T-(538|725)-MALV, Málvinnsla
Þáttunaraðferðir

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

Október 2007

Outline

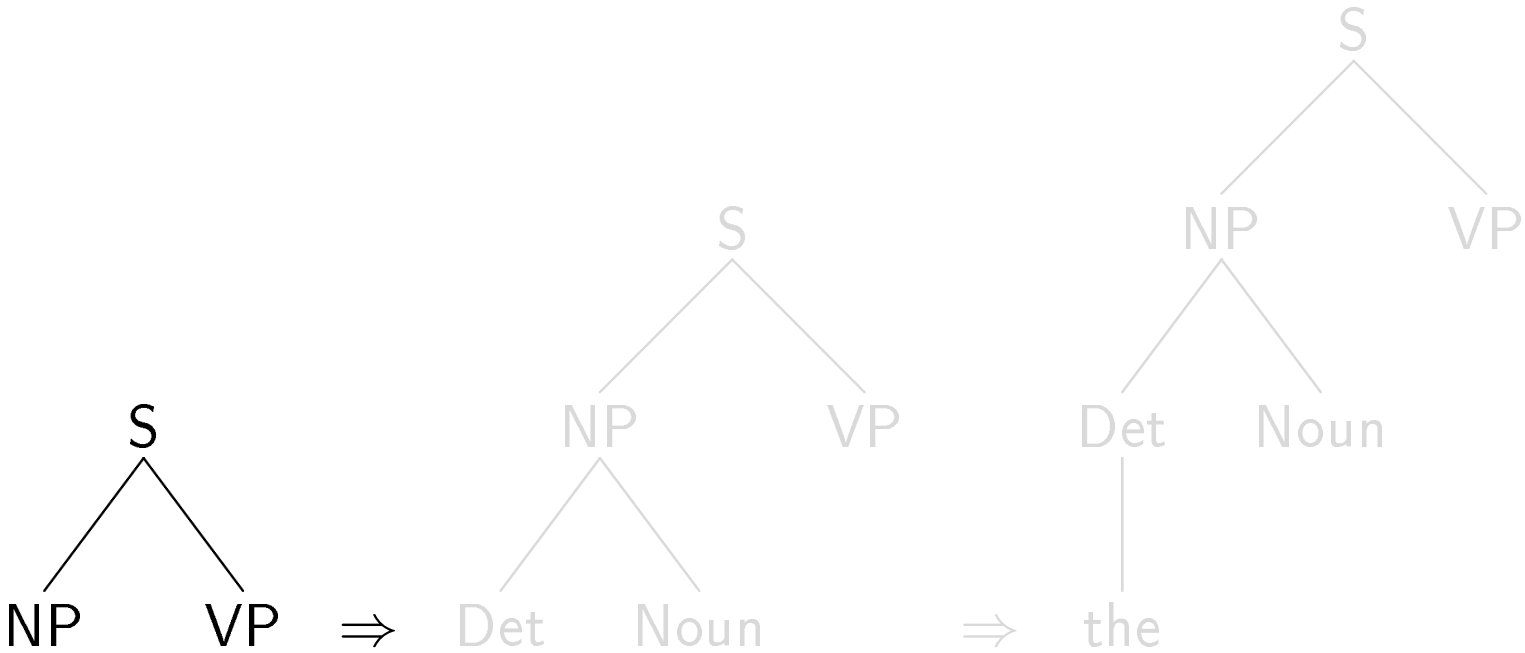
Outline

Ofansækin þáttun (e. top-down parsing)

- Byggir þáttunartré frá rót niður í lauf.
- Byrjað á byrjunareiningu, S.
- Öll hluttré búin til sem hafa S vinstra megin við örina: $S \rightarrow X$
- Síðan haldið áfram í næsta hluttréi (hnúti), X.
- Reglur fundnar sem hafa X vinstra megin við örina og öll hluttré búin til fyrir hægri hliðina.
- Síðan koll af kolli þangað til komið er niður að orðunum (laufunum).
- Trjám sem ekki passa er hent.

