Lingvistiskt uppmärkt text 1

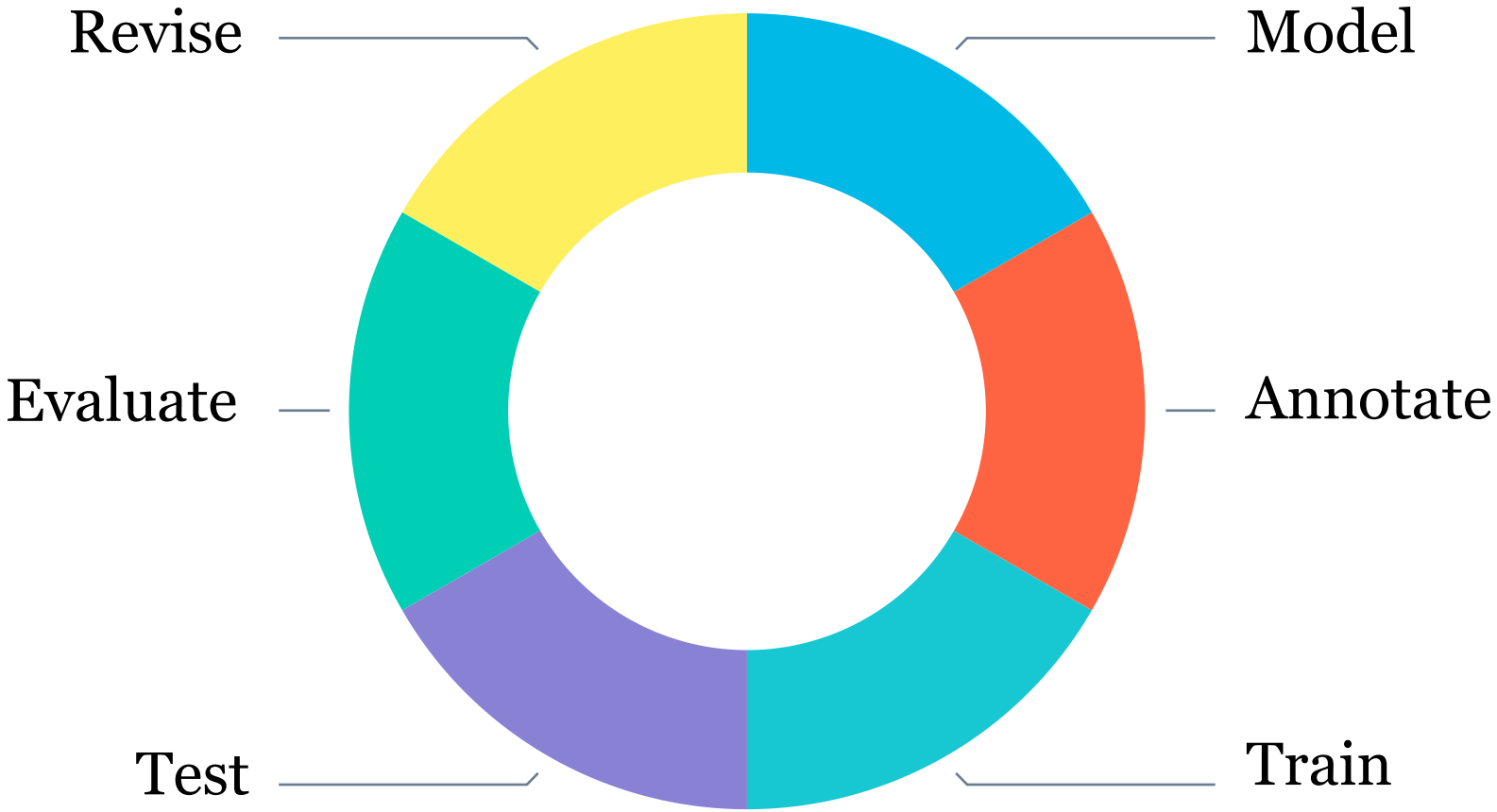
Inför laboration 2

729G49 Språk och Datorer

(baserat på Marco Kuhlmanns bilder från 2020)

LINKÖPINGS UNIVERSITET

Utvecklingscykeln



Källa: Pustejovsky och Stubbs (2013)

Möjliga fel vid tokenisering

Undersegmentering

Den automatiska tokeniseringen missar att segmentera en teckensekvens som enligt guldstandard ska segmenteras

