Schlussprüfung Einführung in die Computerlinguistik I HS 11

Aufgabenstellung: Martin Volk/Simon Clematide

Prüfung vom 19. Januar 2012
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

☐ Multidisziplinfach (ETH)

☐ Andere:

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

Strasse: _____ Hauptfach: _____

PLZ/Ort: _____ E-Mail: _____

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

Aufgabe Nr.:	1	2	3	4	5	6	7	8	9	10	11	12	**Summe**
Punktzahl:	5	10	6	8	16	5	8	19	2	3	3	5	90
Davon erreicht:													

Note SU: _____ Note SP: _____

Endnote: _____ Bestanden: ☐ Ja ☐ Nein

Auf jedes separate Blatt mit Lösungen den Nachnamen schreiben! Viel Erfolg!

Wichtige Hinweise

Punkte-Maximum: 90 pro Punkt sollte ungefähr 1 Minute gebraucht werden
Hinweis: Bitte schreibe in einem überlegten und knappen, aber verbalen Stil (keine Stichwortsammlungen). Bei inhaltlichen Auswahlsendungen, wo einfach mal alles spontan hingeschrieben wird und Falsches wie Korrektes munter vermischt sind, behalten wir uns Abzüge vor. Erlaubtes Hilfsmittel ist die Referenzkarte zum Annotieren. Verlange Zusatzblätter, falls du mehr Platz brauchst.

5 **1. Sprachidentifikation (5 Punkte)** Was braucht es, um die Sprache eines unbekannten Dokuments automatisch zu bestimmen? Wie sieht ein einfaches Verfahren aus?

- Sprachidentifikation läuft typischerweise über den Vergleich von Buchstaben-Folgen (Buchstaben-n-Gramme). n wird normalerweise zwischen 2 und 5 gewählt, typischerweise 3.
- Man zählt in einem Korpus mit gegebener Sprache, welche Buchstaben-3-Gramme wie häufig vorkommen. Für die Sprachidentifikation wählt man die häufigsten 3-Gramme.
- Bei der Sprachidentifikation eines gegebenen Dokuments (oder Textabschnitts) vergleicht man die Häufigkeit der im Dokument vorkommenden 3-Gramme mit den gespeicherten 3-Grammen für verschiedene Sprache (Profil). Die Sprache mit dem ähnlichsten 3-Gramm-Profil ist die Sprache des Dokuments.

2. Wortartenbestimmung (10 Punkte)

6 (a) Welches sind die Wortarten (Part-of-Speech Tags) für das Wort "zu" in den folgenden Sätzen. Gib jeweils die Wortart nach dem Stuttgart-Tübingen-Tagset (gemäss Referenzkarte) an und begründe Deine Entscheidung kurz.

a) Die Amerikaner müssen hart arbeiten, um das Land wieder auf den richtigen Kurs **zu**/PTKZU bringen.

PTKZU: "zu" markiert den Infinitiv des Verbs "bringen".

b) Ausserdem sagen die Europäer Handelserleichterungen für Agrarprodukte **zu**/PTKVZ .

PTKVZ: "zu" ist abgetrenntes Präfix des Verbs "zusagen".

c) Die Ideale, **zu**/APPR denen sich damals die Sachsen und die Tschechen bekannten, waren identisch.

APPR: "zu" ist eine Präposition, die zusammen mit dem Relativpronomen "denen" eine PP bildet.

d) Den einen geht der Schritt **zu**/PTKA weit, den anderen nicht weit genug.

PTKA: "zu" ist ein Partikel, der das Adjektiv "weit" modifiziert.

e) Später behauptete die japanische Regierung jedoch, die vier südlichen Inseln gehörten nicht **zu**/APPR den Kurilen.

APPR: "zu" ist eine Präposition, die zusammen mit der NP "den Kurilen" eine PP bildet und deren Kasus Dativ bestimmt.

f) Die Heilsfront forderte Kafi auf, **zu**/APPR demokratischen Reformen zurückzukehren.

