UPPSALA UNIVERSITET

Grundläggande Textanalys VT 2016

Språkgranskning (1)

Eva Pettersson

eva.pettersson@lingfil.uu.se

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

UPPSALA UNIVERSITET

Typoglycemia

Aoccdrnig to a rseearch taem at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoatnt tihng is taht the frist and lsat ltteer be in the rghit pclae. The rset can be a taotl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Scuh a cdonition is arppoiatrely cllaed Typoglycemia. Amzanig huh? Yaeh and you awlyas thguoht slpeling was ipmorantt.

According to a research team at Cambridge University, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter be in the right place. The rest can be a total mess and you can still read it without a problem. This is because the human mind does not read every letter by itself, but the word as a whole. Such a condition is appropriately called Typoglycemia. Amazing, huh? Yeah and you always thought spelling was important.

Varför behövs stavningsstandard?

- Ofta kan vi förstå felstavade ord, dock:
  - Olika uttal i olika dialekter
  - Undvika onödiga tvetydigheter och missuppfattningar
  - Lättare att söka i olika register, lexikon etc
  - Dessutom underlättar det för maskinell analys av texten ☺

Historisk text utan stavningsnorm

besväradhe sig befallningsmannen Haanss Sivardzson, öffuer Tegelzmora boerne, som hafva så inbärgat Crononess Höö på grufmåssen, att dett mästedeels ähr förskämbt, på huilken skadha han Ratione officij protesterade. Ther till bemälte allmoge svaradhe sig så hårtt vara förbudne, att ingen skulle vijka derifrå, vidh — 40 m, förre ähn ängen bärgat blefve, och derfhöre den eene daghen slogz ängen, och den andre in emoth afton, hadhe the höett i Stack, som uthan tuifvelsmåål intet tårtt var: Blef för det såleedess interlocutorie affsagt, att Påvel Grufvefougdhe, huilkom bemälte Eng för hanss löön gifuess för — 80 daler kopparmynthe skall bemälte höö emodth tagha, så myket han kan någorleedess komma till vägha uthan sin Märkelig skadha ahntagha, och Nembningeman skall veetha beskeedh på dem som det höett bärgadhe, huilket Påvell sig till betallning anammat. Finss sedhan någodt som så illa medhfharidt ähr att han ingalunda kan det vidhertagha, så vardher Rätten förklarandess sig på dem som skäligen bevijsass kunna höet genom försumelse, och vhanrycht och genom otidigt bärgandhe fördärfvat hafva, på huilka och Nembningeman i lijka måtto, beskeedh och underrättellse vetta skall;

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

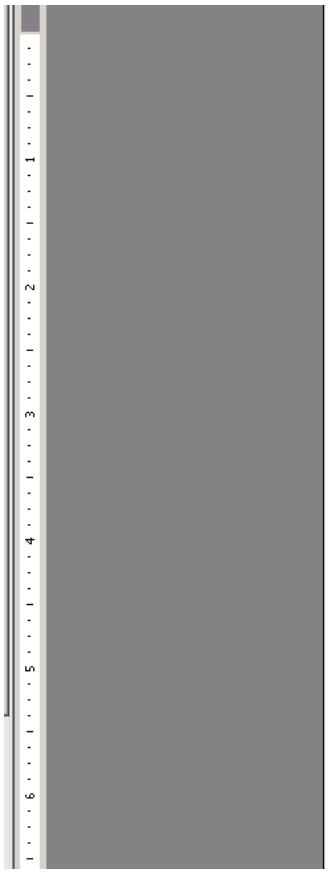
Stavningskontrollens två delar

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

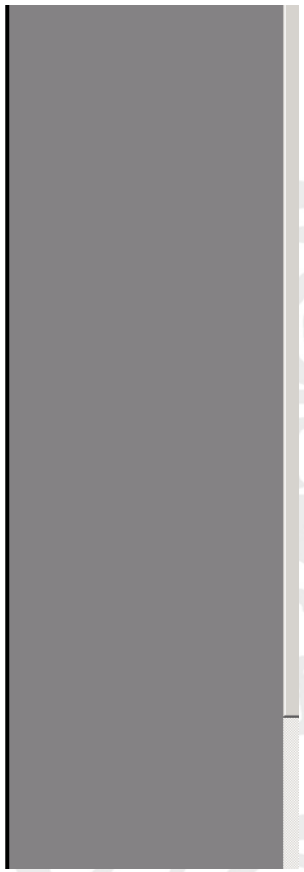
2. Felkorrigering (*error correction*)
   att ge ersättningsförslag

Inte alltid nödvändigt med ersättningsförslag. Ibland räcker det att skribenten görs uppmärksam på att det förekommer en felstavning över huvud taget.

