Elementare statistische Methoden

Vorlesung “Computerlinguistische Techniken”

Alexander Koller

28. November 2014

CL-Techniken: Ziele

- Ziel 1: Wie kann man die Struktur sprachlicher Ausdrücke berechnen?

- Ziel 2: Wie geht das effizient, auch wenn der sprachliche Ausdruck mehrdeutig ist?

- Ziel 3: Wie erkennt man unter allen möglichen Lesarten die richtige?

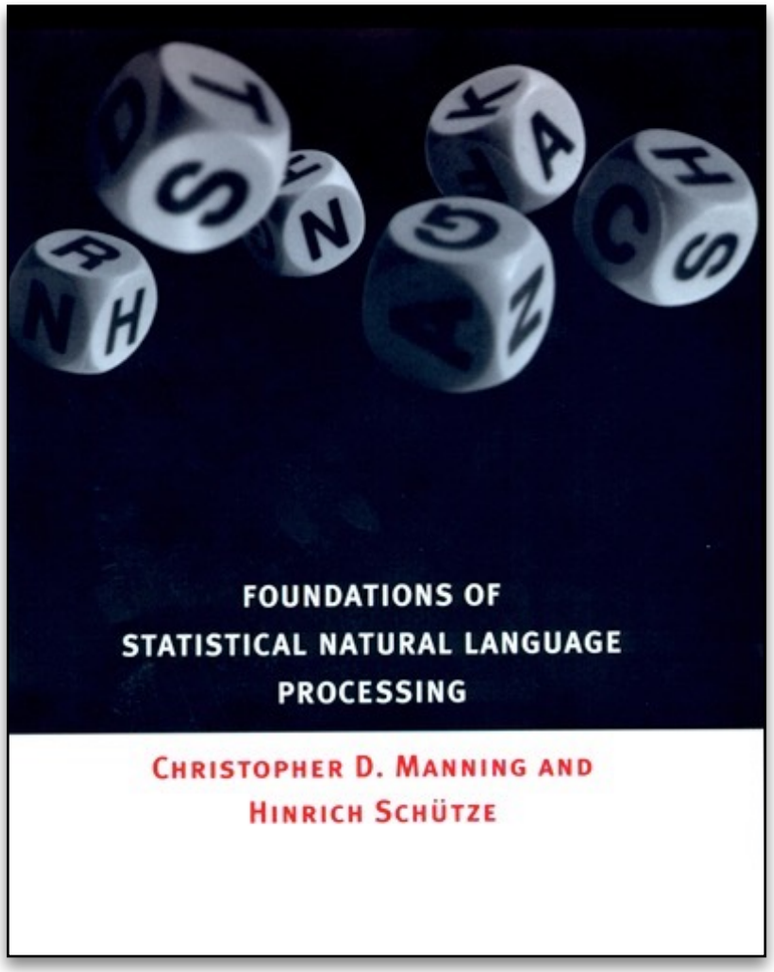
Übersicht Teil 2

- Elementare Wahrscheinlichkeitstheorie
- n-Gramm-Modelle
- Hidden Markov Models
- Probabilistische kontextfreie Grammatiken
- Einfache statistische Modelle von Semantik

Lehrbuch

- Jurafsky & Martin behandelt auch statistische Modelle.

- Etwas mehr in die Tiefe geht
  Manning & Schütze:



Heute

- Sprache als Zufallsprozess
- Elementare Wahrscheinlichkeitslehre
- Statistische Sprachmodellierung

Let’s play a game

- Wir schreiben zusammen einen Satz.
- Jeder von Ihnen diktiert mir ein Wort.
- Sie können dabei alles anschauen, was bisher an der Tafel steht.