APPR: "zu" ist eine Präposition, die zusammen mit der Plural-NP "demokratischen Reformen" eine PP bildet und deren Kasus Dativ bestimmt.

2 (b) Wie gut sind gängige PoS-Tagger für Zeitungstexte auf dem STTS-Tagset? Was bedeutet das für die durchschnittliche Fehlerzahl pro Satz?

Gängige PoS-Tagger erreichen bei Zeitungstexten des Deutschen eine Genauigkeit von 95-97% (+- 1%). Eine Genauigkeit von 95% bedeutet, dass im Durchschnitt 1 Tagging-Fehler pro Satz vorliegt, wenn die durchschnittliche Satzlänge 20 Wörter beträgt. Analog: Eine Genauigkeit von 96% bedeutet 1 Fehler pro Satz bei einer durchschnittlichen Satzlänge von 25 Wörtern, etc.

2 (c) Von welchen Faktoren hängt die Genauigkeit eines statistischen PoS-Taggers ab? Nenne 3 Faktoren.

- Grösse des Trainingskorpus
- Tagging-Qualität des Trainingskorpus
- Nähe des Genre des Trainingskorpus zum zu taggenden Text
- Anzahl unbekannter Wörter im Text
- Grösse des beachteten Kontextfensters
- Anzahl Wörter im Text, die mehrere PoS-Tags haben können (Grad der PoS-Mehrdeutigkeit)
- Eigenschaften der Sprache (Morphologie-Reichtum, Wortstellung, etc.)

6 **3. Morphologie (6 Punkte)** Was ist das Ergebnis der computerlinguistischen Morphologieanalyse (**Lemmatisierung, Kompositasegmentierung, morpho-syntaktische Merkmale**) der folgenden Wortformen? Notiere die morpho-syntaktischen Merkmale entsprechend dem STTS-Tagset. Diskutiere kurz Schwierigkeiten und Alternativanalysen.

a) *einzuschlafen* Morphologische Analyse: –

VVIZU - Lemma: *einschlafen* - Schwierigkeit: Sonderform für Verben mit trennbarem Präfix

b) *Reiheneckhäusern* Morphologische Analyse: Dat.Pl.Neut

NN - Lemma: *Reihe-n#eck#haus* (mit Morphem-Grenzen) Schwierigkeit: Umlaut im Plural; Alternativanalyse: *Reihe#neck#haus*

c) *im* Morphologische Analyse: Neut.Sg.Dat / Masc.Sg.Dat

APPRART - Lemma: *in d(as)* ODER *in d(er)*: Schwierigkeit: Bei flektierten Funktionswörtern ist die Lemmabestimmung nicht so klar gegeben.

4. Recall, Precision und Accuracy (8 Punkte) In einem Studienprojekt musste für alle grossgeschriebenen Wörter eines Text bestimmt werden, ob ein geographischer Name vorliegt oder nicht. Für das Testset von 9000 Wörtern hat Alphonsos Programm insgesamt 1000 positive (d.h. Geo-Name) und 8000 negative (d.h. kein Geo-Name) Klassifizierungen berechnet.

