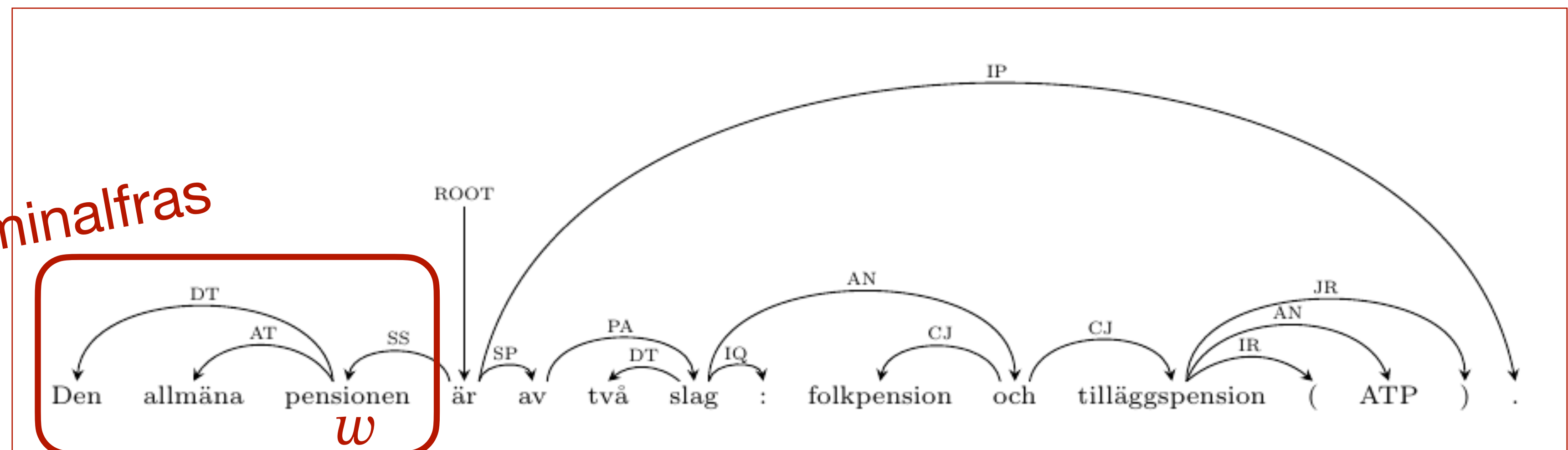
- I en dependensanalys finns inga speciella frasnoder, som i en frasstrukturanalys. Fraser svarar istället mot delträd i en dependensanalys.
 - **Delträd** = en nod och alla noder som nås från den noden.
- Vi kan definiera frasen med ordet w som huvud som den ordsekvens som utgörs av w och alla de ord som är direkt eller indirekt underordnade w .

Fraser svarar mot delträd (2)



- I en dependensanalys finns inga speciella frasnoder, som i en frasstrukturanalys. Fraser svarar istället mot delträd i en dependensanalys.
 - **Delträd** = en nod och alla noder som nås från den noden.
- Vi kan definiera frasen med ordet w som huvud som den ordsekvens som utgörs av w och alla de ord som är direkt eller indirekt underordnade w .

