AI för naturligt språk

Språkliga datamängder

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Vad är en korpus?

- **korpus** s. *-en -ar*, (språkv.) mängd av undersökt textmaterial

  Svenska Akademiens ordlista

- A body of texts, utterances or other specimens considered more or less representative of language; usually stored electronically.

  The Oxford Companion to the English Language

- Inom NLP används begreppet för alla sorters språkdatamängder, inte bara sådana som framtagits i lingvistiska syften.

Korpusar av olika slag

- Generella korpusar

  problem: representativitet

- Specifika korpusar

  inriktning mot en viss genre, författare, historisk period

- Parallela korpusar

  samma text i flera språk, för att studera översättningsfrågor

Några klassiska korpusar

Namn	År	Storlek	Språk och sort
British National Corpus (BNC)	1994	100 miljoner ord	brittisk engelska, talad och skriven
American National Corpus (ANC)	2003	22 miljoner ord	amerikansk engelska, talad och skriven
Corpus of Contemporary American English (COCA)	2008	425 miljoner ord	amerikansk engelska, talad och skriven
Stockholm–Umeå Corpus (SUC)	2006	1 miljon ord	svenska, skriven

Länkar till korpusar och korpusverktyg

- Stockholm–Umeå Corpus (Språkbanken)
- Korp (sökvertyg) (Språkbanken)
- Google Books Ngram Viewer
- Universal Dependencies Project

Från råmaterial till annoterad text

Aktivitet	Beskrivning
Urval	Välja ut de texter som ska vara med i korpusen
Insamling	Samla in texterna, t.ex. genom att ”spindla” webben
Avformatering	Ta bort strukturell uppmärkning, t.ex. HTML, XML
Segmentering	Dela upp textmaterialet i relevanta enheter
Annotering	Lägga till relevant information, t.ex. ordklasser

AI för naturligt språk

Från råmaterial till segmenterad text

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

WIKIPEDIA
Den fria encyklopedin

Huvudsida
Skriv en ny artikel
Deltagarportalen
Bybrunnen
Senaste ändringarna
Slumpartikel
Ladda upp filer
Stöd Wikipedia
Kontakta Wikipedia
Hjälp

Skriv ut/exportera
Skapa en bok
Ladda ner som PDF
Utskriftsvänlig version

Verktyg
Sidor som länkar hit
Relaterade ändringar
Specialsidor
Permanent länk
Sidinformation
Wikidataobjekt
Citera denna sida

På andra projekt
Bilder & media

Trädkrypare

Denna artikel avhandlar arten trädkrypare. För familjen Certhiidae se Trädkrypare (familj)

Trädkrypare (*Certhia familiaris*) är en fågelart inom familjen trädkrypare (Certhiidae).

Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovansida, vitaktig undersida och långa styva stjärtpennor som den använder för kunna balansera upprätt på trädstammar och grenar. Den kan lättast urskiljas från den likartade trädgårdsträdkryparen, som delar en stor del av dess utbredning i Europa, genom skillnaden i sång.

Trädkryparen har nio eller fler underarter som häckar i olika delar av dess utbredningsområde i de tempererade delarna av Eurasien. Arten återfinns i alla sorters skogsmarker, men där den överlappar med trädgårdsträdkryparen i Västeuropa återfinns den med större sannolikhet i barrskogar eller på högre höjder. Den häckar i trädhålor eller bakom barkflagor, och föredrar introducerade mammutträd som häckningsplatser där sådana finns tillgängliga. Honan lägger vanligen fem eller sex vita ägg med rosa prickar i det fodrade boet, men ägg och ungar är sårbara för angrepp från hackspettar och däggdjur, däribland ekorrar.

Trädkryparen är insektsätare och klättrar uppför trädstammar som en mus, för att leta efter insekter som den plockar från skrevor i barken med sin fina böjda näbb. Den flyger sedan till foten av ett annat träd med en distinkt oregelbunden flykt. Fågeln lever ensam på vintern, men kan bilda gemensamma sovplatser

Trädkrypare

Status i världen: Livskraftig (lc)[1]

Status i Sverige: Livskraftig

Systematik

Domän Eukaryoter

Trädkrypare – Wikipedia

sv.wikipedia.org/wiki/Trädkrypare

Reader

Skapa konto Logga in

WIKIPEDIA
Den fria encyklopedin

Artikel | Diskussion | Visa | Redigera | Redigera wikitext | Visa historik | Sök

- Huvudsida
- Skriv en ny artikel
- Deltagarportalen
- Bybrunnen
- Senaste ändringarna
- Slumpartikel
- Ladda upp filer
- Stöd Wikipedia
- Kontakta Wikipedia
- Hjälp

