

Introduktion till språkteknologi

OH-serie 2:

Datorstöd för språkgranskning

oktober 2008

Mats Dahllöf (efter Sofia Gustafson-Capková)

Institutionen för lingvistik och filologi

UPPSALA UNIVERSITET

Huvudpunkter

- Vad är datorstödd språkgranskning
- Skrivprocessen och skrivstöd
- Datoriserade skrivstöd
- Hur används dessa stöd
- Slutsatser och frågeställningar.
- **Läs:** Ola Karlsson: *Kontrollera språkkontrollen* (länk från hemsidan).

Vad är datorstödd språkgranskning?

- En hjälp för användaren att granska och bearbeta en texten.
- Skrivstöd
 - stöd för redigering
- Elektroniska ordböcker
- Språkkontroll
 - Stavning, grammatik och stil

Perfekt språkgranskning

- Att finna alla stavfel och ogrammatiska konstruktioner, och ge (korrekta) förslag på korrigerings, men inte ge några falsklarm.
- (Extraktionsproblem, identifiera och lokalisera "fel".)
- Vad är ett språkfel?
 - Dem – dom, mig – mej
 - större än han/honom
 - söndrig

Vad är viktigast?

- Att programmet hittar så många fel som möjligt (i idealfallet alla)?
- Att programmet ger så få falsklarm som möjligt (i idealfallet inga alls)?
- Konflikt mellan dessa.
- Vilka är användarens behov?

Teckenbaserad kontroll

- Teckenbaserad mönstermatchning.
- Begränsad funktionalitet, kontrollerar främst teckenanvändning som t.ex. stor bokstav efter punkt eller apostrofanvändning.
- Tekniken skulle t.ex. kunna användas av en regel att larva då något matchar mot mönstret "det -a -arna"
 - *det fina fåglarna*
 - *det ska varna*
- T.ex. Skribent, Norstedts ordbok.

Stavningskontroll

- Kontrollen av felstavade ord arbetar för det mesta mot en ordlista.
- Ju mer morfologi ett språk har, dest mer praktiskt att arbeta med grundformer och regler för böjningar, avledningar och sammansättningar.
- Speciellt svårt: sammansättningar (felaktiga särkrivningar) -- "fel" alternativ finns i ordlistan.
 - En *jätte stor* som ett hus.
 - En *jättestor* bil.
- *Han är småskolärare, hon är barnnorska. Nu skall de till Dunmark på semester.*

Grammatikkontroll

- Fenomenbaserat (ytanalys)
 - Kontrollerar en begränsad del av meningen som kan förväntas innehålla fel.
 - Snäv kontext, regler för ordklasssekvenser, t.ex. kan två supinumformer i rad vara en otillåten kombination. Ytlig parsning.
- Grammatikbaserat (djupare analys)
 - Kontrollera om (hela) satsen är grammatisk given en uppsättning grammatikregler -- > ej möjligt.
 - En lösning: "relaxation", dvs att ange vilka regler som inte alltid behöver följas.

Statistiska metoder

- Programmet lär sig vanliga fel genom att ett maskininlärningsprogram tränas på en korpus med vanliga fel.
- Mindre vanligt.
- Statistiska metoder vid analysen (taggning och parsning) kan dock förekomma utan att själva feldetekteringen är statistiskt baserad.

Existerande system för svenska

- Fenomenbaserade:
 - Grammatifix (Word) – Lingsoft
 - Granska – NADA, KTH
 - Grim – Granska anpassat för andraspråksinlärare av svenska.
 - Finite Check – Göteborgs universitet
- Grammatikbaserade:
 - SCARRIE – Uppsala universitet

SCARRIE

- Utvecklat vid Uppsala universitet, korrekturläsningens verktyg för tidningsskribenter.
- Stavningskontroll
 - Lexikonuppslagning
 - Sammansättningsregler (svårigheter)
- Grammatikkontroll
 - Partiell parsning
 - Söker igenom analysen för fel.

SCARRIE-projektet – feltyper

- Feltyper annoterade i en korpus med autentiska fel.
- Hierarkisk uppdelning av feltyper – 500 på den mest fingranulerade nivån.
- Översta nivån av feltyper:
 - Stavfel 43%.
 - Interpunktion 17%.
 - Stil, mening och referens (syftning) 16%.
 - Grammatiska fel 15%.
 - Grafiska fel 9%.

SCARRIE feltyper II

- NP-fel: Ett stort bil.
- PN-fel: Jag såg hon.
- Verbkedjor: Hon börja läsa.
- Särskrivningar: Upplands kusten.
- Ordföljdsfel: Jag undrar vad gör de.

Utvärderingsmått

- Dessa utvärderingsmått används ofta då en språkteknologisk tillämpning skall leta upp någon typ av fenomen i en text.
- **Precision:** andelen utpekade exempel som är verkliga exempel.
- **Täckning (recall):** andelen verkliga exempel som pekas ut.

SCARRIE – resultat

– Stavningskontroll (hitta stavfel):

- Täckning 96,5 %
- Precision 41,3 %

– Grammatikkontroll (hitta grammatikfel):

- Täckning 85,7 %
- Precision 92,3 %

Granska grammatikkontroll

- Domeij (2003), Knutsson (2005).
- arbetar mot ett lexikon samt med hjälp av enkel kontroll av ändelser (om programmet ser "lampan" kommer också "lampor" att accepteras).
- har viss hantering av sammansättningar baserat på en undantagslista.
- tokeniserar – ord och meningar.
- Ordklasstagging och morfologisk analys (statistiskt baserad).
- Feltektering (regelbaserad som använder sig av ord- och ordklassinformation).
- Partiell parsning.