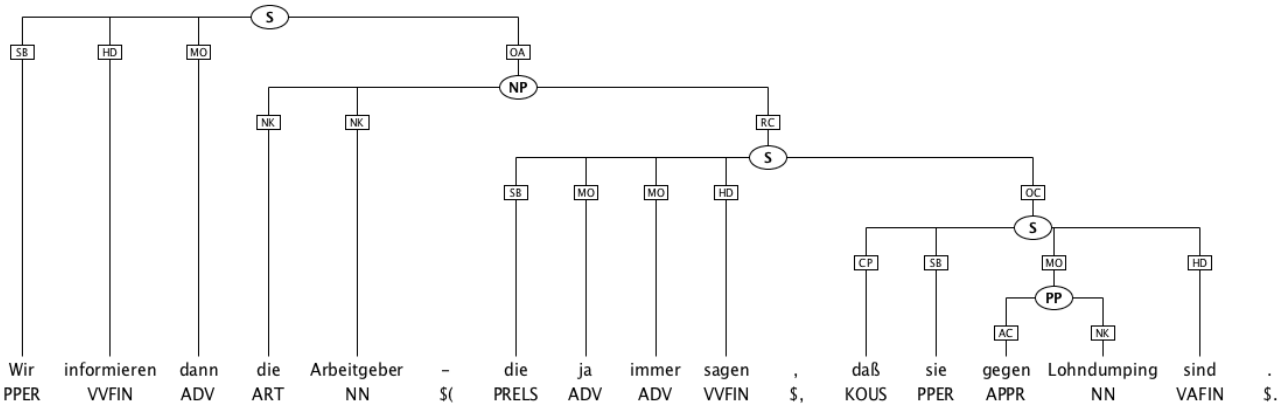
4 (a) Wie hoch ist **Recall**, **Precision** und **Accuracy**, wenn sein Programm in 800 Fällen Geo-Namen und in 7000 Fällen Nicht-Geo-Namen korrekt klassifiziert hat?

Precision= $800/(800+200) = 8/10 = 0.8$
Recall= $800/(800+1000) = 8/18 = 4/9 = 0.44$
Accuracy = $7800/9000 = 78/90 = 39/45 = 13/15$

4 (b) Bei der Durchsicht der Resultate stellt sich heraus, dass das Testset Fehler enthält und 100 Falschpositive eigentlich TP (True Positive) sein müssen. Wie hoch ist **Recall** und **Precision** mit dem korrigierten Testset? Wie hoch ist die **Accuracy**?

Precision= $900/(900+100) = 9/10 = 0.9$
Recall= $900/(900+1000) = 9/19 = 0.47$
Accuracy = $7900/9000 = 79/90$

16 **5. Annotation eines Syntaxbaums (16 Punkte)** Annotiere mit Hilfe der „Referenzkarte zur Annotation“ den folgenden deutschen Satz entsprechend der TIGER-Annotationsrichtlinien (Bitte Baum mit **STTS-Wortartentags**, **Phrasen** und **Funktionen**, aber **ohne** Morphologie zeichnen!)):



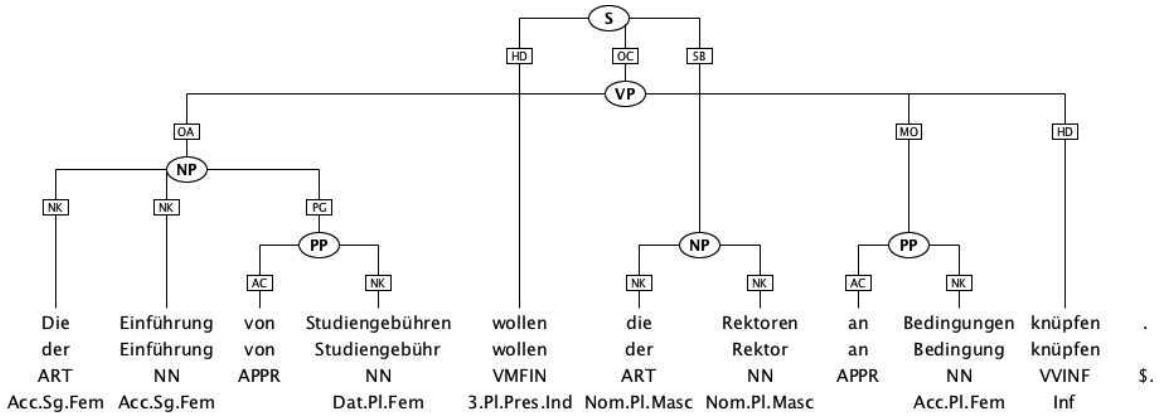
Korrektur-Richtlinien:
Pro STTS-Tag: 0.25 (4.25)
Pro Funktionslabel: 0.5 (9)
Pro Phrase: 0.5 (2.5)
RelativSatz korrekt: +0.25

5 **6. TF und IDF (5 Punkte)** Was bedeutet die Grundformel der Relevanzbestimmung im Information Retrieval (TF * IDF)? Welche Ideen stecken dahinter?

