

Språk och Datorer (729G49)

Språkteknologi 4: Kollokationer

28/4/2025



Innehåll

- Repetition: Lingvistiskt uppmärkt text
- Kollokationer
- n-grammar
- Pointwise Mutual Information (PMI)

Repetition:
Lingvistiskt uppmärkt text

1	Genom	genom	PP	3	AA
2	skattereformen	skattereform	NN	1	PA
3	införs	införa	VB	0	ROOT
4	individuell	individuell	JJ	5	AT
5	beskattning	beskattning	NN	3	SS
6	(	(	PAD	5	IR
7	särbeskattning	särbeskattning	NN	5	AN
8	)	)	PAD	5	JR
9	av	av	PP	5	ET
10	arbetsinkomster	arbetsinkomst	NN	9	PA
11	.	.	MAD	3	IP

Lingvistiska Annotationer

- Graford *skattereformen*
- Lemma *skattereform*
- Ordklass NN (substantiv)
- Huvudord Ord nummer 3 i meningen (*införs*)
- Grammatisk funktion SS (subjekt)
- Morfologiska egenskaper utrum, singular, definit, nominativ
- Betydelse skattereform..nn.1

1	Genom	genom	PP	3	AA
2	skattereformen	skattereform	NN	1	PA
3	införs	införa	VB	0	ROOT
4	individuell	individuell	JJ	5	AT
5	beskattning	beskattning	NN	3	SS
6	(	(	PAD	5	IR
7	särbeskattning	särbeskattning	NN	5	AN
8	)	)	PAD	5	JR
9	av	av	PP	5	ET
10	arbetsinkomster	arbetsinkomst	NN	9	PA
11	.	.	MAD	3	IP

Kollokationen

Kollokationer

- "As mature speakers of a language, we all know which words tend to occur with other words." (Yule, sid. 139)
- Ofta kan man förutsäga vilket ord kommer näst.
 - Varför?

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, ...

Föränderliga och Oföränderliga 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	<i>nästa vecka</i>
substantiv + substantiv	<i>års fängelse</i>
verb + substantiv	<i>äga rum</i>
adverb + adjektiv	<i>mycket bra</i>
verb + adverb	<i>ska inte</i>
verb + partikel	<i>ställa upp</i>
verb + preposition	<i>berätta om</i>

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 överhuvudtaget 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.

Kollokationer som Lexikala Enheter

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

- 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/svårt skadad – seriously injured, schwer verletzt

