UPPSALA UNIVERSITET

Grundläggande Textanalys VT 2015

Språkgranskning (1)

Eva Pettersson

eva.pettersson@lingfil.uu.se

Referatuppgiften

- 10 minuters muntlig presentation av vetenskaplig artikel med 5 minuters efterföljande diskussion
- 1-2 sidors skriftlig sammanfattning i pdf-format till eva.pettersson@lingfil.uu.se senast 5 juni
- Obligatorisk närvaro vid båda presentationstillfällena
- Vid frånvaro: skriftlig sammanfattning av presenterade artiklar
- Tre tider att välja på (vi stryker en):
  - ~~Torsdag 21 maj klockan 10-12~~
  - Torsdag 28 maj klockan 10-12
  - Torsdag 28 maj klockan 13-15

UPPSALA UNIVERSITET

Översikt

- Denna gång
  - Stavningskontroll
    - Allmänt om stavningskontroll
    - Feligenkänning
    - Felkorrigering
    - Samarbetsuppgift
- Nästa gång
  - Grammatikkontroll
  - Stilkontroll
  - Kontrollerat språk
  - Språkgranskningssystem, med fokus på MS Word och Granska

Vad förväntas av det ideala stavningskontrollprogrammet?

- Känna igen och larma för alla felstavade ord (täckning)
- Känna igen och acceptera alla rättstavade ord (precision)
- Ge ett korrekt rättningsförslag för alla felstavade ord

Realistiska förväntningar på stavningskontrollprogrammet

- Känna igen och larma för ~~alla~~ *de mest frekventa och/ eller lättidentifierade* felstavningarna
- Känna igen och acceptera alla rättstavade ord*, som är tillräckligt frekventa i språket*
- Ge ett ~~korrekt~~ *sannolikt* rättningsförslag för alla felstavade ord

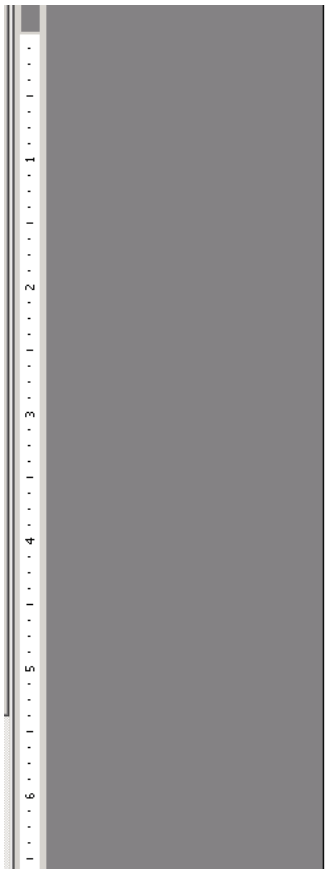
UPPSALA UNIVERSITET

Stavningskontrollens två delar

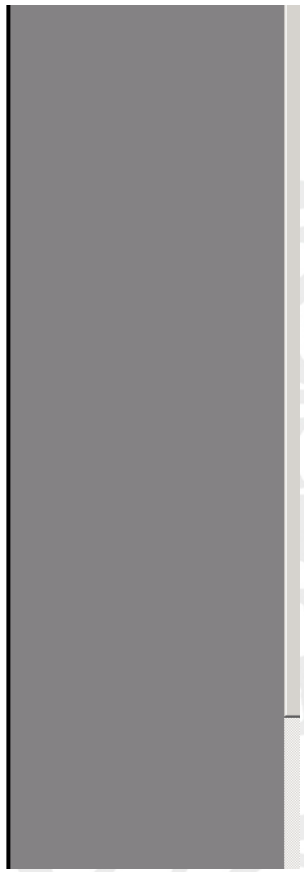
1. Feligenkänning (*error detection*)
   att hitta felen

