

Universität Stuttgart
INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG
Azenbergstraße 12
70174 Stuttgart

STUDIENARBEIT

Ausspracheregeln in der Text-to-Speech Synthese für das Kroatische

DANIEL DURAN
E-Mail: contact@daniel-duran.de

Betreuer: PD Dr. phil. Bernd Möbius
Prüfer: PD Dr. phil. Bernd Möbius
Studienarbeit Nummer: 49
Beginn der Arbeit: 01. 05. 2005
Ende der Arbeit: 19. 12. 2005

Ich möchte zunächst die Gelegenheit nutzen mich bei MATEUSZ WIĄCEK für die gute Zusammenarbeit und seine Kritik im Zusammenhang mit dieser Arbeit zu bedanken. Außerdem danke ich DENIZ ÇALAĞAN sowie meinem Bruder MIHAEL DURAN für Kritik und viele kleine und große Hinweise.

Auch möchte ich mich bei BERND MÖBIUS für die Unterstützung bedanken.

Zu guter Letzt bedanke ich mich bei meinen Eltern für alles.

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe.

Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

Daniel Duran

Esslingen, den 19. Dezember 2005

Zusammenfassung

In dieser Arbeit sollen Ausspracheregeln für ein kroatisches Text-to-Speech System entwickelt werden. Dies geschieht unter zu Hilfenahme des *Festival Sprachsynthesystems*.

Im einleitenden ersten Kapitel erfolgt eine allgemeine Übersicht über die Sprachsynthese als Forschungsgebiet der Computerlinguistik sowie als praktische Anwendung. Es werden verschiedene Konzepte der Sprachsynthese kurz vorgestellt und dann wird eine kurze Beschreibung der Text-to-Speech Synthese gegeben. Im Anschluß daran erfolgt eine einleitende Betrachtung der kroatischen Sprache sowie einer genaueren Begriffsbestimmung des Ausdrucks „Kroatische Sprache“.

Im zweiten Kapitel wird die geschriebene kroatische Sprache ausführlich dargestellt. Besondere Bedeutung wird dabei den Merkmalen und Gegebenheiten der geschriebenen Sprache zukommen, die für die Text-to-Speech Synthese wichtig sind. Die kroatische Orthographie wird zunächst allgemein und in ihrer Gesamtheit betrachtet werden. Danach werden einzelne spezielle Details und Gegebenheiten vorgestellt. Speziell für die Text-to-Speech Synthese relevante Probleme werden ausführlich dargestellt werden. In diesem Kapitel werden auch verschiedene Arten der phonetischen und phonologischen Transkription der kroatischen Sprache vorgestellt.

Das dritte Kapitel befaßt sich ausführlich mit der Grammatik des Kroatischen. Dabei wird nicht die gesamte Grammatik betrachtet, sondern nur die für die Synthese relevanten Teilbereiche, insbesondere der Phonetik, Phonologie und Morphologie. Besonders hervorgehoben werden aus diesem Bereich das Lautinventar des Kroatischen sowie der Wortakzent.

Die im Rahmen dieser Arbeit entwickelten Ausspracheregeln werden im vierten Kapitel vorgestellt und erläutert. Zu diesem Zweck erfolgt zunächst eine kurze Darstellung des Festival Systems und seiner Funktionsweise. Die entwickelten Regeln werden dann im Anschluß im einzelnen ausführlich vorgestellt und auch getestet. Eine abschließende kurze Analyse der Ausspracheregeln stellt die praktischen Ergebnisse dieser Arbeit dar.

Inhaltsverzeichnis

1	Einführung	3
1.1	Text-to-Speech Synthese	3
1.1.1	Anforderungen	3
1.1.2	Komponenten eines TTS-Systems	4
1.1.3	Verschiedene Konzepte der Sprachsynthese	5
1.1.4	Bestehende Lösungen	6
1.2	Eingabedaten eines (kroatischen) TTS-Systems	7
1.3	Linguistischer Überblick	9
1.3.1	Begriffsbestimmung	9
2	Die geschriebene kroatische Sprache	14
2.1	Das Kroatische Alphabet	14
2.2	Graphem-Phonem-Beziehungen	15
2.3	Orthographie	16
2.3.1	Groß- und Kleinschreibung	17
2.3.2	Abkürzungen und Akronyme	18
2.4	Vokale und andere Probleme	19
2.4.1	Häufige Zusatzbuchstaben	20
2.4.2	Sonstige Zusatzbuchstaben	22
2.5	Allgemeiner Schreibstil	23
2.5.1	„ASCII-Schreibweise“	23
2.6	Fremdwörter	24
2.6.1	Fremdwörter, die keine Eigennamen sind	24
2.6.2	Fremde Eigennamen	25
2.7	Zahlen und Zahlwörter	27
2.7.1	Kardinalzahlen	27
2.7.2	Weitere veränderliche Zahlwörter	28
2.8	Homonyme	28
2.9	Phonetische und Phonologische Transkriptionen	31
2.9.1	IPA	31
2.9.2	SAMPA	31
2.9.3	Slavistik und Kroatistik	32
3	Grammatik	34
3.1	Lautinventar	34
3.1.1	Das Lautinventar für diese Arbeit	38
3.2	Wortakzent	39

4	TTS–Ausspracheregeln	44
4.1	Festival und Festvox	44
4.2	Lautinventar	45
4.2.1	Die Vokale	45
4.2.2	Die Konsonanten	46
4.3	Lexikon	47
4.4	Transkriptionsregeln (Letter to sound rules)	48
4.4.1	Die möglichen Transkriptionen	49
4.4.2	Die letter–to–sound rules in Festival	52
4.5	Test und Auswertung	52
4.5.1	Testvorgang und Auswertung	53
4.5.2	Vergleich mit manueller Transkription	56
5	Schlußbetrachtungen	58
A	Anhang: Phoneset	61
B	Anhang: Lexikon	64
C	Anhang: Letter-to-Sound Rules	69
D	Anhang: Testkorpus	73
	Abkürzungsverzeichnis	78
	Literaturverzeichnis	80
	Abstract (English)	83

Kapitel 1

Einführung

„Sprachsynthese: eine unmögliche Aufgabe!“
(Möbius 2002)

1.1 Text-to-Speech Synthese

Als Sprachsynthese bezeichnet man die künstliche Erzeugung gesprochener Sprache (auch Rede genannt) mit Hilfe eines Computers, oder auch anderen Maschinen. Ein Text-to-Speech¹ Sprachsynthesystem ist ein Computerprogramm, das geschriebenen (orthographischen) Text in gesprochene Sprache umwandelt. Im Idealfall stellt ein solches System also einen (unbeschränkten) Vorleseautomaten dar.

In diesem Abschnitt soll zunächst ein kurzer Überblick über die allgemeinen Anforderungen an ein Text-to-Speech System (kurz: TTS) gegeben werden. Anschließend gehe ich auf die einzelnen Komponenten eines TTS ein und werde am Ende dieses Abschnitts noch einige bestehende Lösungen für die kroatische Text-to-Speech Synthese vorstellen.

1.1.1 Anforderungen

Ein TTS-System soll nach Definition einen eingegebenen geschriebenen Text in gesprochene Sprache umwandeln und wiedergeben (vorlesen).

Die *Eingabe* eines TTS-Systems besteht aus geschriebenem Text. So könnte man von einem allgemeinen TTS erwarten, daß es beliebigen geschriebenen Text als Eingabe akzeptiert und in einer angemessenen Weise verarbeiten kann. Die Eingabe sollte eventuell auch nicht nur auf orthographischen, also „korrekten“ Text beschränkt sein. Ein TTS zum Vorlesen elektronischer Post oder SMS Nachrichten z.B. müßte mit Tippfehlern sowie den unterschiedlichsten nicht-orthographischen Zeichenketten zurechtkommen, wie z.B. :-), }:o), LOL, oder

¹Den deutschen Begriff „Text-zu-Sprache“ werde ich in dieser Arbeit nicht verwenden.

