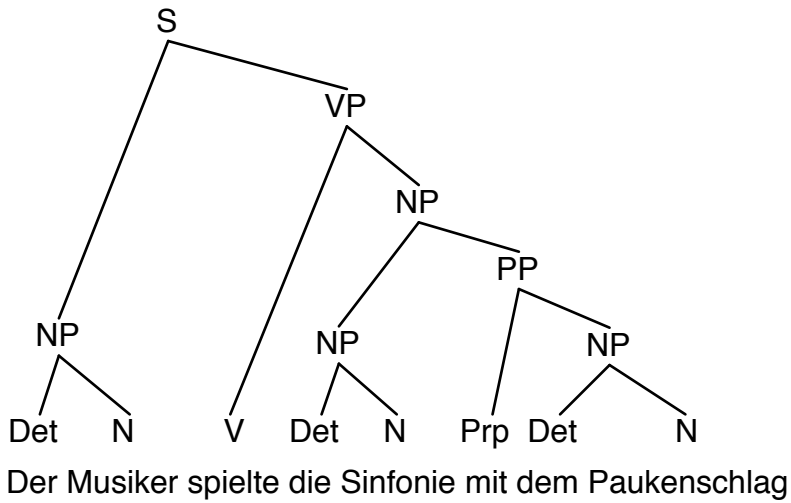
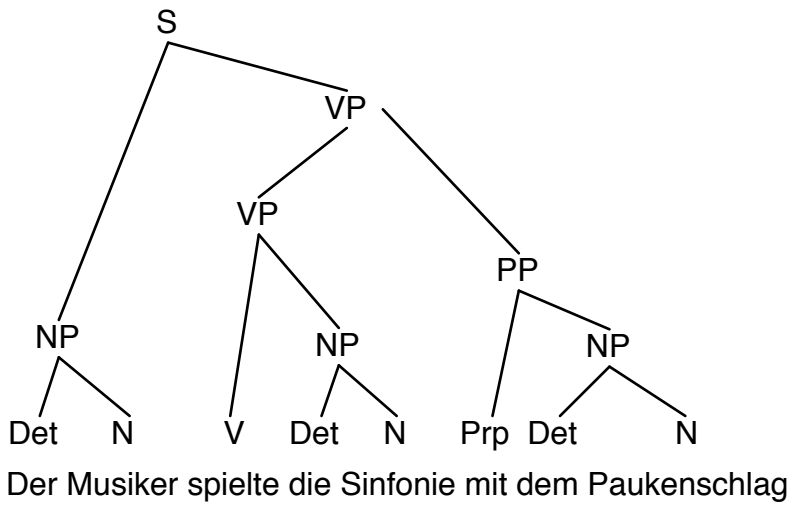


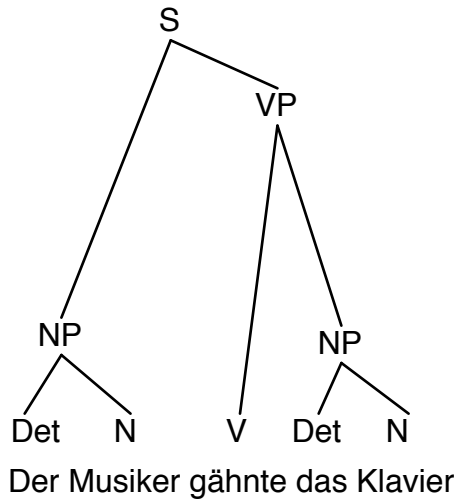
Musterlösung zur Übungsklausur

1. (15 Punkte)

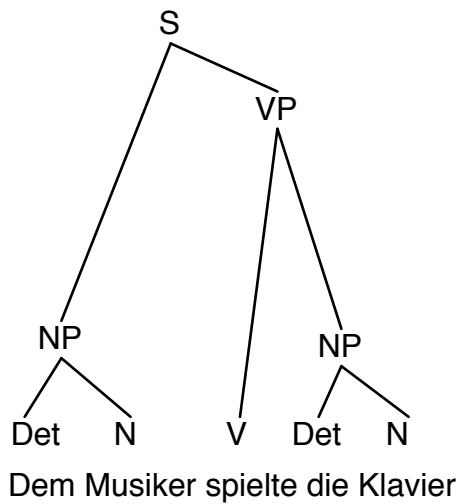


2. (10 Punkte)

Zum Beispiel:



Grammatik weiß nicht über Subkategorisierung Bescheid: erlaubt transitive Verwendung von intransitiven Verben.



