ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung I / 6 HS 2016

Vektorquantisierung – Sprachsynthese

Buch: Kapitel 4.7 und 7

Beat Pfister

Sprachverarbeitung I / 6

Vorlesung: **Vektorquantisierung**

Einführung in die Sprachsynthese

- Ziel der Sprachsynthese
- Zusammenhang Text und Lautsprache
- Konzeption eines Sprachsynthesesystems

Übung: Vektorquantisierung von Sprachsignalen

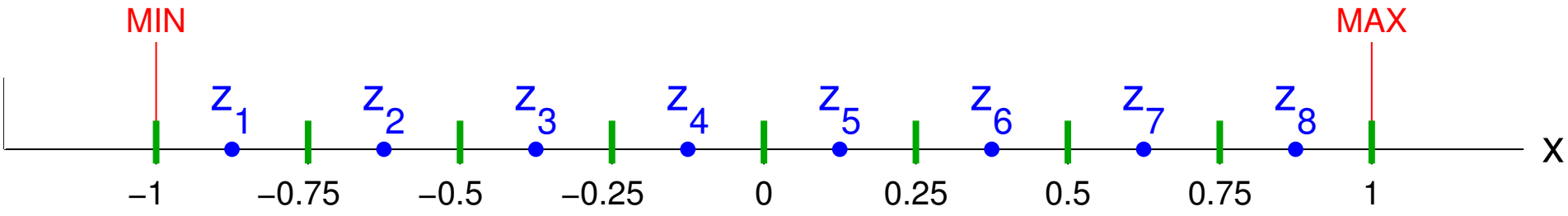
Teil 1:

Vektorquantisierung

Quantisierung

Zuordnung eines Wertes $\mathbf{x}$ aus einer grossen oder unendlichen Wertemenge zu einem der Werte $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_M\}$

Beispiel mit $M = 8$: uniforme Quantisierung des Intervalls [MIN. . . MAX]



Mass für die Güte der Quantisierung: Quantisierungsfehler $|\mathbf{x} - \mathbf{z}_i|$

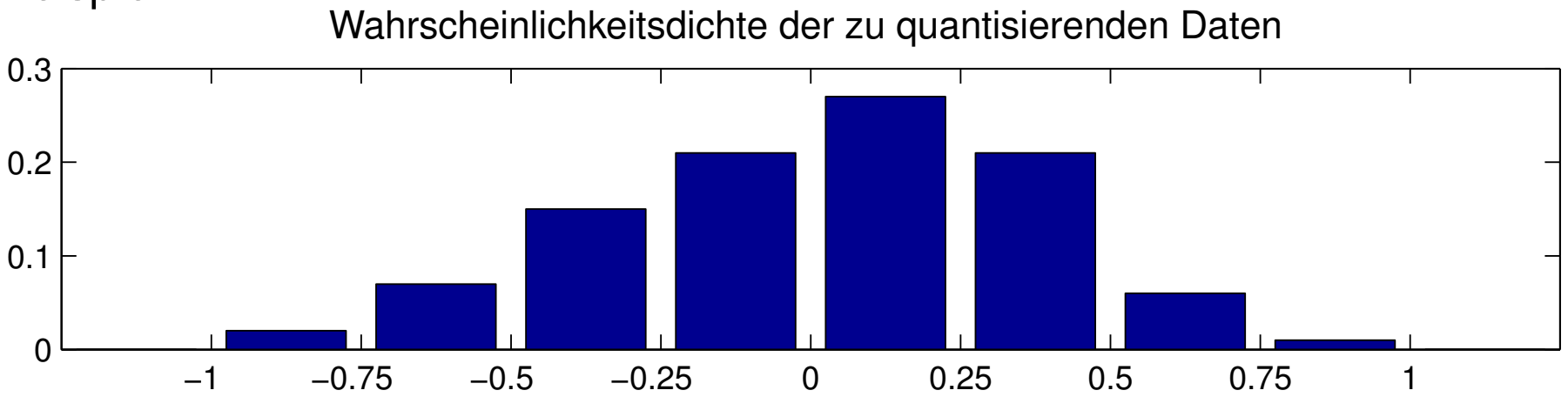
Merke: Uniforme Quantisierung ist nur für **gleichverteilte** Daten optimal !

Datenabhängige Quantisierung

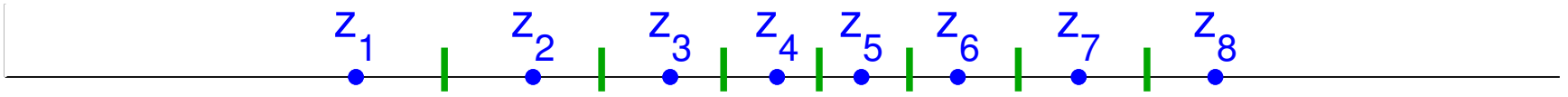
Prinzip: Mittlerer Quantisierungsfehler soll minimal sein

$\longrightarrow$ feinere Quantisierung wo Daten dichter sind

Beispiel:



Optimale Quantisierung:



Skalare Quantisierung vs. Vektorquantisierung

Skalare Quantisierung

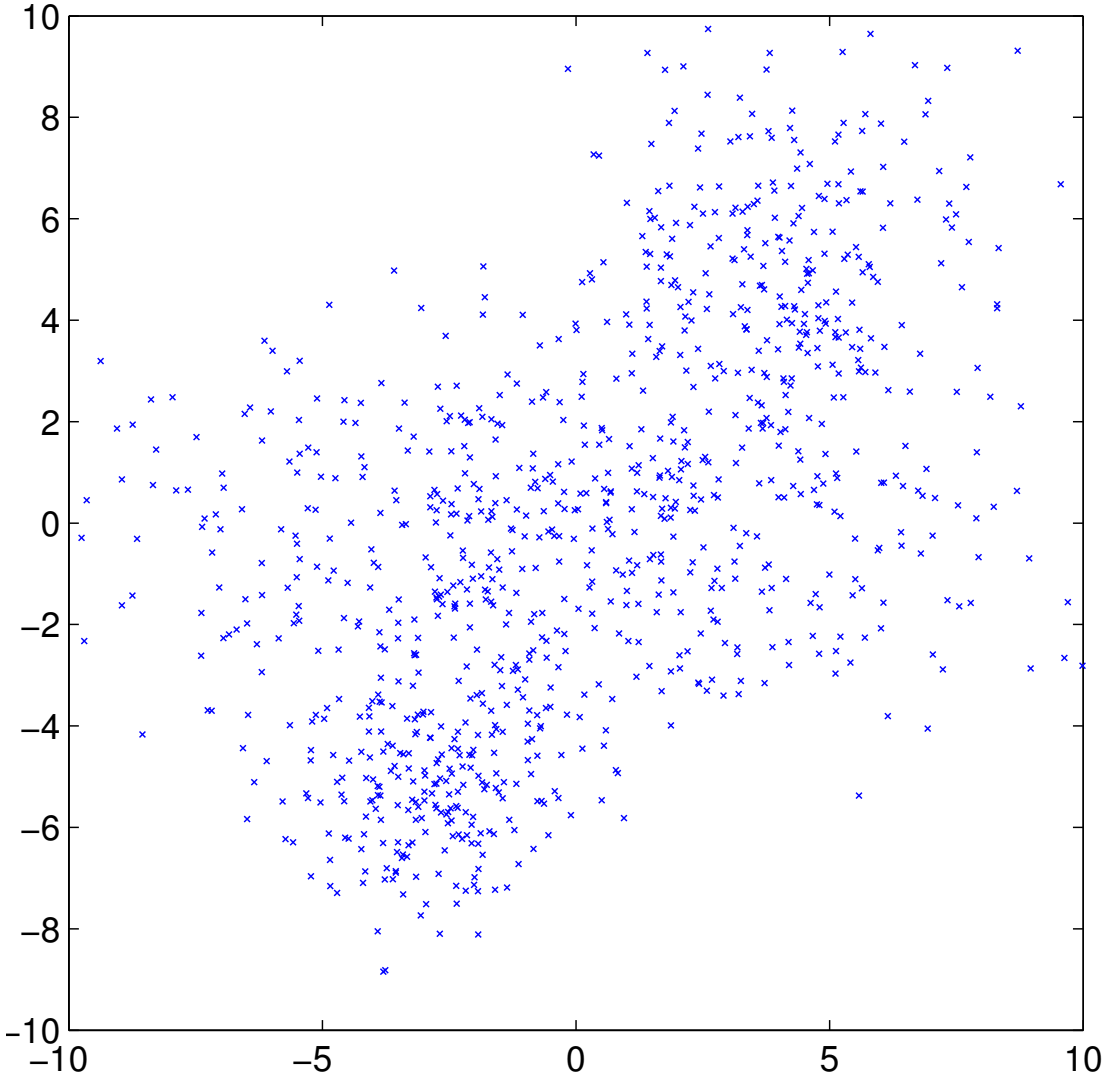
- uniforme Quantisierung: alle Quantisierungs-Intervalle gleich gross
- nicht-uniforme Quantisierung: ungleiche Quantisierungs-Intervalle
  (z.B. datenabhängige Quantisierung)

Vektorquantisierung

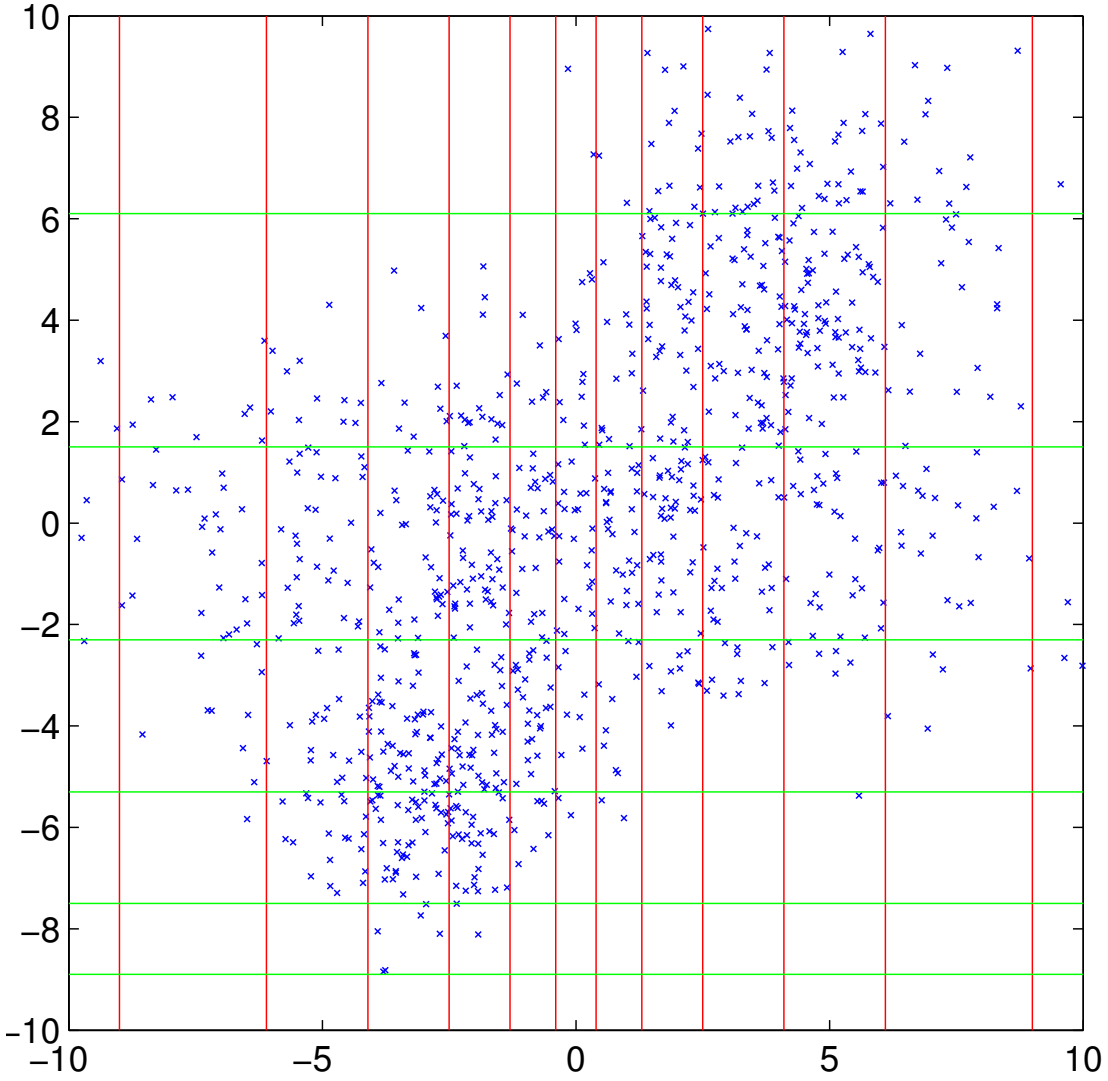
- . . . ?

Quantisierung

2-dimensionale Vektoren

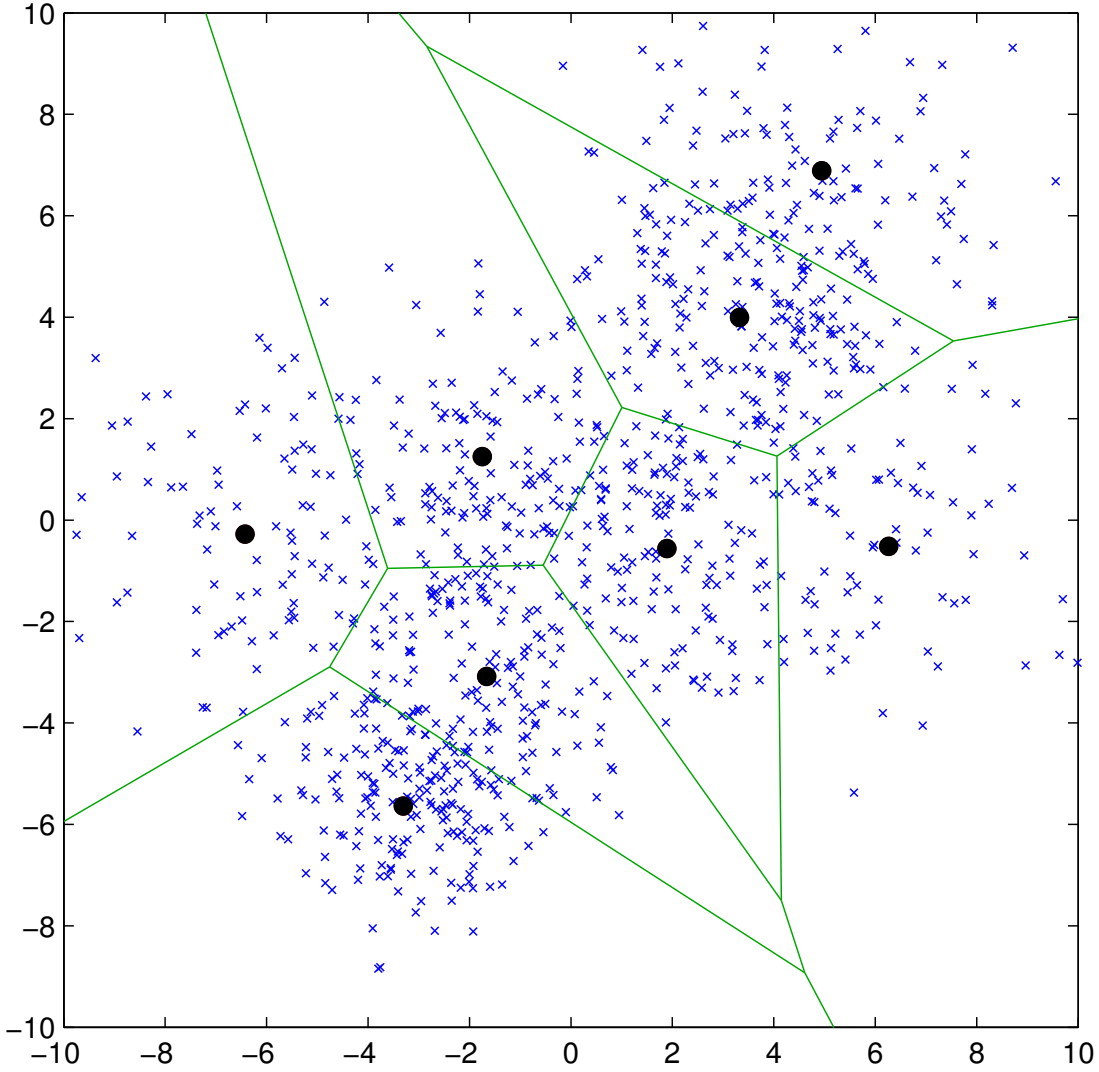


Skalare Quantisierung oder Vektorquantisierung ?



Vektorquantisierung

2-dimensionale Vektoren



Beispielsvektoren

Codebuchvektoren

Partitionsgrenzen
(Quantisierungsintervalle)

Quantisierung

Abbildung eines Wertes $\mathbf{x}$ auf einen der Werte $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_M\}$

a) Codierung: $j = \underset{i}{\operatorname{argmin}}\ d(\mathbf{x}, \mathbf{z}_i)$

b) Decodierung: $\tilde{\mathbf{x}} = \mathbf{z}_j$

Quantisierungsfehler (Distorsion): $E_q = d(\tilde{\mathbf{x}}, \mathbf{x})$

Terminologie

- Codebuch — Menge der Werte $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_M\}$
- $\mathbf{z}_j$ — j-ter Codebuchvektor, Zentroid des j-ten Quantisierungsintervalls
- $d(\mathbf{x}, \mathbf{y})$ — Distanzmass

Vektorquantisierung

Vektorquantisierung eines Wertes $\mathbf{x}$: $\tilde{\mathbf{x}} = \mathbf{z}_j$ mit $j = \underset{i}{\text{argmin}}\ d(\mathbf{x}, \mathbf{z}_i)$

Für die Vektorquantisierung werden gebraucht:

- Codebuch: $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_M\}$
- Distanzmass (positiv definit): $0 < d(\mathbf{x}, \mathbf{y}) < \infty$ für $\mathbf{x} \neq \mathbf{y}$

  $d(\mathbf{x}, \mathbf{x}) = 0$

  z.B. Euklidische Distanz: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\| \mathbf{x} - \mathbf{y} \|^2} = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2}$

Codebuch

Voraussetzungen für die Erzeugung eines Codebuchs:

- Trainingsvektoren (repräsentative Stichprobe für zu quant. Daten)
- Distanzmass (gleiches wie bei Quantisierung)
- Klassierungsprozedur (Bestimmung des nächsten Codebuchvektors)
- Verfahren zur Bestimmung des Zentroids für die Vektoren einer Partition

Verfahren zur Erzeugung eines Codebuches der Grösse M:

Keine analytische Lösung $\longrightarrow$ iterative Verfahren

Generieren eines Codebuches der Grösse M

Iteratives Aufteilen der gegebenen Trainingsvektoren in M Partitionen

K-means-Algorithmus

1. Initialisierung: Auswahl M beliebiger Trainingsvektoren
   $\longrightarrow$ Initialcodebuch >>>
2. Klassierung: für jeden Trainingsvektor den nächsten Codebuchvektor $\mathbf{z}_i$ suchen
   $\longrightarrow$ Index i (und Distorsion)
3. Aufdatieren: aus Trainingsvektoren mit Index i neues Zentroid $\mathbf{z}_i$
   $\longrightarrow$ neue Codebuchvektoren >>>
4. Iteration: Zurück zu 2 falls Abnahme der mittleren Distorsion > Schwellwert

Probleme des K-means-Algorithmus

- Algorithmus bleibt in lokalem Minimum “hängen”
  d.h. Wahl der Initialvektoren beeinflusst das Codebuch
  $\longrightarrow$ Codebuch ist möglicherweise schlecht >>>

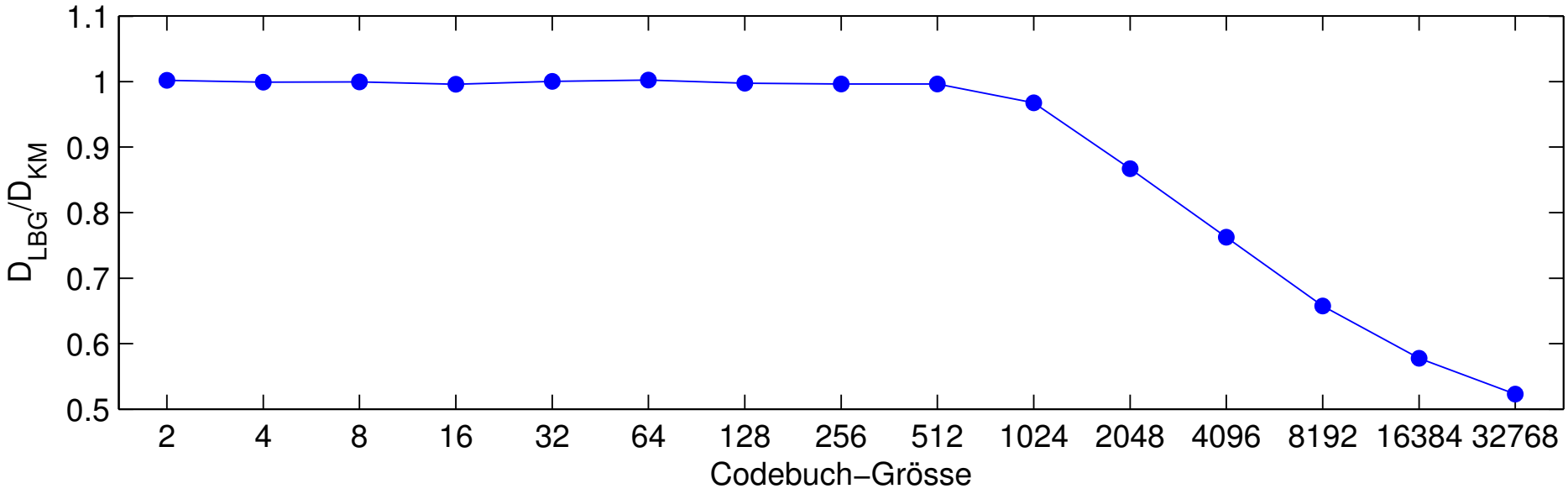
- Algorithmus konvergiert langsam

Der LBG-Algorithmus *(Linde-Buzo-Gray)*

Erzeugung eines Codebuches der Grösse $M = 2^k$ (k ganzzahlig)

1. Initialisierung: Codebuch mit $M = 1$ (Zentroid der Trainingsvektoren) >>>
2. Splitting: Zahl der Codebuchvektoren verdoppeln
   (Addition/Subtraktion eines kleinen Zufallsvektors) >>>
3. Klassierung: für jeden Trainingsvektor den nächsten Codebuchvektor $\mathbf{z}_i$ suchen
   $\longrightarrow$ Index i (und Distorsion)
4. Aufdatieren: aus Trainingsvektoren mit Index i neues Zentroid $\mathbf{z}_i$
   $\longrightarrow$ neue Codebuchvektoren
5. Iteration A: Zurück zu 3 falls Abnahme der Distorsion > Schwellwert >>>
6. Iteration B: Zurück zu 2 falls Codebuch nicht gross genug >>>

Vergleich K-means- mit LBG-Algorithmus Distorsion

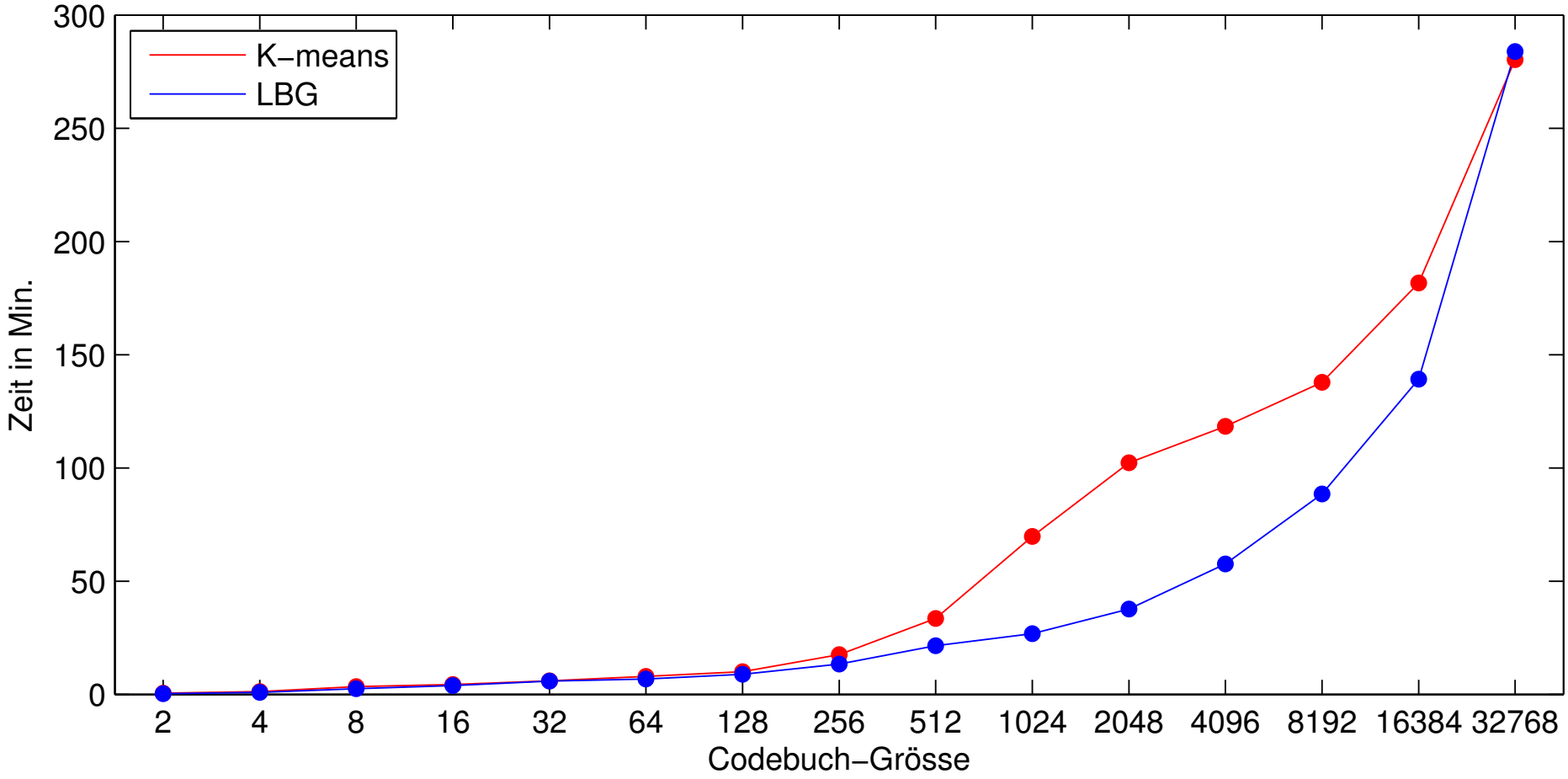


D_{KM} Distorsion für mit dem K-Means-Algorithmus ermittelte Codebücher

D_{LBG} Distorsion für mit dem LBG-Algorithmus ermittelte Codebücher

(Grösse des Trainingssets: 400000 Vektoren)

Vergleich K-means- mit LBG-Algorithmus: Trainingszeit



Vektorquantisierung in der Sprachverarbeitung

Vektoren ergeben sich aus der Merkmalsextraktion:
LPC-Parameter, AKF-Koeffizienten, cepstrale Parameter etc.

Einsatzgebiete:

- Datenreduktion: effiziente Übertragung oder Speicherung
  (z.B. Mobiltelephonie: CELP Codec ITU G.728 mit 16 kBit/s)
- Abbildung reellwertiger Vektoren auf M diskrete Symbole
  (z.B. Spracherkennung mit diskreten Hidden-Markov-Modellen)

Teil 2:

Einführung in die Sprachsynthese

Sprachsynthese

Was ist das Ziel der Sprachsynthese ?

$\longrightarrow$ Sprachsynthese soll . . . >>>

Sprachsynthese

Was ist das Ziel der Sprachsynthese ?

$\longrightarrow$ Sprachsynthese soll Text **korrekt** in Lautsprache umsetzen!

Korrekte Sprachsynthese heisst somit:

Die Sprachsynthese muss das leisten,
was eine Person macht,
wenn sie einen Text richtig vorliest !

Leitbild für die Sprachsynthese

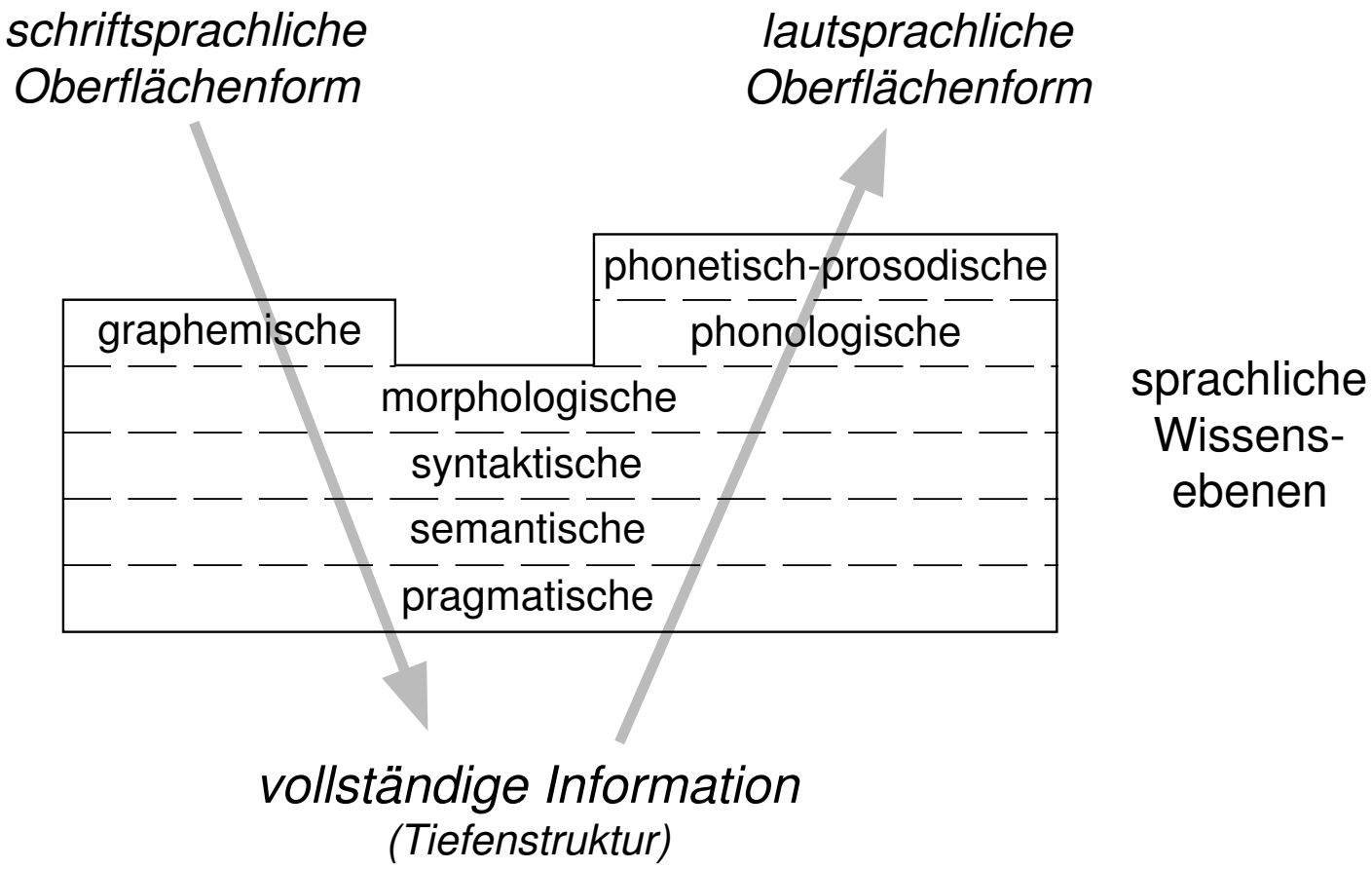
$\longrightarrow$ **Person, die einen Text vorliest**

Kriterien für das richtige Vorlesen:

1. keine Fehler wie:
   a) falsch ausgesprochene Wörter
   b) unpassend gesetzte Satzakzente und sinnwidrige Gruppierungen
2. passender Sprechstil: fröhlich, ernst, kämpferisch, etc. (je nach Situation, Inhalt und Zuhörer)

$\longrightarrow$ Ist nur dann möglich, wenn die Person den Text versteht !

Darstellung des richtigen Vorlesens



Konsequenzen für die Sprachsynthese

Problem: direkte Umsetzung von Text in gesprochene Sprache (bzw. Sprachsignal) ist i.a. nicht möglich

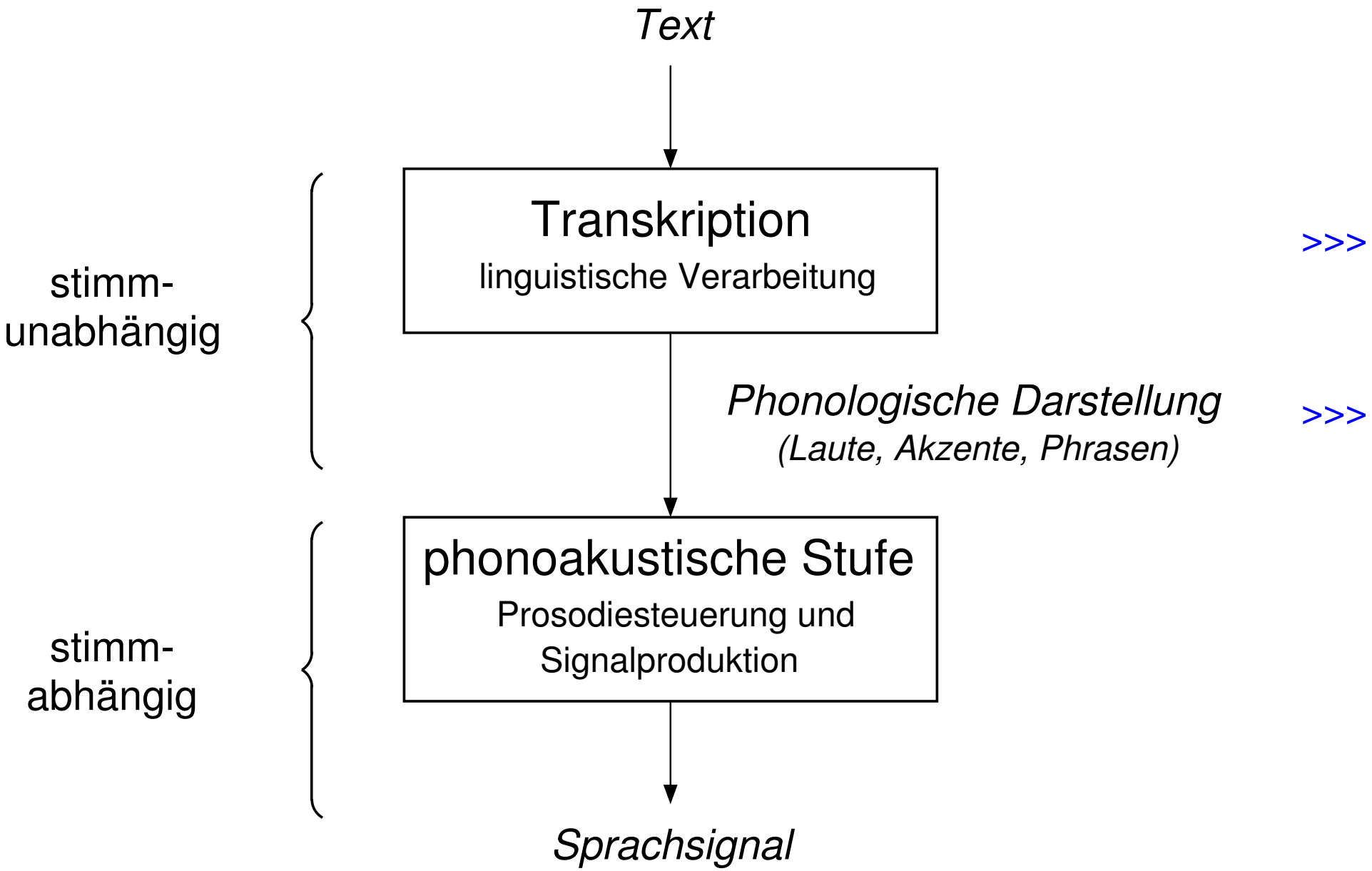
Frage: Was hilft uns nun das Leitbild ?

Antwort: Es ist die “Tiefenstruktur” des Textes zu erzeugen !
(unter Einsatz von Wissen über die Sprache etc.)

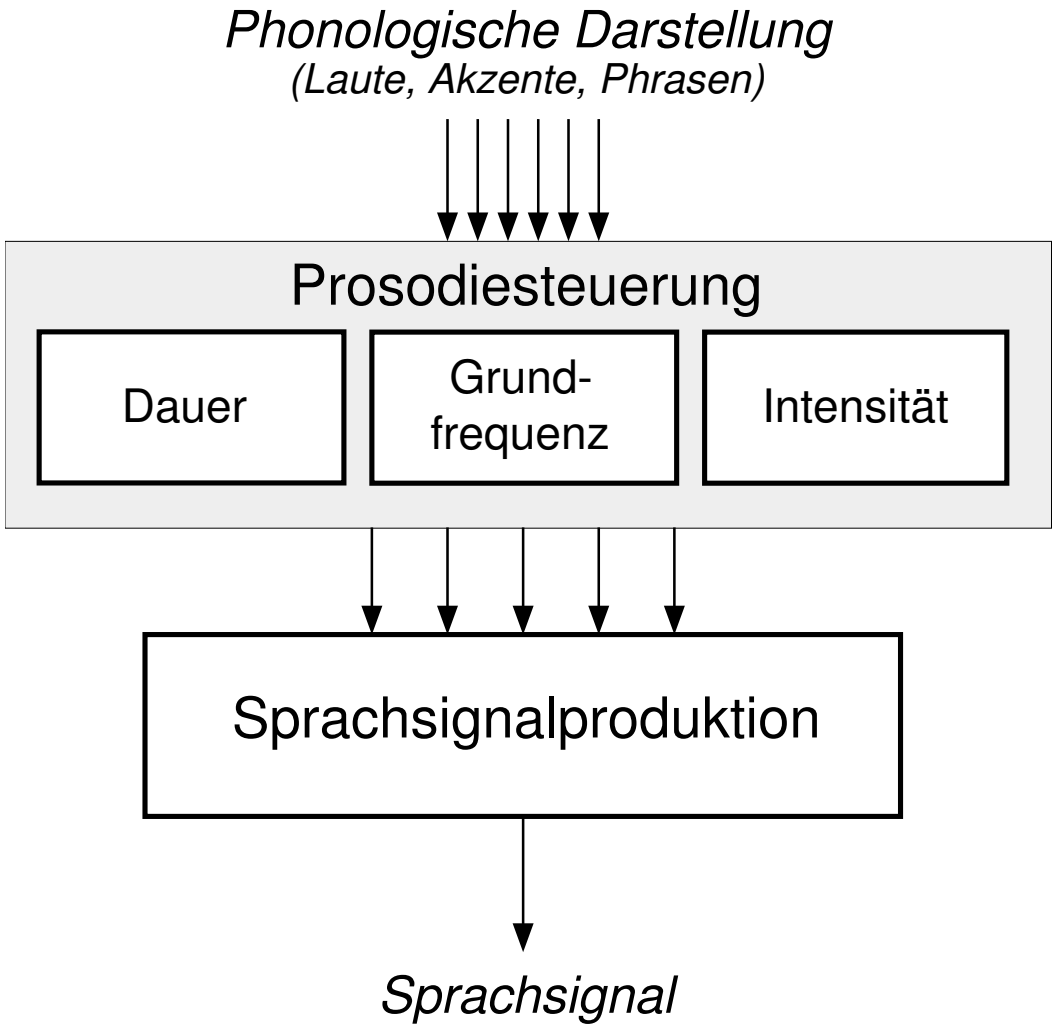
Sprachsynthese ist eine sehr komplexe Aufgabe

Sinnvolle Aufteilung in Teilaufgaben (bzw. Teilsysteme)
mit überprüfbaren Schnittstellen!

stimmunabhängige Systemteile $\Leftrightarrow$ stimmabhängige Systemteile


Text
Transkription
linguistische Verarbeitung
stimm-
unabhängig
>>>
Phonologische Darstellung
(Laute, Akzente, Phrasen)
>>>
phonoakustische Stufe
Prosodiesteuerung und
Signalproduktion
stimm-
abhängig
Sprachsignal

Phonoakustische Stufe



Lautinventar für die Sprachsynthese

Frage: Welche Laute müssen erzeugt werden können?

Hinweise aus der Linguistik:

Phonemik: Inventar der Phoneme (bedeutungsunterscheidende Laute)

$\longrightarrow$ Mindestanforderung für die Sprachsynthese

wegen Unterscheidbarkeit von Wörtern wie /daːtən/ und /taːtən/

Phonetik:

- Lautinventar und Eigenschaften der Laute (Detaillierungsgrad)

  $\longrightarrow$ Grundsatz für die Sprachsynthese:

  grösseres Lautinventar $\longrightarrow$ potentiell bessere Sprachqualität

- Aussprachewörterbuch: (für Deutsch z.B. Duden) >>>

  definiert: "Tuch" $\rightarrow$ [tuːx] bzw. "Bruch" $\rightarrow$ [brʊx]

- nicht alle Aussprachevorschriften in Aussprachelexikon enthalten

  z.B. Aspiration [t]: "Tuch" $\rightarrow$ [t^huːx] bzw. "Stab" $\rightarrow$ [ʃtaːb]

Lautinventar für die Sprachsynthese

Synthese muss korrekte Aussprache produzieren können

Vorgehen:

- Information aus Aussprachewörterbuch wird in Transkription benutzt
  $\longrightarrow$ Lautfolge in der phonologischen Darstellung
- zusätzliche Aussprachevorschriften und feinere Lautdetaillierung[1] werden bei der Sprachsignalproduktion berücksichtigt

[1] Feinere Lautunterscheidungen sind sinnvoll z.B. für den Konsonaten [n], der in “Kahn” schwach und in “kann” stark gesprochen wird, aber im Duden als [kaːn] bzw. [kan] angegeben wird.

Thema der nächsten Lektion:

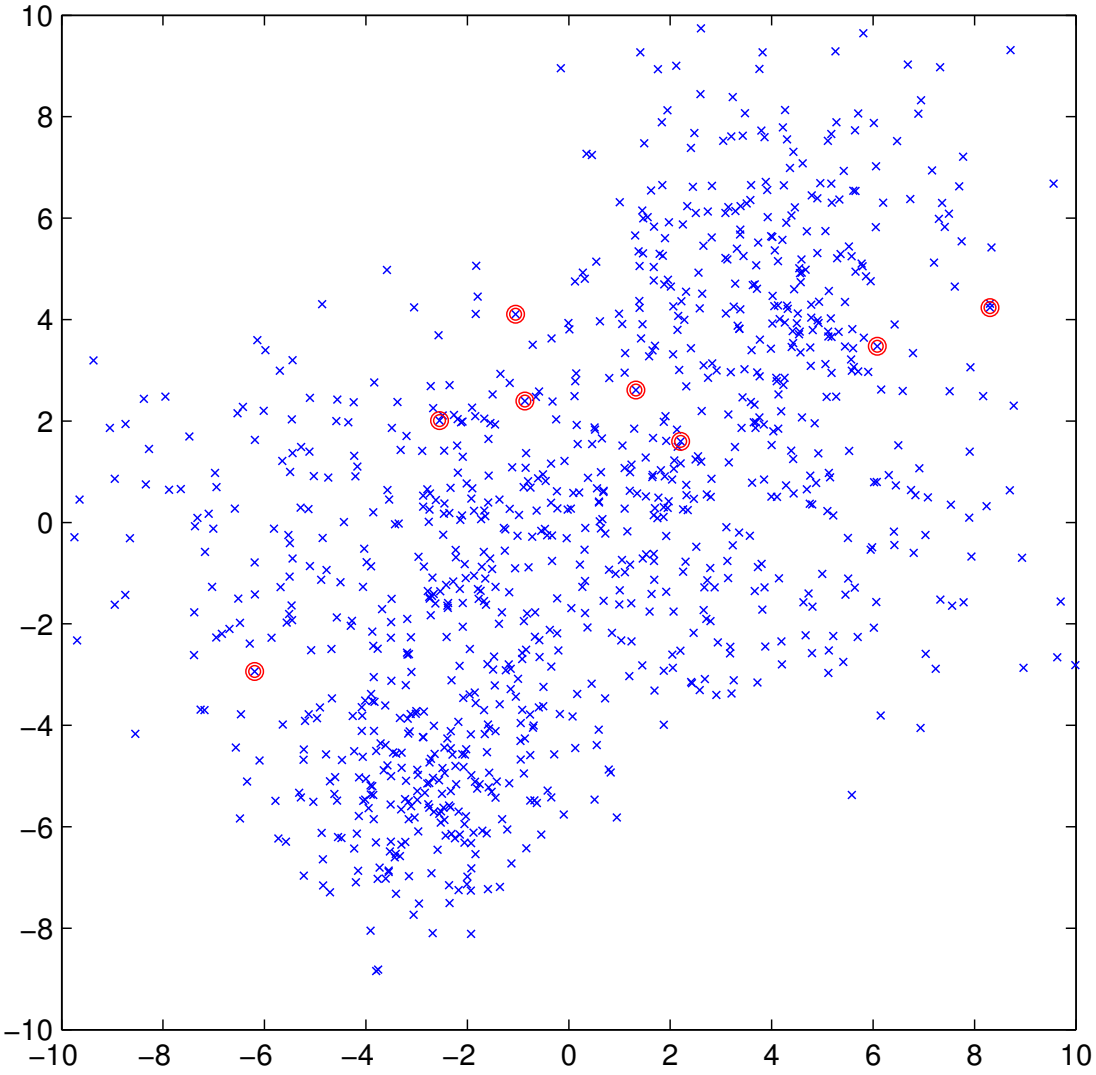
Phonoakustische Stufe

(Methoden der Sprachsignalproduktion)

Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

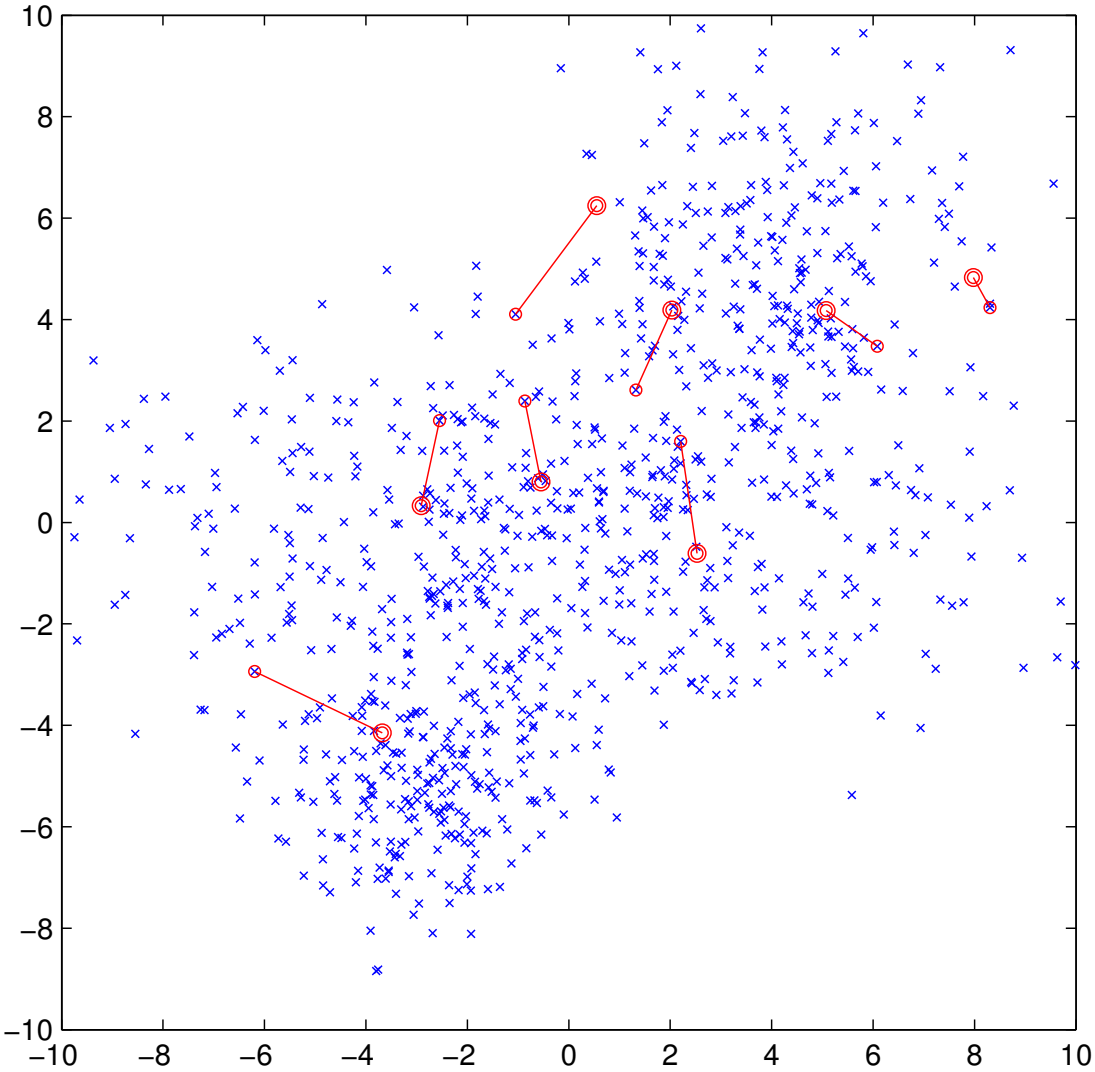
Initialisierung

Codebuchgrösse: 8



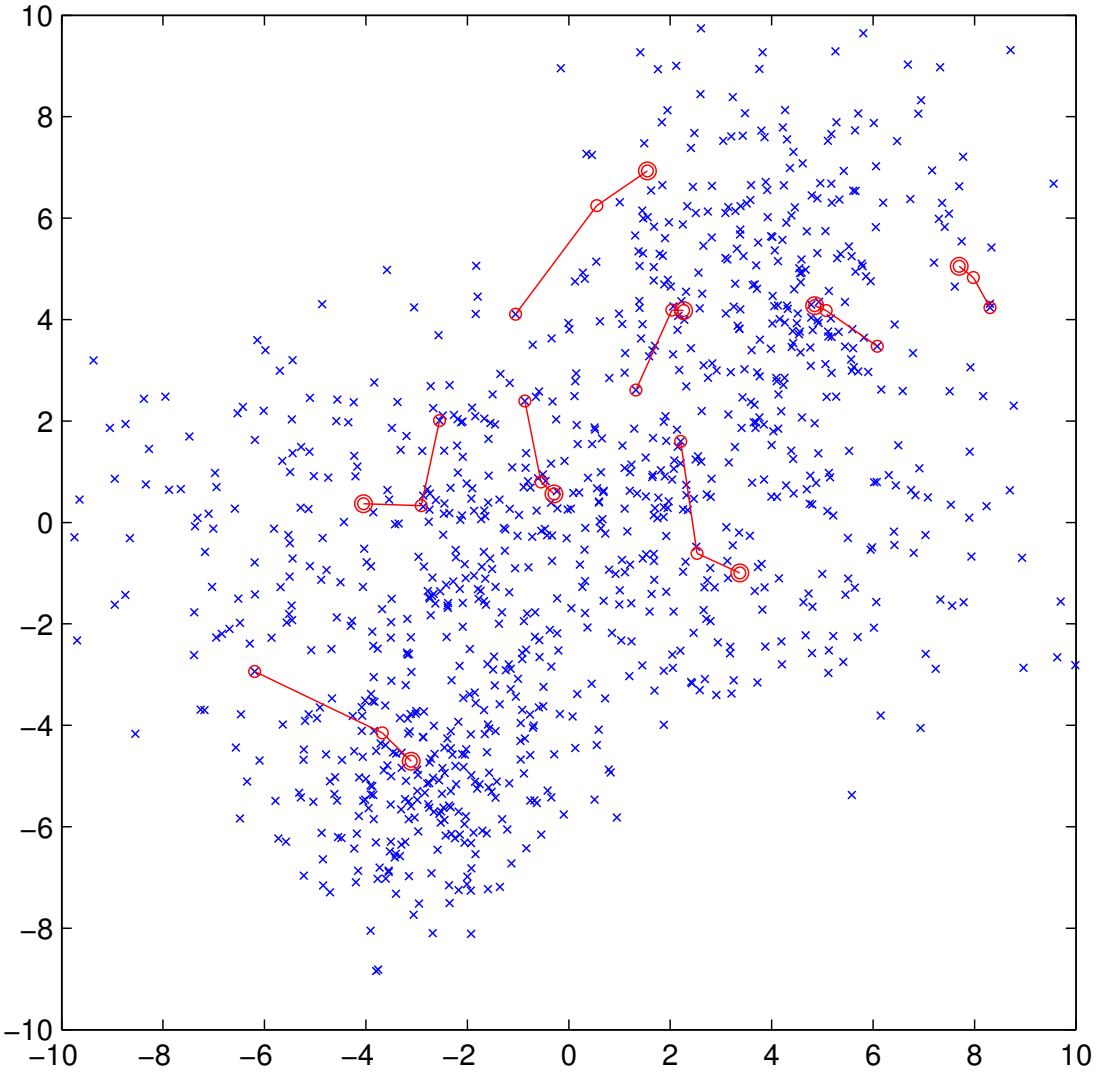
<<<

1. Iteration

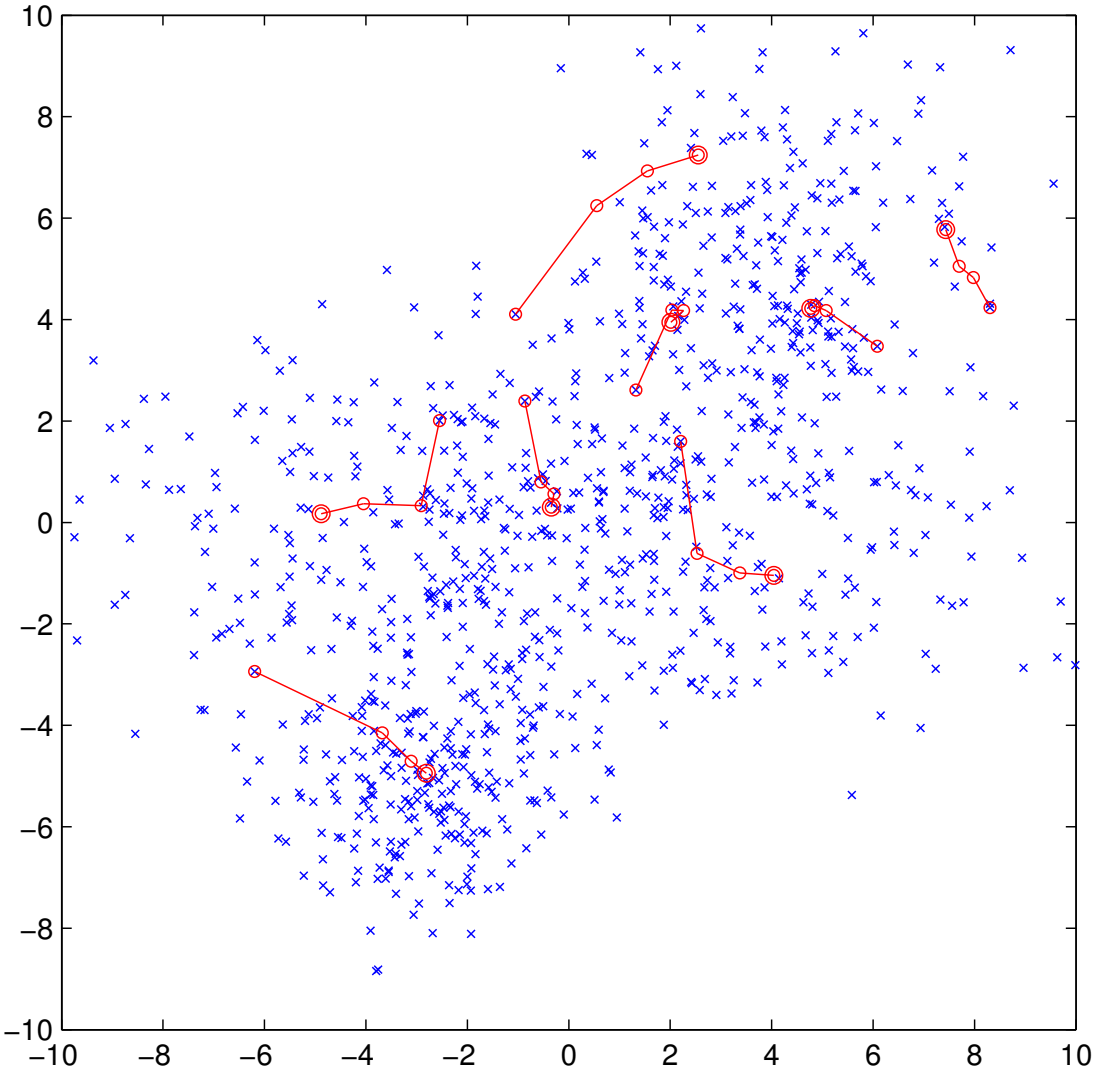


<<<

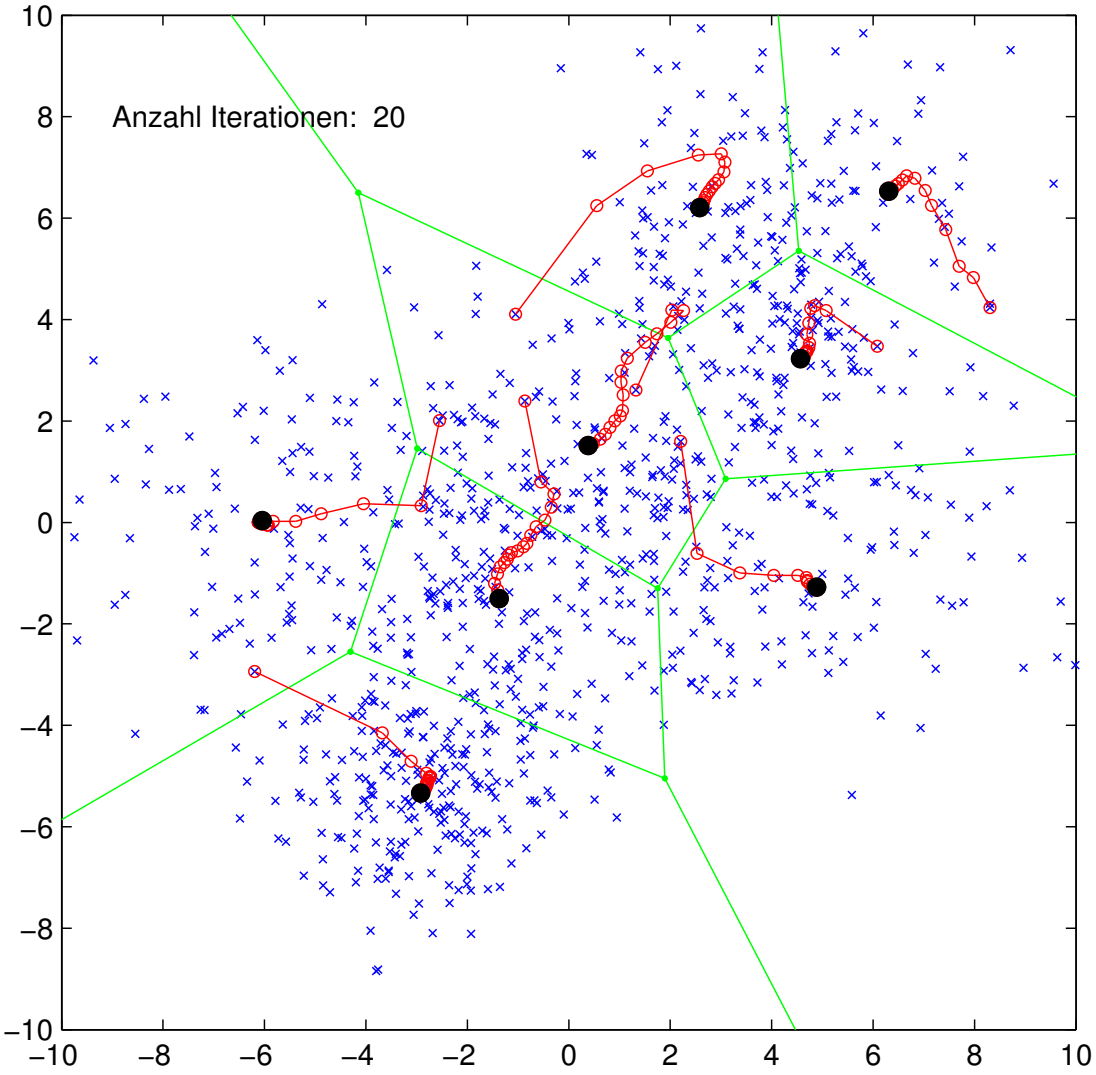
2. Iteration



3. Iteration



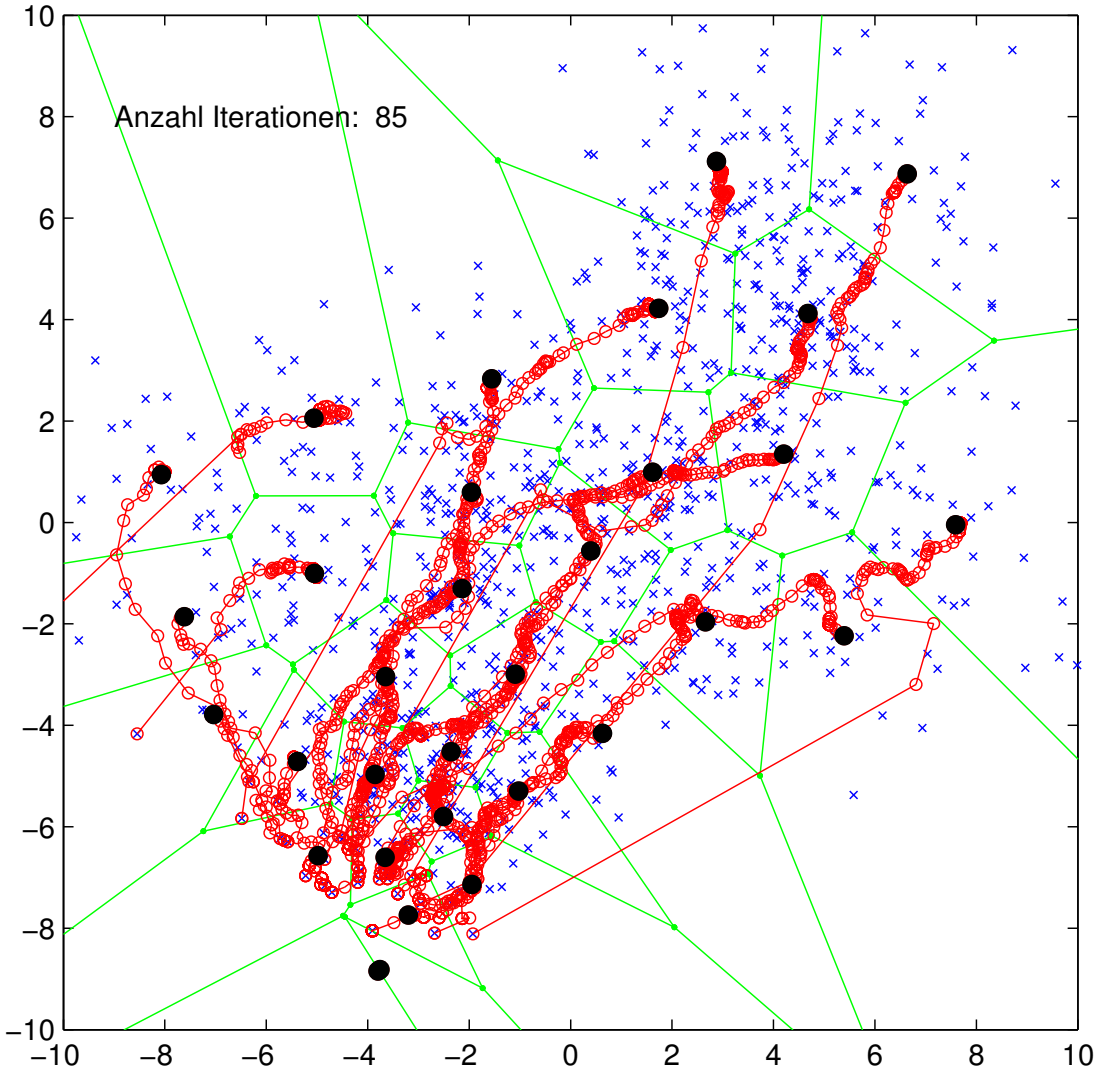
Codebuch der Grösse 8



<<<

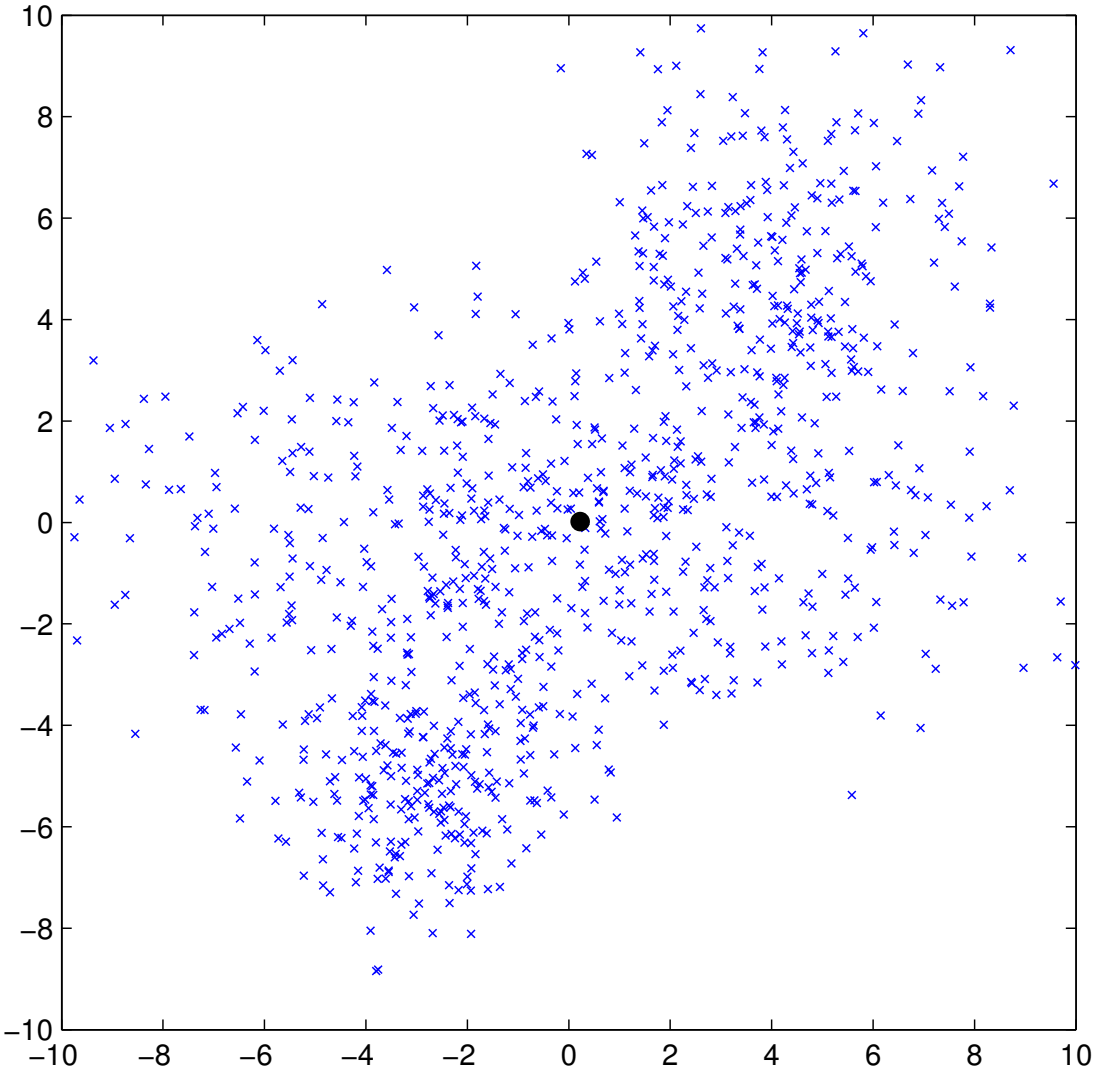
Lokales Minimum des K-means-Algorithmus

($M = 32$)



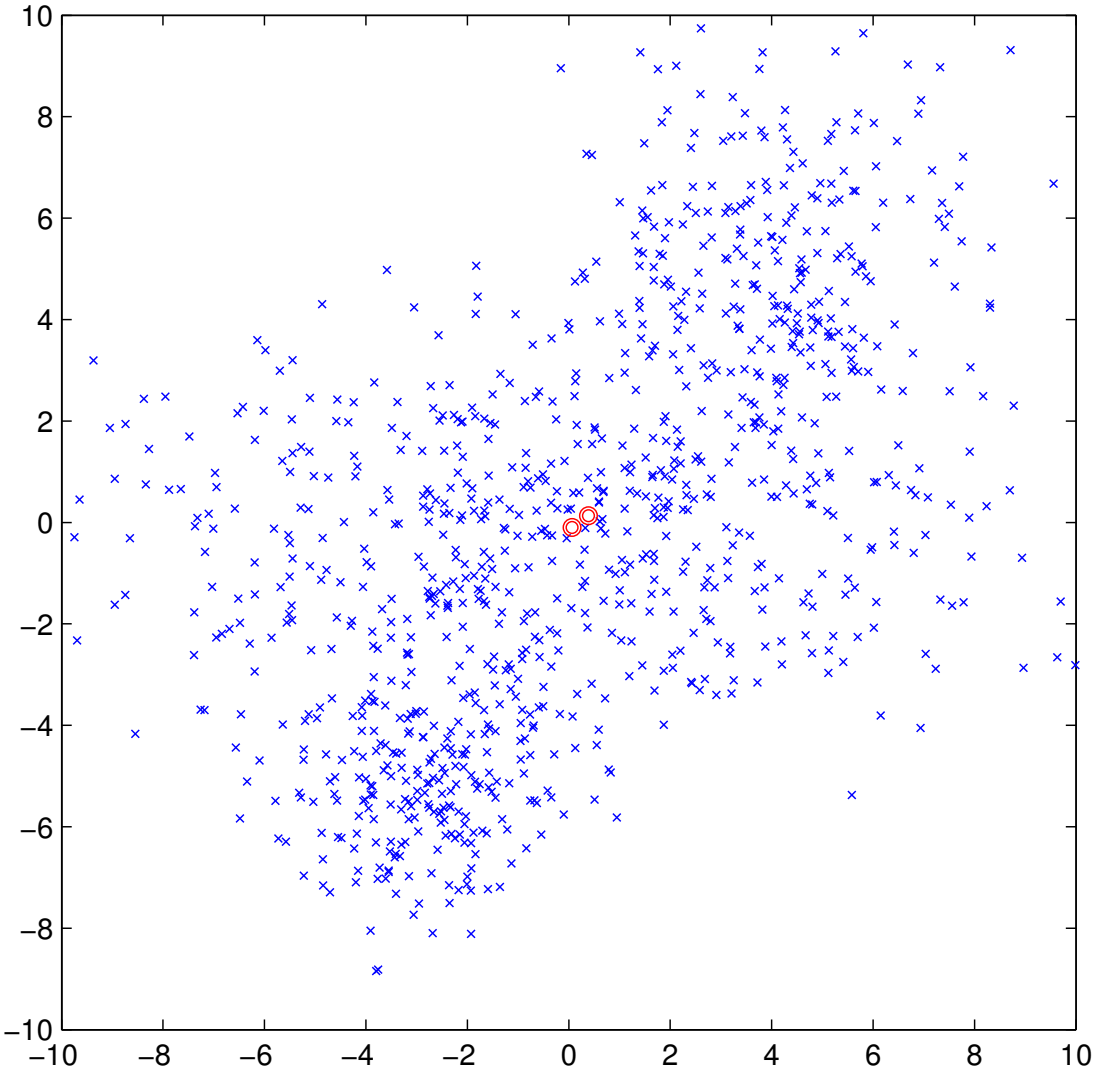
<<<

Initialisierung



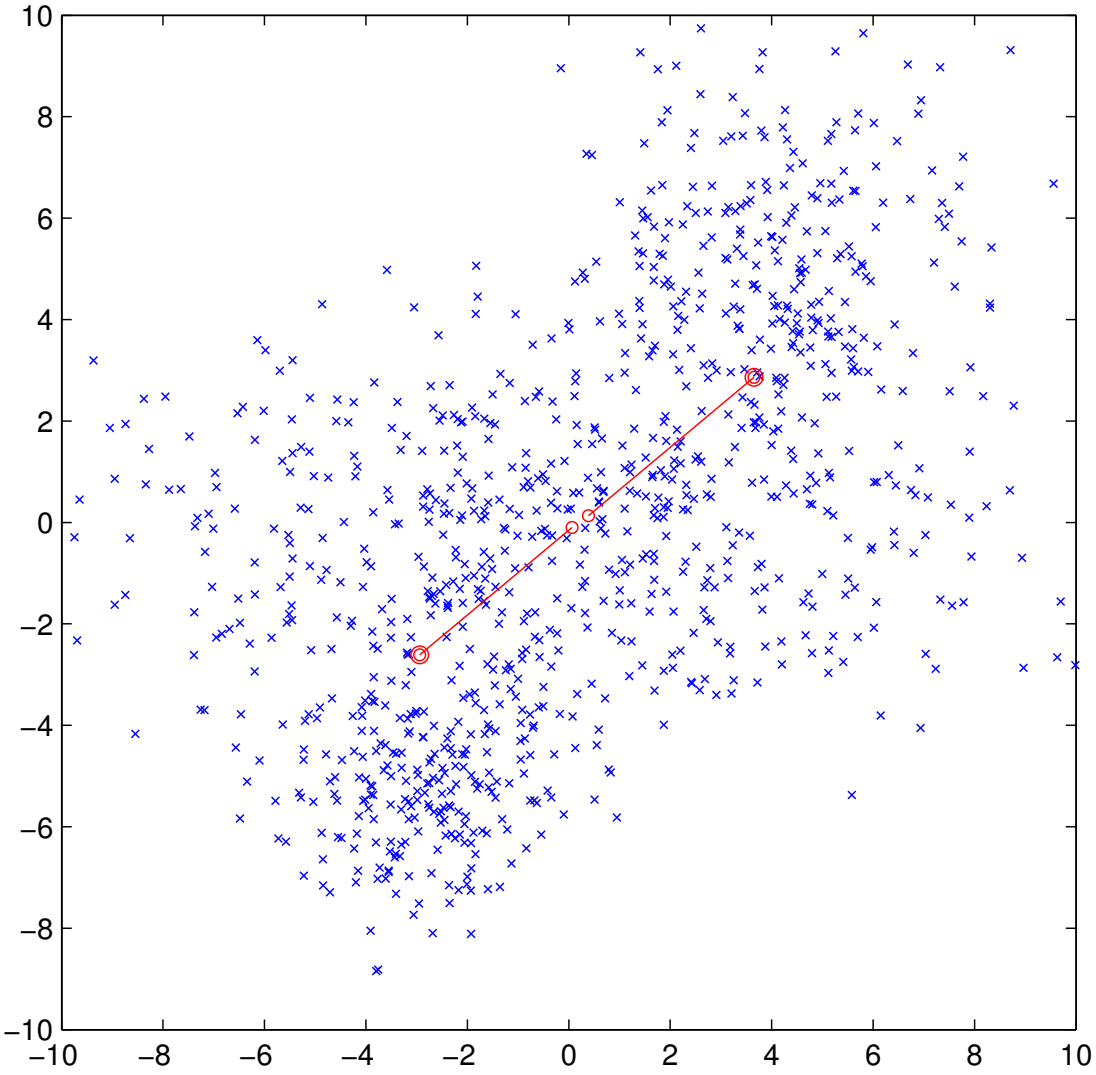
<<<

Splitting

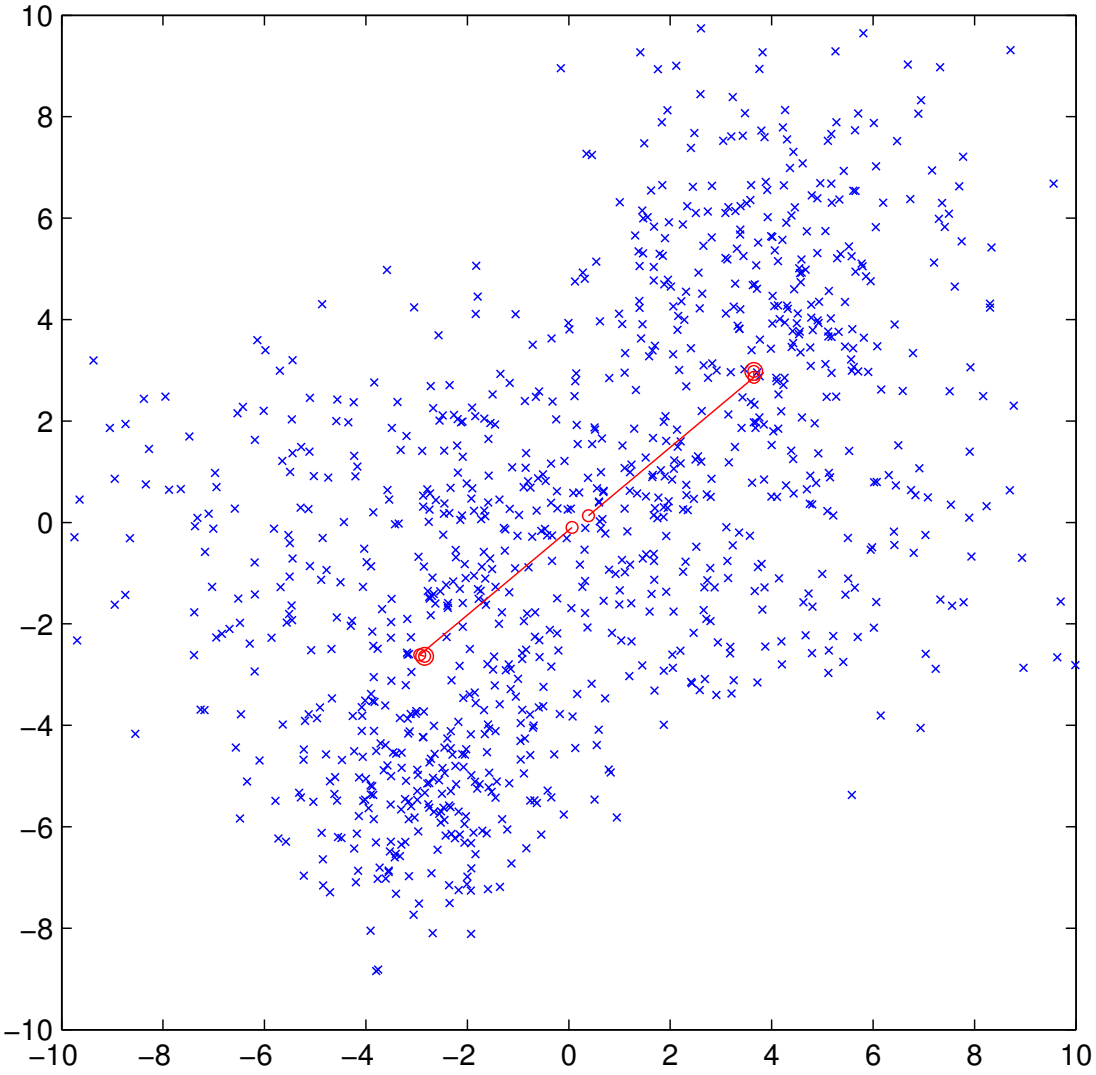


<<<

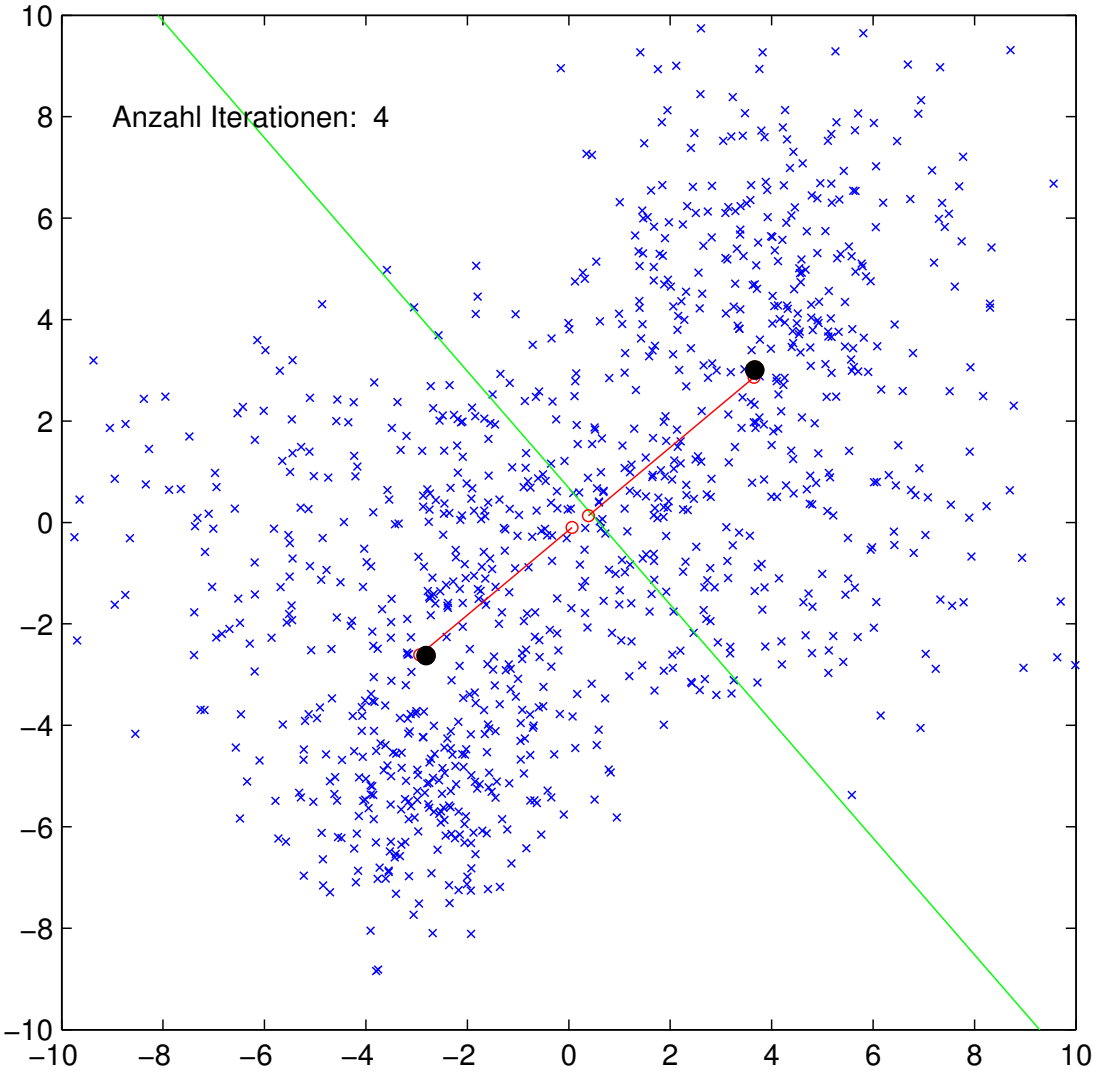
1. Iteration



2. Iteration

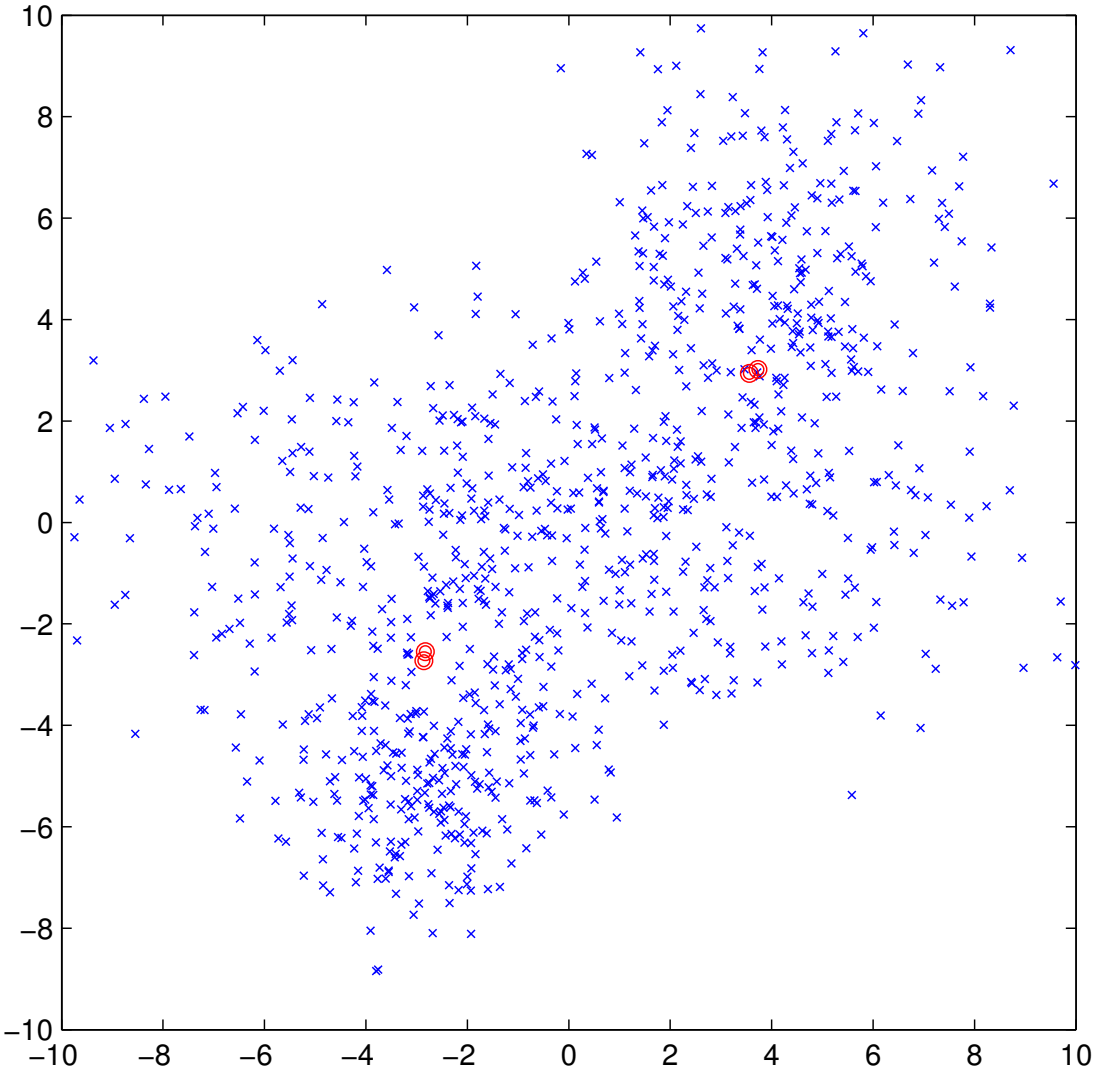


Codebuch der Grösse 2

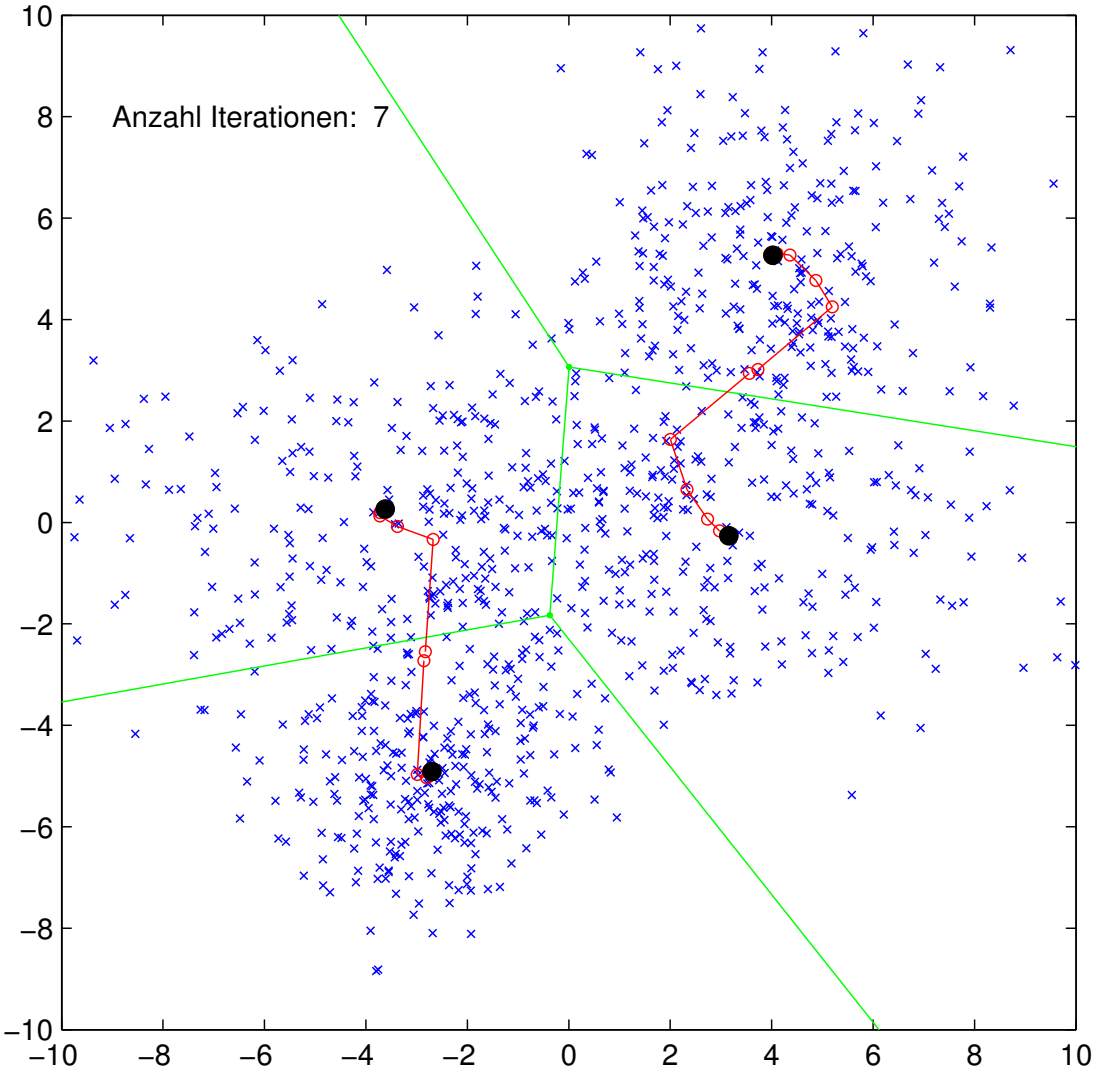


<<<

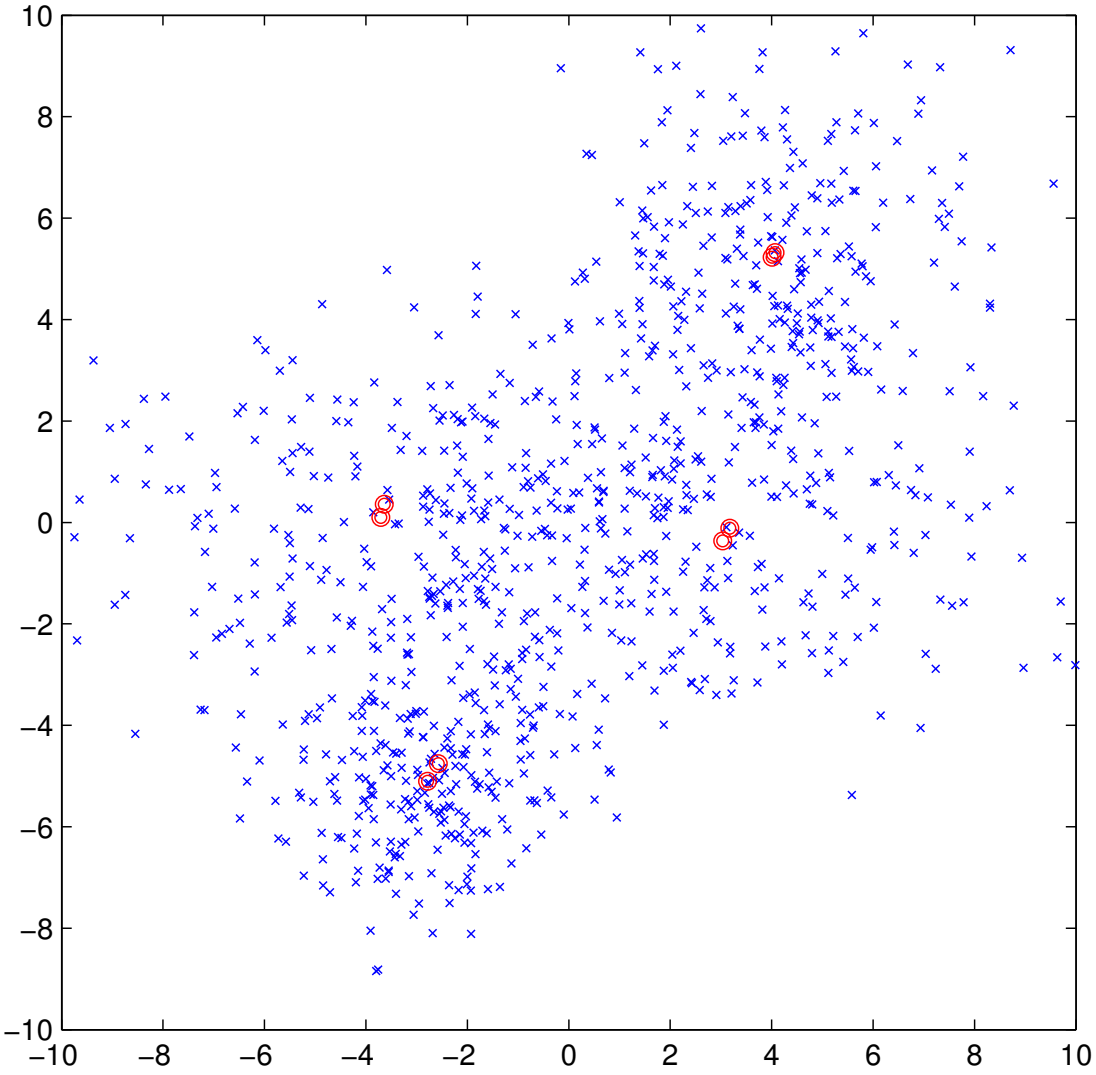
Splitting



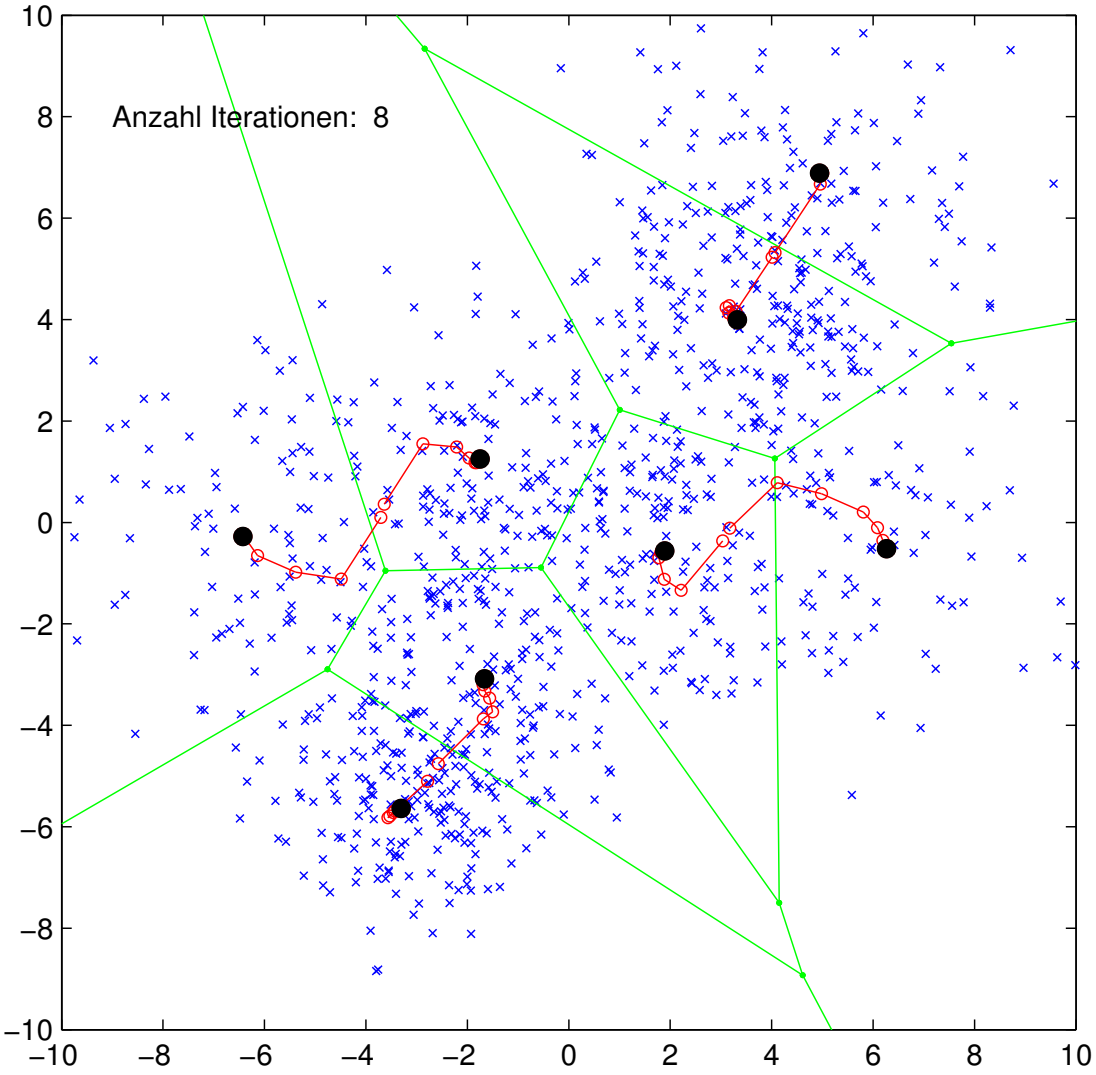
Codebuch der Grösse 4



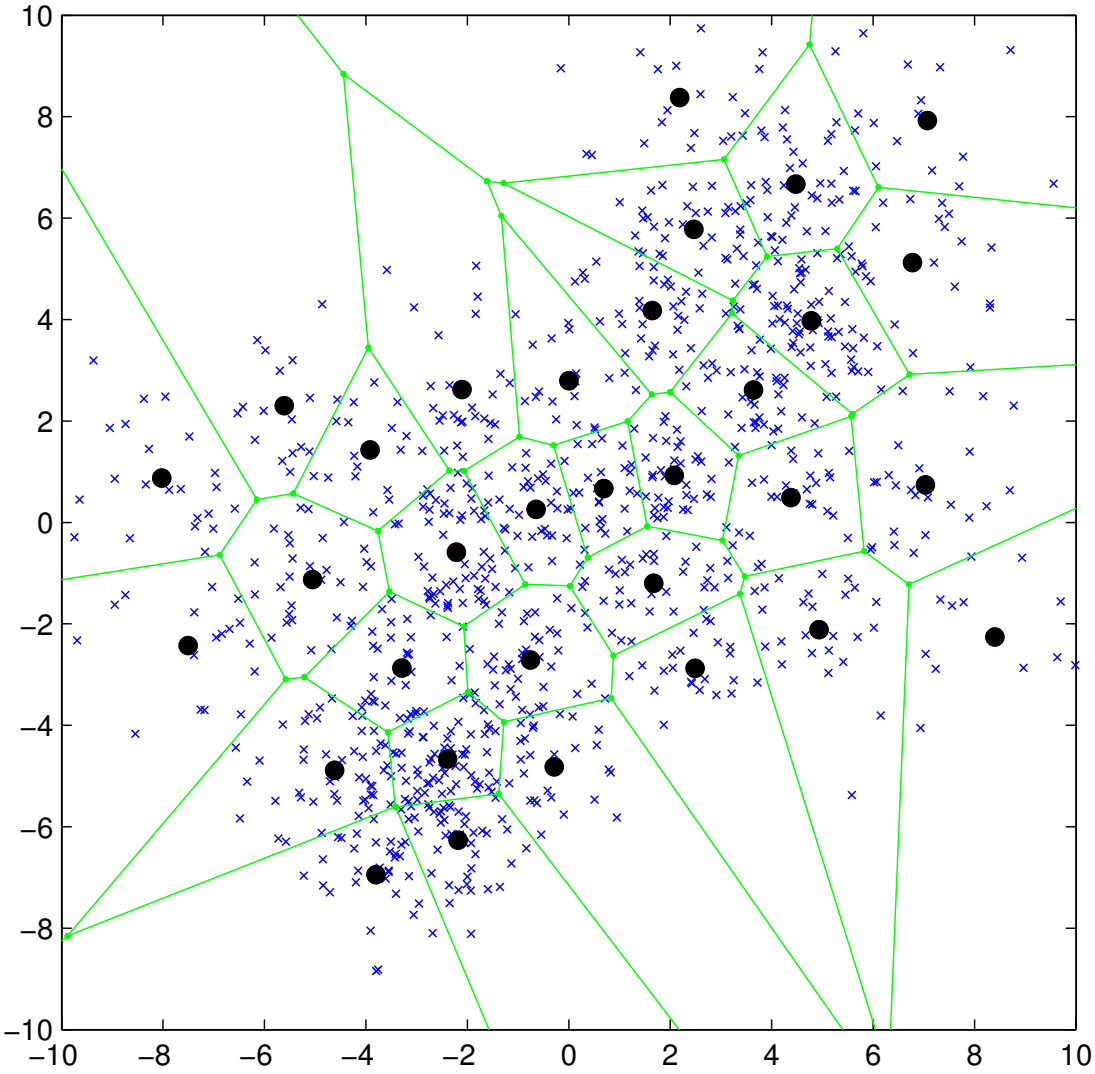
Splitting



Codebuch der Grösse 8



Codebuch der Grösse 32



<<<

Beispiele einer Sprachsynthese

Was fällt Ihnen auf ?

1: Ein Festungsturm überragt alle Häuser der Stadt.

2: Der Redner hat sich thematisch auf das Wesentliche beschränkt.

3: Viele neuere Autos sind mit einem ABS ausgerüstet.

4: Anstelle von “guten Tag” sagt man im Welschland “bonjour”.

5: Die Reise führte von Luzern nach Gersau im Kanton Schwyz.

6: Alten, kranken und invaliden Menschen hilft dies kaum.

7: Hat er sich wirklich eingehend beraten lassen?

8: Sein Kollege hat an der ETH in Zürich studiert.
Seine Freundin hat auch in Zürich studiert.

<<<

Ausgabe der Transkriptionsstufe:

phonologische Darstellung

(stimmunabhängige Beschreibung des zu erzeugenden Sprachsignals)

Eingabetext: Heinrich besuchte gestern die Ausstellung im Kunstmuseum.

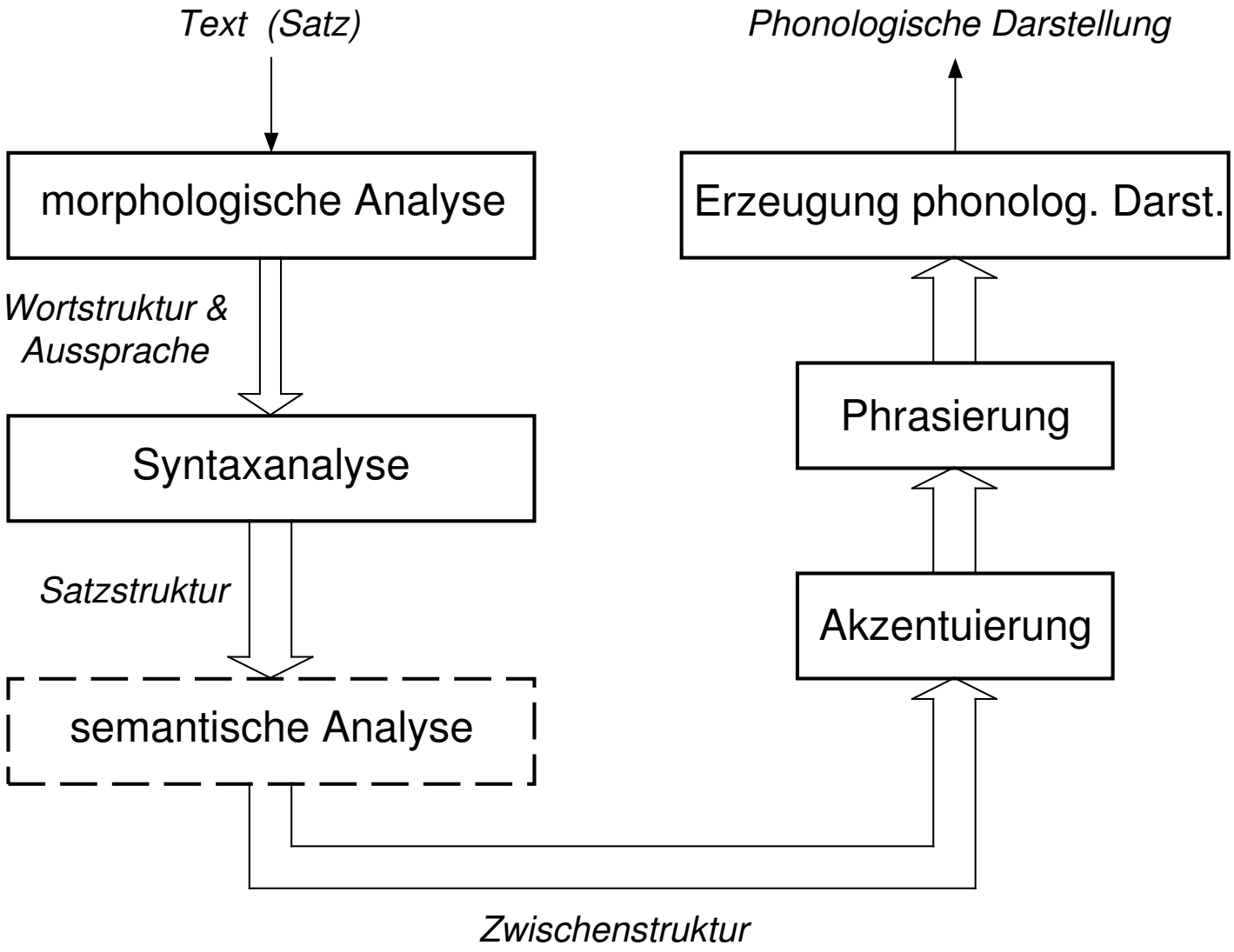
phonol. Darst.: (P) [1]haiˍn-rɪç #{2} (P) bə-[2]zuːx-tə [1]gɛs-tərn #{4}
(T) diː [2]|aʊs-[4]ʃtɛ-lʊŋ |ɪm [1]kʊnst-mu-[4]zeː-ʊm.

in ETHPA-Notation:

(P) [1]ha_in-rIC #{2} (P) b@-[2]zu:x-t@ [1]gEs-t@rn #{4}
(T) di: [2]?a_us-[4]StE-lUN ?Im [1]kUnst-mu-[4]ze:-Um.

<<<

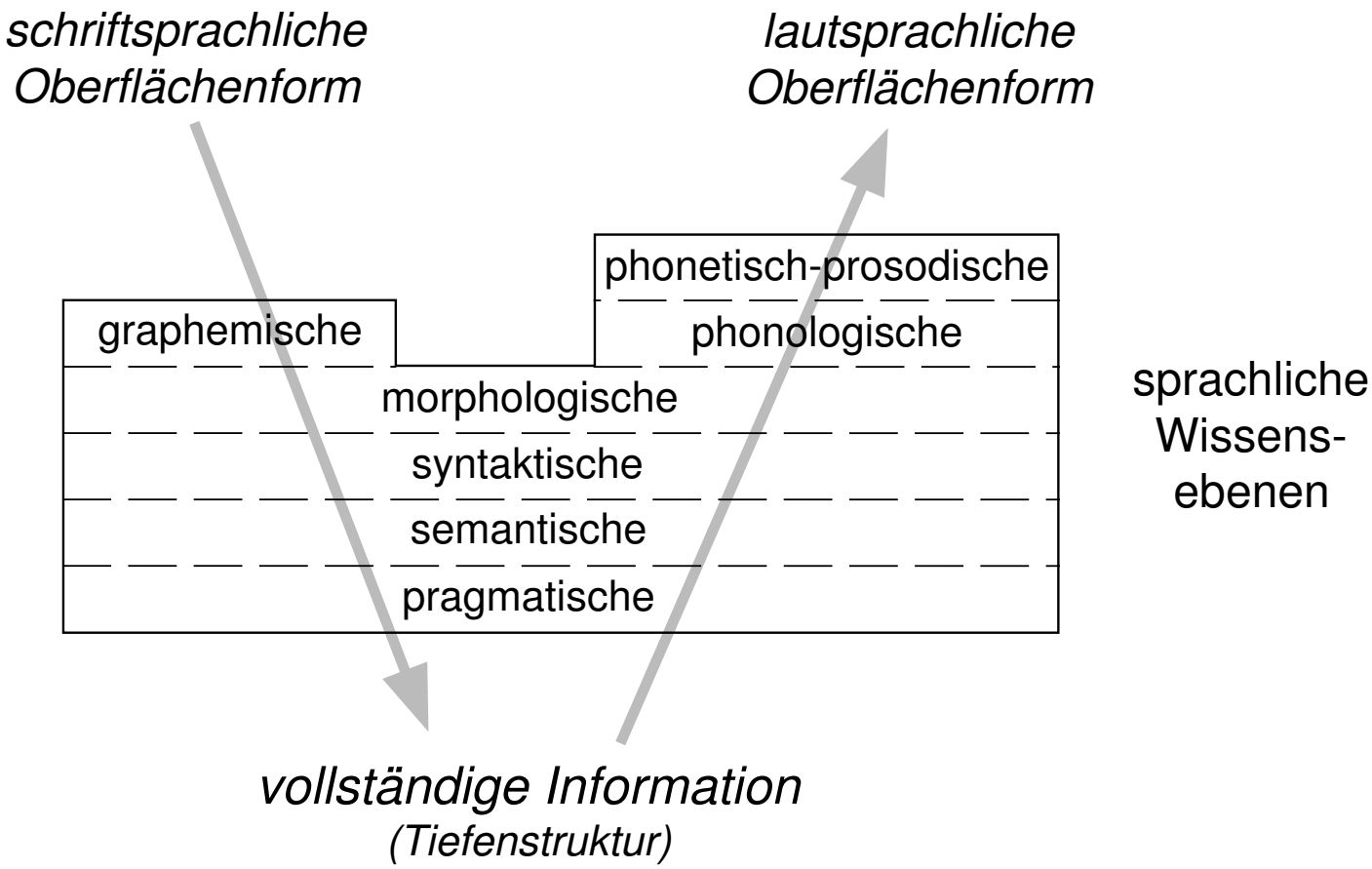
Transkription



>>>

<<<

Darstellung des richtigen Vorlesens



<<<

Aussprachewörterbuch

Ein Aussprachewörterbuch ist für die Sprachsynthese unentbehrlich,
weil es keine Regeln gibt, wie z.B. die folgenden Wörter auszusprechen sind:

mit [a]: Bach, Dach, Krach, schwach ...

mit [aː]: brach, nach, sprach, stach ...

<<<