Översegmentering

Den automatiska tokeniseringen delar på en teckensekvens som enligt guldstandarden inte ska segmenteras

bl. a.
t. ex.
New York
Anna-Lena
mat- och sovklocka

Normalisering

Ändra alla bokstäver till gemener

- windows Windows

Harmonisering av stavningsvarianter

- colour, color; gaol, jail; metre, meter

Avstamning (borttagandet av suffix)

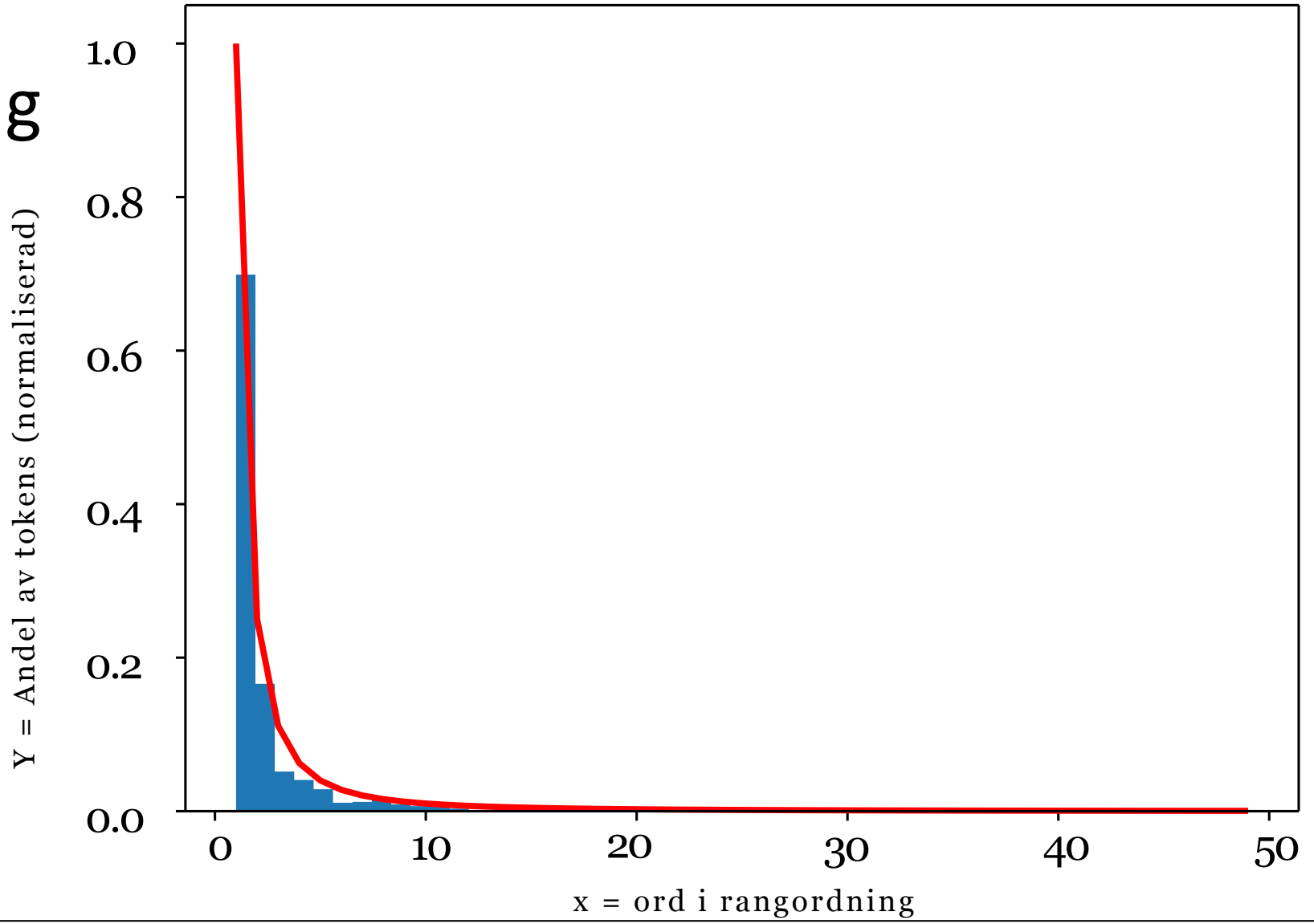
- räknade → räkna, räknaren → räkna

Denna föreläsning

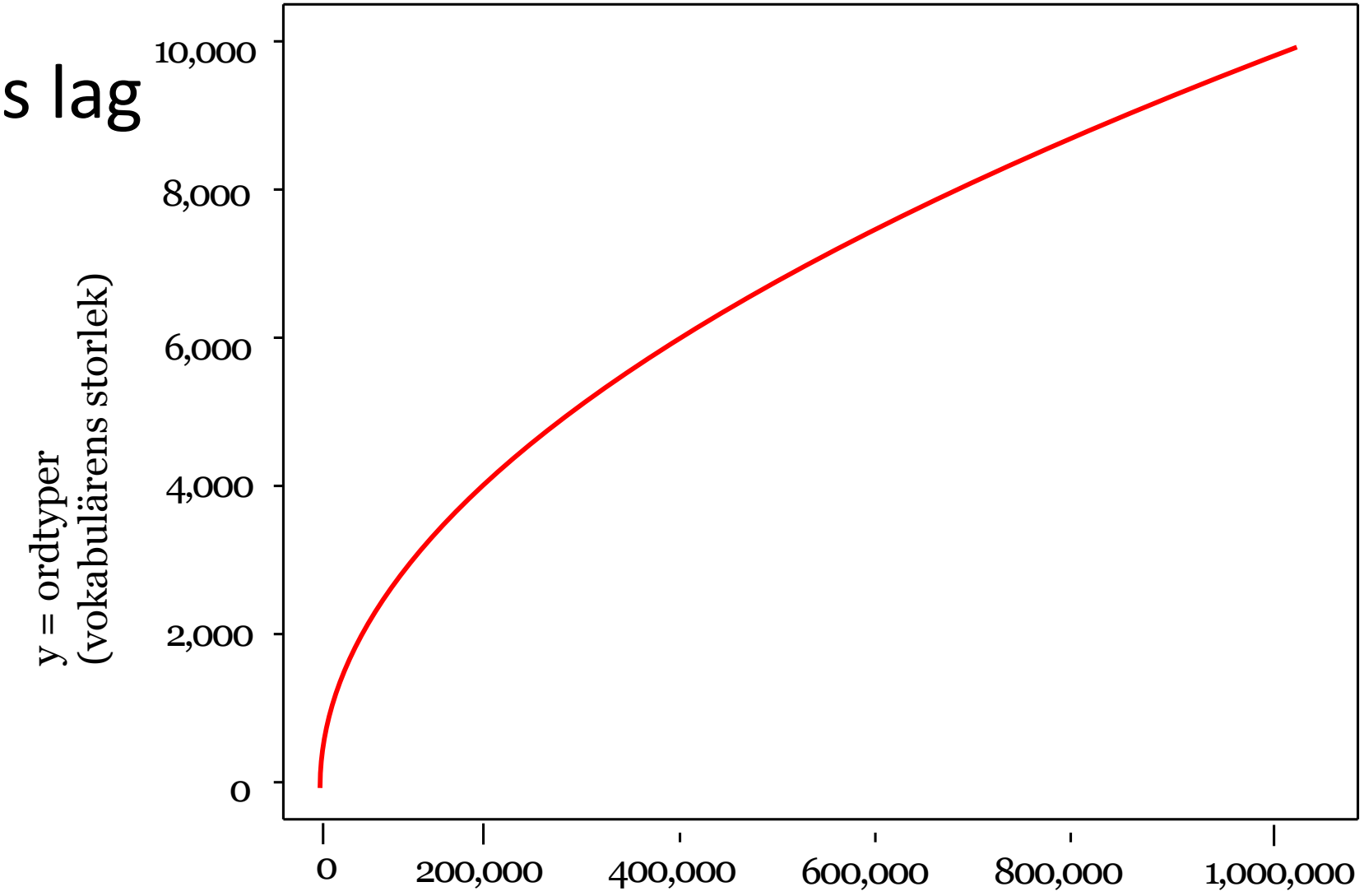
- Frekvensanalys, statistiska egenskaper hos språk
- Lingvistiskt uppmärkt text
- Olika filformat
- Generatorfunktioner

Frekvensanalys, statistiska egenskaper hos språk

Zipfs lag


1.0
0.8
0.6
0.4
0.2
0.0
Y = Andel av tokens (normaliserad)
0
10
20
30
40
50
x = ord i rangordning

Heaps lag


10,000
8,000
6,000
4,000
2,000
0
y = ordtyper
(vokabulärens storlek)
0
200,000
400,000
600,000
800,000
1,000,000
x = antal ord i en korpus

Denna föreläsning

- Frekvensanalys, statistiska egenskaper hos språk
- Lingvistiskt uppmärkt text
- Olika filformat
- Generatorfunktioner

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

Vad finns i en korpus?