Grammatik weiß nicht über Artikel-Nomen-Kongruenz Bescheid: erlaubt Kombination aus Det+N, die nicht in Kasus, Genus, oder Numerus übereinstimmen.

3. (5 Punkte)

- Computerlinguistik als Sprachwissenschaft. Ziel: Systematische vollständige und überprüfbare Modelle sprachlicher Phänomene. Erhebung und Erfassung sprachlicher Daten.
- Computerlinguistik als Ingenieurwissenschaft. Ziel: Realisierung von Computersystemen zur automatischen Verarbeitung menschlicher Sprache. Unterstützung menschlicher Benutzer bei der Suche nach Information, Bedienung von Computersystemen, und Kommunikation.
- Computerlinguistik als Kognitionswissenschaft. Ziel: Erforschung der menschlichen Sprachfähigkeit. Besseres Verständnis des Sprachlernens, der Sprachverarbeitung, und des Verhältnisses zwischen Sprache und sonstigen kognitiven Fähigkeiten.

4. (10 Punkte)

- Antwort 1 ist relevant, aber nicht korrekt. Das System liefert die Antwort auf eine verwandte Frage (nach dem bevölkerungsreichsten Land der EU), die zufällig mit der Antwort der gestellten Frage übereinstimmt.
- Antwort 2 ist weder relevant noch korrekt. Das System liefert einen Satz, der die richtigen Stichwörter erhält, allerdings in einer falschen Konfiguration.
- Antwort 3 ist relevant und richtig. Hier handelt es sich um die gestellte Frage, und die richtige Antwort

5. (10 Punkte)

Zum Beispiel:

- Natürliche Prosodie. Läßt sich nicht ohne weiteres aus dem Schriftbild rekonstruieren, hängt von der Informationsstruktur und der syntaktischen Struktur ab.
- Korrekte Aussprache von Abkürzungen/Akronymen. USA wird als drei Buchstaben ausgesprochen, UNO als ein Wort.
- Koartikulation. Die Aussprache von Buchstaben wird beeinflusst von ihrem Kontext (z.B. Ende des Wortes → Auslautverhärtung). Daher kann man nicht einfach alle Laute einmal aufnehmen und einfach hintereinanderkleben.

6. (10 Punkte)

- ELIZA:
 - i. Wie geht es Ihnen?
 - ii. Mir geht es schlecht.
 - iii. Ihnen geht es also schlecht?
 - iv. Ja. Ich habe Ärger mit meiner Familie.
 - v. Was haben Sie für Ärger mit Ihrer Familie?

ELIZA verwendet ein einfaches pattern matching

(Mustererkennung), um herauszufinden, worüber der Benutzer

spricht, und stellt ihm dann eine weitere Frage zu diesem Thema, die durch einfaches Einfügen dieses Themas in einen Fragesatz erstellt wird. Tiefere sprachliche Analyse (auf Bedeutung) findet nicht statt. ELIZA hat keinen internen Zustand (keine „Welt“), und kann keine Antworten auf Fragen geben.

- SHRDLU:
 - i. Lege die rote Pyramide auf den schwarzen Block.
 - ii. OK.
 - iii. Was liegt auf dem schwarzen Block?
 - iv. Die rote Pyramide.
 - v. Liegt noch etwas auf ihm?
 - vi. Ich nehme an, Sie meinen mit „ihm“ den schwarzen Block.
Nein.
 - vii. Nimm den schwarzen Block.
 - viii. Geht nicht. Die Pyramide steht darauf.SHRDLU führt eine tiefe sprachliche Analyse der Benutzereingaben durch (inklusive Bedeutung) und erlaubt dem Benutzer, eine kleine Welt („Blocksworld“) abzufragen und zu manipulieren. SHRDLU hat also einen komplexen inneren Zustand und weiß zudem über physikalische Regeln bescheid (z.B. dass man kein Objekt anheben kann, auf dem ein anderes Objekt steht). Auch sprachliches Wissen hat SHRDLU viel; im Beispiel löst das Programm das Pronomen „ihm“ korrekt auf.

