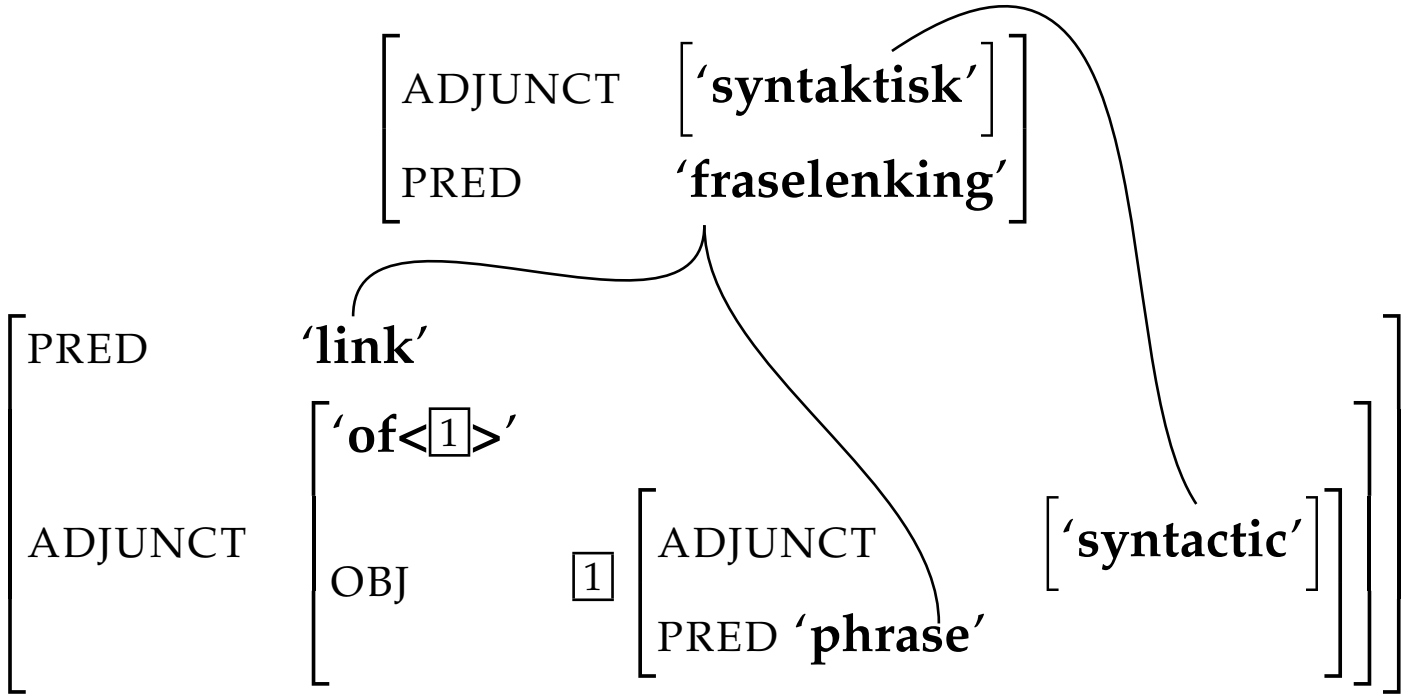


Kevin Brubeck Unhammer

Universitetet i Bergen

10. desember 2010

FRASELENKING

- Kva og kvifor
- Standardmetoden: N-grambasert

LENKINGSKRAV

- F-strukturlenkjer
- Rangering
- C-strukturlenkjer

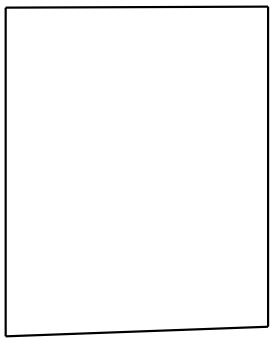
VANSKAR

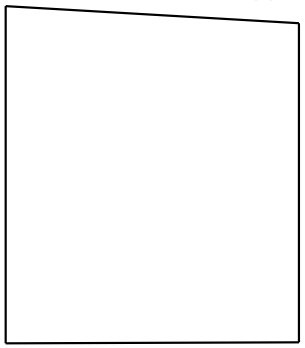
- Dårleg inndata
- Generalisering til nye grammatikkar
- Når predikat ikkje korresponderer 1-til-1

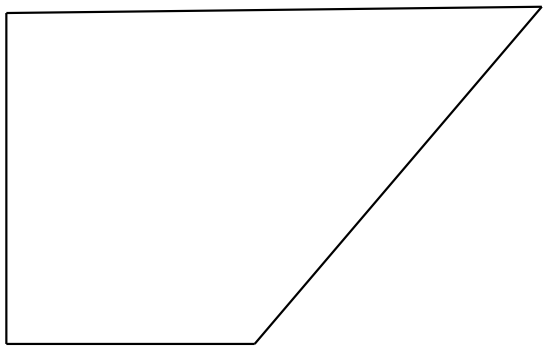
FRASELENKING

- Finne frasar som korresponderer i omsetjingar
  - «frase» = konstituent? dependenseining? syntaktisk funksjon? N-gram? chunk?
- Data: vanlegvis N-gramtabellar frå statistisk samanstilling, reint korpusbasert
- Formål: vanlegvis statistisk maskinomsetjing