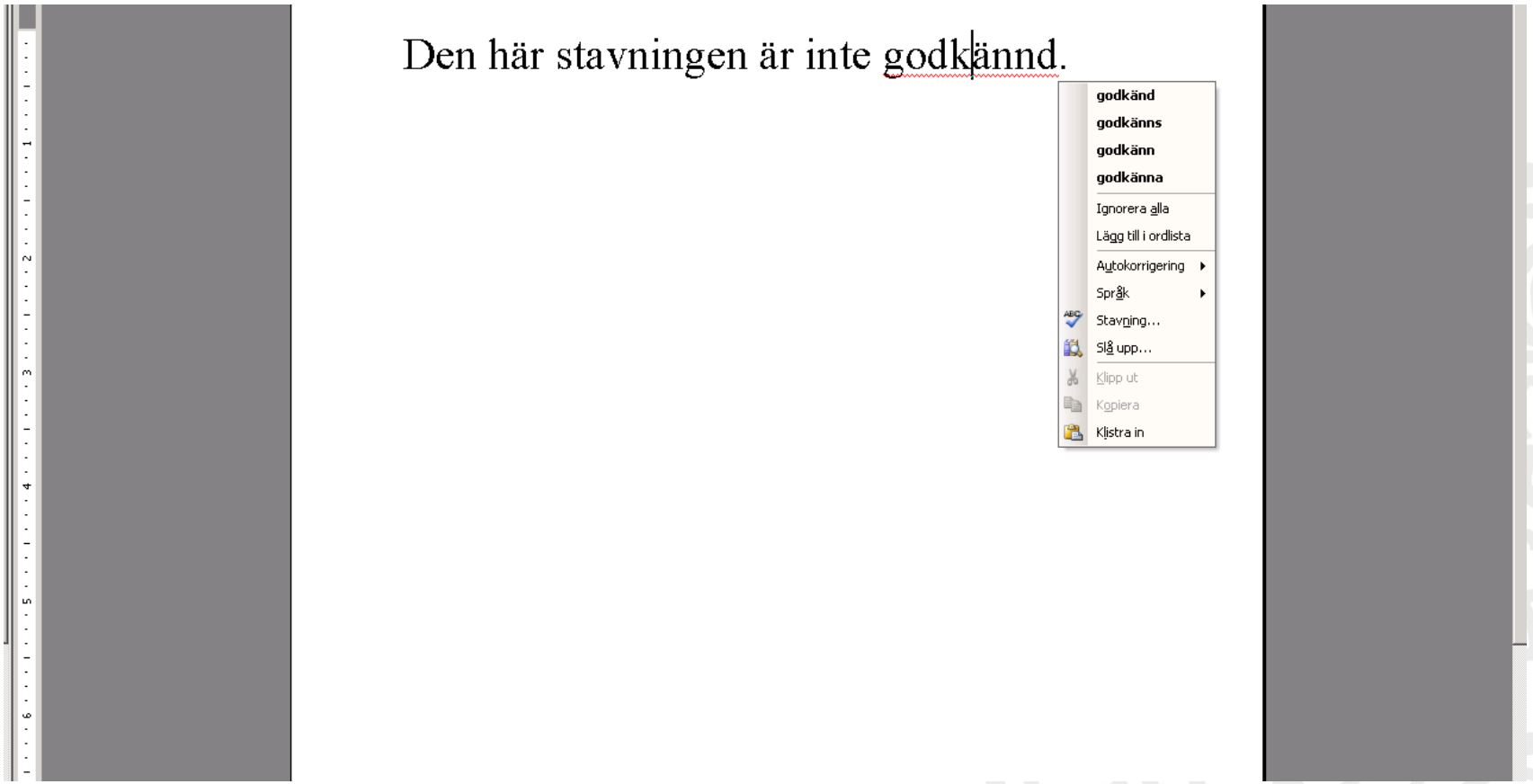
Feligenkänning i Microsoft Word



Den här stavningen är inte godkännd.