Nominalfras



Verbet i Dependensträdet

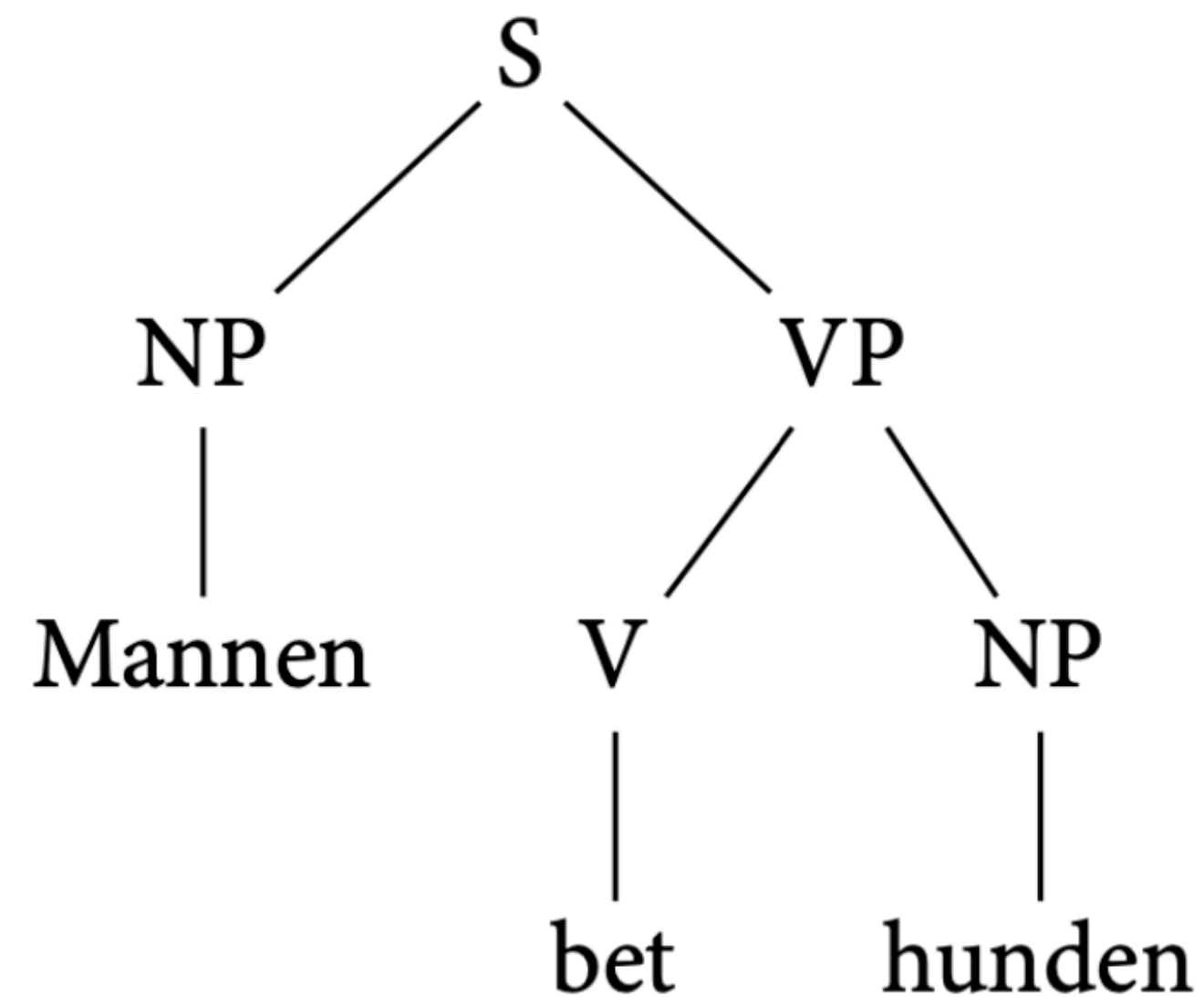
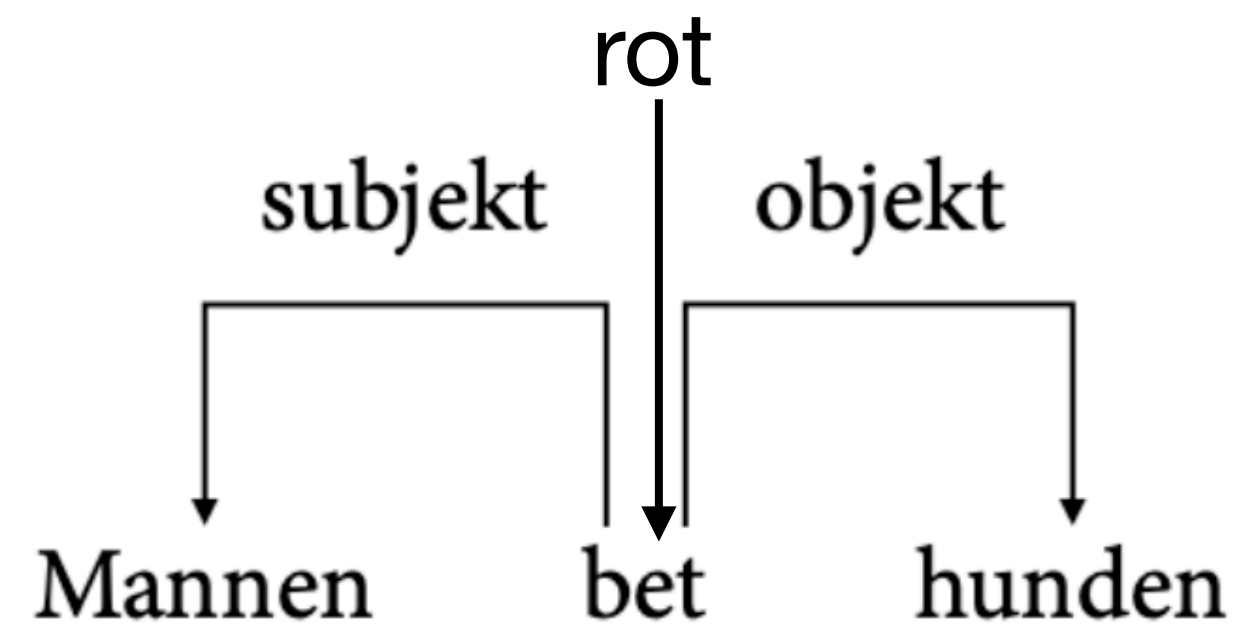


