



Försättsblad till skriftlig tentamen vid Linköpings Universitet

(fylls i av ansvarig)

Datum för tentamen	2011-03-14
Sal	TER 2
Tid	14-18
Kurskod	TDDD01
Provkod	TEN1
Kursnamn/benämning	Språkteknologi
Institution	IDA
Antal uppgifter som ingår i tentamen	11
Antal sidor på tentamen (inkl. försättsbladet)	3
Jour/Kursansvarig	Lars Ahrenberg
Telefon under skrivtid	013-282422
Besöker salen ca kl.	14.45
Kursadministratör (namn + tfnr + mailadress)	Helene Meisinger 281868, helene.meisinger@liu.se
Tillåtna hjälpmedel	Inga
Övrigt (exempel när resultat kan ses på webben, betygsgränser, visning, övriga salar tentan går i m.m.)	
Vilken typ av papper ska användas, rutigt eller linjerat	valfritt
Antal exemplar i påsen	

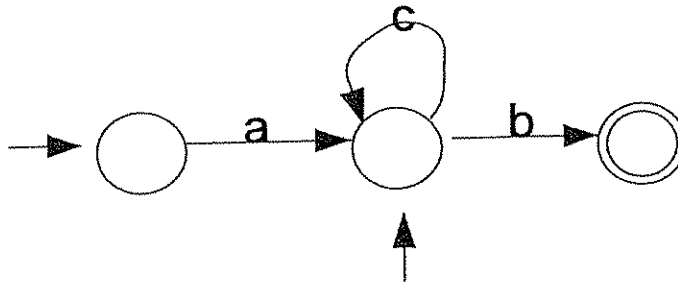
TENTAMEN
TDDD01 Språkteknologi
 måndag 14 mars 2011 kl. 14-18

Inga hjälpmedel är tillåtna.

Maximal poäng finns angiven för varje fråga. Maximal poäng på hela tentan är 24.
 12 poäng ger säkert godkänt.

Besvara alla frågor. Det går bra att besvara flera frågor på samma papper.

1. Förklara vad som menas med att lemmatisera en text och ange något syfte, eller någon tillämpning, där det är motiverat att man gör det. (2p)
2. Den finita automaten här definierar ett språk över alfabetet {a,b,c}. Skriv automatens språk med ett reguljärt uttryck. (1p)



3. Många adjektiv i svenska som slutar på -a kan också vara substantiv (*glada, visa, raka, ...*) eller verb (*visa, viga, råta, ...*). Ange kontextvillkor av den typ som kan användas i system för ordklasstaggning som entydigt eller nästan entydigt visar om en viss förekomst av ett sådant ord är (a) ett substantiv, (b) ett verb, (c) ett adjektiv. (Du ska alltså ange ett villkor per delfråga.) (3p)
4. Definiera måtten recall och precision och förklara varför båda måtten behövs. (3p)
5. Vad menas med redigeringsavstånd (Minimal edit distance, eller Levenshtein distance)? Beskriv kortfattat hur Levenshteinavståndet definieras och avgör vilket eller vilka av följande ord teckensekvensen *svältkost* har kortast avstånd till: *sveciaost, frukost* eller *svältkatastrof*. (2p)
6. Anta att vi har givet en bigrambaserad språkmodell. Vilken av följande två ordsekvenser ges störst sannolikhet av en sådan modell: *Sverige kämpade hårt - Sverige kämpade hårt men förlorade* Svaret ska motiveras. (1p)

7. I en korpus hittades följande frekvenser för några olika ordsekvenser:

många	100
inte	200
är	400
inte många	10
inte är	20
inte många är	2
inte är många	0
många är	80
är många	20

- (a) Beräkna Maximum Likelihood-uppskattningen av sannolikheten för att 'många' kommer efter 'är' dvs $p(\text{många} \mid \text{är})$ utifrån denna korpus. (b) Beskriv något sätt att uppskatta trigramsannolikheten $p(\text{många} \mid \text{inte är})$ till något annat än 0 givet dessa data. (Du behöver inte ge ett värde, det räcker med en beskrivning.) (2p)
8. En top-down chart-parser drivs framåt genom successiva prediceringar av hypoteser och matchningar av hypoteserna mot orden i indata och tidigare verifierade hypoteser. Låt grammatiken G vara given enligt nedan, där S står för kategorin meningar. (a) Vilken är den initiala hypotesen och hur representeras den som en punkterad regel? (b) Vilka prediceringar genereras utifrån denna initiala hypotes och reglerna i G oberoende av indata som ska parsas? (2p)
- $S \rightarrow NP VP$
 $NP \rightarrow DET N$
 $NP \rightarrow PN$
 $VP \rightarrow IV$
 $VP \rightarrow TV NP$
9. En central modul i ett frågebesvarande system är frågeanalysatorn som bestämmer svarstyp. (a) Förklara vad som menas med en svarstyp och ge några exempel. (b) Ge något exempel på att det inte räcker med att känna igen frågeord i en fråga för att kunna bestämma svarstyp. (2p)

10. I system för ordbetydelsebestämning representeras kontexten för ett ord som skall betydelsebestämmas ofta med en s.k. kontextvektor. Förklara vad en kontextvektor är och hur den används för att bestämma betydelse. Du kan illustrera med ordet *panna* i meningen nedan och anta att ordet har tre olika betydelser (ansiktsdel, kärl eller drivanordning).

Han hade glömt pannan på spisen och det gjorde honom riktigt bekymrad. (3p)

11. Ett sätt att identifiera s.k. chunkar, till exempel icke-rekursiva nominalfraser i löpande text baseras på taggar som markerar positionen för ett ord i frasen. Beskriv närmare vilka taggar som används och hur taggningsprocessen går till om man använder transformationsregler. (3p)