<https://sprakbruk.fi/artiklar/fast-fraser-i-svenskan/>

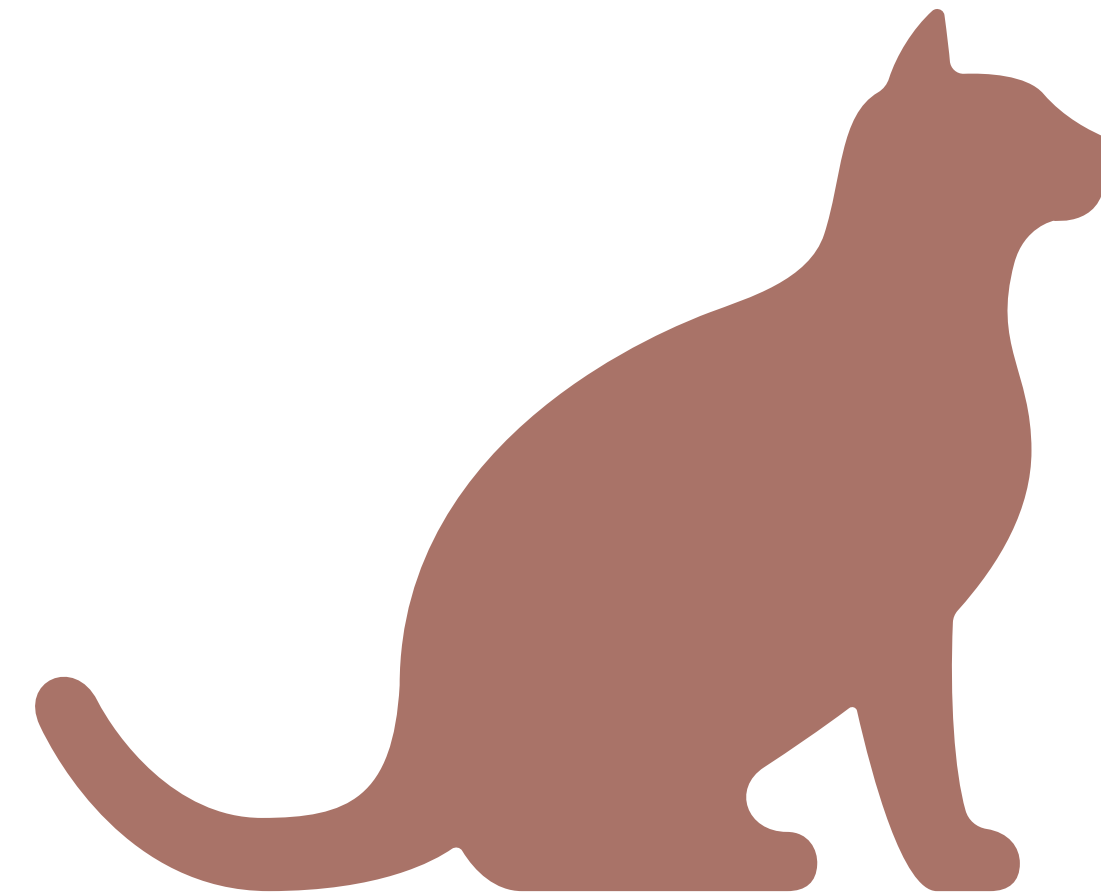