2. Felkorrigering (*error correction*)
   att ge ersättningsförslag

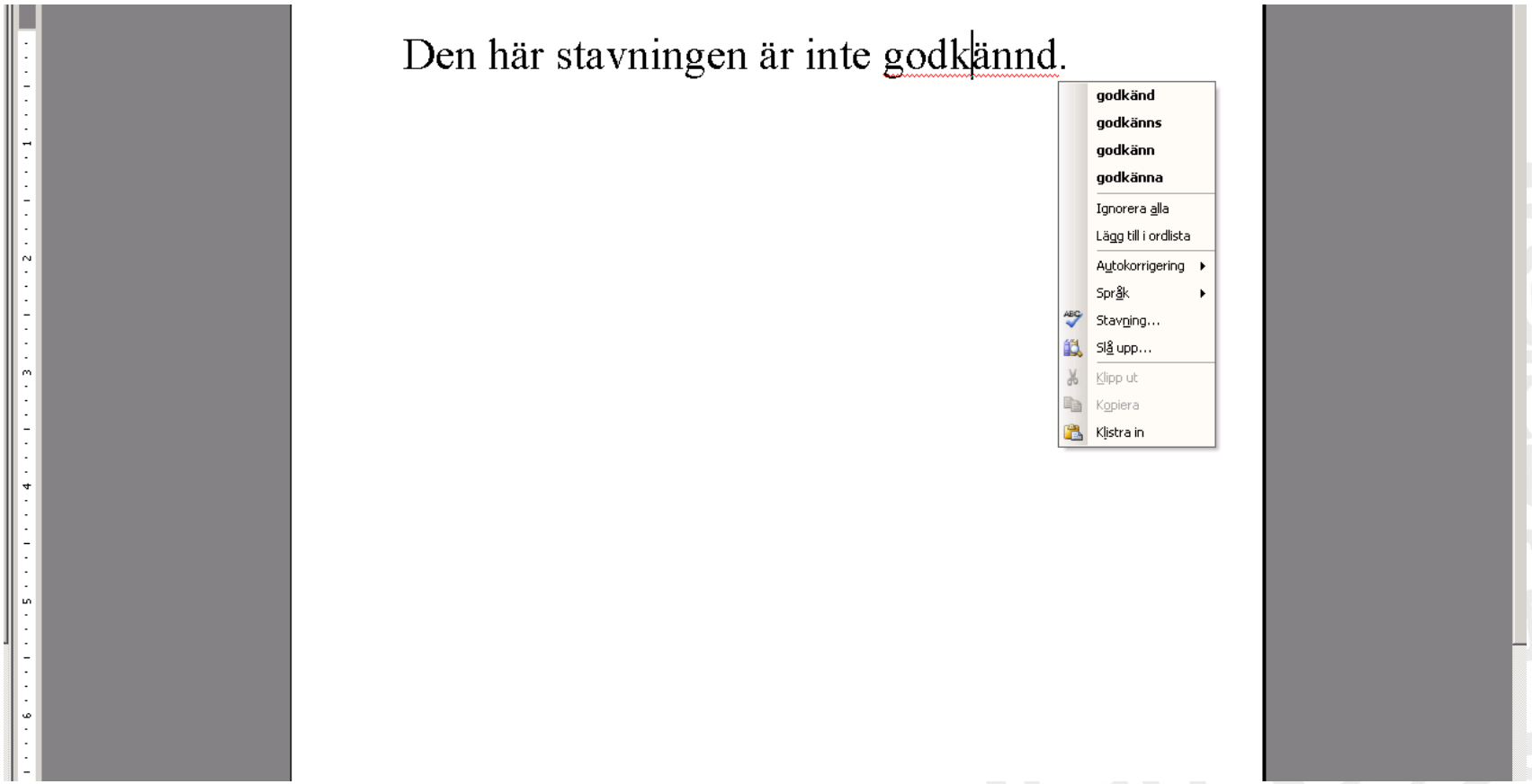
Feligenkänning i Microsoft Word



Den här stavningen är inte godkännd.



