Kollokationer

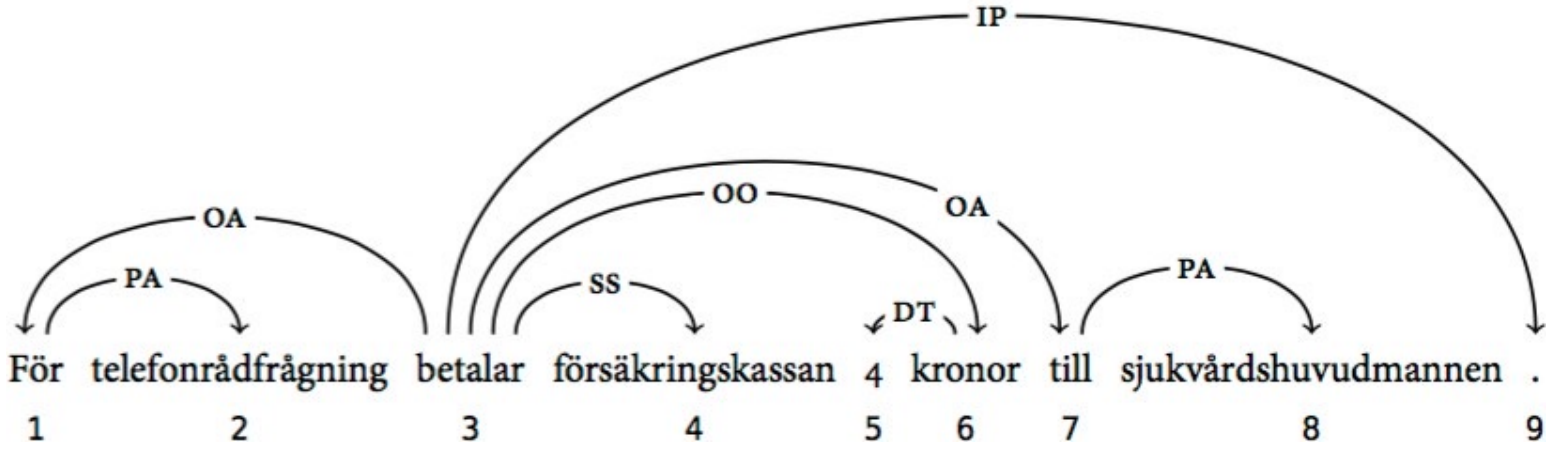
Inför laboration 4

729G49 Språk och Datorer

(baserat på Marco Kuhlmanns bilder från 2020)

LINKÖPINGS UNIVERSITET

Dependensträd



huvudord ⟶ dependent

Dependensträd

1	För	PP	3	OA
2	telefonrådfrågning	NN	1	PA
3	betalar	VB	0	ROOT
4	försäkringskassan	NN	3	SS
5	4	RG	6	DT
6	kronor	NN	3	OO
7	till	PP	3	OA
8	sjukvårdshuvudmannen	NN	7	PA
9	.	MAD	3	IP

Kollokationer

Kollokationer

En kollokation är en fast fras—en sekvens av två eller fler ord som ofta förekommer tillsammans

- sparka _____ ?
- salt och _____?

Andra termer: fasta fraser, fasta uttryck, lexikaliserade fraser, idiom, …

”As mature speakers of a language, we all know which words tend to occur with other words.” (Yule, sid. 139)

Kollokationer

- Vissa kollokationer är oföränderliga:

  trots allt, ont i halsen,…

- Andra kan varieras mer eller mindre:

  sparka boll, ren lögn, salt och peppar,…

Typ	Exempel
adjektiv + substantiv	*nästa vecka*
substantiv + substantiv	*års fängelse*
verb + substantiv	*äga rum*
adverb + adjektiv	*mycket bra*
verb + adverb	*ska inte*
verb + partikel	*ställa upp*
verb + preposition	*berätta om*

En text

Som bekant uppfördes Eiffeltornet till världsutställningen Exposition Universelle i Paris 1889, vilken påpassligt öppnade portarna ett sekel jämt efter den franska revolutionen. Jag vet inte riktigt när det där med världsutställningar kom ur mode. Eller om det över huvud taget fortfarande finns. Har ni hört talas om någon på sista tiden?

Men då, förr alltså, var världsutställningar rejäla kioskvältare. Ett slags OS i teknologiskt och arkitektoniskt skryt.

Kollokationers status i språket

- Kollokationer hör till ett språks lexikon på samma sätt som enstaka ord gör
- Kollokationer bidrar till att språket flyter och blir idiomatiskt
- Kollokationer följer inga regler men modersmålstalare upptäcker genast när man kombinerar ord på ”fel” sätt

Källa: Fasta fraser i svenskan

Kollokationers status i språket

Kunskap om kollokationer och de medföljande lexikala, morfologiska och syntaktiska restriktionerna är viktig för andraspråksinlärning.

- Lina och Per har ont i halsen – *Lina och Per har ont i sina halsar

En andraspråksinlärare måste vanligtvis lära sig fraserna som lexikala helheter, vilket ofta är mödosamt.