AAAAAARRRRGGGHH². Auch müssen verwendete Abkürzungen, Daten, Zahlen, Internet- und E-Mail Adressen sowie auch Personen- und Eigennamen korrekt artikuliert werden. Dies stellt keine triviale Aufgabe für den Computer dar, wie später noch genau beschrieben werden soll.

An die *Ausgabe* eines TTS können unterschiedliche Anforderungen gestellt werden. Neben so grundlegenden Eigenschaften wie der korrekten Verarbeitung der Eingabe, kann von einem TTS vor allem die Produktion verständlicher Sprachausgabe gefordert werden. Oft wird darunter eine „natürliche“ Aussprache verstanden. Die computergenerierte Sprache sollte sich also so menschlich wie möglich anhören. Das erzeugte Audiosignal sollte vom Anwender im Idealfall nicht oder nur schwer von einem vom Menschen gesprochenen oder vorgelesenen Text unterschieden werden können. Daneben spielt in konkreten Anwendungen auch die Art der Stimme selber eine große Rolle: wird eine angenehme Stimme zum Vorlesen längerer Texte gewünscht, eine sachliche oder eine ausgefallene Stimme, eine männliche oder weibliche — oder gar eine erkennbar künstliche „Roboterstimme“? Man kann sich auch fragen, welchem Sprachregister die generierte Sprache entsprechen soll — sachliche, sehr deutlich artikuliert, Bühnensprache oder vielleicht sogar, je nach Einsatzgebiet, dialektal gefärbte Aussprache?

Man sieht also, daß neben linguistischen auch zahlreiche andere Faktoren bei der Implementierung, den möglichen Einsatzgebieten wie auch bei der Bewertung eines TTS-Systems eine Rolle spielen. Ein weiterer wichtiger Punkt ist das Verhalten des Systems zwischen den beiden bereits genannten Punkten, der Eingabe und der Ausgabe: die *Verarbeitung*. Für diese Arbeit stellt dieser Punkt das zentrale Thema dar (genauer genommen nur ein kleiner Teil aus diesem Bereich). Im folgenden Abschnitt werden die Komponenten eines TTS-Systems kurz vorgestellt.

1.1.2 Komponenten eines TTS-Systems

Ein TTS-System muß eine Reihe von Operationen durchführen. Im Allgemeinen können zwei Phasen der Verarbeitung bzw. Komponenten unterschieden werden (egal ob es nun modular aufgebaut ist oder nicht): (1) die linguistische Textanalyse der Eingabe und (2) die Generierung der synthetischen Sprachausgabe (Möbius 2001). Folgende Aufzählung soll diese Gliederung ein wenig mehr im Detail erläutern:

- **Linguistische Textanalyse der Eingabe.** Dieser Teil der Verarbeitung wandelt den eingegebenen Text in eine interne symbolische Repräsentation um. Auf Basis dieser Repräsentation kann im nächsten Schritt eine Audioausgabe erzeugt werden. Dieser erste Teil erstellt also grob gesprochen eine enge phonetische Transkription des Eingabetextes und versieht ihn je nach Implementierung mit zusätzlicher linguistischer Information wie z.B. Wortarten, Pausen usw. Innerhalb dieser Komponente lassen sich, je nach Art und Umfang eines gegebenen TTS-Systems, im Einzelnen noch folgende Teile unterscheiden:
 - Textnormalisierung. Hierbei handelt es sich um eine erste (nicht-linguistische) Vorverarbeitung der Eingabe, wie z.B. der Entfernung überflüssiger (doppelter) Leerzeichen oder Zeilenumbrüche.

²In einem für diese Arbeit erstellten Testkorpus (siehe Anhang D) sind z.B. die Zeichenketten „lol“ 4823 mal, „eek“ 122, „roff“ 82 mal und „aaaargh“ immerhin noch 7 mal vorgekommen.

- Tokenisierung. Der eingegebene Text wird in seine einzelnen kleinsten Teile (Tokens) aufgeteilt. Darunter ist vor allem die Zerlegung in die einzelnen Wörter gemeint, was in Sprachen wie Deutsch oder Kroatisch zum größten Teil anhand der Leerzeichen erfolgen kann, da in diesen Sprachen in der Schrift ein Leerzeichen zwischen zwei Wörtern gesetzt wird. Die Definition von „Wort“ kann natürlich unterschiedlich sein und ist unter Umständen auch ein Problem mit dem man sich beim Entwurf von TTS-Systemen auseinandersetzen muß. Weitere Tokens in einem Text stellen Satzzeichen oder Zahlen dar.
 - Phonologische Analyse. Mit Hilfe von Aussprache- und Syllabifizierungsregeln der betreffenden Sprache wird anhand des geschriebenen Textes die konkrete Aussprache ermittelt.
 - Lexikalische Analyse. Der Text kann für die Zwecke der Synthese auf mehreren Ebenen analysiert werden. So kann z.B. eine lexikalische Analyse mit Hilfe eines *Lexikons* zur Bestimmung des Wortakzentes beitragen. Ein *Part-of-Speech Tagger* kann die Wortart der einzelnen Wörter bestimmen. Eine *morphologische Komponente* kann Informationen über morphologische Eigenschaften und Lemmata liefern (dies kann besonders wichtig sein bei stark flektierenden Sprachen, wie etwa dem Kroatischen). All diese Analysen können besonders dann sehr hilfreich sein, wenn es zu Ambiguitäten kommt und die korrekte Aussprache von mehrdeutigen Wörtern von ihrer Disambiguierung abhängt.
 - Syntaktische Analyse. Ein Parser kann wertvolle Informationen über Satzstrukturen und syntaktische Funktionen liefern, die zur Bestimmung der besten Aussprache (insbesondere der Satzintonation) von Bedeutung sein können.
 - Zahlen, Abkürzungen, Namen. Besondere Analyseschritte sind nötig, wenn die Eingabe Abkürzungen enthält, die expandiert und in der vollen Form ausgesprochen werden sollen („z.B.“ → „zum Beispiel“). Auch Ziffern erfordern eine genauere Analyse bevor sie korrekt in Sprache umgesetzt werden können. Eigennamen stellen in vielen Sprachen Abweichungen von den üblichen phonotaktischen oder orthographischen Regeln dar und können ohne gesonderte Behandlung oft nicht korrekt verarbeitet werden.
- **Generierung der synthetischen Sprachausgabe.** Zur Generierung Sprachausgabe gehört nicht nur die eigentliche Erzeugung des (digitalen) Audiosignals sondern dazu gehören auch all die Prozesse, die zur Bestimmung der Zusammensetzung des zu erzeugenden Signals, notwendig sind. Diese Bestimmung der Zusammensetzung spielt insbesondere bei der so genannten *konkatenativen Synthese* von Sprachsegmenten aus einem vorher aufgenommenen Inventar von Sprachaufnahmen eine große Rolle. Die Bestimmung der idealen Kandidaten, welche für die Sprachausgabe verwendet werden sollen, kann insbesondere bei der *Unit Selection* eine aufwendige Aufgabe darstellen. Darauf werde ich im folgenden Abschnitt 1.1.3 etwas genauer eingehen.

1.1.3 Verschiedene Konzepte der Sprachsynthese

Die *Text-to-Speech* Synthese stellt ein spezielles Konzept der Sprachsynthese dar, und zwar was die Art der Eingabedaten angeht (vgl. Abschnitt 1.2). Daneben existieren auch andere Konzepte wie z.B. *Concept-to-Speech*, bei welcher die Sprachausgabe nicht anhand eines geschriebenen Textes erzeugt wird, sondern anhand von abstrakten Daten und Konzepten.

Auf diese neben dem TTS existierenden Konzepte möchte ich hier nicht eingehen, da sie nicht zum Gegenstand dieser Arbeit gehören.

