Studienarbeit

Thema:

Lautschriftumsetzung und Worttrennung der chinesischen Schriftsprache

Studienarbeiter:	Jürgen Reichert
Betreuerin:	Dipl. Inform. Tanja Schultz
Fakultät:	Informatik
Institut:	für Logik, Komplexität und Deduktionssysteme Prof. Dr. A. Waibel

Inhaltsverzeichnis

1 Einleitung 4

1.1 Aufgabenstellung 4

1.2 Motivation 5

1.3 Gliederung 6

2 Die chinesische Sprache 7

1.4 Entwicklung der Zeichen 7

1.5 Morphologie der chinesischen Sprache 12

1.6 Das gesprochene Chinesisch 13

1.7 Weiter Besonderheiten der chinesischen Sprache 16

1.7.1 Die Schreibrichtung 16

1.7.2 Die Phonetik und Tonalität 17

1.7.3 Chinesisches Zahlensystem 20

1.7.4 Grammatik der chinesischen Sprache 20

1.7.5 Sprichwörter (成语) 20

1.7.6 Sonstiges 21

1.8 Chinesisch und Computer 22

1.8.1 Eingabemethoden 22

1.8.2 Zeichendarstellung und Fonts 23

1.8.3 Codierung 23

1.8.4 Vorteile der chinesischen Schrift 24

1.8.5 Chinesische Systeme 25

3 Relevantes Datenmaterial 26

1.9 Sammeln von Daten 26

1.10 Datenaufarbeitung 27

1.10.1 Hilfetools zur Textbearbeitung 27

1.10.2 SQL-Datenbank und SQL-Skripts 27

1.10.3 Wörterbuchverarbeitung 28

1.10.4 Speicherdatenbank 29

1.10.5 TCL-Tools 30

1.10.6 Handarbeit 33
1.11 Datensatzbeschreibungen 36
1.11.1 Wörterbuch 36
1.11.2 Wortwahrscheinlichkeit 37
1.11.3 Zeichenwahrscheinlichkeit 37
1.11.4 Pinyin-Wortzuordnungs-Unigramm 38
1.11.5 Pinyin-Wortzuordnungs-Bigramm 38

4 Analyse und Entwurf eines Romanisierungssystems 39

1.12 Brute-Force Ansatz 40
1.13 Mehrdeutigkeiten 40
1.14 Funktionsweise des Pinyinprogramm 41
1.15 Eigenschaften des Pinyinprogramm 43
1.16 Implementierung 44
1.16.1 Ablaufdiagramm des Pinyinprogramms 45
1.17 Test und Validierung 46
1.18 Performance 47

5 Zusammenfassung und Ausblick 50

Anhang A Schnittstellenbeschreibung der Speicherdatenbank 50

Anhang B Aufbau des Pinyinprogramms 52

Abbildungsverzeichnis 54

Literaturverzeichnis 55

1 Einleitung

Die vorliegende Arbeit „Worttrennung und Umsetzung der chinesischen Schrift in die Pinyin-Lautschrift“ ist eine Teilaufgabe des „GlobalPhone“-Projekts [16], welches zur Aufgabe hat, Daten von möglichst vielen Sprachen zu sammeln und für die computerbasierte Sprachverarbeitung aufzubereiten. Zum Zeitpunkt der Abfassung dieser Arbeit, werden die Sprachen Arabisch, Chinesisch, Japanisch, Koreanisch, Kroatisch, Portugiesisch, Russisch, Spanisch und Türkisch bearbeitet.
Weiterhin soll diese Arbeit für das Sprach-zu-Sprach-Übersetzungssystem Janus [17] einige Teilaspekte behandeln. Janus ist ein modulares, programmierbares System zur Spracherkennung, und Sprachübersetzung, welches aufgrund seiner Flexibilität die Eigenschaften der unterschiedlichsten Sprachen berücksichtigen kann.

1.1 Aufgabenstellung

Die chinesische Sprache ist in vielerlei Hinsicht von den uns geläufigen Sprachsystemen verschieden. Dies kommt im Volksmund durch Redewendungen wie „das kommt mir aber chinesisch vor“ oder „dies ist Fachchinesisch für mich“ zum Ausdruck. Und tatsächlich enthält sowohl die Schriftsprache als auch gesprochene Sprache Eigenschaften, die eine besondere Betrachtungsweise erfordern.
Diese Arbeit möchte zwei dieser Eigenschaften der chinesischen Sprache näher untersuchen und eine Möglichkeit finden, um die hauptsächlich für europäische Sprachfamilien entwikkelten Sprachverarbeitungsverfahren auch auf die chinesische Sprache anwenden zu können.
Die erste dieser Eigenschaften ist, daß es in der chinesischen Schrift keine Wortgrenzen gibt. Für eine computergerechte Verarbeitung müssen Wortgrenzen somit künstlich eingeführt werden und zwar so, daß die entstehenden Teilabschnitte Sinneinheiten repräsentieren. Dazu müssen die kleinsten Sinneinheiten eines Satzes gefunden werden, wobei es aufgrund grammatischer und semantischer Besonderheiten nicht trivial ist diese Trennung vorzunehmen. In vielen Fällen wird auch pragmatisches Wissen für die Trennung benötigt und oftmals ist eine von Hand durchgeführte Partitionierung mit einem gewissen subjektiven Charakter behaftet, da es nicht immer eindeutige Regeln gibt.
Die zweite in dieser Arbeit behandelte besondere Eigenschaft der chinesischen Sprache ist jene, daß es im Chinesischen keine direkte Bindung der Aussprache an die Schrift gibt, es also erforderlich wird, ein Verfahren zu entwickeln, das die zeichenbasierte Schriftsprache in eine Lautschrift überführt. Diese Aufgabe kann mit einer einfachen Abbildungstabelle bis zu einem gewissen Gerade erfüllt werden. Es gibt allerdings eine nicht zu unterschätzende Anzahl von Zeichen die mehrere, vom Kontext abhängige, Aussprachen haben. Somit sind, um eine befriedigende Qualität der Lautspracheumsetzung zu erhalten, weitere, komplexere Verfahren anzuwenden. In vielen Fällen ist auch der umgekehrte Fall, das Umsetzen der Lautschrift in die zeichenbasierte Schriftsprache, von Interesse. Verfahren für eine solche Umsetzung sind nicht einfach durch Umkehrung obiger Umsetzung ableitbar, da bei ersterer Umsetzung wesentliche Informationen verloren gehen, die nur durch den Kontext und Metawissen wiedergewonnen werden können. Diese Arbeit beschäftigt sich nur mit der Umsetzung der zeichenbasierten Schriftsprache in eine Lautschrift, wenngleich auch einige Grundlagen für die umgekehrte Richtung geschaffen wurden.

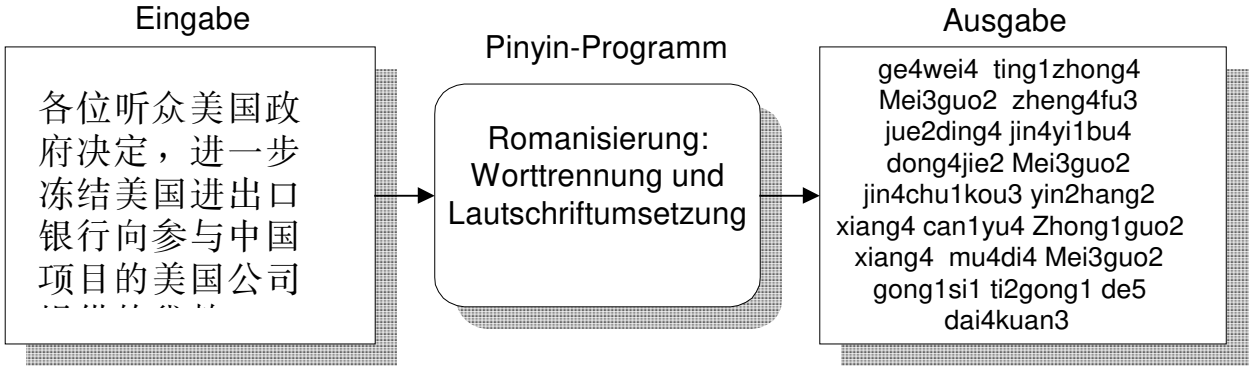


Abbildung 1.1: Worttrennung und Lautschriftumsetzung der chinesischen Schrift

1.2 Motivation

Seit Mitte dieses Jahrhunderts bis Ende der achtziger Jahre hat China eine Politik des Isolationismus geführt. Bis auf wenige Ereignisse, wie die Befreiung/Machtergreifung durch die Kommunisten 1948 und die Kulturrevolution (Säuberung von politisch Andersdenkenden) 1966-76 ist China fast völlig im zeitpolitischen Geschehen untergegangen. Nur wenige beschäftigten sich mit chinesischer Politik, Wirtschaft und Kultur. Dies hatte zur Folge, daß es seit der Öffnung Chinas seit ca. 1979 einen großer Nachholbedarf an Informationen zu befriedigen galt, denn mit zunehmender wirtschaftlicher und politischer Bedeutung Chinas wurde es immer wichtiger neue Beziehungen mit China aufzubauen.
Anlaß sich eingehend mit der chinesischen Sprache zu beschäftigen gibt es somit genug:

- Wirtschaftliche, politische und kulturelle Bedeutung Chinas
- Chinesisch sprechen ca. 20% der Weltbevölkerung, somit ist Chinesisch die meist gesprochene Sprache überhaupt.
- Chinesisch ist diejenige lebendige Sprache, deren Entwicklungsgeschichte am weitesten zurückreicht. Somit ist Chinesisch für die verschiedensten Sprachwissenschaften von besonderer Bedeutung.
- Viele literarische Werke liegen in keiner Übersetzung vor. Auch moderne Fachliteratur in Forschung und Wissenschaft, sowie Recherchedatenbanken, Statistiken und Gesetzesregelungen sind oftmals nur in der Landessprache erhältlich. Literatur zu den, dem westlichen Schach entfernt ähnelnden, Spielen Weiqi (Go) und Xiangqi (chin. Schach), die auch bei uns eine stark wachsende Anhängerschaft finden, ist bei uns nur schwer zu finden. Wobei besonders Weiqi, in der Forschung der künstlichen Intelligenz, dem westlichen Schach den Rang abgelaufen hat.
- Kalligraphie, die Fähigkeit des ausdrucksvollen Schreibens mit dem Pinsel, ist ein fester Bestandteil der chinesischen Kunst. Für ein tieferes Verständnis der chinesischen Kunst, ist somit ein gewisses Gefühl für chinesische Schrift erforderlich.
- Die Konzepte der chinesischen Sprache unterscheiden sich von denen unserer Sprache sehr, so daß Chinesisch ein guter Kandidat ist, um möglichst viele dieser Konzepte kennenzulernen. Insbesondere kann die Flexibilität des Janus-Systems mittels dieser Konzepte validiert werden.

1.3 Gliederung

Diese Arbeit beginnt im Kapitel 2 mit einer kurzen Einführung in die chinesische Sprache. Nach einer Übersicht über die Entwicklung der chinesischen Schrift, wird auf die Besonderheiten der chinesischen Sprache eingegangen. Danach wird die Problematik von Chinesisch auf dem Computer betrachtet.
Im Kapitel 3 wird auf die verschiedenen Arten der Beschaffung von Daten, zum Aufbau von Wörterbüchern, Statistiken und Grammatiken eingegangen. Daran anschließend wird beschrieben, wie die gesammelten Daten bearbeitet wurden. Es wurden hierzu etliche kleine Hilfe-Tools entwickelt, wovon einige stellvertretend vorgestellt werden. Der grösste (*) Anteil der Bearbeitung mußte allerdings von Hand durchgeführt werden, was einen nicht unerheblichen Zeitaufwand erforderte.
In Kapitel 4 wird die Analyse- und Entwurfsphase des in dieser Arbeit angefertigten Pinyinprogramms welches sowohl die Worttrennung als auch die Umsetzung in die Lautsprache Pinyin vornimmt, dargestellt. Einige Details der Implementierung des Pinyinprogramms werden beschrieben, sowie Test- und Validierungsverfahren vorgestellt.
Kapitel 5 faßt schließlich die Arbeit nochmals zusammen und gibt einen Ausblick auf weitere mögliche Entwicklungen und Einsatzgebiete.

(*) Das verwendete Textverarbeitungssystem kann keine ‘ß’ nach einem Umlaut darstellen, da diese Zeichenkombination als chinesisches Zeichen interpretiert würde. Deshalb wird in solchen Fällen ‘ß’ durch ‘ss’ ersetzt.

2 Die chinesische Sprache

Um sich mit der chinesischen Sprache zu beschäftigen, sollte man deren Geschichte und Entwicklung kennen. Auch um Entwurfsentscheidungen der vorliegenden Arbeit nachvollziehen zu können, ist eine gewisse Kenntnis der Entwicklung und Struktur der chinesischen Sprachen von Vorteil. Eine praktische Einführung in die chinesische Sprache geben die Bücher [8] und [9].
Obwohl die gesprochene Sprache vor der Schrift existierte, lässt sich über Lautgebung vor Entwicklung der Schrift nur sehr wenig sagen. Es ist ja gerade die Schrift, die eine Weitergabe von Informationen über viele Generationen hinweg vereinfacht.
In den frühen Epochen war die Entwicklung der Schrift mit der Entwicklung der chinesischen Zeichen nahezu identisch, somit bekommt der Entwicklung der Zeichen eine besondere Bedeutung.

2.1 Entwicklung der Zeichen

Die Triebfeder für die Entwicklung einer Schrift war, wie in vielen anderen Kulturen sicher auch, die Notwendigkeit sich angeeignetes Wissen weiterzugeben. Die chinesische Schrift begann mit bildlichen Aufzeichnungen. Die ersten bekannten Aufzeichnungen waren Abbildungen des täglichen Lebens (oft als Höhlenmalerei noch erhalten). Diese sind aber noch keine wirkliche Schrift, sondern nur eine Vorstufe, denn sie konnten weder durch gesprochene Sprache exakt wiedergegeben werden, noch konnten sie alle Bestandteile der gesprochenen Sprache wiedergeben.
Nach chinesischen Überlieferungen soll ca. 3000 Jahre vor Chr. erstmals der Beamte „Cang Jie“ am Hofe des Kaiser „Huang Di (Gelber Kaiser)“ damit begonnen haben, Symbolaufzeichnungen zu systematisieren und damit den Grundstein für ein Schriftsystem zu legen (s. [12] Modern Chinese Characters). Die Überlieferungen enthalten viele mythische Elemente, so soll Cang Jie z.B. 4 Augen besessen haben. Es ist somit nicht einfach festzustellen, inwieweit die Überlieferungen einen Wahrheitsgehalt haben, wie wohl die Existenz des Beamten „Cang Jie“ doch als sehr wahrscheinlich gilt.
Der älteste chinesische Symbolschriftfund ist wahrscheinlich der in Ton eingeritzte Fund im Dorf Banpo nahe bei Xi`an (s. [12]). Der Fund von 1972 wird auf nahezu 6000 Jahre datiert.

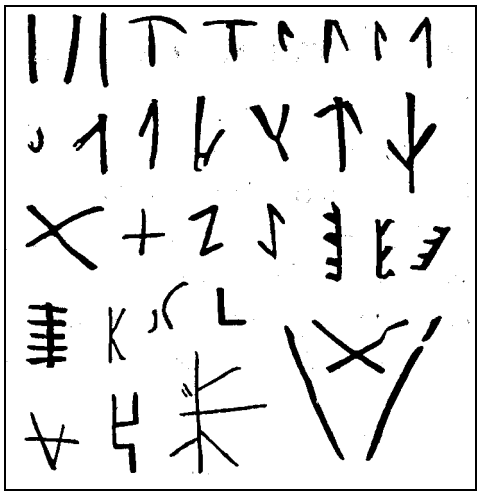


Abbildung 2.1: Erste Schriftfunde auf Tonscherben in Banpo

Diese eingeritzten Symbole haben alle noch eine sehr einfache Struktur, kennzeichnen aber den Beginn der Formung der ersten chinesischen Zeichen. Denn sie haben nicht mehr bloß einen dekorativen Charakter oder stellen eine reine Abbildung dar, sondern lassen eine erste Systematik erkennen. Andere Funde brachten Aufzeichnungen auf Knochen, Schildkrötenpanzern und Steinen zu tage.
Ein anderes im frühen China entwickeltes Aufzeichnungssystem benutzte Knoten verschiedener Grösse und Farbe an Seilen und Schnüren. Dieses eignete sich besonders zum Zählen, für verwaltungstechnische Angelegenheiten und als Kalender für Ereignisse und dergleichen (ein Vergleich zum berühmten Knoten im Taschentuch zwingt sich einem auf).

Für die weitere Entwicklung der chinesischen Schrift lassen sich 6 Kategorien der Zeichenkonstruktion verfolgen (nach Xu Shen):

1. Kategorie der reinen Piktogramme:

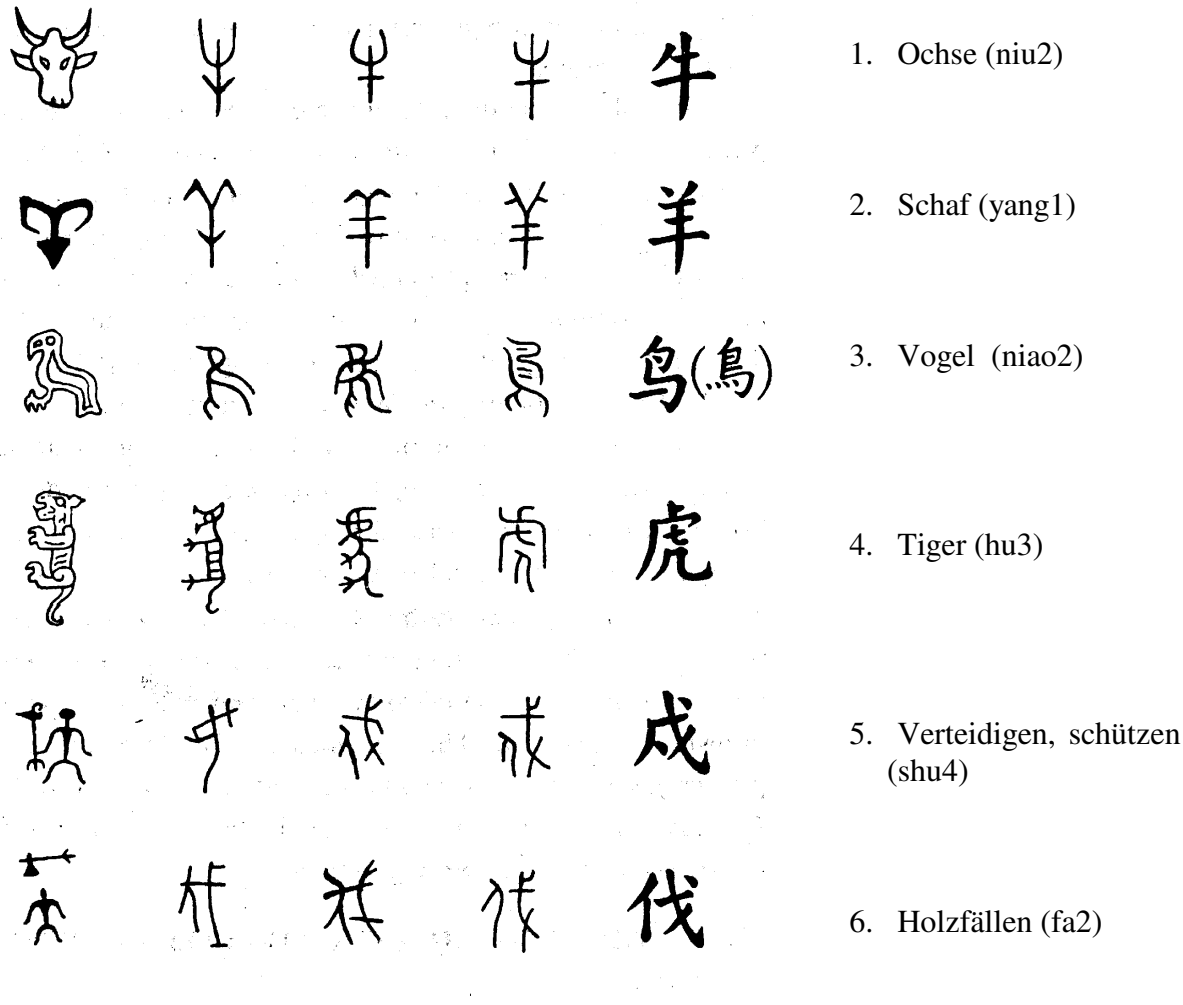


Abbildung 2.2: Entwicklung von der frühen Bilderschrift zu heutigen chinesischen Zeichen

Beim Piktogramm lässt, sich eine Linie von der rein bildlichen Darstellung (links) bis zum heutigen Zeichen (rechts) verfolgen. Oftmals ist es somit in dieser Kategorie möglich, von dem Aussehen des Zeichen auf seine Bedeutung zu schließen, wenn auch sicher ein gehö-

riges Maß von Phantasie und Erfahrung dafür nötig ist. Fast alle Zeichen dieser Kategorie stammen aus der ersten Periode der Sprachentwicklung, nur wenige kamen später hinzu (wie z.B. 伞 Regenschirm, yu3; 凸 konvex, tu1; 凹 konkav, ao1). Alle diese Zeichen haben gemeinsam, daß sie nicht weiter in untergeordnete Sinneinheiten unterteilt werden können, aber als Bildungselement für neue Zeichen dienen können. Nur ca. 4% der chinesischen Zeichen fallen in diese Kategorie.

