# Tentamen i DH2418 Språkteknologi 2011-10-19, kl 9-13

---

Tentamen består av två delar. Till del 1 (uppgift 1-5) är inga hjälpmedel tillåtna. Till del 2 (uppgift 6-10) får kursboken *Speech and Language Processing* av Jurafsky och Martin användas. Lämna in del 1 i separat konvolut innan du tar fram boken och fortsätter med del 2.

Utbytesstudenter får ha med sig lexikon (på både del 1 och del 2), och skriva tentan på engelska om så önskas.

För att bli godkänd måste du ha minst **13 poäng på del 1 inkl. teoretiska bonuspoäng,** och **minst 12 poäng på del 2 inkl. praktiska bonuspoäng.** Betygsgränserna är enligt följande på hela tentan:
**A**: 45 poäng och uppåt, **B**: 40-44 poäng, **C**: 35-39 poäng, **D**: 30-34 poäng, **E**: 25-29 poäng, **Fx**: 24.

---

## Del 1: Teoretiska uppgifter (20p)

Skriv korta svar på dessa fem frågor; en riktlinje är max två sidor för hela del 1.

1. Det sägs ibland att de största språkteknologiska utmaningarna kommer sig av att språkliga konstruktioner kan vara **flertydiga.** Ge exempel på **två olika sorters** flertydighet i språk. För varje exempel, välj en språkteknologisk applikation, och visa hur flertydigheten kan utgöra ett problem. (*4p*)

2. Utvärderingsbegreppen **täckning** och **precision** är användbara inom flera språkteknologiska områden, men vad står de för när en grammatikkontroll skall utvärderas? Beskriv också kortfattat hur en språkteknolog skall gå tillväga för att ta reda på en grammatikkontrolls täckning och precision. (*4p*)

3. Ange tre metoder som en ordklasstaggare kan använda för att gissa rätt tagg för ett ord som inte finns i lexikonet. (*4p*)

4. Vad är det för skillnad mellan en **Boolesk modell** och en **vektorrumsmodell** för informationssökning? Förklara också vad det finns för för- och nackdelar med de olika modellerna. (*4p*)

5. Vad innebär det att ett dialogsystem tillåter **blandat initiativ**? Förklara hur man kan åstadkomma detta. (*4p*)

## Del 2: Problemdel (30p)

Skriv helst maximalt 1 sida per uppgift.

6. Skatteverket vill införa ett automatiskt system för styrning av e-post. Varje e-brev som anländer till verket ska kategoriseras med avseende på vilken avdelning som ska besvara brevet. Kategorierna är ännu inte helt bestämda, men tänkbara kategorier är ”personbevis”, ”inkomstskatt”, ”fastighetsskatt”, ”utlandssvenskar”, ”avdrag”, osv. Om du vore teknisk projektledare för de språkteknologiska delarna av detta system, hur skulle du designa dessa? Vilka delproblem måste lösas, och hur ska systemet utvärderas? (*6p*)

7. Programmet Stava genererar rättelseförslag till ett felstavat ord genom att testa alla tänkbara korrigeringar av ordet mot ordlistan (bloomfiltret). Förslagen rangordnas sedan med hjälp av förfinad felstavningsmetrik och ordfrekvensdata.
    a. Skissa hur man kan göra för att ge rangordnade rättelseförslag utan ordlista och ordfrekvenser, med hjälp av en tredje ordningens markovmodell för bokstäver. (*4p*)
    b. Vad är risken med ett ordlistefritt system för generering av rättelseförslag, med avseende på hur användarna uppfattar systemets trovärdighet? (*2p*)

8. Det finns många svenska deckarförfattare, bland andra Camilla Läckberg, Inger Frimansson, Liza Marklund, Leif GW Persson och Åke Edwardsson. Sture vill undersöka hur lätt det går att skilja deras böcker åt, och har därför tränat en automatisk kategoriserare där han kan mata in ett textstycke och få det kategoriserat som ett av författarnamnen ovan (**CL**, **IF**, **LM**, **LGWP**, eller **ÅE**) eller som klassen **Någon Annan**. Som träningsmaterial har Sture använt hela sitt deckarbibliotek; sju böcker av CL, och två böcker av var och en av de övriga fyra författarna. Närmare bestämt har Sture märkt varje stycke ur de sammanlagt femton böckerna med författarnamn och matat in som indatapunkter. När han testar sin kategoriserare med meningar ur böckerna ser han att han får rätt svar i ungefär 33 % av fallen. Sture är ganska nöjd, eftersom slumpen bara skulle gett rätt i 20 % av fallen.
    a. Deckarentusiasten Kalle köper Stures kategoriserare men kommer snart tillbaka och klagar. Kalle hävdar att det blir mycket sämre resultat än 33 % när han testar, och dessutom kategoriseras dikter av Tomas Tranströmer som skrivna av Leif GW Persson. Vad har Sture gjort för fel? (*3p*)
    b. Kommentera Stures val av baslinje: Är det rimligt att jämföra kategoriseraren mot slumpen i det här fallet? Finns det någon annan baslinje som är lämpligare? (*3p*)

9. Förläggare som ger ut ordböcker och lexikon brukar utse en eller flera redaktörer som sköter årliga uppdateringar. Nya betydelser av ord ska läggas till, medan utgångna betydelser ska förses med etiketten *åldrig* eller rensas bort helt. (För några år sedan kunde man till exempel lägga till en ny möjlig översättning av ordet ”fett” till det engelska ordet ”cool”.) Ge förslag på hur detta arbete delvis skulle kunna automatiseras med hjälp av statistiska metoder. (*6p*)

***V. G. VÄND!***

10. Vi vill skriva ett datorprogram som tar en svensk mening som indata, och som utdata ger en smiley som representerar om meningen har en positiv, neutral eller negativ innebörd; antingen ☺, 😐, eller ☹. Om meningen är ogrammatisk ska resultatet vara ”?”. Här är några exempel på hur det ska fungera:

| **Meningen** | **ska ge som svar:** |
|---|---|
| han är olycklig | ☹ |
| han är inte olycklig | ☺ |
| han är knappast inte olycklig | ☹ |
| det är inte sant att han knappast inte är olycklig | ☺ |
| han är lycklig | ☺ |
| han är inte lycklig | ☹ |
| han var sjuk | ☹ |
| han var sjuk och blev frisk | ☺ |
| han var sjuk och blev aldrig frisk | ☹ |
| han var sjuk blev frisk | ? |

Vi förutsätter att systemet ska ha en begränsad vokabulär, men kunna hantera en så stor mängd meningar att det **inte** går att lägga alla tänkbara inputmeningar i en tabell som den ovan. Förklara med hjälp av begrepp från kursen hur en lösning skulle kunna designas. (*6p*)