Vgl. auch Unterlagen.

- TF = Term Frequency = Die Häufigkeit des Auftretens eines Terms in einem Dokument.
- IDF = Inverse Document Frequency = Der (logarithmisierte) Anteil der Anzahl Dokumente, welche einen Term enthalten, in Bezug auf die Gesamtzahl der Dokumente.
- Idee: Ein Dokument D ist umso relevanter für einen Term T, umso häufiger T in D auftritt und in je weniger Dokumenten aus der Dokumentsammlung T auftritt. Durch die Kombination von TF und IDF erhalten häufige Stoppwörter eine geeignet tiefe Relevanz.

7. Überkreuzende Kanten (8 Punkte)



4 (a) Wieso gibt es bei folgendem Baum eine überkreuzende Kante? Motiviere diese Entscheidung der TIGER-Richtlinien.

- Dieser Satz weicht von der Standard-Wortstellung des Deutschen Subjekt-Verb-Objekt ab. Das Akkusativobjekt ist an den Anfang des Satzes gerückt worden, dadurch verschiebt sich das Subjekt hinter das finite Verb.
- Die Verbalphrase umfasst - nach den TIGER-Richtlinien - nicht das finite Verb sondern nur das infinitive Verb mit seinen Objekten und Ergänzungen. Das Akkusativ-Objekt am Satzanfang muss also in die selbe VP wie das infinite Verb am Satzende.
- Das Subjekt, das finite Verb und die VP werden als Tochterknoten zum Satzknoten markiert. Durch alle diese Bedingungen ergeben sich hier überkreuzende Kanten.

4 (b) Inwiefern bieten überkreuzende Kanten Schwierigkeiten für die maschinelle Verarbeitung (Parsen, Generieren, Suche in Baumbanken)?

- Bäume mit überkreuzenden Kanten können nicht (allein) durch kontextfreie Regeln beschrieben werden. Deshalb werden das Parsen und die Generierung aufwändiger (die effizientesten Verfahren für kontextfreie Grammatiken können nicht eingesetzt werden).
- Obenstehendes bedeutet auch, dass solche Bäume nicht direkt durch einfache geschachtelte XML-Strukturen abgebildet werden können. Man muss Knotenreferenzen (refs) hinzufügen. Das Durchsuchen von Baumbanken mit überkreuzenden

Kanten ist schwieriger, weil es unklare Reihenfolge-Beziehungen gibt. Im Beispielsatz gibt es keine lineare Präzedenz für die Subjekt-NP und die VP. Keine von beiden steht vor der anderen. Die VP beginnt vor der Subjekt-NP, aber endet nach der Subjekt-NP. Man muss deshalb Konventionen für die Präzedenz von Knoten mit überkreuzenden Kanten aufstellen, z.B. eine Konstituente a steht vor einer Konstituente b, falls die am weitesten links stehende Tochterkonstituente von a vor der am weitesten links stehenden Tochterkonstituente von b steht.