2. Kategorie der indikatorischen Zeichen:
   Diese Zeichengruppe benutzt Symbole, um einen abstrakten Sachverhalt auszudrücken (z.B. 上 oben, 下 unten, 一 eins, 二 zwei, 三 drei). Oft werden diese Symbole mit Zeichen der ersten Kategorie überlagert, dabei werden die Zeichen, ähnlich wie zwei Folien, überlagert. (z.B. 木 Piktogramm für Baum + 一 Symbol für eins = 本 Wurzel oder 末 Spitze, Ende). Auch diese Zeichen können nicht weiter in untergeordnete Sinneinheiten unterteilt werden. Sie bilden die kleinste Gruppe mit unter 1% aller Zeichen.

3. Kategorie der assoziativen Zeichen:
   In dieser Gruppe werden Zeichen gebildet, indem Zeichen der vorherigen Gruppen kombiniert werden. Dabei werden die Zeichen verkleinert und neben/übereinander gruppiert, ohne daß Striche der einzelnen Zeichen sich dabei kreuzen. Die Beziehung zwischen den einzelnen Elementen ergibt hierbei dann die neue Bedeutung. Beispiele:
   Piktogramm Mensch 人 + Piktogramm Mensch 人 = 从 folgen (Mensch folgt Mensch);
   Piktogramm Baum 木 + Piktogramm Baum 木 + Piktogramm Baum 木 = 森 Wald
   Piktogramm Mensch 人 + Piktogramm Baum 木 = 休 ausruhen (Mensch lehnt an Baum)
   Diese Gruppe macht ca. 13 % der chinesischen Zeichen aus.

4. Kategorie der Piktogramm-Phonetischen Zeichen:
   Die dritte Kategorie erlaubt schon einen grösseren Wortschatz zu bilden. Aber oft führt aber die reine Assoziation zu sehr komplexen Strukturen, die aus vielen einzelnen Komponenten zusammengesetzt sind. Durch die Abkehr von der reinen Bedeutungsorientierung können sehr einfach neue Zeichen geschaffen werden. Bestehende Zeichen werden um eine phonetische Einheit ergänzt. Z.B. Piktogramm Mund 口 + phonetische Einheit 土 (tu3) ergibt 吐 (er)brechen (Bedeutungseinheit Mund, Aussprache tu3).
   Leider haben die meisten Zeichen im Laufe der Zeit ihre phonetische Bedeutung verloren, so daß Zeichen trotz gleicher phonetischer Einheit verschieden ausgesprochen werden. Genauso haben viele Zeichen eine gravierende Bedeutungswandlung erlebt, so daß die Bedeutungseinheit oft fehl am Platz ist. Der grösste Teil, nämlich ca. 80 % der chinesischen Zeichen fällt in diese Entwicklungskategorie.

5. Kategorie notativer Ähnlichkeit:
   Ähnliche Begriffe werden mit Zeichen, die mit verwandten Bildungsvorschriften generiert wurden, gebildet. Daraus erhält man bei ähnlichen Bedeutungseinheiten ähnlich aussehende Zeichen. Z.B. haben 顶 Spitze und µß Gipfel eine verwandte Bedeutung und enthalten beide auf der rechten Seite das Radikal 页(ca. 1-2 %).

6. Kategorie geliehener Zeichen:
   Für neue Wörter in der gesprochenen Sprache, für die es keine schriftliche Repräsentation gab, wurden Schriftzeichen von andern Wörter mit ähnlicher oder gleicher Aussprache verwendet. Eines oder beide der Zeichen wurden dann, um eine Unterscheidung zu ermöglichen, verändert.

Beispiel: 其 qi2 ihr,sein und 箕 ji1 Kehrrichtschaufel, wobei Kehrrichtschaufel das ältere Zeichen ist, und später von 其 nach 箕verändert wurde, um eine Unterscheidung zu ermöglichen. Dies ist das flexibelste, aber auch dasjenige Verfahren, das die ursprüngliche Bedeutung der Zeichen am wenigsten erhält. Wiederum ca. 1-2 % der Zeichen fallen in diese Kategorie.

Die sich so entwickelten chinesischen Zeichen haben in der vergangenen Zeit allerlei Formänderungen durchlebt. Man unterscheidet hauptsächlich 5 verschiedene Phasen [11], [12]:

1. Knochen- und Schildkrötenpanzer - Inschriften:
   Aus der späten Shang-Dynastie (1711-1066 v. Chr.) sind ca. 150000 Funde von Knocheninschriften bekannt und analysiert worden. 4500 verschiedene Zeichen sind in dieser Zeit verwendet worden, wovon von ca. 900 die Linie bis zu den heutigen Zeichen verfolgt werden kann. Die einzelnen Zeichen tauchten in vielen Varianten auf: verschiedene Strichanzahlen, Orientierung, Grösse, so daß eine sichere Klassifikation nicht immer erfolgen kann.

Abbildung 2.3: Orakelknocheninschrift

2. Bronzeinschriften:
   Während der Zhou-Dynastie (1066-256 v. Chr.) wurden die meisten Schriftfunde auf Bronzetafeln gefunden. Die Zeichenformen variierten im noch stärkeren Maße. Während nach und nach immer weniger reine Piktogramme verwendet wurden.

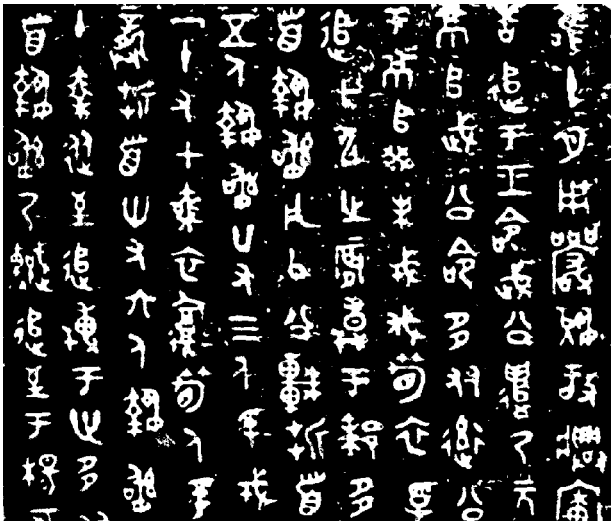


Abbildung 2.4: Bronzeinschrift der Zhou-Dynastie

3. Die Siegelschrift:
   Unter der Qin-Dynastie (221-206 v. Chr.) wurde die Schrift zur Siegelschrift vereinheitlicht. Diese Siegelschrift wurde zum Standard im gesamten Land. Bei den Zeichen herrschten nun lineare, symbolische Formen vor, die noch viele runde Elemente enthielten. Es wurde die einheitliche, quadratische Form der Zeichen vorgegeben.

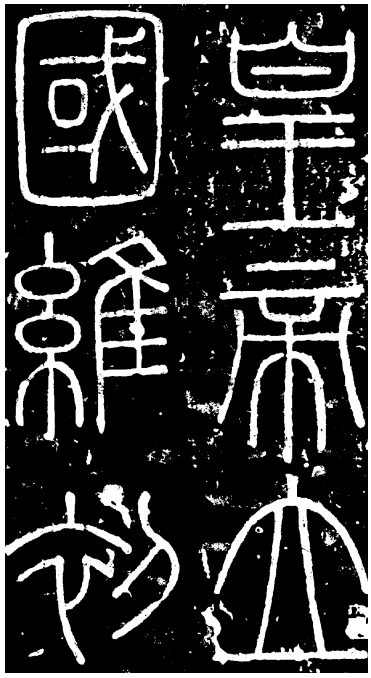


Abbildung 2.5: Siegelinschrift

4. Die offizielle Schrift:
   In der späten Qin- und der Han-Dynastie (206 v. Chr. bis 220 n. Chr.) wurde die offizielle Schrift eingeführt. Ziel war es ein schnelleres Schreiben zu ermöglichen, dazu wurden alle Kurven und Rundungen begradigt. Viele ähnliche Zeichenkomponenten wurden zusammengefaßt und, wo möglich, die Strichanzahl vermindert. Auch die Aussprache der einzelnen Zeichen wurde, wo man es für nötig hielt, geändert. Diese Schriftreform war die grösste seit den frühen Anfängen der chinesischen Schrift. Sie war aber so konsequent, daß sich die Schrift in den nächsten 1800 Jahren bis heute nur noch geringfügig änderte.

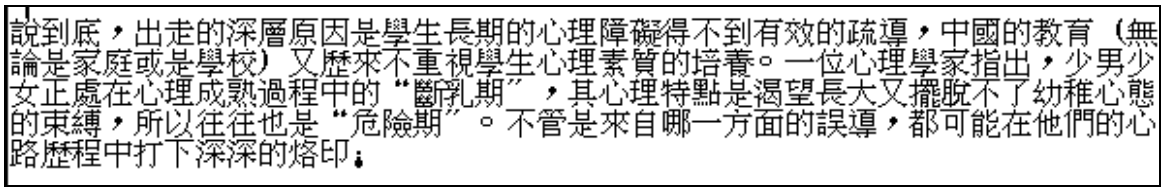

說到底，出走的深層原因是學生長期的心理障礙得不到有效的疏導，中國的教育（無論是家庭或是學校）又歷來不重視學生心理素質的培養。一位心理學家指出，少男少女正處在心理成熟過程中的“斷乳期”，其心理特點是渴望長大又擺脫不了幼稚心態的束縛，所以往往也是“危險期”。不管是來自哪一方面的誤導，都可能在他們的心路歷程中打下深深的烙印。

Abbildung 2.6: Offizielle Schrift wie sie heute noch in Taiwan verwendet wird

5. Standardisierte (vereinfachte) Form:
   Das Komitee zur Reform der Chinesischen Sprache, hatte in der Zeit von 1956-64 eine Reform beschlossen, in der folgende Vereinfachungen vorgenommen wurden:

Verschiedene Zeichen mit derselben Bedeutung wurden meistens eliminiert, wobei das häufiger vorkommende Zeichen oder das einfachere Zeichen übrig geblieben ist und gegebenenfalls noch weiter vereinfacht wurde. Dadurch wurde die Anzahl der häufig verwendeten Zeichen nahezu halbiert.
Diese Änderungen wurden aber nur auf dem Festland China durchgeführt. Hongkong, Taiwan und die meisten der Auslandschinesen benutzen die nicht vereinfachte Form, die auch als „Lang-Zeichen"-Form im Gegensatz zur „Kurz-Zeichen"-Form bezeichnet wird. Es existieren also heute zwei verwendete chinesische Schriftsysteme, wobei die Bedeutung der „Kurz-Zeichen" wahrscheinlich mit der Bedeutung der Volksrepublik China weiter zunehmen wird. Für „Kurz-Zeichen"- Kundige ist es einfacher „Lang-Zeichen"- Texte zu lesen als umgekehrt. Von dieser Reform sind ca. 2500 Zeichen betroffen.
Die Regierung der Volksrepublik China hatte im Jahre 1974 erneut versucht die chinesische Schrift zu reformieren. Die Einführung dieser erneuten Reform stieß allerdings bei der Bevölkerung auf Ablehnung und mußte einige Jahre später wieder aufgehoben werden. (Manchmal findet man neben den standardisierten Zeichen noch Fragmente der erneuten Reform, die einem als Sprachschüler das Leben schwer machen können.)

说到底，出走的深层原因是学生长期的心理障碍得不到有效的疏导，中国的教育（无论是家庭或是学校）又历来不重视学生心理素质的培养。一位心理学家指出，少男少女正处在心理成熟过程中的“断乳期”，其心理特点是渴望长大又摆脱不了幼稚心态的束缚，所以往往也是“危险期”。不管是来自哪一方面的误导，都可能在他们的心路历程中打下深深的烙印!

Abbildung 2.7: Vereinfachte Schrift wie sie heute in der Volksrepublik China verwendet wird. (der dargestellte Text entspricht dem in der vorherigen Abbildung)

Systematik hatte im alten China einen großen Stellenwert, so daß es schon früh Sammlungen (Zeichen-Wörterbücher) der Zeichen gab. Dies wurde mitunter dadurch begünstigt, daß schon früh verschiedene Arten des Druckhandwerks eingeführt wurden. Um die Entwicklung der Anzahl der Zeichen aufzuzeigen, möchte ich nach [12] stellvertretend 4 Wörterbücher erwähnen:

Name des (Zeichen-)Wörterbuchs	**Veröffentlichungs-jahr**	**Anzahl der Zeichen**
Shou1Wen2Jie3Zi4 (说文接字)	ca. 100	9353
Guang3 Yun4 (广韵)	ca. 1008	26000
Kang1xi1 Zi4dian3 (康熙字典)	ca. 1716	47035
Han4yu3 Da4 Zi4dian3 (汉语大字典)	1997	ca. 60000

2.2 Morphologie der chinesischen Sprache

In der frühen chinesischen Sprache repräsentierte ein Zeichen eine Sinneinheit, also genau ein Wort. Dies führte zu einer enormen Ansammlung von Zeichen, von denen einige so speziell waren, daß sie evt. nur wenige Male in der Literatur aufgetaucht sind. Zum Beispiel gab es für den Begriff Pferd mehr als hundert Ausprägungen. Von denen einige Zeichen sehr außergewöhnlich oder speziell waren:

- Pferd mit weißer Stirn

- Pferd mit schwarzen Lippen
- Pferd mit gelber Mähne
- Pferd mit einem weißen Auge
- Pferd, daß sieben Ellen groß ist
- Pferd mit blut-schwarzer Farbe
- ...

Durch diese Problematik, welche es äußerst schwer machte, daß außerhalb der Gelehrten oder Beamten, die allgemeine Bevölkerung die Zeichenschrift lernte, wurde es erforderlich ein weiteres strukturelles Element einzuführen. Wie in einer Schrift die auf einem Alphabet basiert, wurden Begriffe gebildet, indem mehre Zeichen zusammengefaßt wurden. Dies erlaubte es nun einen beliebig großen Wortschatz bei begrenzter Zeichenanzahl zu bilden. Heutzutage sind ca. 90 % aller Wörter aus bis zu 10 morphemen Einheiten (Zeichen) zusammengesetzt, wobei die mittlere Wortlänge ca. 2 Zeichen beträgt.
Beispiele:

- fliegen 飞 + Maschine 机 = Flugzeug 飞机
- schlagen 打 + elektrisch 电 + Sprache 话 = telefonieren 打电话
- knüpfen 结 + Frucht 果 = 结果 Ergebnis

Ein weiteres Beispiel soll anhand eines Satzes von Konfuzius den Unterschied vom früheren zum heutigen, die Morphologie verwendenden Chinesisch, verdeutlichen:

	ungetrennt	Sinneinheiten (Worte) getrennt
Konfuzius	学而是习之，不亦说乎？	学 而 是 习 之， 不 亦 说 乎 ？
heute	有时候学习很有意思？	有时候 学习 很 有意思 ？

(Deutsch: Ist Lernen nicht hin und wieder ein Vergnügen?)

Wie man sieht war in früherer Zeit das Trennen von Bedeutungseinheiten (Wörter) identisch mit dem Trennen von Zeichen, was heute nicht mehr so gilt und womit sich diese Arbeit unter anderem beschäftigt.

2.3 Das gesprochene Chinesisch

Die Aussprache des antiken Chinesisch hat sich bis zur Neuzeit stark gewandelt. Über die gesprochene chinesische Sprache, in der frühen Zeit lässt sich nur sehr wenig sagen. Auch nach Aufkommen und Standardisierung der Schrift änderte sich die Lautgebung sehr stark und Aussagen über die Lautgebung sind schwierig. Hilfreich für die Erforschung der frühen Phonetik ist die Poesie, in der besonders Reime und andere strukturelle Elemente Rückschlüsse auf die Aussprache zulassen. Diese Problematik ist unter anderem darauf zurückzuführen, daß die chinesischen Schriftzeichen sehr wenige Festlegungen über die Aussprache eines Wortes machen. Dies ermöglichte es sogar, daß völlig verschiedene, unabhängige Sprachen sich das chinesische Schriftsystem zu nutze machten. So kommt es zu dem Phänomen, daß sich zwei Sprecher völlig unterschiedlicher Sprachen über das Schriftsystem verständigen können. Japanisch und zum Teil Koreanisch sind Beispiele für solche Sprachen. Die alt-hebräischen Konsonantenschrift, in der die Vokale aus Effizienzgründen weggelassen wurden, ist ein ähnlicher Fall, in dem die Aussprachen von verschiedenen Worten nicht mehr ermittelt werden können.

Neben den Schwierigkeiten die phonetische Entwicklung in der Geschichte zu verfolgen, ergebenn sich weitere, wenn man festlegen möchte, was denn die heute gesprochene chinesische Sprache ist. Aufgrund der Eigenschaften des Chinesischen ist es möglich, daß viele „Dialekte“ das selbe Schriftsystem benutzen können. Dabei haben viele Dialekte grössere sprachliche Unterschiede als z.B. Deutsch und Englisch. Das bedeutet, daß Menschen eines Landes sich nicht sprachlich verständigen können, obwohl sie vielleicht dasselbe Schriftsystem beherrschen. Um ein Land einfach verwalten zu können und effektiv Radio- und Fernsehprogramme produzieren zu können, hat die Regierung der Volksrepublik China schon seit längerem damit begonnen, einen Dialekt, und zwar den in der Umgebung von Peking gesprochenen Mandarin-Dialekt, zur Hochsprache (Putonghua 普通话) zu deklarieren. In Schulen und allen öffentlichen Einrichtungen wird dieser Dialekt ausschließlich verwendet. Trotzdem gibt es noch viele Chinesen, die diese Hochsprache nicht verstehen und noch mehr, die sie nicht sprechen können. Die folgende Tabelle listet die wichtigsten chinesischen Dialekte nach Statistiken [13] auf, wobei die gesamte Anzahl der Sprecher sich bis heute allerdings mehr als verdoppelt hat:

Sprecheranzahl in Millionen

	1953	ca.1988
Nordchinesische Dialekte (4 Gruppen)	387	836
Wu-Dialekte (4 Gruppen)	46	77
Dialekte von Guangdong	27	47
Dialekte von Hunan und Jiangxi	26	56
Hakka-Dialekte	20	27
Dialekte von Süd-Fujian	15	25
Dialekte von Nord-Fujian	7	11
Gesamtzahl	***528***	***1079***

Zusätzlich gibt es noch Sprecher der chinesischen Sprache in Taiwan, Hongkong/Macao, in chinesischen Kolonien und Auslandschinesen in Nordamerika und Europa.

Neben diesen chinesischen Dialekten gibt es aber noch nichtchinesische Nationalitäten mit ihren eigenen Sprache in der Volksrepublik China. Die folgende Tabelle gibt über diese Sprachfamilien eine Übersicht (offizielle Zahlen von 1957):

Name	**Ethno-linguistische Gruppe**	**Siedlungsgebiet**	**Anzahl**
Zhuang	Thai	Yunnan, Guangdong	7800000
Uiguren	Türkische Gruppe	Xinjiang, West-Gansu	3900000
Yi	Tibeto-birmanische Gruppe	Yunnan, Guizhou, Hunan	3260000
Zang	Tibetische Gruppe	Tibet, Qinghai, Sichuan	2770000
Miao	Miao-Yao	SW-Provinzen	2680000
Mandschuren	Tungusische Gruppe	NO, Mongolei, Pekinger Region	2430000
Mongolen	Mongolische Gruppe	Mongolei, NO, Gansu, Qinghai	1640000
Buyi	Thai	Yunnan	1320000
Koreaner	Koreanische Gruppe	Nordosten	1250000

Abbildung 2.8: Die Verteilung der Dialekte in China (aus [13])

In dem Bemühen die Aussprache festzuhalten und niederzuschreiben wurden verschiedene Lautsprachesysteme eingeführt. Das Bopomofo-System benutzt für jedes Phonem ein eigenes Zeichen.

Abbildung 2.9: sechs Phoneme des zeichenbasierten Bopomofo-Systems

Weitere auf dem römischen Alphabet basierende Lautschriftsysteme wurden von Missionaren Anfang des neunzehnten Jahrhunderts, aus der Notwendigkeit die Sprache zu lernen, eingeführt. Wobei das Sprachempfinden jeweils stark von der eigenen Muttersprache geprägt war. Das Pinyin Lautsprachesystem wurde von der Regierung der Volksrepublik China zur Vereinheitlichung der Aussprache entwickelt und als Standard zur Erlernung der Hochsprache (Putonghua 普通话) eingeführt. Die folgende Tabelle vergleicht 3 auf Buchstaben basierende Systeme ausschnittsweise:

Pinyin	Wade-Giles	Yale
nian4	nien4	nyan4
piao4	p'iao4	pyau4
xuan1	hsu:an1	sywan1
zi5	tzu5	dz5
zong1	tsung1	dzung1

In Taiwan werden die Systeme Wade-Giles, Yale und Bopomofo nebeneinander benutzt, während die Volksrepublik das selbst entwickelte Pinyin-System nun mehr ausschließlich benutzt. Die vorliegende Arbeit beschäftigt sich mit dem Pinyin-System. Wobei eine Überführung der einzelnen Systeme ineinander, ohne grössere Schwierigkeiten mittels einfacher Tabellen erfolgen kann.

2.4 Weiter Besonderheiten der chinesischen Sprache

Die chinesische Sprache basiert auf einer großen Anzahl von Zeichen. Diese sind die kleinsten Sinneinheiten und Grundlage zum Aufbau von Wörtern und Sätzen. Diese Zeichen füllen jeweils genau ein Quadrat und haben alle dieselbe Grösse.

