GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Automatisk identifiering av konstruktionskandidater för ett svenskt konstruktikon

Markus Forsberg

Språkbanken
Göteborgs universitet

2013-03-19

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Föredraget

- Föredraget är baserat på en artikel inskickad igår till *NoDaLiDa 2013 workshop: Lexical semantic resources for NLP*.
- *Automatic identification of construction candidates for a Swedish constructicon*
  Linnea Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice, and Emma Sköldberg
- Detta föredrag kommer handla om partiellt schematiska konstruktioner, metodutveckling och språkteknologi.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Ett svenskt konstruktikon

- *Ett svenskt konstruktikon* är ett projekt med fokus på svenska konstruktioner som samlas in, analyseras, beskrivs, och görs allmänt tillgängliga via Språkbanken.
- Projektet är ett samarbete mellan språkteknologer, grammatiker, lexikografer, fraseologer, semantikforskare och L2-forskare.
- Utvecklingsversionen är tillgänglig här: http://spraakbanken.gu.se/resurs/konstruktikon

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Partiellt schematiska konstruktioner

- Detta föredrags fokus är *partiellt schematiska konstruktioner* = flerordsuttryck med både fasta och variabla led.
- Exempel: *X och X* (*bra och bra*)
- Exempel (?): *RG kronor* (*42 kronor*)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Automatisk identifiering av konstruktioner

- Det språkteknologiska arbetet i projektet fokuserar på identifiering av konstruktioner i dubbel bemärkelse:
  - att upptäcka partiellt schematiska konstruktioner i stora mängder text;
  - att lokalisera konstruktikonets konstruktionerna i text.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Automatisk identifiering av flerordsuttryck

- Att automatiskt identifiera flerordsuttryck med språkteknologiska metoder är komplicerat.
- Anledningen till detta är att det är svårt att skilja flerordsuttryck från kollokationer.
- Att identifiera partiellt schematiska konstruktioner ärver detta problem (och lägger till nya).

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Språkbankens forskningsinfrastruktur

- Arbetet har Språkbankens forskningsinfrastruktur som startpunkt.
- Korp – en korpusinfrastruktur
- Karp – en lexikal infrastruktur
- Tillsammans utgör de en plattform som möjliggör att språkteknologiska experiment kan snabbt kan gå från idé till faktiska resultat.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Materialval: SUC 2.0

- Materialval: SUC 2.0, en balanserad korpus för svenska som är manuellt annoterad med grundform och morfosyntaktisk beskrivning.
- Det är ett relativt litet material: 1,17 miljoner token.

word	msd	lemma
Hur	HA	hur
är	VB. PRS. AKT	vara
det	PN. NEU. SIN. DEF. SUB+OBJ	den
då	AB	då
i	PP	i
Mellanöstern	PM. NOM	Mellanöstern
?	MAD	

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Hybrida n-gram

- Experimentet bygger på arbetet med StringNet.
- Ett grundläggande koncept i StringNet är hybrida n-gram.
- Hybrida n-gram är en generalisering av n-gram där vi även inkluderar informationen i annotationslagren.
- Här tar vi bara med grundformer och ordklasser, men många andra alternativ är möjliga.
- *Hur är* $\Rightarrow$ *hur vara*, *hur VB*, *HA vara* och *HA VB*.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Generering av hybrida n-gram

- 2-, 3-, 4-gram
- Med tanke på att vi vill fånga partiellt schematiska konstruktioner, så filtrerar vi bort de helt schematiska (*HA VB*) och helt lexikala (*hur vara*).
- Vi filtrerar också bort hybrida n-gram som innehåller följande ordklasser: MID, MAD, PAD och UO.
- Resultat: 16 miljoner hybrida n-gram varav 8.8 miljoner är unika.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Rangordning

- Nästa steg är att rangordna alla hybrida n-gram.
- Detta kan göras med ett samförekomstmått.
- PMI – point-wise mutual information
- PMI har ett känt problem: måttet föredrar lågfrekventa samförekomster.
- En känd lösning är att multiplicera med den absoluta frekvensen.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Rangordning