Wörter raten

Wörter raten

- Kupfer ist ein

Wörter raten

- Kupfer ist ein
- Kupfer ist ein chemisches

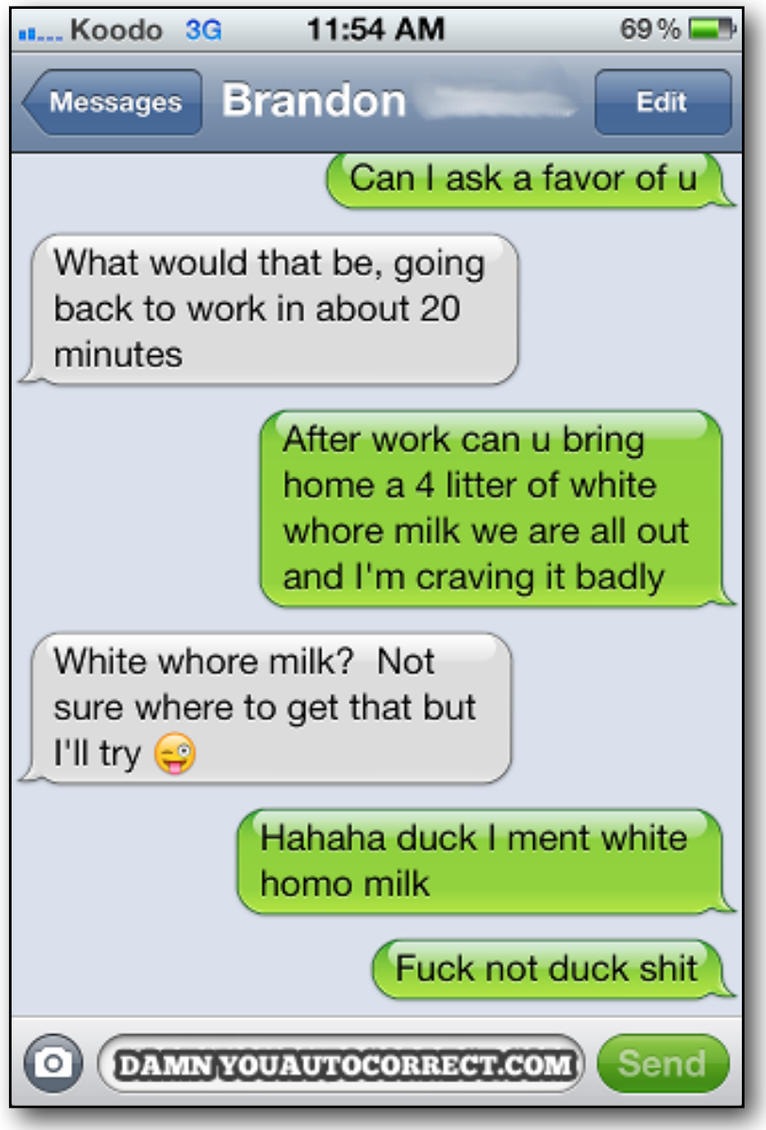
Wörter raten

- Kupfer ist ein
- Kupfer ist ein chemisches
- Hans ruft Maria mit dem Handy

Wörter raten

- Kupfer ist ein
- Kupfer ist ein chemisches
- Hans ruft Maria mit dem Handy
- Nächste Woche wird Hans

Eine Anwendung



Grundidee

- Auftreten von Wörtern, Phrasen, Sätzen in Sprache folgt bestimmten Wahrscheinlichkeiten.

- Viele sprachliche Phänomene sind inhärent “weich” und mit W. gut zu modellieren.
  - z.B. gradierte Grammatikalität; Selektionspräferenzen

- Umgang mit Ambiguitäten.
  - Modell: richtige Lesart = wahrscheinliche Lesart.

Wahrscheinlichkeiten

- Grundlage der W.theorie sind *Ereignisse*, die mit einer bestimmten W. eintreten können.

- Beispiele:
  - Würfel: {1 gewürfelt, ..., 6 gewürfelt}
  - Wetter morgen: {Regen, bewölkt, Sonnenschein}
  - Sprachen: {nächstes Wort ist “Element”, nächstes Wort ist “Metall”, ...}

- Die Menge der Ereignisse heißt *Ereignisraum;* wir nehmen i.a. an, dass er endlich ist.

Zufallsvariablen

- Ereignisse werden dadurch beschrieben, dass eine *Zufallsvariable* einen bestimmten Wert annimmt.
  - Zufallsvariable: Variable, deren Wert vom Zufall abhängt
  - Wert wird bei der Auswertung der Variablen zufällig bestimmt

- Beispiele:
  - Würfel: Werte von X sind 1, ..., 6
  - Wetter: Werte von X sind heiter, bewölkt, ...
  - Sprache: Werte von X sind “Element”, “Metall”, ...

Wahrscheinlichkeiten

- Zufallsvariable X nimmt ihre Werte mit bestimmten *Wahrscheinlichkeiten* an.

- Wir schreiben
  - P(X = a) für die W., dass X den Wert a annimmt
  - P(X) für die *W.verteilung* von X, also die Funktion
    $a \mapsto P(X = a)$

- Wahrscheinlichkeiten sind Zahlen zwischen 0 und 1. Es muss immer gelten:

$$\sum_{a \in A} P(X = a) = 1$$

Wahrscheinlichkeiten

- *Frequentistische* Interpretation: Wenn wir N-mal den zufälligen Wert von X bestimmen, nimmt Variable X etwa $N \cdot P(X = a)$ mal den Wert a an.

