HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Språkkontrollprogram i dag och i morgon. Vad är möjligt i dag, lingvistiskt och tekniskt? Hur ser framtiden ut?

Kimmo Koskenniemi
21.4.2005

Institutionen för allmän språkvetenskap
Humanistiska fakulteten

Språkkontrollprogram i dag och i morgon

- Vad kan man redan hantera?
- Varför är det så svårt att göra allt?
- Vad kunde vi ännu göra?
- Stora, medelstora och små språk
- Hotbilder: (a) ekonomiska, (b) upphovsrättigheter
- Varför är det så viktigt?
- Hur kan vi klara oss, och vad behöver vi?

Vad kan man redan hantera?

- Människan vet instinktivt vad är rätt och vad är fel, men för en dator måste man ha exakta regler, listor och undantag - och även så, blir det varken 100 % täckning eller 100 % precision:
- **stavningskontroll** (ordbok + böjningsregler + avledningsregler + sammansättningsregler)
- **avstavning** (delning av sammansatta ord + avstavningsregler)
- **synonymer** (listor av synonymer)
- **grammatikkontroll** (morfologisk analys, frasmönster, mönster för typiska fel)

Varför är det så svårt att göra allt?

- Språket är stort – större än man vanligen tror
- Språket varierar (under tiden, geografiskt, individuellt, ...)
- Inte helt fasta regler
- Ord och syntaktiska strukturer är ofta flertydiga
- Reglerna är inexakta och betydelserna diffusa
- Språket är ett öppet system (nya begrepp, nya ord, osv. kryper in ständigt)

Vad är tekniskt möjligt?

- Datorerna är kraftiga och stora redan nu (1000 gånger större och snabbare än 20 år sedan)
- Det är svårt och tidskrävande att programmera, särskilt att bygga stora system (dvs. konsten att programmera har utvecklat långsamt)
- Datorerna behöver ännu nu konkreta och entydiga order eller kommandon (de gör vad vi säger, inte vad vi vill)
- Vad än vi kan beskriva exakt
  - manuellt med regler eller listor som man skriver ned eller
  - automatiskt genom maskininlärning

Vad kunde man ännu göra?

- Textklassificering
  - till vilket tema eller område hör texten?
  - skarpare stavningskontroll och bättre förslag av rättningar
- Ontologi
  - man kan beskriva betydelsen delvis om man samlar begrepp och deras relationer
- Språkinlärning och intelligent feedback
  - Stöd för invandrare och andra för vilka språket är deras modersmål
- Naturlig dialog mellan datorer och användaren
- Bättre kontroll av läsbarhet och stil

Stora, medelstora och små språk

- Det är ungefär lika svårt att göra program för avstavning, stavningskontroll osv. för små och för stora språk.
- För stora språk finns det färdiga språkresurser till hands (korpusar, maskinläsbara ordböcker, annoterade korpusar, osv.)
- För små och medelstora språk saknas många resurser.
- För världsspråk produceras moduler för språkkontroll kommersiellt (och det är lönsamt)
- För små språk måste samhället betala allt
- För medelstora måste samhället ta del på något sätt (finansiera forskning, samla språkresurser, osv.).

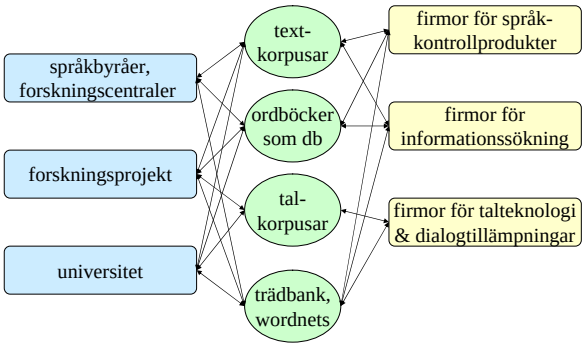
Hur ser framtiden ut? Dystert?

- De största språken är kommersiellt viktiga och lösningar dyker upp även om man bara väntar
- Detsamma går inte med små språk, t.ex. med isländska, grönländska eller nordsamiska
- Mellanstora språk, som finska, svenska osv. ligger däremellan
- Många länder (t.ex. Estland, Island) har konkreta nationella program för att stöda nationella språk, men inte alla

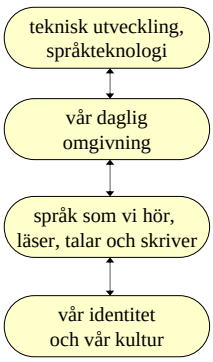
Plan för samarbete

- Vissa språkliga resurser skall vara öppna, dvs. tillgängliga till alla
  - textkorpusar, ordböcker, synonymlistor, talkorpusar, s.k. trädbank
  - Open Source licens och öppna format (XML)
- Den offentliga sektorn skall ta ansvar för att bygga språkresurser - och forskare och näringslivet kan delta
- Man skulle starta flere projekt för att skapa språkresurser som vi ännu saknar (finansierade av EU, nordiskt eller nationellt)

Hur?



Varför?



Språkkontroll för andraspråksskribenter

Microsoft Word 2000 – Svefix 1.0

Material

- 18 texter skrivna av 9 studerande vid Vasa universitet (finska som modersmål)
- besöksrapport / reserapport
- 6136 ord
- 428 ord - 879 ord

Tillvägagångssätt

- Kategorier:

1) syntax
2) ordlära & morfologi
3) rättskrivning & interpunktion
4) stil

Manuell granskning av materialet

- sammanlagt 447 anmärkningar
- syntax (52 %): species, ordföljd och prepositioner
- rättskrivning och interpunktion (24 %)
- Stil (16 %): ordval
- ordlära & morfologi (9 %): genus

Falska alarm

- Word 2000: 42 %
- Svefix: 23 %
  - namn: *Korsholm, Laihela,Veera, Häggman*
  - *stordia,vasabo – vöråbo*
  - inget verb (…”*att jag måste hinna till bussen. Bussen till Ikea.”)*
  - *tecknet* :(

Syntax

- vissa ordföljdsfel (a), speciesfel (b) och kongruensfel (c)
- Word 2000: # konsekvent
- Svefix: bättre
- Verbböjning
  - *+ inredar > inreder*
  - *- hurdant liv som lade bakom*

Ordlära & morfologi

- Genusfel
  - obestämd artikel + substantiv (a) samt ord som i singularis har böjts enligt fel genus (b)
- Svefix: bättre
  - numerus (*den andra försöken)*
  - pluralformer (*ett tjugotal ivriga studeranden*)

Rättskrivning & interpunktion

- fungerar bäst
- ger alternativ; Svefix mångsidigare
- hittar inte allt
- datum på svenska
- klarar inte av att kommentera interpunktion utöver punkter efter siffra & mellanslag