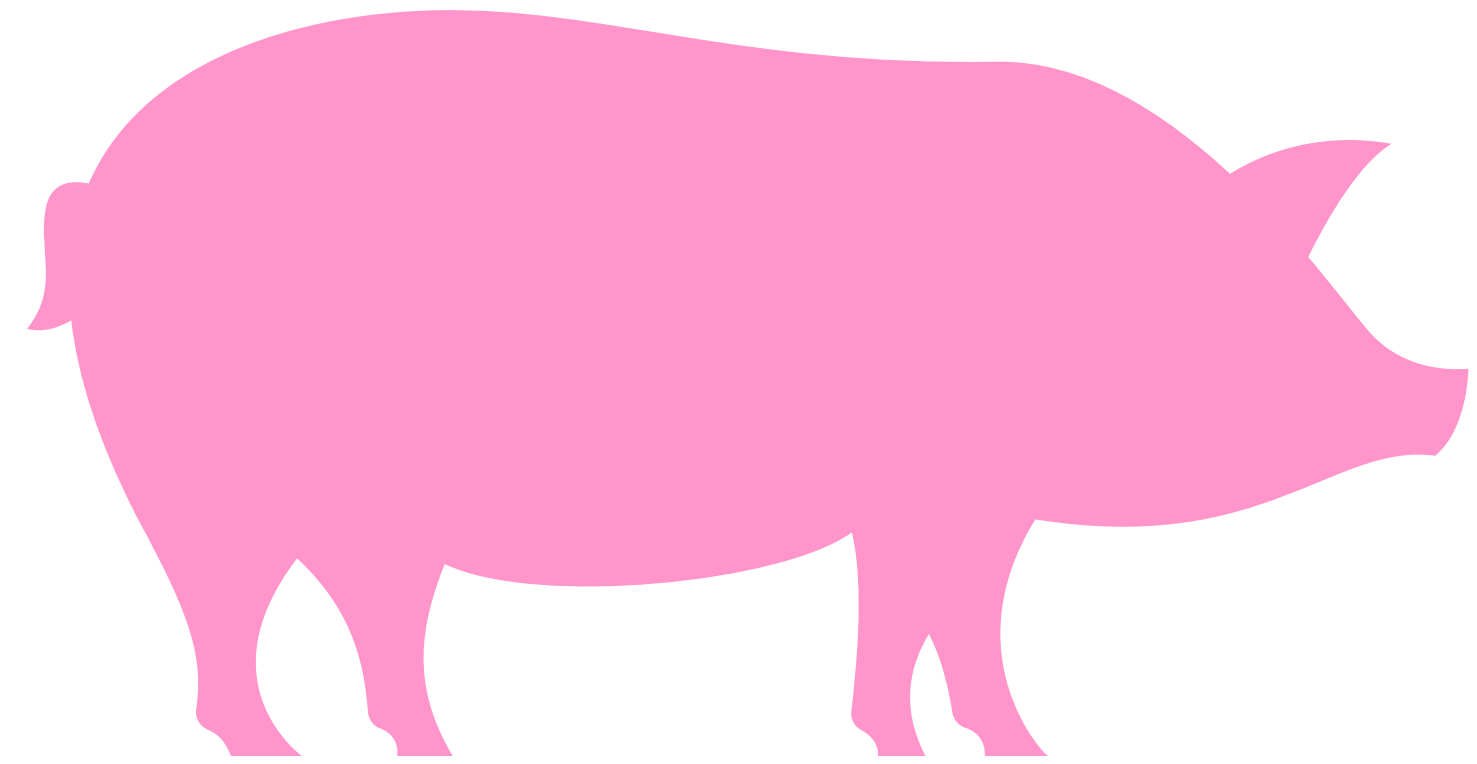
Fasta Fraser