7. (10 Punkte)

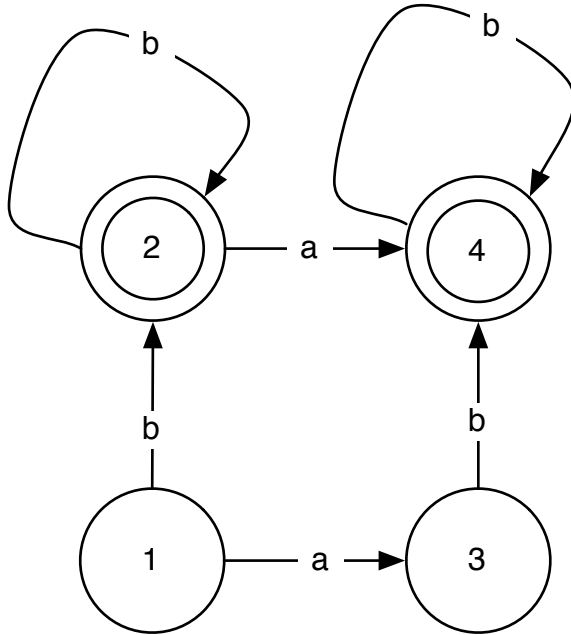
Zum Beispiel:

- Unvollständigkeit entsteht aufgrund von Synonymie: Auf die Anfrage „Kauf Auto“ hin findet die naive Methode nur Dokumente, die „Kauf Auto enthalten“, aber nicht „Auto kaufen“ oder „Fahrzeug kaufen“ oder „Fahrzeug erwerben“.
- Inkorrekte Ergebnisse entstehen aufgrund von Mehrdeutigkeit. Auf die Anfrage „Bank Stuttgart“ hin werden unter Umständen auch Dokumente über Parkbänke in Stuttgart zurückgegeben.

8. (15 Punkte)

Es genügt, sich auf diejenigen Zustände des DEA zu beschränken, die vom Startzustand aus zugänglich sind. Das sind typischerweise sehr viel weniger als die volle Anzahl der Zustände.

9. (15 Punkte)

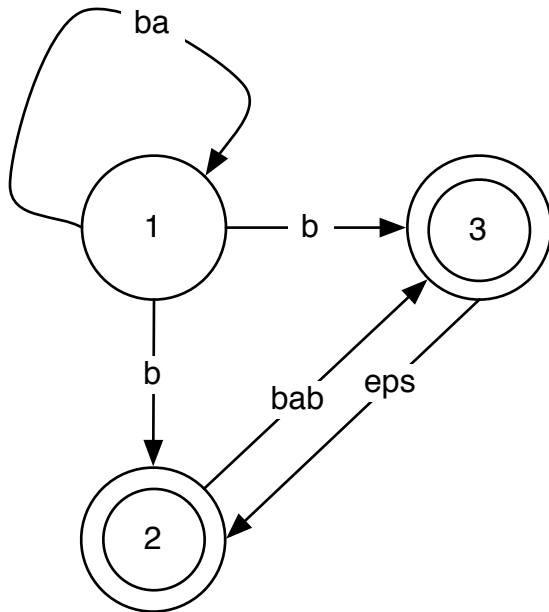


10. (10 Punkte)