Betrachtet man die Ausgabe der Sprachsynthesysteme lassen sich auch verschiedene Konzepte unterscheiden. Sie lassen sich nach der Art und Weise der Erzeugung des akustischen Ausgabesignals einteilen:

- Die *Formantsynthese* versucht die akustischen Eigenschaften des menschlichen Sprachapparates oder aber die akustischen Merkmale der Sprachlaute auf elektronischem Wege so genau wie möglich nachzubilden. Hier können zahlreiche Methoden der digitalen Signalverarbeitung zum Einsatz kommen, oder auch elektronische Verfahren zur analogen Signalerzeugung. Die Qualität von so erzeugten Sprachsignalen ist aber bestenfalls unbefriedigend im Vergleich zur natürlichen menschlichen Sprache.
- Die *artikulatorische Synthese* versucht den menschlichen Sprachapparat und seine physiologischen und akustischen Eigenschaften zu simulieren und so die exakte Lautformung im menschlichen Vokaltrakt nachzuahmen.
- Die *konkatenative Synthese* nutzt ein vorher aufgezeichnetes Audioinventar von Sprachsegmenten zur Generierung des akustischen Signals. Dieses Konzept der Sprachsynthese läßt sich noch weiter untergliedern. Die beiden verbreitetsten sind:
 - *Diphonsynthese*. Als Inventar dient eine Menge von so genannten Diphonen. Dies sind kleine Sprachsegmente die (im Allgemeinen) von der Mitte eines Lautes (eines Phons) bis zur Mitte des folgenden reichen. Das Lautinventar besteht also aus einer Sammlung von allen in einer Sprache möglichen Übergängen zwischen zweier ihrer Laute. Die Qualität der so synthetisierten Sprache ist meistens nicht ideal, aber sollte für alle unterschiedlichen Texte konstant gleich bleiben.
 - *Unit Selection*. Als Inventar dient ein großer Korpus von aufgenommener zusammenhängender gesprochener Sprache. Das heißt das Inventar, auf das bei der Synthese zurückgegriffen wird, besteht aus ganzen Worten, Sätzen oder sogar längeren zusammenhängenden Äußerungen. Die Auswahl der einzelnen Segmente für die Synthese erfolgt während der Laufzeit des System dynamisch in Abhängigkeit des jeweiligen zu synthetisierenden Textes. Die Größe der einzelnen Segmente ist nicht fest und kann im besten aber allgemein eher unwahrscheinlichen Fall eine komplette Äußerung enthalten, wenn diese genau so im Inventar bereits vorhanden ist. Die Qualität der so synthetisierten Sprache kann sehr gut und unter Umständen von natürlicher Sprache nicht zu unterscheiden sein. Allerdings kann die Sprachqualität in Abhängigkeit des eingegebenen Textes variieren, z.B. wenn für die Synthese nur die kleinstmöglichen Segmente (einzelne Laute oder Diphone) genommen werden können, da ein gegebenes Wort nicht im Korpus enthalten ist.

1.1.4 Bestehende Lösungen

Da der Markt für kroatische Sprachsynthesysteme im internationalen Vergleich nur sehr klein ist gibt es bis jetzt nur relativ wenige kommerzielle oder auch nur experimentelle TTS-Systeme.

Für das MBROLA System existiert eine kroatische Stimme. Genauer gesagt handelt es sich dabei „nur“ um eine Diphonbasis für die Diphonsynthese welche nur zu nicht-kommerziellen

und nicht-militärischen Zwecken verwendet werden darf. Das MBROLA Projekt ist im Internet unter der Adresse <http://tcts.fpms.ac.be/synthesis/mbrola.html> zu finden (vgl. Bakran 1998).

In Kroatien ist insbesondere auch der Kroatische Bund der Blinden *Hrvatski savez slijepih*³ an Sprachsynthese Systemen interessiert und auf diesem Gebiet engagiert. Dort wurde auch die Entwicklung einer Linux Distribution für Blinde⁴ initiiert, welche auch kroatische Sprachsynthese bietet (auf Basis von MBROLA). Daneben gibt es auch eine kroatische Version des Tschechischen Sprachsynthesystems *WinTalker Voice* für das Windows Betriebssystem (vgl. Delić & Perčinić).

1.2 Eingabedaten eines (kroatischen) TTS-Systems

Wie ich bereits in Abschnitt 1.1 beschrieben habe nimmt ein TTS-System geschriebenen Text als Eingabe. Diese Texte können je nach Verwendungszweck des TTS-Systems aus ganz unterschiedlichen Quellen stammen und daher ganz unterschiedliche Formate haben. Grundsätzlich sind alle elektronischen Texte als Eingabe denkbar. Dies können vom Anwender direkt eingegebene Texte sein, oder auch E-Mails, Internetseiten, SMS Nachrichten, Texte aus Instant Messaging und Chats, Newsgroups, E-Books usw. Aber auch gedruckte Texte können in Frage kommen, wenn sie durch Texterkennung dem TTS in elektronischer Form verfügbar gemacht werden. Das ist allerdings ein ganz anderes Thema.

Beim Entwurf eines TTS-Systems ist es also notwendig sich über die Rechtschreibung (und eventuell auch alternative Schreibweisen) der zu verarbeitenden Sprache im Klaren zu sein. Die Orthographie des Kroatischen und die damit zusammenhängenden Probleme werden ausführlich in Kapitel 2 dargestellt.

Zeichenkodierung

Ein anderer, aber nicht weniger wichtiger Aspekt der Eingabedaten ist ihre Kodierung. Die einzelnen Zeichen müssen im Computer binär kodiert sein. Das wohl bekannteste System zur Kodierung von Texten bzw. einzelnen Zeichen ist der ASCII Zeichensatz. Hierbei handelt es sich um ein System, daß je 7 Bits verwendet um ein Zeichen zu repräsentieren. So können insgesamt 128 verschiedene Zeichen kodiert werden. Das Problem ist, daß nur die Buchstaben aus dem englischen Alphabet ohne Diakritika enthalten sind. Da es nicht nur englische Texte gibt sind weitere Zeichensätze definiert worden. Die meisten Zeichenkodierungssysteme verwenden 1 Bit mehr, so daß ein einzelnes Zeichen durch 8 Bits, also einem Byte repräsentiert wird.

Schwierigkeiten bei der Verarbeitung elektronischer Texte können sich aus der Tatsache ergeben, daß es viele verschiedene solcher Zeichensätze gibt. Die meisten sind so aufgebaut, daß sich die erste Hälfte, also die Zeichen von 0 bis 127, von (den darstellbaren Zeichen von) ASCII nicht unterscheidet und die zweite Hälfte zusätzliche Zeichen für die verschiedenen Sprachen enthält.

³Internet: <http://www.savez-slijepih.hr>

⁴Internet: <http://www.ipsis.hr/gls/>

Für Kroatische Texte sind dabei vor allem zwei der 1-Byte Zeichensätze von Bedeutung: *ISO 8859-2* (Latin-2) und *Microsoft Windows Codepage 1250* (CP1250). Beide Zeichensätze enthalten die kroatischen Zeichen mit Diakritika *ćđšž* und *ČĆĐŠŽ*, welche in ASCII nicht definiert sind. Abbildung 1.1 stellt die unterschiedlichen Zeichenkodierungssysteme gegenüber. Was auffällt ist, daß sich die Codes für einzelne Zeichen in Latin-2 und CP1250 unterscheiden. Dies führt dazu, daß so kodierte Texte nicht ohne weiteres austauschbar sind und unter Umständen falsch angezeigt werden.