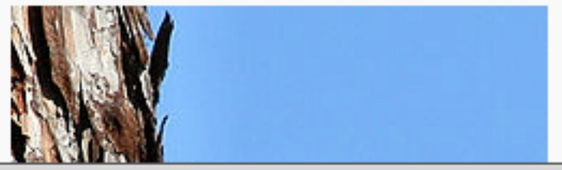
Trädkrypare

Denna artikel avhandlar arten trädkrypare. För familjen Certhiidae se Trädkrypare (familj)

Trädkrypare (*Certhia familiaris*) är en fågelart inom familjen trädkrypare (Certhiidae).

Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovansida, vitaktig undersida och långa styva stjärtpennor som den använder för kunna balansera upprätt på trädstammar och grenar. Den kan lättast urskiljas från den likartade trädgårdsträdkryparen, som delar en stor del av dess utbredning i Europa, genom skillnaden i sång.

Trädkrypare
Status i världen: Livskraftig (lc)[1]
Status i Sverige: Livskraftig



38 | 1 | 0 | 57

Search Resource Content

- Trädkrypare — sv.wikipedia.org
  - Images
  - Scripts
  - Stylesheets
  - Cookies — sv.wikipedia.org
  - Extension Scripts
  - Local Storage — sv.wikipedia.org
  - Session Storage — sv.wikipedia.org

Filter Resource List

div#mw-conte… | p | "Den liknar andra arter inom fam…"

```
<table class="infobox" cellspacing="5" style="width: 22em; text-align: left; font-size:
88%; line-height: 1.5em; font-size:95%;width:20.3788em;padding:.18em;">…</table>
<p>…</p>
<p>
  "Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovansida, vitaktig
  undersida och långa styva stjärtpennor som den använder för kunna balansera upprätt på
  trädstammar och grenar. Den kan lättast urskiljas från den likartade "
  <a href="http://sv.wikipedia.org/wiki/Tr%C3%A4dg%C3%A5rdstr%C3%A4dkrypare" title=
  "Trädgårdsträdkrypare">trädgårdsträdkryparen</a>
  ", som delar en stor del av dess utbredning i "
  <a href="http://sv.wikipedia.org/wiki/Europa" title="Europa">Europa</a>
  ", genom skillnaden i sång."
</p>
<p>…</p>
<p>…</p>
<p></p>
```

Textsegmentering

- **Textsegmentering** är uppgiften att dela upp en text i lingvistiskt meningsfulla enheter, såsom ord, meningar och stycken.
- När enheterna en segmenterar i är ord eller ordliknande enheter kallas textsegmentering för **tokenisering**.

  token = ord, tal, skiljetecken

Tokenisering

Rå text	Tokeniserad text
Den liknar andra arter inom familjen med böjd näbb, mönstrad brun ovansida, vitaktig undersida och långa styva stjärtpennor som den använder för att kunna balansera upprätt på trädstammar och grenar.	Den liknar andra arter inom familjen med böjd näbb , mönstrad brun ovansida , vitaktig undersida och långa styva stjärtpennor som den använder för att kunna balansera upprätt på trädstammar och grenar .

Möjliga fel vid tokenisering

- **Undersegmentering**

  Den automatiska tokeniseringen missar att segmentera en teckensekvens som egentligen ska segmenteras.

- **Översegmentering**

  Den automatiska tokeniseringen delar på en teckensekvens som egentligen inte ska segmenteras.

Undersegmentering och översegmentering

Förväntad tokenisering	Automatisk tokenisering
kamm ,	kamm,
New York (1 token)	New York (2 token)
bl. a.	bl.a.
högskole- eller universitetsutbildning	högskole - eller universitetsutbildning

AI för naturligt språk

Lingvistiska annotationer

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Ett konkret exempel

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	substantiv (NN)
• huvudord	ord nummer 3 i meningen (*införs*)
• grammatisk funktion	subjekt (ss)
• 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)

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**.
- 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 tanka*)