- *borsta tänderna*
 - **putsa tänderna*
 - **tvätta tänderna* (förutom finlandssvenskan som följt av inflytande från finskans *pestä hampaansa*),
- tyska: *die Zähne putzen*, norska: *å pusse tennene*
- franska: *se brosser les dents*, engelska: *to brush your teeth*
- spanska: *lavarse los dientes* (eller *cepillarse los dientes*)
- ... <https://sprakbruk.fi/artiklar/fast-fraser-i-svenskan/>

Idiomer

Svenska:	köpa grisen i säcken
finska:	ostaa sika säkissä
engelska:	buy a pig in a poke
franska:	acheter un chat en poche,
tyska:	die Katze im Sack kaufen
nederländska:	een kat in de zak kopen



<https://sprakbruk.fi/artiklar/fast-fraser-i-svenskan/>

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

n-grammar

n-grammar

- Unigram: Sekvens av 1 token
- Bigram: Sekvens av 2 tokens
- Trigram: Sekvens av 3 tokens
- ...
- n-gram: Sekvens av n tokens

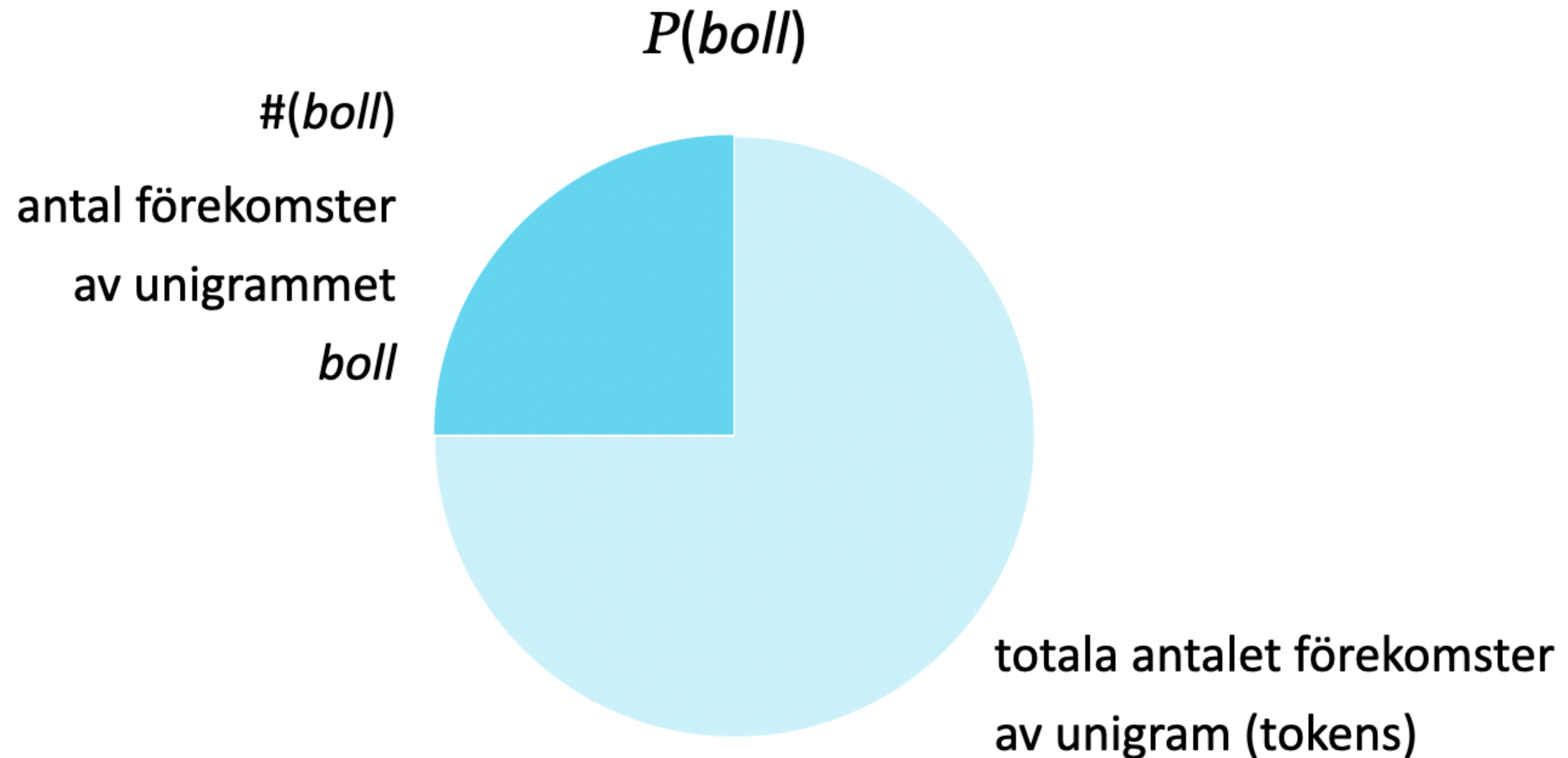
Exempel

Som bekant uppfördes Eiffeltornet till världsutställningen (...)