- Ännu ett problem: klippa-klistra text.
- UIF: unik instansfrekvens

$$\text{PMI-UIF(H)} = \mathit{UIF} * \mathit{log}_2\left(\frac{P(H)}{\prod_{x \in H} P(x)}\right)$$

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Redundansproblemet

- Problem: större delen av de hybrida n-grammen är delsekvenser av andra hybrida n-gram, vilket ger ett resultat med stor redundans.
- Lösning: rensa bort alla hybrida n-gram som är delsekvenser av andra n-gram med högre PMI-UIF.
- Ett något luftigare identitetsbegrepp: $att_{IE} = IE$.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Kandidatlistan

vara$_{VB}$ ute$_{AB}$ och$_{KN}$ VB	*är ute och letar (3)*	15	0.93	52.24
vara$_{VB}$ JJ för$_{PP}$ att$_{IE}$	*är viktiga för att (2)*	26	1.61	52.83
stänga$_{VB}$ av$_{PL}$ NN	*stängt av motorn (1)*	11	0.68	52.25

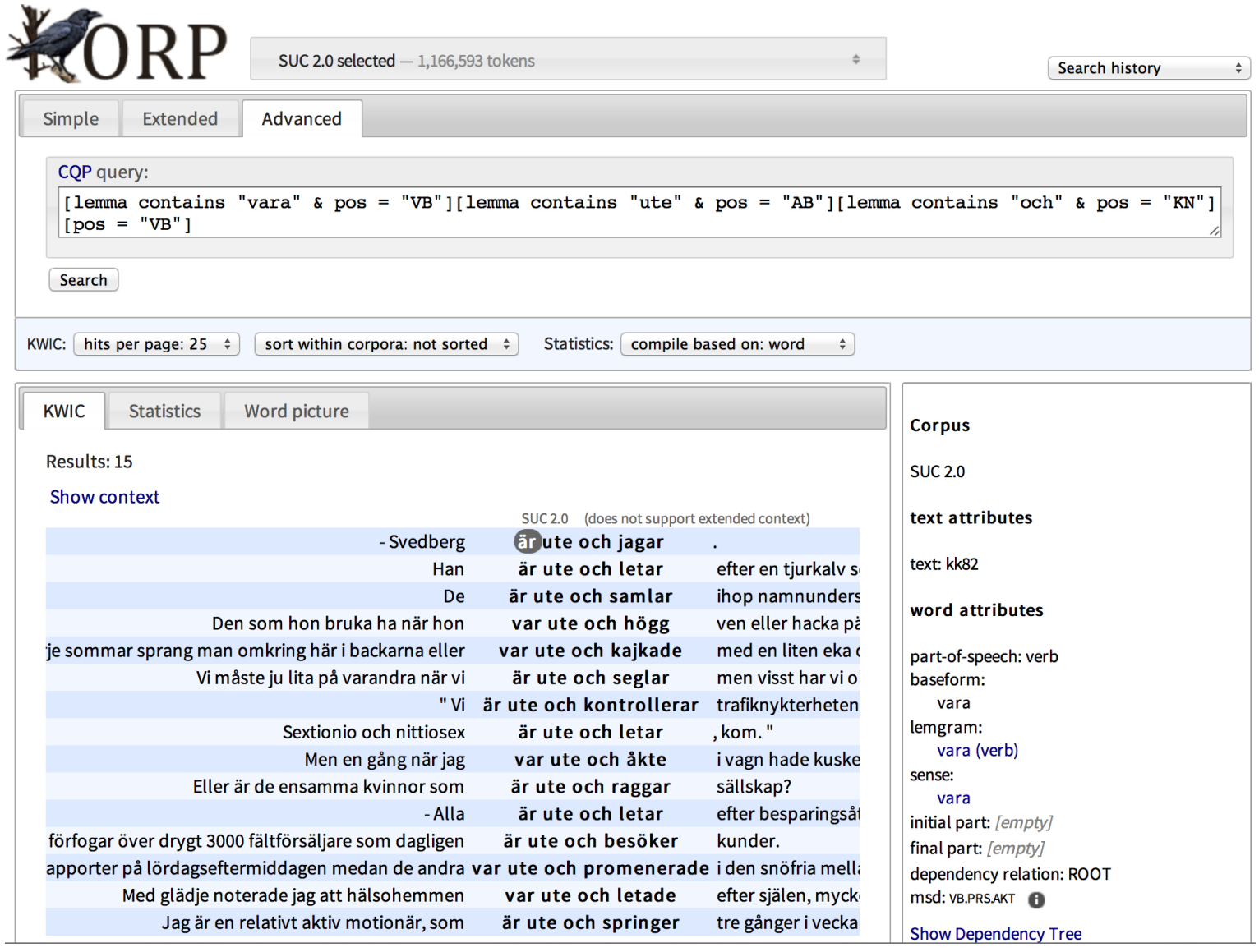
- Top-2500
- <http://http://spraakbanken.gu.se/swe/resurs/konstruktikon/kandidater>
- Det finns även andra kandidatlistor (precis nu bara en) som är baserade på automatiskt annoterat material (via Korps importkedja).

