

# Einführung in die Sprachwissenschaft

## Übung 1

Oktober 2020

Lösen Sie die folgende Übungen individuell. Wenn die Anzahl der Übereinstimmungen in den gewählten Beispielen größer als durch Zufall zu erwarten ist, müssen Sie die Aufgabe erneut einreichen, oder wird diese als nicht-bestanden bewertet.

Laden Sie die Lösung als eine PDF-Datei in Moodle hoch. Bei 40 bis 45 Punkte bekommen Sie eine 1,0; bei 20 Punkte eine 4,0.

1. Wählen Sie ein willkürliches Wort, dass mindestens ein Zeichen enthält, dass nicht zum ASCII-Zeichensatz gehört. Schreiben Sie für jeden Buchstaben dieses Wortes die Zahl in hexadezimal, dezimal und binär auf. Nutzen Sie hierfür Codepage 850, Codepage 852, Codepage 437 oder ISO 646. Die Tabellen finden Sie z.B. auf Wikipedia. (10 Punkte)
2. Schreiben Sie für jede Zahl aus der vorigen Aufgabe auf, wie diese Zahl in ISO 8859-1 interpretiert wird. (10 Punkte)
3. Ist die Folge der Binärzahlen aus der ersten Übung eine korrekte Folge in UTF8? Wenn nein, argumentieren Sie, warum das nicht der Fall ist. Wenn ja, schreiben Sie die Buchstaben auf, die dort stehen, wenn man die Zahlen als UTF8 Reihenfolge liest. (5 Punkte)
4. Suchen Sie einen Text (Oder Chatnachricht, oder E-Mail) in dem ein Zeichen falsch dargestellt ist. Fügen Sie den Text (kurzer Ausschnitt mit dem Fehler reicht) oder ein Screenshot in Ihrer Abgabe ein. Rekonstruieren Sie erstens mit welcher Kodierung der Text ursprünglich geschrieben wurde, und zweitens, wie der Text bei der Darstellung interpretiert wurde. Zeigen Sie hierbei möglichst genau, welche Zahl falsch interpretiert wurde. (10 Punkte)
5. Wie viel Byte pro Buchstaben braucht man etwa zum Speichern eines einfachen englischen Textes, wenn man reines Unicode verwenden würde? Und wie viel zum Speichern eines russischen Textes? Wie viel zum Speichern eines philippinischen Textes (Tagalog)? Siehe auch: <http://www.utf8-zeichentabelle.de/unicode-utf8-table.pl> (5 Punkte)
6. BONUS-AUFGABE: In UTF-8 brauchen Buchstaben eine unterschiedliche Zahl von Bits. Wie würde eine optimale (im Hinblick auf den Speicherbedarf) Kodierung aussehen? Welche wichtige Eigenschaft muss eine solche Kodierung erfüllen? Entwerfen Sie eine optimale Kodierung für eine Sprache mit nur den 5 Buchstaben A,B,C,D,E, wobei A und E häufiger vorkommen als B,C, und D. (5 Punkte)