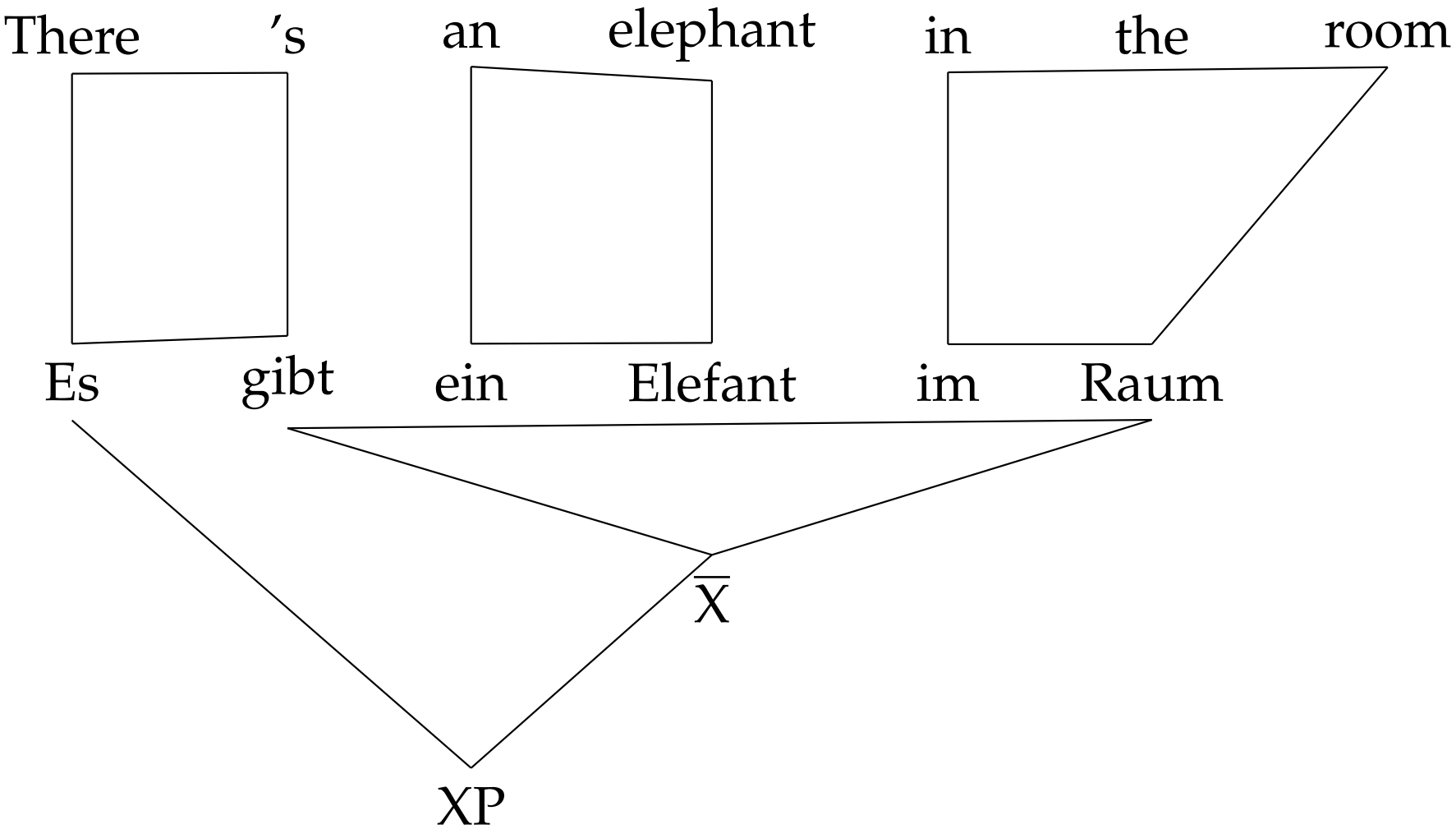
There ’s an elephant in the room

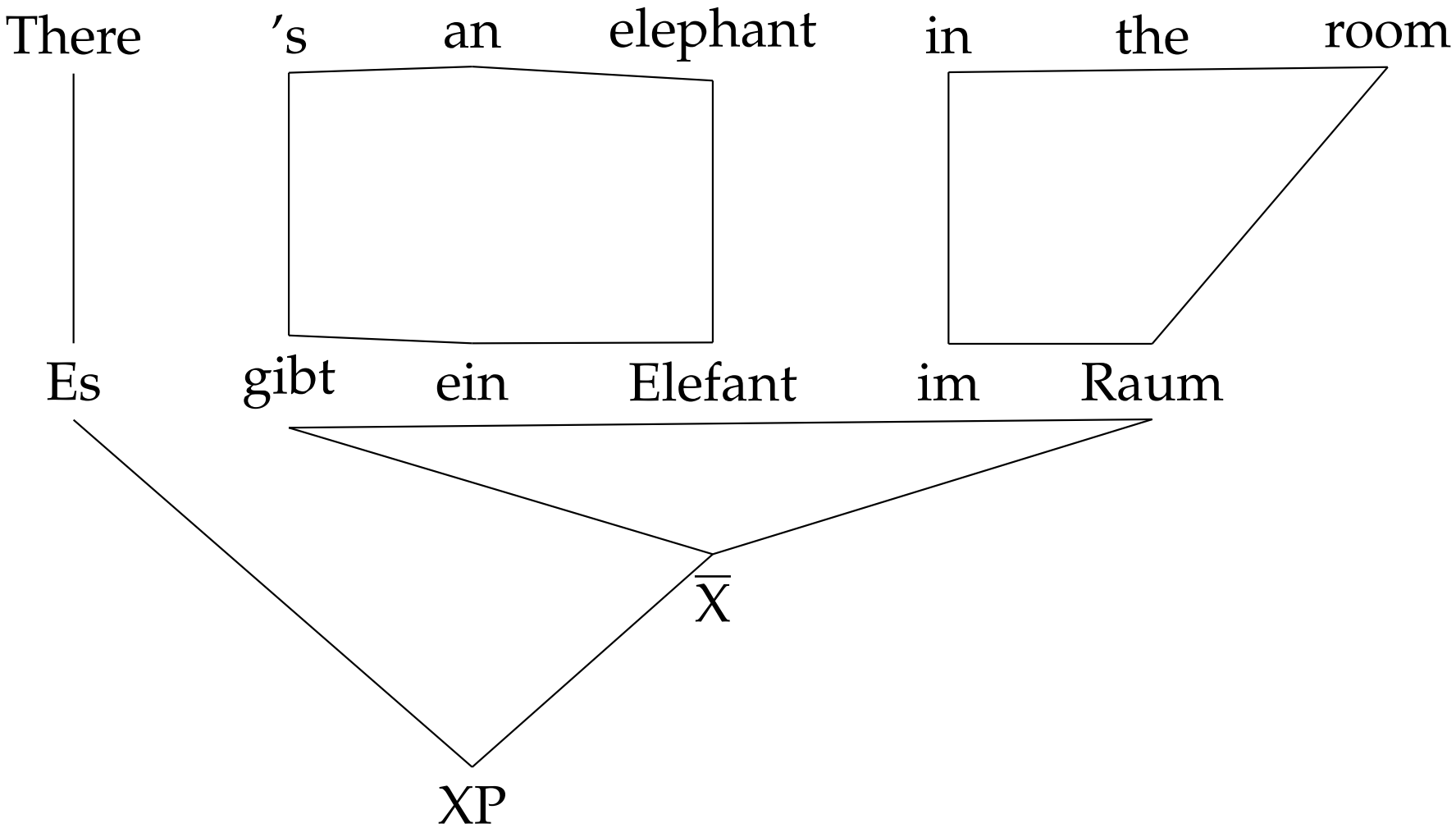






Es gibt ein Elefant im Raum


There
's
an
elephant
in
the
room
Es
gibt
ein
Elefant
im
Raum
X̄
XP


There
's
an
elephant
in
the
room
Es
gibt
ein
Elefant
im
Raum
X̄
XP

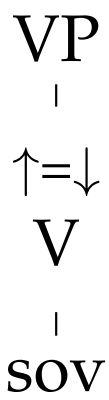
CAT N-GRAMTABELL | FRASELENKING | SYNTAKS

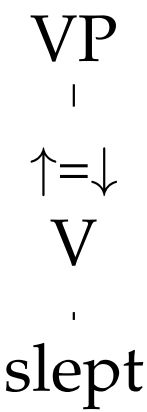