- I frasstrukturträdet är subjekt och predikat sidoordnade.
- I dependensträd är det finita verbet toppnoden. Subjektet blir en dependent, vid sidan av ev. andra led (objekt etc.).
 - Det är verbet som definierar resten av meningen:
 - Vilka andra led måste, eller kan, finnas?

Relationer i Dependensträdet

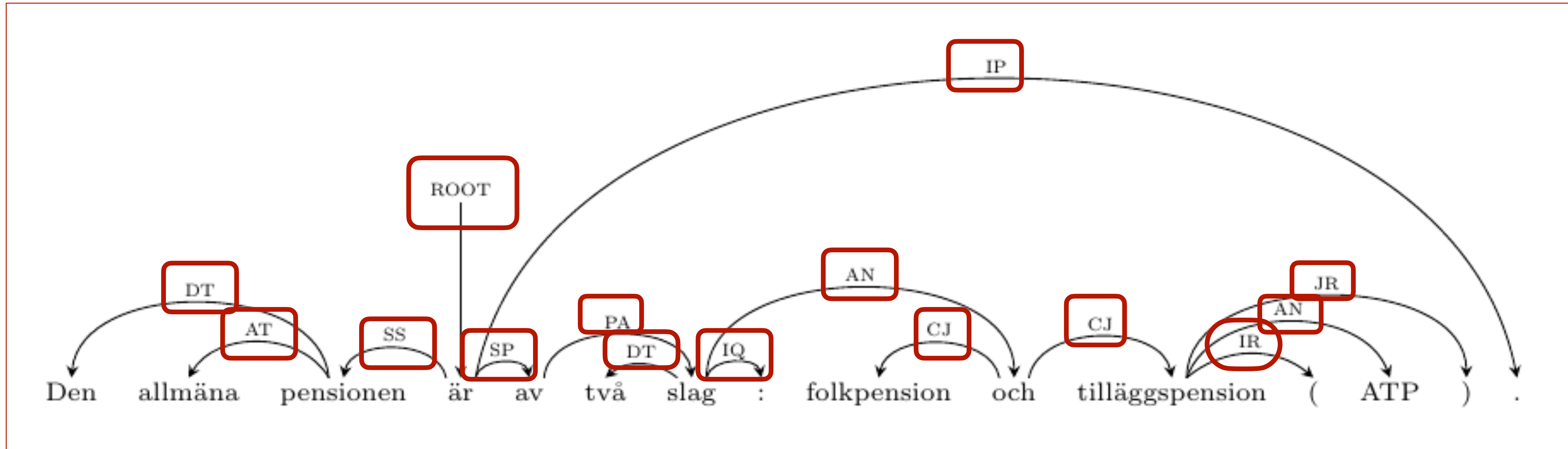


- Relationer i dependensträdet är **1:1**:
 - Varje ord motsvarar 1 led i meningsstrukturen.
- Frasstrukturträdet: Minst +1 led!



Grammatiska Funktioner i Dependensträd

- Varje nod i dependensträdet har en funktion, som har ett namn.



Grammatiska funktioner från Swedish Treebank

(Utval)

Etikett	Grammatisk funktion
SS	Annat subjekt
DT	Determinerare
OO	Direkt objekt
IO	Indirekt objekt
PA	Komplement till en preposition
OA	Objektsadverbial
IP	Interpunktion

Dependensträd i Språkteknologi

- Dependensträd är mycket mer populära i språkteknologi.
- Enklare struktur: Varje ord har exakt 1 huvudord och det inte finns icke-terminala symbol.
- Det gör även parsning enklare:
 - Mer efficients algoritmer.
 - Lättare att träna modeller.
 - Lättare att integrera i system.
- Mer semantiska: Vilket ord är dependent till vilket annat ord?