Stil

- Word 2000:endast en gång (*måtte*)
- Svefix: Fortelius (2003): ”finlandismkontrollen i Svefix är omfattande och visar på ett gediget avvägningsarbete”
- *stan*, *va*, *dom*

Jämförelse mellan mänsklig bedömare och dator

- 68 – 25
- 5/3 falska alarm: namn
  - 20/22 relevanta
- korrigeringarna överensstämmer i 14 (W)/16 (SF) fall = 20 %

Syntax

- 38 – 3
- bestämd form efter *detta*
  **detta beslutet*
- adverbialets placering i två bisatser
  **...att staden skulle fortfarande...*
  **... att historia har aldrig varit...*
- # determinativ konstruktion, *bli* + yrkesbeteckning ordföljden i en huvudsats efter en bisats, indirekta frågesatser som förutsätter *som* samt prepositioner

Ordlära

- 11 – 6 (W)/8 (SF)
- 4 av 7 genusfel och båda böjningsfelen kommenteras.
- Word hittar inte: *den andra försöken, en arbetsrum*
- Svefix hittar inte: adjektivböjning (1), löst sammansatt verb

Rättskrivning

- Bra
- 8 – 5
- Tryckfel: *kalade, bran, invänarna, vat*
- Datum som skrivits med punkt (*den 28. december*)
- Missar
  - formen *on* som använts i stället för prepositionen *om*.
  - felaktig användning av komma
  - nationalbeteckning som skrivits med stor bokstav (*Rysk*)

Stil

- Word 2000: inga kommentarer
- Svefix*: Arbis*
- Ordval: *ledning – regering, lektion – föreläsning, både – båda*)
- Fraser: *tuottaa tulosta, käyttää paljon aikaa*

Automatisk kontroll:

- mellanslag (4 fall)
  - siffra och procenttecken (11 %)
  - *före detta*
- skiljetecken (2 fall)
  - *fel..man*

Slutord

- datoriserad språkkontroll är ett användbart hjälpmedel för blivande språkexperter
- möjligheter – begränsningar
- Svefix 1.0 är bättre än Word 2000 i fråga om:
  - namn
  - ordföljd, kongruens
  - stavningsförslag
  - stilkontroll

Tabell 1. Noterade fel vid manuell genomgång av texterna.

Ord (428-870)	Markeringar	Syntax	Ordlära & morfologi	Rättskrivning & interpunktion	Stil
6136	**447**	**230**	**40**	**105**	**72**

Tabell 2. Meddelanden om fel i de undersökta texterna.

	Meddelanden	Ej förslag	Påpekande	Förslag	Förslag + info
Word	**186**	**71**	**8**	**67**	**40**
Svefix	**207**	**41**	**-**	**67**	**97**

Ex. 1	a) ...eftersom det har nu gått 150 år...> det nu har gått...(T3/1) b) Jussi blev alkoholiserad och hans självkänslan var...> självkänsla (T9/1) c) Namnet var inte heller klar. > klart (T1/1)
Ex. 2	a) ...och också de som var inte intresserade av opera...(T1/2) b) ...som skulle också komma med...(T2/9)
Ex. 3	a) På denna dag höldes första mötet (T1/7) b) Alla vad vana vid gamla namnet Vasa (T1/7)
Ex. 4	a) ...var nästan en begrepp ...(T1/1) b) Teman var mycket populärt och salet ...(T4/1)

Ex. 5	a) tycte > tyckte, tyste, tycke (T1/1) b) svenskaspråkiga > svenskspråkiga (T6/1) c) Jag var så tröt att...(T2/2) d) ...de behövde ett tag över huvudet. (T4/1)

Ex. 6	a) Vasa stad brann den 3. augusti 1852. (T1/1) b) ...till exempel den 2. juni var ... (T1/1) c) Den 24. September 2002 berättade ...(T5/1)

Ex. 7	a) Före vi stannade i Tammerfors ...(T2/2) > *innan/ förrän* b) Det tog inte länge ... (T1/7) > *ta lång tid, dröja länge* c) Lite efter Tammerfors höll vi en paus... (T1/7) > *ta paus, ha paus* d) åldringar > *äldre (person, människa), gamla, pensionär*

Språkkontroll för andraspråksskribenter

Ola Knutsson
KTH Kod, Stockholm

CrossCheck
Språkliga datorstöd och andraspråksinlärning

Språkkontrollens roll i andraspråkskribenters skrivande

- Anpassning till normen? Eller uppnå kommunikationsmål?
- Fokus på form
- Ett verktyg bland flera andra
- Lära sig mer om språket eller att lyckas med skrivuppgiften.

Hur skiljer sig svenska som andra-språk från "normalt språkbruk"?

- Felanalys är ett svårt område: bör felen tolkas som de uppträder eller i ljuset av vi tror skribenten avser?
- Storskaliga korpusundersökningar och jämförelser saknas
- En hel del fel är gemensamma.

Räcker inte reglerna till?

- All grammars leak (Sapir, 1921)
- Kan man hitta alla fel i en text: * *? ?? ?
- Hur skall man analysera en text som är full av fel?
- Bygger inte de flesta språkmodeller på korrekt språk?

Tre metoder för grammatikkontroll

- Granska – handskrivna regler som söker efter på förhand kända feltyper.
- ProbGranska – "modellering" av lokal grammatikalitet med hjälp av statistik.
- SnålGranska – regler skapas automatiskt med hjälp av konstgjorda fel.

•Jämförelse på 10000 ord från SSM-korpusen

	Word	Granska	Prob	Snål	Någon Granska	Totalt
# fel	392	411	102	121	528	592
# falska	21	13	19	19	48	-
# stavfel	334	293	35	26	314	363
# falska stavfel	18	5	-	-	5	-
# grammatikfel	58	118	67	95	214	229
# falska grammatikfel	3	8	19	19	43	-

6

Par		Båda	Endast Granska	Endast Prob Granska	Endast Snål-Granska	Någon Granska (i paret)
Granska+ Prob-Granska	# Fel	17	101	50		168
	Falska	0	8	19		27
Granska+ Snål-Granska	# Fel	44	74		51	169
	Falska	3	5		16	24
Prob-Granska + Snål-Granska	# Fel	11		56	84	151
	Falska	0		19	19	38

7

Granska: slutsatser

- **Fördelar:**
- God kontroll över diagnoser och ersättningsförslag
- "buggar" kan mer uppenbart åtgärdas
- **Nackdelar:**
- Kräver mycket manuellt arbete
- Svårt att kalibrera för olika nivåer på granskningen.

8

ProbGranska: slutsatser

- **Fördelar:**
- Metoden är bra på att identifiera fel som är svåra att skriva regler för.
- Enkelt att ställa in olika tröskelvärden
- **Nackdelar:**
- Ger varken diagnos eller ersättningsförslag.
- Hög precision ger mycket låg täckning.