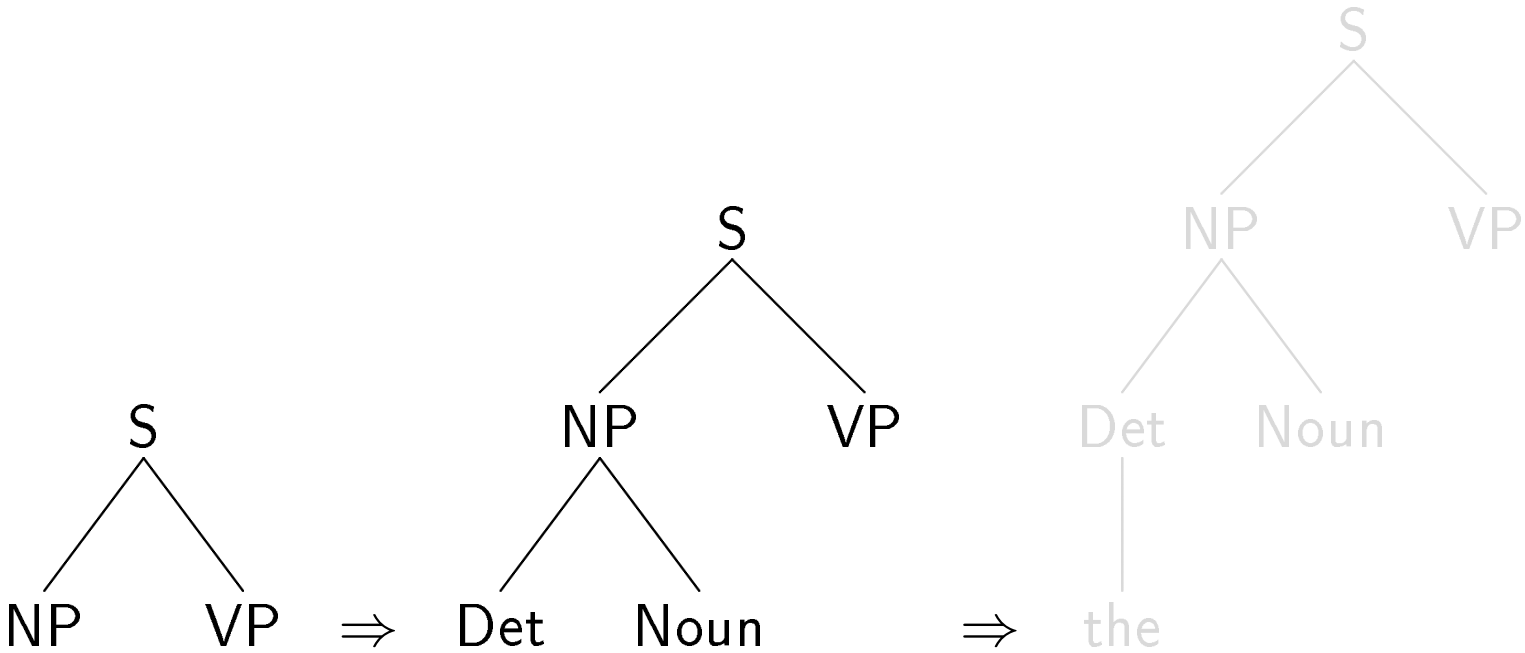
Ofansækin þáttun: Dæmi

The waiter brought the meal



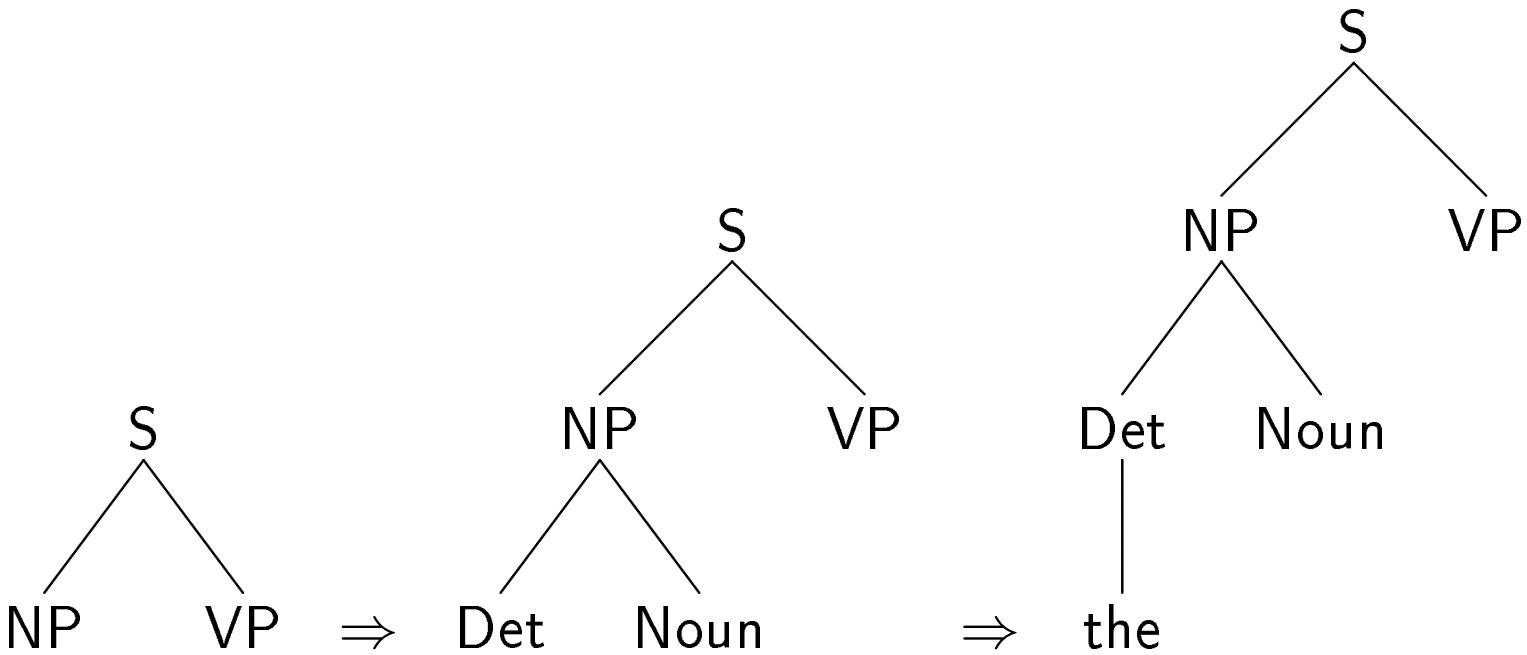
Ofansækin þáttun: Dæmi

The waiter brought the meal



Ofansækin þáttun: Dæmi

The waiter brought the meal



Ofansækin þáttun

- Notað t.d. fyrir DCG (Definite Clause Grammar) í Prolog.
- “Depth-first strategy”:
  - Ræður ekki við vinstri endurkvæmni, t.d.:
  - np –> np, pp.
  - np –> np, conj, np.
- “Backtracking”:
  - Endurþáttun sömu setningaliða getur átt sér stað mörgum sinnum.