- N-grambasert (datadriven/korpusbasert) lenking kan filtrerast med kunnskap
  - berre lenkjer som samsvarer med *syntaktiske nodar* (Samuelsson & Volk, 2007)
  - berre lenkjer som samsvarer med ein *dependensanalyse* (Hearne mfl., 2008)
  - berre lenkjer som samsvarer med ein *f-strukturanalyse* (Graham & Genabith, 2009)

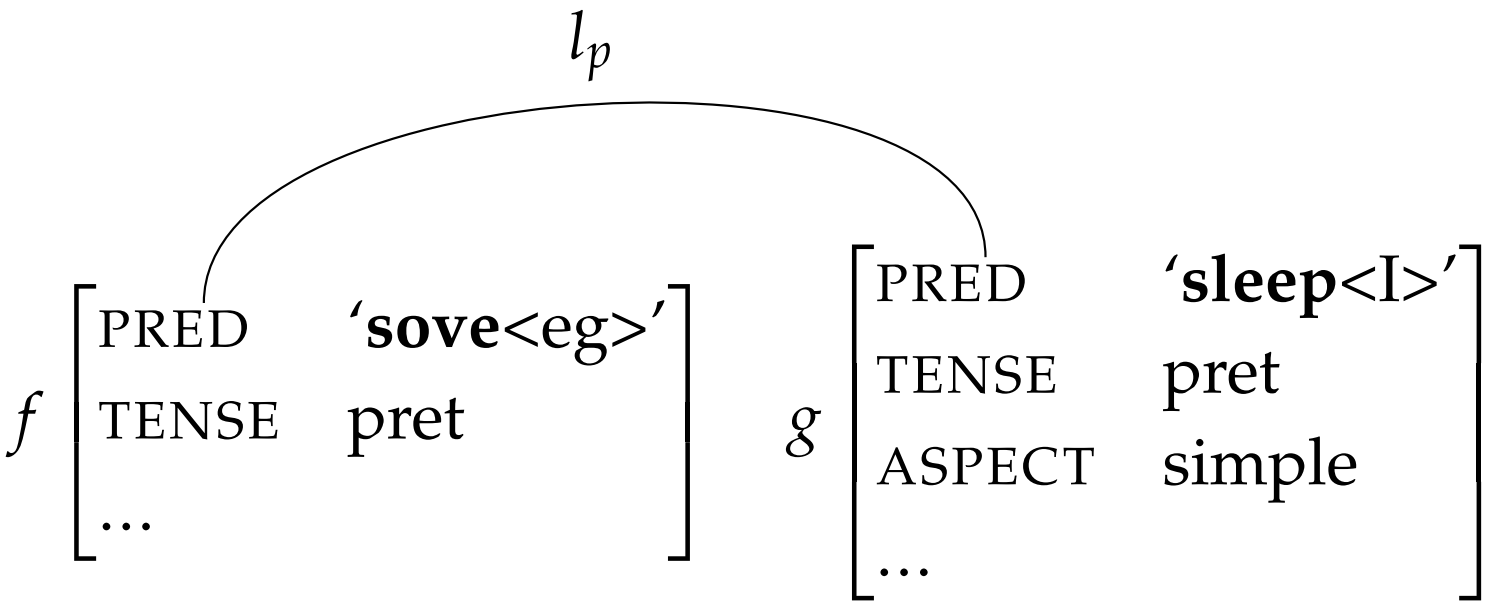
CAT SYNTAKS | FRASELENKING | N-GRAMTABELL