SnålGranska: slutsatser

- **Fördelar:**
- Begränsad manuell insats
- En taggare per feltyp kan ge bra diagnos och även ersättningsförslag (åtminstone för särskrivningar).
- **Nackdelar:**
- Varje feltaggares enskilda falsklarm kan resultera i många falsklarm om man sätter ihop dem i ett verktyg.

Flera verktyg är bättre än ett

- En ensemble av granskare ger fler möjligheter, t.ex. genom röstning
- Andra verktyg är också viktiga för andraspråksinlärare: sökbara lexikon, konkordans-motorer, generell grammatisk information, ordböjning --> Grim
- Flera olika verktyg kompenserar bristerna hos de enskilda verktygen. Verktygslådan ger också en mer mångfacetterad bild av språket
- Olika synsätt: normativa, liberala, kreativa och språkbruksbaserade

11

- Hur kan vi arbeta med mer liberala språkmodeller?
- Skulle mer generella angreppsätt på grammatikalitet vara gångbara? Finns det några nya angreppsätt i sikte?
- Vilka är de vanligaste felen bland andraspråks-talarna? Vilka fel är viktiga att rätta för inlärar-språkets utveckling?
- Hur påverkar automatisk språkkontroll språk-brukarnas syn på språk och språkinlärning?
- Hur kan vi utvärdera bättre?
- Nu: ett språk och en språkbrukare. Framtid:

12

Mer information

- Språkmiljön Grim: www.nada.kth.se/grim
- Projektet CrossCheck:
- www.nada.kth.se/theory/projects/xcheck
- Språkliga datorstöd och andraspråksinlärning:
- www.nada.kth.se/~knutsson/call-en.html

13

Skrivstöd för barn

Robin Cooper, Ylva Hård af Segerstad
och Sylvana Sofkova Hashemi

Institutionen för lingvistik
Göteborgs universitet

Projektet "Att lära sig skriva i IT-samhället"

- Syftet:
  - Undersöka hur och till vad barn i skolåldern använder skrift
  - Hur skrivstöd kan anpassas bättre till barns skrift och skrivande
- Textmaterial:
  - Texter av olika typer (fri berättelse, brev, rapport, saga...) skrivna av 160 barn i skolår 4-8 (dvs. 10-15 år)
  - Skrivna i skolan och på fritiden
  - Med papper + penna, ordbehandlingsprogram, mobiltelefon

Skrivprogram och skrivstöd

- Ordbehandling på dator (både i skolan och hemma)
  - uteslutande *Microsoft Word*
  - stavnings- och grammatikkontroll är ALLTID påslagen

Skrivstöd anpassat till barn saknas

- Baserade på vuxentext
  - offentliga (professionella) texter
  - vuxnas skrivproblem
- Inte anpassade till barn
  - barn gör många fler fel
  - felen är av annan art och komplexitet
  - texterna som helhet ser annorlunda ut
  - stödjer inte skrivutvecklingen

Anpassning efter målgruppens skrivande

- Stödprogram behöver fungera bra för barn
  - Identifiera barns skrivproblem (stavning och grammatik)
  - Ge förklaringar på ett sätt som barn förstår

Ett planerat projekt: "SkrivIT"

- Skrivstöd för barn med svenska som första- eller andraspråk
- IKT som pedagogiska verktyg:
  - Hur kan"alternativa skrivmiljöer" användas som stöd i skrivutvecklingen?
    - T.ex. blog som verktyg för reflektion, chatt som undervisningsmiljö

Grammatikfel hos barn och vuxna

- Skillnad i frekvens
  - Vuxna gör ca 1 grammatikfel per 1.000 ord
  - Barn gör ca 9 grammatikfel per 1.000 ord
- Skillnad i typ av fel
  - **Vuxna:** extra insatta/missade ord, kongruensbrott i substantivfraser och ordval
  - **Barn:** brott i finit verbform, missade ord i satser och ordval
  - finit verbformsfel ca 8 gånger vanligare hos barn

Grammatikkontroll performans

- Vuxentexter:
  - Befintliga skrivstöd täcker i genomsnitt 58% (tidningstext, studentuppsatser)
- Barntexter:
  - Befintliga skrivstöd täcker i genomsnitt 12% i barntexter
  - Detta trots att många feltyper finns definierade hos befintliga grammatikstöd

FiniteCheck

- Ny metod: beskriver korrekt språkbruk
  - mindre antal regler
  - inte nödvändigt att tänka ut möjliga fel
- Baserat på barntexter
- Täcker inte många feltyper ännu
- Jämfört med andra svenska verktyg
  - bra resultat på både barn- och vuxentexter
  - motiverar behovet för anpassning till barn

Hur skriver barn?

- Stavning
  - ca 8 felstavade ord per 100 ord
- Ordsegmentering (isär- och ihopskrivning)
  - ca 2 ord per 100 ord
- Stora bokstäver och interpunktion
  - satsradning, felplacerad interpunktion
- Talspråkspåverkan
  - talspråksvarianter, direkt anföring, ljudhärmning
- Grammatik
  - verbform, extra insatta/missade ord, ordval

Exempel på barntext

*så här **börja** det jag var på mitt land och **bada** då var jag liten plötsligt kom en snok ifösej så hiugger inte snokar i vatten men jag blev allafal **jätte** **räd** för jag kunde inte simma då och snoken jagade mig länre och längre ut då ko min bror med en **gumi** **båt** och tog upp mig då blev jag **jätte** **glad***

(10 år)

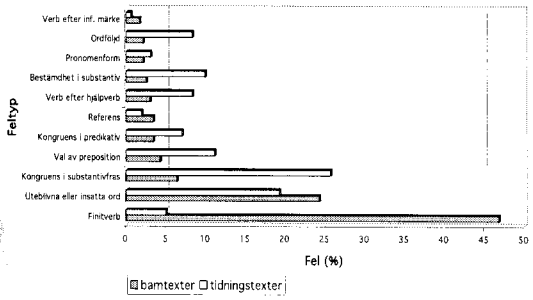
Grammatik hos barn

- Finit verbform:
  `Hon fråga vad det var för nåt.`
- Extra insatta/missade ord:
  `Gunnar var på semester (-) Norge.`
- Ordval:
  `vi var väldigt lika på sättet`
- Kongruens i substantivfras:
  `Det var en räkningen på deras lägenhet.`
- Verbform efter hjälpverb:
  `Men kom ihåg att det inte ska blir någon riktig brand.`

Andra grammatikfel hos barn

- Kongruens i predikativ komplement
- Bestämdhet i enkla substantiv
- Pronomen kasus
- Verbform efter infinitivmärke
- Utelämnat hjälpverb/infinitivmärke
- Ordföljd
- Referens

