UNIVERSITÄT LEIPZIG

Institut für Informatik

Phonetische Suche in neugrichischen Texten

Dr. Dieter Sosna

Report 1 (2007)

ISBN 1430-3701

Phonetische Suche in neugriechischen Texten[1]

Dieter Sosna

Kurzfassung: Es wird ein Verfahren PHONI beschrieben, welches ähnlich dem Soundex-Algorithmus ist und die Besonderheiten des Neugriechischen bei der phonetischen Suche beachtet. Die phonetische Umwandlung eines Wortes erfolgt in bis zu drei Schritten, die Ergebnisse jedes Schrittes können zur Suche mit abnehmender Selektivität verwendet werden.
Der Algorithmus läßt sich leicht an andere Sprachen anpassen. Er wurde im Online-Wrterbuch „Teiresias“(Deutsch-Neugriechisch) zum Test implementiert [WB] [2].

Summary: The paper describes algorithms, we called them PHONI, which are similiar to the Soundex-algorithm and allow a phonetic search in texts written in the newgreek language. It was realized in the online dictionary „Teiresias“[WB].

Περίληψη: Ονομάζεται αλγόριθμος η ΦΩΝΗ της φωνητικής αναζήτησης, που είναι ίδιο με Σουνδεξ και παρατηρεί τις ιδιαιτερότητες της Νέας Ελληνυκής γλώσσας. Η φωνετική μετατροπή μιας λέχης ακολουθεί τρεις φάσεις και τα αποτελέσματα καθεμιάς μπορούν να χρησιμοποιηθούν προς αναζήτηζη με αφαιρετική. Η φωνητική αναζήτηζη πραγματοποίθηκε στο ηλεκτρονικό λεξικό Τειρεσίας [WB].
Ο αλγόριθμος είναι εύκολα προσαρμοστικός και σε άλλες γλώσσες.

[1] Version vom 30. Januar 2007
[2] URL: *http : //teiresias.uni − leipzig.de*

Inhaltsverzeichnis

1 Problemstellung

Im Neugriechischen gibt es sechs Grapheme, die das Phonem **i** bezeichnen und sich in der Aussprache grundsätzlich nicht unterscheiden (s.u.). Sprachwissenschaftler sind der Meinung, daß aus altgriechischen Wurzeln vielfach ableitbar ist, welches Graphem das richtige sei, aber für Lernende eine phonetische Suche in einem Wörterbuch durchaus Sinn mache. Das Problem des Phonem **i** ist besonders augenfällig, ähnliche Probleme bestehen aber auch für die Phoneme **o** und **e**. Bei den Konsonanten fällt die Unterscheidung der Schreibung zwischen stimmhaften und stimmlosen schwer, wenn durch die Umgebung, in der sie gerade stehen, eine Assimilation bewirkt wird. Dies spricht für eine phonetische Suche, die einem derzeit bearbeiteten Diplom-Thema „Wörterbuch Neugriechisch-Deutsch“ eine passende Realisierung finden soll.
Ein Test der Suchmaschine „google“ ergab, daß diese keine phonetische Suche beherrscht. Zwar vermag sie bei Rechtschreibfehlern ein ähnliches Wort zu erkennen, die Ähnlichkeitssuche beruht wahrscheinlich auf der Zahl der Buchstabensubstitutionen. Substituiert man beispielsweise in dem Wort βιβλίο (= Buch) die ι durch ein anderes Graphem οι oder υ zum gleichen Phonem ([3]), greift die Suche ins Leere, auch wenn man die Suchmaschine auf griechische Seiten beschränkt. Google erkennt offensichtlich nicht, daß es sich bei dem Suchwort um ein gleichklingendes Wort handelt, also evt. „nur“ ein orthographischer Fehler vorliegt.
Im Internet berichten die Autoren des „Cypros Law Reports“(siehe [Ano]) darüber, daß der Soundex-Algorithmus die Belange einer phonetischen Suche im Neugriechischen nicht trifft und es keine Arbeit zu diesem Problem gibt. Deshalb benutzen sie ein eigenes, intuitiv erstelltes Verfahren, welches sie leider auch nicht publizierten.

2 Der Soundex-Algorithmus als Lösungsansatz

Der Soundex-Algorithmus realisiert eine Lösung des Problems der phonetischen Suche. Der originale Algorithmus war auf die englische Sprache ausgerichtet und wurde von Robert C. Russel am 2. April 1919 in den USA patentieren lassen. Russel teilte die Buchstaben des Alphabets in verschiedene Kategorien, denen dann jeweils ein Buchstabe oder eine Ziffer zugeordnet