- allvarligt skadad – seriously injured, schwer verletzt

Källa: Fasta fraser i svenskan

Svenska	Engelska
Avsluta en affär	c___ a deal
Be en bön	s___ a prayer
Betala räkningar	f___ the bill
Fatta eld	c___ fire
Göra ett prov	s___ a test
Rasta hunden	w___ the dog
Väcka känslor	s____ emotions

Källa: Snoder (2019), Appendix A

Svenska	Engelska
Avsluta en affär	close a deal
Be en bön	say a prayer
Betala räkningar	foot the bill
Fatta eld	catch fire
Göra ett prov	sit a test
Rasta hunden	walk the dog
Väcka känslor	stir up emotions

Källa: Snoder (2019), Appendix A

Hur kan vi automatiskt hitta kollokationer i text?

Hur kan vi automatiskt hitta kollokationer i text?

Att bara välja ut de ordpar som har högst antalet förekomster leder till ointressanta kollokationer.

- *en boll* är mycket vanligare än *sparka boll*

Vi skulle vilja ha ett mått som fokuserar på sammanhang i vilket två ord samförekommer oftare än ”väntat”.

Tärningskast

- Vi kastar en vanlig sexsidig tärning. Vad är sannolikheten för händelsen ”jämnt antal prickar”?

- Vi kastar tärningen två gånger. Vad är sannolikheten för händelsen ”jämnt antal prickar vid första kastet *och* jämnt antal prickar vid andra kastet”?

Oberoende händelser

Två händelser A och B kallas **oberoende** om och endast om sannolikheten för att båda ska inträffa är produkten av deras enskilda sannolikheterna, dvs. om

$$P(A \text{ och } B) = P(A) \cdot P(B)$$

Pointwise Mutual Information

$P(x)$ sannolikheten för att ordet x förekommer i korpusen

$P(y)$ sannolikheten för att ordet y förekommer i korpusen

$P(xy)$ sannolikheten för att orden x och y samförekommer

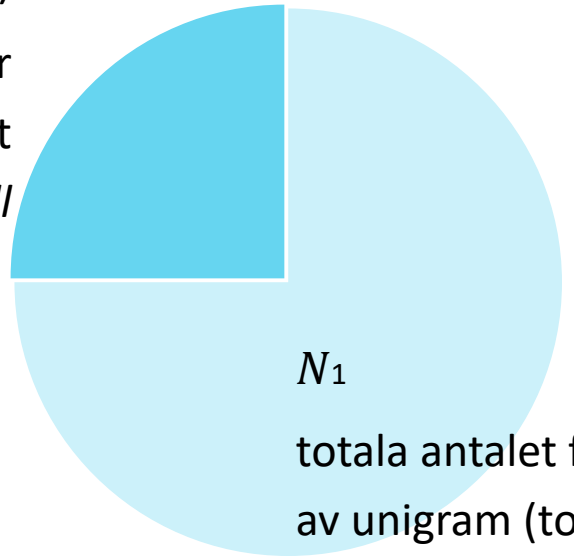
$$PMI(x, y) = \log \frac{P(xy)}{P(x)P(y)}$$

Skattning av unigramsannolikheter

$P(boll)$

$\#(boll)$

antal förekomster av unigrammet *boll*



N_1

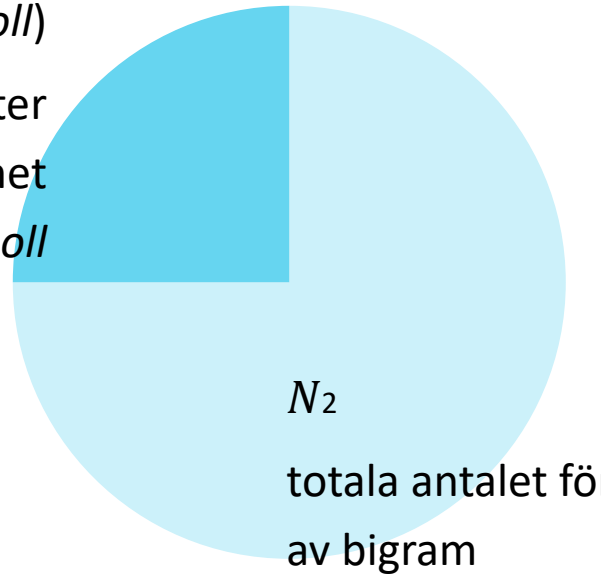
totala antalet förekomster av unigram (tokens)

Skattning av bigramsannolikheter

P(*sparka boll*)

#(*sparka boll*)

antal förekomster av bigrammet *sparka boll*



N_2

totala antalet förekomster av bigram

Pointwise mutual information

$$PMI(x, y) = \log \frac{P(xy)}{P(x)P(y)}$$

- Om x och y är oberoende gäller $P(xy) = P(x)\ P(y)$ och kvoten är lika med 1.
- Om sannolikheten för händelsen att de två orden samförekommer är *högre* än sannolikheten för ”nollhypotesen” att de är oberoende är kvoten *större* än 1.
- Om sannolikheten för att orden samförekommer är *mindre* än sannolikheten för nollhypotesen är kvoten mindre än 1.

Veckans labb

Laboration 4

- Uppskatta sannolikheter för unigram och bigram
- Använda pMI för att hitta kollokationer och bedöma resultatet
- Jämföra pMI med en variant

Tack för idag!

LINKÖPINGS UNIVERSITET