Korpinstanser

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT



GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Analys av konstruktionskandidaterna

- Av 2500 kandidater godtogs 50 som fullvärdiga konstruktioner.
- En indikation på att metoden fångar vad vi vill är att den också fångar konstruktioner som redan finns med i det svenska konstruktikonet.
- Exempel:

$$RG\ NN\ per_{PP}\ NN$$

- *en gång per dygn, 500 kronor per månad*

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Analys av konstruktionskandidaterna

- Ett annat exempel:
  den_{DT} *RO NN*
- *den 1 juli* och *den tionde mars*
- jämför med *PP NN*:
  i mars, *på morgonen* och *på eftermiddagen*.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Analys av konstruktionskandidaterna

- Exempel på konstruktioner som inte finns med i SweCxn:
- *RG $\text{år}_{NN_{GEN}}$ ålder_{NN}*
  - *sju års ålder*
- *varken_{KN} NN eller_{KN} NN*
  - *varken uppehållstillstånd eller arbetstillstånd*
- *vara_{VB} sig_{PN} NN eller_{KN} (NN)*
  - *vare sig fotboll eller ishockey*
- *vara_{VB} sig_{PN} PN VB (eller_{KN} inte_{AB})*
  - *vare sig vi vill eller inte*

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Analys av konstruktionskandidaterna

- *vara*$_{VB}$ *ute*$_{AB}$ *och*$_{KN}$ *VB*
  - *vara ute och jaga*
- Men även metaforiskt:
  - *vara ute och cykla*
  - *vara ute och segla*
  - *vara ute och snurra*

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Felanalys

- Syntaktiska djur:
  - *den*$_{DT}$ *JJ NN VB* (*de nordiska länderna är*)
  - *SN PN VB en*$_{DT}$ (*att det var en*)
- Fragmentariska djur:
  - *vara*$_{VB}$ *sig*$_{PN}$ *NN eller*$_{KN}$ *(NN)*
- lexikala djur:
  - *i*$_{PP}$ *all*$_{DT}$ *fall*$_{NN}$ *VB*
  - *över*$_{PP}$ *huvud*$_{NN}$ *tagen*$_{PC}$ *VB*

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Resultat

- Ur en metodologisk synvinkel har experimentet gett bra resultat, eftersom vi kunnat hitta tidigare obeskrivna partiellt schematiska konstruktioner.
- Precisionen är inte imponerande (typ 2%), men som ett tryck-på-en-knapp-verktyg för att upptäcka nya konstruktioner har det visat sig vara ett värdefullt bidrag till arbetet med ett svenskt konstruktikon.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Problem och möjliga lösningar

- Problem:
  - lexikala djur
  - syntaktiska djur
  - fragmentariska (djur)
- Framtida arbete:
  - att utnyttja Språkbankens lexikala resurser i kombination med Korpimportens syntaktiska analys för att försöka förbättra precisionen.
  - att undersöka hur vi kan göra kandidaterna mer flexibla för att fånga de större enheterna.

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Demonstration?

- Tid över för en snabb demonstration?
- http://spraakbanken.gu.se/swe/resurs/konstruktikon/kandidater