• 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*

Graford och ordtyper

’Rose is a rose is a rose is a rose.’

Gertrude Stein (1874–1946)

Graford och ordtyper

’Rose is a rose is a rose is a rose.’

Gertrude Stein (1874–1946)

Korpus	Antal graford	Antal ordtyper
Shakespeare	ca 884,000	ca. 31,000
Riksmöte 2012/2013	4,645,560	96,114
Google Ngrams	1,176,470,663	13,588,391

Många olika typer av ord

Begreppet ord kan syfta på ett **graford** eller en **ordtyp**

- *Rose is a rose is a rose is a rose*.

Begreppet **lexem** betecknar en mängd ordformer som representerar samma grundläggande betydelse

- Ordformer: *tanke, tanken, tankar, tankarna, tankarnas*. Lexem: TANKE

Begreppet **lemma** betecknar den form av ett lexem som brukar användas för att representera lexemet i t. ex. en ordbok

- för substantiv: nominativ singularis (*tanke*), för verb: infinitiv *(att tänka)*

Skolans nio ordklasser

Tagg	Kategori	Exempel
VB	verb	kasta
NN	substantiv	pudding
PN	pronomen	hon
JJ	adjektiv	glad
AB	adverb	inte
KN	konjunktion	och
PP	preposition	över
RG	räkneord	tre
IN	interjektion	aj

Finkornighet i beskrivningen

Varje korpus måste bestämma sig för ett system med vars hjälp de lingvistiska datan ska beskrivas

- 9 ordklasser eller 22?

Förutom själva korpusen måste det därför finnas en definition av de kategorier som använts.

- ofta tillsammans med konkreta exempel.

Ordklasser i SUC (1)

Tagg	Kategori	Exempel
NN	substantiv	*pudding*
VB	verb	*kasta*
PP	preposition	*över*
AB	adverb	*inte*
JJ	adjektiv	*glad*
PN	pronomen	*hon*
DT	determinerare	*denna*
KN	konjunktion	*och*
PM	egennamn	*Evelina*

https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf

Ordklasser i SUC (2)

Tagg	Kategori	Exempel
PC	particip	*utsänd*
SN	subjunktion	*att*
RG	räkneord (grundtal)	*tre*
HP	frågande/relativt pronomen	*som*
IE	infinitivmärke	*att*
PL	partikel	*ut*

https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf

Ordklasser i SUC (3)

Tagg	Kategori	Exempel
PS	possessivt pronomen	*hennes*
HA	frågande/relativt adverb	*när*
UO	utländskt ord	*the*
RO	räkneord (ordningstal)	*tredje*
IN	interjektion	*ja*
HD	frågande/relativ determinerare	*vilken*
HS	frågande/relativt possessivt pronomen	*vars*

https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf

Denna föreläsning

- Frekvensanalys, statistiska egenskaper hos språk
- Lingvistiskt uppmärkt text
- Olika filformat
- Generatorfunktioner

Olika filformat

Hur representeras korpusdata?

- Tabulerade data
- Extensible Markup Language (XML)
- JavaScript Object Notation (JSON)

Tabulerade data (CoNLL-format)

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

Tabulerade data (CoNLL-format)

- Filen struktureras i rader och kolumner
- Rader skiljs åt med ett nyrad-tecken
- Kolumner skiljs åt med ett separatortecken (ex. tab)