Feltyper – Granska (Knutsson, 2005)

- Kongruens i NP – jag såg en glador.
- Predikativfel (inkongruens) – mannen var glada.
- Särskrivning – gladan tog en åker sork.
- Verbkedjor – jag har spana på en glada en längre tid.
- Objektsform – jag visade gladan för hon.
- Stavning – jag såg en galda.

Granska – resultat

- Stavnings- och grammatikkontroll: Täckning 52%. Precision 53%. (Är det bra?)
- Böjningsfel i verbkedjor: täckning 97%, precision 83%.
- Kongruensfel : täckning 74%, precision 72%.
- Särskrivningar: täckning 40%, precision 67%.

Grim

- Datoriserat skrivstöd för andraspråksinlärare.
- Baserat på Granska, men utökat för att stödja andraspråksinlärare av svenska:
- Visualisering/färgkodning av olika typer av fel.
 - Böjningsformer.
 - Markering av ordklasser, fraser och satser.
 - Lexikon – Lexin.
 - Korpussökning – Parole.

Förhoppningar – resultat, GRIM

- Förhoppningen var att tillgång till flera olika verktyg skulle verka aktiverande på studenten, och hjälpa studenten att själv finna sig tillrätta i språket.

Resultat:

- Problem med:
 - falska alarm (dock inte lika problematiskt med icke detekterade fel).
 - diagnos och rättning av särkrivningar.
 - flera möjligheter till korrektion.

Att jämföra olika system

- Svårt att jämföra siffror från olika utvärderingar:
 - Olika feltyper.
 - Olika texter.
- Komparativ utvärdering:
 - Samma feltyper.
 - Samma texttyper.
 - Synliggör olika systems strategier.

Ännu ett utvärderingsmått: F-score

F-score räknas ut som

$$2 * (\text{täckning} * \text{precision}) / (\text{täckning} + \text{precision})$$

(Ett sätt att väga samman täckning och precision.)

Granska och andra grammatikkontroller

Texter skrivna av grundskoleelever

Granska, F-score 23%

SCARRIE, F-score 18%

MS Word, F-score 14%

Finite Check, F-score 49%

(från Knutsson 2005)

Granska och MS Word

- Undersökningen utförd på inläraarsvenska (Svante)

	Granska (P / R)	Word (P / R)
Typografi	70 / 21	62 / 82
Stavning	86 / 57	38 / 91
Grammatik	80 / 84	86 / 34
Total	82 / 63	47 / 66

(Knutsson, 2005)

Återkoppling

- Hur förklarar man för användaren vad det är för fel med sekvensen som är understruken med rött?
 - Förklaringar använder ofta en rätt teknisk språkvetenskaplig terminologi. Vet alla vad predikativ är?
- Olika återkoppling till olika användare.

Användarperspektivet

- Viktigt att användaren är medveten om programmets begränsningar och har ett kritiskt förhållningssätt.
- Ett sätt att kommunicera programmets begränsningar kan vara att tydliggöra syftet med programmet.
- Vilka är de feltyper som är viktigast att programmet kan upptäcka?
- Hur ger man explicit feedback på bästa sätt?
- Vilka språkteknologiska resurser är praktiska att integrera i en ordbehandlare?

Ersättningsförslag

- Ersättningsförslag kan vara till hjälp, om de ofta är de avsedda.
- Vad krävs av användaren?
- Osäkra skribenter kan ha svårt att välja.

Hur används datorstöden?

- Främjar en ytlig granskningsstrategi.
- Djupare bearbetning av text handlar om arbete av annat slag (t.ex. omstrukturering, sovring, klargöranden, preciseringar).
- Fungerar kanske för en typ av användare, men inte för andra.

Slutsatser

- Datorstödd språkgranskning kan klart hjälpa oss att undvika slarvfel med felslag etc.
 - Ju mer problemet ligger i sammanhang och presentation av budskapet, desto svårare för den datorstödda språkgranskningen.
- Viktigt med ”upplysta användare”.
 - Viktigt med ett uttalat syfte för programmet – detta påverkar även vilka feltyper som fokuseras samt vilken typ av återkoppling man ger.

Referenser

- Domeij, R. (2003): Datorstödd språkgranskning under skrivprocessen. Doktorsavhandling, Institutionen för lingvistik, Stockholms universitet.
- Domeij, R. (2005): Datorn granskar språket. Svenska språknämnden.
- Karlsson, O. (2007): Kontrollera språkkontrollen. Språkvård 2/2007.
http://www.sprakradet.se/servlet/GetDoc?meta_id=2331
- Knutsson, O. (2005): Developing and Evaluating Language Tools for Writers and Learners of Swedish. Doktorsavhandling, Institutionen för numerisk analys och datalogi, KTH.
- Sågvall Hein, A. (1998): A Chart-Based Framework for Grammar Checking, Initial Studies. NODALIDA '98 Proceedings Vol. 11. Center for Sprogteknologi, Denmark.
<http://www.lingfil.uu.se/personal/anna/NODALI.pdf>
- Sågvall Hein, A., Olsson, L., Dahlqvist, B., Mats, E. (1999): Evaluation Report for the Swedish Prototype, Deliverable 8.1.3. In: Working Papers in Computational Linguistics & Language Engineering, vol 13, 1999. Dept. of Linguistics, Uppsala University.