- Men me kan au gå andre vegen
  - Data: LFG-analysar
  - Formål: annotert trebank
  - Kunnskapsdriven lenking, kan ev. filtrerast med N-gramtabell (eller omsetjingsordbok)

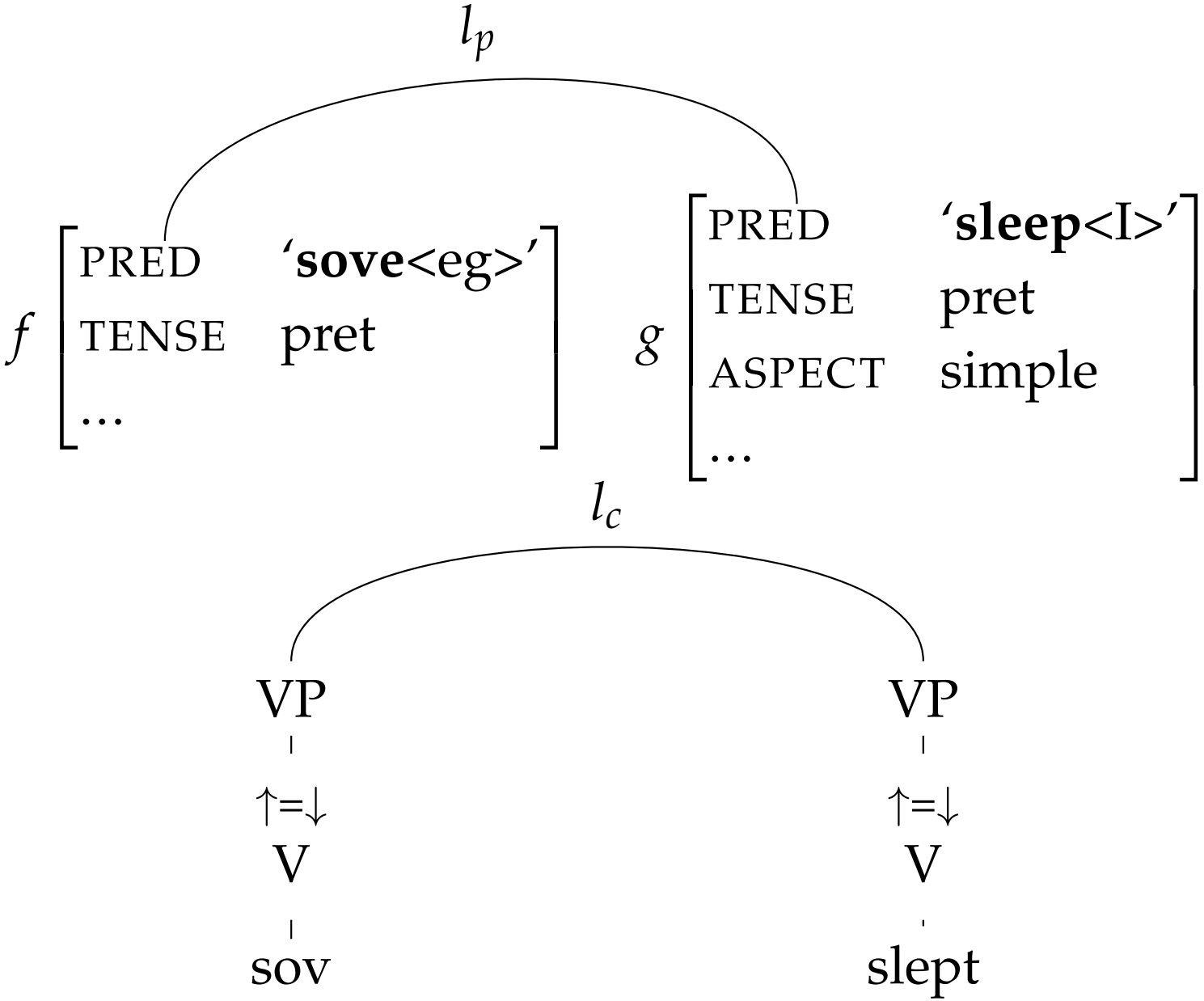
$$f\begin{bmatrix}\text{PRED} & \text{‘}\textbf{sove}\text{<eg>’}\\ \text{TENSE} & \text{pret}\\ \ldots & \end{bmatrix} \qquad g\begin{bmatrix}\text{PRED} & \text{‘}\textbf{sleep}\text{<I>’}\\ \text{TENSE} & \text{pret}\\ \text{ASPECT} & \text{simple}\\ \ldots & \end{bmatrix}$$