Die vierte Spalte der Abbildung 1.1 enthält *Unicode* Werte für die speziellen kroatischen Buchstaben. Unicode ist ein Standard, der entwickelt wurde um einen einheitlichen Zeichensatz für alle Sprachen, genauer gesagt Schriften, zu definieren⁵. Bisher umfasst Unicode schon mehr als 90.000 verschiedene Zeichen. Es ist also nicht mehr möglich die Zeichen in einem reinen 8-Bit System zu kodieren, das ja nur 256 verschiedene Werte und damit maximal auch nur 256 Zeichen zulässt.

Unicode selber stellt kein bestimmtes Zeichenkodierungssystem dar, sondern ist als abstrakter Standard zu verstehen. Es gibt unterschiedliche Systeme um den Unicodezeichen konkrete Bitfolgen zuzuordnen. Sie unterscheiden sich vor allem auch darin, wie viele Bits für die Kodierung eines Zeichens verwendet werden. Für das Internet, und damit möglicherweise auch für TTS-Systeme, ist besonders *UTF-8* von Bedeutung. UTF-8 steht für *Unicode Transformation Format 8*⁶ und stellt ein System zur Kodierung der Unicode Zeichen in Bitfolgen variabler Länge dar. Es ist deshalb so populär, da es ASCII kompatibel ist. Die in ASCII definierten Zeichen werden in UTF-8 mit der gleichen Bitfolge, also ebenfalls mit einem Byte kodiert. Höhere Unicode Zeichen (also Zeichen im Code-Bereich über 127) werden mit bis zu 4 Bytes kodiert. Besonders bei Schriften, die auf dem lateinischen Alphabet beruhen wird so erheblich Speicherplatz gespart — im Vergleich zu anderen Unicode Zeichenkodierungen, wie z.B. UTF-32, das alle Zeichen grundsätzlich mit 32 Bit (4 Byte) kodiert. In Abbildung 1.1 sind in der Spalte UTF-8 die Byte-Werte für die speziellen kroatischen Buchstaben angesehen. Man sieht, daß diese Zeichen mit jeweils zwei Bytes kodiert werden.

Zeichen	CP1250	Latin-2	Unicode	UTF-8
ć	E8	E8	010D	C4 8D
ć	E6	E6	0107	C4 87
đ	F0	F0	0111	C4 91
š	9A	B9	0161	C5 A1
ž	9E	BE	017E	C5 BE
Č	C8	C8	010C	C4 8C
Ć	C6	C6	0106	C4 86
Đ	D0	D0	0110	C4 90
Š	8A	A9	0160	C5 A0
Ž	8E	AE	017D	C5 BD

Abbildung 1.1: Vergleich der Zeichensätze (Hexwerte)

Soll nun ein TTS-System zum Vorlesen von Internetseiten entwickelt werden, muss man damit rechnen, daß verschiedene Zeichensätze verwendet werden. So verwenden z.B. die

⁵siehe (Unicode Standard 2005)

⁶Internet: <http://www.utf-8.com>

Internetseiten des *Večernji List* (Tageszeitung) www.vecernji-list.hr oder der *Wikipedia* hr.wikipedia.org die UTF-8 Zeichenkodierung; die Seiten des *Vjesnik* (Tageszeitung) www.vjesnik.hr oder die meisten Internetseiten der kroatischen Ministerien und staatlichen Organe, wie z.B. des Parlaments *Sabor* www.sabor.hr, verwenden den Windows Zeichensatz CP1250 und die Seiten des staatlichen Fernsehens *HRT* www.hrt.hr schließlich verwenden den Latin-2 Zeichensatz (ISO 8859-2).

Vor der eigentlichen Verarbeitung des Textes in einem TTS sollte also gegebenenfalls noch eine Zeichensatznormalisierung stattfinden um die verschiedenen Zeichensätze verarbeiten zu können. Dies gilt insbesondere dann, wenn die zu synthetisierenden Texte aus verschiedenen Quellen stammen (können).

1.3 Linguistischer Überblick — Begriffsbestimmung

Kroatisch ist eine slavische Sprache der indoeuropäischen Sprachfamilie und gehört zum westlichen Zweig der südslavischen Untergruppe. Als Amtssprache wird Kroatisch in der Republik Kroatien und (neben Bosnisch und Serbisch) in Bosnien-Herzegowina verwendet. Außerdem stellt Kroatisch im österreichischen Burgenland, sowie in der Vojvodina in Serbien und einigen Orten im italienischen Molise und weiteren kleineren Ortschaften in den Nachbarländern eine regionale Minderheitensprache einer autochtonen kroatischen Bevölkerungsgruppe dar. Ein detaillierter Überblick über die linguistische, philologische oder politische Situation des Kroatischen kann und soll im Rahmen dieser Arbeit nicht erfolgen. Da Kroatisch in Deutschland aber nicht zu den allgemein wohlbekannten und häufig beschriebenen Sprachen zählt, soll hier zu Beginn ein kurzer Gesamtüberblick gegeben werden. Eine ausführliche Darstellung der linguistischen Merkmale erfolgt so weit, wie es für das Thema dieser Arbeit, der textbasierten Sprachsynthese, von Bedeutung ist. Zunächst soll nun ein kurzer Überblick und eine genauere (dialektologische) Begriffsbestimmung erfolgen (Abschnitt 1.3.1), gefolgt einer etwas ausführlicheren Beschreibung der Orthographie (Abschnitt 2.3) sowie von einigen Worten zur Grammatik (Kapitel 3).

1.3.1 Begriffsbestimmung

Im vorigen Abschnitt wurde bereits erwähnt, daß Kroatisch zu den südslavischen Sprachen gehört. Der Begriff „Kroatisch“ kann aber ganz unterschiedlich interpretiert werden. Es ist deshalb sinnvoll an dieser Stelle zunächst eine kurze Begriffsbestimmung vorzunehmen.

Der Begriff „Kroatisch“ könnte (unter anderem) folgende unterschiedliche Konzepte bezeichnen:

- die Amtssprache Kroatiens
- die in Kroatien gesprochene Volkssprache
- die Standardsprache der Kroaten
- die Gemeinsprache in Kroatien

Zu jedem dieser vier Punkte will ich nun eine Erklärung geben sowie einige Anmerkungen machen, bevor ich meine für diese Arbeit relevante Interpretation des Begriffs „Kroatisch“ angebe.

Amtssprache

Als *Amtssprache* wird eine Sprache (oder einer Varietät der Sprache) bezeichnet, welche offiziell, per Gesetz, von staatlichen Behörden verwendet werden soll.

Die offizielle Amtssprache Kroatiens wird in der Verfassung schlicht als „*hrvatski jezik*“, also wörtlich übersetzt als *kroatische Sprache* bezeichnet.

Članak 12.

U Republici Hrvatskoj u službenoj je uporabi hrvatski jezik i latinično pismo.
[...]

Quelle: <http://www.usud.hr/htdocs/hr/ustavrh.htm>

Article 12

The Croatian language and the Latin script shall be in official use in the Republic of Croatia. [...]

Quelle: http://www.usud.hr/htdocs/en/the_constitution.htm

Die korrekte Bezeichnung für die *Amtssprache* Kroatiens ist also *Kroatisch*.

Volkssprache(n)

Als *Volkssprache* wird eine bestimmte Varietät der Sprache bezeichnet, welche der Bevölkerung innerhalb eines bestimmten Gebietes als Umgangssprache dient. Diese Varietät kann sich von der allgemeinen, überregional verwendeten Sprache unterscheiden und wird im allgemeinen Sprachgebrauch auch als „Dialekt“ bezeichnet.

Es ist auch möglich die *Volkssprache* der Kroaten als „Kroatisch“ zu bezeichnen. Dies kann aus dialektologischer Sicht aber nur als Oberbegriff verstanden werden. Der kroatische Sprachraum ist nicht einheitlich, sondern gliedert sich in drei Hauptdialekte (oder Dialektgruppen): *Kajkavisch*, *Čakavisch* und *Štokavisch*. Die Bezeichnungen für diese drei Hauptdialekte leiten sich vom jeweiligen Interrogativpronomen für „was“ in diesen drei Dialekten ab: *kaj* im Kajkavischen, *ča* im Čakavischen und *što* im Štokavischen. Dieses Wort stellt aber keineswegs den einzigen Unterschied zwischen diesen drei Dialekten dar. Sie unterscheiden sich teilweise so stark voneinander, daß sich ihre Sprecher nur schwer mit den Sprechern eines der anderen beiden Dialekte verständigen können.

