

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

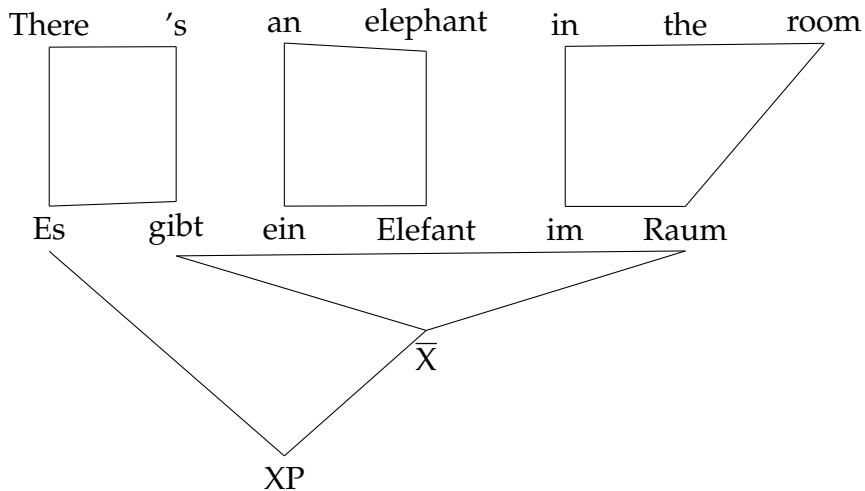
Generalisering til nye grammatikkar

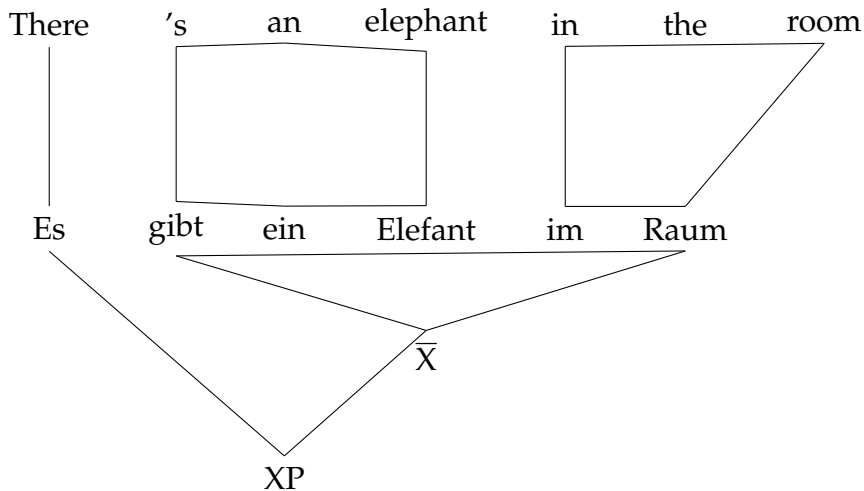
Når predikat ikkje korresponderer 1-til-1

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

im Raum





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)

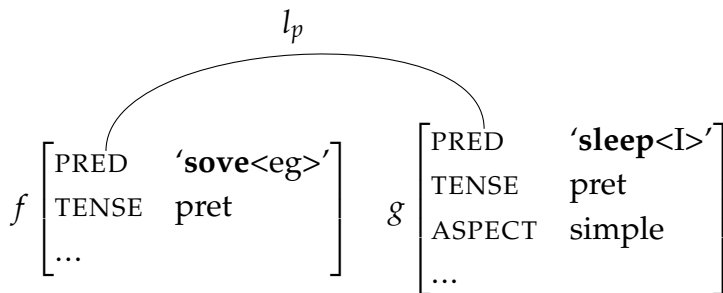
CAT SYNTAKS | FRASELENKING | N-GRAMTABELL

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

$$f \left[\begin{array}{ll} \text{PRED} & \text{'sove<eg>'} \\ \text{TENSE} & \text{pret} \\ \dots & \end{array} \right] \quad g \left[\begin{array}{ll} \text{PRED} & \text{'sleep<I>'} \\ \text{TENSE} & \text{pret} \\ \text{ASPECT} & \text{simple} \\ \dots & \end{array} \right]$$

