# Zählen & Zerlegen

Ein Algorithmus zur morphologischen Segmentierung von Texten

FELIX GOLCHER betreut von ANKE LÜDELING

Humboldt-Universität zu Berlin — Institut für deutsche Sprache und Linguistik — Korpuslinguistik

## Worum geht's?

- Die Zerlegung eines Textes in seine Einzelteile ist Voraussetzung jeder maschinellen Verarbeitung.
- Leerzeichen trennen Worte, aber nicht Morpheme.
- Nicht alle Sprachen verwenden Leerzeichen.

⇒ Das Problem ist wichtig und keineswegs trivial.

Hier wird kein Lexikon verwendet, nur unannotierter Trainingstext.

## Grundprinzip

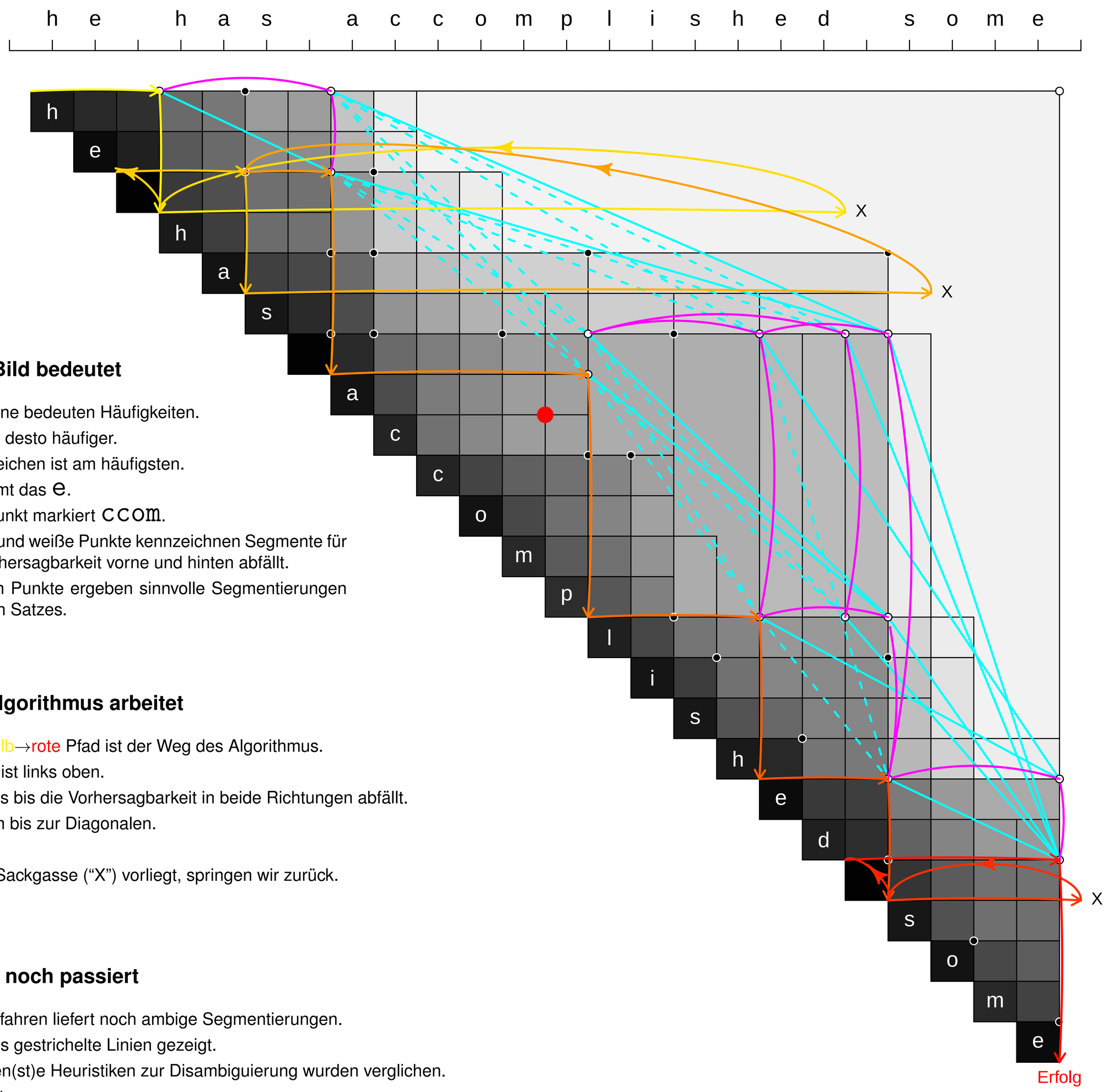
**Nach vorne**

- Das `h` nach `accomplis` ist vorhersagbar.
- Das dann folgende `e` viel weniger.

**Nach hinten**

- Das Leerzeichen vor `accomplish` ist vorhersagbar.
- Das `s` davor viel weniger.

⇒ Der Satz wird vollständig in Zeichenketten mit dieser Eigenschaft zerlegt.




## Was das Bild bedeutet

- Die Grautöne bedeuten Häufigkeiten.
- Je dunkler, desto häufiger.
- Das Leerzeichen ist am häufigsten.
- Dann kommt das `e`.
- Der rote Punkt markiert `ccom`.
- Schwarze und weiße Punkte kennzeichnen Segmente für die die Vorhersagbarkeit vorne und hinten abfällt.
- Die weißen Punkte ergeben sinnvolle Segmentierungen des ganzen Satzes.

## Wie der Algorithmus arbeitet

- Der von gelb→rote Pfad ist der Weg des Algorithmus.
- Startpunkt ist links oben.
- Nach rechts bis die Vorhersagbarkeit in beide Richtungen abfällt.
- Nach unten bis zur Diagonalen.
- Von vorne.
- Falls eine Sackgasse ("X") vorliegt, springen wir zurück.

## Was dann noch passiert

- Dieses Verfahren liefert noch ambige Segmentierungen.

  ⇒ Im Bild als gestrichelte Linien gezeigt.
- Verschieden(st)e Heuristiken zur Disambiguierung wurden verglichen.
- Übrig bleibt:

```
([he ][ has ])([{accomp}{lish}][ed ])(some)
```

## Literatur

Bebel, August (2004). *Aus meinem Leben – Erster Teil*[1]. URL: http://www.gutenberg.org/files/12267/12267-8.txt (besucht am 17. 11. 2006).

Harris, Zellig S. (1955). "From Phoneme to Morpheme". In: *Language* 31.2. reprinted in Hiż 1970, S. 190–222.

Hiż, Henry, Hrsg. (1970). *Papers in Structural and Transformational Linguistics*. Holland: Dordrecht.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.