Innerhalb dieser drei Hauptdialekte lassen sich noch weitere Aufteilungen vornehmen. Ein besonders für das Štokavische wichtige, in der Literatur zu findende Kriterium zur Unterteilung stellt das Entwicklungsergebnis des altslavischen Phonems dar, das mit *Jat* bezeichnet wird (in der Literatur meist als *ě* geschrieben) (siehe Abschnitt 3). Dieser Laut hat sich in einigen Gebieten zu einem [ɛ] entwickelt, in anderen zu einem [i], und wieder anderen zu

einem komplexen System von abwechselndem [iɛ] - [jɛ] - [ɛ] - [i]. Nach dieser Entwicklung spricht man von *ikavischen*, *ekavischen* und (*i*)*jekavischen* Dialekten. Abbildung 1.2 soll diese Einteilung anhand einiger weniger Beispiele verdeutlichen.

Jekavisch	Ikavisch	Ekavisch	Deutsch
brijeg ¹	brig	breg	<i>Berg N sg</i>
bregovi / brjegovi	brigovi	bregovi	<i>Berge N pl</i>
mlijeko	mliko	mleko	<i>Milch N sg</i>
mljekar (/mʎɛkar/)	mlikar	mlekar	<i>Milchmann</i>
snijeg	snig	sneg	<i>Schnee</i>
sjeći	sići	seći	<i>schneiden</i>
siječem	sičem	sečem	<i>schneiden 1sg</i>
rijeka	rika	reka	<i>Fluß</i>

¹ Die Zeichenkette „ije“ steht für das so genannte „lange Jat“, das in der heutigen (jekavischen) Sprache dem Diphthong /iɛ/ entspricht. Siehe Abschnitt 3.1 zum Lautinventar und Abschnitt 2.2 für die Graphem-Phonem-Beziehungen des Kroatischen.

Abbildung 1.2: Vergleich Jekavisch — Ikavisch — Ekavisch

Das Štokavische wird außerdem noch in Neuštokavisch und Altštokavisch geteilt. Diese Unterteilung wird anhand der Ausprägung des Wortakzents vorgenommen. In den verschiedenen Dialekten gibt es eine Vielzahl an Akzentsystemen mit teilweise bis zu fünf unterschiedlichen Wortakzenten. Beim Neuštokavischen hat sich ein System von vier Wortakzenten herausgebildet.

Das Štokavische ist der einzige Dialekt, der sowohl von Kroaten, als auch von Bosniern und Serben gesprochen wird und diente (wohl genau aus diesem Grund) als Basis für die Standardsprache. Kajkavisch, das einige Ähnlichkeiten mit dem benachbarten Slowenisch aufweist⁷, und Čakavisch werden nur von Kroaten gesprochen.

Die Dialekte teilen den kroatischen Sprachraum geographisch wie folgt auf:

- Ikavisches Neuštokavisch wird in weiten Teilen Dalmatiens und entlang der Küste gesprochen, jekavisches Neuštokavisch im Süden bei Dubrovnik, entlang der bosnischen Grenze und in Teilen Slawoniens (im Norden Kroatiens), wo auch Altštokavisch vorkommt.
- Ekavisches Neuštokavisch wird nur im äußersten Nordosten gesprochen.
- Kajkavisch kann als überwiegend ekavisch bezeichnet werden, obwohl das Jat hier teilweise auch eine komplexere Entwicklung erfahren hat. Gesprochen wird dieser Dialekt im Nordwesten Kroatiens, mit der Hauptstadt Zagreb.

⁷Natürlich weist Kajkavisch aufgrund der Zugehörigkeit zur slavischen Sprachfamilie Ähnlichkeiten mit allen slavischen Sprachen, insbesondere den südslavischen, auf. Mit *Ähnlichkeiten* sind hier solche Gemeinsamkeiten gemeint, die speziell beim Kajkavischen und dem Slowenischen zu beobachten sind.

- Čakavisch hat sowohl ikavische wie jekavische Varianten und wird in Istrien, rund um die Kvarner Bucht, sowie in einem schmalen Streifen entlang der Küste und auf den Adriainseln gesprochen.

Es wird also deutlich, daß der Begriff „Kroatisch“ als Bezeichnung für die Volkssprache der Kroaten zu allgemein ist und nur als Oberbegriff verwendet werden könnte.

Standardsprache

Als *Standardsprache* wird eine Sprache mit vorgeschriebener (standardisierter) Grammatik und Vokabular bezeichnet, welche in allen Bereichen der sprachlichen Kommunikation und Literatur einheitlich verwendet werden kann.

So kann als weitere Möglichkeit der Begriff „Kroatisch“ auch verwendet werden um die *Standardsprache* zu Bezeichnen. In der linguistischen Literatur findet man dazu sehr häufig die Bezeichnung „*hrvatski književni jezik*“ (wörtlich: kroatische Literatursprache). Obwohl alle drei Hauptdialekte im Laufe der Geschichte als Schrift- und Literatursprachen dienten, basiert die heutige Standardsprache im wesentlichen auf der jekavischen Variante des Neuštokavischen. Die Standardisierung des Kroatischen begann etwa im 16./17. Jahrhundert und wurde im wesentlichen Ende des 19. Jahrhunderts abgeschlossen.

Insbesondere zu Zeiten Jugoslawiens wurden dann nicht nur die Literatur- sondern vor allem auch die Standardsprache „*književni jezik*“ genannt und auch heute noch ist dieser Begriff für die Standardsprache gebräuchlich. DRAGUTIN RAGUŽ bemerkt dazu im Vorwort zu seiner Grammatik (Raguž 1997):

„... tek s osamostaljenjem hrvatske države (1991. godine) to ime, *hrvatski jezik*, postaje i službenim. Stoga nema više razloga taj jezik zvati *hrvatskim književnim jezikom*, nego naprosto *hrvatskim jezikom*, onako kako to čine i drugi narodi za svoje jezike.“⁸

Einen detaillierten Überblick über die Entwicklung der Standardsprache bieten z.B. MOGUŠ (in Moguš 1991) oder die Einführung sowie das Kapitel über die sprachgeschichtlichen Veränderungen in der Grammatik von BARIĆ et. al. (Barić et. al. 1995).

Die kroatische Eigenbezeichnung für die (Standard-) Sprache ist also „*hrvatski jezik*“, oft auch „*hrvatski književni jezik*“ oder einfach nur „*hrvatski*“.

In dieser Arbeit soll der Begriff Kroatisch daher als Bezeichnung für die kroatische Standardsprache dienen.

⁸„... erst mit dem selbständig werden des kroatischen Staates (im Jahre 1991) wird dieser Name, *kroatische Sprache*, auch offiziell. Daher gibt es keinen Grund mehr diese Sprache *kroatische Literatursprache* zu nennen, sondern einfach nur *kroatische Sprache*, so wie dies auch andere Nationen für ihre Sprachen tun.“ (Übersetzung D. Duran, Hervorhebungen im Original)

Gemeinsprache

Schließlich kann noch gesagt werden, daß mit dem Begriff „Kroatisch“ aber auch die kroatische *Gemeinsprache* bezeichnet werden kann, also diejenige Sprache, in der „alle“ sprechen (allgemeine Alltags- oder Umgangssprache). Man könnte auch von einem *gefühlten Standard* sprechen (vgl. 3). Dies wäre die vierte Möglichkeit den Begriff „Kroatisch“ zu interpretieren. In wieweit sich diese Sprache von den vorhergehenden Beschreibungen unterscheidet, insbesondere von der in der Literatur beschriebenen Standardsprache, wird dann an entsprechender Stelle in dieser Arbeit wieder angesprochen werden. Eine endgültige Analyse zu diesem Thema kann hier aber nicht erfolgen.