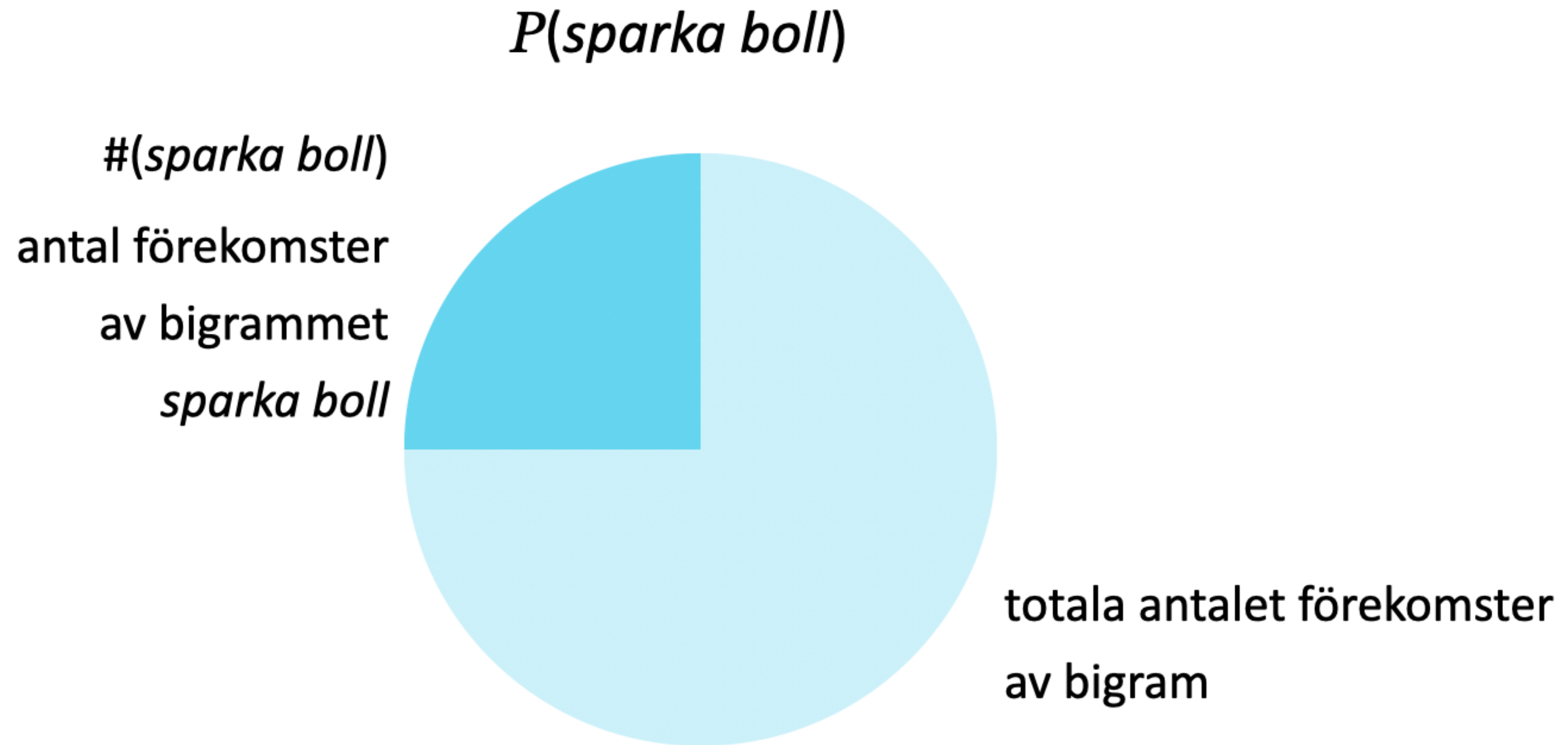
Unigram	Bigram	Trigram
Som	Som bekant	Som bekant uppfördes
bekant	bekant uppfördes	bekant uppfördes Eiffeltornet
uppfördes	uppfördes Eiffeltornet	uppfördes Eiffeltornet till
Eiffeltornet	Eiffeltornet till	Eiffeltornet till världsutställningen