Syntaktisk Parsning

- En **parser** är ett datorprogram (eller algoritm) som tar **en mening** som indata (t.ex. en lista med ord), och returnerar dess **syntaktiska struktur**:
 - Identifiera vilket ord som är huvudord,
 - vilka andra leder den finns (tills alla ord har exakt ett huvudord),
 - och vilken grammatiskt relation dem har.
- Moderna parsers använder **maskininlärning**.
 - De är tränade på stora korpusar som har **annoterade dependensstrukturer** (t.ex. Universal Dependencies).

Syntaktisk Tvetydighet

Syntaktisk Tvetydighet

- Vilka dependensträd kan vi rita till denna mening?

Jag såg mannen med kikaren.

Exempel

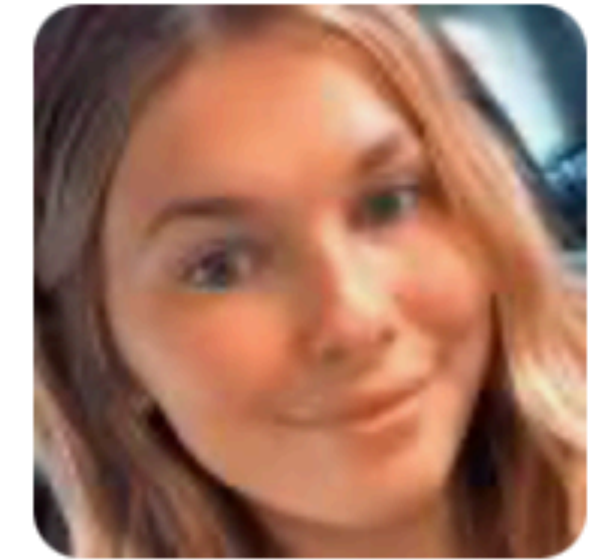


Daily Mail

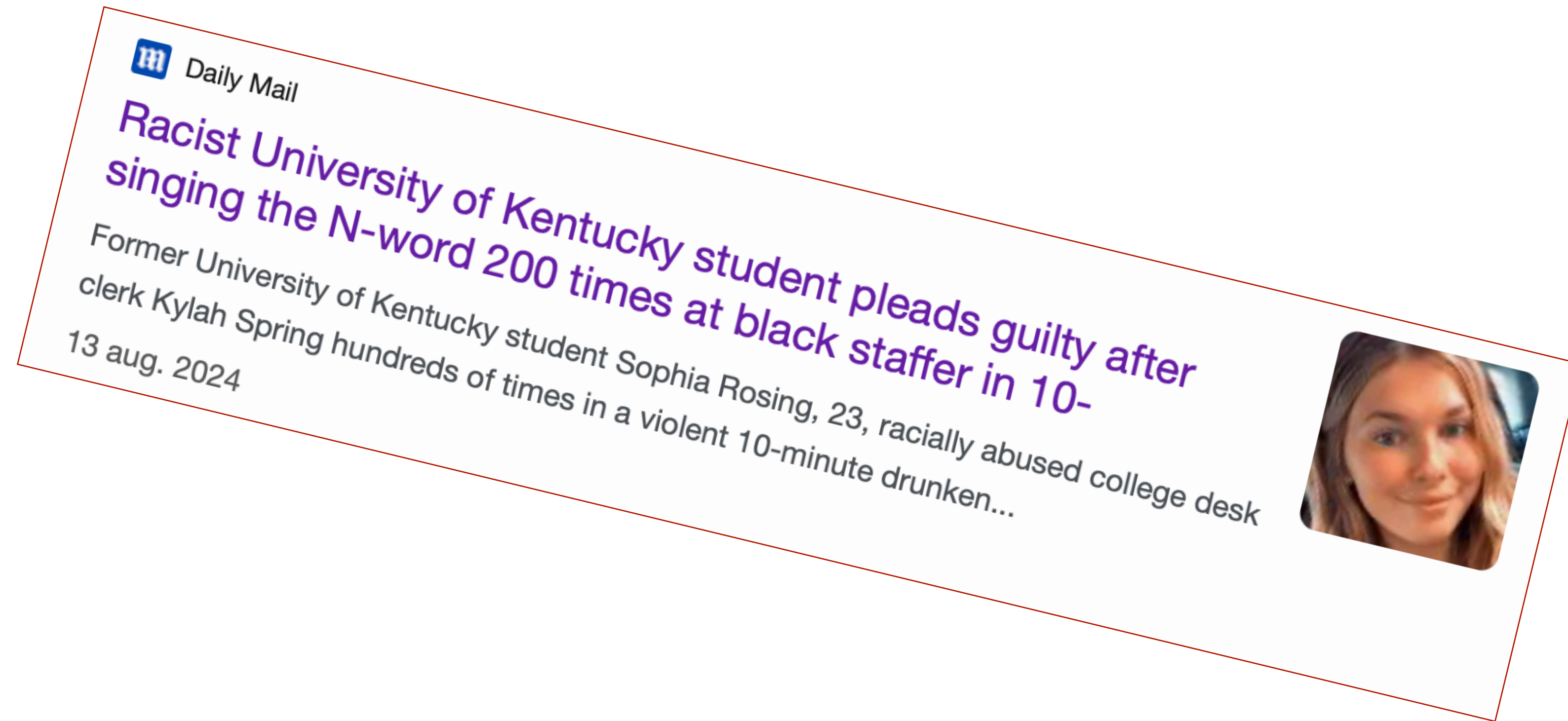
Racist University of Kentucky student pleads guilty after singing the N-word 200 times at black staffer in 10-