[3] vergleiche Seite 10

wurde. Durch Regeln wird die entstehende Zeichenkette bei der Zuordnung vereinfacht. Es werden z.B. Vokale im Wortinneren teilweise und Folgen von Konsonanten in einer Kategorie bis auf einen Vertreter eliminiert ([Mok97]). Die Idee wurde später in den USA weiterentwickelt und wird auch derzeit von Verwaltungsdienstellen angewendet.
Durch die Orientierung an der Phonetik des Englischen ist Verfahren in seiner ursprünglichen Form nur bedingt in anderen Sprachen anwendbar. Mit dem Daitch-Mokotoff-Soundex-System wurde eine Verallgemeinerung der Methode auf osteuropäische Sprachen (z.B. Russisch) und auf die deutsche Sprache 1986 publiziert (siehe auch [Mok97]). Als Anpassung an die sprachlichen Eigenheiten wurde die Zahl der Kategorien erhöht und eine Kodierung eines Buchstabens durch eine Folge von Kategorien, die dem Klangbild des Buchstabens folgt, genutzt. Mit anderen Worten: Grapheme, die die Folge von mehreren Phonemen repräsentieren, werden durch eine Folge von Kategorien dargestellt. Weiter werden Kombinationen von zwei Graphemen, die gemeinsam ein Phonem beschreiben, bei der Umkodierung gemeinsam einer Kategorie zugeordnet. Ist ein Wort zu lang, gehen nur die ersten Zeichen in die Bildung des Soundex-Wertes ein, er besteht bei dem Verfahren von Daitch-Mokotoff aus 6 Ziffern (im Originalverfahren aus 1 Buchstabe, 3 Ziffern).
Das Neu-Griechische gehört zu den Sprachen, die mit einem Graphemen jeweils eine Folge von Phonemen kodieren können, zum Beispiel mit ξ die Folge ks und mit ψ ps. Folglich ergeben sich ähnliche Probleme wie in den osteuropäischen Sprachen.

3 Algorithmen für das Neugriechische

Von Soundexverfahren kann die Idee übernommen werden, dass nach bestimmten Regeln jedes Wort einer Klasse „ähnlich klingender“ Wörter zugeordnet wird. Auch ein zu suchendes Wort wird nach diesen Regeln in die Menge dieser Klassen abbildet. Das Ergebnis der phonetischen Suche besteht aus allen Worten, die in derselben Klasse wie das gesuchte Wort liegen. Nur müssen die Regeln zur Klassenbildung den phonetischen Eigenheiten des Neu-Griechischen angepasst werden. In der vorliegenden Arbeit sollen zwei Regelmengen vorgestellt und ihre Entstehung begründet werden.

3.1 Allgemeines

Die griechische Schriftsprache basiert auf Folgen von Buchstaben, die (von links nach rechts) sequentiell gelesen werden. Der Autor von [Rug01] schreibt dort unter der Überschrift „Vom Buchstaben zum Phonem“:
(Zitat:) Obwohl die griechische Orthographie historisch ist ..., kann man fast immer die Aussprache eines geschriebenen Wortes erschließen. Unsicherheit besteht nur bei den Kombinationen μπ, ντ und γκ, bei denen die korrekte Aussprache sprachhistorische Kenntnisse voraussetzt (und folglich nur von einer Minderheit aller Griechischsprechenden geleistet werden kann). Ein ebenfalls unbedeutender Unsicherheitsfaktor besteht bei den Graphemen, die dem Phonem **i** entsprechen, nämlich wenn diese unbetont vor einem Vokal steht: ...(Ende des Zitats)

Die erste Feststellung dieses Zitats soll noch dahingehend präzisiert werden, dass der Lautwert eines Graphems entweder von diesem allein, d.h. unabhängig vom Kontext der ihn umgebenden Grapheme, oder von ihm und dem nachfolgenden Graphem oder aber von ihm und den beiden folgenden oder von ihm, dem Nachfolger und dem Vorgänger bestimmt werden. Dies bedeutet die Auswertung von Dreiergruppen.

Damit kann die Grundstruktur der Transformation eines Wortes in seine phonetische Äquivalenzklasse für das Neugriechische beschrieben werden:
Ein Wort der griechischen Sprache wird durch eine Folge von Graphemen (eine Zeichenkette) beschrieben. Diese Zeichenkette wird zeichenweise sequentiell umgewandelt. Bei dieser Umwandlung ist jeweils eine Gruppe von maximal drei aufeinanderfolgenden Zeichen gemeinsam zu betrachten([4]).
Von der Aussprache auf Folge der Grapheme zu schließen, ist unter alleiniger Kenntnis des Lautbildes nicht möglich, wie in Abschnitt 1 am Beispiel des **i**-Phonems beschrieben wurde. Der griechische Philologe S. Stampoulou (Universität Leipzig) merkt an, dass nahezu alle Tatsachen der Orthographie sprachwissenschaftlich und sprachhistorisch erklärbar sind ([5]). Da aber beim Nutzer eines Wörterbuch diese Kenntnisse im allgemeinen nicht vorhanden sind, wird die Implementierung einer phonetischen Suche motiviert.

3.2 Varianten der phonetischen Suche

Ein Verfahren zur phonetischen Suche enthält im Neugriechischen grundsätzlich Gestaltungsmöglichkeiten durch die Wahl der Umformungsregeln zur Bil-

[4] soweit dies am Ende der Zeichenkette noch möglich ist.
[5] Persönliche Gespräche des Autors mit S. Stampoulou

dung der Äquivalenzklassen. Offensichtlich ist eine Suche auf der Basis des Vergleiches der Äquivalenzklassen schwächer, d.h. weniger selektiv, als ein direkter Vergleich der Zeichenketten, liefert also im Vergleich zur exakten Übereinstimmung verfahrensbedingt falsch positive Ergebnisse. Grundsätzlich gilt: wird die Zahl der möglichen Äquivalenzklassen verringert, wird Präzision der Suche schlechter, liefert also mehr falsch positive Ergebnisse. Deshalb ist es sinnvoll, im Suchalgorithmus auf die Zahl der Klassen ähnlich klingender Worte Einfluss zu nehmen.