Felkorrigering i Microsoft Word



UPPSALA UNIVERSITET

FELIGENKÄNNING

UPPSALA UNIVERSITET

Feligenkänning

- Isolerade ord
  - Skrivfel som resulterar i icke-ord: *och* → *coh*
- Ord i kontext
  - Stavfel som resulterar i riktiga ord (real word errors)
    jag er dålig på att stava
    språk teknologi är kul
  - Bättre korrigeringsförslag
    *det är **sårt** att stava*
    Word föreslår: såret
    svårt
    sårat
    såt
    sått

UPPSALA UNIVERSITET

Feligenkänningsstrategier

- Trigram av tecken
  - Larmar för ovanliga teckenkombinationer
  - Används främst inom OCR
- Lexikon
  - Ord som saknas i lexikonet larmas för som felstavningar
  - Fullformslexikon eller stamlexikon

Svagheter med lexikonmetoden

- För stort lexikon ger låg täckning
  - många fel missas (t.ex. *verv, boke*)
- För litet lexikon ger låg precision
  - många falska alarm
  - kan lura skribenten att till exempel särskriva
- Omöjligt att lista alla ord i lexikonet – språket är produktivt

UPPSALA UNIVERSITET

Komplement till lexikonmetoden

- Morfologiska regler för avledningar
  svamp-ig (jmf. *svamp-ar*)
  be-bo (jmf. *bo-r*)
  vattn-a (jmf. *vattn-et*)
- Regler för att hantera sammansättningar
- Egennamnsigenkänning
- Domänspecifika lexikon
- Tillåta användaren att lägga till egna ord i lexikonet

UPPSALA UNIVERSITET

Feligenkänning av sammansättningar

- Basstrategi: sammansättning = ord + ord
  - *dator* + *lingvistik* = *datorlingvistik*
- Förfinad strategi: sammansättning = förled + ord
  - *flicka* + *klänning*= *flickklänning*
  - *äpple* + *paj* = *äppelpaj*
  - *kvinna* + *parti* = *kvinnoparti*
  - *cigarr* + *rök* = *cigarrök*

UPPSALA UNIVERSITET

För- och nackdelar med sammansättningsanalys

- Minskar antalet falska alarm (bättre precision)

- Kan öka antalet missade fel (sämre täckning)
  Missade fel i Word97 (åtgärdat i senare versioner):

kotakt	*kontakt*
makelera	*makulera*
medalg	*medalj*
cykelsäll	*cykelställ*
särkskilt	*särskilt*

Att fundera kring

- Hur göra med sällsynta/fackspråkliga ord? Kan ligga nära felskrivningar av frekventa ord…
  - *verv/värv* (verv = kraft, livfullhet, glöd enligt SAOL)
  - *boke/boken* (boke = bokvirke)
- Dialektala ord?
  däven, tyket, hurven, tölig
- Slang?
  keff, gola, kurra, impa, guss, deppa, dagis
- Talspråk? Hur sträng bör man vara?
  mej, direktörn, idag

UPPSALA UNIVERSITET

FELKORRIGERING

Felkorrigeringens två delar

- Ta fram ett antal korrigeringskandidater
- Rangordna korrigeringskandidaterna

Feltyper

- Kompetensfel (spelling confusion)
  - Fonetiska fel: *restaurang → resturang*
  - Homofonfel: *gott → gått*
- Performansfel (typographical errors/typos)
  - Insättning *språkteknologi**ii***
  - Borttagning *s**på**kteknologi*
  - Substitution *språkte**l**nologi*
  - Transposition *sp**år**kteknologi*

Feltyper (forts)