2.4.1 Die Schreibrichtung

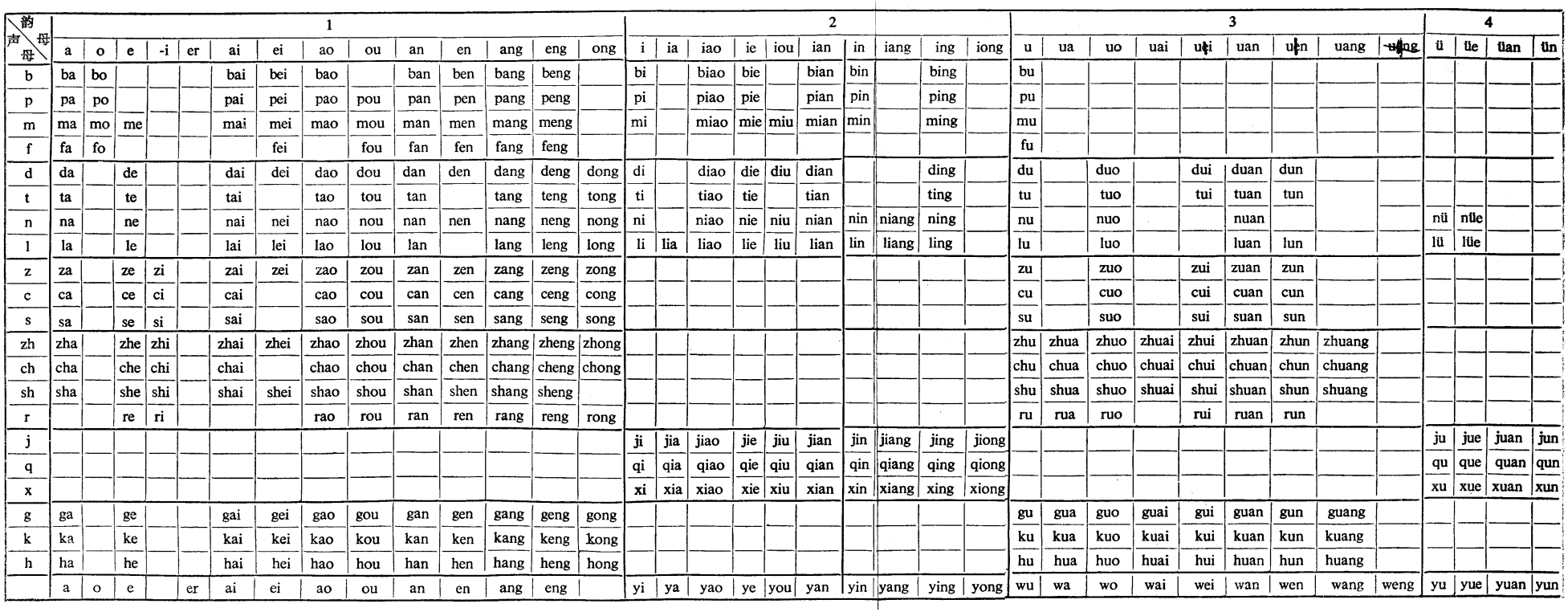
Da jedes Zeichen für sich eine unveränderliche, abgeschlossene Einheit darstellt, ist der Anknüpfungspunkt für das nächste zu schreibende Zeichen beliebig. Es gab in der Vergangenheit deshalb auch die unterschiedlichsten Schreibrichtungen nebeneinander. In Zeitungen wurden sogar auf der selben Seite unterschiedliche Schreibrichtungen eingesetzt. Die Volksrepublik China verwendet nahezu nur noch die uns geläufige Schreibrichtung (links-nach-rechts, oben-nach-unten, vorne-nach-hinten). In Taiwan sind noch andere Schreibrichtungen gebräuchlich, wobei durch Anforderungen der besseren Computerverarbeitbarkeit, immer mehr die uns geläufige Schreibrichtung verbreitet wird.

Abbildung 2.10: Beispiel einer in Taiwan gebräuchlich Schreibrichtungen
1. Mose 1 einer taiwanesischen Bibel
(oben-nach-unten, rechts-nach-links, hinten-nach-vorne)

2.4.2 Die Phonetik und Tonalität

Wörter der chinesische Hochsprache können in Aussprachesilben unterteilt werden, dabei ist es bemerkenswert, daß dabei von wenigen Ausnahmen abgesehen, jedes Zeichen genau eine Silbe repräsentiert. Wobei allerdings ein Zeichen je nach Kontext unterschiedliche Aussprachesilben repräsentieren kann.
Die Aussprachesilben bestehen aus einem Konsonanten (oder Affrikat = Verschlußlaut mit anschließendem Reibelaut) und einem Vokalkonstrukt. Wenige auch nur aus einem Vokalkonstrukt.
Menge der Konsonanten = {b, p, m, f, d, t, n, l, z, c, s, zh, ch, sh, r ,j, q, x, g, k, h, y, w}
Menge der Vokalkonstrukte = {a, o, e, -i, er, ai, ei, ao, ou, an, en, ang, eng, ong,
i, ia, iao, ie, iou, ian, in, iang, ing, iong,
u, ua, uo, uai, ui, uan, un, uang, ü, üe, üan, ün}

Die Tabelle auf der folgenden Seite zeigt die erlaubten Kombination von Konsonanten und Vokalkonstrukten.



韵母 / 声母	1														
	a	o	e	-i	er	ai	ei	ao	ou	an	en	ang	eng	ong
b	ba	bo				bai	bei	bao		ban	ben	bang	beng	
p	pa	po				pai	pei	pao	pou	pan	pen	pang	peng	
m	ma	mo	me			mai	mei	mao	mou	man	men	mang	meng	
f	fa	fo					fei		fou	fan	fen	fang	feng	
d	da		de			dai	dei	dao	dou	dan	den	dang	deng	dong
t	ta		te			tai		tao	tou	tan		tang	teng	tong
n	na		ne			nai	nei	nao	nou	nan	nen	nang	neng	nong
l	la		le			lai	lei	lao	lou	lan		lang	leng	long
z	za		ze	zi		zai	zei	zao	zou	zan	zen	zang	zeng	zong
c	ca		ce	ci		cai		cao	cou	can	cen	cang	ceng	cong
s	sa		se	si		sai		sao	sou	san	sen	sang	seng	song
zh	zha		zhe	zhi		zhai	zhei	zhao	zhou	zhan	zhen	zhang	zheng	zhong
ch	cha		che	chi		chai		chao	chou	chan	chen	chang	cheng	chong
sh	sha		she	shi		shai	shei	shao	shou	shan	shen	shang	sheng	
r			re	ri				rao	rou	ran	ren	rang	reng	rong
j														
q														
x														
g	ga		ge			gai	gei	gao	gou	gan	gen	gang	geng	gong
k	ka		ke			kai	kei	kao	kou	kan	ken	kang	keng	kong
h	ha		he			hai	hei	hao	hou	han	hen	hang	heng	hong
	a	o	e		er	ai	ei	ao	ou	an	en	ang	eng	

韵母 / 声母	2										
	i	ia	iao	ie	iou	ian	in	iang	ing	iong
b	bi		biao	bie		bian	bin		bing	
p	pi		piao	pie		pian	pin		ping	
m	mi		miao	mie	miu	mian	min		ming	
f										
d	di		diao	die	diu	dian			ding	
t	ti		tiao	tie		tian			ting	
n	ni		niao	nie	niu	nian	nin	niang	ning	
l	li	lia	liao	lie	liu	lian	lin	liang	ling	
z										
c										
s										
zh										
ch										
sh										
r										
j	ji	jia	jiao	jie	jiu	jian	jin	jiang	jing	jiong
q	qi	qia	qiao	qie	qiu	qian	qin	qiang	qing	qiong
x	xi	xia	xiao	xie	xiu	xian	xin	xiang	xing	xiong
g										
k										
h										
	yi	ya	yao	ye	you	yan	yin	yang	ying	yong

韵母 / 声母	3									
	u	ua	uo	uai	uei	uan	uen	uang	ueng
b	bu									
p	pu									
m	mu									
f	fu									
d	du		duo		dui	duan	dun			
t	tu		tuo		tui	tuan	tun			
n	nu		nuo			nuan				
l	lu		luo			luan	lun			
z	zu		zuo		zui	zuan	zun			
c	cu		cuo		cui	cuan	cun			
s	su		suo		sui	suan	sun			
zh	zhu	zhua	zhuo	zhuai	zhui	zhuan	zhun	zhuang		
ch	chu	chua	chuo	chuai	chui	chuan	chun	chuang		
sh	shu	shua	shuo	shuai	shui	shuan	shun	shuang		
r	ru	rua	ruo		rui	ruan	run			
j										
q										
x										
g	gu	gua	guo	guai	gui	guan	gun	guang		
k	ku	kua	kuo	kuai	kui	kuan	kun	kuang		
h	hu	hua	huo	huai	hui	huan	hun	huang		
	wu	wa	wo	wai	wei	wan	wen	wang	weng	

韵母 / 声母	4			
	ü	üe	üan	ün
b				
p				
m				
f				
d				
t				
n	nü	nüe		
l	lü	lüe		
z				
c				
s				
zh				
ch				
sh				
r				
j	ju	jue	juan	jun
q	qu	que	quan	qun
x	xu	xue	xuan	xun
g				
k				
h				
	yu	yue	yuan	yun

Abbildung 2.11: Tabelle der Pinyin - Sprachlaute des Mandarin Dialekts

Die Anzahl der verfügbaren Aussprachesilben, wie in Abbildung 2.11 ersichtlich, beträgt nur ca. 400. Diese Zahl ist im Gegensatz zur Gesamtanzahl der Zeichen von ca. 60000 sehr niedrig, so daß im Mittel 150 Zeichen dieselbe Aussprache haben. Um diese große Anzahl gleich ausgesprochener Zeichen doch ein wenig besser zu differenzieren, enthält jede Aussprachesilbe eine bestimmten Tonverlauf, d.h. eine Veränderung der Tonhöhe während der Aussprache. In der Hochsprache ist dies einer von 5 möglichen Tonverläufen, wobei der fünfte Tonverlauf einer Nichtbetonung entspricht. Andere chinesische Dialekte enthalten bis zu 13 Tonverläufe. Der Tonverlauf ist bedeutungstragend, somit ist bei gleicher Aussprachesilbe und anderem Tonverlauf ein völlig anderes Zeichen mit unterschiedlicher Bedeutung gemeint.

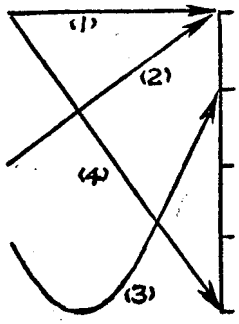


Abbildung 2.12: Die Tonverläufe der chinesischen Hochsprache

Die Pinyin-Lautschrift gibt beides wieder, die Aussprachesilbe und den Tonverlauf. Die in der Literatur verwendete Kennzeichnung des Tonverlaufs (oder kurz Ton genannt) könnte auf Computern nur unter zu Hilfenahme von speziellen Zeichensätzen erfolgen und wird deshalb der Einfachheit halber durch Zahlen gekennzeichnet (dies erleichtert auch die Computereingabe).

Beispiel: die 5 Töne einer Aussprache-Silbe mit einer Auswahl von Bedeutungen:

- mā = ma1 : Mutter 妈, streicheln 摩, Dämmerung 麻, wischen 抹, ...
- má = ma2 : Hanf 麻, Anästhesie Âé, Kröte 蟆, ...
- mǎ = ma3 : Pferd 马, Nummer 码, stapeln 码, Yard 码, Ameise 蚂, ...
- mà = ma4 : schimpfen 骂, Heuschrecke 蚂, ...
- ma = ma5 : Fragepartikel 吗, Bewußtseinskennzeichnung 嘛, ...

Trotz der Tonalität der Sprache ergeben sich für eine Aussprachesilbe mit Tonverlauf oftmals noch Dutzende von möglichen Zeichen und damit auch Bedeutungen. Das heißt, daß beim Vorlesen, also der Transformation von Schrift in gesprochene Sprache, Information verloren gehen, die nur aus dem Satzkontext und Metawissen wiedergewonnen werden kann. Chinesische Zeichen und auch Wörter sind also sehr stark vom Kontext abhängig, was im Deutschen nur sehr viel seltener der Fall ist.

Beispiel: Auch im Deutschen kann Kontextwissen erforderlich sein:
Er geht zur Bank.

➔ Kann doppeldeutig sein: Geht er zum Geldinstitut oder zur Sitzbank?

Er sitzt auf der Bank, um sich auszuruhen | Er geht zur Bank, um ein Konto zu eröffnen.

➔ Bedeutung kann mit großer Wahrscheinlichkeit aus dem Kontext erschlossen werden.

Die Tonalität selbst ist wiederum geringfügig vom Kontext abhängig. Aber die Regeln der Änderung von Tönen je nach Kontext weichen bei unterschiedlichen Lehrbüchern wieder voneinander ab. Sie werden in dieser Arbeit nicht berücksichtigt.

2.4.3 Chinesisches Zahlensystem

Das chinesisches Zahlensystem unterscheidet sich in zweierlei Hinsicht von unserem. Erstens werden statt der arabischen Ziffern 0123456789 die chinesischen Zeichen 零(〇)一二三四五六七八九十百千万 usw. verwendet und zweitens basiert das Zusammenfassen großer Zahlen nicht auf der Basis 1000 sondern auf der Basis 10000.

Beispiel: Hundertmillionen (100 * 1000 * 1000) entspricht 千万 (1000 * 10000)

Dies führt manchmal beim Kontakt mit Chinesen zu Verwechslungen. Heute werden aber auch immer mehr arabischen Ziffern verwendet. Besonders in der Wissenschaft, als Artikelnummern und zur mathematischen Berechnung.

2.4.4 Grammatik der chinesischen Sprache

Die chinesische Grammatik benutzt keine so ausdrucksstarken Methoden wie Konjunktion und Deklination, sondern muß viele grammatische Strukturen über Funktionalzeichen und Konstrukte realisieren.
Eine Behandlung der unerschöpflichen chinesischen Grammatik würde mir Sicherheit den Rahmen dieser Arbeit sprengen. Eine gute Einführung ist in [6] und [7] zu finden.

2.4.5 Sprichwörter (成语)

In der gehobenen chinesischen Ausdrucksweise kommt der Verwendung von Sprichwörtern eine besondere Bedeutung zu. Diese Sprichwörter (³ÉÓï cheng2yu3) haben ihren Ursprung meist in der klassischen chinesischen Literatur und stammen dort von einer speziellen Geschichte oder einer Anekdote. Sie bestehen meist aus genau vier Zeichen, die, ohne einen Satz zu bilden, abstrakt eine Aussage über einen Sachverhalt wiedergeben.
Um gehobenes Chinesisch verstehen oder sprechen zu können, ist es nötig eine grössere Auswahl der mehrere tausend Sprichwörter zu kennen.

Beispiel einer Anekdote mit abgeleitetem Sprichwort:

=> Sprichwort: 杯弓蛇影 (Glas Bogen Schlange Schatten)

晋朝的时候，有一个人叫乐广，他很会说话， 很能用道理说服人。
乐广有一个好朋友，两个人常常在一起喝酒，谈天。可是后来，那个朋友有一个多月没到乐广家来了。乐广就派人去了解情况。派去的人回来说，那个朋友病了。原来上一次他在乐广家喝酒，看见酒**杯**里有一条小**蛇**，可是酒已经喝下去了，有什么办法呢！他当时心里很不舒服， 回到家就病了。
乐广听了，觉得很奇怪，酒**杯**里怎么会有小**蛇**呢？他走到上次喝酒的地方，仔仔细

细地看了一遍，忽然看见墙上挂着一张**弓**，他立刻明白了。于是，他又派人去请那个朋友来喝酒，而且还说他能治好他的病。
那个朋友开始恨不愿意来，最后他还是来了。乐广还请他喝酒，让他坐老地方。那个朋友本来就很不妨心，他往酒**杯**里一看，嘿，那条小**蛇**还在酒**杯**里呢！他吓得出了一身冷汗。乐广指着墙上的**弓**笑着说：“酒**杯**里没有什么**蛇**，这是墙上**弓**的**影**子。”
他把墙上的**弓**拿下来，酒**杯**里的小**蛇**立刻不见了。他朋友这才明白是怎么一回事，病也就好了。

Übersetzung:

In der Jin-Dynastie gab es einen Mann namens Yueguang, der sehr redegewandt war und mit Argumenten überzeugen konnte.
Yueguang hatte einen guten Freund mit dem er oft zusammensaß und beim Schnapstrinken über Gott und die Welt redete. Aber eines Tages blieb jener Freund mehr als einen Monat aus. Yueguang war besorgt und schickte jemanden um in Erfahrung zu bringen, was denn los ist. Als der Bote zurückkam, berichtete er, daß jener Freund seitdem er das letzte Mal bei Yueguang Schnaps trinken war, krank ist. Weil in dem Schnapsglas eine kleine Schlange war, aber als er sie entdeckte war es schon zu spät, denn er hatte schon davon getrunken - er konnte also nichts mehr machen. Von dieser Zeit ging es ihm dann immer schlechter.
Als Yueguang das hörte, war ihm seltsam zu Mute. Wie konnte in den Schnaps eine kleine Schlange geraten? Er ging an den Ort an dem sie das letzte Mal zusammen getrunken hatten zurück und untersuchte alles einmal sorgfältig. Plötzlich sah er an der Wand einen Bogen hängen und sofort wurde ihm einiges klar. Wieder schickte er jemanden zu seinem Freund, daß er wieder zum Schnapstrinken kommen soll - und wenn er kommt, daß er ihn auch von seiner Krankheit heilen könne.
Jener Freund wollte anfangs überhaupt nicht kommen. Am Ende kam er trotzdem, weil er aus Verzweiflung nicht mehr wußte was er tun soll. Yueguang bot ihm wieder Schnaps am selben Ort wie letztes Mal an. Der Freund fühlte sich immer unwohler in seiner Haut. Als er dann ins Schnapsglas schaute sah er wieder die kleine Schlange und erschrakt fürchterlich, daß ihm der kalte Schweiß aus allen Poren kam.
Yueguang zeigte zur zum Bogen an der Wand und beruhigte ihn: „Die Schlange in deinem Glas ist nur der Schatten des Bogens an der Wand.“ Er nahm den Bogen von der Wand und sofort verschwand die Schlange im Glas. Jetzt begriff der Freund endlich die gesamte Sache und seine Krankheit war auch geheilt.

Bedeutung und Anwendung:
Dieses Sprichwort wird für Menschen benutzt, die alles überängstlich ohne Grund anzweifeln.
(这个成语用来比喻有人怀疑这个，怀疑那个，实际并没有那么一回事。)

Die Bedeutung und Schwierigkeit chinesischer Sprichwörter geht weit über die Bedeutung der deutschen Sprichwörter hinaus, da sie

a) häufiger benutzt werden (im gehobenen Chinesisch)
b) selten auf den Sinn geschlossen werden kann (im Deutschen ist der Bezug zur Bedeutung meist direkter z.B. Lügen haben kurze Beine, Morgenstund hat Gold im Mund, ...)
c) aus klassischem Chinesisch abgeleitet sind
d) keine grammatische Struktur besitzen
e) meist nur aus Wortbruchteilen bestehen

2.4.6 Sonstiges

- Städtenamen: Manche Städtenamen sind sehr alt und benutzen sehr seltene Zeichen, die oft vielen Chinesen unbekannt sind oder in manchen Computerzeichensätzen nicht vorrätig sind.

- Personennamen: Der Nachname kommt vor dem Vornamen bei der Anrede. Üblich ist es immer sowohl Nachname als auch Vorname zu nennen, das gilt sogar für Freunde und teilweise auch Familienangehörige. Namenszeichen können auch sehr selten sein.
- Chinesischer Kalender: Der chinesische Kalender basiert auf den Mondphasen. Dadurch ist der Jahresbeginn um eine sich ändernde Anzahl von Tagen zu unserem Jahresbeginn verschoben. Neben dem chinesische Kalender wird auch der westliche Kalender verwendet.

2.5 Chinesisch und Computer

Seit Einführung der Drucktechnik, die viel früher als in der westlichen Welt eingeführt wurde, war man bemüht diese zu verbessern. Dies führte im letzten Jahrhundert dann zur Entwicklung von chinesischen Schreibmaschinen, die sich aufgrund der Komplexität, Grösse und aufwendigen Bedienung nie durchsetzten konnten. Es konnte immer nur ein Teil der Zeichen direkt gedruckt werden (max. 2000). Alle anderen, selteneren Zeichen mußten von Hand aus einem Kasten herausgesucht und auf eine Vorrichtung aufgesteckt werden. In den letzten Jahrzehnten, mit Einführung des Computers, änderte sich diese Situation dann schlagartig. Der kostengünstigere Computer ermöglichte es, alle Zeichen gleichartig zu behandeln und mittels Matrixdruckern auszugeben. Trotz dieser Vorteile ist die Verarbeitung der chinesischen Schrift mittels der, hauptsächlich im Westen, entwickelten Computer nicht so ohne weiteres möglich. Es stellen sich die in den folgenden Unterkapiteln behandelten Fragen nach der Eingabemethode und der Codierung chinesischer Zeichen.

2.5.1 Eingabemethoden

In China werden Computer mit einer Standardtastatur ausgeliefert, denn eine spezielle Tastatur mit allen chinesischen Zeichen wäre zu teuer, zu unübersichtlich und zu groß.
Es müssen deshalb Verfahren entwickelt werden, um die chinesischen Zeichen auf die wenigen Tasten der Tastatur abzubilden. Neben Verfahren, die die Tastatur als Eingabeinstrument benutzen existieren auch andere Hilfsmittel. Insgesamt haben sich eine beträchtliche Anzahl von Eingabemethoden für die unterschiedlichsten Einsatzgebiete entwickelt.