Quelle für Regeln sind die in allen Wörterbüchern und Grammatiken des Neugriechischen gegebenen Hinweise zur Aussprache. Besonders detailiert ist die Darstellung in [Rug01], die für die weitere Arbeit genutzt werden soll. Eine Analyse dieser Ausspracheregeln zeigt, dass sie sich in zwei Typen teilen lassen:

- Einmal sind es Regeln, die beschreiben, wie einem Graphem oder einer Gruppe von Graphemen ein Phonem oder eine Folge von Phonemen zugeordnet wird. Mit den Regeln dieses Typs kann u.a. ausgedrückt werden, wenn die Ausprache eines Graphems davon abhängt, welche Grapheme ihm in der Zeichenkette, die ein Wort oder eine Wortgruppe beschreibt, benachbart sind.

- Durch das Auftreten von Allophonen und Sandi ergeben sich Variationen in der Aussprache und damit das Problem, inwieweit diese Erscheinungen in der phonetischen Suche beachtet werden müssen, unbeachtet bleiben können oder auch unbeachtet bleiben müssen. Die sich so ergebenden Unsicherheiten in der Schreibung sollen gerade durch die Regeln eines zweiten Typs kompensiert werden. Dieses Vorgehen kommt dem Ansatz des Soundex-Algorithmus näher.

Die Entscheidungen, welche Regel implementiert werden sollen, hängen auch von der Zweckbestimmung der Suche ab:
Sollen nur quasi unhörbare Unterschiede in der Schreibung oder auch phonetische Varianten kompensiert werden? Wie stark müssen die phonetischen Abweichungen zweier Worte sein, damit sie als Unterschiede angesehen werden und zur Einordnung der Worte in unterschiedliche Äquivalenzklassen führen?
Da diese Fragen unterschiedlich beantwortet werden können, werden in der vorliegenden Arbeit drei Varianten vorgeschlagen, die aufeinander aufbauen.

1. *Fast-orthographische Suche:*
   Es werden nur die unhörbaren Unterschiede in den Schreibweisen ausgeglichen.

2. *Suche nach dem Klangbild:*
   Im Anschluß an an die Umformung des Suchstrings nach der ersten Variante wird jetzt das Klangbild erzeugt, vereinfacht und zur Grundlage der Suche gemacht.

3. *Soundex-artige Suche:*
   Auf das Ergebnis der vorigen Bearbeitung wird ein Soundex-Algorithmus angewendet.

Ein Vergleich dieser Vorschläge soll durch Test entsprechender Implementierungen erfolgen.

Eine dritte Gruppen von Regeln muß an dieser Stelle noch genannt werden, die in einem erweiterten Sinn der phonetischen Suche zugeordnet werden, die aber auch bei einer streng orthographischen Suche Beachtung finden sollten. In der Entwicklung des Neugriechischen sind Wandel für einige Konsonantverbindungen zu beobachten. :

Tabelle 1: **Lautwandel**

von		nach	von		nach
$\pi\tau$	$\mapsto$	$\varphi\tau$	$\varphi\vartheta$	$\mapsto$	$\varphi\tau$
$\varkappa\tau$	$\mapsto$	$\chi\tau$	$\chi\vartheta$	$\mapsto$	$\chi\tau$
$\sigma\chi$	$\mapsto$	$\sigma\varkappa$	$\nu\delta$	$\mapsto$	$\nu\tau$
$\sigma\vartheta$	$\mapsto$	$\sigma\tau$			

Für eine Reihe von Wörtern mit diesen Konsonantkombinationen sind beide Schreibungen als korrekt akzeptiert, bei anderen nur eine Form. Da sich keine allgemeine Regel für die Korrektheit angeben läßt, sind bei einer Suche i.A. beide möglichen Schreibungen in Betracht zu ziehen.
Der Lautwandel führt auf die folgenden Regeln, wobei unterstellt wird, dass ein Wort zeichenweise (von links nach rechts) bearbeitet wird. Wie die oben gegebene Tabelle vermuten läßt, müssen für diese Regeln nur Paare von Graphemen ausgewertet werden .

Tabelle 2: **Umkodierungen - Lautwandel**

Nr.	Z	nZ	Variante	Schritt
101	π	τ	φτ	+
102	ϰ	τ	χτ	+
103	σ	χ	σϰ	+
104	σ	ϑ	στ	+
105	φ	ϑ	φτ	+
106	χ	ϑ	χτ	+
107	ν	δ	ντ	+

([6])

3.3 Fast-orthographische Suche

Diese Form der Suche geht davon aus, dass bei korrekter Aussprache ein Zuhörer mit guten Vorkenntnissen fast immer aus dem Klangbild auf die Orthographie schließen kann. Mit den hier vorgestellten Regeln sollen zwei Gruppen von Ausnahmen bearbeitet werden:

- **Unsicherheiten, die durch Lautverschiebung entstehen** (s.o.)