Barn vs. vuxna



Hur barn använder stavnings- och grammatikkontroll

- Alltid påslagen, både i skolan och hemma
- Inställt på kontroll medan man skriver
- Uppenbara fel rättas direkt, annars väljer man bland alternativen
- De ignorerar
  - när korrekta ord markeras
  - vid osäkerhet
  - vissa låter markeringen stå kvar
- De lägger sällan till egna ord i lexikonet

Stavningskontroll

+ Hjälper till att hitta:
  - stavfel som bildar nonsensord
  - ihopskrivningar

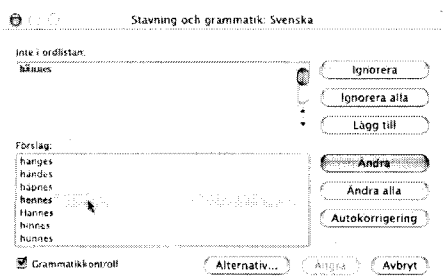
– Hanterar inte:
  - stavfel som sammanfaller med existerande ord
  - grova/fonologiska stavfel listas långt ner eller inte alls

Hur hanteras isärskrivningar?

- Barnen chansar eller ignorerar

Många stavfel kvar!

Stavningskontroll



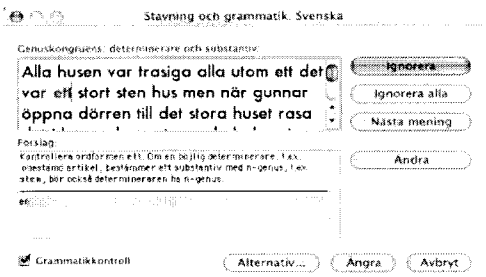
Grammatikkontroll

+ Hjälper till att hitta:
  - teckenfel
  - inledande versal
  - vissa enkla grammatiska fel

– Har svårt med:
  - talspråksformer
  - verbfel, t.ex. reduktion av verbändelser
  - andra grammatiska fel (enkla, komplexa)

- Ger obegripliga förklaringar/kommentarer
  - ignoreras oftast av barnen

Grammatikkontroll



FiniteCheck: teknik

- Nätverk av finite-state transducers
- Xerox Finite State Tool (kompilator)
- UNIX, Emacs
- Positiva grammatikregler med olika detaljnivåer
- Subtrahering av grammatiker som språkgranskning (Karttunen, L. et al. 1997)

Felgranskning

- Subtraktion av *narrow grammar* från *broad grammar*

```
define VCerror ["<vc>" [VC - [VC1 | VC2]]
"</vc>"];

Men <vp><vpHead>kom ihåg
</vpHead></vp> att <np>det </np><vp>
<vpHead> inte <Error verb after Vaux>
<vc> ska blir </vc> </Error>
</vpHead><np> någon <ap> riktig </ap>
brand</np> </vp>
```

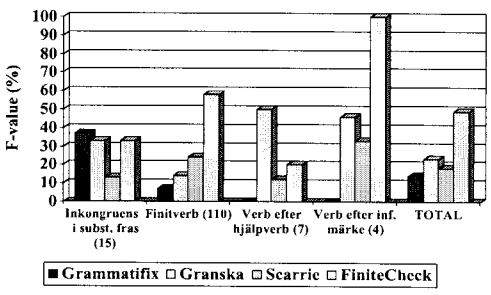
Täckta feltyper

- Kongruensbrott i substantivfraser
  `Det var en räkningen på deras lägenhet.`
- Finit verbform
  `Hon fråga vad det var för nåt.`
- Verbform efter hjälpverb
  `Men kom ihåg att det inte ska blir någon riktig brand.`
- Verbform efter infinitivmärke
  `Glöm inte att stäng dörren.`

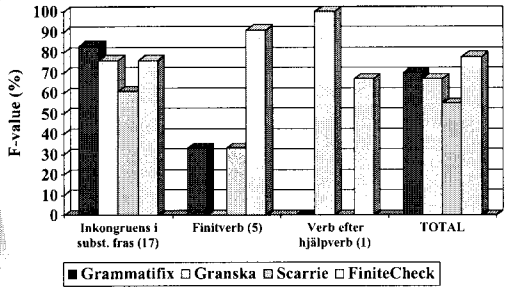
Performanstester

- Barntextkorpusen (träningsdata)
  - 134 hand- och datorskrivna texter
  - 58 skolbarn
  - 9-13 år
  - 29 812 ord (3 373 ordtyper)
- Vuxentext
  - kort berättelse
  - *Granskas* demotext
  - 1.070 ord

Performans: barntexter



Performans: vuxentext



Om projektet

"Att lära sig skriva i IT-samhället":

http://www.ling.gu.se/~sylvana/SkrivaIT/

Kontakt:

Sylvana Sofkova Hashemi
E-post: sylvana@ling.gu.se
Tel. 031-773 1175

Ylva Hård af Segerstad
E-post: ylva@ling.gu.se
Tel. 031-773 4532

Språkteknologiske verktøy fra LingIT

Torbjørn Nordgård

Hvem er LingIT?

- Holder til i Trondheim
- 5 ansatte, derav 3 i full stilling
- Basis i NTNU
- Eiere er ansatte ved NTNU, SIM (lokalt investeringsselskap med privat og statlig kapital) og enkelte andre privatpersoner

LingITs produkter

- Skrivestøtte for dyslektikere
  - LingDys
  - LingRight
- Skrivestøtte for nordmenn som skal skrive engelsk
  - LingRight
- Spørresystem med naturligspråklig grensesnitt

LingDys

- Forprosjekt 1999/2000 med støtte av NFR
  - Resultat: Demonstrator og en del entusiastiske fagfolk
  - Kommersialisering i ny bedrift – LingIT AS
- Stavekontroll
- Oppslag i ordbok (hva betyr ordforslagene)
- Talesyntese (spesielt for tegnsetting)
- Ordprediksjon
- Dialekttilpasning av stavekontroll
- Begge de norske skriftstandarder
- Fungerer i Office-pakken til MS
- OpenOffice-støtte fra ca mai 2005

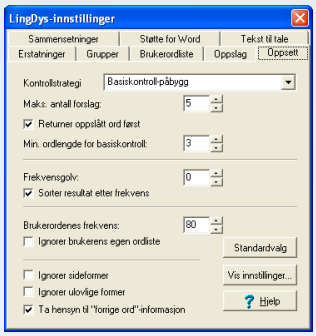
LingITs stavekontroll

- FSA-teknologi i bunnen
- Tilleggsfunksjonalitet
  - Dialekttilpasning
  - Regler som brukerne selv kan definere
  - Legge til egne ord
  - Ulike søkestrategier
  - Manipulere sammensetningshåndtering
  - …
- Salgsargument: stavekontrollen er mer velegnet for målgruppen enn stavekontrollen i for eksempel MS Word