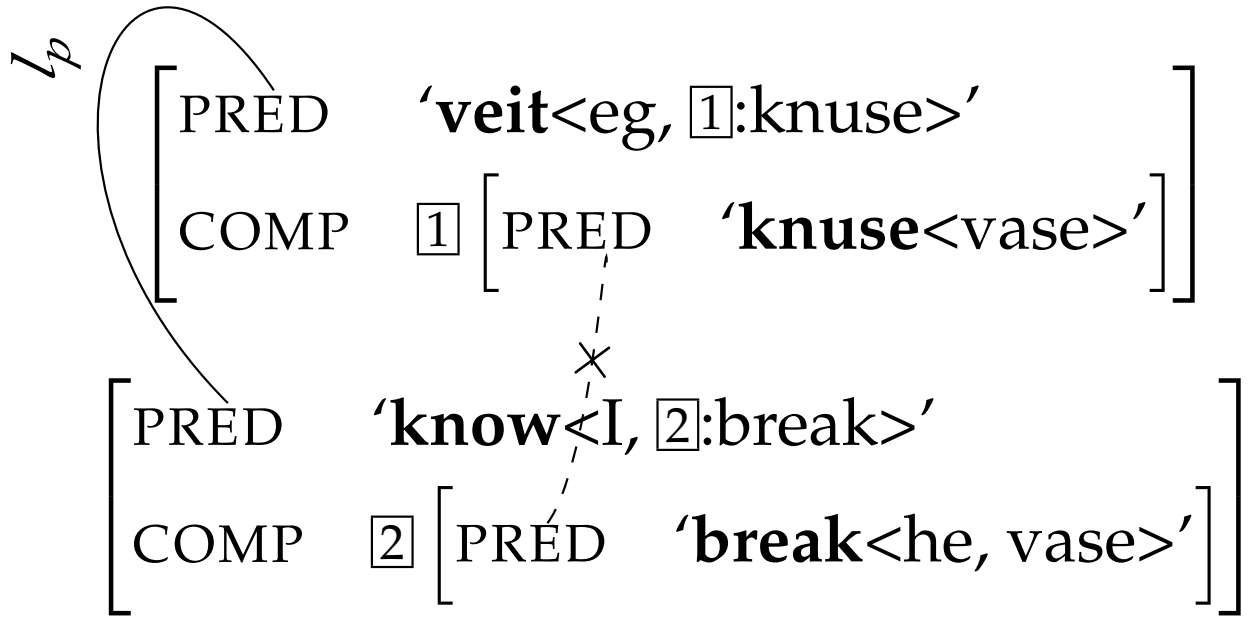





l_p
f
PRED ‘**sove**<eg>’
TENSE pret
…
g
PRED ‘**sleep**<I>’
TENSE pret
ASPECT simple
…


VP
↑=↓
V
sov
VP
↑=↓
V
slept


l_p
f
PRED ‘**sove**<eg>’
TENSE pret
…
g
PRED ‘**sleep**<I>’
TENSE pret
ASPECT simple
…
l_c
VP
↑=↓
V
sov
VP
↑=↓
V
slept



For å lenkje p til q må alle argument av p ha omsetjingar i f-strukturen av q, og omvendt

(1) a. der Tonfall gefällt mir nicht

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{gefallen}\text{<Tonfall, ich}_i\text{>'} \ldots \end{bmatrix}$$

b. jeg liker ikke tonen

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{like}\text{<jeg}_i\text{, tonen>'} \ldots \end{bmatrix}$$

(2) a. Adams veddet en sigarett med Browne på at det regnet.

$$\begin{bmatrix} \text{PRED} & \text{‘}\textbf{vedde}\text{<Abrams, sigarett, Browne, regne>’} \\ \text{ADJUNCT} & \{\} \end{bmatrix}$$

b. abramsi brouns daenajleva sigaretze, rom cvimda.

$$\begin{bmatrix} \text{PRED} & \text{‘}\textbf{da-najleveba}\text{<Abrams, Browne, cvima>’} \\ \text{ADJUNCT} & \left\{ \text{sigareti} \right\} \end{bmatrix}$$

«ORDOMSETJING»

- LPT = Linguistically Predictable Translations
- Her: bottom-up-informasjon (omsetjingsordbøker, 1-gramtabell, ...)