- **phonetisch nicht wahrnehmbaren Unterschiede** möglicher Schreibungen:

  - doppelte Konsonanten bei ππ, ββ, μμ, ττ, νν, σσ, ρρ, ϰϰ, λλ
    werden vereinfacht, da sie phonetisch als einzelne Konsonanten wirken.

  - Grapheme zum Phonem **i**: ι, η, υ, ει, οι, υι
    werden auf das Graphem ι abgebildet.

  - Grapheme zum Phonem **e**: ε, αι
    werden auf das Graphem ε abgebildet.

[6]Die Bearbeitung soll grundsätzlich so erfolgen, dass ein Wort zeichenweise sequentiell umgeformt wird. Allgemein soll für die Bearbeitung eines Zeichens folgendes gelten: Bei fast allen Regeln zur Auswertung eines Zeichens Z muß das folgende Zeichen nZ mit beachtet werden. Wenn die Spalte nZ keinen Wert enthält, dann ist die Umformung unabhängig von dem Zeichen an dieser Position. Ein + in der Spalte „Schritt“ bestimmt, ob dieses Zeichen an nZ bei der Bearbeitung von Z mit transformiert wird, d.h. die weitere Bearbeitung bei dem auf nZ folgenden Zeichen fortsetzt. Andernfalls wird nur die Position Z bearbeitet und danach bei nZ fortgesetzt. Bei allen Regeln wird nicht zwischen Groß- und Kleinschreibung unterschieden.

- Grapheme zum Phonem **o**: ο, ω
  werden auf das Graphem ο abgebildet.

Unsicherheiten in der Aussprache von ντ, μπ, werden in dieser Variante nicht betrachtet. Gleiches gilt für Assimilationen, beispielsweise von stimmhaften Lauten nach stimmlos. Ferner wird vorausgesetzt, daß sowohl das Zeichen für die Betonung (Akut - οχεία) und für die Aussprache von fallenden Diphtonge (Trema - διαλυτικά) zur Auswertung zur Verfügung stehen. Andernfalls liefern die formulierten Regeln teilweise wenig sinnvolle Ergebnisse. Die genannten Zeichen werden sinnfällig auf das Ergebnis der Umformung übertragen, die Kodierung des Ergebnisses durch griechische Buchstaben unterstützt dies, so daß diese Zeichen auch nach Anwendung der Umformung weiter zur Auswertung zur Verfügung stehen.
Alle Regeln finden sich in folgenden Tabelle ([7]).
Die Anwendung der Regeln soll so erfolgen, wie bei Tabelle 3.2 beschrieben. Zu beachten ist, dass Vokale und Vokale mit Akut bzw. Trema im Sinne dieser Regeln unterschiedliche Zeichen sind.

Tabelle 3: **Umkodierungen - kleines Regelwerk**

Nr.	Z	nZ	Bedingung, Bemerkung	Phonem	Code	Schritt
201	α	ι		e	ε	+
202	α	ί		e	έ	+
203	α	υ	nicht bearbeiten	af, av	αυ	+
204	α	ύ	nicht bearbeiten	af, av	αύ	+
205	β	β		v	β	+
206	ε	ι		i	ι	+
207	ε	ί		i	ί	+
208	ε	υ	nicht bearbeiten	ef, ev	ευ	+
209	ε	ύ	nicht bearbeiten	ef, ev	εύ	+
210	η	ι		i	ι	+
211	η	υ	nicht bearbeiten	if, iv	ηυ	+
212	η	ύ	nicht bearbeiten	if, iv	ηύ	+
213	η		sonst	i	ι	

wird fortgesetzt

[7]Nicht in der Tabelle aufgeführte Grapheme werden nicht umcodiert. Einige Regeln sind technischer Natur, sie bewirken selbst keine Umkodierung, sorgen aber dafür, daß andere Regeln, die in der konkreten Situation nicht anwendbar sind, auch nicht zur Anwendung kommen. Das Beispiel αυ $\mapsto$ αυ illustriert dies. Ohne diese Regel würde α $\mapsto$ α und anschließend υ $\mapsto$ ι umcodiert, dies repräsentiert zusammen nicht die Lautverbindung **af** oder **av**. Die Bemerkung „nicht bearbeiten“ weist auf diese technischen Regeln hin.

Fortsetzung: Umkodierungen - kleines Regelwerk

Nr.	Z	nZ	Bedingung, Bemerkung	Phonem	Code	Schritt
214	ή			i	ί	
215	κ	κ		k	κ	+
216	λ	λ		l	λ	+
217	μ	μ		m	μ	+
218	ν	ν		n	ν	+
219	ο	ι		i	ι	+
220	ο	ί		i	ί	+
221	ο	υ	nicht bearbeiten	u	ου	+
222	ο	ύ	nicht bearbeiten	u	ού	+
223	π	π		p	π	+
224	ρ	ρ		r	ρ	+
225	σ	σ|ς		s	σ	+
226	τ	τ		t	τ	+
227	υ	ι		i	ι	+
228	υ	ί		i	ί	+
229	υ		sonst	i	ι	
230	ύ			i	ί	
231	ϋ			i	ΐ	
232	ΰ			i	ΐ́	
233	ω			o	ο	
234	ώ			o	ό	

Eine mögliche Implementierung wird im Pseudocode im Anhang auf Seite **??** vorgestellt. Da es wenige Regeln gibt, die eine Zuordnung zu einer bestimmten Ähnlichkeitsklasse erzwingen und die Anzahl der verschiedenen Klassen im wesentlichen nur durch die Phonemen **e, i** und **o** reduziert wurde, gibt es viele Ähnlichkeitsklassen, jede mit einer geringen Zahl von Elementen, folglich wird die Ergebnismenge einer Suche eine hohe Präzision aufweisen.