Felkorrigering i Microsoft Word



UPPSALA UNIVERSITET

FELIGENKÄNNING

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

UPPSALA UNIVERSITET

Real Word Errors

- Lokala syntaktiska fel
  - Stavfelet går att identifiera genom att titta på ett eller två ord före eller efter det felaktiga ordet
    * *Det er svårt att stava*
- Globala syntaktiska fel (long-distance)
  - För att hitta stavfelet måste man göra en mer utförlig grammatisk analys av meningen
    * *The team that hits the most runs get ice cream*
- Semantiska fel
  - Stavfelet ger upphov till ett befintligt ord som fungerar bra syntaktiskt, men där innebörden av meningen blir konstig
    * *det är svart att stava*

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 *jättetrött → jätte trött*
- 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*

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*	ko-takt
makelera	*makulera*	make-lera
medalg	*medalj*	med-alg
cykelsäll	*cykelställ*	cykel-säll
särkskilt	*särskilt*	särk-skilt

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

Approacher till felkorrigering

1. Interaktiv stavningskontroll
   - Fel detekteras och markeras medan man skriver
   - Skribenten väljer själv om åtgärd behövs
2. Automatisk stavningskontroll
   - Fel detekteras och korrigeras automatiskt, utan inblandning från skribenten

UPPSALA UNIVERSITET

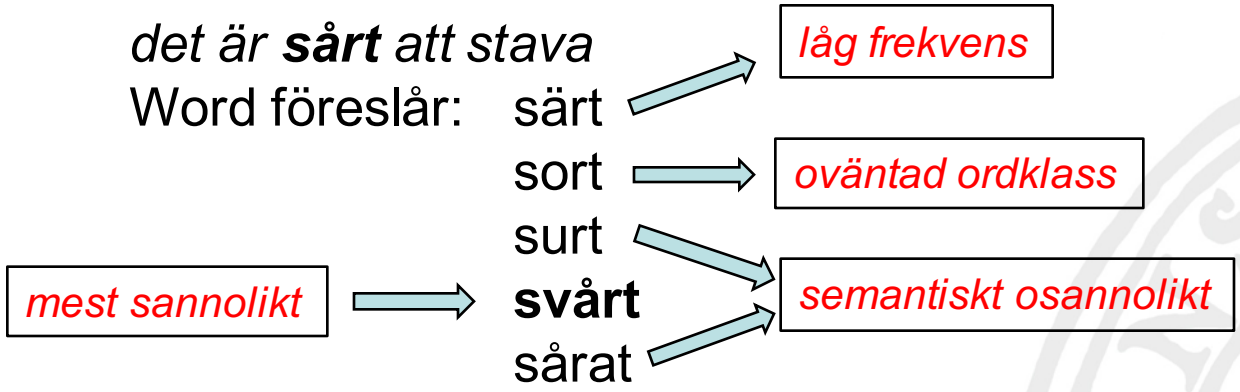
Felkorrigering i MS Word

- Interaktiv felkorrigering
  - Rödmarkering under skrivandets gång, med ersättningsförslag om skribenten högerklickar på det felmarkerade ordet
- Automatisk felkorrigering
  - Vissa vanliga ”säkra” felskrivningar autokorrigeras, såsom:

á la carte	→	*à la carte*
abbonemang	→	*abonnemang*
coh	→	*och*
kasett	→	*kassett*
affich	→	*affisch*
dublett	→	*dubblett*

Felkorrigering

Det finns många faktorer att ta hänsyn till om man vill ge användbara ersättningsförslag



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

Grundat på korpusdata producerat av skribenter som skriver på sitt modersmål och som inte har särskilda skrivsvårigheter.
Kan se annorlunda ut för exempelvis andraspråksinlärare och dyslektiker.

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) + { 0 om *i* = *j* (likhet); 1 annars (substitution) }
}