- Kompetensfel eller performansfel?
  tunnt
- Spelar det någon roll?
- Oftast inte nödvändigt att veta om kompetensfel eller performansfel
- Kan ge bättre korrigeringsförslag om man tar hänsyn till feltyp

hemta kompetensfel: *hämta/hämtade*
performansfel: *hemtam*

Empiriskt grundade iakttagelser

- De flesta felstavningar är performansfel (insättning, borttagning, substitution, transposition)
- De flesta felstavningar påverkar inte ordets längd med mer än en bokstav
- Första bokstaven i ordet är sällan felaktig
- Tangenternas placering påverkar
- Bokstävernas frekvenser påverkar

UPPSALA UNIVERSITET

Korrigeringsstrategier

- Editeringsavstånd (Minimum Edit Distance/Levenshtein Distance)
- Likhetsnycklar
- N-gramsbaserade tekniker
- Regelbaserade tekniker
- Probabilistiska tekniker

UPPSALA UNIVERSITET

Editeringsavstånd

- Stränglikhet
- Minsta antalet editeringsoperationer som behövs för att omvandla en sträng till en annan
- Editeringsoperationer:
  - insättning
  - borttagning
  - substitution (alt. borttagning + insättning)
  - transposition (alt. substitution + substitution)

Editeringsavstånd formel

dist(0,0) = 0
dist(*i*,0) = *i*
dist(0,*j*) = *j*

dist(*i*,*j*) = min

- dist(*i*-1,*j*) + 1 (borttagning)
- dist(*i*,*j*-1) + 1 (insättning)
- dist(*i*-*j*,*j*-1) + (likhet)
  - 0 om *i* = *j*
  - 1 annars (substitution)

där *i* är strängen *s* fram till *i*:te tecknet,
och *j* är strängen *t* fram till *j*:te tecknet