3.4 Suche nach dem Klangbild

Beschreibung einer Phonetischen Suche
Eine phonetische Suche kann wie folgt beschrieben werden: Als Vorbereitung des Verfahrens werden die Phoneme der Sprache ermittelt. Diese werden in Klassen eingeordnet, wobei eine Klasse ein Phonem oder mehrere Phoneme enthält. Danach werden Phoneme und Phonemkombinationen ermittelt, die sich bei der Aussprache von Graphemen oder Graphemkombinationen

ergeben. Diese Zuordnung kann mehrdeutig sein, wenn beispielsweise Unsicherheiten in der Aussprache bestehen. Die Klasseneinteilung der Phoneme wird nun auf die Grapheme übertragen, indem sie den Phonemklassen zuordnet werden, in denen ihre Phoneme liegen. Jeder Phonemklasse wird ein identifizierendes Zeichen zugeordnet, bei Soundex-Algorithmus werden hierfür die in den Worten des Englichen nicht vorkommenden Ziffern genutzt ([8]). Nach diesen Vorarbeiten kann jedem Wort der Sprache eine phonetische Beschreibung zugeordnet werden, indem zu der Folge seiner Grapheme eine Folge von Phonemklassen zugeordnet wird. Die Einzelschritte dieser Zuordnung ergeben sich aus einer sequentiellen Zuordnung von Graphemen und Graphemgruppen zu Phonemklassen. Bei der phonetischen Suche nach einem Wort wird dessen phonetische Beschreibung gebildet. Diese dient als Suchkriterium. Zum Ergebnis der Suche gehören genau die Worte des zu durchsuchenden Bestandes, deren phonetische Beschreibung mit dem Suchkriterium übereinstimmt.

Bestimmung der Phoneme und Bildung der Phonemklassen
Jede Sprache bestimmt ihre eigenen Phoneme, deshalb ist zu erwarten, dass eine phonetische Suche für eine Sprache auf einem Wortbestand einer anderen Sprache nicht korrekt arbeitet.

Tabelle 4: **Griechische Phoneme: Konsonanten**

Artikulationsstelle	Artikulationsart: momentan, Klusile, stimmlos	momentan, Klusile, stimmhaft	koninuierlich, Frikative, stimmlos	koninuierlich, Frikative, stimmhaft	koninuierlich, Sonoranten, stimmhaft
Labiale	**p**	**b**	**f**	**v**	**m** (Nasal)
Dentale	**t**	**d**	**ϑ**	**δ**	**n** (Nasal)
Alveolare	$\mathbf{t}^s$	$\mathbf{d}^z$	**s**	**z**	**r** (Tremulant)
Velare	**k**	**g**	**x, c**	**γ (j)**	**l** (Lateral)

Hierzu treten noch fünf vokalische Phoneme: **a, e, i (j), o, u** ([9]).

Die Ermittlung des Phonembestandes des Neugriechischen war nicht Gegenstand der hier vorgestellen Untersuchungen, vielmehr wurden Literaturquel-

[8] Allerdings erfährt dabei der erste Buchstabe eines Wortes eine mehr differenzierende Sonderbehandling

[9] Einklammerungen geben Allophone an, **x** bezeichnet den (a)ch-, **c** den (i)ch-Laut.

len ([Kar98] und [Rug01]) und die online-Enzyklopädie Wikipedia ([10]) ausgewertet. Dabei zeigten sich leichte Unterschiede in Detailfragen. Diese sind hier von untergeordnetem Interesse, da bei der oben beschriebenen Klassenbildung gerade leichte Unterscheide überdeckt werden sollen. Als Grundlage der weiteren Arbeit wurde [Rug01] gewählt. Das Ergebnis der Zuordnung zu den Phonemklassen stellt die folgende Tabelle dar:

Tabelle 5: **Griechische Phonemklassen: Konsonanten**

Artikulationsstelle	Artikulationsart: momentan Klusile stimmlos	momentan Klusile stimmhaft	koninuierlich Frikative stimmlos	koninuierlich Frikative stimmhaft	koninuierlich Sonoranten
Labiale	**p** *p*	**b** *p*	**f** *f*	**v** *f*	**m** (Nasal) *m*
Dentale	**t** *t*	**d** *t*	**ϑ** *t* und *s*	**δ** *s*	**n** (Nasal) *n*
Alveolare	$\mathbf{t}^s$ *ts*	$\mathbf{d}^z$ *ts*	**s** *s*	**z** *s*	**r** (Tremulant) *r*
Velare	**k** *k*	**g** *k*	**x, c** *x,s*	**γ (j)** ?	**l** (Lateral) *l*

Bei Benutzung dieser Einteilung ergeben sich die Klassen *p, t, k, s, f, m, n, r* und *l*. Im Gegensatz zum Soundexalgorithmus werden die fünf vokalischen Phoneme beibehalten und bilden jeweils eine weitere eigene Klasse: *a, e, i, o, u*.
Im Ergebnis dieser Transformationen wird insbesondere nicht mehr zwischen stimmhaften und stimmlosen Konsonanten unterschieden, es wird dort in jedem Falle Stimmlosigkeit unterstellt, d.h. sie werden auf *p*, *t* bzw. *k* abgebildet Dadurch werden auch die oben (Seite 7) genannten Unsicherheiten bei μπ, ντ und γχ umgangen. Auch die Besonderheiten der „gelehrten “ Worte (vgl. [Rug01], Seite 19ff) werden ignoriert.