- kaffi $=_{LPT}$ coffee
- kaffi $\neq_{LPT}$ tea
- $\text{han}_i =_{LPT} \text{Joe}_i$
- kaffi $\neq_{LPT}$ Bob

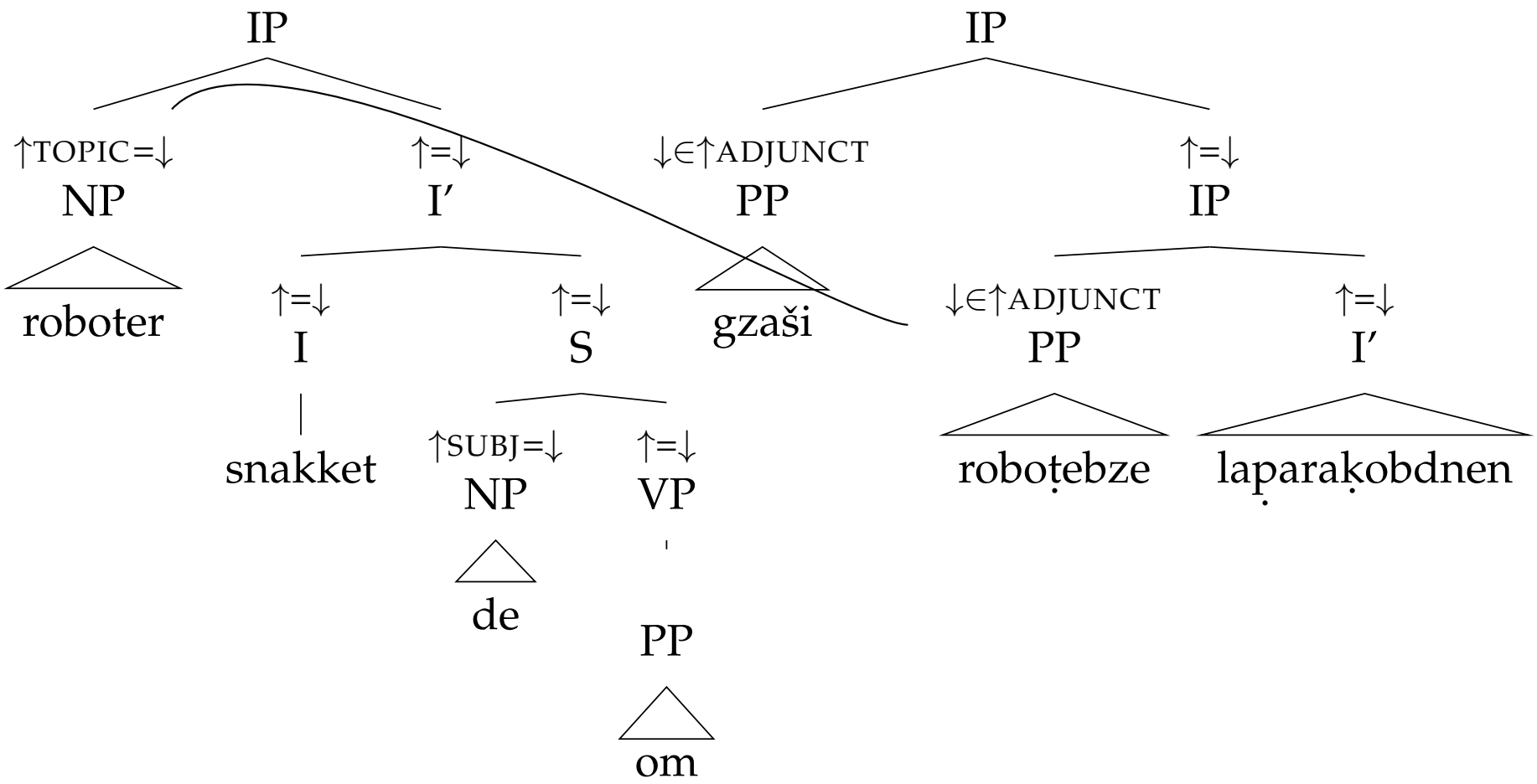
- (kaffi $=_{LPT}$ Joe)

