Aufgaben aus Kegli 2 (Korpuslinguistik)

Kapitel 1

1. Überprüfen Sie mithilfe einer beliebigen Suchmaschine den Sprachgebrauch im Internet. Finden Sie die Formen *wegen dem Regen* bzw. *wegen des Regens* und *Praktika* bzw. *Praktikas*? Wenn ja, welche der Alternativen wird häufiger verwendet? Um wie viel häufiger wird sie verwendet?
2. Bitten Sie zehn Personen, den Satz *Ich habe gestern noch die Datei...* mit dem Verb *downloaden* zu vervollständigen. Überprüfen Sie anschließend mithilfe einer beliebigen Suchmaschine, ob die genannten Formen auch im Internet belegt sind. Stimmen die Ergebnisse Ihrer Befragung mit den Ergebnissen Ihrer Internetrecherche überein?
3. Überlegen Sie sich, welche Art von Texten ein Korpus enthalten sollte, das dazu dient, die Fachsprache des Rechts zu untersuchen. Welche Textsorten sollten enthalten sein? Von wem sollten die Texte stammen?
   Wie unterscheidet sich dieses Korpus von einem Korpus der Jugendsprache, wie von einem Korpus der gesamten deutschen Sprache?
4. Versuchen Sie möglichst genau zu beschreiben, welche Äußerungen unter den Begriff Jugendsprache fallen. Versuchen Sie anschließend, objektive Kriterien festzulegen, die es ermöglichen, jugendsprachliche Texte von anderen Texten abzugrenzen.
5. Stellen Sie sich vor, Sie sollen mit minimalem Aufwand ein Korpus zur Fachsprache der Medizin mit 10.000 Textwörtern aufbauen. In der Wahl der Texte sind Sie frei. Welche Art von Texten würden Sie in Ihr Korpus aufnehmen, in welchem Umfang und warum? Welche Texte würden Sie nicht aufnehmen? Warum?
6. Versuchen Sie, ein repräsentatives Korpus der Jugendsprache zu entwerfen. Wie würden Sie die einzelnen Textsorten bzw. Texte von verschiedenen Sprechergruppen gewichten?
7. Ist Ihnen das doppelte Perfekt oder die Rheinische Verlaufsform bekannt? Verwenden Sie diese grammatischen Formen?
   Überprüfen Sie mithilfe einer Suchmaschine im Internet, ob diese Verbformen im Deutschen verwendet werden. Schlagen Sie anschließend in zwei Grammatiken nach, ob Sie eine Beschreibung dieser Verbformen finden.

Kapitel 2

8. Versuchen Sie, die in den Aufgaben 3, 5 und 6 konzipierten Korpora zur Fachsprache des Rechts bzw. der Medizin und der Jugendsprache sinnvoll in Teilkorpora zu gliedern (vgl. Kapitel 1.3, 1.4).
9. Das International Corpus of English (ICE) enthält insgesamt 20 Millionen Textwörter. Das Korpus besteht aus zwanzig Teilkorpora aus Ländern, in denen Englisch die einzige oder eine der offiziellen Nationalsprachen ist. Jedes der Teilkorpora enthält zu 60% gesprochene und zu 40% geschriebene Sprache, die nach denselben Kriterien erhoben wurden.
   Handelt es sich beim ICE Ihrer Meinung nach um ein einsprachiges oder ein mehrsprachiges Korpus? Handelt es sich bei den Teilkorpora um parallele oder vergleichbare Korpora? Bitte begründen Sie Ihre Ansicht.

Kapitel 3

10. Wie viele Textwörter enthält der oben stehende Ausschnitt aus dem Mainzer Zeitungskorpus? Bitte ermitteln Sie die Zahl der Types (Lemma-Types) und Tokens für die im Ausschnitt enthaltenen Nomen. Zählen Sie Eigennamen zu den Nomen.
11. In einem Korpus A finden sich 80 Belege für das Wort BLUMENTOPF, in Korpus B 100 Belege für dasselbe Wort. Korpus A und Korpus B enthalten je eine Million Textwörter. Korpus C enthält ebenfalls 100 Belege für BLUMENTOPF, aber anderthalb Millionen Textwörter. In welchem der drei Korpora finden sich die meisten Belege für das Wort BLUMENTOPF?
12. Unten finden Sie die Ergebnisse aus dem Erlanger Dürer-Korpus (Müller 1993) und dem Würzburger Korpus der Wissensliteratur (Brendel et al. 1997) zur Wortbildung in frühneuhochdeutschen Fachtexten.
    In welchem der beiden Korpora finden sich die meisten Nominalisierungen mit den Suffixen *er*, *-heit/-keit* und *-ung* (Types, Tokens)? Bitte berechnen Sie zudem das Type-Token-Verhältnis für die einzelnen Suffixe und vergleichen Sie die Ergebnisse miteinander.