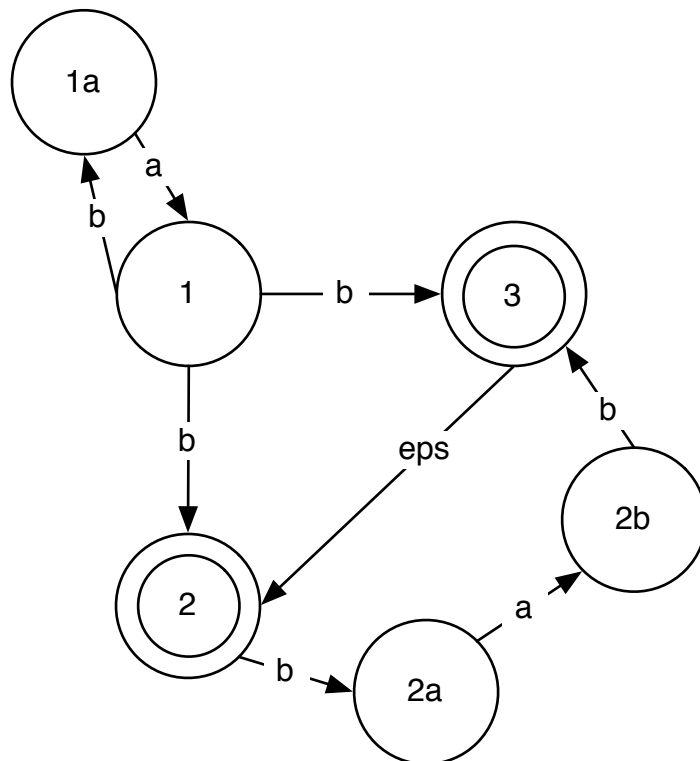
1. Ziehen einer Stichprobe aus einer Menge an Textdateien bzw. Digitalisierung oder sonstige Akquisition von Daten
2. Vorverarbeitung: Tokenisierung, Entfernung von Fehlern, Säuberung, evtl. automatische Lemmatisierung / POS Tagging
3. Eigentliche Annotation: Manuelle Annotation von syntaktischer Annotation. Ggf. (falls nicht im 2. Schritt erledigt) manuelle Lemmatisierung und Wortartenbestimmung.

11. (15 Punkte)

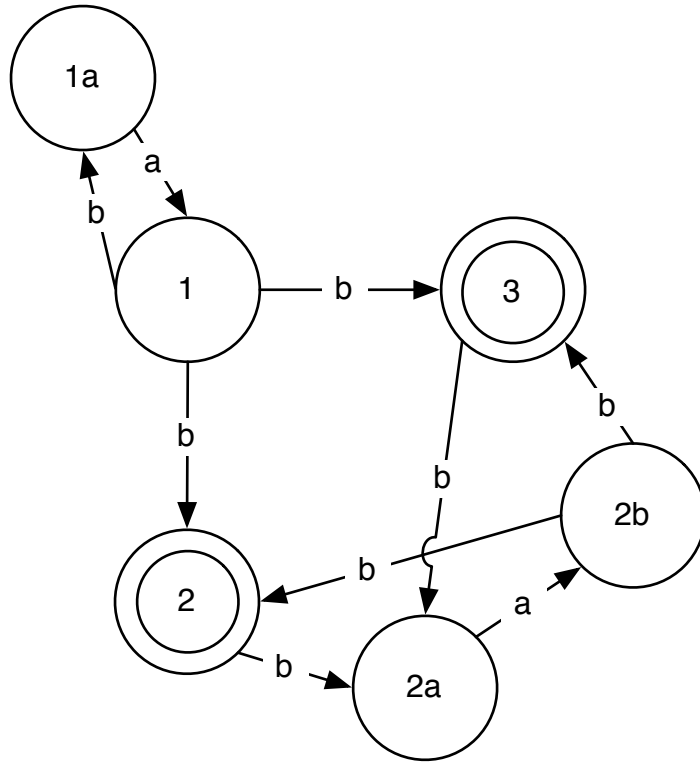
Per Konvention sei bei allen Automaten Zustand 1 Anfangszustand.
Originalautomat:



1. Elimination der Mehrsymbolkanten



2. Elimination der Epsilon-Kante (buchstabierender Automat)



3. Potenzautomat