Kapitel 2

Die geschriebene kroatische Sprache

„TTS systems conventionally treat text as a simple string of characters that correspond to the writing system of the language“
(Möbius 2001)

In diesem Kapitel werde ich die Orthographie des Kroatischen beschreiben. Zunächst soll eine kurze Beschreibung des Alphabets erfolgen, da es im Vergleich zum englischen oder deutschen einige Besonderheiten enthält. Im Anschluß soll dann das „phonologische Prinzip“ der Orthographie mit den eindeutigen Graphem–Phonem–Zuordnungen erläutert werden. Nach dieser Einführung werde ich dann spezielle Probleme der Orthographie im Zusammenhang mit der Text-to-Speech Synthese aufführen.

2.1 Das Kroatische Alphabet

Kroatisch wird mit einer Variante des lateinischen Alphabets geschrieben. Für die typisch slavischen Laute, die keine Entsprechung im Lateinischen haben, wurden im Verlauf der Geschichte unterschiedliche Schreibweisen verwendet. Dabei orientierte man sich auch am Ungarischen oder Deutschen. Seit dem 19. Jahrhundert folgt die Standardorthographie auf Initiative von LJUDEVIT GAJ hin dem Prinzip der möglichst ausnahmslosen und eindeutigen Zuordnung von Graphemen und Phonemen — also der eindeutigen Zuordnung von Buchstaben und Lauten der Sprache. Zu diesem Zweck wurden spezielle Buchstaben aus dem Tschechischen (*č, š, ž*) und Polnischen (*ć*) entlehnt und neue kreiert¹. Nach GAJ wird das kroatische Alphabet (wie übrigens auch das Slovenische) auch als „*gajica*“ bezeichnet. Ein etwas ausführlicherer Überblick über die Entwicklung des Alphabets und der Schriftsprache insgesamt ist z.B. in (Barić et. al. 1995, Uvod) oder (Moguš 1991) zu finden.

¹Der ursprüngliche Vorschlag Gajs sah die Zeichen *č̃, đ̃, ģ̃, ĺ̃, ñ̃, š̃* und *ž̃* anstelle der heutigen *č, đ, dž, lj, nj, š* und *ž* vor, fand in dieser Form allerdings keinen Zuspruch und konnte sich nicht durchsetzen. Aus computerlinguistischer Sicht wäre dies die einfachere Lösung gewesen, da so die Probleme mit den Digraphen vermieden worden wären (siehe Abschnitt 2.4).

Das heute verwendete kroatische Alphabet setzt sich aus 30 Graphemen zusammen, die wie in Abbildung 2.1 gezeigt angeordnet sind.

a	b	c	č	ć	d	dž	d	e	f	g	h	i	j	k
A	B	C	Č	Ć	D	Dž	Đ	E	F	G	H	I	J	K
l	lj	m	n	nj	o	p	r	s	š	t	u	v	z	ž
L	Lj	M	N	Nj	O	P	R	S	Š	T	U	V	Z	Ž

Abbildung 2.1: Das kroatische Alphabet

Die einzelnen Grapheme werden alle als eigenständige Zeichen angesehen und auch so verwendet: In Sortierungen (wie etwa in Wörterbüchern oder Telefonbüchern) werden die Buchstaben mit diakritischen Zeichen, nicht wie im Deutschen wie ihre Grundbuchstaben verwendet, sondern nach ihrer Position im Alphabet einsortiert. Dies gilt insbesondere auch für die Digraphen *dž*, *lj* und *nj*, die jeweils nach *d*, *l* und *n* einsortiert werden. Das bedeutet z.B. das Wort *polje* (dt. Feld) ist nach *polovica* (dt. Hälfte) und das Wort *stanje* (dt. Zustand) nach *stanovnik* (dt. Einwohner) im Wörterbuch zu finden.

Wie die einzelnen Grapheme jeweils den kroatischen Phonemen zugeordnet werden wird im nächsten Abschnitt erläutert.

2.2 Graphem–Phonem–Beziehungen

Wie bereits im vorigen Abschnitt erwähnt, folgt die kroatische Orthographie dem Prinzip der möglichst ausnahmslosen und eindeutigen Zuordnung von Graphem und Phonem. So lässt sich jedem Graphem aus dem Alphabet eindeutig ein Phonem aus dem kroatischen Phoneminventar zuordnen.

Das im vorherigen Abschnitt angesprochene Prinzip der eindeutigen Phonem–Graphem–Zuordnung würde implizieren, daß das Kroatische mindestens über 30 Phoneme verfügt. Im Kapitel 3 werden die Schwierigkeiten bei der Definition des Phoneminventars genauer beschrieben werden. Die Entsprechungen zwischen Graphemen und Phonemen, wie sie traditionellerweise in der Literatur beschrieben werden, soll die Tabelle in Abbildung 2.2 verdeutlichen. Die entsprechenden Laute sind mit IPA–Symbolen dargestellt.

a	b	c	č	ć	d	dž	d	e	f
/a/	/b/	/ts/	/tʃ/	/tɕ/	/d/	/dʒ/	/d̥/	/e/	/f/
g	h	i	j	k	l	lj	m	n	nj
/g/	/x/	/i/	/j/	/k/	/l/	/l̥/	/m/	/n/	/ɲ/
o	p	r	s	š	t	u	v	z	ž
/o/	/p/	/r/	/s/	/ʃ/	/t/	/u/	/v/	/z/	/ʒ/

Abbildung 2.2: Die Graphem–Phonem–Zuordnung

2.3 Orthographie

Im vorangehenden Abschnitt wurde die Grundlage der kroatischen Orthographie dargestellt — das so genannte phonologische Prinzip. Da sich die Orthographie aber zu einer Zeit entwickelte, zu der noch nicht strikt zwischen Phonetik und Phonologie unterschieden wurde, spielen auch andere Gesichtspunkte bei der konkreten Umsetzung des allgemeinen Prinzips im Einzelfall eine Rolle.

„Današnji je pravopis hrvatskoga književnoga jezika *fonološki*, što znači da se temelji na prije spomenutom načelu da grafemi predstavljaju razlikovne jedinice — foneme (osim nekih vrlo rijetkih odstupanja [...]), a ne njihove izgovorne realizacije.“² (Barić et. al. 1995, S. 66)

IVO ŠKARIĆ bemerkt dazu, daß die kroatische Orthographie „intuitiv“ auf den Prinzipien der Phonologie basiere, aber „bewußt“ auf denen der Phonetik. Und er schreibt weiter:

„Zbog te svjesne odluke da se »piše kao što se govori« [...], naš pravopis ima stanovitih nedosljednosti koje ga opterećuju.“³ (Škarić 1991, S. 338)

So schafft es die Orthographie in Kroatien auch hin und wieder als Thema in die Massenmedien. Die einzelnen kroatischen Orthographien verschiedener Autoren unterscheiden sich teilweise in einigen Punkten und jede neu herausgegebene Orthographie ruft neue öffentliche Diskussionen oder wenigstens mediales Interesse hervor. Da man aus computerlinguistischer Sicht aber ohnehin nicht allein mit *einer* einzigen Standardorthographie rechnen darf spielen diese Probleme keine große Rolle. Die Benutzer eines Text-to-Speech Systems werden in der Regel erwarten, daß das System ihren eingegebenen Text synthetisiert und „vorliest“ und nicht, daß es ihnen ihre falsche Rechtschreibung vorhält. Also egal ob nun *greška* oder *grješka* (dt. Fehler), *podaci* oder *podatci* (dt. Daten), *neću* oder *ne ću* (dt. (ich) will/werde nicht) dem orthographischen Standard entspricht (um nur einige strittige Beispiele zu nennen), eine TTS-Synthese fürs Kroatische muß mit Sicherheit mit beiden Formen umgehen können. Von den hier aufgezählten Beispielen werden jeweils beide Varianten in der geschriebenen Sprache verwendet.

