Försättsblad till skriftlig tentamen vid Linköpings Universitet

Datum för tentamen	2013-03-11
Sal (2) Om tentan går i flera salar ska du bifoga ett försättsblad till varje sal och ringa in vilken sal som avses	G37 TER3
Tid	14-18
Kurskod	TDDD01
Provkod	TEN1
Kursnamn/benämning **Provnamn/benämning**	Språkteknologi Skriftlig tentamen
Institution	IDA
Antal uppgifter som ingår i tentamen	11
Jour/Kursansvarig Ange vem som besöker salen	Lars Ahrenberg
Telefon under skrivtiden	013-282422
Besöker salen ca kl.	14.45
Kursadministratör/kontaktperson (namn + tfnr + mailaddress)	Liselotte Lundberg ank 1278, liselotte.lundberg@liu.se
Tillåtna hjälpmedel	inga
Övrigt	
Vilken typ av papper ska användas, rutigt eller linjerat	Valfritt
Antal exemplar i påsen	

Lars Ahrenberg

Institutionen för datavetenskap
Linköpings universitet

TENTAMEN

TDDD01 Språkteknologi

måndag 11 mars 2013 kl. 14-18

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 28.
14 poäng ger säkert godkänt.

Det går bra att besvara flera frågor på samma papper.

Besvara alla frågor

1. Förbearbetning av texter i språkteknologiska sammanhang innebär t.ex. att man separerar skiljetecken som komma, punkt etc. som egna tokens och omvandlar alla stora bokstäver till små. Varför gör man det? (2p)

2. (a) Skriv ett reguljärt uttryck, utan att använda disjunktionsoperatorn |, som identifierar orden i mängden {lä, lät, lätt, lätta} och inga andra ord; (b) Ange en tillståndsautomat (eng. *finite-state automaton*), även kallat bokstavsträd, som accepterar precis samma ord. (c) Visa sedan hur automaten kan utvidgas för att även accepera orden i mängden {rät, rätt, rätta} men inga fler. (3p)

3. Vad gör en morfologisk parser? (1p)

4. I en svensk korpus finner vi följande frekvenser för några utvalda ord och ordsekvenser:

   på grund av: 250
   på grund: 260
   grund av: 271
   på: 32,900
   grund: 400
   av: 24,500

   (a) Vad är den uppskattade sannolikheten för p(av | på grund), om vi (a) använder trigramsannolikheter, (b) en omskrivning med hjälp av bigramsannolikheter? För båda fallen använd Maximum Likelihood-uppskattningen. (2p)

5. Utjämning (eng. *smoothing*) är en teknik som används när man bygger statistiska språkmodeller (eng. *language models*). Förklara vad det innebär och varför man använder det. Beskriv därefter den utjämningsteknik som kallas Add-1 eller Laplace. (3p)

Institutionen för datavetenskap
Linköpings universitet

LINKÖPINGS UNIVERSITET

e-post:
lars.ahrenberg@liu.se

6. I modellen 'den brusiga kanalen' (eng. *Noisy Channel*) ser formeln för att välja mest sannolika sträng ut så här:

$$S^* = \underset{S}{\operatorname{argmax}}\, p(S|O)$$

(a) Förklara formeln i ord. (b) Ofta skriver man om sannolikheten p(S|O) med hjälp av Bayes regel. Visa resultatet av omskrivningen och förklara varför man gör den. (3p)

7. (a) Vad innebär chunkning? (b) Vilka mått är relevanta för att mäta prestanda hos ett chunkningssystem? Motivera ditt svar. (3p)

8. Vid s.k. top-down chart-parsning av kontextfri grammatik används tre grundläggande operationer för att generera och bekräfta hypoteser: predicering, scanning och kombinering (eng. *completion*). Beskriv vad dessa operationer innebär. Du kan använda den enkla grammatiken nedan för att illustrera operationerna: (3p)

S → NP VP	ADJ → tyska, dyra
NP → N	N → bilar, kameror
NP → ADJ N	V → är, suger
VP → V	
VP → V ADJ	

9. Naturliga språk har många ord som är flertydiga, dvs har olika betydelser i olika kontexter. Svenska flertydiga ord är t.ex. *kurs, känna, ljus, mask.* Ange tre olika typer av indikatorer som kan användas för att bestämma vilken betydelse ett ord har i en given kontext, t.ex. med en metod som Naive Bayes. (3p)

10. Ett moment i ett frågebesvarande system är att identifiera textutdrag (eng. *passages*) där svaret på en given fråga sannolikt kan hittas. Ange minst två kriterier som använts för att identifiera sådana textutdrag (tre stycken räcker för full poäng). (2p)

11. I flera språkteknologisa tillämpningar kan det vara svårt att mäta recall på det vanliga sättet, dvs utifrån ett facit med ett rätt svar för varje datapunkt som systemets utdata sedan jämförs med och bedöms som korrekt eller fel. Ange minst två sådana tillämpningar och för var och en av dem, något sätt man använt för att utnyttja befintliga korrekta utdata på andra sätt. (3p)