	Korpus	Dürerkorpus	Wissensliteratur
	Textwörter	444.000	1.073.000
-er	Types	93	510
	Tokens	700	4.505
-heit/-keit	Types	76	454
	Tokens	326	6.575
-ung	Types	193	1.025
	Tokens	2.443	5.213

13. Welche Bedeutungen haben die Wörter *Karte*, *decken* und *grün*? Wie werden sie verwendet? Im Zusammenhang mit welchen anderen Wörtern werden sie häufig verwendet? Gibt es feste Redewendungen? Welches der Wörter wird am häufigsten verwendet? Überprüfen Sie Ihre Intuition anhand eines Wörterbuchs und mithilfe einer Suchmaschine im Internet.
14. Bitte erstellen Sie eine Konkordanz für das Wort *Absatz*. Benutzen Sie dazu ein beliebiges Korpus mit Konkordanzfunktion wie das DWDS-Korpus oder ein Programm wie Cosmas II oder WebConc. Überprüfen Sie anschließend 50 Treffer. Welche Bedeutungsvarianten für *Absatz* finden Sie?
15. Erstellen Sie eine Konkordanz für die Wörter *Mann* und *Frau*. Benutzen Sie dazu ein beliebiges Korpus mit Konkordanzfunktion wie das DWDS-Korpus oder ein Programm wie Cosmas II oder WebConc. Überprüfen Sie anschließend je 50 Treffer. Welche Kollokationen finden Sie? Welche Eigenschaften werden Männern, welche Frauen zugeschrieben?
16. Bitte vergleichen Sie die Wortliste des IDS mit der Wortliste des Projekts Deutscher Wortschatz in Leipzig. Zu welchem Ergebnis kommen Sie?
17. Bitte erstellen Sie eine alphabetisch geordnete Frequenzliste (Wortform-Types) für den Textausschnitt aus dem Mainzer Zeitungskorpus (vgl. Abbildung 3, Kapitel 3.1).

Kapitel 4

18. Rekapitulieren Sie die Fragestellungen, die den Untersuchungen von Blaha et al. (2001), O'Halloran (2002) sowie Steyer (2002) zugrunde liegen. Verwenden Sie dazu die Informationen aus den Kapiteln 1.5, 3.2 und 3.6.
19. Bitte nehmen Sie für folgende Sätze eine manuelle Lemmatisierung vor.
    a. Die Katze hat viele Mäuse gefangen.
    b. Er war mit seinen Freunden in der Oper.
    c. Bist du gestern nach dem Kino bei Claudia gewesen?
20. Bitte nehmen Sie für die Sätze a-c aus Aufgabe 19 ein grammatisches Tagging vor. Bestimmen und taggen Sie die Wortarten sowie die Flexionsmerkmale. Benutzen Sie dazu folgende Tags aus dem Morphy-Tagset:
    - Wortarten: Adjektiv (ADJ), Adverb (ADV), definiter Artikel (ART-DEF), Hilfsverb (VER-AUX), indefiniter Artikel (ART-IND), Nomen (SUB), Präposition (PRP), Personalpronomen (PER), Possessivpronomen (POS), Vollverb (VER)
    - Genus: Femininum (FEM), Maskulinum (MAS), Neutrum (NEU)
    - Numerus: Plural (PLU), Singular (SIN)
    - Kasus: Akkusativ (AKK), Dativ (DAT), Genitiv (GEN), Nominativ (NOM)
    - Person: 1. Person (1), 2. Person (2), 3. Person (3)
    - Infinite Verbformen: Infinitiv (INF), Partizip Perfekt (PA2)
21. Ermitteln Sie alle Belege für nominale *-er*-Derivate aus folgenden Sätzen aus dem Mainzer Zeitungskorpus. Ordnen Sie den einzelnen Tokens den jeweiligen Type zu. Welche Probleme ergeben sich bei der Zuordnung?
    a. Sambstags hat der Babst etlich schreiben / in Teutschland / an Keys. M. Jhr Durchl. zu Greetz / Bayern und andere catholische Fürsten abgehen lassen / sich den Protestirenden und andern Anhengern zu wiedersetzen (Teilkorpus 1609, Ausgabe 30, Seite 7, Zeile 11).
    b. wie sie dann verschidene Personen wegen dero Spillen und Fluchen ergriffen / da dann die Spiller nach dem Zuchthause / die andere aber mit Geld gestraft (Teilkorpus 1700, Ausgabe 42, Extrablatt, Seite 5, Zeile 19).
    c. Der Unternehmer zeigte sich bei dieser Gelegenheit abermals nicht nur als einen der geübtesten Orgelspieler, sondern auch als einen vorzüglichen Componisten. (Teilkorpus 1850, Ausgabe 220, 2. Beilage, Seite 2, Spalte 1, Zeile 65)
    d. Vorläufig kommen vier moderne Omnibusse zum Einsatz, die bei Bedarf mit Anhänger versehen werden. (Teilkorpus 1950, Ausgabe 18, Seite 11, Spalte 3, Zeile 116)
22. Dokumentieren und klassifizieren Sie die Belege aus Aufgabe 21 in Kapitel 4.6 nach dem Schema in Abbildung 13.
23. Bitte berechnen Sie für die einzelnen Teilkorpora die Produktivität und das Type-Token-Verhältnis (vgl. Kapitel 3.2). Wie passen Ihre Ergebnisse zu der Beobachtung, dass sich das Wortbildungsmuster diachron ausgebreitet hat?
24. Bitte vergleichen Sie die Ergebnisse des Mainzer Zeitungskorpus mit den Ergebnissen zum Frühneuhochdeutschen im Erlanger und im Würzburger Korpus (vgl. Aufgabe 12, Kapitel 3.3). Was stellen Sie fest?
25. Nach der Auszählung des Mainzer Zeitungskorpus ergaben sich die unten stehenden Zahlen. In welchem der Teilkorpora finden sich die meisten Personenbezeichnungen (Types, Tokens), in welchem die meisten Objektbezeichnungen (Types, Tokens)?