Beispiele:
- Pinyin-Eingabemethode: Bei dieser Eingabemethode gibt man über die Tastatur die Aussprache eines Zeichens in der Pinyin-Lautschrift plus Ton ein, der Computer generiert dann eine Liste von in Frage kommenden Zeichen, aus denen der Benutzer dann das richtige auswählen kann. Wenn man die Pinyin-Lautschrift beherrscht, kann man diese Eingabemethoden sehr schnell erlernen. Statistische oder neuronale Methoden helfen bei manchen Systemen die Auswahlliste möglichst optimal zu sortieren, so daß oft das erste vorgeschlagene Zeichen benutzt werden kann und der Suchaufwand sich so stark reduziert. Trotzdem ist die Eingabegeschwindigkeit von Texten relativ niedrig. Bei Zeichen, deren Aussprache nicht genau bekannt ist , muß zusätzlich noch in einem Aussprachewörterbuch nachgeschlagen werden.
- Zeichenzerlegende Eingabemethoden: Von diesem Typ gibt es sehr viele unterschiedliche Eingabemethoden. Meist wird aus der Struktur der Zeichen ein Zahlen- oder Zeichencode

erzeugt, der das Zeichen repräsentiert. Diese Form der Eingabe erfordert einen großen Lernaufwand von Codes, die je nach der verwendeten Methode teils eine sehr schnelle Eingabe von geübten Personen erlauben.

- Pen-Eingabe: Eine grössere Bedeutung hat die Eingabe mittels eines Pens erlangt. Bei der Zeichen auf ein spezielles Tableau geschrieben werden, welche mittels Software ausgewertet werden. Dieses Verfahren erfordert allerdings zusätzliche Hardware und Übung im gleichmässigen und computerleserlichen Schreiben. Nach einer Trainingsphase ist die Erkennungsrate für Gelegenheitsschreiber im allgemeinen hoch genug. Neben der Bitmap des eigentlichen Zeichens werden auch Strichrichtung, Strichreihenfolge und Strichanzahl ausgewertet, somit ist es möglich ein einzelnes sorgfältig von Hand geschriebenes Zeichen zu erkennen.
- OCR-Eingabe: Mittels eines Scanners werden Texte als Bilder zum Computer übertragen, wo sie dann in Zeichen umgesetzt werden, d.h. in einzelne Teile partitioniert und diese dann Zeichen zu geordnet werden. Dieses Verfahren eignet sich nur für gedruckte Texte.
- Spracherkennung: Die Spracherkennung ist sicher die natürlichste Eingabemethode und würde es erlauben die Benutzerschnittstelle für alle Sprachen einheitlich zu gestalten. Allerdings erfordert dieses Eingabeverfahren beträchtliche Hardwareanforderungen. Die Verfahren zur Spracherkennung haben sich in den letzten Jahren erheblich verbessert, so daß in naher Zukunft dieses Verfahren sicher stark an Bedeutung gewinnen wird. Allerdings muß genauso wie bei der Pinyin-Eingabemethode die Mehrdeutigkeit der Abbildung der Laute auf die Zeichen behandelt werden. Dies erfordert Kontextinformationen, grammatikalisches und semantisches Wissen.

2.5.2 Zeichendarstellung und Fonts

Die Zeichendarstellung auf dem Bildschirm und anderen Ausgabegeräten wie Druckern, erfordert besondere Maßnahmen:

- spezielle Grafiktreiber die mehr als 128/256 Zeichen darstellen können
- hochauflösende Bildschirme und Drucker
- Fontgrössen von evt. mehreren Megabyte müssen verarbeitet werden können (TTF 1-4 MB)
- ein Zwei-Bytecode ist erforderlich um alle Zeichen codieren zu können
- ein Zeilenumbruch darf nur immer nach einer geraden Anzahl von Bytes geschehen
- grafikfähige Drucker mit genügend Hauptspeicher (mehr als 1 MB bei Laserdruckern erforderlich)

2.5.3 Codierung

Für die chinesische Schriftsprache existieren verschiedene Zwei-Bytecodierungen. Sie unterscheiden sich nach Anwendungsgebiet und Menge der darstellbaren Zeichen.

- Telegraphiecode: 1881 wurde in China der Telegraphiecode festgelegt, bei dem ein Codewort aus 4 Ziffern von 0000 bis 9999 besteht. Dieser Telegraphiecode enthält 10000 Zeichen und wird bis heute noch verwendet.
- GuoBiao: Offizielle Codierung der Volksrepublik China. Im Jahr1981 eingeführt, enthält die Kodierung 7783 chinesische Zeichen. Dieser Standardschriftzeichencode (Guo2-Biao1) existiert in verschiedene Varianten, da er mehrmals ergänzt und korrigiert wurde, was aber

in der Praxis nur wenige Zeichen betrifft. Außer den chinesischen Zeichen enthält er zusätzlich die Kodierung der japanischen Lautschrift, Hiragana und Katakana, der kyrillischen Schrift, sowie der lateinischen Schrift. Insbesondere ist der GuoBiao Code eine Obermenge des Standard ASCII-Codes. Alle nicht ASCII-Zeichen entsprechen einem Wert grösser 127, somit lassen sich 16384 nicht ASCII 2-Byte Zeichen darstellen. Somit ist es möglich problemlos ASCII Zeichen mit chinesischen Zeichen in einem Text darzustellen. Wenn zwei aufeinander folgende Zeichen einen ASCII-Wert grösser als 127 haben, werden sie als ein chinesisches Zeichen interpretiert, sonst als normale Zeichen. Dies führt in der deutschen Sprache dazu, daß zwar einzelne Umlaute dargestellt werden können (ä, ü, ö, ß) aber nicht zwei aufeinander folgende Umlaute. Deshalb werden in dieser Arbeit auch die Schreibweisen „grösser, lässt, ...“ statt „gr蓥er, l溥t, ...“ verwendet.

- Big5: Der Big5-Code wird in Taiwan und Hongkong benutzt.
- Unicode: Der Unicode codiert mit einem Zwei-Byte Code die gebräuchlichsten Alphabete der Welt, wie das lateinische, griechische, kyrillische, hebräische, arabische, thailändische, mongolische, und viele andere. Unter anderem auch einige exotische Alphabete, wie z.B. das Tibetische. Die asiatischen Schriftzeichencodes werden in der Unicode Terminologie als CJK-Gruppe (für China, Japan, Korea) bezeichnet. Der Unicode ermöglicht es in einem Dokument beliebige Schriften gemeinsam zu nutzen, er ermöglicht es sogar die Schreibrichtung innerhalb eines Dokument zu wechseln (z.B. von oben nach unten) was mittels Steuerkommandos realisiert wird.
  Der Unicode wird bis jetzt noch nicht oft angewendet, aber aufgrund seiner Flexibilität und Einfachheit wird er die länderspezifischen Kodierungen mittelfristig sicher ablösen. Auch die Verwendung von Unicode in Java und Windows NT werden die Unicodeverbreitung fördern.
- hz: Die hz-Kodierung ist eine Codierung, in der die chinesischen Zeichen auf ASCII-Zeichen abgebildet werden, um problemlos Emails verschicken zu können. ‘~{‘ entspricht dabei der Kennzeichnung, daß alle folgenden Zeichen bis ‘~}~’ als chinesische Zeichen aufgefaßt werden sollen.

Um die einzelnen Codierungen ineinander zu überführen gibt es im Public Domain sowie im kommerziellen Bereich entsprechende Konverter.

2.5.4 Vorteile der chinesischen Schrift

Neben den Eigenschaften, die im Vergleich zu unserem Schriftsystem die Verwendung der chinesischen Schrift erschweren, gibt es auch einige Vereinfachungen und Vorteile:

- Keine Silbentrennung: Da es im Chinesischen weder Leerzeichen zwischen zwei Zeichen, noch Wortgrenzen gibt, kann nach jedem Zeichen die Zeile umgebrochen werden. Es müssen also keine Regeln für eine Silbentrennung gelernt, noch in einer Textverarbeitung implementiert werden.
- Einheitliche Zeichenbreite: Da jedes Zeichen die gleichen Ausmaße hat, gibt es keine Notwendigkeit Verfahren für einen Blocksatz zu implementieren. Proportionalschriften mit Ausrichtungsproblemen oder sich verbreiternde Zeichen bei Fettdruck gibt es nicht.
- Keine Groß-Klein-Schreibung
- Speichereffizienz: Ein mittels chinesischen Schriftzeichen geschriebener Text nimmt im Mittel weniger als 60 % des Speicherplatzes als ein vergleichbarer englischer oder deut-

scher Text ein. Selbst nach Kompression beider Texte (pkzip) sind chinesische Texte noch 20 % kompakter als englische oder deutsche Texte.

2.5.5 Chinesische Systeme

Als Chinesische Systeme bezeichnet man Softwareprodukte, die es ermöglichen chinesische Schrift auf dem Computer zu verwenden. Neben der reine Darstellung der Schrift ist auch die Möglichkeit chinesische Schrift einzugeben vorhanden. Man kann folgende Klassen von Systemen unterscheiden.

- Chinesisches Betriebsystem: z.B. chinesisches Windows, ZWDOS, CCDOS, KC,
- Spezielle Programme (Editoren):z.B. Nanji-Star, byx-edit, DingDang, chinese Word, chin.TEX / emacs ,....
- Chinesische Systeme, die ein englisches/deutsches Windows-Betriebsystem ergänzen, so daß chinesische Schrift verarbeitet werden kann: z.B. RichWin, Chinese Star [13], TwinBridge, Union Way, ...
- Unicode Systeme: z.B. Java, Windows NT, ... Wobei noch keine Eingabemethoden definiert sind.

3 Relevantes Datenmaterial

Nachdem im vorausgegangenen Kapitel Aspekte und Schwierigkeiten der chinesischen Sprache behandelt wurden, soll in diesem Kapitel beschrieben werden, wie benötigtes Datenmaterial für die Lautschriftumsetzung und Worttrennung, unter Berücksichtigung der Probleme der Beschaffung und notwendigen Korrektur, gesammelt und weiterverarbeitet wurde. Im folgenden Kapitel wird ein Verfahren, welches mit teilweise noch fehlerbehafteten Daten eine möglichst gutes Romanisierungsergebnis liefert, dargestellt. Eine möglichst einfache Portierung auf Unix soll als weiteres Entwurfsziel berücksichtigt werden.

Das Romanisierungssystem soll möglichst korrekte Entscheidungen zur Worttrennung und Pinyinumsetzung treffen. Dazu benötigt das Romanisierungssystem Daten unterschiedlicher Art, wobei eine hohe Qualität und eine große Quantität der Daten für eine befriedigende Romanisierung wichtigste Voraussetzung sind. Dem Sammeln genügend vieler Daten kommt also eine wesentliche Bedeutung zu.

Das Romanisierungssystem benötigt, um einen Satz in Wörter zu trennen und diese dann kontextbezogen in die Lautschrift zu überführen, im wesentlichen die folgenden 3 Datenarten:
- Wörterbücher: Wort-Pinyin-Umsetzung, Zeichen-Pinyin-Umsetzung
- Statistiken: Zeichenwahrscheinlichkeit, Wortwahrscheinlichkeit
- Kontextinformationen: spezielle Bigramme mit Wort-Pinyin-Abbildung (ein Bigramm ist ein Maß für die Wahrscheinlichkeit, daß zwei Worte/Zeichen direkt aufeinander folgen)

Leider waren die meisten dieser Datenarten nicht direkt verfügbar, sondern mußten indirekt erzeugt werden. Einzig chinesische Texte waren in ausreichender Anzahl zu erhalten.

3.1 Sammeln von Daten

Als Datenquellen wurden die folgenden verwendet:
- Internet:
  - viele Magazine, Berichte und andere Online-Texte
  - VOA (Voice of America) bietet Texte mit Pinyinübertragung online an
  - Wortlisten
  - Statistiken
  - Software
- Chinesische Wörterbücher
- Computer Wörterbücher
- Hilfe durch Chinesen bei Problemen, bei denen Unterstützung durch „Muttersprachler“ nötig war
- Systematische Bücher über die chinesische Sprache
- Zeitungstexte und Bücher in Verbindung mit einem Scanner und chinesischer OCR-Software

Die Suche im Internet nach relevanten Daten hat neben der Unüberschaubarkeit des Internets folgende Schwierigkeiten aufgeworfen:
- Das Internet ist in China erst seit kurzer Zeit etabliert

- Staatliche Zensur aller Veröffentlichungen ⇒ viel Propaganda
- Der Begriff „Public Domain / Freeware“ ist auf dem Festland China nahezu unbekannt ⇒ wenig frei verwendbares Material
- Verschiedene Kodierungen Big5/GB, Langzeichen/Kurzzeichen
- Verschiedene Lautsprachesysteme
- Auslandschinesen, die im Internet sehr aktiv sind, benutzen häufig nicht die chinesische Hochsprache (Mandarin) sondern sprechen Kantonesisch
- Die Qualität der meisten Rohdaten wie Wortlisten, Pinyin-Transkriptionen war sehr schlecht.
- Themenbezogenheit: Es mußte darauf geachtet werden, daß die gesammelten Daten nicht nur aus speziellen, engen Themengebieten stammen, sondern allgemein repräsentativ sind.

3.2 Datenaufarbeitung

Nachdem Daten in ausreichender Anzahl gesammelt wurden, mußten die Daten aufbereitet werden. Die Aufbereitung umfaßt dabei die Teilaufgaben:

- Zusammenführen verschiedener Datenquellen
- Daten ergänzen
- Daten korrigieren
- Daten formatieren
- Statistik- und Kontext-Informationen generieren

Für die obigen Aufgaben wurden verschiedene Hilfsmittel eingesetzt, welche in den folgenden Unterkapiteln kurz beschrieben werden:

3.2.1 Hilfetools zur Textbearbeitung

Mittels der 4GL-Sprache SAL (Scalable Application Language) wurden einige Tools zur Textzerlegung und Formatänderungen erstellt. Diese werden aufgrund ihrer Kurzlebigkeit (manche wurden nur einmal verwendet), ihrem steten Wandel und ihrer Einfachheit nicht weiter beschrieben.

3.2.2 SQL-Datenbank und SQL-Skripts

Zur Verwaltung der gesammelten Daten wurde die Datenbank SQLBase6.1 von Centura eingesetzt. Eine SQL-Datenbank eignete sich besonders um die Daten zusammenzuführen, einfache Formatänderungen vorzunehmen, Statistiken zu erzeugen, eine Basis-Pinyinumsetzung vorzunehmen, bequem den Datenbestand zu sichern, doppelte Daten zu eliminieren und Datenbestände zu sortieren. Um diese Aufgaben effizient durchführen zu können, wurden SQL-Skripts für das SQLTalk-System entwickelt, so daß wiederkehrende oder ähnliche Aufgaben durch Aufruf des Skriptes erledigt werden konnten.

Die Datenbank eignete sich auch als allgemeines Suchwerkzeug, wobei Suchanfragen frei in SQL definiert werden konnten und somit flexibel nahezu alle Aufgaben der Vorverarbeitung und Teile der Weiterverarbeitung ermöglicht wurden.

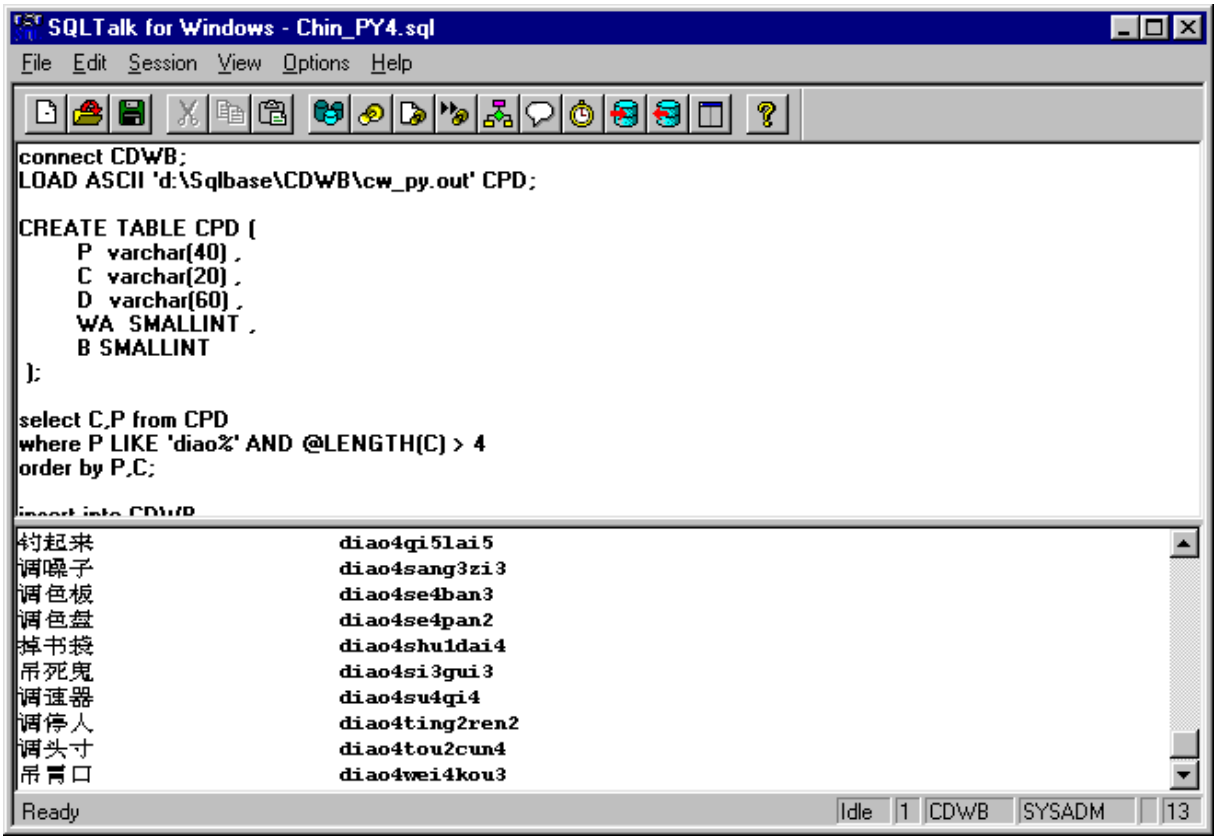


Abbildung 3.1: SQL-Skripts unter SQLTalk

3.2.3 Wörterbuchverarbeitung

Bei der Benutzung von SQL-Skripts zeigten sich allerdings hauptsächlich während des Korrigierens und Ergänzens von Daten im Wörterbuch beim Bedienkomfort und der Handhabung eindeutig Schwachpunkte. Somit wurde es nötig, ein auf die Datenbank aufsetzendes Programm, mit einer Benutzeroberfläche und einer Operationslogik auszustatten.

Mittels der Entwicklungsumgebung Centura Team Developer 1.0 Beta8 wurde dann für die Weiterverarbeitung der Daten ein Hilfsprogramm zur Wörterbuchbearbeitung und verschiedene andere Operationen entworfen und implementiert.
Dieses Programm ist in mehrere Module gegliedert:

- Wörterbuchbearbeitung: Editieren, Löschen und Hinzufügen von Einträgen
- Import und Export von Listen, freie Formatierung, Datenbank Import/Export
- Wortextraktion: aus Texten nicht bekannte chinesische Wörter extrahieren und eine Basis-Pinyin-Zuordnung vornehmen

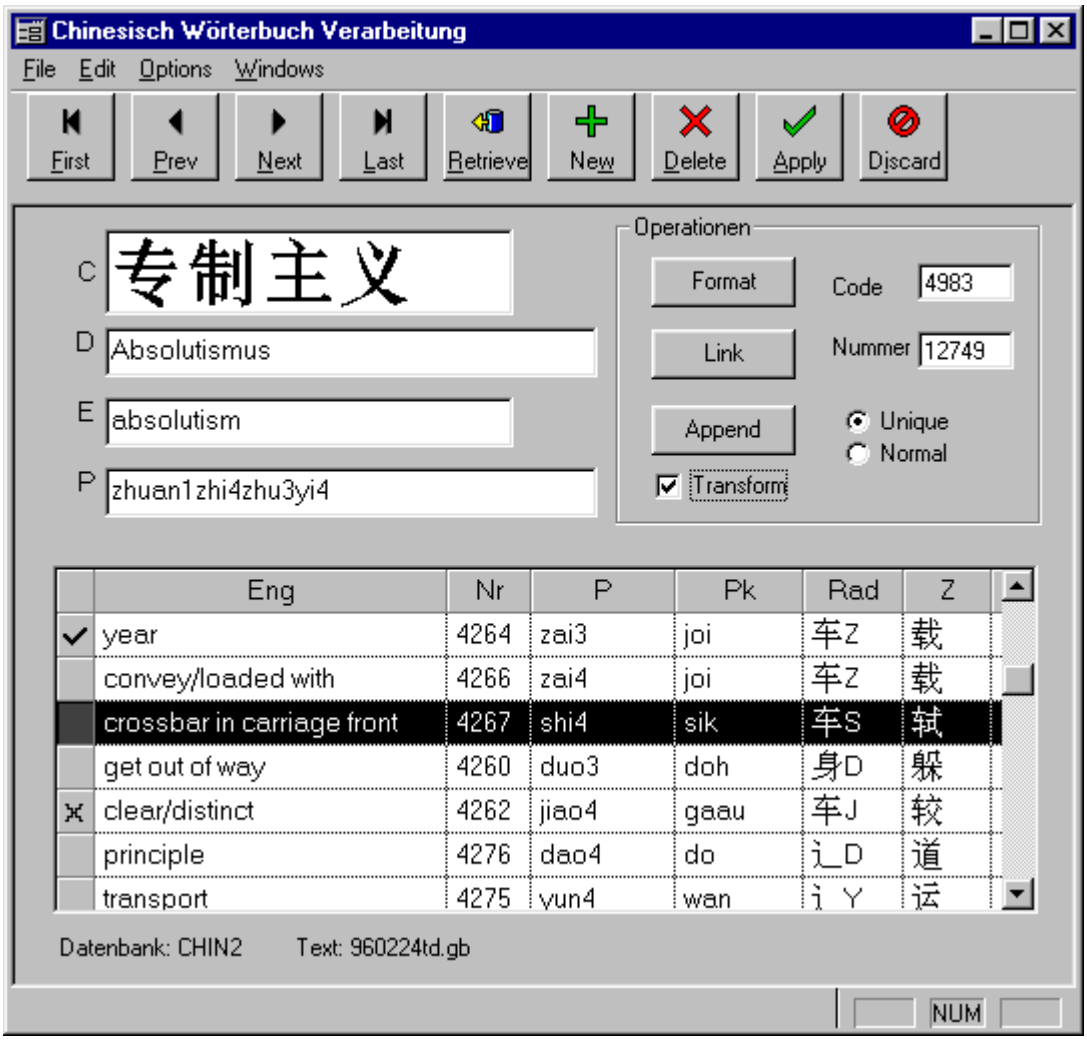


Abbildung 3.2: Wörterbuchverarbeitung mit Benutzeroberfläche

3.2.4 Speicherdatenbank

Die SQL-Datenbank war während der Datensammelphase und der Vorverarbeitung gut einsetzbar. Aber zur weiteren Datenbearbeitung und zur Analyse von Romanisierungsverfahren zeigten sich verschiedene Probleme beim weiteren Einsatz der Datenbank. Denn eine relationale Datenbank ist mehr darauf spezialisiert wenige, komplexe und umfangreiche Anfragen auszuführen, als tausende einfacher Anfragen pro Sekunde zu bearbeiten. Neben diesem Geschwindigkeitsproblem konnten auch einige Anfragen nicht effizient in SQL definiert werden, was den Einsatz eines Embedded SQL Systems erfordert hätte. Dies wiederum hätte eine Portierung auf UNIX-Systeme erschwert.
Aus diesen Gründen wurde ein eigenes Datenbanksystem, welches ausschließlich Speicherstrukturen verwendet, entwickelt. Es ist damit speziell auf die Anforderung „sehr viele Anfragen auf sehr großen Datenbeständen in Echtzeit“ spezialisiert und durch die Implementierung als C++ Klasse lassen sich flexibel beliebige Suchanfragen und andere Operationen formulieren.