Umkodierung in eine Folge von Phonemklassen
Es wird dabei vorausgesetzt, daß zuvor die Umformungen der Variante 1 (Abschnitt 1) erfolgt sind. Als für die weitere Bearbeitung wichtiges Resultat sind dadurch alle Grapheme, die ein i-Phonem beschreiben, auf das Zeichen ι mit seinen Varianten ί, ϊ, ΐ, d.h. unter Kennzeichnung von Betonung und Sprech-

[10]URL:*http://de.wikipedia.org/wiki/Neugriechisch*

silbentrennung abgebildet. Entsprechendes gilt für die **e**- und **o**-Phoneme.
Differenzierte Betrachtung werden für die Grapheme verschiedenen Vokal- und Konsonantkombinationen angestellt, wobei ein Wort, wie schon besprochen, sequentiell umgeformt wird. Einige der Bearbeitungsschritte sollen jedoch erläutert werden, die folgende Tabelle gibt eine Auflistung aller Regeln an. Die Grapheme ξ, ψ werden auf die Sequenz von Phonemklassen *ks* bzw. *ps* abgebildet. Das auf ξ, ψ folgende Graphem wird dann behandelt, als ob es auf ein **s** folgte.
Es erfolgt weiter die Bearbeitung des i-Phonems und der Kombinationen „Konsonant + i-Phonem “, auch von γ+i-Phonem und γ+e-Phonem. Weitere Regeln betreffen u.a. Kombinationen mit ϰ.
Akut und Trema sind nach dieser Überführung in Phonemklassen nicht mehr verfügbar.
Wie schon angemerkt, wird die tabellarische Darstellung der Aussprache in [Rug01] dem vorgeschlagenen Regelsystem zu grunde gelegt. Im Unterschied zu den detailierten Ausspracheregeln dort haben die Regeln hier das Ziel, eine Reihe von Unterschieden bewußt nicht zu beachten.
Für die Grapheme δ und ϑ gibt es im Hochdeutschen keine adäquaten Phoneme. Beide werden grob vereinfachend über das Phonem **s** umgesetzt: δ ↦ *s* und ϑ ↦ *s*. Mit Rücksicht auf deutsche Sprechgewohnheiten wird als Ausnahmeregelung ϑ zusätzlich auf τ abgebildet.

Um die Regeln übersichtlich zu präsentieren, werden Boolsche Funktionen vereinbart. Diesen wird die Position eines Zeichens der zu kodierenden Zeichenkette übergeben. Alle Funktionen werten das an dieser Position stehende Zeichen aus.

Tabelle 6: **Boolsche Funktionen**

Nr.	**Funktion**	:	**Bedingung für Rückgabe des Wertes** *wahr*
301	?ini	:	das Zeichen steht am Wortanfang
302	?vok	:	das Zeichen ist ein Vokal (α, ά, ε, έ, η, ή, ι, ί, ϊ,ΐ, ο, ό, υ, ύ, ϋ, ΰ, ω, ώ)
303	?kon	:	das Zeichen ist ein Konsonant (alle Zeichen des Alphabets außer den Vokalen)
304	?k-g	:	das Zeichen ist ein Konsonant außer γ und außer χ
305	?nV	:	das Zeichen folgt nach einem Vokal
306	?nK	:	das Zeichen folgt nach einem Konsonanten
307	?I	:	ab dieser Stelle ist das Phonem **i** kodiert und ist es unbetont?

wird fortgesetzt

Fortsetzung

Nr.	**Funktion**	:	**Bedingung für Rückgabe des Wertes** *wahr*
308	?'I	:	desgl., aber betont
309	?E	:	ab dieser Stelle ist das Phonem **e** kodiert und ist es unbetont?
310	?'E	:	ab dieser Stelle ist das Phonem **e** kodiert und ist es unbetont?
311	?AOU	:	ab dieser Stelle ist eines der Phoneme a, o oder u kodiert?
312	?'AOU	:	ab dieser Stelle ist eines der Phoneme a, o oder u kodiert und dieses betont?

Wie alle logischen Funktionen können die Funktionen in Tabelle 6 mit logischen Operatoren verknüpft werden (& Konjunktion,| Disjunktion, ! Negation).

Die eigentliche Umkodierung wird in Tabelle 7 ([11],[12]) beschrieben. Wenn in Spalte 6 eine Aufzählung erfolgt, sind das Alternativen, die dann jeweils beide weiter verfolgt werden.) Eine Erklärung bedarf die erste Regel. Sie betrifft nicht ein bestimmtes Graphem, sie wurde aufgenommen, um eine entsprechende Regel nicht bei jedem Konsonanten (außer γ und χ notieren zu müssen. Ihr Geltungsbereich tangiert den einiger anderer Regeln, was jedoch umproblematisch ist, da ihre konkrete Wirkung in solchen Fällen gleich der der anderen ist.