Rammebetingelser

- Nesten alle benytter Microsofts produkter
  - Word
  - Outlook
- API-ene til MS legger begrensninger på hva man kan gjøre, og hvordan man kan lage løsninger
  - Ulike API-er for Word, Outlook, Excel, Powerpoint, …
- Staving, grammatikk, oversettelse er ulike API-er i Word
- Egne grensesnitt utenpå Office-rammene er et alternativ, men er kanskje litt teknisk risikabelt

Demo


LingDys-innstillinger
Sammensetninger
Støtte for Word
Tekst til tale
Erstatninger
Grupper
Brukerordliste
Oppslag
Oppsett
Kontrollstrategi:
Basiskontroll-påbygg
Maks. antall forslag:
5
Returner oppslått ord først
Min. ordlengde for basiskontroll:
3
Frekvensgolv:
0
Sorter resultat etter frekvens
Brukerordenes frekvens:
80
Ignorer brukerens egen ordliste
Standardvalg
Ignorer sideformer
Ignorer ulovlige former
Ta hensyn til "forrige ord"-informasjon
Vis innstillinger...
Hjelp

Produktutveckling av språkgranskningsprogram

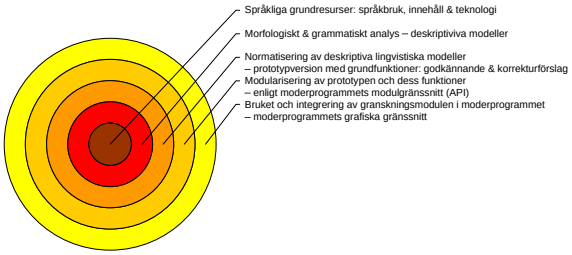
Antti Arppe
Helsingfors universitet
den 21 april, 2005

Innehållet

- Utvecklingsprocessen
- Utvecklingsprinciper
- Resultat
- Begränsningar
- Framtida utvecklingsmöjligheter

Utvecklingsprocessen

- Lökmodellen



Utvecklingsprocessen - lökmodellen

- Traditionella och modernare språkliga grundresurser
  - Textkorpora, felkorpora, nyordlistor
  - Grund-, synonym och terminologiordböcker, grammatikbeskrivningar, stilmanualer, lexikala databaser (t.ex. WordNet) eller ontologier
  - i form av pappersböcker, elektroniska ordbehandlingsprogramfilar (WordPerfect/Word) eller databaser
- Morfologiskt eller grammatiskt analys
  - Känner igen brett alla möjliga olika strukturer, dvs. är deskriptiv, men är inte på första hand normativ, dvs. kan vara övergenererande
  - I fall av Lingsoft tvånivåmodellen (TWOL) för morfologi och restriktionsgrammatik (CG) för disambiguering och ytgrammatiskt analys

Felbeskrivningar

401 Gender: masculine – feminine:
(EE/JB)

Examples:
1. den trötte flickan
2. det slöe teveprogrammet

Detection (method/accuracy/impact): (**CG/high/medium**)
1. UTR-MASC + <FEM>
2. UTR-MASC + NEU

Correction:
1-2. A UTR-MASC DEF SG => A UTR DEF SG

Corrected examples:
1. den trötta flickan
2. det slöa teveprogrammet

Analysprogram

```
kivi$ tw-swe -##
 SWETWOL 951027
Copyright (C) Lingsoft, Inc. 1994
 TWOL 1999/04/23
Copyright (C) K. Koskenniemi and Lingsoft, Inc. 1983-1999
Two-Level Compiler
Copyright (C) 1994, Xerox Corporation. All rights reserved.
.. save file loaded
komission
"<komission>"
        "ko#mission"  <N> # N UTR INDEF SG NOM
 tiedostonhallinta
"<tiedostonhallinta>"
        "tie#dos#ton#hall#inta"  <NUM> # <N> # <N> # <N> # V ACT IMP
        "tie#dos#ton#hall#inta"  <NUM> # <N> # <N> # <N> # V ACT INF
 ytgrammatiskt
"<ytgrammatiskt>"
        "yt#grammatisk"  <N> # A NEU INDEF SG NOM
        "yt#gram#matisk"  <N> # <Measure> <N> # NDER-isk A NEU INDEF SG NOM
språkgranskningverktyg
"<språkgranskningverktyg>"
språkgranskningsverktyg
"<språkgranskningsverktyg>"
        "språk#granskning#verktyg"  <N> # VDER/-ning <N> # N NEU INDEF SG/PL NOM
        "språk#granskning#verk#tyg"  <N> # VDER/-ning <N> # <N> # N NEU INDEF SG/PL NOM
```

Lökmodellen …

- Normatisering av deskriptiva modeller → prototyp
  - Stavningskontroll:
    - Begränsningen av öppna egenskap i modeller vilka gör möjligt deras höga igenkännande
    - Sammansättning : kompensering av förminskning i igenkännande genom listande av mest frekventa sammansättningar
  - Grammatikkontroll:
    - Övergång från utgångspunkt med antagande av text som rättskrivet till tvivlande gällande textens korrekthet
    - En mening med en felaktig struktur kan ha ett teoretiskt möjligt, grammatiskt analys
    - Det är relativt lätt att skapa regler för att beskriva felaktigheter i olika grammatiska strukturer, med den största utmaningen är att försäkra att ordsekvensen under granskning egentligen hör till strukturen i frågan

Prototyp och dess funktionaliteter

```
kivi$ swespell
Orthografix for Swedish - Speller Demo
Copyright (C) Lingsoft, Inc. 1996-1997
Loading /usr/local/lib/lingsoft/orthografix/lssp_sv.lex...
rw-r----- 2033/1007 2053919 Jun 30 14:48 1997 speller.sav
rw-rw---- 2033/1007   3850 Jun 30 14:48 1997 cancase.bts
rw-rw---- 2033/1007   8998 Jun 30 14:48 1997 firstsugg.bts
rw-rw---- 2033/1007   9613 Jun 30 14:48 1997 presugg.bts
rw-rw---- 2033/1007   7342 Jun 30 14:48 1997 postsugg.bts
rw-rw---- 2033/1007   4378 Jun 30 14:48 1997 lastsugg.bts
> språkgranskningverktyg
språkgranskningverktyg ( "språkgranskningsverktyg" )
>
```