RANGERING

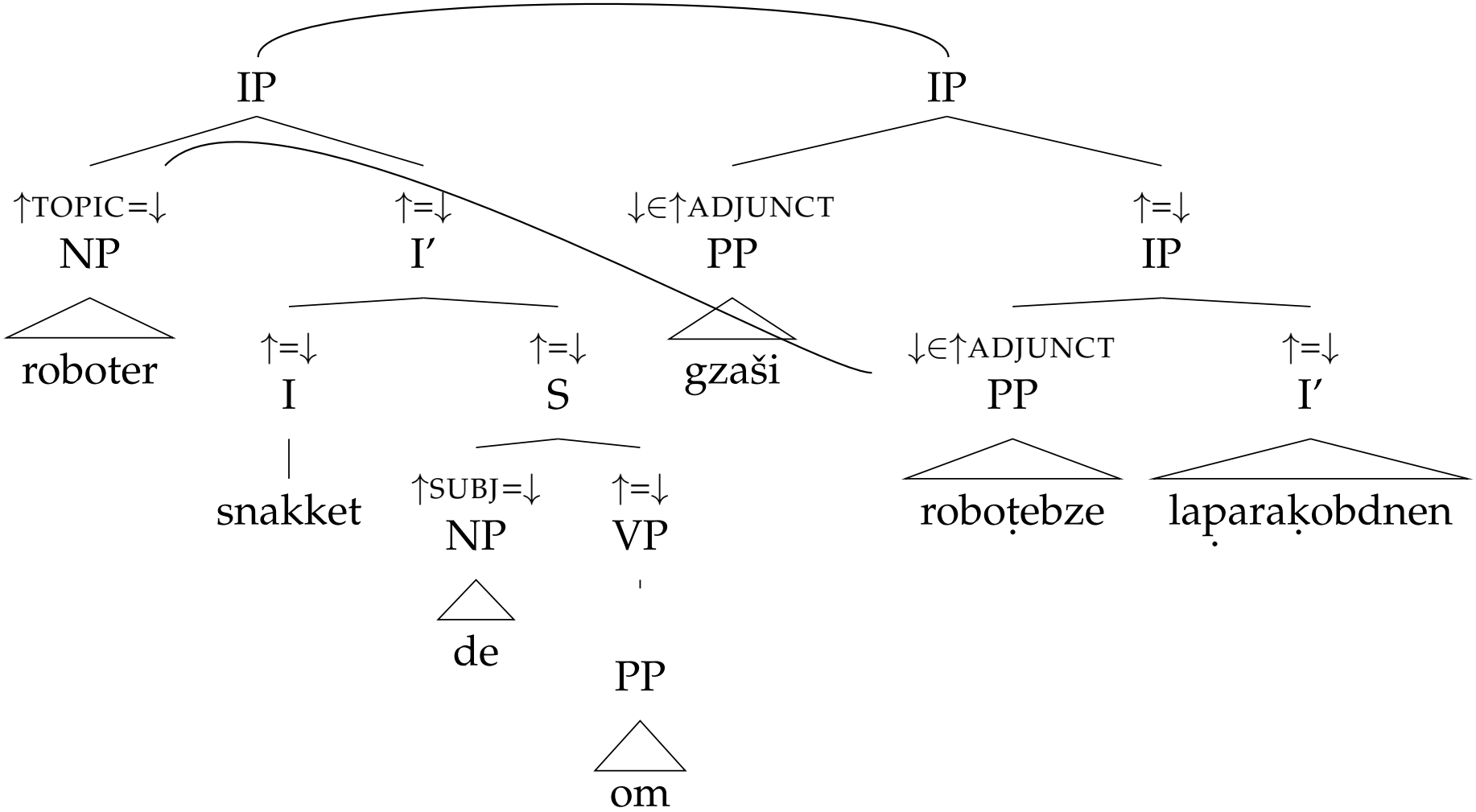
- LPT-informasjon
- Lik følgje i argumentstruktur
- Djupaste lenking:

$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{veit}\text{<eg,}\boxed{1}\text{:knuse>'} \\ \text{COMP} & \boxed{1}\begin{bmatrix} \text{PRED} & \text{`}\textbf{knuse}\text{<han,vase>'} \end{bmatrix} \end{bmatrix}$$

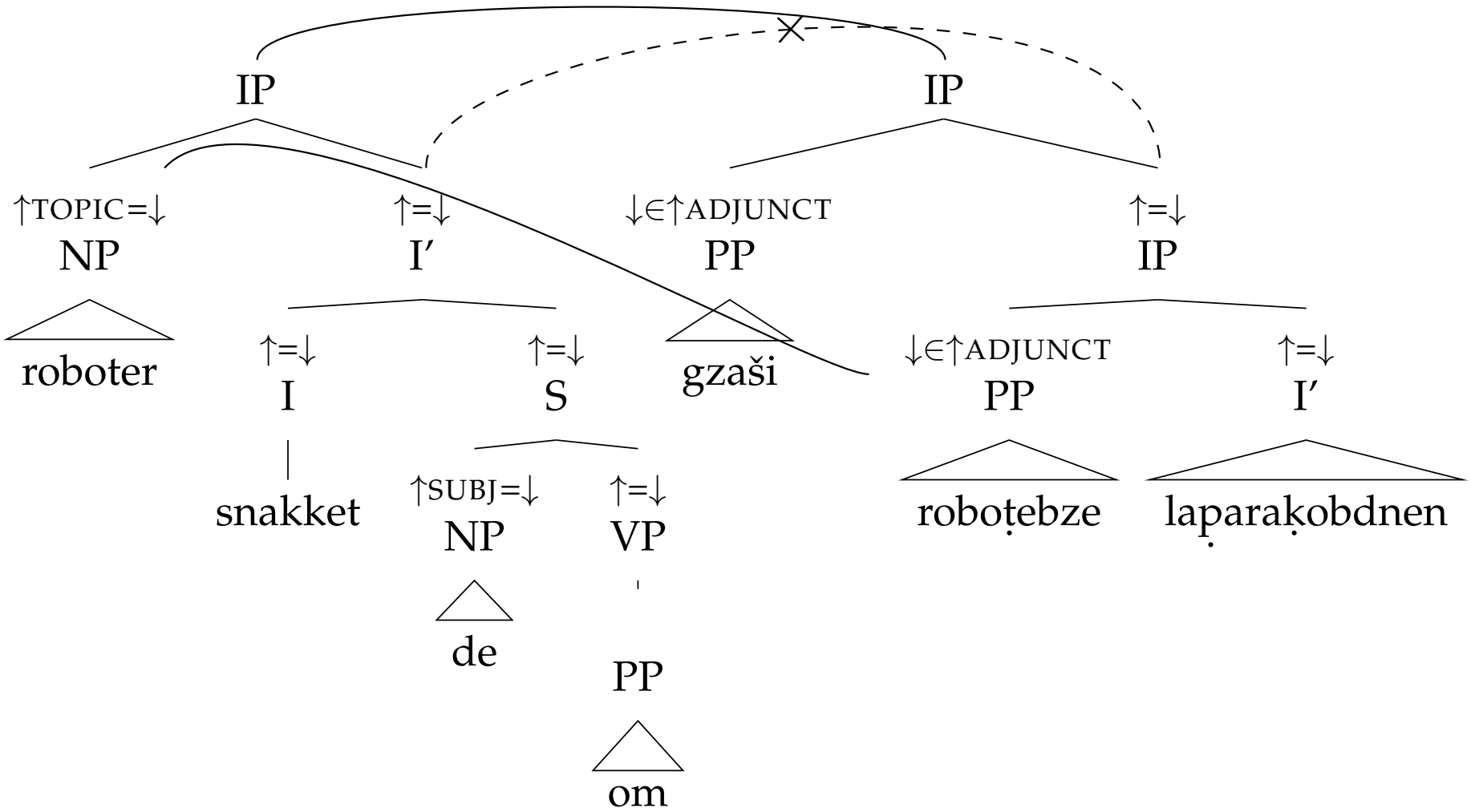
$$\begin{bmatrix} \text{PRED} & \text{`}\textbf{know}\text{<I,}\boxed{2}\text{:break>'} \\ \text{COMP} & \boxed{2}\begin{bmatrix} \text{PRED} & \text{`}\textbf{break}\text{<he,vase>'} \end{bmatrix} \end{bmatrix}$$