Ord- och Bigramsannolikheter,

Skattning av unigramsannolikheter



Skattning av bigramsannolikheter



🔍

Linköping University,Uppsala University,Umeå University

✕

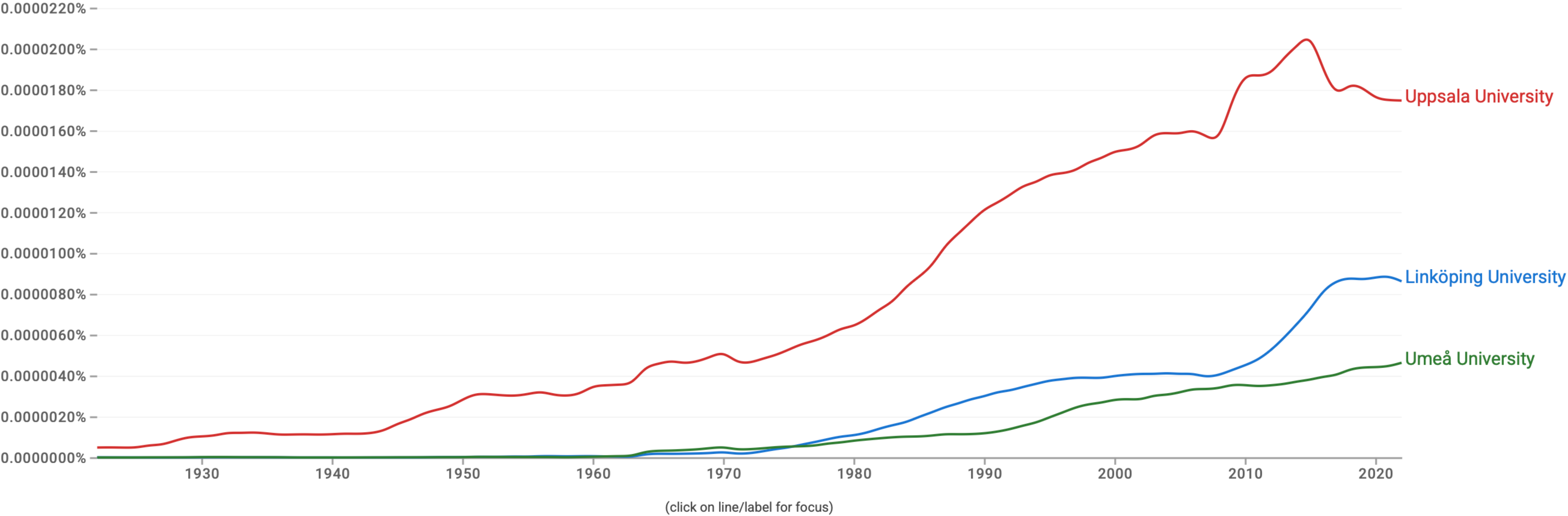
?

1922 - 2022 ▾

English ▾

Case-Insensitive

Smoothing ▾



Pointwise Mutual Information (PMI)

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 sannolikheter, dvs. om

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

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

$P(x)$	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
$P(x) P(y)$	oberoende sannolikheten för att ordet x förekommer och att ordet y förekommer

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

- 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.

Problem med PMI

- 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**.

PMI-värden i 50M bigrammar i Wikipedia

(bara bigrammar med >1000 förekomster)

<i>word 1</i>	<i>word 2</i>	<i>count word 1</i>	<i>count word 2</i>	count of co-occurrences	PMI
puerto	rico	1938	1311	1159	10.0349081703
hong	kong	2438	2694	2205	9.72831972408
los	angeles	3501	2808	2791	9.56067615065
carbon	dioxide	4265	1353	1032	9.09852946116
prize	laureate	5131	1676	1210	8.85870710982
san	francisco	5237	2477	1779	8.83305176711
nobel	prize	4098	5131	2498	8.68948811416
ice	hockey	5607	3002	1933	8.6555759741
star	trek	8264	1594	1489	8.63974676575
car	driver	5578	2749	1384	8.41470768304
it	the	283891	3293296	3347	-1.72037278119
are	of	234458	1761436	1019	-2.09254205335

[https://en.wikipedia.org/wiki/Pointwise mutual information](https://en.wikipedia.org/wiki/Pointwise_mutual_information)

PMI-värden i 50M bigrammar i Wikipedia

(bara bigrammar med >1000 förekomster)

<i>word 1</i>	<i>word 2</i>	<i>count word 1</i>	<i>count word 2</i>	count of co-occurrences	PMI
puerto	rico	1938	1311	1159	10.0349081703
hong	kong	2438	2694	2205	9.72831972408
los	angeles	3501	2808	2791	9.56067615065
carbon	dioxide	4265	1353	1032	9.09852946116
prize	laureate	5131	1676	1210	8.85870710982
san	francisco	5237	2477	1779	8.83305176711
nobel	prize	4098	5131	2498	8.68948811416
ice	hockey	5607	3002	1933	8.6555759741
star	trek	8264	1594	1489	8.63974676575
car	driver	5578	2749	1384	8.41470768304
it	the	283891	3293296	3347	-1.72037278119
are	of	234458	1761436	1019	-2.09254205335

[https://en.wikipedia.org/wiki/Pointwise mutual information](https://en.wikipedia.org/wiki/Pointwise_mutual_information)

L4

Tack för idag!