LUNDS UNIVERSITET
Lunds Tekniska Högskola

Kursplan för kalenderåret 2006

SPRÅKBEHANDLING OCH DATALINGVISTIK EDA171

Language Processing and Computational Linguistics

Antal poäng: 5. **Betygskala:** UG. **Valfri för:** C4, D4, E4, F4. **Kursansvarig:** Professor Pierre Nugues, Pierre.Nugues@cs.lth.se, Inst f datavetenskap. **Förkunskapskrav:** EDA027 Algoritmer och datastrukturer. **Prestationsbedömning:** Godkända inlämningsuppgifter, projekt och godkända uppsatser krävs för godkänt betyg på kursen. Inlämningsuppgifter och projekt utförs normalt i grupper om två eller tre studenter men kan även utföras individuellt. De tre inlämningsuppgifterna behandlas under sammanlagt 6 laborationstillfällen. Dessutom skall 3 individuella uppsatser om artiklar inom området skrivas. Projektet handleds av en lärare och redovisas i slutet av kursen. Om projektet inte godkänns vid detta tillfälle ges en möjlighet att inom en månad förbättra resultatet och presentera på nytt. **Övrigt:** Kursen kan komma att ges på engelska. Begränsat deltagarantal (58). Högst 6 till utbytesstudenter. Återstående 52 platser fördelas i proportion till antalet sökande vid LTH och Nat fak. Urvalsgrund för studenter vid LTH: 1. Antal poäng som uppnåtts vid LTH. 2. Antal poäng i kurser i Datavetenskap. **Hemsida:** http://www.cs.lth.se/EDA171.

Mål

Kursen introducerar teorier och tekniker som tillämpas i maskinell behandling av naturligt språk (Natural Language Processing, NLP). Den täcker hela fältet, från taligenkänning och talsyntes till semantik och dialogsystem.

Kursen inriktar sig på tillämpningar i industriell skala eller på försöksnivå, såsom dokumentsökning på Internet, informationsextrahering, talande agentprogram och verbal interaktion i virtuella världar. Centrala algoritmer beskrivs i Prolog eller med reguljära uttryck.

Innehåll

- *En översikt över NLP:* Presentation av NLP, tillämpningar, lingvistikens delområden, exempel.
- *Korpuslingvistik:* Reguljära uttryck, ändliga automater, introduktion till Perl, konkordanser, tokenisering, frekvenslistor, kollokationer.
- *Morfologi och ordklasstaggning:* Morfologi, transduktorer, ordklasstaggning.
- *Frasstrukturgrammatiker i Prolog:* Konstituenter, syntaxträd, prolog för analys av naturligt språk, DCG-regler, variabler, syntaktisk parsning, semantisk parsning med kompositionell analys.
- *Syntaktiska formalismer:* Konstituens och dependens, chart parsing, statistisk parsning,

funktioner, dependensparsning.

- *Semantik:* Formell semantik, lambda-kalkyl, kompositionalitet; substantiv, verb, determinerare, ord och betydelse, lexikal semantik, kasusgrammatiker, semantiska grammatikor.
- *Diskurs och dialog:* diskurs och retorik, anaforer, dialogstruktur, RST (rhetorical structure theory), dialog, ändliga automater, adjacency pairs, talhandlingar, multimodalitet.
- *Översikt över talsyntes och taligenkänning.*

Litteratur

En bok som kommer att publiceras hösten 2005 kommer att rekommenderas som kurslitteratur. Närmare uppgifter i oktober.

Rekommenderad referenslitteratur:

Manning & Schutze: Foundations of Statistical Natural Language Processing, MIT Press 1999. ISBN: 0-262-13360-1.