LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

MODULTEILPRÜFUNG ZUR VORLESUNG ,,KORPUSBEARBEITUNG IN DER COMPUTERLINGUISTIK“ SS 2016

DR. MAXIMILIAN HADERSBECK

KLAUSUR AM 11.07.2016

VORNAME:	
NACHNAME:	
MATRIKELNUMMER:	
STUDIENGANG:	B.Sc. Computerlinguistik, B.Sc. Informatik, Magister

Bitte unterstreichen Sie den für Sie zutreffenden Studiengang!

Die Klausur besteht aus **4 Aufgaben**. Die Punktzahl ist bei jeder Aufgabe angegeben. Die Bearbeitungsdauer beträgt **60 Minuten**. Bitte überprüfen Sie, ob Sie ein vollständiges Exemplar erhalten haben. **Tragen Sie die Lösungen in den dafür vorgesehenen Raum im Anschluss an jede Aufgabe ein. Falls der Platz für Ihre Lösung nicht ausreicht, benutzen Sie bitte nur die ausgeteilten Zusatzblätter!**

Aufgabe	mögliche Punkte	erreichte Punkte
1. Zeichensatz/Satzende	13	
2. UNIX Befehle	24	
3. UNITEX	9	
4. XML	9	
Summe	55	
Note		

Einwilligungserklärung

Hiermit stimme ich einer Veröffentlichung meines Klausurergebnisses in der Vorlesung ,,Korpusbearbeitung in der Computerlinguistik“ vom 11.07.2016 unter Verwendung meiner Matrikelnummer im Internet zu.

Datum: _____ Unterschrift: _____

NAME:

1 Zeichensatz/Satzende

Aufgabe 1 Zeichensatz/Satzende

1. Nennen Sie den `UNIX` Befehl, der die `ISO-Latin1` Datei `sz.txt` nicht überschreibt, sie aber in eine `UTF-8` Datei mit dem Namen `sz_utf8.txt` konvertiert?

(2 Punkte)

2. Wird bei der Konvertierung von `UTF-16BE` nach `UTF-16LE` die Datei eher größer oder kleiner.

(1 Punkt)

3. Welche Kodierung der Datei `text.txt` liegt vor, dessen octal-dump sie hier sehen? Welches Wort ist hier gespeichert? Handelt es sich um eine UNIX oder eine Windows Datei?

```
max@linux> od -a text.txt
0000000   k   C   6   n   n   e   n  cr  nl
0000010
```

(2 Punkte)