Lökmodellen …

- Modularisering av granskningsverktygsprototypen
  - Implementering av prototypens språkliga funktionaliteter enligt moderprogrammets gränssnitt (*Application Program Interface* – API)
  - Betyder programmeringsarbete och omfattande testning i samband med moderprogrammet
- Bruket av granskningsmodulen som en integrerad komponent av moderprogrammet (t.ex. ordbehandlare)
  - Påverkat i stor grad av design och implementering av de grafiska gränssnitten som moderprogrammet innehåller
  - Språkprogrammens natur som inbäddade moduler, som av slutanvändarsynpunkt kan och borde se ut som en sömfri komponent av moderprogrammet
  - Det kan finnas praktiska begränsningar i gränssnittet gällande hur mycket språklig hjälpinformation kan erbjudas till brukaren, t.ex. infobubblan i Microsoft Words grammatikkontroll

Modularisering enligt moderprogrammets API-gränssnitt

The engine versions 1.x support the following functions as specified in the CSAPI 3.0, unless otherwise noted below:

SpellerVvvvvv

SpellerKkkk

SpellerSssOoooooo

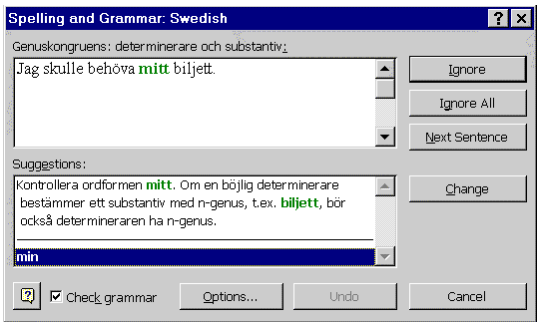
The following options are supported:

[...]

SpellerGggOoooooo

[…]

Bruket som en inbäddad modul av moderprogrammet



Utvecklingsprinciper

- Medvetande av det tillgängliga teknologins starka och svaga punkter samt dess begränsningar
  - om tvånivåmodellen är stark i analys av sammansatta ord, det kan leda till det att med den man genererar kanske alldeles för effektivt morfologiskt möjliga korrigeringsförslag som verkar konstiga för brukare utan lingvistisk utbildning
  - om restriktionsgrammatikens formalism bygger på ordens närkontekst och därmed är mest effektiv i att känna igen fraser i ytgrammatiska strukturer, det löner sig inte att använda den för att känna igen fel som berör längre avstånd.

Utvecklingsprinciper …

- Undersökning och förståelse av fenomenet själv (rättstavnings- och grammatiska fel) och deras orsaker och källor
  - forskning av skrivandet på dator har visat att man gör särskilt vissa tryckfel som är svåra att beteckna med bara ögonen, t.ex. överflödiga mellanslag (Severinsson-Eklundh)
  - klipp-och-klistra –metoden och att det är lätt att editera text – då man inte behöver skriva sekventiellt och därmed tänka meningar igenom innan man skriver dem – kan leda till att man inte orkar hela tiden kontrollera ordens grammatiska korrekthet eller helt enkelt inte bara ser sådana fel.

Utvecklingsprinciper …

- Undersökning av vilka är typiska fel i just det språk för vilket man skapar språkkontrollprogram
  - felkategorier som är typiska för engelska, t.ex. kongruens mellan subjekt och objekt, gäller inte för alla andra språk, t.ex. skandinaviska
- Fokusering pga. egentligt språkbruk, utgående från korpora:
  - då ser man vilka äkta fel skrivarna gör och i hurdana varierande strukturella kontekster dessa fel görs (med vilket man kan kontrollera att man upptäcker äkta fel och inte korrekta strukturer)

Resultat - stavningskontroll

- En balans mellan igenkännandet av rättskrivna ord med hjälp av sammansättningssregler och felskrivna ord som kan tolkas som sammansättningar
- Begränsning av sammansättning i fall av korta ord, t.ex. svenska *ko, te, ton, vis* och *å,* antingen överhuvudtaget eller i vissa positioner i sammansättningar
- Som kompensation av det förminskade igenkännandet listande (pga. korpora) av de mest frekventa sammansatta ord som består av dessa begränsade komponenter, t.ex. *is#te*

Resultat - grammatikkontroll

- Val av feltyper i vilka antalet falska larm är högst 30% i äkta språkbruk (utvecklingskorpus)
- Feltyper i svensk grammatikkontroll
  - Nominalfraser
  - Subjekt–predikativ komplement-strukturer
  - Verbkedjor
  - Ordföljd i bisatser

Begränsningar

- människans språk är mycket mer mångsidigare än man någonsin kommer att tänka sig då man startar utvecklandet av ett språkkontrollverktyg
  - det räcker med bara några CG-regler för att beskriva en felformerad nominalfras i svenska
  - men det behövs tiodubbelt mera regler för att försäkra jämfört med konteksten att det egentligen är en nominalfras som man granskar
- stavningskonventioner och ordförråd i människans språk innehåller i praktiken många undantag och är i allmänhet inte fullständigt eniga konsekventa system

Begränsningar i praktiken

- Lingvistiskt analys baserat på strukturella drag hittar inte semantiska eller pragmatiska konstigheter
  - *En sol skiner* [solen i något solsystem]
  - *Solen skiner* [solen i vårt solsystem]
- Grammatiska fel kan leda till strukturellt möjliga analys, vilka kan vara väldigt svåra att detektera ← frasgränser
  - *Finns det [slutgiltiga siffror] någonstans?*
  - *Finns [de slutgiltiga siffrorna] någonstans?*
- Fel i komplicerade eller sällsynta strukturer kan kvarstå odetekterade pga. den valda utvecklingsstrategin
  - granska ord bara då man är säker att hör till samma struktur
  - *De i stadgarna, paragraf 4, nämnda ärendet om kallelse till hedersordförande behandlas.*

Framtida utvecklingsmöjligheter

- Anpassning av grammatik- och stavningskontroll till grupper med speciella behov – med SveFix (FCIS) eller Norsk grammatikkontroll (UiO) som exempel och mönster
  - Olika språkdomäner: teknik, medicin, juridik, EU-språk
  - Andraspråkskribenter av nordiska språk
  - Dyslektiker
  - …
- Marknaderna i Norden är för små för rent kommersiellt lönsam utveckling i fall av skräddarsydda versioner för de större nationella språk eller allmänna versioner för de mindre nationella språk
- Grund för gemensamma nordiska projekt?
  - Först – prov av konceptens genomförbarhet för ett nordiskt språk
  - Sedan – duplicering av koncepten till andra nordiska språk
  - T.ex. grammatikkontroll för svenska (1997-1999) → finska, danska och norska (bokmål) (2000-2001) → finlandsvenska (2001-)

Slut och tack – frågor?

antti.arppe@helsinki.fi

Stilnormering i norsk

Presentasjon på seminar om språkkontrollprogram, arrangert av Nordens Språkråd.

Pargas, Finland, 21-22. april 2005.

