Beispiele

Die **.sentences.txt*-Dateien sind alle http://corpora2.informatik.uni-leipzig.de/download.html entnommen. Die Datei *webseite_badw.html* enthält den Quellcode der Startseite von http://www.badw.de vom 16.07.15.

Für reine Suchen können Sie AntConc oder einen Texteditor (z. B. Notepad++) verwenden.
Zum Suchen und Ersetzen ist AntConc nicht geeignet. Verwenden Sie hierfür einen Texteditor.

In AntConc bleiben Sie zum Suchen bitte im ersten Karteireiter „concordance“, in welchem sich unten das Suchfeld befindet (siehe auch Präsentation). In Notepad++ gelangen Sie über STRG F ins Suchfenster bzw. über das Fernglas-Symbol in der Leiste.

In Notepad++ sollte die Option „am Ende von vorn beginnen“ (engl.: wrap around) aktiviert sein.
Die Option „Reguläre Ausdrücke“ bzw. „Regex“ ist in beiden Programmen obligatorisch!

Beispiel 1:

a) Wie lautet der Ausdruck *b{2}|[^b]{2}* in natürlicher Sprache? Versuchen Sie es auf Englisch.
b) Suchen Sie alle Wortbildungen auf *–keit* bzw. *–heit* aus der Beispieldatei *deu_wikipedia_2010_10K-sentences.txt*
c) Suchen Sie nach den Wörtern *Laus*, *Maus*, *Haus* und wenden Sie dabei 3 unterschiedliche Lösungen an (Tipp: Alternativen).

Beispiel 2:

Konvertieren Sie die Ortsangaben aus *OrterNeu.csv* in SQL-Statements. Die erforderlichen Werte stehen in den ersten drei Spalten der csv-Datei; die Spalten sind jeweils durch ein Semikolon getrennt. Der Rest der Zeile ist irrelevant. Die Ausgabe soll folgendermaßen aussehen:

INSERT INTO tbl_orte (Ort,Kreis,Planquadrat) VALUES ('X','X','X');

X steht hier jeweils für den variablen Wert (Treffer) aus der regex-Suche, die aus den ersten drei Spalten gelesen werden müssen über ein entsprechendes Suchmuster im regex.

Beispiel 3:

Entfernen Sie alle Kommentare (eingeschlossen in *<!--* und *-->*), JavaScript-Blöcke (eingeschlossen in *<script…>* und *</script>*) sowie HTML-Tags (eingeschlossen in spitzen Klammern *<* und *>)* der Datei *webseite_badw.html* sowie alle unnötigen Zeilenumbrüche und Leerzeilen, sodass der reine Text der Webseite in ca. 150 Zeilen angezeigt wird, jedoch dennoch strukturiert lesbar ist (d.h.: es darf nicht alles in einer einzigen Zeile stehen!).

Beispiel 4:

Dieses Beispiel beschäftigt sich mit dem Türkischen. Auch ohne Türkischkenntnisse ist die Aufgabe mit den nachfolgenden Angaben lösbar, da diese „nur" in einen entsprechenden regex umgesetzt werden müssen.
Im Türkischen findet die Verbalnegation mit dem vokalharmonischen Morphem *–mX* statt, das die Allomorphe *–mi, -mu, -mü* und *–mı* besitzt (obiges X steht für die vier unterschiedlichen Vokale).
Finden Sie alle Präsensformen, die nicht negiert sind. Präsensformen enthalten das Morphem *–yor*, nachdem die Person sowie u. U. weitere Morpheme stehen können (anders gesagt: nach *yor* kann noch etwas stehen).
Der Bindestrich in den Angaben ist der in der Linguistik übliche Morphemtrenner. Er ist NICHT Teil des Wortes und somit auch nicht im regex enthalten!
Tip: benutzen Sie einen *lookaraound*.

Beispiel 5:

Erstellen Sie eine Wortliste aus *deu_wikipedia_2010_10K-sentences.txt*, die ausschließlich Wörter enthält: Entfernen Sie alles, was kein Wort sein kann (v.a. Interpunktionszeichen und Ziffern). Dabei dürfen Sie gern großzügig sein und Vorkommnisse wie z. B. *3D-Drucker* oder *20er-Jahre* missachten (wenige Wörter enthalten Ziffern).
Wenn in der Datei nur noch Wörter stehen (getrennt durch Leerzeichen bzw. Zeilenumbrüche), sind diese alle durch einen einzelnen Zeilenumbruch voneinander zu trennen. Fügen Sie die so entstandene Wortliste in Excel o.ä. ein und sortieren Sie die Spalte alphabetisch. Kopieren Sie das Ergebnis wieder in Notepad++ und entfernen Sie alle doppelten Einträge mit einem letzten regulären Ausdruck (Tipp: Verwenden Sie hierfür zwei Gruppen).