AI för naturligt språk

Dataformat för språkliga datamängder

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

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

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

  tabulator, komma

Extensible Markup Language (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>
```

Extensible Markup Language

- 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> har ett attribut pos med värdet NN
- Ett element kan innehålla såväl text som andra element.

Fördelar och nackdelar

Tabulerade data

- enkelt och platseffektivt
- implicit representation

  För att ta ut en ordklass behöver man veta vilken kolumn den ligger i.

Extensible Markup Language (XML)

- komplext och platskrävande
- explicit representation

  För att ta ut en ordklass kan man direkt efterfråga motsvarande attribut.

AI för naturligt språk

Statistiska egenskaper hos språkdata

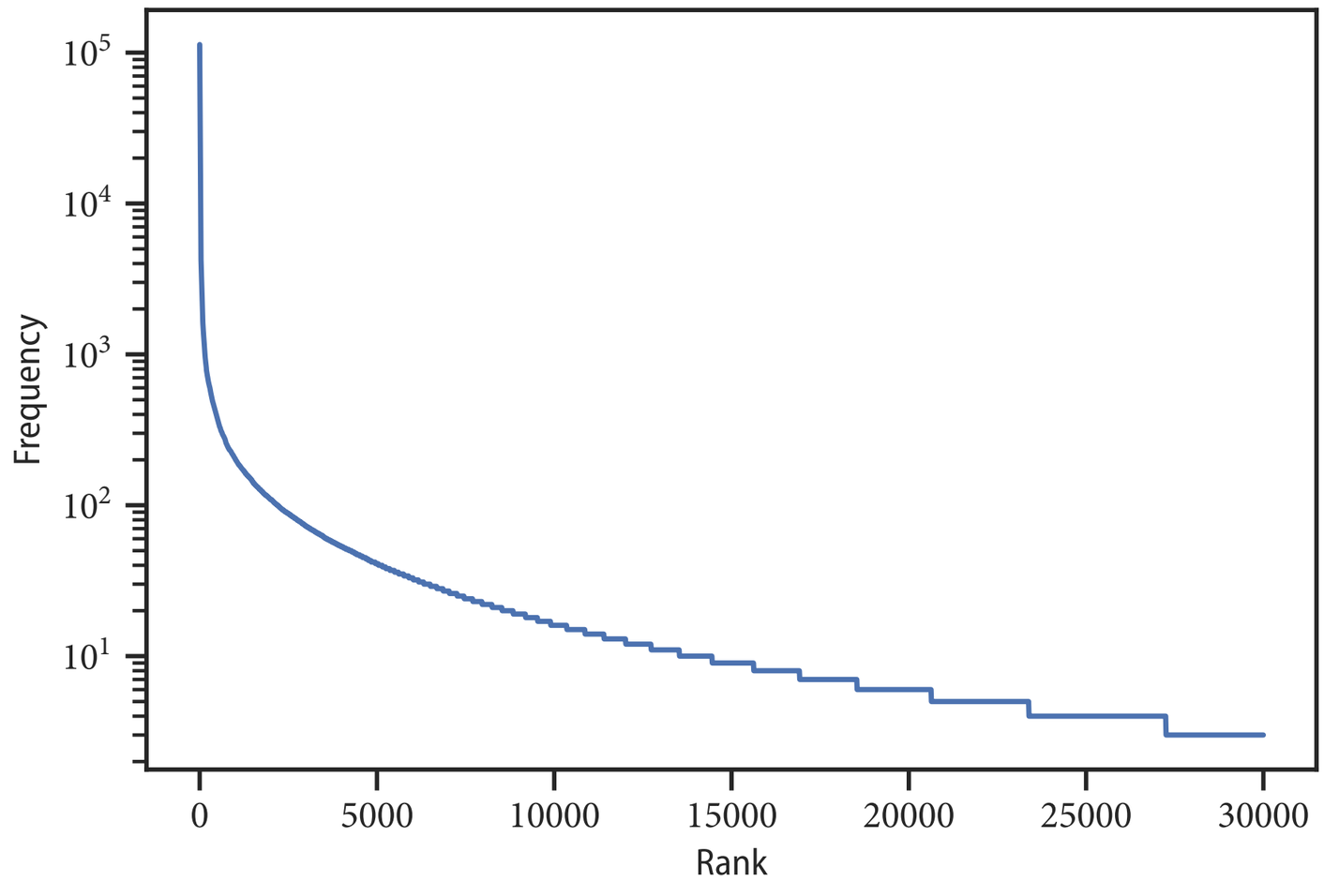
Marco Kuhlmann
Department of Computer and Information Science