Bjørn Seljebotn

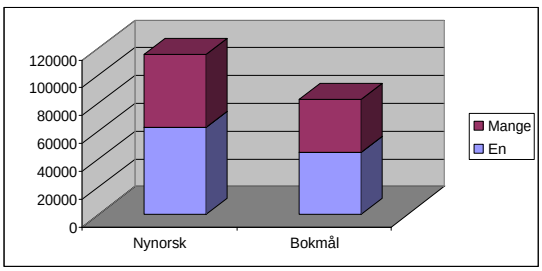
Stilnormering i norsk

- Norsk har store variasjonar i lovlege ord- og bøyingsformer
- To likestilte målformer, bokmål og nynorsk
- Hovudformer og sideformer
- Mange ulike dialektar gir mange parallelle former
- Problem med konsekvent språkbruk

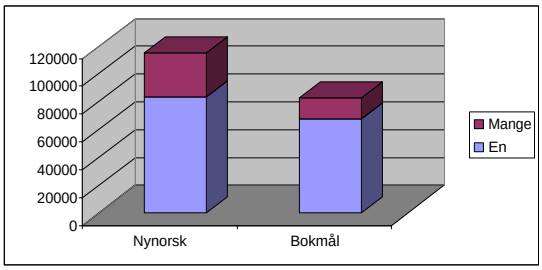
Stilnormering i norsk

- Nær halvparten av alle ord på nynorsk og bokmål har variasjon i bøying.
- Nær ein tredel av alle ord på nynorsk har alternativ skrivemåte.
- Noko mindre variasjon i bokmål.

Bøyingsvariasjon i norsk



Variasjon i ordform i norsk



Eksempel normering i nynorsk

- Eksempelet er henta frå Nynorskordboka.
- Vi viser først aktuelle variablar i orddel eller bøying.
- Deretter viser vi aktuelle ordformer.
- Til slutt appliserer vi normeringskriteria i norsk rettskriving.

ORDARTIKKEL FRÅ NYNORSKORDBOKA

draumebilete

draum~ eller **draume~** [**drømme~**] / **bilete** [**~bilde**] [**drømme|bilete** eller **~bilde]**

idealbilete *lage seg eit d- av noko*(*n*)

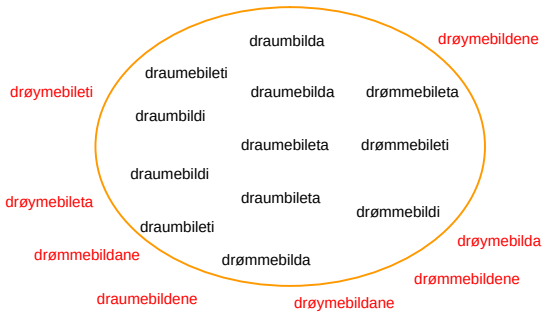
Aktuelle variablar *draumebilete*

Ord	**Draumebileta**				
Språk	Norsk				
Målform	Nynorsk				
Form	Nøytrum	bestemt form pluralis			
		Ord	Fuge	Ord	Bøying
Variablar	draumebileta	draum	e	bilete	a
		drøyme	/0	bilde	**i**
		drøm			**ene**
		drømme			**ane**
Grøn	Normert sideform				
Raud	Illegal form nøytrum nynorsk				

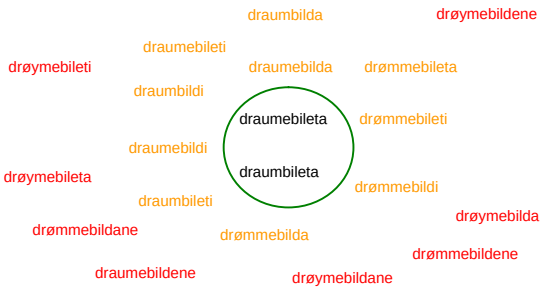
Aktuelle ordformer *draumebilete*

draumbilda drøymebildene draumebileti drøymebileti draumebilda drømmebileta draumbildi draumebileta drømmebileti drøymebileta draumebildi draumbileta drømmebildi draumbileti drøymebilda drømmebilda drømmebildene draumebildene drøymebildane drømmebildane

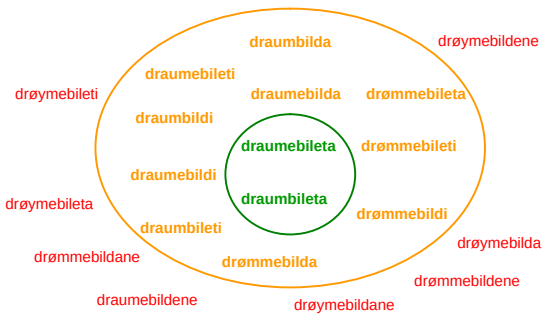
Norm: Nynorsk sideformer



Norm: Nynorsk hovudformer



Norm: Nynorsk rettskriving



Stilnormering i norsk

- Eksempel på bruk av språkmalar eller stilmalar i bokmål.
- Eksempelet er henta frå Bokmålsordboka.

ORDARTIKKEL FRA BOKMÅLSORDBOKA

oppmykningsøvelse
~s|øvelse eller ~s|øving
gymnastikkøvelser som myker opp kroppen

oppmjukingsøvelse
~s|øvelse eller **~s|øving**
gymnastikkøvelser som myker opp kroppen

Stilnormering i norsk

- I Nynodatas system er alle rettskrivingsvariantar kategoriserte og samla i språkmalar.
- Vi har ein kategori for kvar variabel. Kvar kategori er kopla til relevante ord, i ein avansert databasestruktur.
- I systemet ligg det totalt ca 160 kategoriar for nynorsk og ca 120 for bokmål.
- For kvar språkmalkategori kan brukaren definere sin preferanse.
- I ordet *oppmykningsøvelse* er det fire ulike variablar.

Aktuelle variablar *oppmykningsøvelse*

Ord	**Oppmykningsøvelsen**				
Språk	Norsk				
Målform	Bokmål				
Form	Bestemt form sing.				
		Orddel	Orddel	Orddel	Bøying
Variablar	oppmykningsøvelsen	myk	ning	else	en
		mjuk	ing	ing	a

Aktuelle ordformer *oppmykningsøvelse*

oppmjukningsøvelsen
oppmjukningsøvinga
oppmjukingsøvelsen
oppmjukingsøvinga
oppmjukingsøvingen
oppmykningsøvingen
oppmykingsøvelsen
oppmykningsøvelsen
oppmykingsøvinga
oppmykningsøvinga
oppmykingsøvingen
oppmjukningsøvingen

Ord: oppmykningsøvelse
Norm: Bokmål rettskriving

oppmjukningsøvelsen
oppmjukningsøvinga
oppmjukingsøvelsen
oppmjukingsøvinga
oppmjukingsøvingen
oppmykningsøvingen
oppmykingsøvelsen
oppmykningsøvelsen
oppmykingsøvinga
oppmykningsøvinga
oppmykingsøvingen
oppmjukningsøvingen

Stilnormering i norsk

- Det er fire ulike variablar i ordet *oppmykningsøvelse.*
- Systemet lagar ulike sett med ord for kvar av variablane.
- Brukarpreferansen for kvar språkmalkategori avgjer kva sett som skal brukast.
- Systemet kontrollerer preferansen for kvar variabel inntil det står att berre eitt alternativ.