4. Was sagen diese Statements am Anfang einer Datei über die Datei aus?

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TEI>
<TEI>
```

(2 Punkte)

5. Was gibt folgender Befehl aus?
   `bash>echo "abcäöü123" | LC_ALL=en_US tr -d '[:digit:]'`

(1 Punkt)

6. Was gibt folgender Befehl aus?
   `bash>echo -e "Mann\nMünchen\nMächte\nMaus\nMist" | LC_ALL=de_DE.UTF-8 sort`

(2 Punkte)

7. Nennen sie den Unterschied zwischen einem Wort und einem Token. Bei der Bestimmung von Satzende ist es wichtig Abkürzungen zu erkennen. Nennen Sie zwei Strategien zur Erkennung von Abkürzungen.

(3 Punkte)

13 PUNKTE

2 UNIX Befehle

Aufgabe 2 UNIX Befehle

1. Mit welchem UNIX Befehl kann man den nachfolgenden Geheimschrift-Text von Ludwig Wittgenstein (hier nur Kleinbuchstaben) lesbar machen. Nennen sie den Namen des UNIX Befehl's und die Idee, die dahinter steckt.
   `zo hxsdvihgvn zyvi ufspv rxs wvn emidfiu wvi uvrtsvrg`

(2 Punkte)

2. In ihrem `daten` Verzeichnis sind 5000 Text-Dateien mit den Namen `1.txt .... 5000.txt`. Wie lautet der UNIX Befehl, der die Dateien mit den Nummern von 2000 bis 2999 in das Verzeichnis `done` verschiebt?

(2 Punkte)

3. Welche Ausgabe erzeugt folgender Befehl: `paste f1.txt f2.txt`, wenn in der Datei `f1.txt` die Zahlen `1,2,3` und in der Datei `f2.txt` die Buchstaben `a,b,c` zeilenweise gespeichert sind.

(2 Punkte)

4. Wie können nur die letzte Zeile der Datei `f1.txt` und die letzte Zeile der Datei `f2.txt` in die Datei ende.txt mit **zwei** UNIX Befehlen kopiert werden.

(2 Punkte)

5. Entwickeln Sie eine Frequenzliste aller Wörter (ohne Interpunktionszeichen) die in der Datei `taz.txt` vorkommen. Wie lauten die UNIX Befehle in einer Pipe, die folgende Aufgaben erfüllen: Die Frequenzliste soll der Häufigkeit nach absteigend sortiert und in einer Datei mit dem Namen `sorted.txt` gespeichert werden. Die Groß/Kleinschreibung der Wörter soll beim Zählen und Sortieren nicht beachtet werden.

(5 Punkte)

6. Sie haben eine komprimierte `gzip` Datei `daten.txt.gz`. Wie können Sie, ohne die Datei zu entpacken, die ersten 1KB der Datei `daten.txt.gz` in der Datei `first.txt` speichern?

(3 Punkte)

7. Sie sollen alle Dateien im Verzeichnis `seminar` in einem `tar` Archiv `seminar.tar` sichern.

(2 Punkte)

8. Mit welchem UNIX Kommando kann man alle Zeilen eines Textes ausgeben, in denen ein Wort aus exakt 3 Großbuchstaben vorkommt? Wie lautet der UNIX-Befehl?

(1 Punkt)

9. Welche Funktionalität hat das Unix Kommando `agrep` ? Das erste Argument dieses Kommandos definiert die sogenannten "Kosten". Was versteht man darunter und welche 3 Arten von Kosten werden unterschieden? Welchen Wert muss die Option `cost` im nachfolgenden Befehl haben, damit folgender Befehl eine Ausgabe gibt:
`echo "Schweinsteiger"  | agrep -cost "Schleimsteiger"`

(5 Punkte)
24 PUNKTE

3 Fragen zu UNITEX

Aufgabe 3 UNITEX

1. Was ist eine lokale Grammatik?

(2 Punkte)

2. Was macht das Preprocessing bei UNITEX?

(2 Punkte)

3. Zeichnen Sie den Graphen einer lokalen Grammatik, die folgende Wortfolgen findet: Ein Artikel (DET) wird von einem Nomen (N), einem Verb (V) und beliebig vielen Adverbien (ADV) gefolgt. Die Adverbien werden mit Kommas getrennt und vor dem letzen Adverb steht die Konjunktion `und`
   z.B. `"die Sonne leuchtet hell"`
   z.B. `"die Sonne leuchtet hell, gelb und schön"`
   z.B. `"die Sonne leuchtet stark, hell, rot und schön"`

(2 Punkte)

4. Was ist ein DELA? Ein Lexikoneintrag im DELA kann pro Wort bis zu 5 Einträge speichern. Nennen Sie mindestens drei Einträge. Wie könnte der DELA-eintrag mit drei Einträgen zu den beiden Worten: `schöner` und `Häuser` aussehen?

(3 Punkte)
9 PUNKTE

4 XML

Aufgabe 4 XML

1. Wann nennt man ein XML-Dokument wohlgeformt? Nennen Sie 5 Kriterien.

(5 Punkte)

2. Worin widerspricht das folgende XML-Fragment der XML-Wohlgeformtheit? Bitte korrigieren Sie im Textfragment:

```
<TEI>
<Header>

    <author>Ludwig Wittgenstein<audor/>

 </teiHeader>

         <ab n="Ts-213,61r[3] abnr="1">

            <S "Ts-213,61r[3]_1">(Wenn wir sagen, Satz ist jedes Zeichen, womit

wir etwas meinen, so könn<lb rend="shyphen">te man fragen: was meinen wir und

<emph rend="space">wann</emph> meinen wir es?</S>

            <s n="Ts-213,61r[3]_2" >Während wir das</lb> Zeichen

geben? <abbr type="abb">u.s.w.</abb>)</s>

        </ba>

</body>

</TIE>
```

(4 Punkte)
9 PUNKTE

es folgt ein Zusatzblatt für weitere Bemerkungen/Notizen