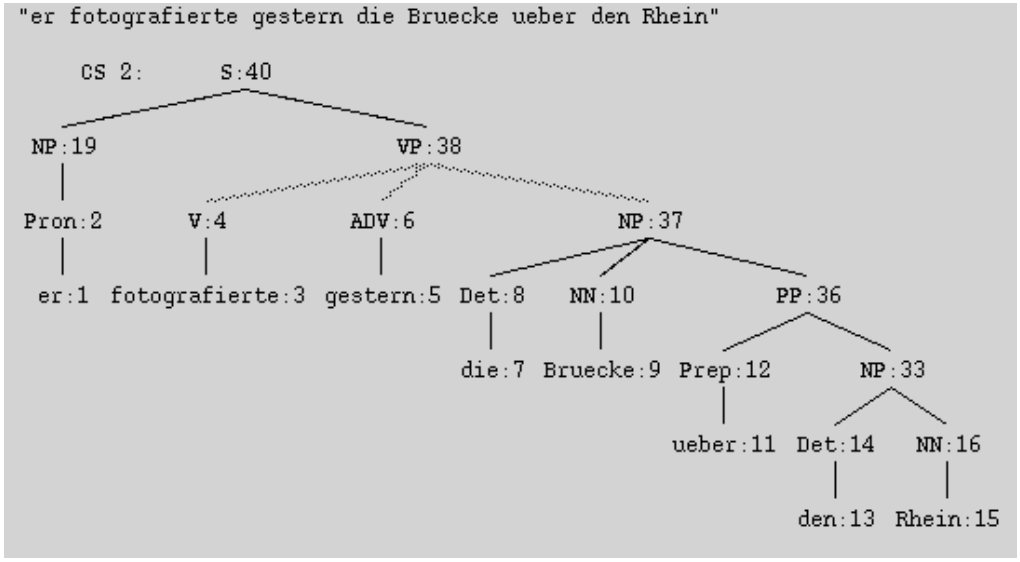
- Der Transfer von Bäumen mit überkreuzenden Kanten in andere Sprachen (bei regel-basierter MÜ) ist aufwändiger.

8. LFG (19 Punkte) Gegeben seien die folgenden Grammatikregeln mit passendem Lexikon:

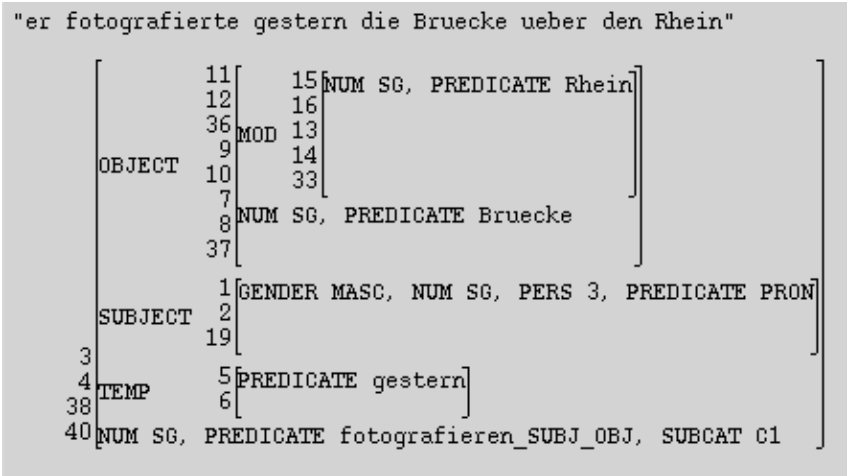
```
"Ein Satz besteht aus einer Subjekt-NP und einer VP"
S --> NP: (^ SUBJECT) = ! ;
      VP: ^ = !
          (^ SUBJECT NUM) = (! NUM) .

"Eine NP besteht aus einem Determiner und einem Nomen"
"Eine NP besteht aus einem Determiner und einem Nomen und einer PP"
"Eine NP besteht aus einem Pronomen"
NP --> { Det
         NN
         (PP)
       | Pron
       }.

"Eine VP besteht aus einem Verb (mit optional ADV und/oder PP)"
"Eine VP besteht aus einem Verb, einem ADV und einer NP (mit optional einer PP)"
VP --> { V: ^=!
            (^ SUBCAT) = C0 ;
       (ADV: (^TEMP) = ! )
       (PP)
       |
       V: ^=!
          (^ SUBCAT) = C1 ;
       ADV: (^TEMP) = ! ;
       NP: (^OBJECT) = ! ;
       (PP)
       }.

"Eine PP besteht aus einer Präposition und einer NP"
PP --> Prep
       NP: (^ MOD) = ! .
```

12 (a) Notiere die C-Struktur und die F-Struktur für den Satz *„Er fotografierte gestern die Brücke über den Rhein“*?

