

Schlussprüfung Einführung in die Computerlinguistik I HS 10

Aufgabenstellung: Simon Clematide

Prüfung vom 20. Januar 2011
Institut für Computerlinguistik, Universität Zürich

Vorname _____ Matrikelnummer _____

Nachname _____

Für Studierende der folgenden Studiengänge:

- ☐ BA - Studiengang Computerlinguistik (Phil. Fakultät)
- ☐ BA - Studiengang Computerlinguistik und Sprachtechnologie (Phil. Fakultät)
- ☐ BA-Studierende (Wirtschaftswissenschaftliche Fakultät)
- ☐ Studierende des Nebenfachs Informatik mit Studienbeginn ab WS 04/05
- ☐ Multidisziplinär (ETH)
- ☐ Andere:

Nur für Lizentiatsstudierende der Computerlinguistik als ein Fach aus der Phil. Fakultät:

Strasse: _____ Hauptfach: _____

PLZ/Ort: _____ E-Mail: _____

Die Prüfungsergebnisse von ECL und PCL von den Lizentiatsstudierenden werden (zusammen) per Post verschickt. *Bitte Adresse nicht vergessen!*

Aufgabe Nr.:	1	2	3	4	5	6	7	8	Summe
Punktzahl:	24	8	8	12	12	10	10	6	90
Davon erreicht:									

Note SU: _____ Note SP: _____

Endnote: _____ Bestanden: ☐ Ja ☐ Nein

Auf jedes zusätzliche Blatt mit Lösungen den Nachnamen schreiben!

Viel Erfolg!

Wichtige Hinweise

Punkte-Maximum: 90 (pro Minute 1 Punkt)

Hinweis: Bitte schreiben Sie in einem überlegten und knappen, aber verbalen Stil (keine Stichwort-sammlungen). Bei inhaltlichen Auswahlendungen, wo einfach mal alles spontan hingeschrieben wird und Falsches wie Korrektes munter vermischt sind, behalte ich mir Abzüge vor. Erlaubtes Hilfsmittel ist die Referenzkarte zum Annotieren.

1. Morphologische und syntaktische Annotation (24 Punkte)

- 18 (a) Zeichnen Sie den TIGER-Baum im Stil des Annotate-Werkzeugs mit **Wortart** (STTS **ohne** Morphologieanalyse), **Phrasenkategorien** und **syntaktischen Funktionen** zu folgendem Satz (auf der letzten Prüfungsseite ist der Satz bereits im Querformat aufgedruckt):
Eine Bank, die Gelder von Ben Ali und seiner Entourage verwaltet, muss laut Gesetz die Konten sperren.
- 3 (b) Wo liegt für Sie die syntaktische Hauptschwierigkeit im obigen Satz? Begründen Sie kurz.
- 1 (c) Geben Sie für die folgenden Wortformen aus obigem Satz die detaillierte **morphologische Analyse** nach STTS-Schema:
- a) seiner: _____
- b) Gesetz: _____
- 2 (d) Welche Wortarten sind für die Wortform „laut“ im Deutschen denkbar? Begründen Sie linguistisch (!), welche Wortart im obigen Satz vorliegt.

2. Spam-Filter (8 Punkte) In einem Testset von 1000 E-Mails hat es 400 echte Spam-Mails.

- 4 (a) Student Bruno hat einen Spam-Filter programmiert, der Spam mit einer Precision von 80% erkennt. Wieviele False-Positives (FP) entstehen, wenn sein Filter insgesamt 300 Positive ergibt? _____
- 4 (b) Studentin Gerda hat ihren selbst programmierten Filter so eingestellt, dass er gleich viele Positive wie Negative erzeugt. Wie hoch ist ihre Precision, wenn sie einen Recall von 90% hat? _____

- 8 **3. Multilingualität (8 Punkte)** Inwiefern ist Multilingualität bereits für einfache Text-Segmentierung ein Problem? Wie lässt es sich lösen?

4. Evaluation (12 Punkte)

6 (a) Worum handelt es sich beim BLEU-Score? Was wird damit gemessen und anhand wovon?

6 (b) Wieso sind Evaluationen in der CL im Allgemeinen wichtig? Wozu dienen sie?

12 **5. Maschinelle Übersetzung (12 Punkte)**

Vergleichen Sie den transferbasierten Ansatz mit dem statistischen. Welche wichtigen Vorteile bzw. Nachteile gibt es?

- 10** 6. **Sprachsynthese (10 Punkte)** Welche Verarbeitungsstufen sind typisch für ein modernes Sprachsynthesesystem wie Mary?

7. Kontextfreie Regel-Grammatiken (10 Punkte)

Gegeben sei eine englische Mini-Grammatik $G = \langle \{IP, I', VP, NP, V_i, PN, N, D\}, L, R_{syn} \cup R_{lex}, IP \rangle$ im Stil der IP-Analysen (IP=inflectional phrase, d.h. Phrase mit flektiertem (Hilfs-)Verb als Kopf):

Syntaxregeln: $R_{syn} = \left\{ \begin{array}{l} IP \rightarrow NP I', \quad I' \rightarrow I VP, \quad VP \rightarrow V_i, \quad VP \rightarrow V_t NP, \\ NP \rightarrow D N, \quad NP \rightarrow PN, \end{array} \right\}$

Lexikonregeln: $R_{lex} = \left\{ \begin{array}{l} I \rightarrow \text{might}, \quad I \rightarrow \text{should}, \quad V_i \rightarrow \text{sleep}, \quad V_t \rightarrow \text{meet}, \quad V_i \rightarrow \text{hit}, \\ D \rightarrow \text{the}, \quad D \rightarrow \text{her}, \quad PN \rightarrow \text{Lucy}, \quad N \rightarrow \text{friend}, \quad N \rightarrow \text{dog}, \end{array} \right\}$

- 3 (a) Schreiben Sie die Linksableitung für den Satz „Lucy should meet her friend“ in der Form:
 $IP \Rightarrow \dots \Rightarrow \text{Lucy should meet her friend}$

- 7 (b) Ergänzen Sie die Grammatikregeln und das Lexikon, damit auch Sätze erkannt werden mit satzwertigen Objekten wie: „Lucy might say that her friend should meet“ oder „her friend might think that Lucy should say that her dog should sleep“, aber nicht *„that her friend might think“.

8. Merkmalstrukturen (6 Punkte)

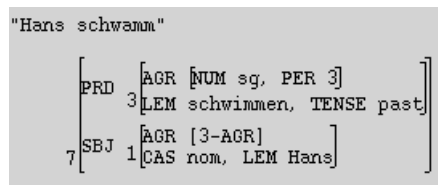
$$M_1 = [\text{NUM} \quad \text{sg}]$$

$$M_2 = [\text{AGR} \quad [\text{NUM} \quad \text{pl}]]$$

$$M_3 = [\text{AGR} \quad []]$$

- 1 (a) Können die Merkmalstrukturen M_1 die Merkmalstruktur M_2 unifiziert werden?
☐ Ja ☐ Nein ☐ undefiniert
- 1 (b) Können die Merkmalstrukturen M_2 die Merkmalstruktur M_3 unifiziert werden?
☐ Ja ☐ Nein ☐ undefiniert

- 4 (c) Zeichnen Sie die folgende koreferente XLE-Merkmalstruktur als Graphen auf:



(Bitte Baum mit STTS-Wortartentags, Phrasen und Funktionen, aber ohne Morphologie zeichnen!)

Eine Bank, die Gelder von Ben Ali und seiner Entourage verwaltet, muss laut Gesetz die Konten sperren.