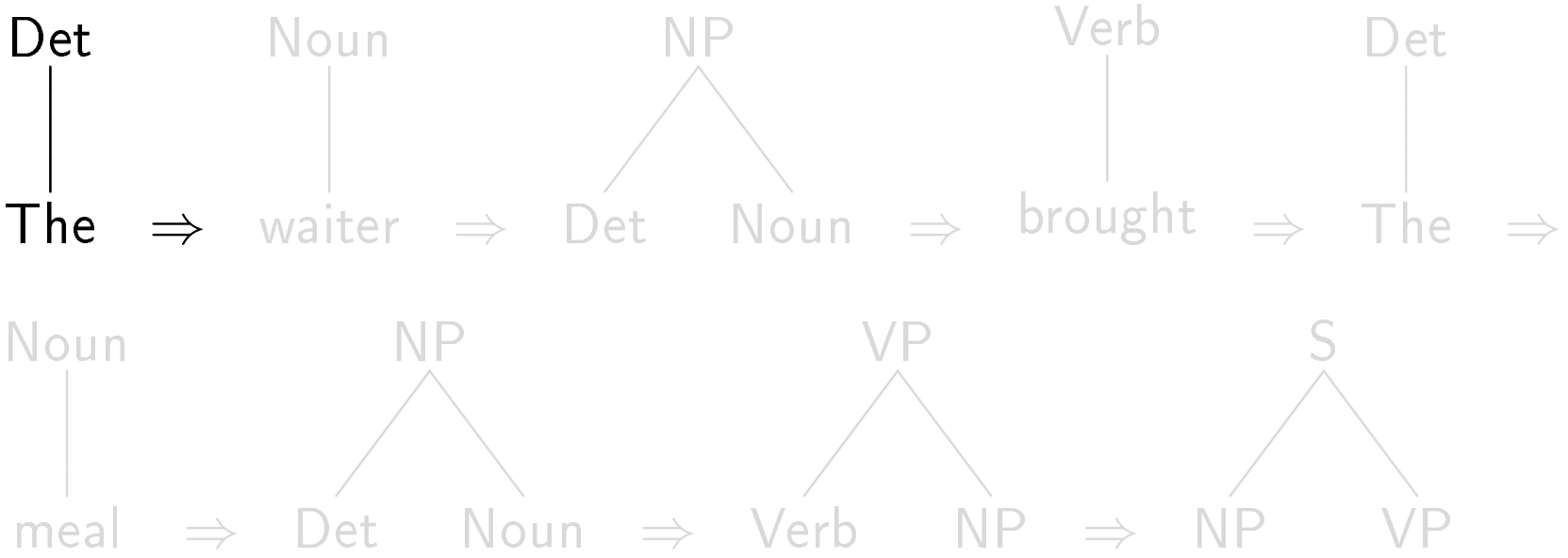
Outline

Neðansækin þáttun (e. bottom-up parsing)

- Byggir þáttunartré frá laufum upp í rót.
- Byrjað á orðunum.
- Athugað hvort orð W kemur fyrir hægra megin við ör í einhverri reglu: $X \rightarrow W$
- Hægri hlið skipt úr fyrir vinstri hlið, X.
- Síðan koll af kolli þangað til komið er upp í rót.
- Trjám sem ekki leiða upp í rót er hent.