Former University of Kentucky student Sophia Rosing, 23, racially abused college desk clerk Kylah Spring hundreds of times in a violent 10-minute drunken...

13 aug. 2024



Exempel



- Racist University of Kentucky student pleads guilty [...].
- Racist University of Kentucky student pleads guilty [...].

Tolka Tvetydighet

- Ofta är det **kontexten** som avgör vilken tolkning är mer sannolikt:
 - Sammanhang i texten: *"Jag stod uppe i utkikstornet och spanade ut över ängen. Då fick jag syn på någon som rörde sig mellan träden långt borta. Jag såg mannen med kikaren."*
 - Sammanhang i världen: Ganska osannolikt att det finns en *"Racist University of Kentucky"*.
- Ibland är en tolkning mer syntaktiskt rimligt än andra.
 - *"The old man the boats."*

Tvetydighet i Språkteknologi

- För att tolka meningar på ett korrekt sätt behöver en modell alltså:
 - Tillräckligt med kontext: En mening i taget, som det är vanligt i syntaktisk parsning, räcker inte för alla uppgifter och meningar.
 - Kunnskap om världen och domänen.
 - Kunnskap om sannolikheter i syntaktiska strukturer.

Tvetydighet i Maskinöversättning

- Syntaktisk tvetydighet är inte bara ett problem när man ritar dependensträd. Det kan också påverka andra applikationer, t.ex. maskinöversättning:
 - *"The old man the boats."*
 - ~~*"Den gamla mannen båtarna."*~~
 - *"De gamla bemannar båtarna."*
 - *"Jag såg mannen med kikaren."*
 - *"I saw the man through the binoculars."* (kanske???)

Ordföljder i Språkteknologi

I Labb 3

Labb 3, Uppgift 4

✓ Uppgift 7: Ordföljd (reflektion)

I lingvistikdelen lär ni er att den "raka" ordföljden i svenskan följer mönstret subjekt–verb–objekt, men att det förutom denna mest vanliga ordföljd även finns alternativa ordföljder. I reflektionsuppgiften ska ni försöka hitta empiriska belägg för olika ordföljder.

Till er hjälp får ni en funktion `lab3.problem7()` som skriver ut meningar i korpusfilen vars uppbyggnad följer ett visst mönster. Detta mönster anges som ett reguljärt uttryck som kan referera till grammatiska relationer. Mönstret i nedanstående kodcell t.ex. matchar meningar som består av exakt ett subjekt (`SS`), ett verb (`VB`) och ett direkt objekt (`OO`), i exakt den ordningen och utan andra satsdelar eller interpunktion:

```
[ ] lab3.problem7('^SS VB OO IP$')
```

Funktionen lab3.problem7()

- Hur ser funktionen ut?
- Tyvärr för komplex för att vara en uppgift i labbet :(
- Varför?

Funktionen lab3.problem7()

- Hur ser funktionen ut?
- Tyvärr för komplex för att vara en uppgift i labbet :(
- Varför?
- Gå igenom träden, hitta & spara delträd.
 - Varför är det viktigt att ha koll på delträd?