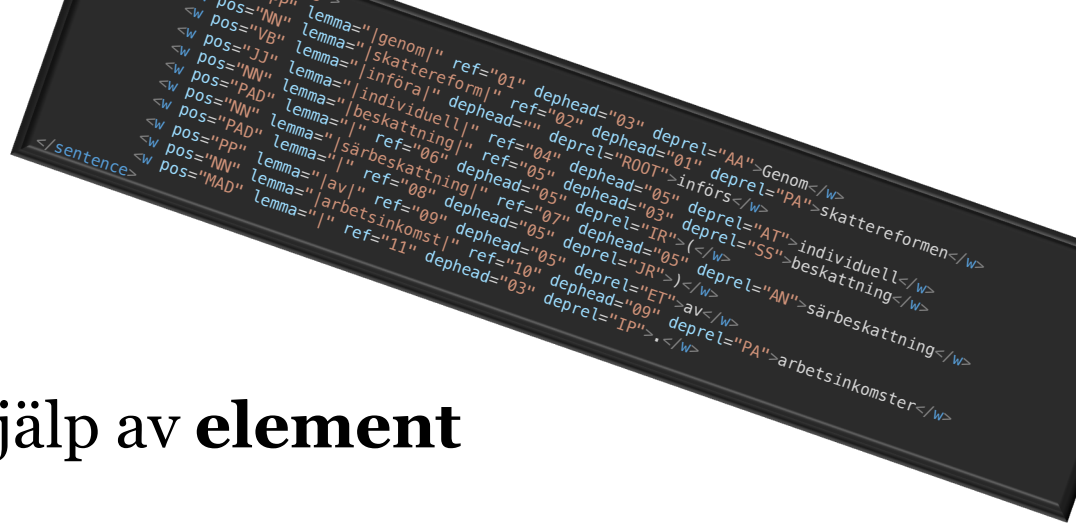
CoNLL-format

- CoNLL-formatet har använts i flera tävlingar i samband med International Conference for Natural Language Learning
- Formatet representerar en lista med meningar
- Slutet på en mening markeras med en blankrad

XML

```
<sentence id="8f74-8115">
    <w pos="PP" lemma="|genom|" ref="01" dephead="03" deprel="AA">Genom</w>
    <w pos="NN" lemma="|skattereform|" ref="02" dephead="01" deprel="PA">skattereformen</w>
    <w pos="VB" lemma="|införa|" dephead="" deprel="ROOT">införs</w>
    <w pos="JJ" lemma="|individuell|" ref="04" dephead="05" deprel="AT">individuell</w>
    <w pos="NN" lemma="|beskattning|" ref="05" dephead="03" deprel="SS">beskattning</w>
    <w pos="PAD" lemma="|" ref="06" dephead="05" deprel="IR">(</w>
    <w pos="NN" lemma="|särbeskattning|" ref="07" dephead="05" deprel="AN">särbeskattning</w>
    <w pos="PAD" lemma="|" ref="08" dephead="05" deprel="JR">)</w>
    <w pos="PP" lemma="|av|" ref="09" dephead="05" deprel="ET">av</w>
    <w pos="NN" lemma="|arbetsinkomst|" ref="10" dephead="09" deprel="PA">arbetsinkomster</w>
    <w pos="MAD" lemma="|" ref="11" dephead="03" deprel="IP">.</w>
</sentence>
```

XML



- Information struktureras hierarkiskt med hjälp av **element**
- Ett element består av en starttagg och en sluttagg

  <w>särbeskattning</w>
- Varje element kan dessutom ha ett antal attribut-värde-par

  <w pos=”NN”>särbeskattning</w>
- Ett element kan innehålla såväl text som andra element

Fördelar och nackdelar

Tabulerade data

- enkelt och platseffektivt
- implicit representation
  - För att ta ut en ordklasstagg behöver man veta vilken kolumn den är i

XML

- komplext och platskrävande
- explicit representation
  - För att ta ut en ordklasstagg kan man direkt efterfråga motsvarande attribut

Denna föreläsning

- Frekvensanalys, statistiska egenskaper hos språk
- Lingvistiskt uppmärkt text
- Olika filformat
- Generatorfunktioner

Generatorfunktioner

Vad är en generatorfunktion i Python?

- Ett sätt att använda data utan att data lagras
- Ett sätt att minska utnyttjandet av minne
- Under *Material* på kurshemsidan för L2 finns en notebook med exempel och förklaringar

Vad är en generatorfunktion i Python?

- När den kallas på så returnerar den en objekt men den exekveras inte automatiskt
- Använder nyckelordet *yield* för att leverera relevanta data. Kan innehålla ett eller fler *yield*.
- När funktionen når *yield* så pausas den och lokala variabler och deras tillstånd behålls i minnet mellan anropen

```
def file_reader(file_name):
    file = open(file_name)
    result = file.read().split('\n')
    return result
```

```
def file_reader(file_name):
    for row in open(file_name, 'r'):
        yield row
```

```
generated_text = file_reader('some_big.txt')

row_count = 0
for row in generated_text:
    row_count += 1

print('Row count is ' + str(row_count))
```

Tack för idag!

LINKÖPINGS UNIVERSITET