	1609	1650	1700	1750	1800
Textwörter (in Tsd.)	98,9	98,3	98,0	102,8	101,1
Person Types	153	87	156	207	214
Person Tokens	640	303	574	688	608
Objekt Types	5	8	14	19	19
Objekt Tokens	52	48	47	54	24

	1850	1900	1950	2000
Textwörter (in Tsd.)	108,6	149,5	136,5	137,4
Person Types	313	446	498	687
Person Tokens	847	1.417	1.435	1.740
Objekt Types	20	74	68	56
Objekt Tokens	55	116	113	80

26. Bitte berechnen Sie das Type-Token-Verhältnis in den einzelnen Teilkorpora. Vergleichen Sie die Ergebnisse mit den Ergebnissen aus Aufgabe 23. Was stellen Sie fest?

Kapitel 5

27. Bitte erstellen Sie eine Internetkonkordanz für die Wörter *Mittel* und *Schein*. Überprüfen Sie anschließend 50 Treffer. Welche unterschiedlichen Bedeutungen finden Sie? Vergleichen Sie Ihre Ergebnisse mit den Wörterbucheinträgen von Mittel und Schein. Was stellen Sie fest?
28. Welche Pluralform ist üblich: *Kontos*, *Konti* oder *Konten*? Bitte überprüfen Sie die Pluralformen von Konto anhand des DWDS-Kernkorpus. Lassen sich Unterschiede in der Verwendung der Formen feststellen, was die Zeit und die Textsorte betrifft?
29. Probieren Sie die einfachen und komplexen Suchausdrücke in Tabelle 4 und 5 aus. Zu welchen Ergebnissen führen die Abfragen?
30. Formulieren Sie Suchabfragen für folgende Aufgabenstellungen. Zu welchen Ergebnissen kommen Sie?
    a. Suchen Sie in den belletristischen Texten nach Sätzen, die sowohl das Lexem *Buch* als auch das Lexem *lesen* enthalten.
    b. Suchen Sie nach Sätzen, in denen zwischen *große* und *Augen* maximal ein weiteres Wort steht.
    c. Suchen Sie alle Belege für das Lexem gehen sowie für die Wortformen *gehen*, *ging*, *gegangen* aus den Gebrauchstexten der Jahre 1980 bis 1990.
    d. Suchen Sie alle Sätze in Zeitungstexten, die das Lexem *Bundesrepublik*, aber nicht das Lexem *Deutschland* enthalten.
    e. Suchen Sie alle Belege der Jahre 1950 bis 1960, die mit der Buchstabenfolge *dy* beginnen bzw. enden.
31. Ermitteln Sie die Kollokationen für *Hund*, *Katze* und *Maus*. Bitte vergleichen Sie die Ergebnisse. Was fällt auf?
32. Unternehmen Sie im Grimm-Korpus eine Suche nach dem Wortende *-chen (*chen)* Bitte vergleichen Sie anschließend die ersten 20 Treffer mit den Ergebnissen in (34). Was fällt auf?

33. Probieren Sie die Suchausdrücke in Tabelle 6 und 7 im Korpus der Belletristik/Trivialliteratur aus. Wie viele Types und Tokens finden Sie?
34. Formulieren Sie Suchabfragen für folgende Aufgabenstellungen. Zu welchen Ergebnisse kommen Sie?
    a. Suchen Sie in allen öffentlich zugänglichen Korpora nach der doppelten Pluralform *Praktikas*.
    b. Suchen Sie in allen öffentlich zugänglichen Korpora nach den Lexemen *downloaden* und *updaten*. Erstellen Sie ausgehend von Ihren Ergebnissen ein Flexionsparadigma für die beiden Verben.
    c. Ermitteln Sie im Bonner Zeitungskorpus alle Wortformen, die mit *ein-* beginnen und mit *-ung* enden.
    d. Suchen Sie in allen öffentlich zugänglichen Korpora nach Wortformen, die den Wortbestandteil *linguistik* enthalten.