LINKÖPINGS UNIVERSITET

Zipfs lag

- **Zipfs lag** är ett empiriskt visat samband som beskriver en statistisk egenskap hos lingvistiska datamängder.
- Den säger att ett ords frekvens (antal förekomster i textmängden) är omvänt proportionell mot ordets rankning i frekvenstabellen.
- Zipfs lag är uppkallad efter den amerikanske lingvisten George Kingsley Zipf (1902–1950).

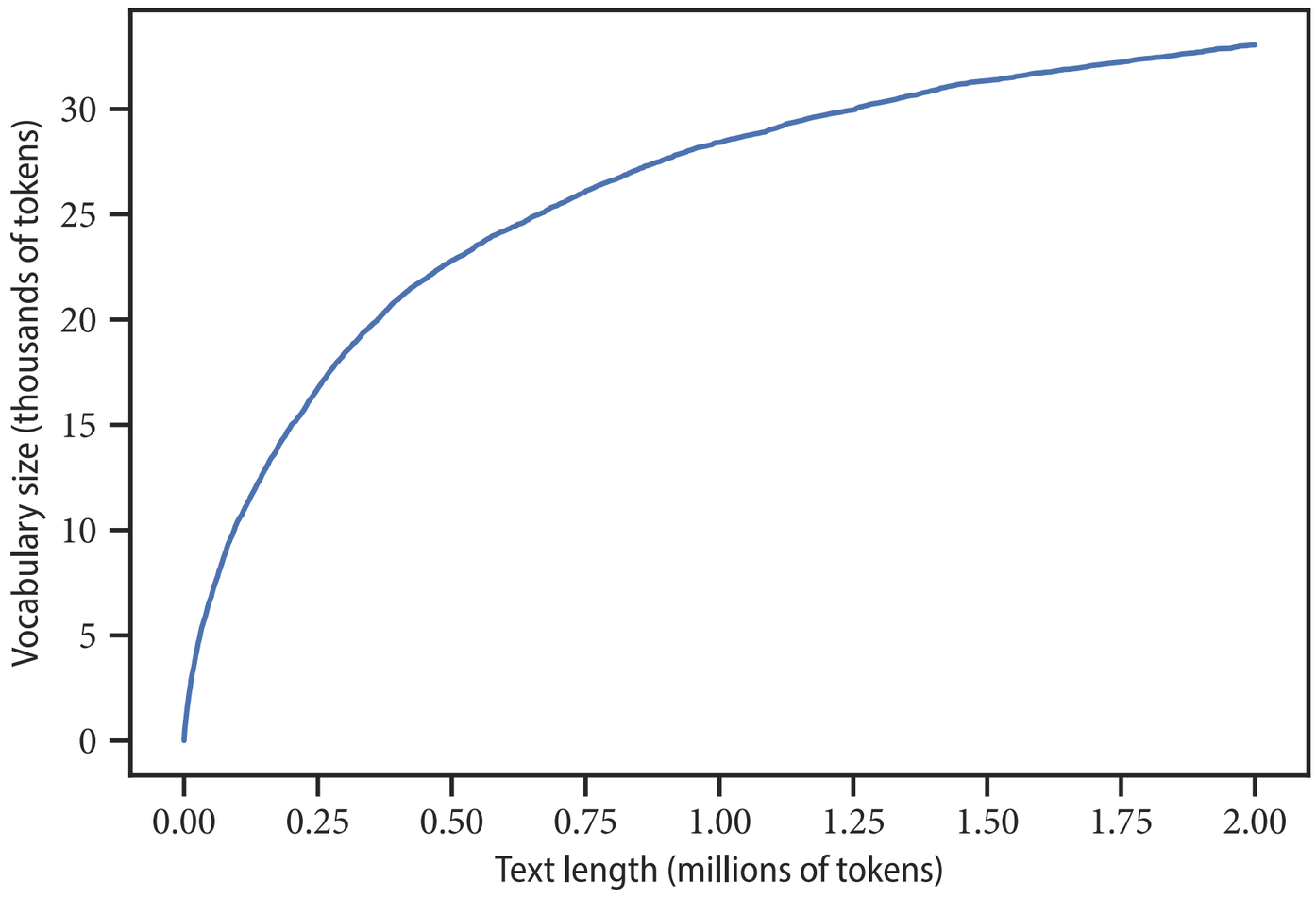
Zipfs lag


10^5
10^4
10^3
10^2
10^1
Frequency
0
5000
10000
15000
20000
25000
30000
Rank

Heaps' lag

- Även **Heaps' lag** är ett empiriskt visat samband som gäller stora textdatamängder.
- Den beskriver antalet unika ord (ordtyper) som en funktion av textdatamängdens storlek.
- Ju mer text vi ser, desto färre nya unika ord kommer vi upptäcka, samtidigt som vi aldrig slutar upptäcka nya unika ord.

Heaps lag


Vocabulary size (thousands of tokens)
0
5
10
15
20
25
30
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Text length (millions of tokens)