UPPSALA UNIVERSITET

Editeringsavstånd illustrerat

r ä n g n a

r e g n a r

Editeringsavstånd illustrerat

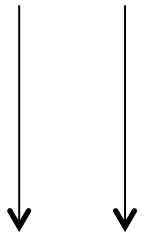
r ä n g n a

r e g n a r

Editeringsavstånd: 0

Editeringsavstånd illustrerat

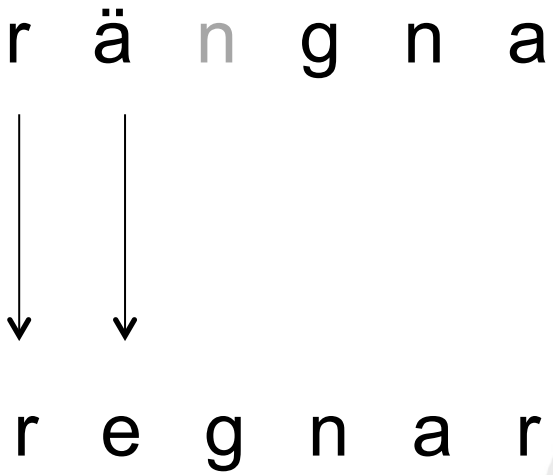
r ä n g n a



r e g n a r

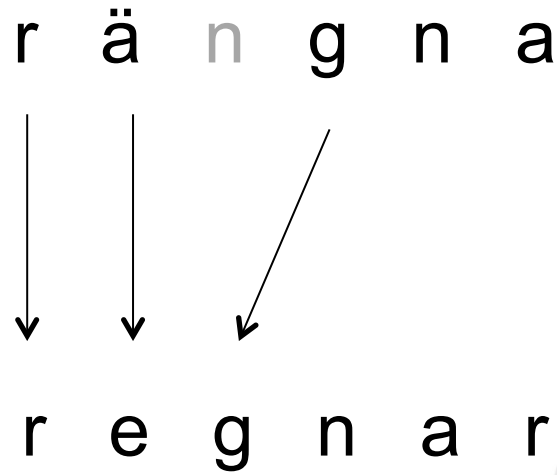
Editeringsavstånd: 1
substitution

Editeringsavstånd illustrerat



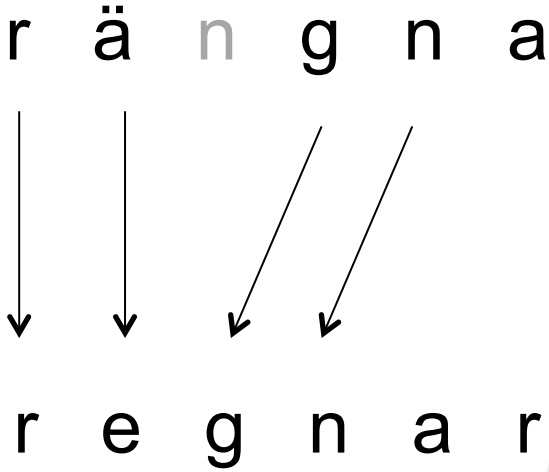
Editeringsavstånd: 2
substitution + borttagning

Editeringsavstånd illustrerat



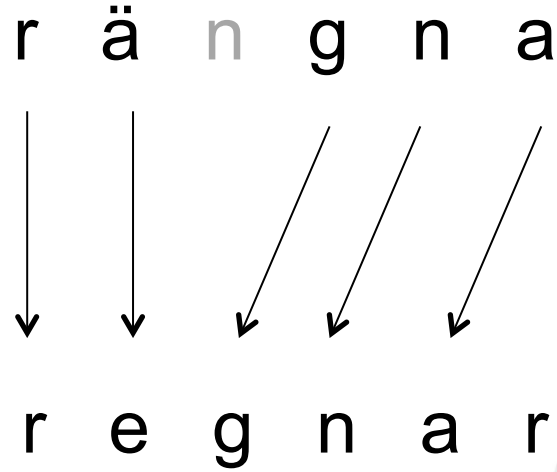
Editeringsavstånd: 2
substitution + borttagning

Editeringsavstånd illustrerat



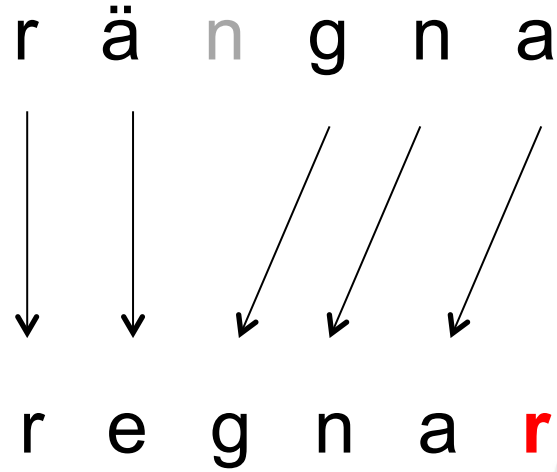
Editeringsavstånd: 2
substitution + borttagning

Editeringsavstånd illustrerat



Editeringsavstånd: 2
substitution + borttagning

Editeringsavstånd illustrerat



Editeringsavstånd: 3
substitution + borttagning + insättning

UPPSALA UNIVERSITET

Editeringsavstånd: dynamisk programmering

		r	**e**	**g**	**n**	**a**	**r**
	0	1	2	3	4	5	6
r							
ä							
n							
g							
n							
a							

UPPSALA UNIVERSITET

Editeringsavstånd: dynamisk programmering

		r	**e**	**g**	**n**	**a**	**r**
	0	1	2	3	4	5	6
r	1						
ä	2						
n	3						
g	4						
n	5						
a	6						

Editeringsavstånd: dynamisk programmering

		r	e	g	n	a	r
	0	1	2	3	4	5	6
r	1	0					
ä	2						
n	3						
g	4						
n	5						
a	6						

kostnaden för att komma hit från min övre vänstra granne **substitution**	kostnaden för att komma hit från min övre granne **borttagning**
kostnaden för att komma hit från min vänstra granne **insättning**	minimum av de tre möjliga “dragen”, dvs det billigaste sättet att komma hit