VP
|
↑=↓
V
|
SOV

VP
|
↑=↓
V
|
slept



VP

|

 $\uparrow=\downarrow$

V

|

SOV

VP

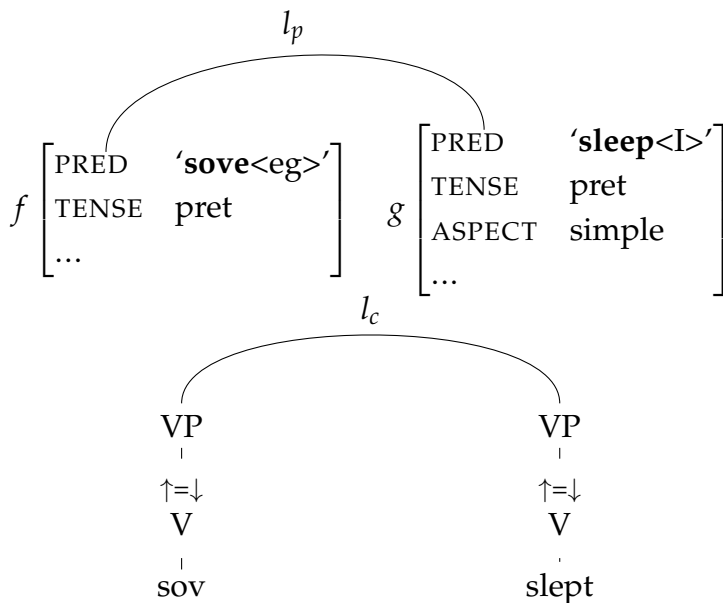
|

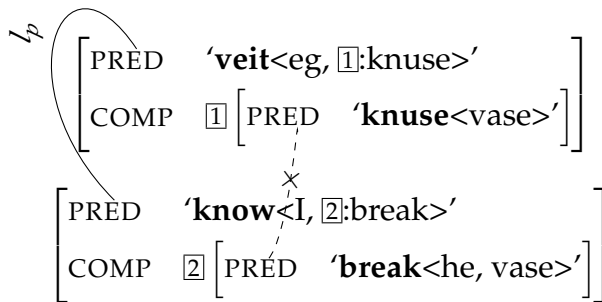
 $\uparrow=\downarrow$

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

[PRED 'gefallen<Tonfall, ich_i>' ...]

b. jeg liker ikke tonen

[PRED 'like<jeg_i, tonen>' ...]

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

| | |
|---------|---|
| PRED | ' vedde <Abrams, sigarett, Browne, regne>' |
| ADJUNCT | { } |

- b. abramsi brouns daenajleva sigaretze, rom cvimda.

| | |
|---------|--|
| PRED | ' da-najleveba <Abrams, Browne, cvima>' |
| ADJUNCT | { sigareti } |

«ORDOMSETJING»

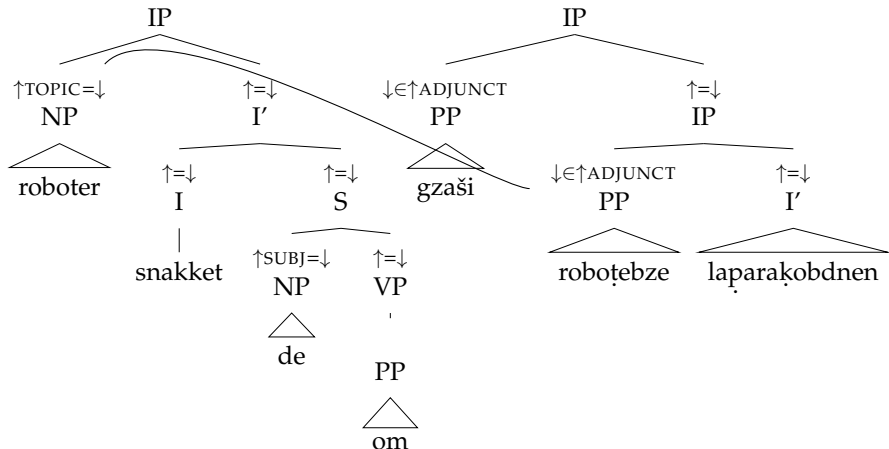
- ▶ LPT = Linguistically Predictable Translations
- ▶ Her:
bottom-up-informasjon
(omsetjingsordbøker,
1-gramtabell, ...)
- ▶ kaffi =_{LPT} coffee
- ▶ kaffi ≠_{LPT} tea
- ▶ han_i =_{LPT} Joe_i
- ▶ kaffi ≠_{LPT} Bob
- ▶ (kaffi =_{LPT} Joe)

