

# Grundlagen der Analyse von Sprachdatenbanken

Jonathan Harrington

# Die Laborsprache

Fast alle Untersuchungen in der Experimentalphonetik im 20. Jahrhundert befassen sich mit der Laborsprache. Warum?

‘Competence/performance’ Trennung in Chomsky & Halle (1968). Nur die Laborsprache ist relevant für die phonologische Darstellung von Wörtern...daher kaum eine Befassung im ‘SPE’ mit Intonation, Prosodie, Spontansprache.

Synthese u. Perzeptionsexperimente (Haskins Labs.). Seit 1950 eine Beschäftigung mit den **minimalen akustischen Unterschieden** (perceptual cues) zwischen Phonemen.

Fortschritte in der Sprachsynthese: durch die Analyse von isolierten, laborgesprochenen Silben.

# Warum Die Laborsprache?

Technische Überlegungen. Bis mitte der 80er Jahre:

- Kaum Möglichkeiten mehr als 15 Sekunden digitaler Sprache zu speichern
- Festplatten meistens nicht größer als 500 mB für das gesamte Labor.
- 3 – 4 Minuten um **einen** Spektrogramm digital darzustellen.

# Der Bedarf an Sprachdatenbanken

- Entsteht in den 80er Jahren wegen der Entwicklung von Systemen, die die gesprochene Sprache **automatisch** erkennen sollen
- Um solche Systeme zu erstellen, muss die **Variabilität** der gesprochenen Sprache (Änderungen wegen der Sprechgeschwindigkeit, Sprecherunterschiede) modelliert werden.
- Jedoch kann die Variabilität (wegen Sprecherunterschiede, Spontansprache usw.) nur **in einem sehr umfangreichen Sprachdatenbank** modelliert werden.

# Sprachdatenbanken

Was sind Sprachdatenbanken?

Sammlung von Sprechdaten

Hierzu können gehören:

Zeitsignale (digital): akustische, artikulatorische

Aus dem Zeitsignal abgeleitete Signale z.Bsp.  
spektrale Daten

Etikettierungen der Signale: Verknüpfung von  
Signalteilen an Symbole

# Probleme in der Entwicklung von Sprachdatenbanken seit 1985

Was soll angenommen werden ? (sogar hier konnten die Phonetiker kaum helfen, wegen ihrer Beschäftigung mit der Laborsprache)

**Keine Infrastruktur.** Die Erstellung von Sprachdatenbanken ist arbeitsintensiv und verlangt Kooperationen zwischen Laboren. Aber es gab weder Software noch Hardware um umfangreiche S-Datenbanken miteinander zu entwickeln (und kein Internet!).

**Keine Richtlinien für die Etikettierung.** Was sollte man etikettieren und wie (welche Einheiten usw.)?

**Theorie-neutral.** Viele Phonetiker hielten sich davon fern, weil der Aufbau von Sprachdatenbanken als ein uninteressantes, theorie-neutrales Problem betrachtet wurde.

# Sprachdatenbanken heutzutage

1. Mehrere Sprachdatenbanken sind heutzutage vorhanden, z.B.

- Linguistic Data Consortium (LDC)
- Bavarian archive for speech signals
- Tones and Break Indices (für prosodische Etikettierung)
- IPDS Kiel corpus of spontaneous speech

2. 'Open-source' Software für die Analyse von Sprachdatenbanken und für die digitale Sprachverarbeitung kann jetzt vom Netz heruntergeladen werden.

Jedoch:

- Sind nur die wenigsten Sprachdatenbanken etikettiert worden
- Ist der größte Mehrheit von S-datenbanken für die englische Sprache
- kaum allgemeine Richtlinien, wie Sprachdatenbanken etikettiert werden sollen (daher werden etikettierte S-Datenbanken zu wenig zwischen Laboren ausgetauscht).

# Aufbau & Funktion von Sprachdatenbanken

## Erstellung

Praat, EMU

Digitale Zeitsignale  
akustisch,  
artikulatorisch

Signalverarbeitung

abgeleitete Signale  
(z.B. spektrale Daten)

Etikettieren

Verknüpfung mit  
Symbolen

Signale

Etikettierung

## Abfrage

EMU

Die Signale davon

Abfrage

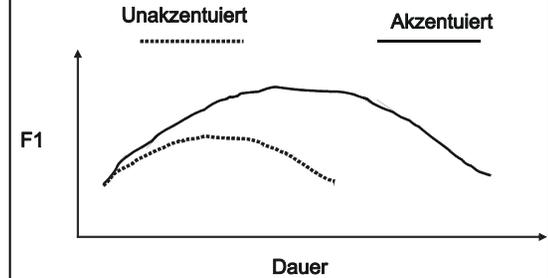
Eine Liste  
ausgewählter  
Symbole mit  
Zeiten

Abfrage

## Analyse

EMU-R

Graphik



Statistik

**p << 0.001**

Weitere  
Signalverarbeitung

# Aufbau einer Sprachdatenbank: Software

Praat: Doing Phonetics by Computer.

Phonetik Institut, Amsterdam, 1998-

<http://www.praat.org>

EMU: Ein System für die Analyse von  
Sprachdatenbanken 1988-

Macquarie Universität, Sydney und IPDS, Kiel

<http://emu.sourceforge.net/>

R: Eine allgemeine Programmiersprache. Das  
'open-source' Äquivalent von der 'S' ('Splus')  
Programmiersprache. 1988-

<http://www.r-project.org/>

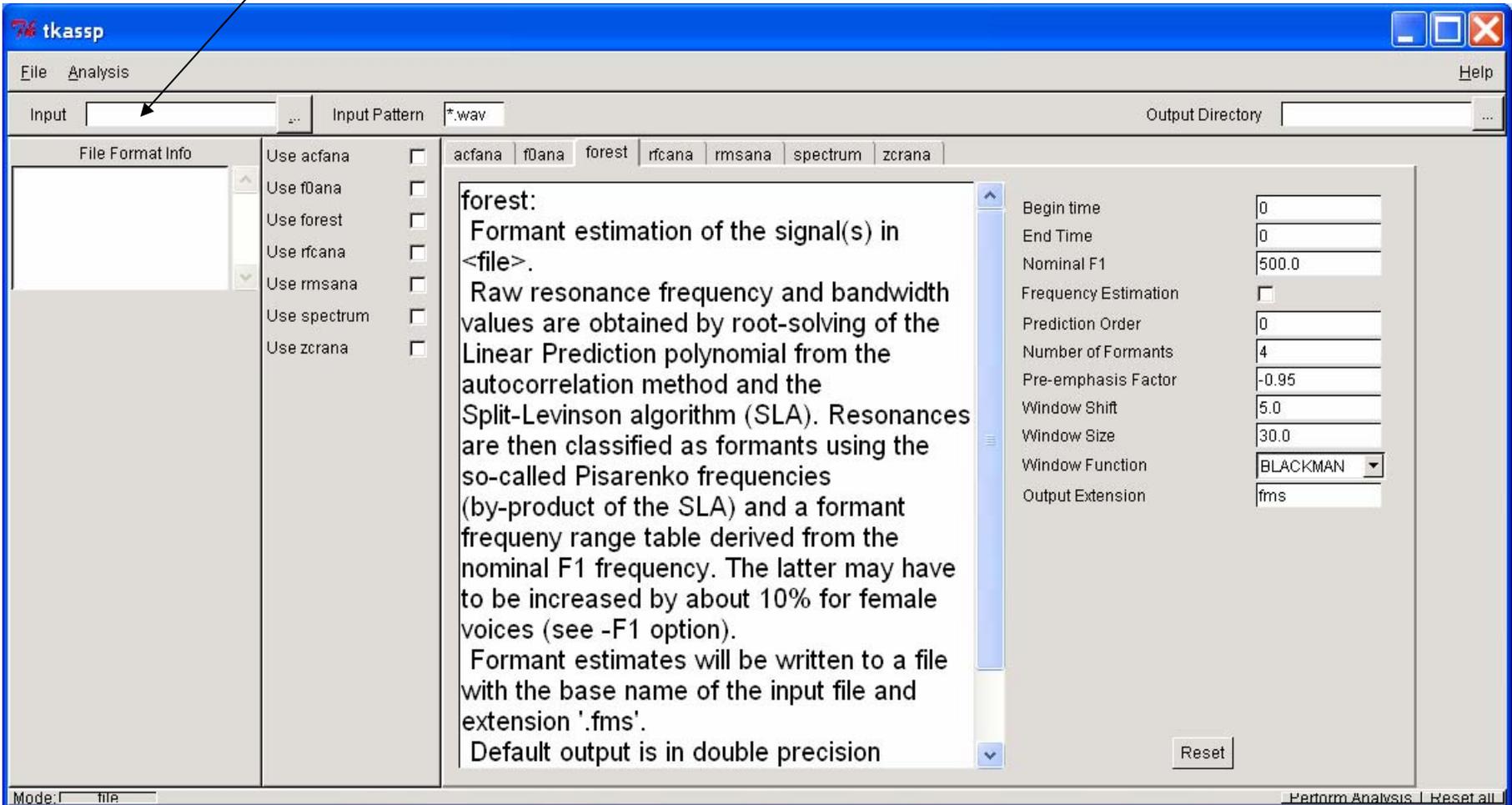




# 1. Sprachdatenbank Erstellung

Signalverarbeitung: tkassp (ASSP in EMU)

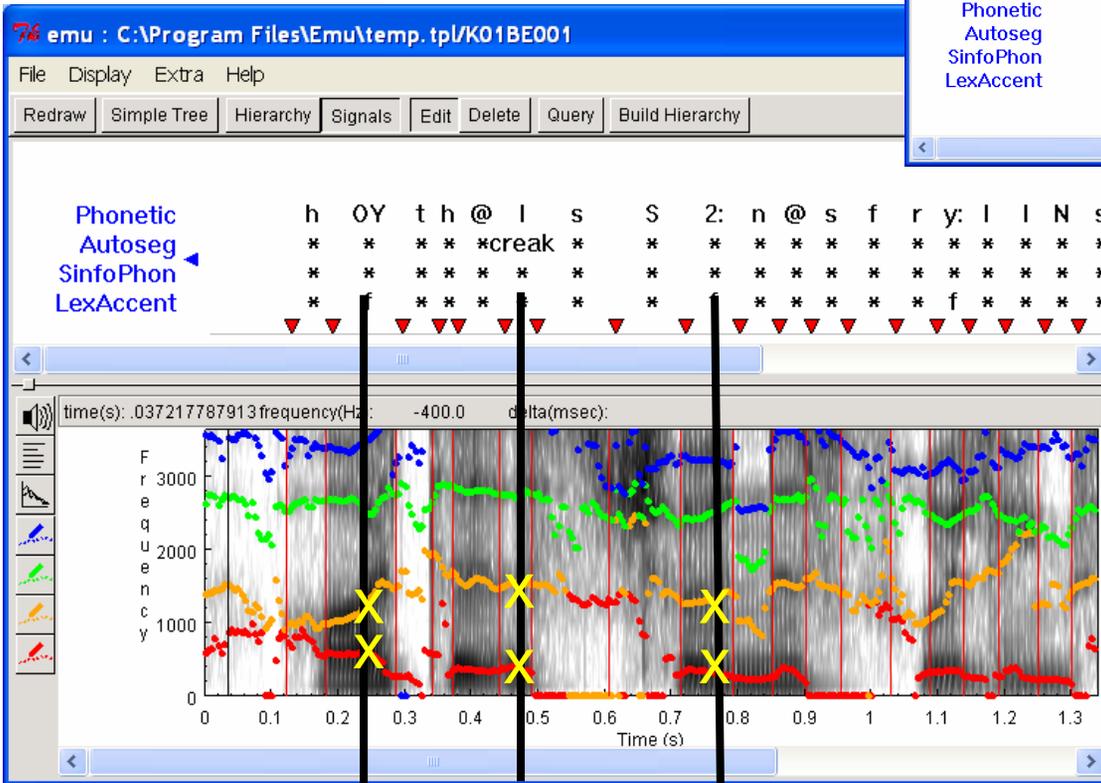
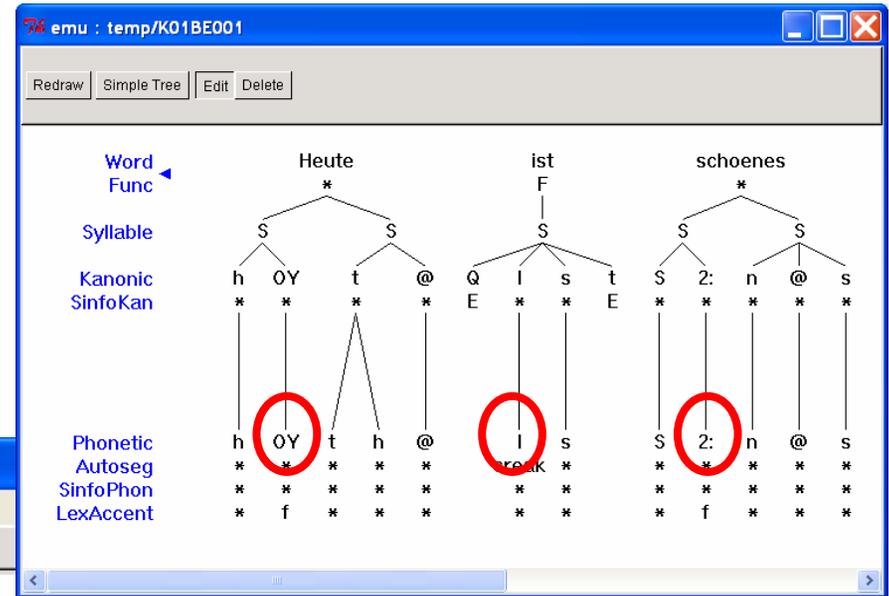
z.B. alle wav-Dateien einer Sprachdatenbank



# 2. Sprachdatenbank Abfrage

## EMU

z.B. Alle Vokale in Wortinitialen Silben und deren ersten 2 Formanten zum zeitlichen Mittelpunkt finden



# 3. Sprachdatenbank Analyse

R Programmiersprache, und in R-EMU Funktionen

z.B. F1 x F2 Verteilung von [ɪ a u] in einer Sprachdatenbank