Editeringsavstånd: dynamisk programmering

		r	e	g	n	a	r
	0	1	2	3	4	5	6
r	1	0	1				
ä	2	1	1				
n	3						
g	4						
n	5						
a	6						

kostnaden för att komma hit från min övre vänstra granne **substitution**	kostnaden för att komma hit från min övre granne **insättning**
kostnaden för att komma hit från min vänstra granne **borttagning**	minimum av de tre möjliga “dragen”, dvs det billigaste sättet att komma hit

Editeringsavstånd: dynamisk programmering

		r	e	g	n	a	r
	0	1	2	3	4	5	6
r	1	**0**	1	2			
ä	2	1	**1**	2			
n	3	2	2	**2**			
g	4						
n	5						
a	6						

kostnaden för att komma hit från min övre vänstra granne **substitution**	kostnaden för att komma hit från min övre granne **insättning**
kostnaden för att komma hit från min vänstra granne **borttagning**	minimum av de tre möjliga “dragen”, dvs det billigaste sättet att komma hit

UPPSALA UNIVERSITET

Editeringsavstånd: dynamisk programmering

		r	e	g	n	a	r
	0	1	2	3	4	5	6
r	1	0	1	2	3		
ä	2	1	1	2	3		
n	3	2	2	2	2		
g	4	3	3	2	3		
n	5						
a	6						

kostnaden för att komma hit från min övre vänstra granne **substitution**	kostnaden för att komma hit från min övre granne **insättning**
kostnaden för att komma hit från min vänstra granne **borttagning**	minimum av de tre möjliga “dragen”, dvs det billigaste sättet att komma hit

UPPSALA UNIVERSITET

Editeringsavstånd: dynamisk programmering

		r	e	g	n	a	r
	0	1	2	3	4	5	6
r	1	**0**	1	2	3	4	
ä	2	1	**1**	2	3	4	
n	3	2	2	**2**	2	3	
g	4	3	3	2	**3**	3	
n	5	4	4	3	2	**3**	
a	6						

kostnaden för att komma hit från min övre vänstra granne **substitution**	kostnaden för att komma hit från min övre granne **insättning**
kostnaden för att komma hit från min vänstra granne **borttagning**	minimum av de tre möjliga “dragen”, dvs det billigaste sättet att komma hit

UPPSALA UNIVERSITET

Editeringsavstånd: dynamisk programmering

		r	e	g	n	a	r
	0	1	2	3	4	5	6
r	1	**0**	1	2	3	4	5
ä	2	1	**1**	2	3	4	5
n	3	2	2	**2**	2	3	4
g	4	3	3	2	**3**	3	4
n	5	4	4	3	2	**3**	4
a	6	5	5	4	3	2	**3**

kostnaden för att komma hit från min övre vänstra granne **substitution**	kostnaden för att komma hit från min övre granne **insättning**
kostnaden för att komma hit från min vänstra granne **borttagning**	minimum av de tre möjliga “dragen”, dvs det billigaste sättet att komma hit

UPPSALA UNIVERSITET

Likhetsnycklar

- Ord matchas mot nycklar istället för ord
- Ord som stavas på liknande sätt har likadana (eller nästan likadana nycklar)

Likhetsnycklar: Soundex

- SOUNDEX = Indexing on Sound
- Odell & Russel, 1918 (!)
- Fonetisk likhet
  - vokaler ignoreras
  - konsonanter grupperas tillsammans om de liknar varandra fonetiskt
- Användning: Flygbokningssystem (Davidson 1962)

UPPSALA UNIVERSITET

Soundex – algoritm

1. Behåll det första tecknet
2. Ersätt efterföljande tecken enligt nedan:

a, e, i, o, u, y, h, w	0
b, f, p, v	1
c, g, j, k, q, s, x, z	2
d, t	3
l	4
m, n	5
r	6