Leistungsdaten der Speicherdatenbank:

- ca. 100000 Suchanfragen pro Sekunde möglich (je nach Rechnertyp/Datenbankgrösse)
- beliebige Anzahl Relationen, Felder
- beliebige Feldgrösse (Maximalwert muß angegeben werden)
- kein ungenutzter Speicher, nur tatsächliche Feldgrösse wird allokiert.
- Standardoperationen: delete, insert, update, select
- beliebige Anzahl von frei definierbaren Mehrfachindizes pro Relation
- Jeder Index hat Dirty-Flag und wird nur bei Bedarf aktualisiert
- Sortierverfahren basiert auf effizienter QuickSort-Implementierung
- Suchverfahren basiert auf effizienter BinarySearch-Implementierung
- Vergleichsfunktion für Suchen/Sortieren kann überschrieben werden
- im Result-Set kann navigiert werden -> Joins können über Schleifenbearbeitung des Result-Set realisiert werden
- Relationen können importiert und exportiert werden, dabei können Trenn- und Begrenzungszeichen definiert werden
- Tabelleninformationen samt Indizes können mit importiert und exportiert werden
- Daten-Übernahme/Übergabe an SQL-Datenbank ist möglich
- Die Speicherstrukturen reorganisieren sich selbst, wenn definierbare Grenzwerte überschritten werden

Auf folgende Features wurde zwecks Performance und Einfachheit verzichtet:
- Transaktionsverwaltung
- Datenbankmanagement
- andere Datentypen als String/Binary
- eigenständige Auslagerung von Speicher auf Festplatte, um bei Hauptspeichermangel die Betriebsystems-Speicherauslagerung zu umgehen.
- SQL-Parser, Variablenbindungen, komplexe SQL-Befehle

Die Speicherdatenbank wurde eingesetzt, um Datenmanipulationen und spezielle Anfragen zur Analyse und Kontrolle der Daten vorzunehmen, die auf einer SQL-Datenbank nicht effizient möglich gewesen wären. Das C-Programm Manipulate umfaßt mehrere solcher Routinen, die auch Testverfahren und Konsitenzprüfung enthalten. Neben dieser Routinensammlung baut auch das Romanisierungssystem „Pinyin“ auf der Speicherdatenbank auf.

3.2.5 TCL-Tools

Die frei verfügbare Sprache TCL ist eine interpretierte Skriptsprache mit leistungsfähigen Textverarbeitungsfunktionen, die für alle gängigen Plattformen erhältlich ist. Diese Vorteile haben dazu beigetragen, daß auch einige Tools mit TCL erstellt wurden. Zwei dieser Tools werden in den folgenden Unterkapiteln kurz vorgestellt.

3.2.5.1 Konvert

Das Skript **Konvert.tcl** konvertiert spezielle HTML-Seiten von Voice of America (VOA) in eine reine Textform. Dabei werden alle HTML-Kommandos ignoriert und nur die in Tabellendarstellung vorhandenen chinesischen Texte mit ihren Pinyin-Transkriptionen extrahiert.

Abbildung 3.3: Website mit Pinyintranskription

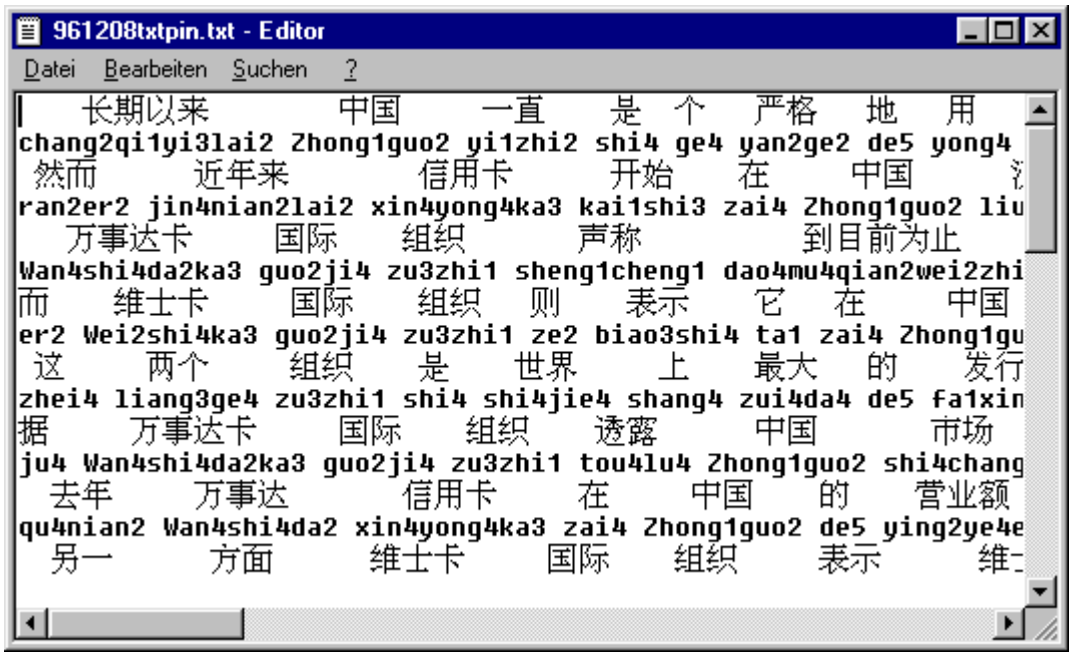


Abbildung 3.4: In Plaintext umgesetztes HTML

Das Skript **Konvert.tcl** erwartet neben optionalen Flags die Eingabedatei (HTML) und die Ausgabedatei (Plaintext) als Parameter.
Folgende Optionen/Flags können angegeben werden:

- z: chinesische Zeichen werden ausgegeben (s.o. Abb.)
- p: Pinyin wird ausgegeben (s.o. Abb.)
- f: Zeichenbreite mittels Leerzeichen an Pinyin anpassen (s.o. Abb.)
- l: Zeilen durch Leerzeile trennen
- s: Synopsis (englischer Kommentar) wird mit ausgeben
- u: Zeilenumbruch nach n Zeichen (z.B. u80)
- r: Pinyin vor Zeichen ausgeben (sonst Zeichen vor Pinyin)
- n: Zeilennummern ausgeben
- o: Satzzeichen ignorieren (s.o.Abb.)
- v: Testdaten generieren (nur Zeichenfolge ohne Leerzeichen) zu Validierung des Pinyinsystems

3.2.5.2 Bigramm

Das Skript **Bigramm.tcl** erzeugt Bigramme und Unigramme nach der NIST-Definition. Neben der NIST-Form sind noch andere Formen möglich, besonders Bigramm-Zeichen-Pinyin-Zuordnung und Unigramm-Zeichen-Pinyin-Zuordnung, die in der Lautschriftumsetzung verwendet werden.

Folgende Optionen/Flags können angegeben werden:

- l: logarithmischen Score ausgeben
- x: Typauswahl (1 Unigramm, 2 Bigramm, 0 NIST-Form)
- d: Doppeltzeile (Zeichenzeile dann Pinyinzeile)

3.2.6 Handarbeit

Neben dem Einsatz verschiedener Hilfstools, ist immer wieder sehr viel Handarbeit notwendig, um die Qualität der Daten auf ein akzeptables Maß zu heben. Neben der systematischen Korrektur der Daten, ist es möglich, daß das Pinyinsytem seine automatische Romanisierung mit einer von Hand vorgenommenen Romanisierung vergleicht und ein Fehlerprotokoll aller Abweichungen generiert. Dieses Fehlerprotokoll muß dann abgearbeitet werden, wobei dann alle Fehler, die auf Grund von Fehlern in der Datenbasis auftraten, in der Datenbank korrigiert werden müssen. Mittels dieses Vorgehens kann das Pinyinsytem sukzessive verbessert werden.
Dabei können 2 Typen von Fehlern pro chinesisches Wort auftreten:

- Wort ist überhaupt nicht in der Datenbasis
- Eine oder mehrere Pinyinabbildung pro Wort sind falsch, fehlen oder sind zuviel

Diese Fehler müssen von Hand korrigiert werden, da keine zuverlässige, computerbasierte Datenquelle verfügbar ist. Oftmals bedeutet dies, daß in einem chinesischen Wörterbuch zeitaufwendig nachgeschlagen werden muß. Dabei wurden hauptsächlich die Wörterbücher [1], [2], [3], [4], [5], [10] und [14] verwendet.

3.2.6.1 Wörterbuchsuche

Da die chinesische Schrift kein Alphabet als Grundlage hat, richtet sich die Anordnung der Wörter in einem Wörterbuch nach der zugehörigen Pinyin-Transkription. Wobei das chinesische Wort für jedes zugehörige Pinyin an verschiedenen Stellen im Wörterbuch meist mit unterschiedlichen Bedeutungen auftaucht. Die Zeichen mit gleichen Pinyin werden dann nach Ton und nach Häufigkeit des Gebrauchs sortiert aufgelistet. Sobald also die Pinyin-Transkription bekannt ist, gestaltet sich die Suche einfach. Wenn man aber zu einem chinesischen Wort dessen Aussprache suchen möchte, gestaltet sich sie Suche aufwendiger, denn auf der Menge der chinesischen Zeichen lässt sich nur schwer eine Ordnungsstruktur aufprägen. Sehr viele Forschungsarbeiten befassen sich mit diesem Thema und es sind sehr viele Ansätze diskutiert worden, von denen sich einige durchsetzen konnten, und je nach Aufgabenstellung nebeneinander verwendet werden. Für die Suche in den verschiedensten Wörterbüchern hat sich ein Verfahren behauptet, bei dem versucht wird die Zeichen nach Strukturmerkmalen (Radikalen) in Klassen einzuteilen. Je nach Wörterbuch werden so ca. 200 - 240 Klassen definiert, wobei eine Resteklasse alle nicht zuordbaren Zeichen enthält. Um nun ein Zeichen im Wörterbuch zu finden wird zunächst die zugehörige Klasse bestimmt, welche einen Verweis auf eine Liste aller zugehörigen Zeichen enthält. Diese Liste ist nach der Anzahl der im Zeichen verwendeten Striche sortiert, so daß ein schnelles Auffinden möglich wird. Zusätzlich enthält diese Liste pro Zeichen jeweils die Seitennummern, unter der es im Wörterbuch einen Eintrag enthält.

Beispiel: Gesucht werden soll die Pinyin-Transkription von 行李 (Gepäck) [1]:

1. Bestimmen des Radikals von 行: Hier eindeutig 彳 (da der rechte Teil von 行 nicht als radikal existiert)
2. Bestimmen der Strichanzahl des Radikals 彳 : 3 Striche

3. Suchen des Radikals in der Radikaltabelle unter 3 Strichen:

一 画		35	又	70	彐(彐⺕)	105	中	140	业	175	缶	209	金
1	丶	36	廴	71	弓	106	贝	141	目	176	来	210	鱼
2	一	37	厶	72	己(巳)	107	见	142	田	177	舌	**九 画**	
3	丨	38	口	73	女	108	父	143	由	178	竹(⺮)	211	音
4	丿	39	匕	74	子(孑)	109	气	144	申	179	臼	212	革
5	亠	**三 画**		75	马	110	牛(⺧)	145	罒	180	自	213	是
6	冂	40	氵	76	幺	111	手	146	皿	181	血	214	骨
7	乙(乛乚)	41	忄	77	纟(糸)	112	毛	147	钅	182	舟	215	香
二 画		42	犭(⺨)	78	巛	113	攵	148	矢	183	羽	216	鬼
8	亻	43	亡	79	小(⺌)	114	片	149	禾	184	艮(⾒)	217	食
9	亠	44	广	**四 画**		115	斤	150	白	**七 画**		**十 画**	
10	讠	45	氵	80	灬	116	爪(⺥)	151	瓜	185	言	218	高
11	二	46	门	81	心	117	尺	152	鸟	186	辛	219	鬲
12	十	47	辶	82	斗	118	月	153	皮	187	辰	220	髟
13	厂	48	工	83	火	119	殳	154	癶	188	麦	**十一画**	
14	扩	49	土(士)	84	文	120	欠	155	矛	189	走	221	麻
15	匚	50	艹	85	方	121	风	156	疋	190	赤	222	鹿
16	卜(卜)	51	廾	86	户	122	氏	**六 画**		191	豆	**十二画**	
17	阝	52	大	87	礻	123	比	157	羊(羊⺶)	192	束	223	黑
18	冖	53	尢	88	王	124	丰	158	类	193	酉	**十三画**	
19	冂	54	寸	89	主	125	水	159	米	194	豕	224	鼓
20	卩	55	扌	90	天(夭)	**五 画**		160	齐	195	里	225	鼠
21	亻	56	弋	91	韦	126	立	161	衣	196	足	**十四画**	
22	厂	57	巾	92	歹	127	疒	162	亦(亦)	197	采	226	鼻
23	人(入)	58	口	93	廿(⺾)	128	穴	163	耳	198	豸	—	
24	八(丷)	59	口	94	木	129	礻	164	臣	199	谷	227	余类
25	乂	60	山	95	不	130	夫	165	戋	200	身		
26	勹	61	中	96	犬	131	玉	166	西(西)	201	角		
27	刀(刂)	62	礻	97	歹	132	示	167	束	**八 画**			
28	力	63	彡	98	瓦	133	去	168	亚	202	青		
29	儿	64	夕	99	牙	134	艹	169	而	203	卓		
30	几(几)	65	夂	100	车	135	甘	170	页	204	雨		
31	宀	66	丸	101	戈	136	石	171	至	205	非		
32	冂	67	尸	102	止	137	龙	172	光	206	齿		
33	阝(在左)	68	忄	103	日	138	戊	173	虍	207	龟		
34	阝(在右)	69	氵	104	曰	139	灬	174	虫	208	隹		

Radikale mit 3 Strichen

Die gesuchte Klassennummer ist 62

Abbildung 3.5: Radikaltabelle

4. Suchen des Zeichens 行 in der Zeichenliste 62

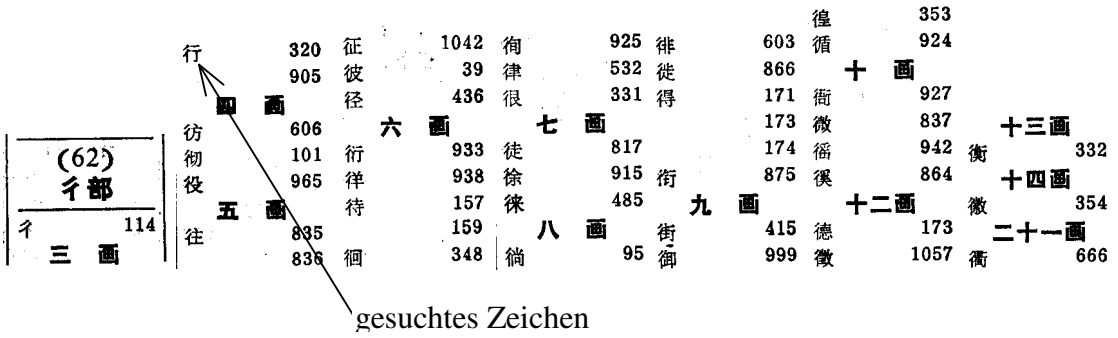
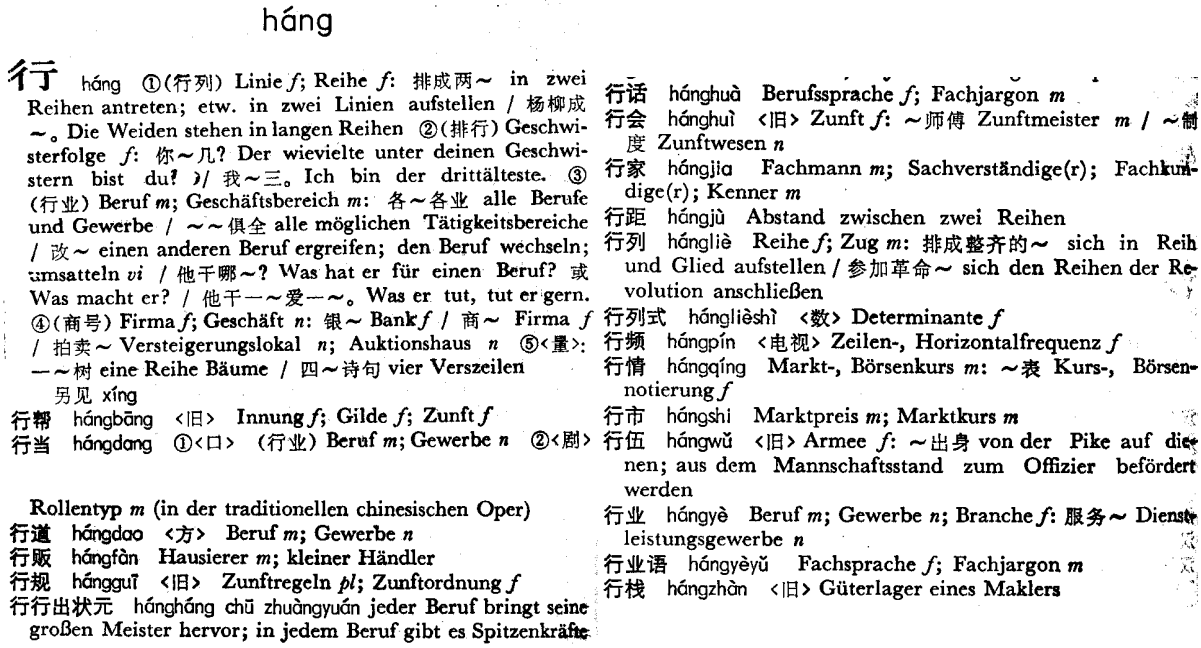


Abbildung 3.6: Zeichenliste zum Radikal 彳 (62)