In dem für diese Arbeit erstellten Testkorpus (siehe: Anhang D) sehen die Häufigkeiten z.B. so aus: *greška*: 2938, *grješka*: 87, *podaci*: 32485 und *podatci*: 1087. Die Ausspracheregeln müssen also eventuell auch Fälle behandeln, die nach der (einen oder anderen) Standardorthographie nicht vorkommen sollten. Die von mir entwickelten Ausspracheregeln werden in Kapitel 4 vorgestellt.

Daß die kroatische Literatur zum Thema Orthographie nicht einheitlich ist bedeutet aber nicht, daß in der kroatischen Schriftsprache orthographisches Chaos herrscht. Die Orthographie folgt weitestgehend dem phonologischen Prinzip. Dies bedeutet, daß die Wörter nach

²„Die heutige Rechtschreibung der kroatischen Sprache ist *phonologisch*, was bedeutet, daß sie auf dem bereits erwähnten Prinzip beruht, daß die Grapheme die unterscheidenden Einheiten repräsentieren — die Phoneme (außer einiger sehr seltener Ausnahmen [...]), und nicht ihre artikulatorischen Realisierungen.“ (Übersetzung DD)

³„Wegen dieser bewußten Entscheidung daß man „schreibt wie man spricht“ [...], hat unsere Rechtschreibung bestimmte Inkonsistenzen, die sie belasten.“ (Übersetzung DD)

der in Abbildung 2.2 gezeigten Graphem–Phonem–Zuordnung geschrieben werden. Einige Beispiele für dieses Prinzip: /ɔ̌zakovo/ = Đakovo, /gospit̩c/ = Gospić, /kɔ̌rtʃula/ = Korčula, /kr̩k/ = Krk, /ʃib̩enik/ = Šibenik, /zagreb/ = Zagreb usw. (alles Städtenamen).

Auf Wortebene fließen aber auch phonetische Phänomene in die Schrift ein. Wo es in der Flexion bei Konsonanten zu einer Angleichung der Stimmhaftigkeit kommt wird dies (in vielen Fällen) auch in der Schrift wiedergegeben. So schreibt man z.B. *izgraditi* (dt. ausbauen — mit dem Morphem /iz-/ als Präfix) und *iskopati* (dt. ausgraben — was mit dem selben Morphem /iz-/ gebildet wurde), *vrabac* und *vrapca* (nicht: **vrabca*, dt. Spatz, Nsg und Gsg), *sladak* und *slatko* (nicht: **sladko*, dt. süß, m und n), *težak* und *teško* (nicht: **težko*, dt. schwer, m und n).

Auch die Angleichung der Artikulationsstelle wird in der Schrift (in einigen Fällen) wiedergegeben: *lišće* < *list* (nicht: **lišce*, dt. Blätter, Laub — Blatt), *nošnja* < *nositi* (nicht: **nosnja*, dt. Tracht — tragen), *grožđe* < *grozd* (nicht: **grozde*, dt. Trauben — Traube), *oraščić* < *orah* (nicht: **oraščić*, dt. Nüßchen — Nuß) usw.

Wenn nach der Flexion oder anderen morphologischen Prozessen zwei gleiche Konsonanten aufeinander treffen würden fällt in einigen Fällen einer weg und dies wird auch so geschrieben: *bezvučan* < *bez + vučan* (nicht: **bezzvučan*), *odahnuti* (nicht: **oddahnuti*) usw.

Auch Kombinationen verschiedener (phonetischer) Regeln können im Schriftbild vorkommen, wie z.B. in *bežični* < **bežični* < **bežični* < *bez + žični* (dt. -los + Kabel) oder *engleski* < **englesski* < **englezski* < *enlgez + -ski* (dt. englisch) usw.

Betrachtet man nur diese wenigen Beispiele könnte man den Eindruck gewinnen, daß eine phonologische Transkription kroatischer Texte eine triviale Aufgabe darstellt. Daß dem aber nicht so ist soll in den folgenden Abschnitten gezeigt werden.

2.3.1 Groß- und Kleinschreibung

Im Kroatischen werden nur Eigennamen und Wörter am Satzanfang sowie Wörter zur Kennzeichnung von Höflichkeit (Titel, Anrede in Briefen usw.) groß geschrieben.

Von mehrteiligen Eigennamen wird nur der erste Teil (das erste „Wort“) groß geschrieben, wie etwa die Namen *Ministarstvo znanosti, obrazovanja i športa* (dt. Ministerium für Wissenschaft, Bildung und Sport), *Hrvatsko društvo za primijenjenu lingvistiku* (dt. Kroatische Gesellschaft für angewandte Linguistik) oder *Hrvatski savez slijepih* (dt. Kroatischer Blindenverband). Eine Ausnahme stellen solche Bestandteile von mehrteiligen Eigennamen dar, die auch allein stehend groß geschrieben würden, wie z.B. *Hrvatske* in *Socijaldemokratska partija Hrvatske* (dt. Sozialdemokratische Partei Kroatiens). Die Bestandteile mehrteiliger Eigennamen, die für Länder, Städte oder andere Lokalitäten stehen (Toponyme), werden alle mit großem Anfangsbuchstaben geschrieben: *Donja Saska* (dt. Niedersachsen), *Sjedinjene Američke Države* (dt. Vereinigte Staaten von Amerika) oder *Velika Britanija* (dt. Großbritannien). Eine Ausnahme von dieser Regel stellen Konjunktionen und Präpositionen in solchen Eigennamen dar, wie z.B. in *Srbija i Crna Gora* (dt. Serbien und Montenegro), *Sveti Petar u Šumi* (Gemeinde in Istrien, dt. Sankt Peter im Wald) oder *Biograd na Moru* (Stadt in Kroatien, wörtlich: Biograd am Meer).

Alle anderen Wörter werden im Kroatischen klein geschrieben.

2.3.2 Abkürzungen und Akronyme

Abkürzungen sind Textbestandteile, die nicht in ihrer vollen Form ausgeschrieben werden. Sie werden, wie auch im Deutschen, im Kroatischen gewöhnlich durch einen Punkt gekennzeichnet. Damit gibt es neben dem Punkt als Zeichen für das Satzende und als Zeichen für Ordinalzahlen drei verschiedene Bedeutungen für dieses Zeichen, das ein TTS-System unterscheiden können muß. (Vgl. Abschnitt 2.7 zu den Ordinalzahlen)

Es gibt eine Reihe von üblichen Abkürzungen, die immer wieder in Texten zu finden sind. Zu den Abkürzungen mit einem Punkt gehören z.B. *čl.* (*članak*) (dt. Artikel (Gesetz)), *g.* (*godina / gospodin*) (dt. Jahr / Herr), *itd.* (*i tako dalje*) (dt. und so weiter), *npr.* (*na primjer*) (dt. zum Beispiel), *str.* (*strana*) (dt. Seite), *tj.* (*to jest*) (dt. das heißt) oder *ul.* (*ulica*) (dt. Straße). Abkürzungen für physikalische Maßeinheiten (*kg, km, m* usw.) oder die Abkürzung für die kroatische Währung Kuna (*kn*) sowie einige traditionelle Abkürzungen (*gđa* (*gospoda*) (dt. Frau) oder *fra* (*fratar*) (dt. Ordensbruder)) werden normalerweise ohne Punkt geschrieben.