Tabelle 7: Umkodierungen - Klangbild

	1:	1.1	2:	3:	4: Bedingung/	5:	6:	7:
Nr.	Z	Bed.	nZ	nnZ	Bemerkung	Phonem	Code	Schritt
400	?k-g		ι	?vok		(j)	?	+
401	α		υ		μαυσωλείο	af, av	*af*	+
402	α		ύ		μαύρος	'af, 'av	*af*	+
403	α					a	*a*	

wird fortgesetzt

[11]In der Spalte 5 („dt. Phonem“) wird keine exakte Umschrift der Phonetik gegeben, diese Spalte beschreibt, wie etwa im Deutschen der Lautwert sich darstellt, in einigen Fällen wird explizit darauf hingewiesen (A), daß hier schon Allophone zusammengefaßt wurden. Alle diese Unterschiede werden bei der Kodierung nach Spalte 6 dann in ein und diesselbe Zielklasse abgebildet. Diese Spalte wird vorbereitend bereitgestellt, falls die Eingabe zur phonetischen Suche mit lateinischen (ASCII-) Zeichen erfolgen soll. Die Verwendung der Wörterbüchern üblichen Notation der Lautschriftwird erscheint wegen der dem Problem innewohnenden Vereinfachungen im Lautbild nicht notwendig.

[12]Die Bedeutung der Spalte 7 („Schritt“) wurde beim kleinen Regelwerk erläutert.

Fortsetzung: Klangbild

Nr.	1: Z	1.1	2: nZ	3: nnZ	4: Bedingung/ Bemerkung	5: dt. Phonem	6: Code	7: Schritt
404	ά					’a	*a*	
405	β				βήτα	v	*f*	
406	γ		γ		stets	(A)	*k*	+
407	γ		κ		stets	(A)	*k*	+
408	γ		χ	?AOU		ng, g	*k*	
409	γ		?I	?AOU	γιος	j	*i*	+
410	γ		?I	!?AOU	γυναίκα	j	*i*	
411	γ		?E ⋁ ?’E		γέρος	j	*i*	
412	γ		?AOU		γάλα	*x* [13] , g	*k*	
413	γ				sonst	g	*k*	
414	δ					delta	*s*	
415	ε		υ		λευκός	ef, ev	*ef*	+
416	ε		ύ			’ef, ’ev	*ef*	+
417	ε [14]				sonst	e	*e*	
418	έ					’e	*e*	
419	ζ					z	*s*	
420	η		υ			iv, if	*if*	+
421	η		ύ			’iv, ’if	*if*	+
	η				s. **i**-Phon.			
422	ϑ					theta	*s, t*	
	i [15]							
423	?’I				bet. i-Phon.	i	*i*	
424	?I				unbet.	i	*i*	
425	ϊ				**i** a. Wortanf.	i	*i*	
426	ΐ				dsgl., betont	i	*i*	
427	κ		?EI	?AOU	κιάλια	k(c), kj	*k*	+
428	κ				sonst	k (A)	*k*	
429	λ					l (A)	*l*	
430	μ		μ			m	*m*	+
431	μ	?ini	π		μπορώ	b	*p*	+
432	μ	?nV	π	?vok	τζάμπα			
					λάμπα	mb, b	*p*	+
433	μ				sonst	m	*m*	
434	ν		ν			n	*n*	+

wird fortgesetzt

[13] x ... ach-Laut, c ... ich-Laut

[14] wegen des ersten Schritts wurden alle **e**-Phon.e als ε kodiert

[15] wegen des ersten Schritts wurden alle **i**-Phoneme als ι kodiert, dennoch werden wegen der klaren Darstellung die Boolschen Funktionen genutzt.

Fortsetzung: Klangbild

Nr.	1: Z	1.1	2: nZ	3: nnZ	4: Bedingung/ Bemerkung	5: dt. Phonem	6: Code	7: Schritt
435	ν	?ini	τ		νταλμαδάκια	d	*t*	+
436	ν	?nV	τ	?vok	έντεκα			
					μαντείο	nd, d	*t*	+
437	ν				sonst	n (A)	*n*	
438	ξ				wie *σ*	ks	*ks*	
439	ο		ϊ		getrennt	o	*o*	
440	ο		υ			u	*u*	+
441	ο		ύ			'u	*u*	+
442	ο		ϋ		getrennt	o	*o*	
443	ο [16]				sonst	o	*o*	
444	ό					'o	*o*	
445	π					p	*p*	
446	ρ				νερό	r	*r*	
447	*σ*				σας	s	*s*	
448	τ				sonst	t	*t*	
	υ				s. **i**-Phon.			
449	φ					f	*f*	
450	χ		?AOU ⋁		χωριό			
			?kon		νύχτα	(ach) x	*k*	
451	χ		?EI		χίλιοι	(i)c(h)	*s*	
452	ψ				wie *σ*	ps	*ps*	
	ω				siehe ο			

Bei der Anwendung dieser Regeln können Zeichenfolgen entstehen, in denen ein Zeichen zweimal unmittelbar hintereinander auftritt. Bei diesen Zeichenverdoppelungen wird des zweite Auftreten des Zeichens wieder entfernt. Im Ergebnis gibt es nur noch die 14 Phonemklassen *a,e,i,o,u, f, k,l m, n, p, r, s, t.*

4 Implementierung

Das kleine und das große Regelwerk wurden im Rahmen elektronischen Wörterbuches „Teiresias“(Neugriechisch - Deutsch) getestet und unter dem Gesichtspunkt der praktischen Anwendbarkeit verfeinert.
Probleme ergaben sich dabei aus der die Kombination der phonetischen Suche mit den Möglichkeiten des Erkennens von Wortstämmen aus finiten Formen.

[16] nach dem ersten Schritt werden alle **o**-Phon.e hier erfaßt.

Da die Aussprache aus der Schreibung bestimmt werden kann, terminiert bei einen vollständigen Wort die Auswertung der Regeln stets. Die Formänderungen bei Konjugation bzw. Deklination implizieren auch Aussprachänderungen: (η μάχη - die Schlacht, Gen.Pl. μάχων). Nach dem Abtrennen der Flexionsendungen entsteht der Wortstamm μάχ, aus ihm kann die Aussprache des χ nicht mehr festgestellt werden. Dies äußert sich auf der technischen Ebene dadurch, dass ein Wort bei der Bearbeitung zu Ende ist und gleichzeitig noch Regeln, die sich schon in der berprfung befinden, zu ihrer vollstndigen Anwendung weitere Zeichen erfordern (hängende Regeln). Eine Lösung dieses Problems wurde darin gefunden, dass in einem solchen Fall die Gültigkeit der betroffenen Regeln stets als *wahr* angenommen wird, mit der vertretbaren Konsequenz einiger falsch positiver Suchergebnisse, was im Konzept dieser toleranten Suche akzeptabel ist.

5 Zusammenfassung

Es wird ein Vorschlag gegeben, wie die phonetische Suche in neugriechischen Texten realisiert werden kann, indem eine Abbildung eines Wortes auf eine Folge von Phonemklassen, die für diese Srache als relevant erkannt wurden, vorgenommen wird. Mit den drei vorgestellten Varianten der phonetischen Suche wird eine unterschiedliche Präzision der Suche erreicht. Die erste Stufe der phonetischen Umwandlung ist für den Nutzer interessant, der schon gute Grundkenntnisse des Zusammenhangs von Schreibung und Aussprache hat und in einer konkreten Detailfrage Unsicherheiten überbrücken will. Die zweite Sufe der Umwandlung hingegen versucht, eine Suche nach phonetischen Gesichtspunkten zu realisieren und abstrahiert aber dabei noch von feinen Unterschieden im Bereich der Allophone. Im Gegensatz zum originalen Soundex-Algorithmus werden die Vokale mit in die Suche einbezogen.
Diese beiden Verfahren erscheinen geeignet für die Verwendung zur phonetischen Suche in einem Wörterbuch. Sie wurden in der Beispielimplementierung [WB] realisiert und können dort getestet werden. Das dritte Verfahren, welches auf das Ergebnis der zweiten Variante den Soundex-Algorithmus anwendet, erscheint für den vorgesehenen Anwendungsbeich des Wörterbuchs als zu grob.
Der vorgestellte Ansatz, welcher die Idee des Soundex-Algorithmus flexibel verallgemeinert, kann auch auf andere Sprachen angewendet werden. Um diese Flexibilität zu demonstrieren, wurde in der o.g. Testimplementierung eine

Suche nach griechischen Worten hinzugefügt, bei der die Eingabe mit dem deutschen Zeichensatz und dem Hintergrundwissen einer deutschen Schreibung und Phonetik griechischer Wörter erfolgen kann.

Literatur

[Ano] Anonym. Cypros law reports.
URL: *http* : *//www.cylawreports.com/LRep.dll/HelpPg*.

[BSS03] *Search Algorithms*, March 2003. DRTC Workshop on Digital Libraries: Theory and Practice, DRTC, Bangalore
URL: *https* : *//drtc.isibang.ac.in/retrieve/26/E__Searchalgo__brijesh.pdf*.

[HC92] Henrich, Günther S. and Chrisomalli-Henrich, Kiriaki. *Langenscheidts Eurowörterbuch - Griechisch*. Langenscheidt Verl., 1992.

[Kar98] Dimitrios Karagiannakis, editor. *PONS - Kompaktwörterbuch für alle Fälle*. Klett-Verlag, 1997, 2. Nachdruck 1998. ISBN 3-12-518010-4.

[Kar03] Alexandros Karakos. *Journal of the American Society for Informationscience and Technology*, 54(11):1069 – 1074, 2003.

[Mok97] Gary Mokotoff. Soundexing and genealogy.
URL: *http* : *//www.avotaynu.com/soundex.html*, 1997.

[Rug01] Hans Ruge. *Grammatik des Neugriechischen*. Romio-sini Verlag, Köln, 2001. 3. neubearb. Auflage, 208 S.

[WB] Teiresias: Online-Wörterbuch Neugriechisch-Deutsch.
URL: *http* : *//teiresias.uni − leipzig.de/*.