5. Nachschlagen unter der gegebenen Seitennummer
a) Seite 320:

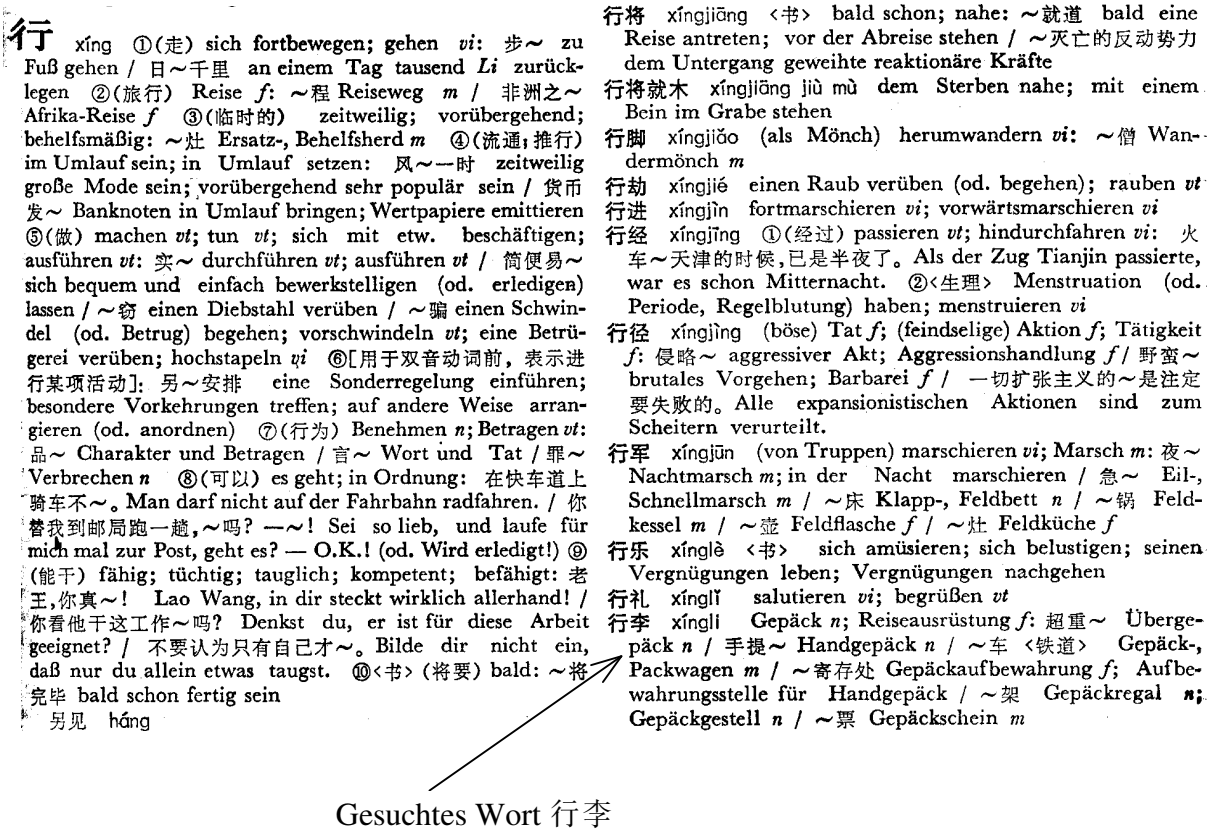

háng

行 háng ①(行列) Linie *f*; Reihe *f*: 排成两～ in zwei Reihen antreten; etw. in zwei Linien aufstellen / 杨柳成～。Die Weiden stehen in langen Reihen ②(排行) Geschwisterfolge *f*: 你～几? Der wievielte unter deinen Geschwistern bist du?)/ 我～三。Ich bin der drittälteste. ③(行业) Beruf *m*; Geschäftsbereich *m*: 各～各业 alle Berufe und Gewerbe / ～～俱全 alle möglichen Tätigkeitsbereiche / 改～ einen anderen Beruf ergreifen; den Beruf wechseln; umsatteln *vi* / 他干哪～? Was hat er für einen Beruf? 或 Was macht er? / 他干一～爱一～。Was er tut, tut er gern. ④(商号) Firma *f*; Geschäft *n*: 银～ Bank *f* / 商～ Firma *f* / 拍卖～ Versteigerungslokal *n*; Auktionshaus *n* ⑤<量>: 一～树 eine Reihe Bäume / 四～诗句 vier Verszeilen
另见 xíng
行帮 hángbāng <旧> Innung *f*; Gilde *f*; Zunft *f*
行当 hángdang ①<口> (行业) Beruf *m*; Gewerbe *n* ②<剧> Rollentyp *m* (in der traditionellen chinesischen Oper)
行道 hángdao <方> Beruf *m*; Gewerbe *n*
行贩 hángfàn Hausierer *m*; kleiner Händler
行规 hángguī <旧> Zunftregeln *pl*; Zunftordnung *f*
行行出状元 hánghángchū zhuàngyuán jeder Beruf bringt seine großen Meister hervor; in jedem Beruf gibt es Spitzenkräfte
行话 hánghuà Berufssprache *f*; Fachjargon *m*
行会 hánghuì <旧> Zunft *f*: ～师傅 Zunftmeister *m* / ～制度 Zunftwesen *n*
行家 hángjia Fachmann *m*; Sachverständige(r); Fachkundige(r); Kenner *m*
行距 hángjù Abstand zwischen zwei Reihen
行列 hángliè Reihe *f*; Zug *m*: 排成整齐的～ sich in Reih und Glied aufstellen / 参加革命～ sich den Reihen der Revolution anschließen
行列式 hánglièshì <数> Determinante *f*
行频 hángpín <电视> Zeilen-, Horizontalfrequenz *f*
行情 hángqíng Markt-, Börsenkurs *m*: ～表 Kurs-, Börsennotierung *f*
行市 hángshi Marktpreis *m*; Marktkurs *m*
行伍 hángwǔ <旧> Armee *f*: ～出身 von der Pike auf dienen; aus dem Mannschaftsstand zum Offizier befördert werden
行业 hángyè Beruf *m*; Gewerbe *n*; Branche *f*: 服务～ Dienstleistungsgewerbe *n*
行业语 hángyèyǔ Fachsprache *f*; Fachjargon *m*
行栈 hángzhàn <旧> Güterlager eines Maklers

Abbildung 3.7: Eintrag zu 行 auf Seite 320

⇨ kein Eintrag zu 行李 vorhanden.

b) Seite 905



行 xíng ①(走) sich fortbewegen; gehen *vi*: 步～ zu Fuß gehen / 日～千里 an einem Tag tausend *Li* zurücklegen ②(旅行) Reise *f*: ～程 Reiseweg *m* / 非洲之～ Afrika-Reise *f* ③(临时的) zeitweilig; vorübergehend; behelfsmäßig: ～灶 Ersatz-, Behelfsherd *m* ④(流通;推行) im Umlauf sein; in Umlauf setzen: 风～一时 zeitweilig große Mode sein; vorübergehend sehr populär sein / 货币发～ Banknoten in Umlauf bringen; Wertpapiere emittieren ⑤(做) machen *vt*; tun *vt*; sich mit etw. beschäftigen; ausführen *vt*: 实～ durchführen *vt*; ausführen *vt* / 简便易～ sich bequem und einfach bewerkstelligen (od. erledigen) lassen / ～窃 einen Diebstahl verüben / ～骗 einen Schwindel (od. Betrug) begehen; vorschwindeln *vt*; eine Betrügerei verüben; hochstapeln *vi* ⑥[用于双音动词前，表示进行某项活动]: 另～安排 eine Sonderregelung einführen; besondere Vorkehrungen treffen; auf andere Weise arrangieren (od. anordnen) ⑦(行为) Benehmen *n*; Betragen *vt*: 品～ Charakter und Betragen / 言～ Wort und Tat / 罪～ Verbrechen *n* ⑧(可以) es geht; in Ordnung: 在快车道上骑车不～。Man darf nicht auf der Fahrbahn radfahren. / 你替我到邮局跑一趟，～吗？—～！ Sei so lieb, und laufe für mich mal zur Post, geht es? — O.K.! (od. Wird erledigt!) ⑨(能干) fähig; tüchtig; tauglich; kompetent; befähigt: 老王，你真～！ Lao Wang, in dir steckt wirklich allerhand! / 你看他干这工作～吗？ Denkst du, er ist für diese Arbeit geeignet? / 不要认为只有自己才～。Bilde dir nicht ein, daß nur du allein etwas taugst. ⑩〈书〉(将要) bald: ～将完毕 bald schon fertig sein

另见 háng

行将 xíngjiāng 〈书〉 bald schon; nahe: ～就道 bald eine Reise antreten; vor der Abreise stehen / ～灭亡的反动势力 dem Untergang geweihte reaktionäre Kräfte

行将就木 xíngjiāng jiù mù dem Sterben nahe; mit einem Bein im Grabe stehen

行脚 xíngjiǎo (als Mönch) herumwandern *vi*: ～僧 Wandermönch *m*

行劫 xíngjié einen Raub verüben (od. begehen); rauben *vt*

行进 xíngjìn fortmarschieren *vi*; vorwärtsmarschieren *vi*

行经 xíngjīng ①(经过) passieren *vt*; hindurchfahren *vi*: 火车～天津的时候，已是半夜了。Als der Zug Tianjin passierte, war es schon Mitternacht. ②〈生理〉 Menstruation (od. Periode, Regelblutung) haben; menstruieren *vi*

行径 xíngjìng (böse) Tat *f*; (feindselige) Aktion *f*; Tätigkeit *f*: 侵略～ aggressiver Akt; Aggressionshandlung *f* / 野蛮～ brutales Vorgehen; Barbarei *f* / 一切扩张主义的～是注定要失败的。Alle expansionistischen Aktionen sind zum Scheitern verurteilt.

行军 xíngjūn (von Truppen) marschieren *vi*; Marsch *m*: 夜～ Nachtmarsch *m*; in der Nacht marschieren / 急～ Eil-, Schnellmarsch *m* / ～床 Klapp-, Feldbett *n* / ～锅 Feldkessel *m* / ～壶 Feldflasche *f* / ～灶 Feldküche *f*

行乐 xínglè 〈书〉 sich amüsieren; sich belustigen; seinen Vergnügungen leben; Vergnügungen nachgehen

行礼 xínglǐ salutieren *vi*; begrüßen *vt*

行李 xíngli Gepäck *n*; Reiseausrüstung *f*: 超重～ Übergepäck *n* / 手提～ Handgepäck *n* / ～车 〈铁道〉 Gepäck-, Packwagen *m* / ～寄存处 Gepäckaufbewahrung *f*; Aufbewahrungsstelle für Handgepäck / ～架 Gepäckregal *n*; Gepäckgestell *n* / ～票 Gepäckschein *m*

Gesuchtes Wort 行李

Abbildung 3.8: Eintrag zu 行 auf Seite 905 (Auszug)

⇨ die Pinyin-Transkription von 行李 lautet „xing2li5“ !

Das Nachschlagen von Pinyin-Transkriptionen bei gegebenen chinesischem Wort in einem Wörterbuch wurde hier ausführlich dargestellt, da es einen nicht unerheblichen zeitlichen Anteil an der Studienarbeit hatte.

Das sorgfältige Datensammeln und die Datenaufarbeitung sind von enormer Wichtigkeit und eine Verbesserung des in dieser Arbeit vorgestellten Verfahrens hat nur Sinn, wenn die Datenbasis eine genügend hohe Güte besitzt.

3.3 Datensatzbeschreibungen

Im folgenden werden die einzelnen erzeugten und im Pinyinprogramm verwendeten Datenrelationen beschrieben:

3.3.1 Wörterbuch

Das Wörterbuch CW_PY enthält die Zuordnung von einem chinesischen Wort in Schriftzeichen zu seiner Pinyin-Transkription. Für jede Zuordnung existiert ein Eintrag. Das Wörterbuch hat ca. 90 000 Einträge.

Auszug:
er5nv3qing2chang2;儿女情长;
er5nv3zhai4;儿女债;
er5shi2;儿时;
er5sun4;儿孙;
er5sun4zi4you3er5sun4fu2;儿孙自有儿孙福;
er5tong2;儿童;
er5tong2du2wu4;儿童读物;
er5tong2jie2;儿童节;
er5tong2jie2mu4;儿童节目;

3.3.2 Wortwahrscheinlichkeit

Die Relation CWS enthält für ca. 30 000 Worte deren Auftrittshäufigkeit in einem Korpus von ca. 5,4 Mio. Worten. Die Wortwahrscheinlichkeit erhält man, indem man die Auftrittshäufigkeit durch die Korpusgrösse teilt. Die Verwendung der Auftrittshäufigkeit hat Geschwindigkeitsvorteile und vermindert Rundungsfehler.

Auszug:
群众;7527
现在;7507
方面;7352
它的;7241
运动;7192
这点;7132
提高;7123

3.3.3 Zeichenwahrscheinlichkeit

Die Relation ZWS enthält für ca. 7 000 Zeichen deren Auftrittshäufigkeit aus einem Korpus von ca. 162 Mio. Zeichen. Die Zeichenwahrscheinlichkeit erhält man indem man die Auftrittshäufigkeit durch die Korpusgrösse teilt. Die Verwendung der Auftrittshäufigkeit hat Geschwindigkeitsvorteile und vermindert Rundungsfehler.

Auszug:
败;38512
斑;2623
班;68002
扳;1356
般;90599
颁;2815
版;171116

3.3.4 Pinyin-Wortzuordnungs-Unigramm

Das Pinyin-Wortzuordnung Unigramm gibt die Wahrscheinlichkeit an, mit der ein Wort auftritt und auf eine spezielle Pinyin-Transkription abgebildet wird. Die Relation enthält die 3000 wichtigsten Wörter, mehr wortgetrenntes, pinyinumgesetztes Datenmaterial von VOA war zu diesem Zeitpunkt nicht verfügbar.

Auszug:
Auszug:
1;冰山;bing1shan1
2;病菌;bing4jun1
16;并;bing4
14;并且;bing4qie3
1;播出;bo1chu1
2;波恩;Boen
7;博吉尔;Bo2Ji2er3
1;博士;bo2shi4
3;伯恩斯;Bo2en1si1
72;不;bu4
2;不安;bu4an1

3.3.5 Pinyin-Wortzuordnungs-Bigramm

Das Pinyin-Wortzuordnungs-Bigramm gibt die Wahrscheinlichkeit an, mit der ein Wort nach einem gegebenen Wort mit Pinyin-Transkription auftritt und auf eine spezielle Pinyin-Transkription abgebildet wird. Die Relation enthält nur ca. 10000 solcher Bigramme, da nicht mehr wortgetrenntes, pinyinumgesetztes Datenmaterial von VOA verfügbar war.

2;报道;bao4dao4;了;le5
2;报道;bao4dao4;美国;Mei3guo2
20;报道;bao4dao4;说;shuo1
1;报道;bao4dao4;台北;Tai2bei3
2;报道;bao4dao4;台湾;Tai2wan1
1;报道;bao4dao4;太;tai4
4;报道;bao4dao4;援引;yuan2yin3
1;报道;bao4dao4;中;zhong1
6;表示;biao3shi4;美国;Mei3guo2
1;表示;biao3shi4;日本;Ri4ben3
2;表示;biao3shi4;如果;ru2guo3
2;表示;biao3shi4;他;ta1

4 Analyse und Entwurf eines Romanisierungssystems

Das einzige, mir bekannte Programm, welches eine Umsetzung eines Textes nach Pinyin vornimmt, ist „C2T“ von Tommi Kaikkonen aus Finnland (1992). Allerdings nimmt dieses Programm keine Worttrennung vor und bildet mittels einer einfachen Tabelle jedes einzelne Zeichen auf eine Pinyin-Transkription ab. Bei einem gewissen Prozentanteil ist diese Abbildung allerdings, wie bereits beschrieben, mehrdeutig und ohne statistische Informationen und Kontextwissen nicht auflösbar. Das „C2T“- Programm hat in diesem Fall nach Worttrennung eine Wortfehlerrate von ca. 20%, da es bei Mehrdeutigkeiten die Entscheidungen sehr willkürlich treffen muß. Um eine Pinyin-Transkription, die den Anforderungen in der Sprachverarbeitung genügt, zu erreichen, ist eine höhere Umsetzungsqualität und Worttrennung erforderlich.
Das zu diesem Zwecke entworfene Pinyinprogramm versucht diese beiden Hauptforderungen zu erreichen. Man sollte dabei beachten, daß ähnlich wie bei der maschinellen Übersetzung manche Entscheidungen trotz statistischer Informationen und Kontextwissen nicht ohne semantisches und pragmatisches Wissen korrekt erfolgen können und somit eine bestimmte Fehlerrate ohne Einbeziehung dieses Wissens nicht unterschritten werden kann (< 1%).
In der Analysephase wurde untersucht inwieweit die beiden Aufgaben Worttrennung und Lautschriftumsetzung voneinander entkoppelt werden können. Anhand von Satzbeispielen (s. Unterkapitel Mehrdeutigkeiten) und Untersuchung von grammatischen Eigenschaften ([6] [7]) zeigte sich, daß die Lautschriftumsetzung am besten auf der Worttrennung aufsetzt, da nur eine vernachlässigbar geringe Abhängigkeit der Worttrennung von Lautschriftumsetzung besteht. Es ergibt dadurch folgender Aufbau des Pinyinprogramms.

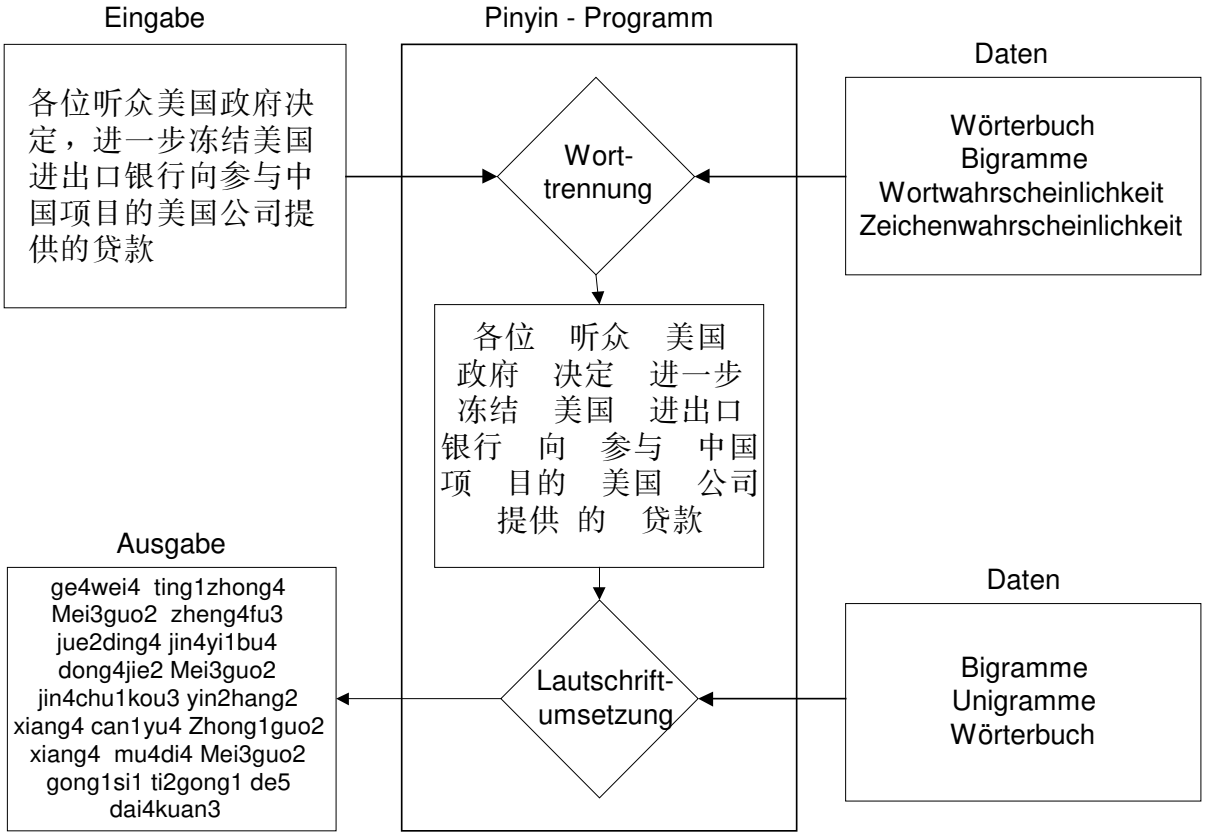


Abbildung 4.1: Blockdiagramm des Pinyinprogramms

4.1 Brute-Force Ansatz

Als erster Ansatz bot sich folgende Vorgehensweise an, bei der Worttrennung und Lautschriftumsetzung völlig unabhängig betrachtet werden:

a) Worttrennung: Immer nach erster im Wörterbuch enthaltener Zeichenkette trennen
   ⇨ Ergebnisse der Wortfehlerrate liegen bei 30% - 40% (zuviel/falsch getrennt)
b) Lautschriftumsetzung: Zeichen einzeln mittels Tabelle auf Pinyin abbilden (wie C2T)
   ⇨ Ergebnisse der Pinyinfehlerrate liegen bei ca. 20% (bezogen auf Wortfehler statt Zeichenfehler)

Nach diesem ersten Ansatz wurden Möglichkeiten gesucht und getestet die hohen Fehlerraten zu verbessern. Mögliche Verbesserung wurden dabei unter Einsatz von statistischen Daten und Kontextinformationen in Form von speziellen Bigrammen erprobt. Grammatiken wurden wegen ihrer Komplexität und ihrem meist unflexiblen Verhalten bei unbekannten grammatischen Strukturen nicht in Erwägung gezogen. Der Einsatz von Trigrammen kam aufgrund von nicht ausreichender von Hand nach Pinyin transkriptierter Texte auch nicht in Frage.

4.2 Mehrdeutigkeiten

Die hohe Fehlerrate des Brute-Force Ansatzes folgt daraus, daß besonders 2 Arten von Mehrdeutigkeiten auftreten.
Die hohe Fehlerrate des Brute-Force Ansatzes folgt daraus, daß besonders 2 Arten von Mehrdeutigkeiten auftreten.

1. Mehrdeutigkeiten bei der Worttrennung:
   Eine chinesische Phrase ist nicht immer einfach in Worte zu trennen, da die meisten, aus mehreren Zeichen bestehenden Worte, sich weiter in existierende Worte unterteilen lassen. Als Verbesserung des Brute-Force Ansatzes könnte man in Betracht ziehen, immer das längste im Wörterbuch vorhandene Wort als Hypothese zu nehmen, da es im allgemeinen wahrscheinlicher ist, als daß die Teilworte aufeinanderfolgen. Allerdings können sich durchaus sinnvolle Konstrukte ergeben, wenn die Worte weiter unterteilt werden.
   Z.B.:
   生气 (sheng1qi4) sich ärgern oder 生 气 leben Luft
   交通 (jiao1tong1) Verkehr oder 交 通 abgeben Verbindung
   卫生 (wei4sheng1) Hygiene oder 卫 生 beschützen Leben
   对不起 (dui4bu4qi3) Entschuldigung oder 对 不 起 richtig nicht aufstehen
   不好意思 (bu4hao3yi4si5) schüchtern oder 不 好 意思 nicht gute Bedeutung

   Da es in der chinesischen Sprache weder Konjunktion noch Deklination gibt und viele Zeichen mehrere Wortarten und Bedeutungen darstellen können, gibt es oft viele Möglichkeiten wie ein Satz/Phrase interpretiert werden kann. Dies kann beispielsweise dann geschehen, wenn bei zwei aufeinanderfolgenden Worten das letzte Zeichen des ersten Wortes auch als erstes Zeichen des zweiten Wortes dienen kann. Weiter noch kann dieses „Zeichenverschieben“ sich über mehrere Worte hinziehen, so daß es nötig wird immer eine gesamte Phrase zu betrachten.
   Z.B.:
   他 陪同 学生 学习 中国 文学 (ta1 pei2tong2 xue2sheng1 xue2xi2 zhong1guo2 wen2xue2)
   Er begleitet die Studenten beim Lernen chinesischer Literatur.