3. Ta bort alla nollor
4. Ta bort alla på varandra följande dubbletter
5. Spara de tre första siffrorna

UPPSALA UNIVERSITET

Soundex – exempel

disapont → D215
disappoint → D215

1. Behåll det första tecknet
2. Ersätt efterföljande tecken enligt nedan:

a, e, i, o, u, y, h, w	0
b, f, p, v	1
c, g, j, k, q, s, x, z	2
d, t	3
l	4
m, n	5
r	6

3. Ta bort alla nollor
4. Ta bort alla på varandra följande dubbletter
5. Spara de tre första siffrorna

Soundex – exempel

disapont → D215
disappoint → D215

Ersättningsförslag för **disapont**:
disband, disbands, disbanded, disbanding, disbandment, disbandments, dispense, dispenses, dispensed, dispensing, dispenser, dispensers, dispensary, dispensaries, dispensable, dispensation, dispensations, deceiving, deceivingly, despondent, despondency, despondently, disobeying, ***disappoint****, disappoints, disappointed, disappointing, disappointedly, disappointingly, disappointment, disappointments, disavowing*

UPPSALA UNIVERSITET

N-gramsbaserade tekniker

- Stränglikhet, dvs andelen gemensamma n-gram (vanligen trigram)

  likhet(i,j) = 2C/(n+n')

  där *n* är antalet trigram i *i*
  och
  n' är antalet trigram i *j*
  och
  C är antalet trigram gemensamma för *i* och *j*

N-gramsbaserade tekniker (forts)

Hur lika är *concider* och *consider*?

##c #co con onc nci cid ide der er# r##
##c #co con ons nsi sid **ide der er# r##**

C (antalet gemensamma trigram) = 7
n (antalet trigram i *concider*) = 10
n' (antalet trigram i *consider*) = 10

likhet(*concider*,*consider*) = 2C/(n+n') = 14/20 = **0,70**

N-gramsbaserade tekniker (forts)

Hur lika är *concider* och *cider*?

##c #co con onc nci cid ide der er# r##
##c #ci **cid ide der er# r##**

C (antalet gemensamma trigram) = 6
n (antalet trigram i *concider*) = 10
n' (antalet trigram i *cider*) = 7

likhet(*concider*,*cider*) = 2C/(n+n') = 12/17 ≈ **0,71**

N-gramsbaserade tekniker (forts)

Modifierat likhetsmått:

likhet(i,j) = 2C/(n+n') → *likhet(i,j) = C/max(n,n')*

där *n* är antalet trigram i *i*
och
n' är antalet trigram i *j*
och
C är antalet trigram gemensamma för *i* och *j*

likhet(*concider*,*consider*) = 7/10 = **0,70**
likhet(*concider*,*cider*) = 6/10 = **0,60**

UPPSALA UNIVERSITET

Samarbetsuppgift

Antag att en skribent av misstag har skrivit in *käran*. Antag vidare att ett stavningskontrollprogram har kommit fram till att det rör sig om en felstavning och att möjliga ersättningsförslag är *tjäran* och *kärran*.

Hur skulle dessa ersättningsförslag rangordnas enligt det modifierade n-gramsbaserade måttet?

Hur skulle ersättningsförslagen rangordnas i termer av editeringsavstånd?

Kommentera resultatet: Är rangordningen lika bra för alla typer av skribenter?

Skulle båda alternativen finnas med bland ersättningsförslagen om man istället hade använt sig av likhetsnycklar på samma sätt som i SOUNDEX? Varför/varför inte?

UPPSALA UNIVERSITET

Referenser

- Markus Dickinson, Chris Brew & Detmar Meurers, 2013 (kapitel 2), *Language and Computers*
- Daniel Jurafsky & James H. Martin, 2000 (avsnitt 5.1-5.6), *Speech and Language Processing*
- Karen Kukich, 1992, *Techniques for Automatically Correcting Words in Text*
- Roger Mitton, 1996, *Spellchecking by Computer*
- Stina Nylander, 2000, *Statistics and Phonotactical Rules in Finding OCR Errors*