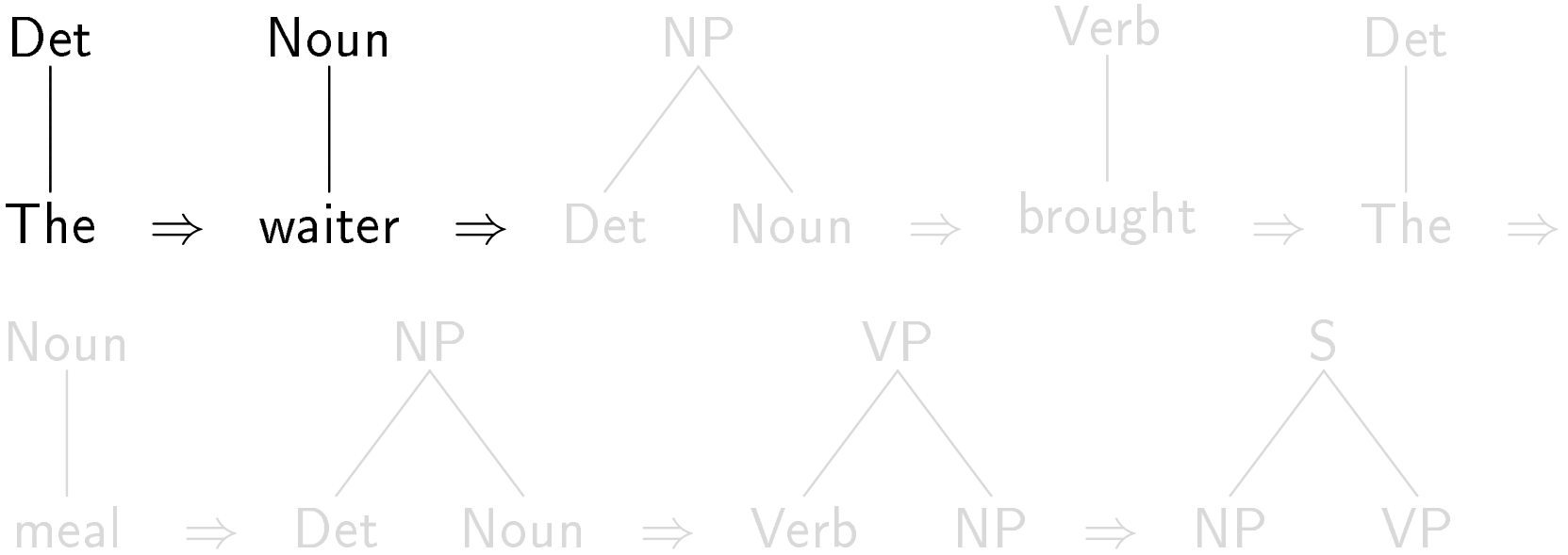
Neðansækin þáttun: Dæmi

The waiter brought the meal



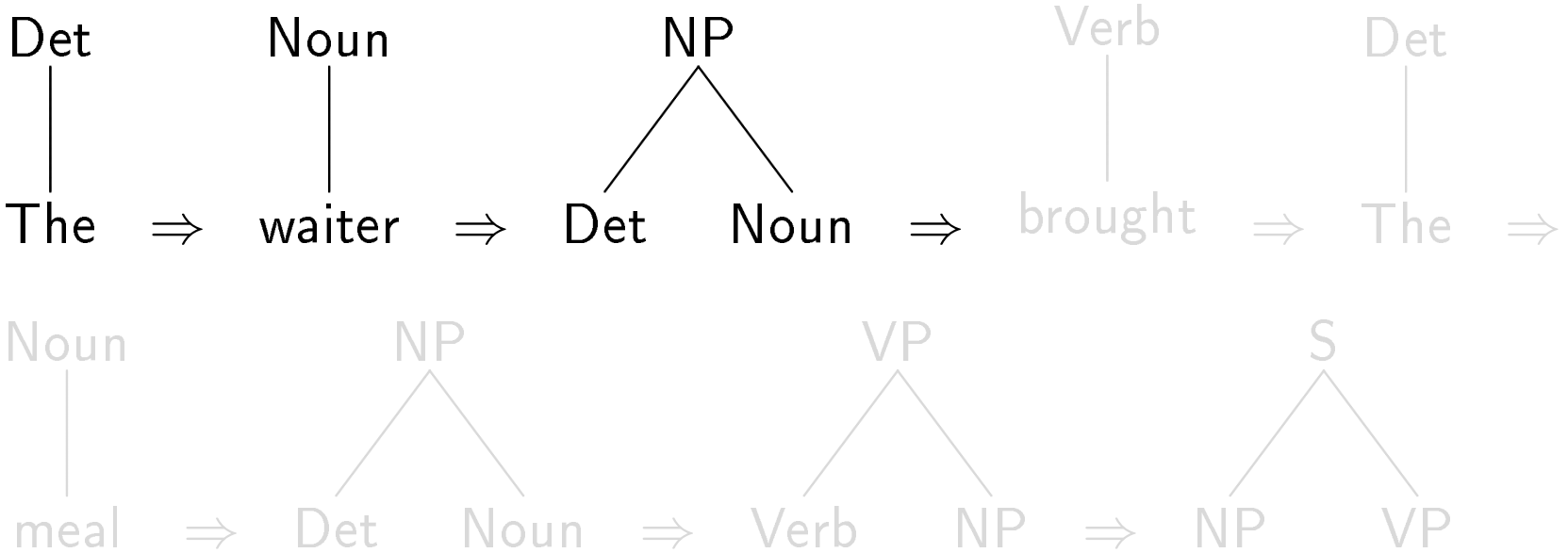
Neðansækin þáttun: Dæmi

The waiter brought the meal



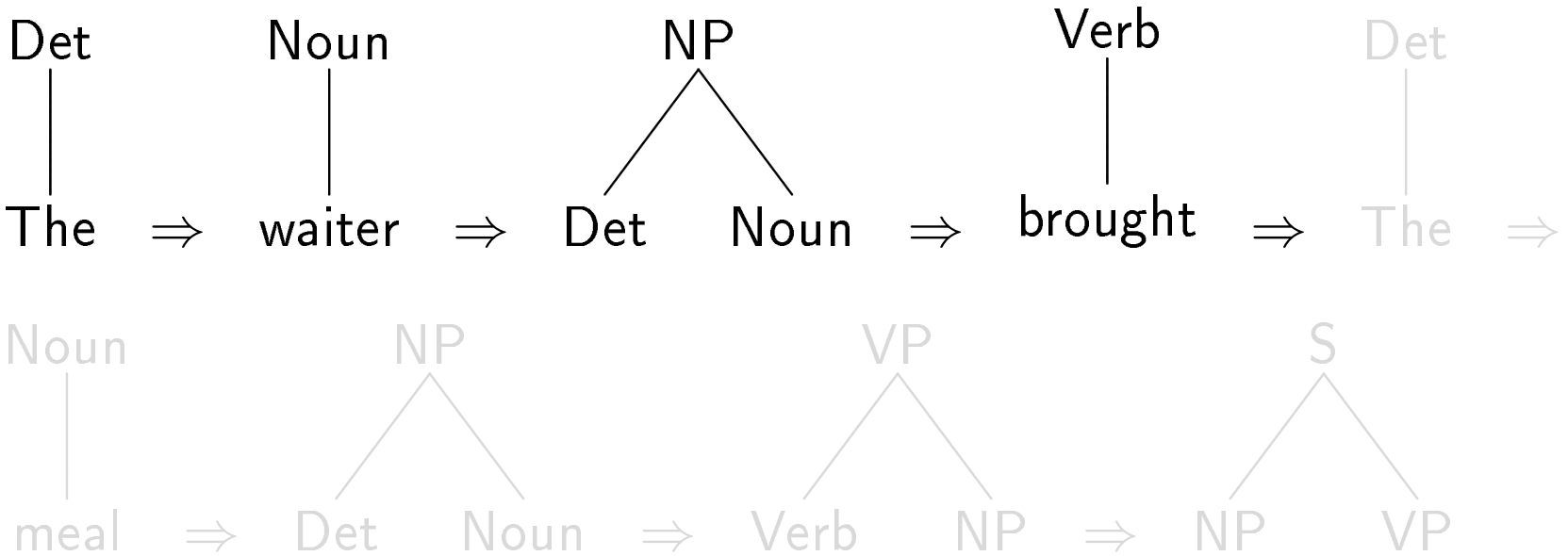
Neðansækin þáttun: Dæmi

The waiter brought the meal



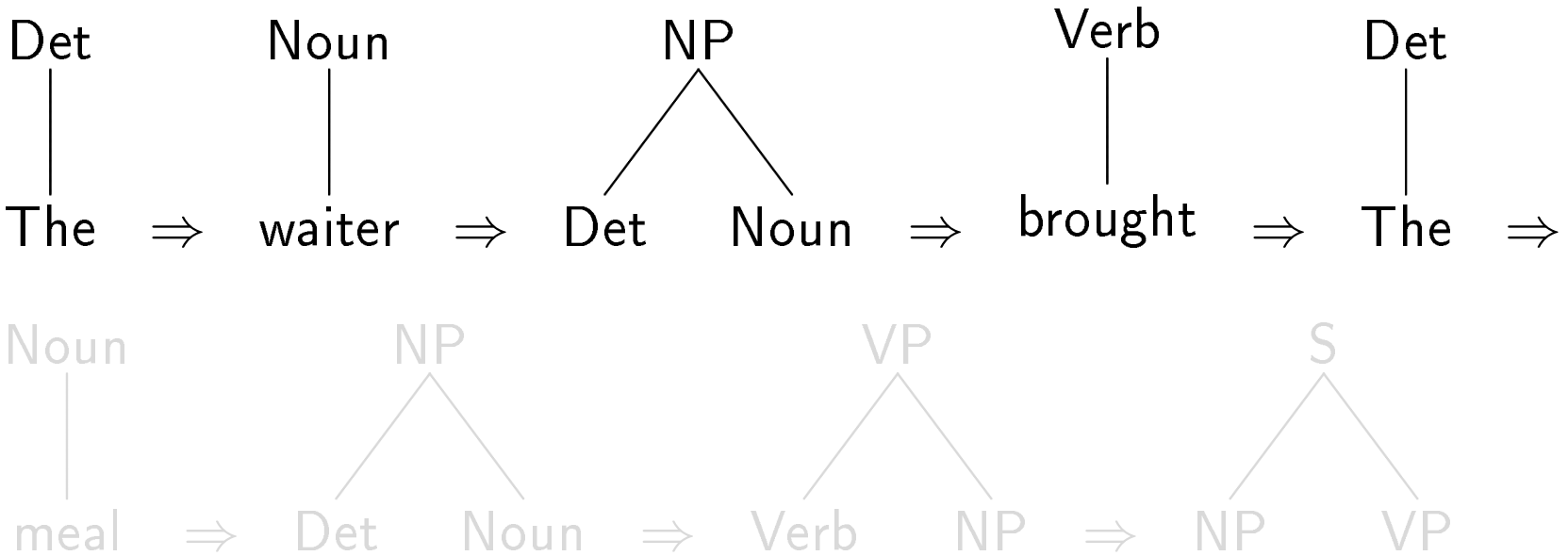
Neðansækin þáttun: Dæmi

The waiter brought the meal



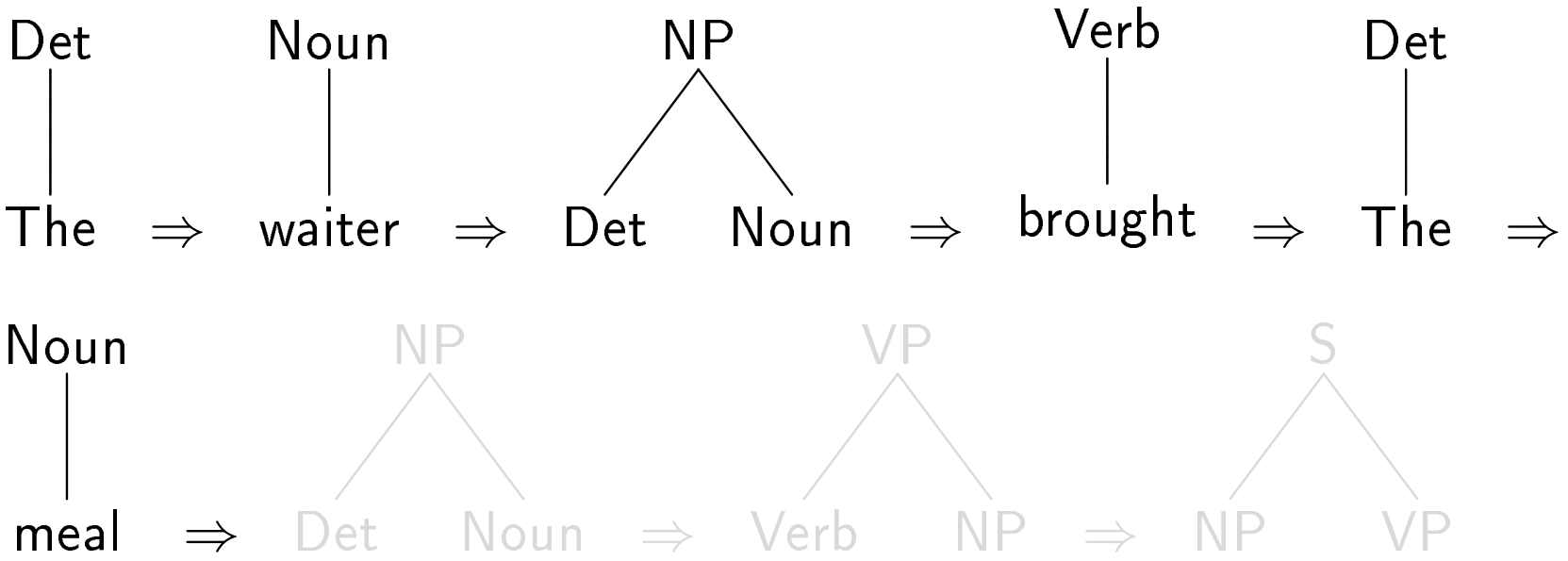
Neðansækin þáttun: Dæmi

The waiter brought the meal



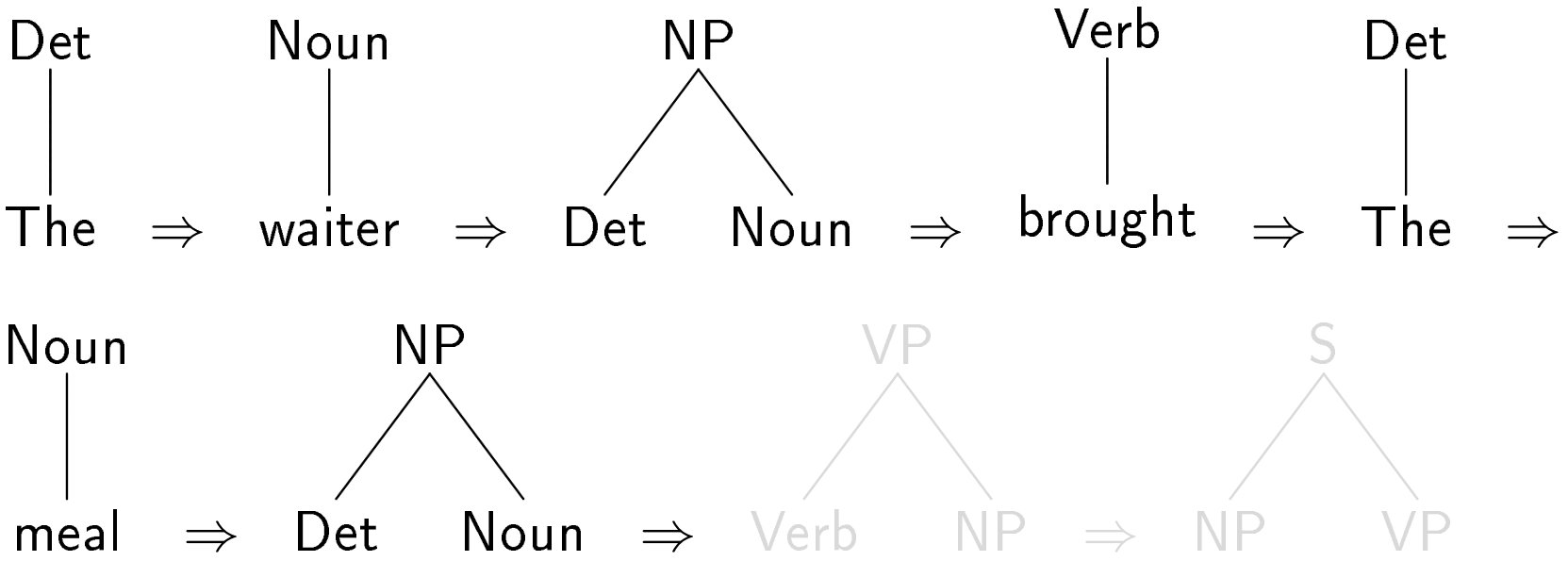
Neðansækin þáttun: Dæmi

The waiter brought the meal



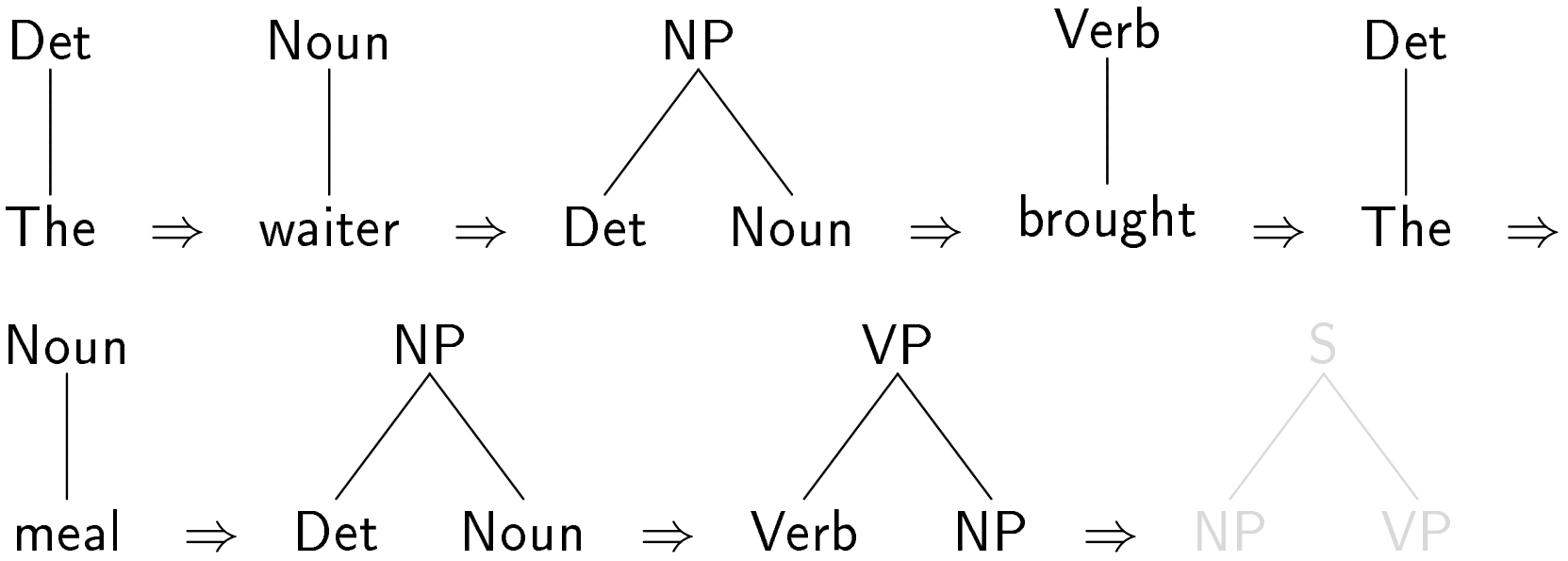
Neðansækin þáttun: Dæmi

The waiter brought the meal



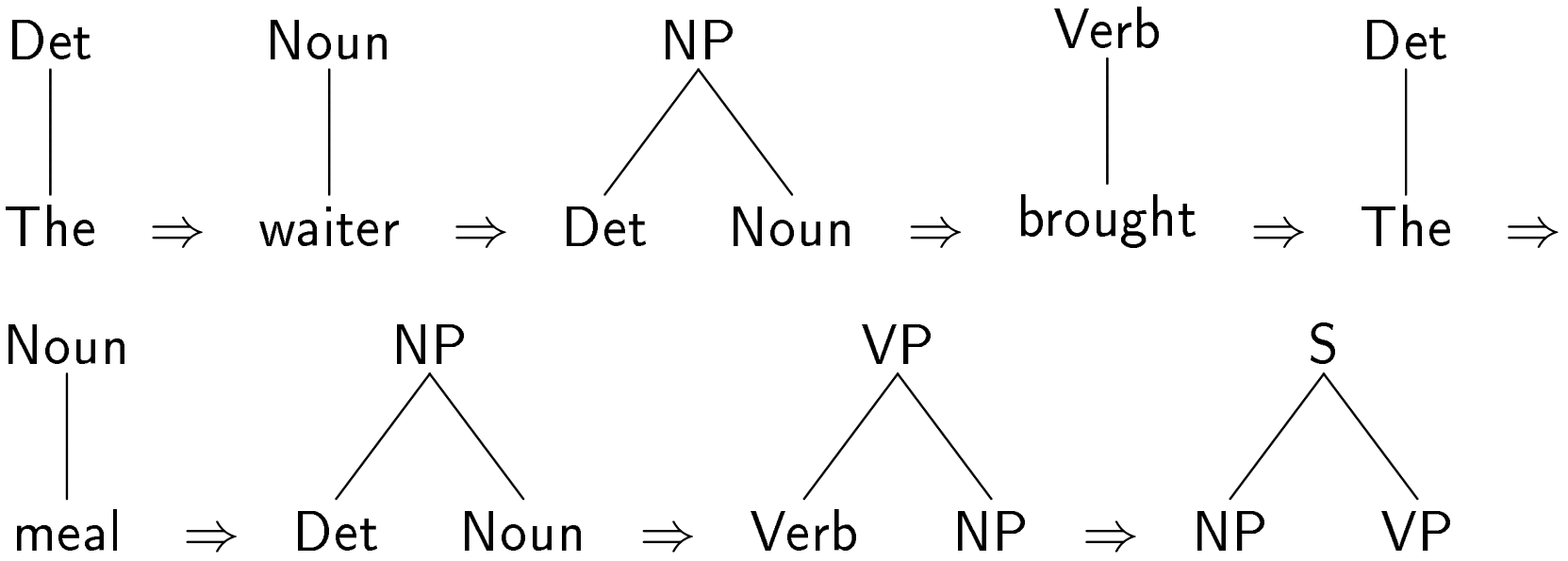
Neðansækin þáttun: Dæmi

The waiter brought the meal



Neðansækin þáttun: Dæmi

The waiter brought the meal



Shift-reduce þáttun

- Tegund neðansækinnar þáttunar:
  1. **Shift** a word from the phrase or sentence to parse onto a stack.
  2. Apply a sequence of grammar rules to **reduce** elements of the stack.
- Lykkjan endurtekin þangað til engin fleiri orð eru í listanum og staflanum hefur verið breytt (e. reduced) í byrjunareininguna.

Shift-reduce þáttun: Dæmi

It.	Stack	S/R	Word list
0			the waiter brought the meal
1	the	shift	waiter brought the meal
2	det	reduce	waiter brought the meal
3	det waiter	shift	brought the meal
4	det noun	reduce	brought the meal
5	np	reduce	brought the meal
6	np brought	shift	the meal
7	np verb	reduce	the meal
8	np verb the	shift	meal
9	np verb det	reduce	meal
10	np verb det meal	shift	
11	np verb det noun	reduce	
12	np verb np	reduce	
13	np vp	reduce	
14	s	reduce	

Ofansækin vs. neðansækin þáttun

Ofansækin þáttun

- Eyðir aldrei tíma í tré sem geta ekki endað í S.
- Notar ekki orðin til að stýra þáttuninni og því verður til fjöldi trjáa sem ekki falla að orðunum.
- Ræður við tóma liði (e. null constituents) (t.d. det –> [].)

Neðansækin þáttun

- Notar orðin til að stýra þáttuninni.
- Eyðir ekki tíma í tré sem ekki falla að orðunum.
- Notar ekki S til að stýra þáttuninni, $\Rightarrow$ býr til hluttré sem aldrei geta orðið að heilu tré.
- Ræður við vinstri endurkvæmni.

Ofansækin vs. neðansækin þáttun

Ofansækin þáttun

- Eyðir aldrei tíma í tré sem geta ekki endað í S.
- Notar ekki orðin til að stýra þáttuninni og því verður til fjöldi trjáa sem ekki falla að orðunum.
- Ræður við tóma liði (e. null constituents) (t.d. det –> [].)

Neðansækin þáttun

- Notar orðin til að stýra þáttuninni.
- Eyðir ekki tíma í tré sem ekki falla að orðunum.
- Notar ekki S til að stýra þáttuninni, $\Rightarrow$ býr til hluttré sem aldrei geta orðið að heilu tré.
- Ræður við vinstri endurkvæmni.

Outline

Endurþáttun setningaliða

- “Backtracking”, sem einkennir ofansækna þáttara, leiðir oft til endurþáttunar á setningaliðum:
  - np –> npx.
  - np –> npx, pp.
  - npx –> det, noun.
  - pp –> prep, np.
- Endurþáttun á *npx* á sér stað fyrir streng eins og: “The meal of the day”?
- (Hægt að leysa með svokallaðri vinstri þáttun (e. left factoring) en þá þarf að breyta mállýsingu; sjá námskeiðið Þýðendur.)

Töfluþáttun (e. chart parsing)

Hvað er?

- Aðferð til að sleppa við endurþáttun á sömu liðum.
- "Chart" er gagnaskipan sem geymir allar mögulegar hlutgreiningar fyrir tiltekna stöðu í setningu.
- Þegar næsta orð er sótt þá er undanfarandi hlutgreining líka sótt í stað þess að þátta upp á nýtt.
- Getur geymt hluttré, fullt tré (sjá Fig. 11.4) eða lýsingu á setningalið sem verið er að þátta (sjá Fig. 11.7).