Delträdens Betydelse i Uppgiften

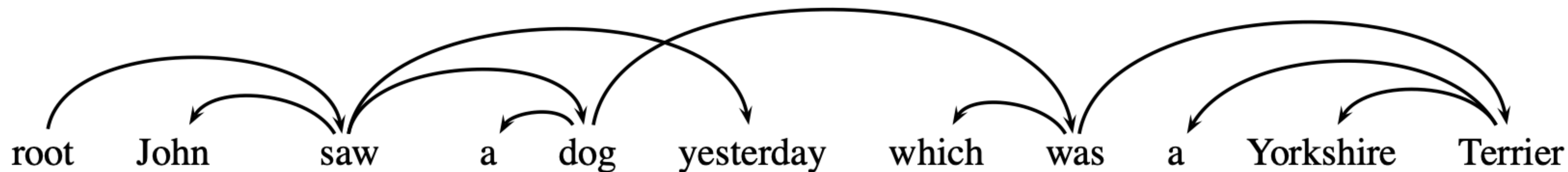
- 
- Jag **är** på vägen men tåget **är** snart framme!
 - Om vi bara går igenom meningen linjärt kan vi inte alltid tolka ordföljden korrekt.

Ordföljd i Olika Syntaktiska Trädstrukturer

Ordföljd i Frasstrukturträd

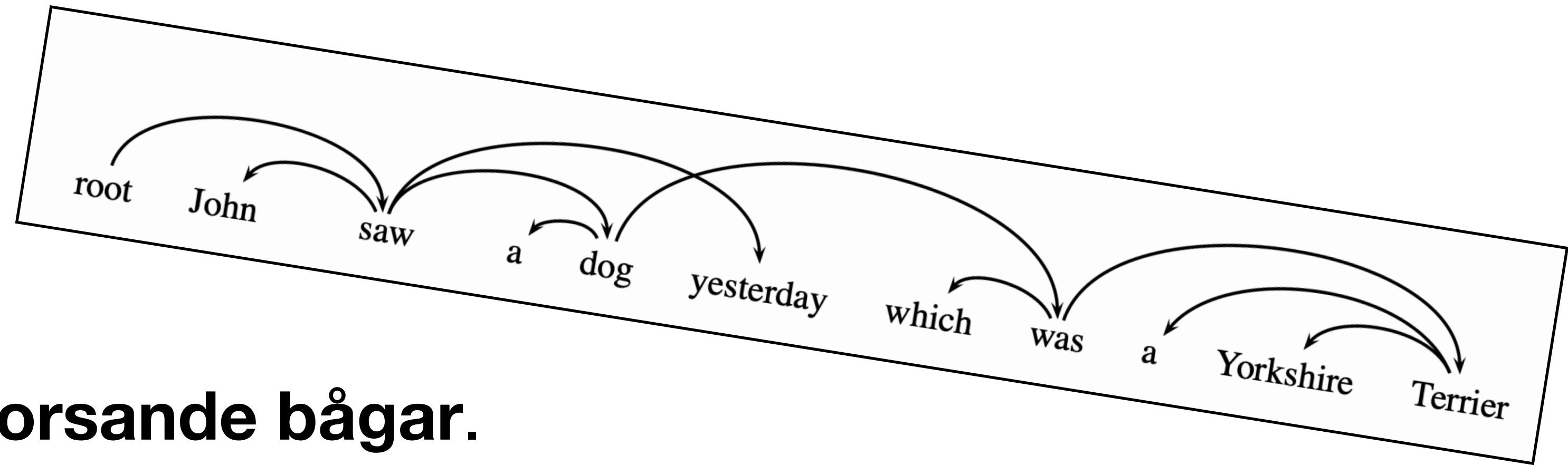
- Frasstrukturträd är anpassade till språk med ganska strikt ordföljd, t.ex. engelska och svenska.
 - Fraser måste vara sammanhängande sekvenser av ord.
 - De kan bara modellera nästlade dependenser utan korsande bågar.
- Många språk har dock fri(are) ordföljd...
- John saw **a dog** yesterday **which was a Yorkshire Terrier**.

Dependensträd är Flexibla!



Non-Projective Dependency Parsing using Spanning Tree Algorithms (McDonald et al., HLT-EMNLP 2005)

Dependensträd är Flexibla!



- I dependensträd kan man ha **korsande bågar**.
 - Sekvensen av ord i ett delträd kan innehålla "luckor".
 - Bättre för flerspråkiga system och analyser.
- Ett träd som innehåller inga korsande bågar kallas för **projektiv**.
- Ett träd som innehåller korsande bågar kallas för **icke-projektiv**.

L3

Tack för idag!