Der Satz ist syntaktisch mehrdeutig. Die intendierte Lesart hat folgende Struktur:



Der genaue Inhalt der F-Struktur ist teilweise vom Lexikon abhängig (hier z.B. die morphologischen Merkmale sowie die Werte des Merkmals PREDICATE). Eine mögliche Lösung wäre:



5 (b) Warum wird der Satz *„Er fotografierte die Brücke“* von einem Parser mit dieser Grammatik nicht akzeptiert?

Die obige Grammatik fordert entweder ein Adverb und eine Objekt-NP oder kein Objekt. Da hier aber ein Akkusativ-Objekt und kein Adverb vorliegt, passt keine der beiden VP-Regeln.

2 (c) Warum verwendet man Grammatikregeln mit Merkmalstrukturen?

- um Grammatikregeln kompakter und verallgemeinernd schreiben zu können.
- um Kongruenz zwischen Konstituenten explizit ausdrücken zu können.
- um Unifikation als grundlegende Operation bei der Syntaxanalyse zu ermöglichen.

2 **9. Merkmalstrukturen (2 Punkte)** Gegeben sind die folgenden beiden Merkmalstrukturen:

```
M1 = [SUBJECT [AGR  [NUM=PL  PERS=3]]]
M2 = [SUBJECT [AGR]  OBJECT [NUM=SG] ]
```

Was ist das Ergebnis der Unifikation der beiden?

Die Einrücktiefe gibt die Schachtelung an.

```
[SUBJECT
   [AGR
      [NUM=PL
       PERS=3]]
 OBJECT
   [NUM=SG]]
```

3 **10. Parallele Korpora (3 Punkte)** Warum sind parallele Korpora interessant für die Sprachtechnologie? Gib drei Gründe.

- als Trainingskorpora für die statistische maschinelle Übersetzung.
- als Evaluationskorpora für regel-basierte und statistische maschinelle Übersetzung.
- als Korpora für die Disambiguierung von Wort-Lesarten (z.B. auch Erkennung von Eigennamen).
- als Korpora für die Extraktion von bilingualen Wörterbüchern.
- als Korpora für verbesserte Wissensextraktion.

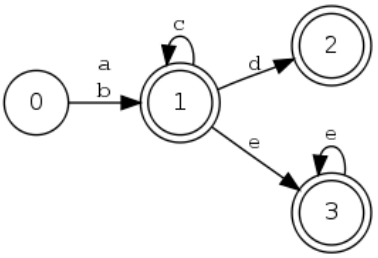
3 **11. Sprachmodelle (3 Punkte)** Wie wird ein statistisches Sprachmodell aufgebaut? Welche Rolle spielt es bei der Statistischen MÜ?

- Ein statistisches Sprachmodell wird erstellt, indem die Wortfolgen (Wort-n-Gramme) in grossen Korpora ausgezählt und zu Wahrscheinlichkeiten verrechnet werden.
- Das Sprachmodell dient dazu, Sätze mit typischen (idealerweise: natürlichen) Wortreihenfolgen zu finden.
- Das Sprachmodell ist - grob gesagt - ein statistischer Grammatikprüfer.
- Bei der Statistischen MÜ wird das Sprachmodell eingesetzt, um die vom Übersetzungsmodell gelieferten Varianten bzgl. der Wortreihenfolge in der Zielsprache zu gewichten.

5 **12. Reguläre Ausdrücke und endliche Automaten (5 Punkte)** Zeichne ein Zustandsübergangsdiagramm (deterministisch oder nicht-deterministisch) eines endlichen Automaten, welcher dieselbe Sprache umfasst wie folgender regulärer Ausdruck R:

```
R = (a|b)c*(d|e*)
```

Als deterministischer Automat (kondensierte Beschriftung):



Als ein nicht-deterministischer Automat: