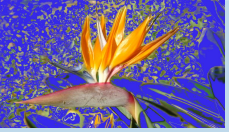
Vorlesung: Bio-Datenbanken

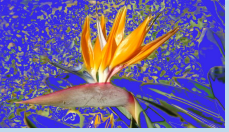
Kapitel 3: Gleichheit / Ähnlichkeit

Dr. Dieter Sosna

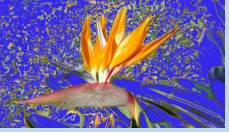
8. November 2007



Kapitel 3: Gleichheit / Ähnlichkeit

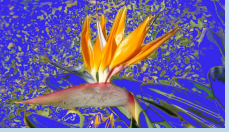


Allgemeines

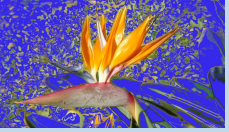


Zwischenstand

- Bisher Integration auf Instanzbasis: Gegenen eine Instanz in Quelle A, welche Einträge aus Quelle B gehören semantisch dazu?
  Lösung durch Vergleich charakterischer Werte (in der Art der Schlüsselkandidaten).
  Noch zu diskutieren: Behandlung kleiner Abweichungen.
- Ziele dieses Kapitels:
  Diskussion des Gleichheitsbegriffs
  Abschwächung zur Ähnlichkeit
  Ähnlichkeit, die auf Gleichheit beruht.



Gleichheit



Gleichheitsbegriff

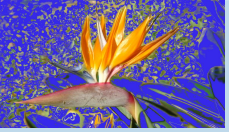
- Spez. Äquivalenzbegriff, muss für jede Kategorie neu definiert werden
- Eigenschaften:

  Definition: Seien x, y, z Elemente einer Kategorie $\mathcal{K}$.

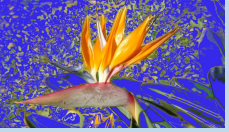
  Eine Relation über $\mathcal{K} \times \mathcal{K}$ ist ein Äquivalenzbegriff in $\mathcal{K}$, wenn sie die folgenden Eigenschaften hat:

$$
\begin{array}{rcll}
x & = & x & \text{Reflexivität (R)} \\
x = y & \leftrightarrow & y = x & \text{Symmetrie (S)} \\
x = y \text{ und } y = z & \leftrightarrow & x = z & \text{Transitivität (T)}
\end{array}
$$

  *

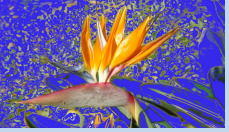


Gleichheit - Beispiele



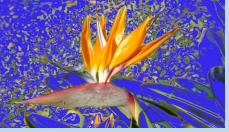
Beispiele: Mengen

- Mengen:
  Zwei Mengen $\mathcal{A}, \mathcal{B}$ sind gleich g.d.w. sie die gleichen Elemente haben.
  (d.h. Identität)
  Beispiel: $\{1, 2, a\} = \{1, 1+1, a\}$, aber $\{1\} \neq \{\{1\}\}$.
  Semantische Heterogenität - versch. Abstraktionsgrad
- Nachweis der Gleichheit zweier Mengen:
  $\mathcal{A} = \mathcal{B} \Leftrightarrow$ (für alle $x \in \mathcal{A}$ gilt $x \in \mathcal{B}$) $\wedge$ (für alle $x \in \mathcal{B}$ gilt $x \in \mathcal{A}$) oder
  m.a.W. $\mathcal{A} = \mathcal{B} \Leftrightarrow \mathcal{A} \subseteq \mathcal{B} \wedge \mathcal{B} \subseteq \mathcal{A}$.
  Nachweis der Gleichheit zweier Mengen $\mathcal{A}, \mathcal{B}$ wird geführ durch Verifizierung der beiden Teilmengenbeziehungen:
  $\overline{(\mathcal{R} \cup \mathcal{S})} \Leftrightarrow \overline{\mathcal{R}} \cap \overline{\mathcal{S}}$.



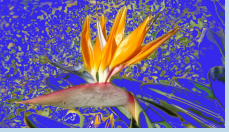
Beispiele: Natürliche Zahlen

- Natürliche Zahlen:
  Zwei natürliche Zahlen sind gleich g.d.w. sie durch gleichviele Nachfolgerbildungen aus der ersten natürlichen Zahl 1 hervorgehen
  Beispiel:
  Es gelten $1' =^{p.d.} 2$ und $1 + 1 =^{p.d.} 1'$.
  Unterschiedliche Kodierung
- Die Eigenschaft *gleichviele Nachfolgerbildungen* kann zu wahr oder falsch bewertet werden, ohne dass das Rechnen definiert ist!



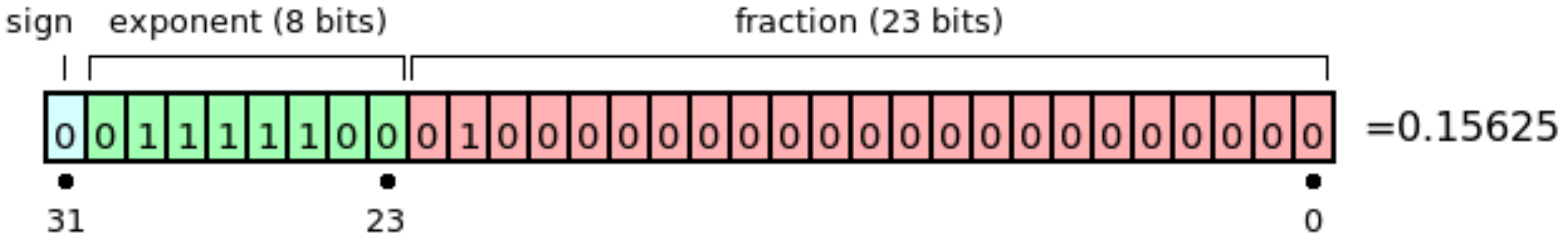
Beispiele: Brüche

- Gemeine Brüche:
  a, b, c, d reell, $bd \neq 0$: $\frac{a}{b} = \frac{c}{d} \leftrightarrow ad = bc$,
  d.h. Zurückführung auf ganze Zahlen.
- Reelle Zahlen (grob: Klassen äquivalenter Cauchy-Folgen):
  Zwei Cauchy-Folgen $\{a_i\}, \{b_i\}, i = 1, 2, ...\infty$ sind äquivalent g.d.w. die Mischfolge $\{a_i, b_i\}, i = 1, 2, ...\infty$ auch Cauchfolge ist.
- Die Darstellung der reellen Zahlen durch Dezimalbrüche beliebiger Stellenzahl ist eindeutig mit Ausnahmen in Verbindung mit 0 und 9.
  Beispiel: $1,0 = 0,\bar{9}$,
  denn es gilt:
  $1,0 = 1 + \sum_{i=1}^{\infty} \frac{0}{10^i} = 0 + \frac{9}{10} \times \sum_{i=0}^{\infty} \frac{1}{10^i}$



Relle Zahlen - Probleme

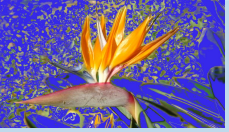
Probleme: (ANSI/IEEE Std 754-1985)
Beispiel : Fließkommazahl - 32-bit.



hat in der Mantisse nur endlich viele Stellen.
Diskrete Zahlendarstellung im Rechner
$\longrightarrow$ Konvertierungsfehler
Fehlerfortpflanzung bei Verarbeitung
$\longrightarrow$ Fehler bei vorverarbeiteten Daten.
von Fließkommazahlen: gleiche Bitfolge.

Realisierung des Tests auf Gleichheit

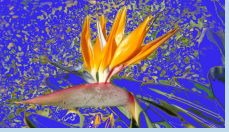
Literatur u. Bildquelle: Wikipedia, IEEE 754



Vektoren

- Definition: Sei $\mathcal{M}$ eine Menge, $n \in N$.
  Dann heißt das n-Tupel $\vec{a} = \{a_i \| a_i \in \mathcal{M}, \rangle = \infty, \in, ..., \backslash\}$ **Vektor** der Dimension n über $\mathcal{M}$. (Beachte Überschied zur math. Definition.)

  Gleichheit: $\vec{a}, \vec{b}$ zwei Vektoren über $\mathcal{M}$ mit Dim. n.
  $\vec{a} = \vec{b} \Leftrightarrow a_i = b_i, i = 1, 2, ..., n$
- Unterschird zu Mengen: Reihenfolge, Homogenität
- Anwendungen in der Informatik:
  Felder (array);
  Listen, Zeichenketten gleicher Länge (wenn die Länge unterschiedlich: nicht vergleichbar, gilt als ungleich).
- Allgemeine Tupel: Komponentenweise andere Grundmengen.
- Mengen, Vektoren, Tupel Grunddatenstrukturen in der Informatik.



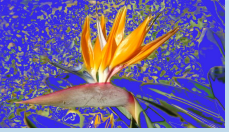
Gleichheit - Zeichenketten

- Vergleiche beruhen (meist) auf bin. Codierung des Textes im gewählten Zeichensatz, ASCII kleinster gemeinsamer Inhalt in vielen europ. Sprachen (trad. Codierung). Problem: nationale Sonderzeichen unterschiedl. codiert.
- Spracherkennung (charakt. Häufigkeiten von Zeichen, Bi- und Trigrammen (später mehr).)
  $\rightarrow$ Codierungserkennung (Sonderzeichen der Sprache in vermuteter Codierung darstellen)
  $\rightarrow$ Überführung in gemeinsame Codierung - UNICODE

Sprachliche Besonderheiten - Umlaute, Betonung, Trema, ..., Ligaturen, ...

- Umlaute: Codierungsproblem, Sortierproblem (s.u.)
- Betonung: im UNICODE-Zeichensatz Unterscheidung: GR $iota$: ι, ί, ϊ, im Telefonbuch: Gleichbehandlung.
- Trema: auch im DE relevant - ($Haiti$ - $Ha\ddot{i}ti$, Asteroid, ...)
- Ligaturen (in Dt. ß $\leftarrow$ s+z), Hindi

DBVS: Verhalten kann festgelegt werden.

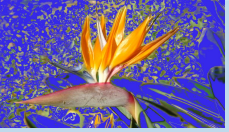


Zeichenketten - lexikographische Sortierung

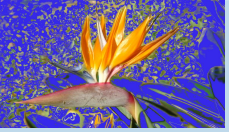
- Zeichenketten - lexikographischer Vergleich
  Algorithmus: *selbst nachtragen*
- Unterschiedliche Einsortierung der Umlaute:
  - ignorieren: ä wie a (DE: Lexika) - DIN 5007-1
  - in DE: ä wie ae, DIN 5007-2: *Kuciak - Kudies - Küchler* (Telefonbuch)
  - in OE: ä nach az (österr. Telefonbuch)
- Beispiel

DIN 5007-1	DIN 5007-2	
Lexika	Telefonb.	Österr.Sort
Göbel	Göbel	Goethe
Goethe	Goethe	Goldmann
Goldmann	Götz	Göbel
Götz	Goldmann	Götz

Quelle: *Wikipedia.Stichwort: Alphabetische Sortierung*.



Welt - Modell - RID

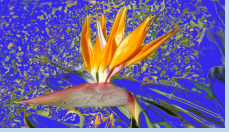


Einschub: Modellierung

Anwendungsmodellierung der Informatik:

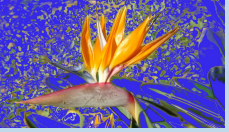
	Welt		Wissenschaft		RID
			Ontologien Theorien	$\Rightarrow$	Metadaten, Daten Datenstrukturen
	Reale Dinge	$\Rightarrow\Leftrightarrow$	ideale Objekte	$\Rightarrow\Leftrightarrow$	Daten
	Interaktionen	$\Rightarrow\Leftrightarrow$	Berechnungen	$\Rightarrow\Leftrightarrow$	Algorithmen
theoret.		Isom.		Isom.	
prakt.		?		?	

Modellbildung abstrahiert von (im Moment scheinbar) unwesentlichen Eigenschaften.
Interesse der Informatik: Wissenschaft $\Leftrightarrow$ RID.
Arbeit im Modell bzw. RID - Interpretation in Welt bzw im Modell und dann in Welt. Bei der Interpretation der Ergebnisse braucht der Informatiker den Fachwissenschaftler.

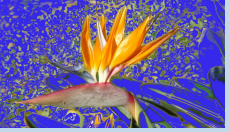


Komplexe Objekte

- I.A. individuelle Definition des Gleichheitsbegriffs,
- Meist auf Grundvergleiche zurückführbar
  In RID ergibt sich Gleicheitsdefinition aus Datenstruktur in Verbindung mit semantikbedingtem Gleichheitsbegriff für Elementarbestandteile (i.A. konjunktiv verknüpft)
- Student(NachName, MatrikelNr, Universität, Imma-Jahr, Vorname, Geb.Datum, ...)
- Gleichheitsdefinition semantisch verwandt mit Festlegung eines Primärschlüssels.



Geom. Aehnlichkeit

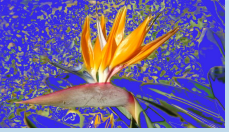


Ähnlichkeit

Motivation: Gleichheit oft zu restriktiv.

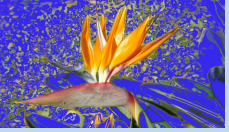
- Zahlen: Messfehler, Fehlerfortpflanzung in num. Algorithmen, instabile Algorithmen, Fehler durch Abschreiben, Ablesen, ...
- Suche in Bildern
- Einige Fehler erkennbar oder korrigierbar (fehlererkennende , -korrigierende Kodierung)- Redundanz,
  Vergleich mit theoretischen Werten.

Definition *Ähnlichkeit* verschieden möglich:
Semantische Stufe: i.a. bessere Ergebnisse
Syntaktische Stufe: formal, funktioniert auch ohne semantisches Wissen, i.a. schwächere Ergebisse.



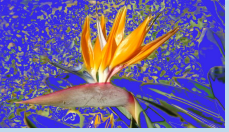
Anwendungsszenarien

- Gesichtserkennung: markante Punkte und Verhältnisse der Entfernungen zwischen diesen, ggf. unter Beachtung von Projektionen.
- Bilder: Farbhistogramme, ...
- Klänge: Spectrum (Fourier-Analyse), ggf. zeitlicher Verlauf.
- Zeichenketten:(flexible Hilfsstruktur der Informatik)
  Soundex - Vergleich
  n-Gramm-Analyse
  edit-distance (Levenstein - Distance)



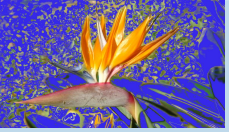
Ähnlichkeit - Definition

- Ursprünglich: Geometrischer Begriff für ebene Dreiecke.
  Definition 1: Zwei ebene Dreiecke sind ähnlich g.d.w. einander entsprechende Stücke proportional sind. *
  Definition 2: Zwei ebene Dreiecke sind ähnlich g.d.w. sie in zwei Winkeln übereinstimmen.*
  Verallgemeinerung: ebene, durch Polygonzüge berandete Objekte (Triangulierung).
- Was bedeutet *Übereinstimmung in zwei Winkeln* α, β ?
  Vereinbarung: α, β befinden sich an den Ecken A, B, wenn das Dreieck in der Reihenfolge ABCA umlaufen wird, befinde es sich links vom Rand.
  $[\ \alpha, \beta\ ]$ ist eine Liste: Reihenfolge relevant:
  Ähnlichkeit abstrahiert von Skalierung, Drehung, Verschiebung $\{\ \alpha, \beta\ \}$ ist eine Menge: Reihenfolge nicht relevant:
  Ähnlichkeit abstrahiert von Skalierung, Drehung, Verschiebung und Spiegelung.



Bemerkungen

- Definition 2 zeigt, dass die Ähnlichkeit eine Äquivalenzrelation ist. Denn:
  Ähnlichkeit $\leftrightarrow$ Gleichheit der einander entsprechenden Winkel.
  Die Gleichheit induziert die Ähnlichkeit. Der zugrunde liegende Gleichheitsbegriff ist eine Äquivalenzrelation
- Allgemeine Beschreibung:
  Gegeben zwei Mengen $\mathcal{A}, \mathcal{B}$ von Objekten.
  Auf $\mathcal{B}$ gibt es eine Gleichheitsbeziehung $=$
  $\varphi : \mathcal{A} \to \mathcal{B}$ eine Abbildung von $\mathcal{A}$ in $\mathcal{B}$.
  dann wird durch φ eine Ähnlichkeitsbeziehung $\sim$ in $\mathcal{A}$ iduziert:
  $x \sim y \leftrightarrow \varphi(x) = \varphi(y)$, $x, y \in \mathcal{A}$.



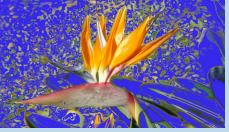
Beispiel: Soundex-Algorithmus

- Entstehung
  Robert C. Russel - 2.April 1918 - Patent Nr. 1 261 167
- Algorithmus
  - Code $\Leftarrow$ 1.Buchstabe + 3 Ziffern
  - Streiche ab dem 2.Buchstaben alle
    a, e, i, o, u, h, w, y
    und füge 3 Ziffern nach Tabelle hinzu.
  - Tabelle:

1	$\Leftarrow$	b,f,p,v	labial	Beachte folgende
2	$\Leftarrow$	c, g, j, k,	heterogen: frikativ; plosiv,	
		q, s, x, z	velar	
3	$\Leftarrow$	d, t	plosiv, dental/alveolare	
4	$\Leftarrow$	l	lateral	
5	$\Leftarrow$	m, n	nasal	
6	$\Leftarrow$	r	Vibranten	

  Regeln:

  1. 2 aufeinanderfolgende Buchstaben mit demselben Kode
     $\rightarrow$ nur 1x

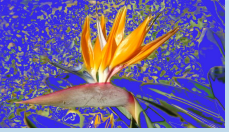


Soundex-Algorithmus (2)

- Beispiele:

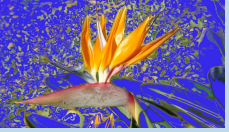
Name	Code	Bemerkung
Miller	M460	Auffüllen (Regel 1), Doppel-l (Regel 2)
Peterson	P362	3 verschiedene Konsonanten
Peters	P362	
Moskovitz	M232	
Moskowitz	M213	Fehlkodierung nichtenglischer Namen
ψ, ξ	?	

- Definition: Phonetische Ähnlichkeit nach dem Soundex-Verfahren: Zwei Zeichenkette über dem ASCII-Zeichensatz sind phonetisch nach den Soundex-Verfahren ähnlich, wenn sie die gleiche Soundex-Kodierung besitzen. *



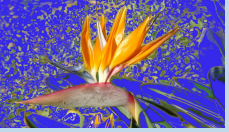
Idee hinter Soundex

- Begriffe *Phone, Phoneme* einer Sprache
- Phonemfeststellung: zwei Worte einer Sprache, die sich nur in einem Phon unterscheiden.
  Beispiel: ehren - lehren ← das Phon (l) ist ein Phonem.
- Phoneme können in Klassen eingeteilt werden nach der Art und dem Ort ihrer Entstehung beim Sprechen.
- Phoneme werden in Schriftsprache durch Grapheme dargestellt - M:N - Abbildung.
- Aussprache eines Graphems kann kontextabhängig sein.
  Internationales Phonet. Alphabet (z.B. 28 a-Varianten)
- Phoneme (und zugeordnete Grapheme variieren mit Sprache), d.h. phonetische Suche muss sprachspezifisch sein.



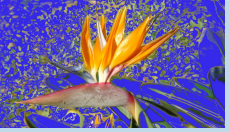
Entwicklungsschritte

- Analyse des Phonembestandes
- Klasseneinteilung der Phoneme, sinnvolle Vereinfachung des Klassensystems. Jede Reduzierung der Klassenzahl macht die Suche unschärfer, aber fehlertoleranter.
- Voraussetzung: Sprache wird durch Folgen von Graphemen beschrieben. Beschreibung der Zuordnung von Graphemen und Graphemfolgen zu Phonemen und Phonemfolgen. Zwei Zeichenketten sind phonetisch ähnlich, wenn sie auf die gleichen Ketten von Phonemklassen abgebildet werden.
  M:N Abbildung
  Konstruktion eines Automaten, der die Umwandlung vornimmt.



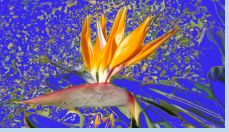
Besonderheiten des Neugriechischen

- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber γγ $\mapsto$ ng.
  Weitere Kombinationen: αυ, ευ, ου $\mapsto$ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, ξ $\mapsto$ Phonemfolge ps, ks
- Mehrere Schreibungen ein Phonem:
  Phonem **i**: ι, η, υ, ει, οι, υι; Phonem **e**: ε, αι; Phonem **o**: ο, ω
- Wechselwirkung der Aussprache mit Betonung, Silbenanfang:
  βάκιλοι - v'akili
  φιλοί - fil'i
  ρολόι - rol'oi
- Unsicherheit bei μπ, ντ und γκ,
- Für die Grapheme δ und θ gibt es im (Hoch-) Deutschen keine adäquaten Phoneme.
- Lautverschiebung: z.B. κτ $\mapsto$ χτ, σθ $\mapsto$ στ.



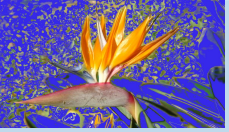
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.
  2. Nachbilden der Phonetik (Lautfolgen, Zeichenfolgen),
     Reduktion (stimmhaft $\mapsto$ stimmlos), **i**-Varianten:
     ca. 60 weitere Regeln,
     Transcription auf Folgen von Phonemen aus ca. 15 Phonemklassen.
     Klassen durch Zeichen des lateinischen Alphabets beschrieben.
  3. (Soundex auf dem Ergebnis von (2) für Spezialfälle).
- Kein Weglassen der Vokale, keine Längenbegrenzung
- Kontextsensitive Regeln in jeder Stufe, Datenstrom.
  Muster: (Zeichen, nächst. Zeichen) $\mapsto$ (Aktion, Fortsetzungspunkt)
- Dreistufiger Ansatz auf andere Sprachen verallgemeinerungsfähig.
  Dritte Stufe dem Thema Wörterbuch nicht angemessen.



Beispiele aus den Regeln der 1. Sufe

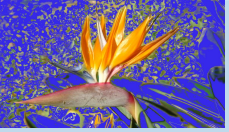
Z	nZ	Bedingung, Bemerkung	Phonem	Code	Schritt
α	ι		e	ε	+
α	ί		e	έ	+
α	υ	nicht bearbeiten	af, av	αυ	+
α	ύ	nicht bearbeiten	af, av	αύ	+
β	β		v	β	+
ε	ι		i	ι	+
η			i	ι	
ο	ι		i	ι	+
ο	ί		i	ί	+
ο	υ	nicht bearbeiten	u	ου	+
υ			i	ι	
ύ			i	ί	
ϋ			i	ϊ	
ΰ			i	ΐ	
ω			o	ο	



2. Stufe

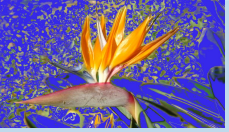
- - Buchstabenkombinationen auflösen: αυ, ευ, μπ, ντ
  - Kombinationen γγ, γκ, κκ bearbeiten
  - **i**-Allophone
  - Stimmhafte Konsonanten → stimmlose, z.B. **z** (ζ) → s
  - ψ, χ → ps, ks
  - Vokalverdopplungen entfernen
- Beispiele:

φιλοξενία	filoksenia
οινοποιείο	inopj'io → inopio



Phoneme für Konsonanten

	Artikulationsart				
	momentan		koninuierlich		
	Klusile		Frikative		Sonoranten
	stimmlos	stimmhaft	stimmlos	stimmhaft	
Labiale	p	b	f	v	m (Nasal)
Dentale	t	d	θ	δ	n (Nasal)
Alveolare	t^s	d^z	s	z	r (Tremulant)
Velare	k	g	(i/a)ch	γ (j)	l (Lateral)

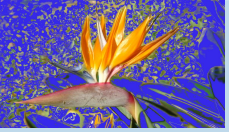


Phoneme für Konsonanten - Klassen

	Artikulationsart				
	momentan		koninuierlich		
	Klusile		Frikative		Sonoranten
	stimmlos	stimmhaft	stimmlos	stimmhaft	
Labiale	p p	b p	f f	v f	m (Nasal)
Dentale	t t	d t	θ t?s	δ s	n (Nasal)
Alveolare	t^s ts	d^z [2] ts	s s	z s	r (Tremulant)
Velare	k k	g k	(i/a)ch c,h	γ(j) [1] j→i	l (Lateral)

[1] γιορτάζω - jort'azo → jortaso iortaso

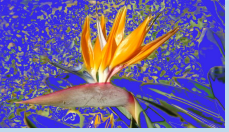
[2] τζατζίκι - dzadz'iki → tsatsiki



Konsonanten - Probleme

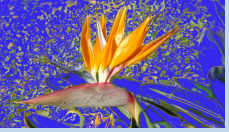
- Stimmhaftes b (μπ) und d (ντ)

λάμπα	-	l'amba	→	lampa
μπαμπάς	-	bab'as	→	papas



Konsonanten - Probleme

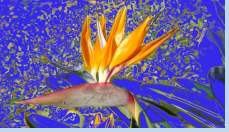
- Stimmhaftes b (μπ) und d (ντ)
  λάμπα - l'amba → lampa warum nicht lapa
  μπαμπάς - bab'as → papas warum nicht pampas
- Unbetontes Phonem **i** nach Konsonanten oder vor Vokal
  Verschiedene Möglichkeiten: i, j, j-ähnlich, (schwach) (i)ch, Ausfall
  ↦ Ursache vieler Regeln
- Wort - phonetische Beschreibung M:N
  μπαμπάς - bab'as → papas und pampas
  παπάς - pap'as → papas
  για - ja → ia
  γεια - ja → ia



Konsonanten - griechische Schreibung

	Artikulationsart				
	momentan		koninuierlich		
	Klusile		Frikative		Sonoranten
	stimmlos	stimmhaft	stimmlos	stimmhaft	
Labiale	π p p	μπ b p	φ f (α,ε,ι)υ f	β v (α,ε,ι)υ f	μ m μ(π) m(p)
Dentale	τ t t	ντ d t	θ t?s	δ s	ν n, ντ nt γ(γ,κ) n^k
Alveolare	τσ t^s ts	τζ d^z [2] ts	σ,ς s s	ζ z s	ρ r
Velare	κ k k	γ(α,ο,ου) g k	χ(ι,α) (i/a)ch c,h	γ(ε,ι)(j) j→i	λ l

16 Phonemklassen: **p, t, k, f, s, (i)c(h), (ac)h, m, n, r, l; a, e, i, o, u**

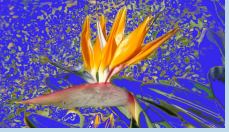


Realisierung

- Diplomarbeit abgeschlossen
- Prototyp:

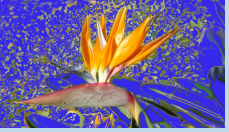
URL

http://teiresias.uni-leipzig.de



Probleme und Ausblick

- Fließkomma-Zahlen: s.o.
- Die prakt. Messung einer Größe ist fehlerbehaftet.
  Kann jetzt noch von Gleicheit gesprochen werden.
- Solche Probleme auch bei Ähnlichkeit:
  Bei trigonometrischen Vermessungen werden die Innenwinkel von Dreiecken ermittelt. Bis zu welchen Meßfehlern sollen die Dreiecke noch als ähnlich gelten?



Zusammenfassung

- Gleichheit / Ähnlichkeit sind Begriffe im Sinne eines Äquivalenzbegriffs.
- Test auf Gleichheit bei Integergrößen, Zeichenketten
- Unter dem Einfluß der Möglichkeit von geringen Abweichungen von wahren Werten (Meßfehler, ...) ist der praktische Nutzen eingeschränkt (Schwächere Kriterien nötig).
- Anwendungsbeispiel des math. Ähnlichkeitbegriffs: Phonetische Ähnlichkeit.
- Phonetische Suche muß sprachspezifisch erfolgen.