

Lösungsvorschlag für das Übungsblatt 5.

Aufgabe 1.

N – Nomen: Lehrer

ADJ – Adjektiv: groß

V – Verb: lernen

DET – Determinator: der

PRON – Pronomina: wir

ADV – Adverb: bald

PART – Partikel: doch

PRAEP – Präposition: auf

KONJ – Konjunktion: und

INTJ – Interjektion: ah

VPART – inne (halten)

XINC – nicht klassifizierbar: toi, toi, toi.

Die regulären Nomen mit nominaler Deklination:

	<i>Singular</i>	<i>Plural</i>
<i>Nom.</i>	der Lehrer	die Lehrer
<i>Gen.</i>	des Lehrers	der Lehrer
<i>Dat.</i>	dem Lehrer	den Lehrern
<i>Akk.</i>	den Lehrer	die Lehrer

Die Nomen mit adjektivischer Deklination:

<i>Schwache Deklination</i>			<i>Gemischte Deklination</i>	
	<i>Singular mask.</i>		<i>fem.</i>	
<i>Nom.</i>	der Verwandte	die Verwandte	ein Verwandter	eine Verwandte
<i>Gen.</i>	des Verwandten	der Verwandten	eines Verwandten	einer Verwandten
<i>Dat.</i>	dem Verwandten	der Verwandten	einem Verwandten	einer Verwandten
<i>Akk.</i>	den Verwandten	die Verwandte	den Verwandten	eine Verwandte
<i>Plural</i>				
<i>Nom.</i>	die Verwandten		Verwandte	
<i>Gen.</i>	der Verwandten		Verwandter	
<i>Dat.</i>	den Verwandten		Verwandten	
<i>Akk.</i>	die Verwandten		Verwandte	

Quantorenähnliche Nomen (auch *bisschen, paar, wenig*):

	<i>Singular mask.</i>		<i>Plural</i>
<i>Nom.</i>	das Kilo	ein Kilo	zwei Kilo
<i>Gen.</i>	des Kilos	eines Kilos	zwei Kilo
<i>Dat.</i>	dem Kilo	einem Kilo	zwei Kilo
<i>Akk.</i>	das Kilo	ein Kilo	zwei Kilo

VST – starke Verben: tragen

VSTT – starke Verben mit abtrennbarem Präfix: abgeben

VSW – schwache Verben: lernen

VSWT – schwache Verben mit abtrennbarem Präfix: aufzählen

VUNR – unregelmäßige Verben: werden

VUNRT – regelmäßige Verben mit abtrennbarem Präfix: anhaben.

Aufgabe 2.

Determinativkomposita (Nichtkopf bestimmt den Kopf näher; in welche Richtung die Bedeutung spezialisiert wird, ist weitgehend offen): Rinderschnitzel, Fassbier, Wandschrank, Zigarettenraucher, Bürogebäude.

Possessivkomposita (Das Kompositum bezeichnet (oft pars pro toto) prominente Eigenschaft des Gemeinten): Milchgesicht, Langbein, Spitzbauch, Trotzkopf, Langfinger.

Kopulativkomposita (Glieder sind gleichrangig; koordinierte Beziehung zwischen Denotaten beziehen sich auf gleiche Entität, gleiche Wortklasse): süßsauer, nasskalt, Dichterkomponist, spritzgießen, graurot.

Aufgabe 3.

(Als Fugenform bezeichnet man Wortformen, die als Vorderglieder in Komposita auftreten.)

Der Algorithmus gibt alle möglichen Zerlegungen eines Strings aus.

Sobald der erste Teilstring (T1) als FF erkannt wird (falls im Lexikon vorkommt), wird der restliche Teilstring (T2) auf die Lexikalität geprüft. Falls der sich als gültige Kopfkonstituente eignet, ist die Zerlegung fertig. Die Teilstrings (T1, T2) werden durch "_" in die Datenstruktur geschrieben und es wird nach weiterem Erstglied gesucht. Falls es nicht der Fall ist, wird der Reststring (T2) rekursiv weiter zerlegt, bis es sich wieder eine gültige Konstituente findet. Dabei wird zu jedem Element aus möglichen Zerlegungen von T2 "T1_" am Anfang angehängt. Diese werden dann zur Datenstruktur hinzugefügt.

Für die gegebenen Beispielen ergibt sich dann:

Verbraucherschutz: Verb_raucherschutz,
 Verbraucher_schutz.

Erzherzog: Erz_herzog

Wie es sich zeigt, werden auch nicht sinnvolle Zerlegungen ausgegeben, wie *Verb_raucherschutz*, was gar nichts mit "Verbraucherschutz" zu tun hat.

Aufgabe 4.

Ein Text (eine Zeichenkette) wird in Token (Einheit der Textsegmentierung) zerlegt. Ein Token ist durch Leerraumzeichen (engl. *white space*) oder Interpunktion begrenzt:

- Wort-Zeichen: /[a-zA-Z0-9]/
- Einzel-Zeichen: /[.,:;!)"' ?-]/
- Grenz-Zeichen: /\s/.

Es gibt aber Fälle, wo ein Lexem aus mehreren Wörtern besteht (Mehrwortlexeme) bzw. Tokens können sowohl Einzel- wie Grenz-Zeichen enthalten:

1. Zusammengesetzte Wörter, deren Elemente durch space getrennt sind (sogenannte phrasal verbs):

z.B. take off

2. Zusammengesetzte Wörter, zwischen Elementen deren andere Einheiten stehen.

3. Stemming (Grundformen bei Verben).

4. Durch Bindestrich/space getrennte oder zusammengeschriebene Wörter (Hyphenation): *rookie-of-the-year, sciencefiction, roleexperimentation*

5. Eigennamen.

Um diese Probleme zu lösen, werden verschiedene Techniken angesetzt, wie Erstellung von Lexika, statistische Analyse (z.B. automatische Phrasenindexierung, Algorithmen für Kollokationenextraktion).

Solche Schwierigkeiten treten auch im Deutschen auf:

1. Verben mit abtrennbarem Präfix (Er steht um 7.00 Uhr auf).

2. Phraseologismen (ins Gras beißen).

3. Stemming (sein – war – gewesen).

4. Hyphenation (falschgeschriebene Wörter).

5. Eigennamen (New York).