RANGERING

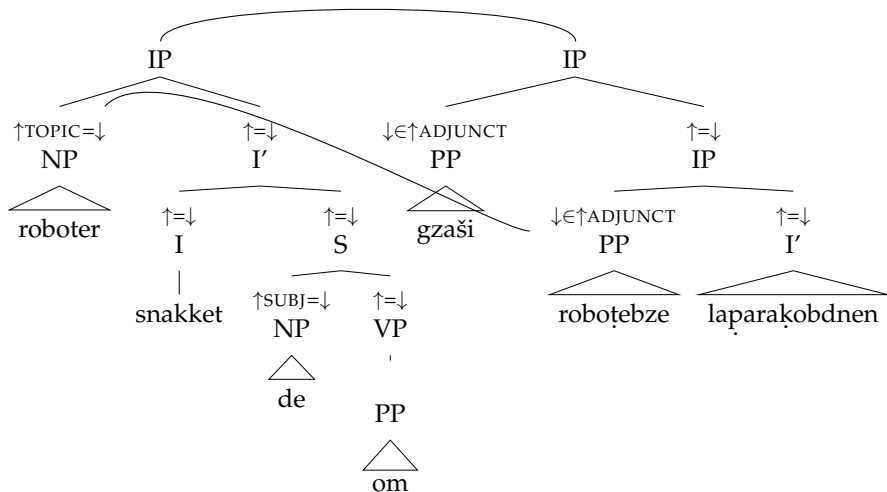
- ▶ LPT-informasjon
- ▶ Lik følge i argumentstruktur
- ▶ Djupaste lenking:

$$\left[\begin{array}{ll} \text{PRED} & \text{'veit<eg, } \boxed{1} \text{:knuse>'} \\ \text{COMP} & \boxed{1} \left[\text{PRED} \quad \text{'knuse<han, vase>'} \right] \end{array} \right]$$

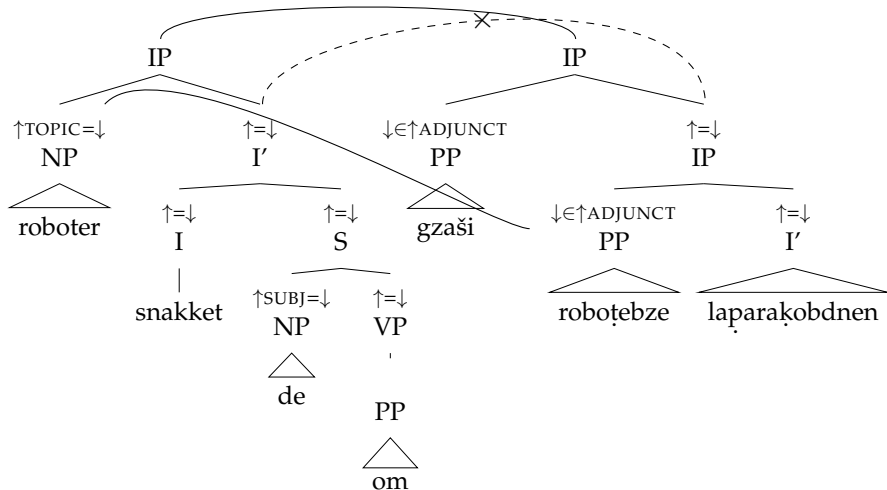
$$\left[\begin{array}{ll} \text{PRED} & \text{'know<I, } \boxed{2} \text{:break>'} \\ \text{COMP} & \boxed{2} \left[\text{PRED} \quad \text{'break<he, vase>'} \right] \end{array} \right]$$



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



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



- (3)a. **gza-ši roboteb-ze laparakobdnen**
 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

$$\left[\begin{array}{ll} \text{PRED} & \text{'perf}<\boxed{1}:\text{snakke*om}>\boxed{2}:\text{de'} \\ \text{XCOMP} & \boxed{1} \left[\text{PRED} \quad \text{'snakke*om}<\boxed{2}:\text{de, robot}>' \right] \end{array} \right]$$

$$\left[\begin{array}{ll} \text{PRED} & \text{'lapara}\dot{\text{x}}i<\text{pro}>' \\ \text{ADJUNCT} & \left\{ \begin{array}{l} \left[\text{PRED} \quad \text{'gza'} \right], \\ \left[\text{PRED} \quad \text{'robo}\dot{\text{t}}i' \right] \end{array} \right\} \end{array} \right]$$

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

b. **gza-ši roboṭeb-ze laparaḵobdnen**
 veg.DAT-til robotar.DAT-på snakka.3PL
 'På vegen hadde dei snakka om robotar'

$$\left[\begin{array}{ll} \text{PRED} & \text{'la<ho,skrive>[1]:eg'} \\ \text{XCOMP} & [1] \left[\text{PRED} \quad \text{'skrive<[1]:eg,brev>'} \right] \end{array} \right]$$

$$\left[\text{PRED} \quad \text{'daacerineb<pro,pro,cerili>'} \right]$$

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.

LISENSAR

Denne presentasjonen 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/>