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

Editeringsavstånd illustrerat

r ä n g n a

r e g n a r

Editeringsavstånd illustrerat

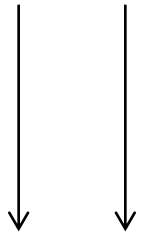
r ä n g n a

r e g n a r

Editeringsavstånd: 0

Editeringsavstånd illustrerat

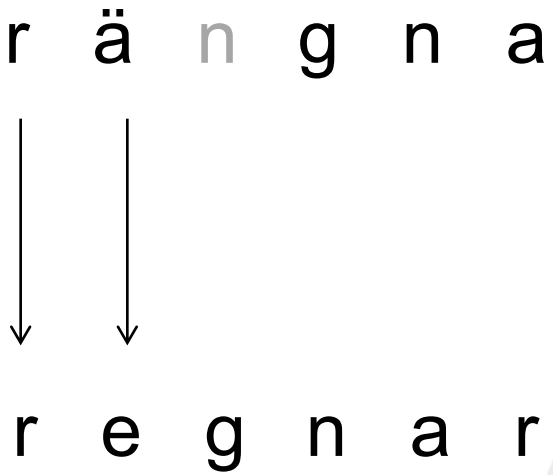
r ä n g n a



r e g n a r

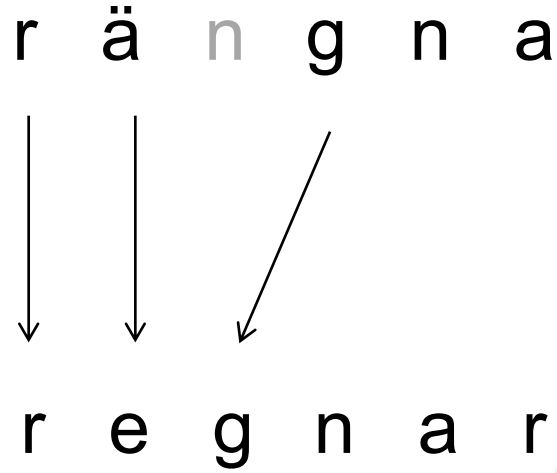
Editeringsavstånd: 1
substitution

Editeringsavstånd illustrerat



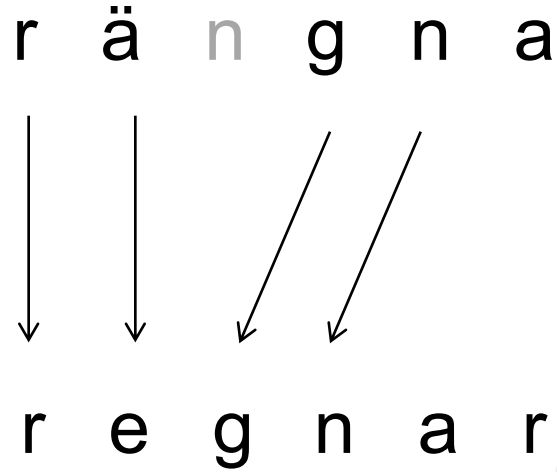
Editeringsavstånd: 2
substitution + borttagning

Editeringsavstånd illustrerat



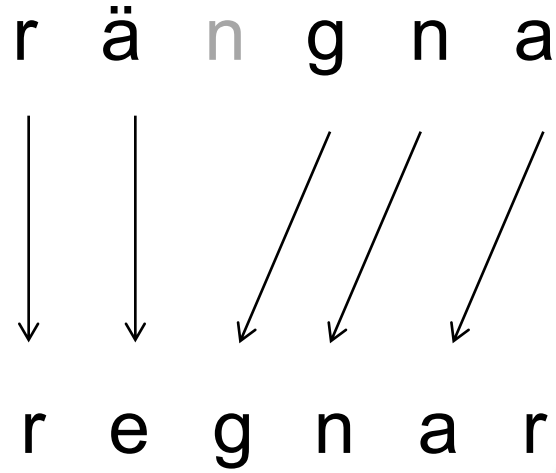
Editeringsavstånd: 2
substitution + borttagning

Editeringsavstånd illustrerat



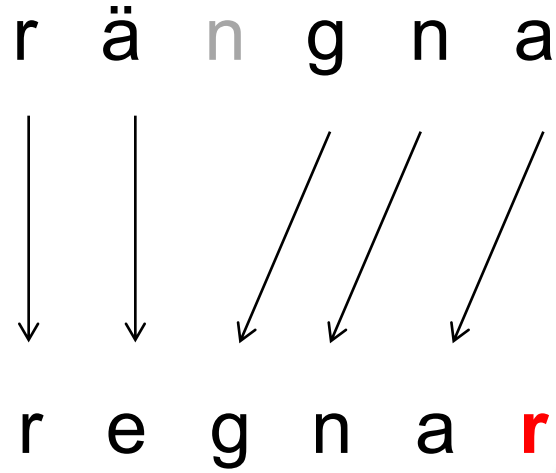
Editeringsavstånd: 2
substitution + borttagning

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						
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							
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 **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##**	concider
##c #co con ons nsi sid **ide der er# r##**	consider

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##**	concider
##c #ci **cid ide der er# r##**	cider

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

likhet(*concider*,*cider*) = 2C/(n+n') = 12/17 ≈ **0,71**
likhet(*concider*,*consider*) = 2C/(n+n') = 14/20 = **0,70**

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

Referenser

- Markus Dickinson, Chris Brew & Detmar Meurers, 2013 (kapitel 2), *Language and Computers*
- 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*