(3)a. **gza-ši roboṭeb-ze laṗaraḳobdnen**
veg.DAT-til robotar.DAT-på snakka.3PL
'På vegen hadde dei snakka om robotar'



(3)a. **gza-ši roboṭeb-ze laṗaraḳobdnen**
veg.DAT-til robotar.DAT-på snakka.3PL
'På vegen hadde dei snakka om robotar'



(3)a. **gza-ši** **roboṭeb-ze** **laṗaraḳobdnen**
veg.DAT-til robotar.DAT-på snakka.3PL
'På vegen hadde dei snakka om robotar'

- Kanskje eit problem å generalisere til grammatikkar utanfor Xpar-prosjektet...
- Ikkje alle predikat finst i argument/adjunkt av rotpredikatet
- Fragmentariske analysar sjølvsagt vanskeleg
  - Forventa inndata er manuelt disambiguerte analysar

$$\begin{bmatrix} \text{PRED} & \text{‘}\mathbf{perf}\langle\boxed{1}\text{:snakke*om}\rangle\boxed{2}\text{:de’} \\ \text{XCOMP} & \boxed{1}\begin{bmatrix}\text{PRED} & \text{‘}\mathbf{snakke*om}\langle\boxed{2}\mathbf{:de, robot}\rangle\text{’}\end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{PRED} & \text{‘}\mathbf{laparaki}\langle\text{pro}\rangle\text{’} \\ \text{ADJUNCT} & \left\{ \begin{bmatrix}\text{PRED} & \text{‘}\mathbf{gza}\text{’}\end{bmatrix}, \begin{bmatrix}\text{PRED} & \text{‘}\mathbf{roboti}\text{’}\end{bmatrix} \right\} \end{bmatrix}$$

(4)a. **de hadde snakket om roboter**

b. **gza-ši roboteb-ze laparakobdnen**
veg.DAT-til robotar.DAT-på snakka.3PL
‘På vegen hadde dei snakka om robotar’

$$\begin{bmatrix} \text{PRED} & \text{`}\mathbf{la}\text{<ho,skrive>}\boxed{1}\text{:eg'} \\ \text{XCOMP} & \boxed{1}\begin{bmatrix} \text{PRED} & \text{`}\mathbf{skrive}\text{<}\boxed{1}\mathbf{:eg,brev}\text{>'} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{PRED} & \text{`}\mathbf{daacerineb}\text{<pro,pro,cerili>'} \end{bmatrix}$$

Takk for merksemda!

LITTERATUR

Dyvik, H., Meurer, P., Rosén, V. & Smedt, K.D. (2009). Linguistically Motivated Parallel Parsebanks. I M. Passarotti, A. Przepiórkowski, S. Raynaud & F.V. Eynde (red.), *Proceedings of TLT8* (s. 71–82). Milano: EDUCatt. Tilgjengeleg frå `http://tlt8.unicatt.it/allegati/Proceedings_TLT8.pdf#page=83`

Graham, Y. & Genabith, J. van. (2009). An Open Source Rule Induction Tool for Transfer-Based SMT. *The Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for Machine Translation*, *91*, 37–46.

Hearne, M., Ozdowska, S. & Tinsley, J. (2008). Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. I *TALN '08.* Avignon, France. Tilgjengeleg frå `http://www.computing.dcu.ie/~mhearne/publications.html`

Samuelsson, Y. & Volk, M. (2007). Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment. I *Proceedings of TLT7.* Bergen, Norway.

- Program/kjeldekode er tilgjengeleg frå https://github.com/unhammer/lfgalign under GNU GPL
- Desse lysarka kan distribuerast under lisensane GNU GPL, GNU FDL og CC-BY-SA.
  - GNU GPL v. 3.0 http://www.gnu.org/licenses/gpl.html
  - GNU FDL v. 1.2 http://www.gnu.org/licenses/gfdl.html
  - CC-BY-SA v. 3.0 http://creativecommons.org/licenses/by-sa/3.0/