Töfluþáttun

```
s --> vp.          s --> np.              vp --> v, np, pp.
vp --> v, np.      np --> det, noun.      np --> det, adj, noun.
np --> np, pp.     pp --> prep, np.
```

Regla með punkti (e. **dotted rule**)

- Stendur fyrir það sem hefur verið þáttað hingað til.
- np –> det noun • (inactive arc)
- np –> det • noun (active arc)
- np –> • det noun (active arc)

Reglur með punktum: Dæmi

Rules	Arcs	Constituents
s –> • vp	[0,0]	• Bring the meal
vp –> v • np	[0,1]	Bring • the meal
np –> • det noun	[1,1]	• the meal
np –> • np pp	[1,1]	• the meal
np –> det • noun	[1,2]	the • meal
np –> det noun •	[1,3]	the meal •
np –> np • pp	[1,3]	the meal •
vp –> v np •	[0,3]	Bring the meal •
s –> vp •	[0,3]	Bring the meal •

Taflan sýnir ekki allar mögulegar reglur.

Töfluþáttun

Earley algrímið

- An efficient context-free parsing algorithm (1970)
  - `http://portal.acm.org/citation.cfm?id=362035`
- Ofansækið algrím.
- Ræður við vinstri endurkvæmni og tóma liði.
- Notar þrjár aðgerðir:
  - Predictor
  - Scanner
  - Completer

Earley algrímið

Predictor

- Velur þær reglur sem hægt er að beita á virka leggi (e. active arcs)
- Fyrir reglu: lhs –> $c_1 c_2 \ldots \bullet c \ldots c_n$
- verða til nýjar reglur: c –> $\bullet x_1 x_2 \ldots x_k$

Scanner

- Samþykkir ný orð úr inntakinu.
- Orðflokkur (pos) hægra megin við punktinn er borinn saman við orð í inntakinu.
- Setur reglu pos –> word • inn í töfluna.

Earley algrímið

Predictor

- Velur þær reglur sem hægt er að beita á virka leggi (e. active arcs)
- Fyrir reglu: lhs –> $c_1 c_2 \ldots \bullet c \ldots c_n$
- verða til nýjar reglur: c –> $\bullet x_1 x_2 \ldots x_k$

Scanner

- Samþykkir ný orð úr inntakinu.
- Orðflokkur (pos) hægra megin við punktinn er borinn saman við orð í inntakinu.
- Setur reglu pos –> word $\bullet$ inn í töfluna.

Earley algrímið

Completer

- Notar þær reglur sem Scanner myndar til að færa punktinn á virkum leggjum sem gerðu ráð fyrir að fá orðflokk næst.
- Leitar fyrst að reglum sem hafa punktinn aftast: c –> $x_1 x_2 \ldots x_n \bullet$
- Síðan að reglum lhs –> $c_1 c_2 \ldots \bullet c \ldots c_n$ og
- býr til nýja reglu lhs –> $c_1 c_2 \ldots c \bullet \ldots c_n$

Earley algrímið: Dæmi

Chart#	Rules	Arcs	Module	Constituents
0	s –> • np	[0,0]	Start state	• the meal of the day
0	np –> • det noun	[0,0]	Predictor	• the meal of the day
0	np –> • det adj noun	[0,0]	Predictor	• the meal of the day
0	np –> • np pp	[0,0]	Predictor	• the meal of the day
1	det –> the •	[0,1]	Scanner	• meal of the day
1	np –> det • noun	[0,1]	Completer	• meal of the day
1	np –> det • adj noun	[0,1]	Completer	• meal of the day
2	noun –> meal •	[1,2]	Scanner	• of the day
2	np –> det noun •	[0,2]	Completer	• of the day
2	np –> np • pp	[0,2]	Completer	• of the day
2	pp –> • prep np	[2,2]	Predictor	• of the day

Earley algrímið: Dæmi (framhald)

Chart#	Rules	Arcs	Module	Constituents
3	prep –> of •	[2,3]	Scanner	• the day
3	pp –> prep • np	[2,3]	Completer	• the day
3	np –> • det noun	[3,3]	Predictor	• the day
3	np –> • det adj noun	[3,3]	Predictor	• the day
3	np –> • np pp	[3,3]	Predictor	• the day
4	det –> the •	[3,4]	Scanner	• day
4	np –> det • noun	[3,4]	Completer	• day
4	np –> det • adj noun	[3,4]	Completer	• day
5	noun –> day •	[4,5]	Scanner	
5	np –> det noun •	[3,5]	Completer	
5	pp –> prep np •	[2,5]	Completer	
5	np –> np pp •	[0,5]	Completer	
0	s –> np •	[0,5]	Completer	

Outline

Tölfræðileg þáttun (e. probabilistic parsing)

- Viljum stundum geta fundið líklegasta þáttunartréð . . .
- . . . í stað þess að búa til öll möguleg tré.
- Getum gert það ef trjábanki (e. treebank) er til.
  - Trjábanki er setningagreind málheild.

PCFG

- PCFG – **P**robabilistic **C**ontext **F**ree **G**rammar (Collins 1996; Charniak 1997)
- Samhengisfrjáls mállýsing þar sem hverri málreglu fylgir líkindi $P(lhs \rightarrow rhs|lhs)$
- Raunveruleg líkindi eru fengin úr trjábönkum.

Tölfræðileg þáttun

- Nálgun með sennileikalíkum (e. maximum likelihood estimation):

$$P(lhs \rightarrow rhs_i | lhs) = \frac{Count(lhs \rightarrow rhs_i)}{\Sigma_j Count(lhs \rightarrow rhs_j)}$$

- Líkurnar á því að setning S hafi þáttunartré T er skilgreint sem margfeldi líkinda sem tengjast sérhverri reglu sem notuð er til að mynda tréð:

$$P(T, S) = \prod_{rule(i) Producing T} P(rule(i))$$

- Sjá útreikning á líkum fyrir tvö þáttunartré á bls. 296.