- Beispiele:
  - Würfel: $P(X = 1) = ... = P(X = 6) = 1/6$
  - Wetter: $P(X = \text{Regen}) \approx$ Anteil von früheren Tagen mit vergleichbarem Wetter, an denen es danach geregnet hat
  - Sprache: das wollen wir hier herausfinden

Rechenregeln

- Die Wahrscheinlichkeit, dass X *nicht* den Wert a annimmt, ist 1 - P(X = a).

- Wahrscheinlichkeit des *sicheren* Ereignisses (X nimmt irgendeinen Wert an) ist 1, des *unmöglichen* Ereignisses also 0.

- Wenn Ereignisse X = a und Y = b sich gegenseitig ausschließen, gilt

  P(X = a oder Y = b) = P(X = a) + P(Y = b).

Komplexe Ereignisse

- Wir interessieren uns oft dafür, wie w. es ist, dass mehrere Variablen gleichzeitig bestimmte Werte annehmen.

- *Gemeinsame* Wahrscheinlichkeit:
  $P(X_1 = a_1, X_2 = a_2)$

- Beispiele:
  - X_1 = nächster, X_2 = übernächster Würfelwurf
  - X_1 = Regen, X_2 = Temperatur

Marginalisierung

- Aus gemeinsamer W.verteilung kann man ZV durch *Marginalisierung* entfernen:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

- D.h. über alle möglichen Werte von Y summieren.

- Beispiel:
  P(X = Regen) =
  P(X = Regen, Y = warm) + P(X = Regen, Y = kalt)

Gemeinsame W.

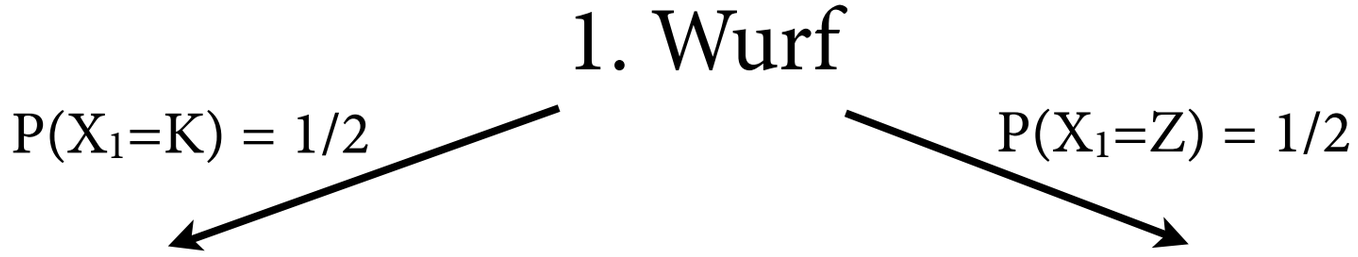
- In vielen Fällen gilt:
  $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$

- Beispiel Münzwurf:

Gemeinsame W.

- In vielen Fällen gilt:
  $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$

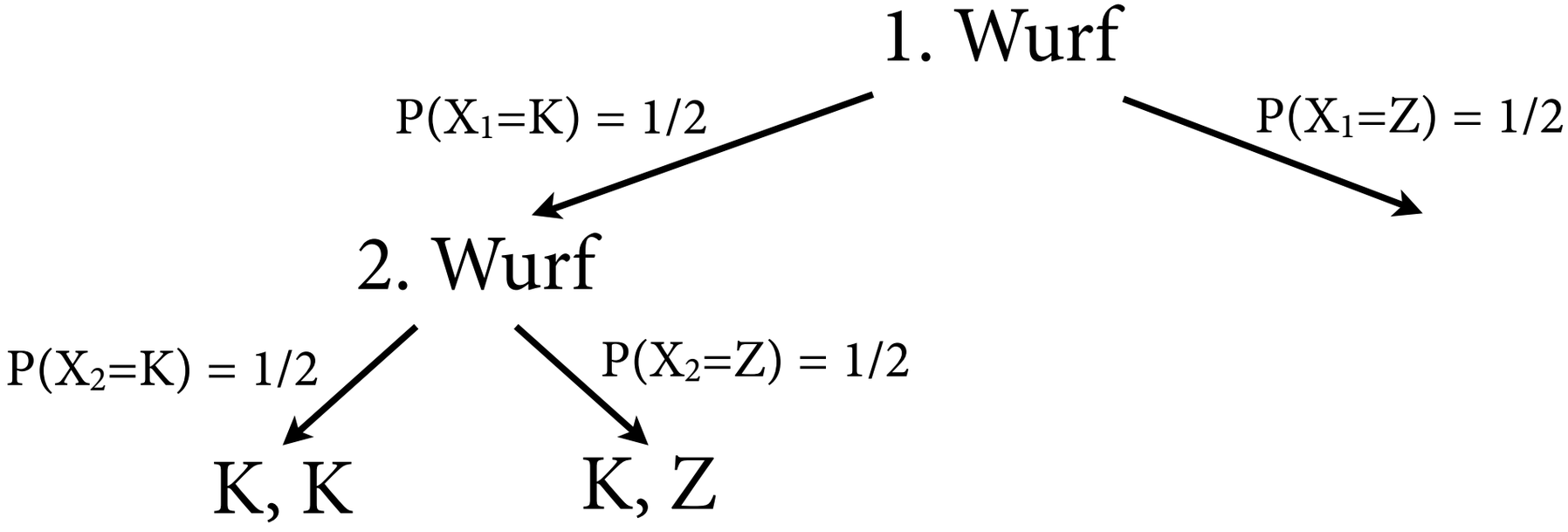
- Beispiel Münzwurf:



Gemeinsame W.

- In vielen Fällen gilt:
  $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$

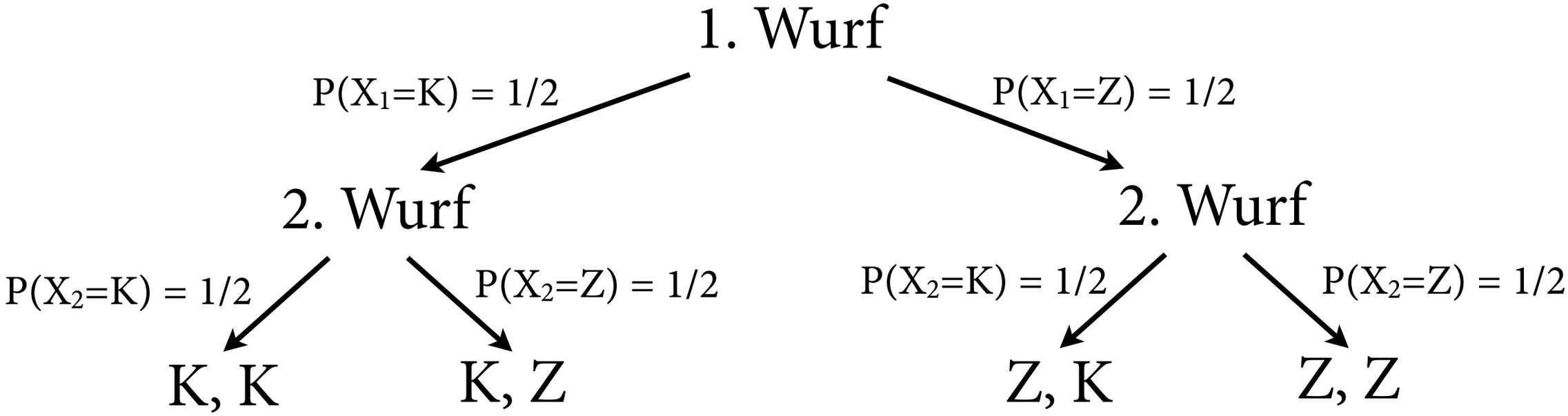
- Beispiel Münzwurf:



Gemeinsame W.

- In vielen Fällen gilt:
  $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$

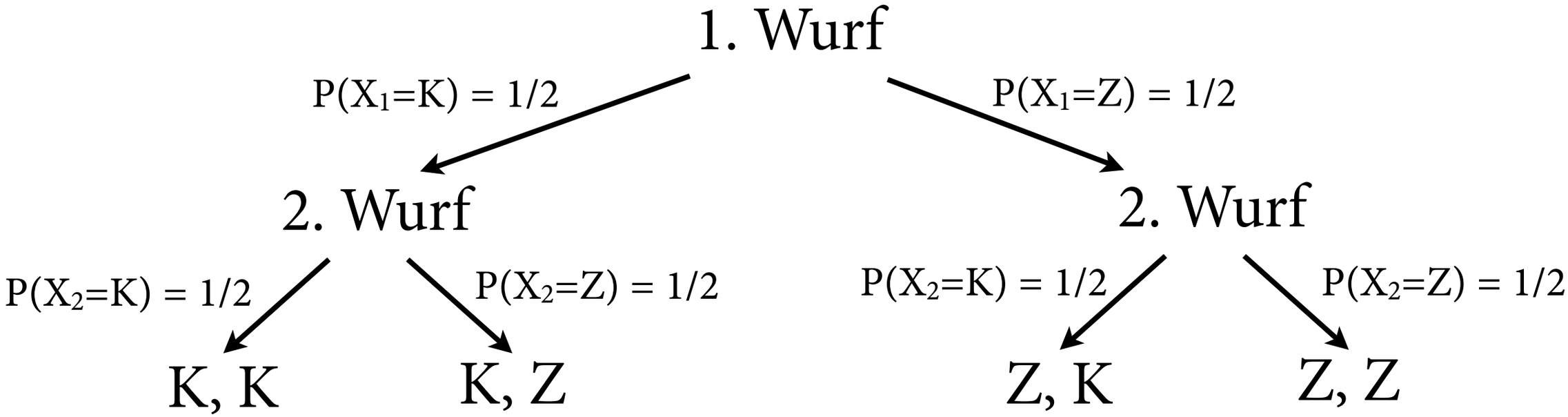
- Beispiel Münzwurf:



Gemeinsame W.

- In vielen Fällen gilt:
  $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$

- Beispiel Münzwurf:

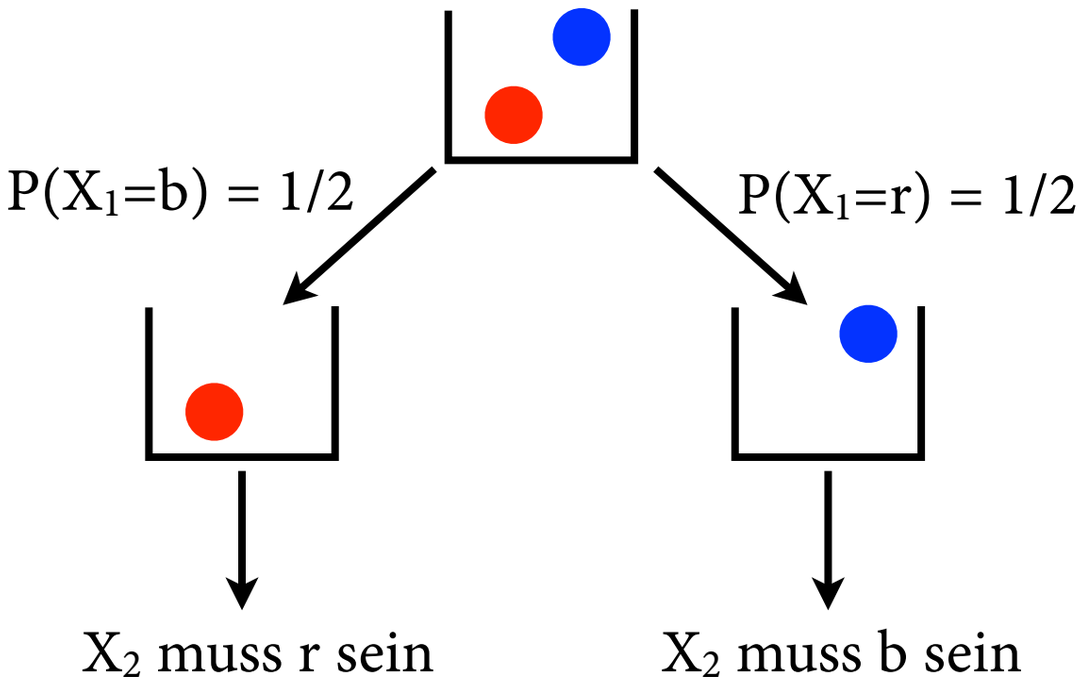


$$P(X_1=K, X_2=K) = 1/4 = P(X_1 = K) \cdot P(X_2 = K)$$

Gemeinsame W.

- $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$ gilt aber nicht für alle Zufallsvariablen X_1, X_2!

- Beispiel: Ziehen ohne Zurücklegen.



P(X	1/2
P(X	1/2

P(X	1/2
P(X	1/2

P(X	0
P(X	1/2
P(X	1/2
P(X	0

Statistische Unabhängigkeit

- Zwei Variablen X_1, X_2 heißen *statistisch unabhängig* gdw für alle a_1, a_2 gilt:
  $P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$

- "Abhängig" heißt: Ausgang von X_1 beeinflusst Ausgang von X_2 irgendwie.