Språkmalkategoriar for *oppmykningsøvelse*

ID	Malkategori	Eksempel	Bjørns mal
14	Vokal y/(j)u	myk	X
		mjuk	
23	Suffiks på -else /-ing	-else	
		-ing	X
25	Variasjon -ning / -ing	-ning	
		-ing	X
61	Genus maskulinum/femininum	boken	
		boka	X

Malkategori: y/ju

Malkategori: -ning/-ing

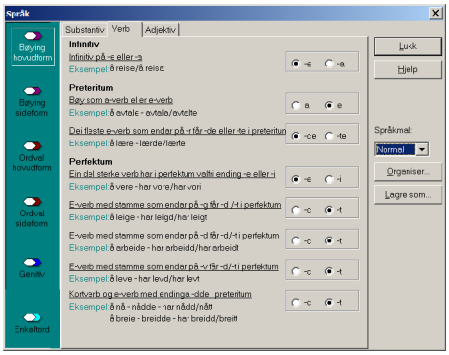
Malkategori: -else/-ing

Malkategori: mask./fem.

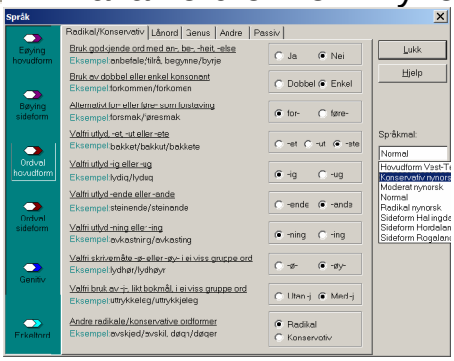
Stilmalar i Nyno

- Ein eldre versjon av systemet er implementert i omsettingsprogrammet Nyno.
- Her er det 3 generelle stilmalar, i tillegg til geografiske malar.
- Den enkelte bedrift kan lage sin eigen mal.
- Vi viser nokre eksempel på konfigurering og bruk av språkmalane i Nyno.

Bøying av verb i Nyno

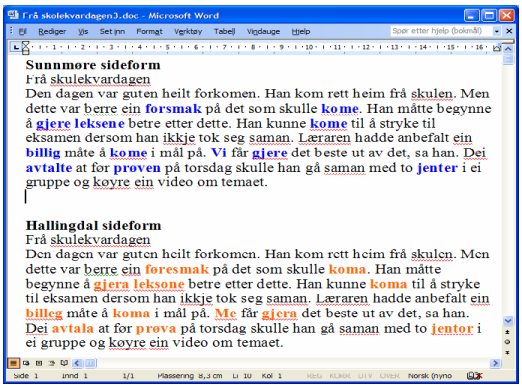


Val av ordformer i Nyno



Eksempel stilmal

Eksempel geografisk språkmal



Stilnormering i norsk

- Ein oppdatert versjon av språkmalane blir ferdigstilt våren 2005.
- Systemet har som mål å vere komplett og inkludere alle variantar i norsk rettskriving.
- Språkmalane kan implementerast i avanserte korrekturprogram for bokmål og nynorsk.

April 2005

Bjørn Seljebotn
Dagleg leiar Nynodata AS

bjorn@nynodata.no

www.nynodata.no

Sprogteknologi i Grønland

Per Langefjord

Behovet

- Grønland er et moderne samfund i en moderne globaliseret verden
- Grønlandsk er det eneste officielle sprog i Grønland og skal som sådan kunne klare alt hvad man forventer af et nationalt sprog i et moderne samfund i en moderne globaliseret verden.

Ressourcerne

- Sprogteknologisk tænkning er så ny i Grønland, at den politiske forståelse kunne være bedre. Der er derfor meget få penge til rådighed.
- Der er meget få grønlændere med en relevant sproglig uddannelse, og de få, der findes, har for travlt til at kunne investere i den lange ekstrauddannelse som teknologien kræver
- Der findes ingen taggede corpora, kodede ordlister el.lign. på et så højt niveau, at de umiddelbart kan anvendes som inddata for sprogteknologien.

En underlig historie

- Bevillingsgiverne har hidtil vurderet, at Oqaasileriffiks listebaserede stavekontrol med 350.000 ord er alt for ambitiøs. Den er blevet kaldt en ”Rolls Royce model” og har ikke fået bevillinger.
- I stedet har man investeret i en stavekontrol på basis af en liste med 17.999 enkeltord. Man kalder den en ”Skoda model” og synes, at den er mere passende til Grønlands begrænsede ressourcer.

Lige nu

- Med den øgede grønlandicering vokser behovet.
- Med vægten af en række underdimensionerede og uanvendelige projekter i rygsækken vokser forståelsen for, at grønlandsk ikke kan klare sig med små løsninger bare fordi grønlandsk er et ”lille” sprog.
- Med de første spæde resultater vokser den politiske forståelse.
- Der uddannes nu unge grønlændere med sprogteknologisk interesse og der opbygges i disse år rådata, der vil muliggøre helt anderledes målrettet virksomhed

Fremtiden

- Der kommer en nogenlunde anstændig stavekontrol til Office allerede i år.
- Oqaasileriffiks on-line database med pt. omkr. 65.000 ord er under editering og får samtidig indlagt ordklassekoder. Arbejdet vil være helt eller næsten afsluttet i 2005.
- Der forventes afsat ½ årsværk i år til at (påbegynde) udviklingen af værktøj til tagging og parsing af grønlandsk tekst.
- Fra omkr. årsskiftet vil nye ord komme på internettet straks efter registrering og politisk behandling.

Det er dyrt at være fattig

- Vi har spildt en masse ressourcer fordi vi ikke har haft råd til at gøre tingene ordentligt fra begyndelsen. En billig EDB-løsning, som har ædt alle vore kræfter, en outsourcet hjemmeside, som aldrig har fungeret, og en listebaseret stavekontrol, som nu bliver lavet helt om.
- Men det var måske alligevel den nødvendige øjenåbner. Den politiske bevågenhed og dermed pengene.
- Vi tror ikke længere på nemme løsninger og har lært, at træerne ikke vokser ind i himlen nord for trægrænsen.

Credo

- Men vi arbejder trøstigt videre i tillid til, at solen begynder at skinne lige rundt næste kurve.