他 陪 同学 生 学习 中 国文学 (ta1 pei2 tong2xue2 sheng1 xue2xi2 zhong1 guo2wen2 xue2)
Er begleitet Kameraden Leben lernen mittlere Landessprache studieren -> ohne Sinn
离 开车时间 还有 五 分钟 (li2 kai1che1 shi2jian1 hai2you3 wu3 fen1zhong1)
Bis zur Abfahrtszeit sind es noch 5 Minuten.
离开 车 时 间 还有 五 分钟 (li2kai1 che1 shi2 jian1 hai2you3 wu3 fen1zhong1)
Verlassen Auto Zeitraum Lücke noch 5 Minuten. -> ohne Sinn

2. Mehrdeutigkeiten bei der Pinyin-Transkription:
   Die meisten chinesischen Zeichen haben eine einzige Aussprachemöglichkeit, nur ca. 13% besitzen 1-4 weitere Aussprachevarianten. Allerdings sind darunter einige sehr häufig gebrauchte Zeichen, so daß um eine akzeptable Umsetzung zu erhalten, Kontext und statistische Informationen heranzuziehen sind. Wenn mehrdeutige Zeichen innerhalb eines Wortes auftreten, genügt es meist den Wortkontext anhand des Wörterbuches heranzuziehen, um die korrekte Transkription zu finden. Bei einzeln auftretenden Zeichen muß anhand der benachbarten Wörter eine korrekte Umsetzung gefunden werden. Falls dies nicht möglich ist, sollte die wahrscheinlichste Umsetzung zur Fehlerminimierung verwendet werden.

Z.B.:
行 hat die Aussprachevarianten: xing2 (gehen,bewegen), hang2(Linie, Reihe).
没 hat die Aussprachevarianten: mei2 (nicht), mo4 (tauchen)
单 hat die Aussprachevarianten: dan1 (einzeln), shan4 (‘Name’), chan2 (Hunnen).
乐 hat die Aussprachevarianten: le4 (Freude), yue4 (Musik)
和 hat die Aussprachevarianten: he2 (mild), he4 (einstimmen), huo2 (verrühren), huo4 (mischen), hu2 (‘Name’)
觉 hat die Aussprachevarianten: jue2 (empfinden, fühlen), jiao4 (Schlaf)
机会计 hat die Aussprachevarianten: ji1hui4 ji4 (die Möglichkeit kalkulieren), ji1 kuai4ji4 (maschinelles Buchhalten)

4.3 Funktionsweise des Pinyinprogramm

Das Pinyinprogramm unterteilt den Romanisierungsvorgang in 4 Phasen:

1. Satzzeichentrennung
   In der ersten Phase wird der zu transformierende Text in Teilsätze getrennt, indem alle chinesischen Schriftzeichen, die zwischen zwei Satzzeichen liegen, zusammengefaßt werden. Diese Teilsätze werden dann bei der Romanisierung als unabhängig betrachtet. Neben dieser Partitionierung werden auch die Satzzeichen und Sonderzeichen auf entsprechende Zeichen des ASCII-Zeichensatzes abgebildet.
   Menge der chin. Satz-/Sonderzeichen = { 。 ， 、 ； ： ？ ！)(「」 『』 “” 《》 }
   Menge der korrespondierenden ASCII-Zeichen = { . , ; : ? !) (' ' ' ' ' ' ' ' }
2. Wörterbuchtrennung
   Die Wörterbuchtrennung hat zur Aufgabe, die Teilsätze noch weiter aufzutrennen, da die Anzahl der zu betrachtenden erlaubten Trennungen exponentiell mit der Teilsatzlänge zunimmt. Schon Teilsätze mit ca. 30 Zeichen würden Millionen von erlaubten Trennungen ergeben, die einzeln zu untersuchen wären.
   Es muß also eine Stelle im Teilsatz gefunden werden, an der garantiert ein Wortübergang stattfindet. Wenn man von der Annahme ausgeht, daß alle Wörter im Wörterbuch enthalten sind, ist eine Trennung bei dem Zeichenpaar möglich, das in keinem Wort des Wörterbuchs vorkommt. Für jedes Zeichenpaar des Teilsatzes muß also geprüft werden, ob es in keinem

Wort des Wörterbuches vorkommt, wenn dies der Fall ist, kann an dieser Stelle der Teilsatz aufgesplittet werden. Das Wörterbuch hat mit über 90000 Wörtern eine Überdeckung von mehr als 99 % eines normalen Textes. Einzig Personen- und Städtenamen bereiten im allgemeinen Schwierigkeiten, da sie meist nicht im Wörterbuch enthalten sind, und deshalb in ihre einzelnen Bestandteile getrennt werden.
Dieser Suchvorgang ist sehr zeitaufwendig, denn es muß jedes Wort des Wörterbuchs untersucht werden. Um diesen Suchvorgang zu optimieren, wird dynamisch beim Start des Pinyinprogramms eine sortierte Relation mit allen Zeichenpaaren des Wörterbuches aufgebaut, somit kann das Nichtvorhandensein eines Zeichenpaares schnell mittels Binärsuche entschieden werden. Durch die dynamische Generierung bei jedem Programmstart, muß diese Relation bei Änderung des Wörterbuchs nicht nachgeführt werden und der Zeitaufwand des Einlesens von externen Datenträgern, welcher grösser als die automatische Generierung insgesamt ist, entfällt.

3. Bewertungstrennung
   a) Anhand des Wörterbuches werden alle möglichen Trennungen des betrachteten Teilsatzes rekursiv erzeugt und in eine Speicherdatenbank geschrieben.
   b) Dann wird jeder einzelne Teilsatz bewertet.
      Die Bewertung beruht auf einem 3-stufigen Back-off Verfahrens, bei dem von qualitativ hochwertigen Bewertungen durch Bigramme, bei Fehlen eines Datensatzes auf schwächer gewichtete Bewertungen durch die Wortwahrscheinlichkeit mit niedriger Qualität zurückgeschaltet wird. Bei fehlender Wortwahrscheinlichkeit wird weiter auf die Verwendung von Zeichenwahrscheinlichkeiten zurückgeschaltet.
      Bigramme $\Rightarrow$ Wortwahrscheinlichkeiten $\Rightarrow$ Zeichenwahrscheinlichkeiten
      Beim Zurückschalten nimmt die Bewertungsqualität ab, was durch eine geringere Gewichtung berücksichtigt wird. Die Backoff-Faktoren wurden dabei bestimmt, indem bei verschiedenen Tests auf denselben Texten des Testsets die Backoff-Faktoren variiert wurden. Es wurde dann schließlich der Backoff-Faktorensatz genommen, bei dem die Fehlerrate am kleinsten war.
      Die Zeichenwahrscheinlichkeiten existieren für nahezu alle GB-Zeichen, so daß immer ein aussagekräftiger Score berechnet werden kann.
   c) die beste Teilsatztrennung, d.h. die mit dem höchsten Score wird ausgewählt
4. Pinyin-Zuordnung
   Für jedes getrennte Wort wird unter Berücksichtigung der beiden benachbarten Worte, die Pinyin-Transkription ermittelt. Dabei wird wieder ein 3-stufiges auf Pinyinwort-Zeichenwort-Bigrammen, Pinyinwort-Zeichenwort-Unigrammen und einem Pinyin-Wörterbuch basierendes Back-off Verfahren eingesetzt. D.h. für ein mehrdeutiges Wort wird zuerst der linke und rechte Kontext (jeweils ein Wort) betrachtet, bei Vorhandensein von passenden Bigrammeinträgen wird das Wort entsprechend dem Bigramm mit dem grössten Score umgesetzt, wobei linkseitiges und rechtseitiges Bigramm gleich gewichtet wird. Falls kein Bigrammeintrag existiert, wird unter den passenden Unigrammeinträgen derjenige mit dem höchsten Score ausgewählt. Wenn auch hier kein Eintrag vorhanden ist, wird die erste passenden Umsetzung des Wörterbuches verwendet.

   Beispiel:

```
WortY soll im betrachteter Teilsatz: ... WortX WortY WortZ ...
in Pinyin umgesetzt werden.

Bigramm:
WortX -> PinyinumsetzungX, WortY -> PinyinumsetzungY1    ScoreY1
```

```
WortX -> PinyinumsetzungX, WortY -> PinyinumsetzungY2    ScoreY2
WortX -> PinyinumsetzungX, WortY -> PinyinumsetzungY3    ScoreY3
...
WortY -> PinyinumsetzungYn  , WortZ -> PinyinumsetzungZ  ScoreYn
WortY -> PinyinumsetzungYn+1, WortZ -> PinyinumsetzungZ  ScoreYn+2
WortY -> PinyinumsetzungYn+2, WortZ -> PinyinumsetzungZ  ScoreYn+3


Unigramm:
WortY -> PinyinumsetzungYm       ScoreYm
WortY -> PinyinumsetzungYm+1     ScoreYm+1
WortY -> PinyinumsetzungYm+2     ScoreYm+2
...

Wörterbuch:
WortY -> PinyinumsetzungYp       (ScoreYp = 0)

Wähle diejenige Pinyinumsetzung deren mit Backoff-Faktoren gewichteter
Score maximal ist.
```

4.4 Eigenschaften des Pinyinprogramm

Das Pinyinprogramm wurde so entworfen, daß die Ausgabeformatierung möglichst jedem Verwendungszweck genügt. So kann festgelegt werden, ob nur in Worte getrenntes Chinesisch oder nur Pinyin, oder beides zusammen ausgegeben werden soll. Wenn sowohl Pinyin, als auch Zeichenausgabe erfolgen soll, kann die Reihenfolge in der dies geschehen soll, bestimmt werden. Um dabei leicht einem chinesischen Wort das Pinyin zuordnen zu können, kann einen spezielle Formatierung verwendet werden, die die Zeichenzwischenräume mit Leerzeichen auffüllt, so daß das chinesische Wort immer genau über/unter der Pinyin-Transkription steht.
Im Chinesischen verwendetete und nur mit chinesisischen Zeichen darstellbare Satz- und Sonderzeichen können entfernt oder mittels einer Abbildungstabelle auf Zeichen des ASCII-Zeichensatzes abgebildet werden.
Am Anfang der Zeile stehende Zeichen wie Zeilennummern und dergleichen können unterdrückt werden.
Um mehrere Texte auf einmal bearbeiten zu können, können diese in einer Liste dem Pinyin - Programm übergeben werden. Mittels des Unix/DOS-Befehls ‘dir * > list ‘ kann auch eine Liste eines gesamten Verzeichnisses erzeugt werden und dem Pinyinprogramm übergeben werden.

Das Pinyinprogramm verzichtet auf eine grafische Oberfläche. Der Quellcode ist damit portabel. Das Pinyinprogramm wird über folgende Command Line Parameter gesteuert:

Pinyin.exe [-flags] In Out [Vgl/Liste]
- In : Eingabetext/Eingabeverzeichnis
- Out : Ausgabetext/Ausgabeverzeichnis
- Vgl : Vergleichstext
- Liste : Liste von Dateien die bearbeitet werden sollen
- Flags/Optionen:
  - s (S)atzzeichen auf ASCII-Zeichensatz abbilden
  - z vorhandene (Z)eilennummern nicht unterdrücken
  - f (F)ormatierte Ausgabe verwenden

c (C)hinesisch ausgegeben
p (P)inyin ausgegeben
r (r)everse Ausgabe (zuerst Pinyin dann Chinesisch)
v (V)alidate (Vergleich von zwei Dateien mit Fehlerprotokoll und Statistik)
n (N)oExecute (keine Romanisierung nur in Verbindung mit v sinnvoll)
d gesamtes (D)irectory/Liste bearbeiten
-> In, Out werden als Verzeichnisse interpretiert
-> Vgl/Liste wird als Liste interpretiert
uxxx (U)mbruch nach xxx ASCI-)Zeichen
x Uni(x) ->Pfadkonventionen
g Debu(g): Zwischendateien nicht löschen

Beispiele:

Pinyin.exe -dcp speech speech speech\list	Die Texte der Liste 'list' aus dem Verzeichnis speech romanisieren und in dasselbe Verzeichnis schreiben.
Pinyin.exe -fcps Text.in Text.out	Text.in romanisieren, formatierte Ausgabe, Satzzeichenabbildung

Pinyin -fcpgvx Test/970321txtpin.in Test/970321txtpin.out Test/970321txtpin.vgl

....

4.5 Implementierung

Zur Implementierung des Pinyinprogramms wurde die Programmierumgebung Visual C++ 4.0 von Microsoft unter Windows NT 4.0 / 95 verwendet. Visual C++ besitzt neben einem sehr schnellen Compiler, einen sehr leistungsfähigen Debugger, der es ermöglichte Variablenwerte chinesisch darzustellen und sogar mittels chinesischer Eingabeverfahren zu verändern.

Nach erfolgreicher Testphase wurde eine Portierung auf Unix in Angriff genommen.

Das Pinyinprogramms besitzt folgende Grundstruktur:

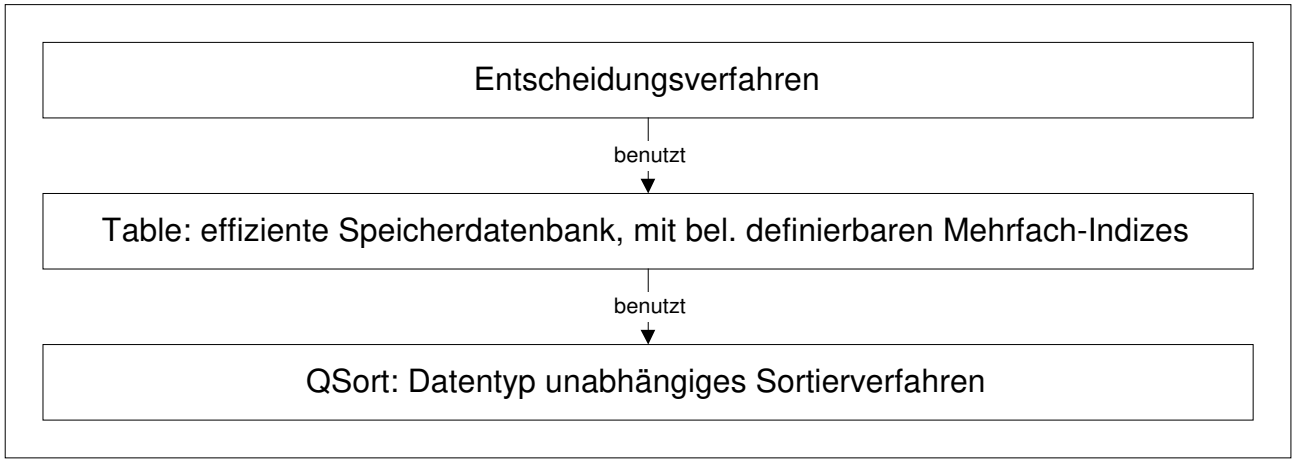


Abbildung 4.2: Grundstruktur des Pinyinprogramms

4.5.1 Ablaufdiagramm des Pinyinprogramms

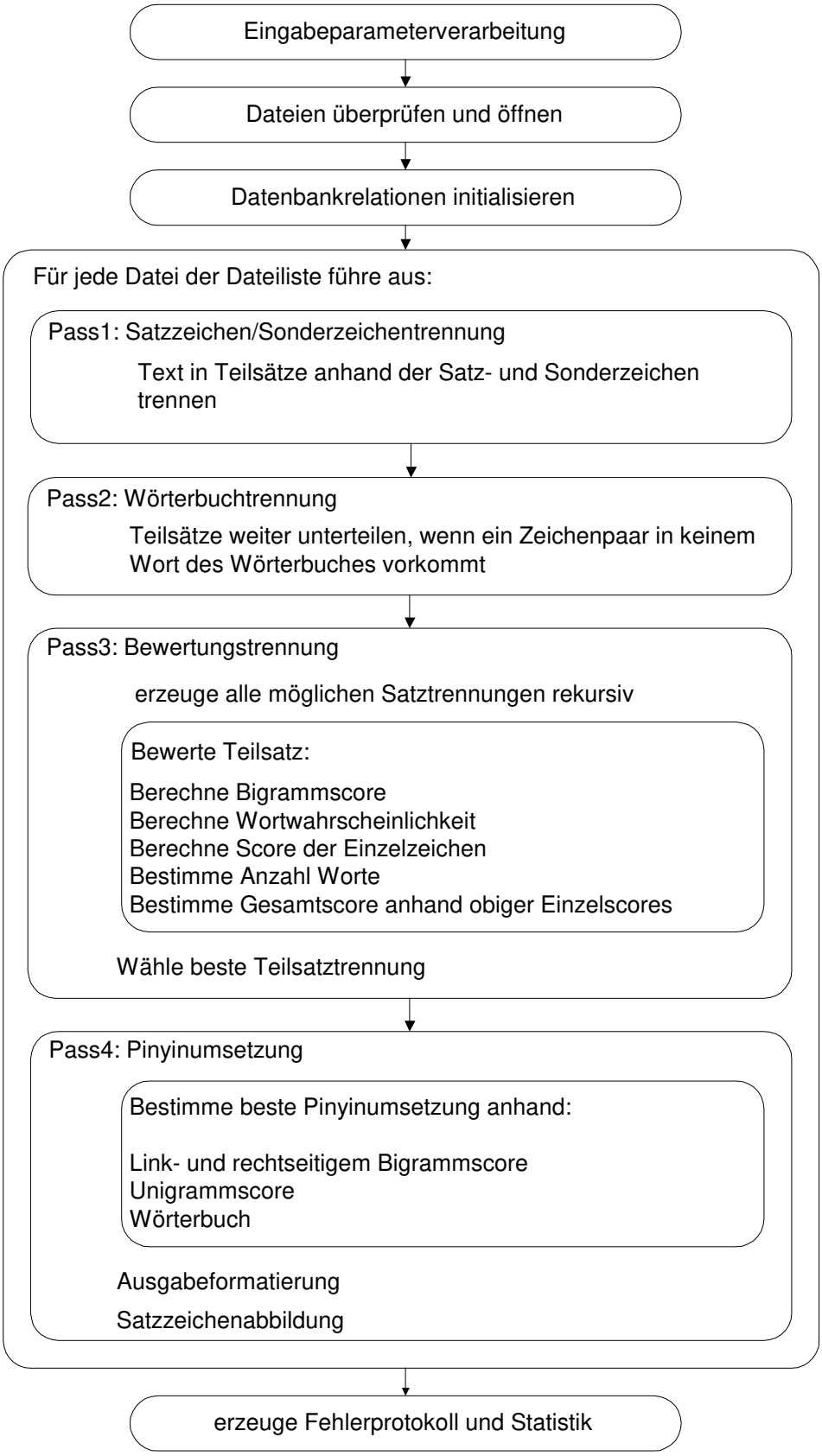


Abbildung 4.3: Ablaufdiagramm des Pinyinprogramms

Bemerkungen: Alle Suchfunktionen müssen speziell an 2 Bytecodes angepaßt werden, so daß die Suche immer nur an geraden Bytepositionen beginnen darf.
Nach jedem Pass wird eine Kontrolldatei geschrieben, so daß Fehler genau zurückverfolgt werden können.

4.6 Test und Validierung

Das Pinyinprogramm besitzt die Möglichkeit den automatisch romanisierten (in Worte getrennten und Pinyin umgesetzten) Text mit einem von Hand romanisierten Text zu vergleichen. In einer übersichtlichen Darstellung werden alle Abweichungen, sowohl in der Worttrennung als auch in der Pinyinumsetzung angezeigt. Wenn man davon ausgeht, daß die Romanisierung von Hand fehlerfrei war, hat man so ein Fehlerprotokoll erhalten mittels dessen die internen Datenstrukturen korrigiert werden können. Neben diesem Fehlerprotokoll erhält man auch einen Statistik über die Anzahl der Trennfehler, der Pinyinumsetzungsfehler und den entsprechenden Fehlerraten.