- Beispiele:
  - Münzwurf: X_1, X_2 sind statistisch unabhängig
  - Ziehen ohne Zurücklegen: X_1, X_2 nicht unabhängig

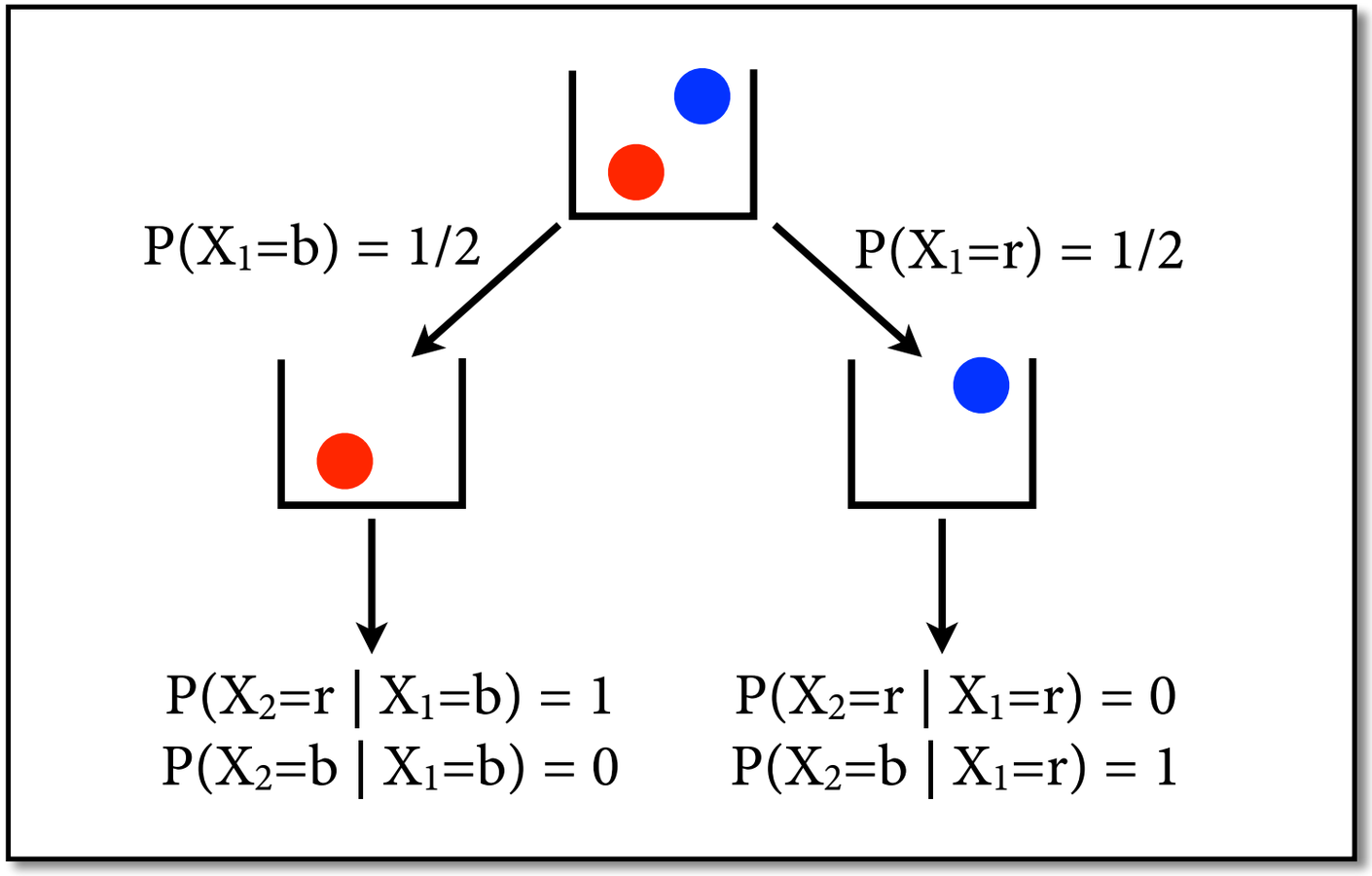
Bedingte Wahrscheinlichkeiten

- Einfluss von einer Variable X auf eine andere Y kann man mit *bedingten W.* darstellen.

- $P(Y = a \mid X = b)$: W., dass Y den Wert a annimmt, wenn X den Wert b annimmt.

Bedingte Wahrscheinlichkeiten

- Beispiel “Ziehen ohne Zurücklegen”:



- Allgemein gilt auch für bedingte W. für jedes b:

$$\sum_{a \in A} P(X_2 = a | X_1 = b) = 1$$

Bedingte Wahrscheinlichkeiten

- Es gilt für alle ZVen X, Y die Rechenregel:
  $P(X = a, Y = b) = P(X = a \mid Y = b) \cdot P(Y = b)$
  $P(X = a, Y = b) = P(Y = b \mid X = a) \cdot P(X = a)$

- Statistische Unabhängigkeit sieht so aus:
  $P(X = a \mid Y = b) = P(X = a \mid Y = b')$ f.a. b, b'

- Bedingung kann mehrere Ereignisse enthalten:
  $P(X = a \mid Y = b, Z = c)$

Die Bayes'sche Regel

- Wichtige Konsequenz ist *Bayes'sche Regel:*

$$P(X = a \mid Y = b) = \frac{P(Y = b \mid X = a) \cdot P(X = a)}{P(Y = b)}$$

- Bayes erlaubt uns, Bedingung umzudrehen, z.B.:
  - Jemand erzählt uns, er hat mit einer Person mit langen Haaren (L) gesprochen; wie w., dass es eine Frau war (F)?
  - Sagen wir P(L|F) = 0.75, P(L|M) = 0.15; natürlich P(F) = 0.5.
  - Mit Bayes ausrechnen: P(F|L) = 0.83.

Abgekürzte Schreibweise

- Schreibweise “$P(X = a)$” ist oft unhandlich.
- Wenn die ZV klar ist, lassen wir sie weg:
  - $P(a)$ statt $P(X = a)$
  - $P(a \mid b, c)$ statt $P(X = a \mid Y = b, Z = c)$
  - $P(w_1, w_2)$ statt $P(X_1 = w_1, X_2 = w_2)$
- Es ist wichtig, sich klarzumachen, was für ZV in dieser Schreibweise gemeint sind.

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

X_1

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

X_1

↓

Kupfer

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

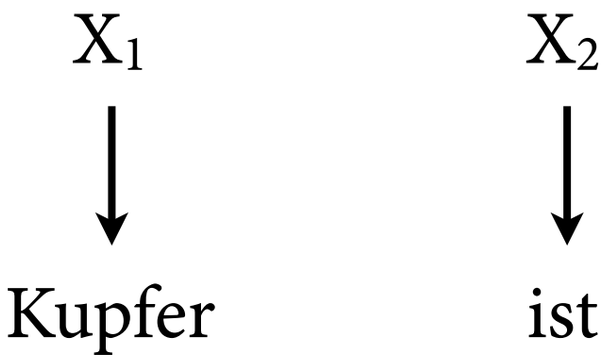
X_1 X_2

Kupfer

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

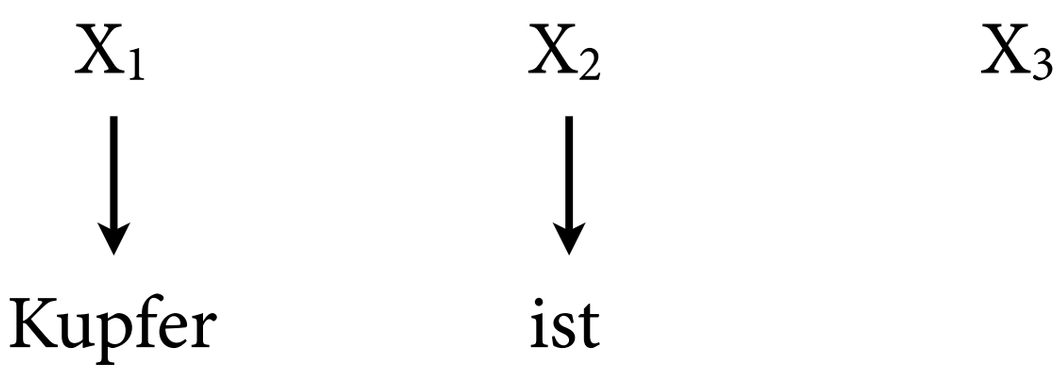
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

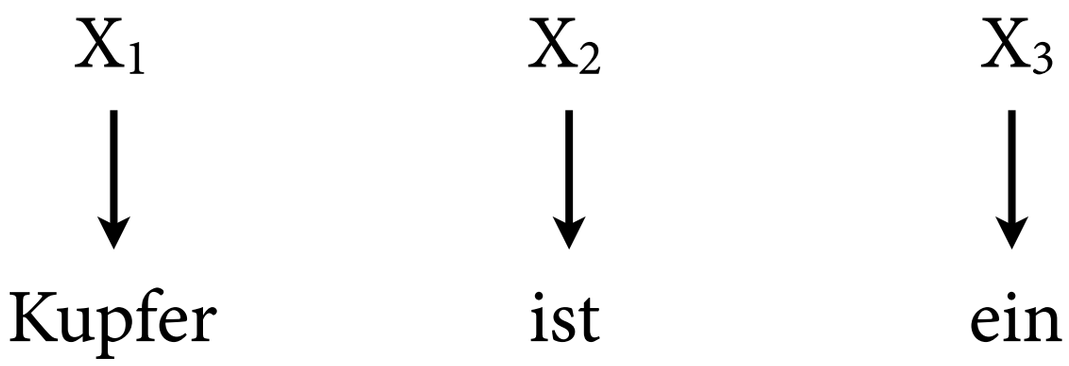
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

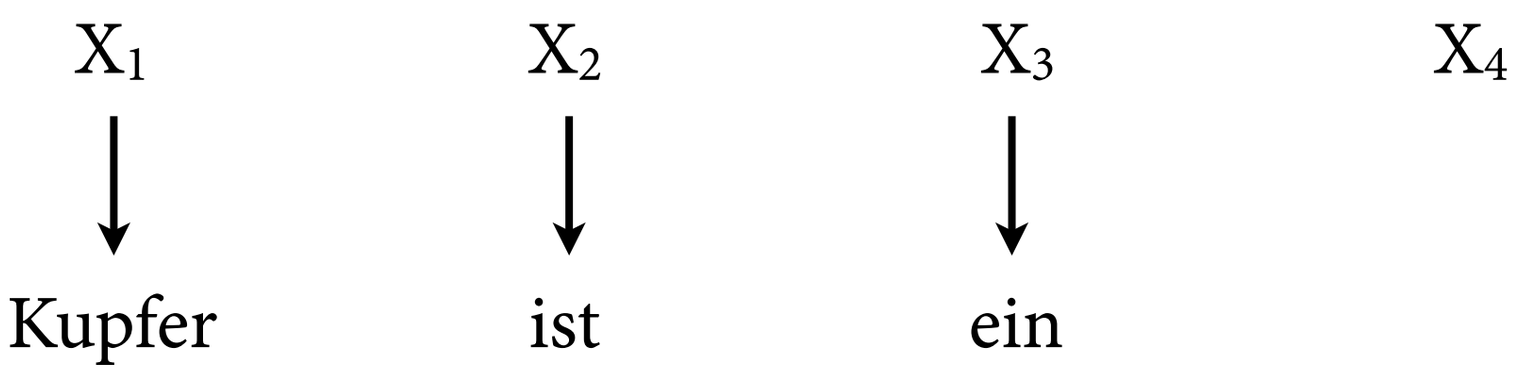
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

Sprache als Zufallsprozess

- Zufallsprozess wird durch W.verteilungen für die einzelnen Positionen definiert:
  - erstes Wort: $P(X_1 = w_1)$
  - zweites Wort: $P(X_2 = w_2 \mid X_1 = w_1)$
  - t-tes Wort: $P(X_t = w_t \mid X_1 = w_1, ..., X_{t-1} = w_{t-1})$

- W. des ganzen Satzes:
  $P(w_1 ... w_n) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_1, w_2)$
  $\cdot ... \cdot P(w_n \mid w_1, ..., w_{n-1})$

W. schätzen

- Zentrales Problem der statistischen CL:
  Was sind die $P(w_t \mid w_1, ..., w_{t-1})$?

- W.verteilung aus *Korpora* schätzen.

- Idee: $P(X_t = w)$ durch Häufigkeit von w in einem Korpus approximieren.
  - *Absolute Häufigkeit:* Anzahl C(w) der Auftreten (Tokens) des Wortes w im Korpus
  - Wenn Korpus N Tokens enthält, hat w *relative Häufigkeit* f(w) = C(w) / N.

Einige wichtige Korpora

Korpus	Tokens	Types
Brown (NLTK)	1.1 Mio	56057
Switchboard (Englisch, gesprochen)	2.4 Mio	ca. 20.000
Penn Treebank (syntaktisch annotiert)	ca. 5 Mio	
Gigaword Corpus	1.7 Mrd	
DWDS-Korpus (deutsch)	100 Mio	
Tiger-Korpus (deutsch, syn. annotiert)	900.000	64.485

Das Zipfsche Gesetz

- Beobachtung: Die meisten Wörter kommen im Korpus selten vor.
  - in Tiger durchschnittlich 13 Tokens pro Type
  - aber: ca. 55% aller Wörter kommen nur einmal vor, ca. 70% höchstens zweimal

- Zipfsches Gesetz:
  - sortiere Wörter nach ihren absoluten Häufigkeiten
  - trage für jedes Wort absolute Häufigkeit in Graph ein
  - wenn beide Achsen logarithmisch sind, bekommt man eine Gerade

Zusammenfassung

- Sprache als Zufallsprozess
- Elementare W.theorie
- Korpora