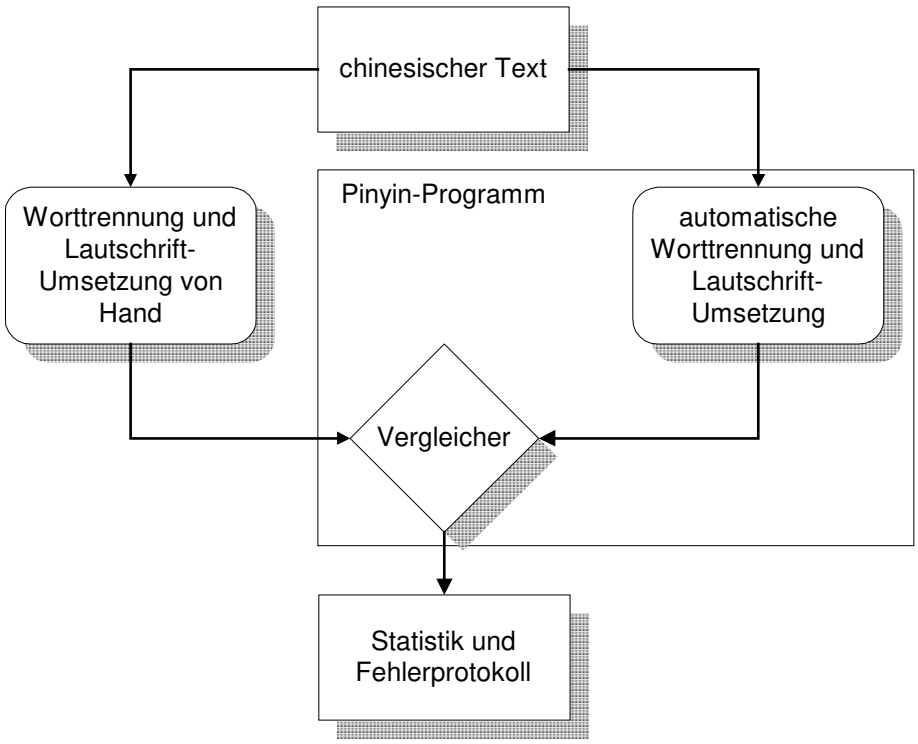


Abbildung 4.4: Test und Validierungsablauf

Fehlerprotokollbeispiel:

Korrekte Zeile:

```
  各位        听众        美国        政府        决定        进一步

ge4wei4 ting1zhong4 Mei3guo2 zheng4fu3 jue2ding4 jin4yi1bu4
```

Fehlerhafte Zeile:

```
  法新社        还报  道       说      进出口         银 行在
                还 报道                              银行   在
fa3xin1she4 hai2bao4 dao4 shuo1 jin4chu1kou3 yin2 xing2zai4
            hai2 bao4dao4 shuo4              yin2hang2  zai4
```

Trennfehler Pinyinfehler Trennfehler und Pinyinfehler

Validierung/Statistik-Beispiel:

Anzahl Woerter: 384
Anzahl Trennfehler: 18 (04.69%)
Anzahl Pinyinfehler: 8 (02.08%)
Trenn-Performance: (95.31%)
Pinyin-Performance: (97.92%)

4.7 Performance

Da das Pinyinprogramm eine Komponente zur Selbstvalidierung enthält, beschränkt sich der Ablauf der Performancemessung auf die Beschaffung eines Vergleichtextes, bei dem die Pinyinumsetzung und Worttrennung von Hand durchgeführt wurde. Da VOA(Voice-of-America)-Texten schon mit einer Romanisierung versehen waren, eigneten sie sich auch gut für Performanceuntersuchungen. Da VOA-Texte auch für das Training und das Testen des Pinyinsystems verwendet wurden, mußte um die Ergebnisse nicht zu beeinflussen, die vorhandene Menge von VOA-Texten in 3 Mengen: eine Trainingsmenge, eine Testmenge und eine Validierungsmenge unterteilt werden. Aus der Trainingsmenge wurden die Wort-Pinyinzuordnungs-Bigramme und die Wort-Pinyinzuordnungs-Unigramme, aufgrund denen das Programm unter anderem seine Entscheidungen trifft, extrahiert. Mittels der Testmenge wurden Fehler im Datenbestand und Algorithmus gesucht und verbessert. Die Validierungsmenge soll nun objektiv, ohne daß deren Daten schon vorher verwendet wurden, eine Bewertung des Systems ermöglichen.
Die mittels des Validierungsmoduls des Pinyinprogrammsso ermittelte Fehlerrate ist höher als die tatsächliche Anzahl Fehler, da im Chinesichen erlaubte Alternativen dort als Fehler be-

trachtet werden. Z.B. hat 这 die beiden korrekten Pinyintranskriptionen zhe4 und zhei4 mit jeweils gleicher Bedeutung oder 那 gar drei korrekte Pinyintranskriptionen na4, ne4 und nei4, welche Überbleibsel von verschiedenen Dialekten sind. Verschiedene Pinyintranskriptionen treten auch bei Eigennamen auf, wo z.B. manchmal gar keine Tonalität verwendet wird. Auch bei der Worttrennung sind manchmal Alternativen, wie schon im Kapitel 2 erwähnt, möglich. Aus diesen Gründen ist in den folgenden Tabellen, neben der vom Pinyinprogramm ausgegebenen Fehlerrate, auch eine bereinigte Fehlerrate angegeben.

Fehlerraten auf VOA-Texten:

	Trainingsmenge	Validierungsmenge	bereinigte Fehler auf Validierungsmenge
Worttrennung	2,6 %	5,4 %	4,3 %
Pinyinumsetzung	1,8 %	3,3 %	2,0 %

Bei VOA-Texten ist zu berücksichtigen, daß die Validierungsmenge aus der selben Datenquelle stammt wie die Trainingsdaten, und somit die Ergebnisse nur für VOA-Texte repräsentativ sind. Aus diesem Grunde wurde auch eine Validierung auf anderen Texten vorgenommen.

Fehlerraten auf „China Daily“-Texten:

	unbereinigt	bereinigt
Worttrennung	6,4 %	4,9 %
Pinyinumsetzung	3,4 %	2,1 %

Insgesamt ergibt sich eine je nach Text stark schwankende Fehlerrate. Besonders Eigennamen von Personen und Orten und komplexe, mehrdeutige Konstrukte bereiten der automatischen Romanisierung Probleme.

Bei den oben genannten Fehlerraten für das Pinyinprogramm, zeigt sich bei Untersuchung der Fehlerprotokolle, daß die auftretenden Fehler meist durch Fehleinträge in den mehr als 100000 zugrunde liegenden, maschinell erzeugten Datensätze verursacht wurden, und nicht durch Fehlentscheidungen aufgrund der verwendeten Methoden des Programms. Eine weitere Korrektur der Daten könnte somit die Gesamtfehlerrate wahrscheinlich unter 1-2 % senken.

Wortfehlerraten verschiedener Ansätze:

	C2T	Brute-Force	Brute-Force mit Verbesserungen	Pinyin
Worttrennung	---	ca. 40 %	ca. 8 %	4,9 %
Pinyinumsetzung	20,1 %	ca. 20 %	ca. 6 %	2,1 %

C2T: Pinyinumsetzung von Tommi Kaikkonen aus Finnland (1992).
Brute-Force: C2T erweitert um eine einfache Worttrennung anhand eines Wörterbuchs
Brute-Force mit Verbesserungen:

- Trennung immer nach längstem Wort im Wörterbuch
- Einbeziehung von Wahrscheinlichkeiten bei Worttrennung und Pinyinumsetzung
- Wortweise Pinyinumsetzung

Pinyin: Pinyinprogramm

5 Zusammenfassung und Ausblick

Das erstellte Pinyinprogramm liefert, unter Benutzung umfangreicher aufwendig vorverarbeiteter Daten, eine gute Romanisierung der chinesischen Schriftsprache. Es ist bei ausreichender Geschwindigkeit (ca. 1 Kilobyte pro Sekunde), flexibel und einfach zu Bedienen.

Als direkteste Anwendungsmöglichkeit für das erstellte Pinyinprogramm kommt die Sprachsynthese in betracht, bei der chinesischer Text romanisiert werden muß, bevor er an einen Sprachsynthesizer weitergeleitet werden kann.
Auch zur Bereitstellung von Trainingsdaten für die Spracherkennung ist eine Romanisierung nötig, da Spracherkennungsverfahren meist auf Lautschriftdaten trainiert werden.
Um die von einem Spracherkenner erkannten Sätze von der Lautschrift wieder in die Schriftform zurück zu wandeln, ist eine Abbildung von Pinyin auf die chinesischen Zeichen notwendig. Diese komplexe Abbildung ist die Rücktransformation von der Abbildung, die das Pinyinsystem leistet. Sie wird am besten mittels lernenden Systemen wie z.B. neuronale Netze oder Fuzzyentscheidern realisiert. Diese wiederum benötigen eine große Anzahl von Beispieltexten mit den zugehörigen Pinyintranskriptionen zum Training und genau die kann das Pinyinsystem in beliebiger Anzahl zur Verfügung stellen.
Weiterhin ist es für einen Chinesisch Lernenden sehr hilfreich, neben einem chinesischen Text auch dessen Pinyintranskription zu haben, so kann einfach die Aussprache gelernt werden und das Nachschlagen von Wörtern in einem Wörterbuch wird wesentlich vereinfacht.
Eine chinesische Computerverarbeitung ohne ein chinesisches System wird möglich, wenn man statt des chinesischen Textes dessen Pinyintranskription verwendet. Da Pinyin-Texte ausschließlich aus ASCII-Zeichen bestehen, können so chinesische Texte auf jedem Computer dargestellt werden und einfach per Email verschickt werden.

Anhang A Schnittstellenbeschreibung der Speicherdatenbank

Die Klassen CQSort und CTable der Speicherdatenbank:

```
Class CQSort: (implementiert Quicksort)
      void QSort(long Anz);   // sortiert beliebige Datenstruktur mittels
                              // Quicksort, wobei Anz die Anzahl der
                              // Elementen ist.
                              // (Aufwand im Mittel O(n log n),
                              //  Worstcase O(n^2))
      BOOL BSearch(long &i);  // sucht mittels Bool'scher Suche das Element,
                              // welches gleich dem Schlüsselelement ist.
                              // liefert TRUE und Elementnummer i, falls
                              // gefunden sonst FALSE zurück.
                              // (Aufwand O(log(n)))

  zu überschreibende (zu implementierende) Methoden sind:
      virtual int CompareSKey(long i);    // vergleicht Element i mit dem
                                          // Schlüsselelement
      virtual void SetSKey(long i);       // setzt Element i als Schlüssel
      virtual void Swap(long i1, long i2);// vertauscht Elemente i1 und i2


Class CTable: (implementiert effektiven Datenzugriff und Datenhaltung
               durch eine rudimentäre beliebig indizierbare
               Speicherdatenbank)

CTable(                               // Erzeugt leere Tabelle anhand
     // Spalten- und Indizesbeschreib.
    char* tbname_para,                // Tabellenname
     char* SpaltenString,             // Spaltenbeschreibung
    char* IndizesString,              // Indizesbeschreibung
    char Trennzeichen,                // Trennzeichen für Spaltennamen,
                                      // Indizes und Daten
    char Begrenzungszeichen,          //  optionales  Begrenzungszeichen  für
                                      // Daten, '\0' kein BGZ
    BOOL Ende_Trennzeichen,           //  definiert,  ob  Zeile  mit  Trennz.
                                      // abgeschlossen werden soll
    char Kopf_KNZ,                    // Kennzeichnet Kopfzeilen für
                                      // Spalten-Indizesbeschreib.
    char IndexTrennzeichen);          //  Trennzeichen  zwischen  Unterindizes
                                      // (z.B. s1,s2;s3,s1;)

CTable(                               // Erzeugt Tabelle anhand Datei, und
                                      // liest evt. vorhandene Daten ein.
    char* tbname_para, char* Datei,   // Tabellenname, Dateiname, ...
    char Trennzeichen,
    char Begrenzungszeichen,
    BOOL Ende_Trennzeichen,
    char Kopf_KNZ,
    char IndexTrennzeichen);

void Set(                             // Ändert die bei der Erzeugung
                                      // gesetzten Merkmale
            char Trennzeichen_,       //...
            char Begrenzungszeichen_,
            BOOL Ende_Trennzeichen_,
            char Kopf_KNZ_,
            char IndexTrennzeichen_);

void read(char* Datei, TWHERE where);// Liest Zeile aus Datei, falls
                                      // Funktion where TRUE.
```

```
                                    // Überliest evt. vorhand. Kopfzeilen
void write(char* Datei, TWHERE where,// Schreibt Zeile in Datei, falls
         long Index, BOOL Kopf)     // Funktion where TRUE.
                                    // Sortiert nach Index(Nummer).
                                    // Schreibt Kopf falls Kopf TRUE
// Unter Windows SQLBase //
void readDB(char* table, TWHERE where);   // liest Zeilen aus DB für die
                                          // Funktion where TRUE ist.
void writeDB(char* table, TWHERE where); // schreibt Zeilen in DB für die
                                          // Funktion where TRUE ist.

long insert(char** Werte);          // Fügt Zeile (Datensatz) in
                                    // Speicherdatenbank
                                    // Werte ist Stringarray der Spalten

void update_at(long Zeile, char** Werte); // Überschreibt Datensatz
                                          // Zeile(nummer) mit Stringarray
void update_atSpalte(long Zeile,    // Überschreibt Spalte(nummer) des
      long Spalte, char* Wert);     // Datensatz Zeile(nummer) mit
                                    // String Wert
void update_where(TWHERE where,     // Führt für die Zeile für die die
      TUPDATE update);              // Funktion where erfüllt ist,
                                    // die Funktion update aus

void delete_at(long Zeile);         // löscht Datensatz Zeile(nummer)
void delete_where(TWHERE where);    // löscht Zeile für die die Funktion
                                    // where erfüllt ist

char** select_at(long Zeile);       // liefert Stringarray der entspr.
                                    // Zeile(nummer)
char** select_where(long& StartZeile,    // liefert erstes Stringarray ab
        TWHERE where);                   // Startzeile, für das die
                                         // Funktion where erfüllt ist.
                                         // Setzt Startzeile neu.
char** select_Index(long& StartZeile,    // liefert erstes Stringarray ab
        long Index, char* Wert);         // Startzeile (-1), dessen Index
                                         // den Wert 'Wert' hat. (Bool'sche
                                         // Suche anhand Index)
                                         // Wert == NULL: liefert
                                         // Zeile(Index(StartZeile))
char* get_tbname();                 // liefert Tabellename
char* get_Spaltenname(long Spalte); // liefert Spaltenname von 'Spalte'
long get_Spaltennummer              // liefert Spaltennummer anhand
       (char* Spaltenname)          // des Spaltennamens
long get_AnzSpalten(char* Spaltenname);  // liefert Anzahl der Spalten
long get_AnzZeilen(char* Spaltenname);   // liefert Anzahl der Zeilen
long get_AnzIndizes(char* Spaltenname);  // liefert Anzahl der Indizes


//*** Hilfreiche interne Funktionen ***//

long StrToAnz(char* string);        // Anzahl von Spalten in einem
                                    // Datensatzstring (Zeile)
char** StrToArray(char* string);    // Wandlung von Datensatzstring zu
                                    // Stringarray
char* ArrayToStr(char** Werte,      // Wandlung von
       long AnzSpalten);            // Stringarray zu Datensatzstring
BOOL readZeile(FILE* fp, char* buf); // Hilfsfunktion: eine Zeile einlesen
```

Anhang B Aufbau des Pinyinprogramms

Im folgenden werden die Hauptfunktionen und deren Aufrufstruktur beschrieben. Um die Funktionsabhängigkeiten zu erfassen, sollte sich wie bei einem C-Programm von unten nach oben bewegt werden.

```
Bewerte {
    Berechne Bigrammscore;
    Berechne Score Einzelworte;
    Berechne Σ Score der Einzelzeichen;
    Bestimme Anzahl Worte;

    Bestimme Gesamtscore der Satztrennung anhand obiger Einzelscores;
}


Trenne_WB {
    ∀ Zeichenpaare ZZ des betrachteten Satzteils
      Teste ob ZZ in keinem Wort des Wörterbuchs als Teilwort existiert, dann trenne Satz-
teil
}

Trenne_BW {
    Trenne Satzteil rekursiv und trage alle Möglichkeiten in SpeicherDB RESULT ein;
    ∀ Satztrennungen aus RESULT;
        Bewerte(Satztrennung);
    Wähle beste Trennung;
}

Pinyin{
    Score = max (aller linkseitiger und rechtseitiger Bigrammscores);
    if Score >0  Pinyinabbildung mittels entsprechendem Bigramm; return;

    Score = max (aller Unigrammscores);
    if Score >0  Pinyinabbildung mittels entsprechendem Unigramm; return;

    Pinyinabbildung mittels Wörterbuch;
}

Execute {
    // Pass1: Satzzeichen/Sonderzeichentrennung
    while readZeile (Zeile)
        while readSatzteil (Zeile, Satzteil)
            write (Satzteil);

    // Pass2: Wörterbuchtrennung
```

```
      while ReadSatzteil (Satzteil)
            Trenne_WB (Satzteil, Satzteil_WB);
         write (Satzteil_WB);

      // Pass3: Bewertung_Trennung
      while ReadSatzteil (Satzteil)
            Trenne_BW (Satzteil, Satzteil_BW);
         write (Satzteil_BW);

      // Pass4: Pinyin_Umsetzung
      while ReadSatzteil (Satzteil)
         Pinyin(Satzteil, Satzteil_next, Satzteil_prev, Satzteil_Pinyin);
         Format(Satzteil_Pinyin);                    // spezielle Ausgabeformatierung
              Abb_Satzzeichen(Satzteil_Pinyin);      //  Satz- und Sonderzeichenabbildung
         write (Satzteil_Pinyin);
}


main  {
         Eingabeparameterverarbeitung;
         Dateienüberprüfen und öffnen;
         Datenbankrelationen initialisieren;

         if Listenverarbeitung
            while ((Listelement = read_next(List))!= NULL)
               Execute (Listelement);
         else
            Execute (Datei);

        if Fehlerprotokoll
       Validate (Datei , Vergleichsdatei);
}
```

Abbildungsverzeichnis

Abbildung 2.1: Erste Schriftfunde auf Tonscherben in Banpo 7
Abbildung 2.2: Entwicklung von der frühen Bilderschrift zu heutigen chinesischen Zeichen ... 8
Abbildung 2.3: Orakelknocheninschrift 10
Abbildung 2.4: Bronzeinschrift der Zhou-Dynastie 10
Abbildung 2.5: Siegelinschrift 11
Abbildung 2.6: Offizielle Schrift wie sie heute noch in Taiwan verwendet wird 11
Abbildung 2.7: Vereinfachte Schrift wie sie heute in der Volksrepublik China verwendet wird. (der dargestellte Text entspricht dem in der vorherigen Abbildung) 12
Abbildung 2.8: Die Verteilung der Dialekte in China (aus [13]) 15
Abbildung 2.9: sechs Phoneme des zeichenbasierten Bopomofo-Systems 15
Abbildung 2.10: Beispiel einer in Taiwan gebräuchlich Schreibrichtungen 1. Mose 1 einer taiwanesischen Bibel (oben-nach-unten, rechts-nach-links, hinten-nach-vorne) 17
Abbildung 2.11: Tabelle der Pinyin - Sprachlaute des Mandarin Dialekts 18
Abbildung 3.1: SQL-Skripts unter SQLTalk 28
Abbildung 3.2: Wörterbuchverarbeitung mit Benutzeroberfläche 29
Abbildung 3.3: Website mit Pinyintranskription 31
Abbildung 3.4: In Plaintext umgesetztes HTML 32
Abbildung 3.5: Radikaltabelle 34
Abbildung 3.6: Zeichenliste zum Radikal 彳 (62) 35
Abbildung 3.7: Eintrag zu 行 auf Seite 320 35
Abbildung 3.8: Eintrag zu 行 auf Seite 905 (Auszug) 36
Abbildung 4.1: Blockdiagramm des Pinyinprogramms 39
Abbildung 4.2: Grundstruktur des Pinyinprogramms 44
Abbildung 4.3: Ablaufdiagramm des Pinyinprogramms 45
Abbildung 4.4: Test und Validierungsablauf 46

Literaturverzeichnis

[1] Xu Zhenmin, Chen Huiying, Dr. Rainer Kloubert et al.: Das neue Chinesisch-Deutsche Wörterbuch (新汉德词典), Peking 1993.

[2] Yang Yezhi, Zhai Lilin, Pan Zaiping et al.: Deutsch-Chinesisches Wörterbuch (德汉词典), Shanghai 1987.

[3] Xinhua Zidian (新华词典), Peking 1975

[4] Andreas Guder-Manitius: Chinesisch-Deutsches Lernwörterbuch, Verlag Ute Schiller, Berlin 1991

[5] 许素芳，施其南 et al.: Handwörterbuch Deutsch-Chinesisch Chinesisch-Deutsch (精选德汉汉德词典), The Commercial Press, Peking und Langenscheidt KG, Berlin und München, 1994

[6] Li Dejin, Cheng Meizhen: Praktische chinesische Grammatik für Ausländer (外国人实用汉语语法)，Sinolingua Peking 1993

[7] Ly Ping-Chien, Monika Motsch: Kurze Grammatik der modernen chinesischen Hochsprache, Dürr & Kessler Rheinbreitbach

[8] He Peihui, Xiong Wenhua, Mei Xiuxian:Practical Chinese Reader (实用汉语课本) Book I - VI, The Commercial Press, Peking 1991

[9] Li Peiyuan, Ren Yuan, Zhao Shuhua, Prof. Käthe Zhao et al.: Grundkurs der chinesischen Sprache (基础汉语课本) Band 1 - 4, Sinolingua Peking 1979-88

[10] Klaus Stermann, Katherin Reiser-von Loh, Ho Wan-yee, Li Jian-ming, Zhao Yongxin: Langenscheidts Sprachführer Chinesisch, Langenscheidt KG, Berlin und München 1994

[11] Sun Baoyong, Chen Feng, Cui Jianying, Bing Xiuli: Learning Chinese Through Group-Character Anlyses (分沮识字学汉语) Band 1 - 3, Sinolingua Peking 1993

[12] Yin Binyong: Modern Chinese Characters (现代汉子), Sinolingua Peking 1994

[13] Jacques Gernet: Die chinesische Welt, Insel Verlag 1989

[14] Chinese Star 2.0 (中文之星), SunTendy 1994

[15] Tao Hongyi, Zhao Tangshou: Handwörterbuch der Gegenwartssprache (现代汉德德汉词典), Peking 1994

[16] The GlobalPhone Project: Multilingual LVCSR with Janus3 2nd SQEL-Workshop, Plzen, Tschechien 1997.

[17] A. Lavie et al.: JANUS-III: Speech-to-Speech Translation in Multiple Languages. To appear in ICASSP 1997.