Abkürzungen werden beim Lesen aufgelöst, d.h. sie werden durch die Ausdrücke ersetzt, für welche sie stehen. Schwierigkeiten, mehr für den Computer als für den menschlichen Leser, ergeben sich dann, wenn Abkürzungen für Ausdrücke mit veränderlichen Wortarten stehen. Diese müssen nicht nur aufgelöst, sondern gegebenenfalls auch korrekt dekliniert werden.

Akronyme sind Textbestandteile, die aus den jeweiligen Anfangsbuchstaben von mehrteiligen Ausdrücken zusammengesetzt sind, für welche sie stehen. Sie werden in Großbuchstaben geschrieben. Akronyme stehen in der Regel für Eigennamen und gehören damit im Grunde zu den veränderlichen Wortarten im Kroatischen. Allerdings werden Akronyme beim Lesen nicht wie die Abkürzungen in jedem Fall aufgelöst. Beim Vorlesen gibt es zwei übliche Vorgehensweisen: (1.) die Akronyme werden „buchstabiert“ oder (2.) sie werden wie gewöhnliche Wörter behandelt.

Buchstabenweise gelesen werden in der Regel Akronyme, die nach den Regeln der kroatischen Phonetik keine kroatischen Lautfolgen darstellen können, da sie z.B. nur aus Konsonanten bestehen. Einige Beispiele hierfür sind im oberen Teil der Abbildung 2.3 zu sehen. Betont werden solche Akronyme, entgegen der standardsprachlichen Norm, auf der letzten Silbe (vgl. Barić et. al. 1995; Škarić 1991): *SAD* > /ɛsa'dɛ:/.

Als Wörter werden Akronyme gelesen, wenn dies die Phonetik des Kroatischen zuläßt (wenn etwa Vokale enthalten sind). Einige Beispiele hierfür sind im unteren Teil der Abbildung 2.3 zu sehen. Es gibt aber auch Ausnahmen, wie z.B. *BiH* (*Bosna i Hercegovina*) (dt. Bosnien und Herzegowina) was normalerweise nicht als /bix/ sondern als /beixa/ gelesen wird, oder das oben schon erwähnte *SAD* (*Sjedinjene Američke Države*) (dt. Vereinigte Staaten von Amerika).

Es gibt zwei verschiedene Arten, die Flexion bei Akronymen zu kennzeichnen. Dies geschieht einerseits durch das Anhängen des jeweiligen Kasus-Suffixes mit einem Bindestrich an das Akronym (*PDV* > *PDV-a, PDV-u* usw.), oder durch ausschreiben der flektierten Form (*SDP* > *esdepeju, esdepeovci* usw.).

Akronym	(mögliche) Bedeutung	Aussprache	Übersetzung
BiH	Bosna i Hercegovina	/bɛi'xa:/	Bosnien und Herzegowina
HDZ	Hrvatski dijalektološki zbornik	/xadɛ'zɛ:/	Kroatische dialektologische Arbeiten (Zeitschrift)
PDV	Porez na dodanu vrijednost	/pɛdɛ've:/	Mehrwertsteuer
RH	Republika Hrvatska	/ɛr'xa:/	Republik Kroatien
SAD	Sjedinjene Američke Države	/ɛsa'dɛ:/	Vereinigte Staaten von Amerika
SDP	Socijaldemokratska partija Hrvatske	/ɛsdɛ'pɛ:/	Sozialdemokratische Partei Kroatiens
HAZU	Hrvatska akademija znanosti i umjetnosti	/xazu/	Kroatische Akademie der Wissenschaften und Künste
HINA	Hrvatska informativna agencija	/'xina/	Kroatische Nachrichtenagentur
MUP	Ministarstvo unutarnjih poslova	/'mup/	Innenministerium
NATO	(Organizacija sjevernoatlantskog ugovora)	/'natɔ/	NATO

Abbildung 2.3: Einige Akronyme

2.4 Vokale und andere Probleme

Wie im Abschnitt 3.2 genauer dargestellt werden wird, gibt es im (standard-) Kroatischen vier lexikalische Wortakzente. Trotz des phonologischen Prinzips als Grundlage der Orthographie werden diese Akzente in der Schrift aber nicht wiedergegeben. Weder die Akzentstelle (welche im Kroatischen nicht fest ist) noch die Akzentart werden in der Schrift markiert. Eine derartige Schreibweise mit markierter Akzentstelle und Akzentart würde die fünf Vokale, den Diphthong /iɛ/ und das silbische /r/ betreffen, da dies diejenigen Phoneme sind auf welche der Wortakzent fallen kann. Bakran schreibt dazu folgendes (Bakran 1996a, Seite 249):

„Četiri normativno različita akcenta riječi u hrvatskom nemaju dovoljno važnu razlikovnu funkciju i, možda baš zbog toga, u pismu se ne bilježe“⁴

Die Vokale sowie das silbische [r] werden nur mit ihren einfachen Graphemen geschrieben (und das [r] sogar mit demselben Graphem wie das nicht-silbische [r]). Weder die Stelle oder der Ton des Wortakzents noch die Länge werden in der Schrift wiedergegeben. Dies führt insbesondere dann zu Ambiguitäten in der Schrift, wenn sich Wörter nur im Akzent oder der Länge unterscheiden. In solchen Fällen kommt es zu so genannten Homographen (siehe Abschnitt 2.8).

Ein weiteres Problem entsteht da, wo das Prinzip der eins-zu-eins Zuordnung von Phonemen und Graphemen nicht vollständig eingehalten wurde. Dies betrifft wie schon erwähnt die Laute [ɾ] und [r], aber auch die zusammengesetzten Grapheme *dž*, *lj* und *nj* für /dʒ/, /ɫ/ und /ɲ/. Die Grapheme *dž*, *lj* und *nj* stehen laut Literatur bis auf einige wenige Ausnahmen immer für die entsprechenden Phoneme:

⁴„Die vier normativ unterschiedlichen Wortakzente haben im Kroatischen keine ausreichend wichtige distinktive Funktion und werden, vielleicht gerade deswegen, in der Schrift nicht wiedergegeben.“ (Übersetzung DD)

„Teškoća u čitanju nema jer u pismu nema skupine *lj* koju bi trebalo čitati *l-j*, a *n-j* i *d-ž* nalaze se samo na sastavu složenica...“⁵ (Babić et. al. 1996)

Nur an Morphemgrenzen kann es zu den Folgen /nj/ oder /dž/ kommen, wie z.B. in *izvanjezični* < *izvan* + *jezični* (dt. außersprachlich), *izvanjadranski* (dt. außeradriatisch) oder *podžupan* < *pod* + *župan* (dt. wörtl. Unterkreisrat).

Allerdings gibt es auch Beispiele in der Literatur nach denen es in einigen Fällen auch eine Folge /lj/ geben kann. Dies seien ausschließlich Fremdwörter, wie z.B. *reljef* = /rɛl.jɛf/, *ateljé* = /a.tɛl.jɛ/. Ebenfalls in Fremdwörtern kann es außerdem auch noch zur Folge /nj/ kommen, wie etwa in *injekcija* = /in.jɛk.tsi.a/, *konjugacija* = /kɔn.ju.ga.tsi.a/ oder *konjunktura* = /kɔn.jun.k.tu.ra/.

Ebenso problematisch ist die Schreibweise des Diphthongs /iɛ/ als Trigraph *ije*. Hier tritt dasselbe Problem auf, wie bei den zusammengesetzten Graphemen *dž*, *nj* und *lj*: in einigen Fällen steht die Zeichenfolge *ije* für die Phonemfolge /i.jɛ/ bzw. /i.ɛ/ und in anderen Fällen ist sie als Diphthong /iɛ/ auszusprechen. Einige Beispiele, die die Schwierigkeiten verdeutlichen sollen sind in Abbildung 2.4 dargestellt.

einsilbiges /iɛ/	zweisilbiges /ijɛ/
bijel	kutije
mlijeko	nijedan
snijeg	pijem
svijet	županije

Abbildung 2.4: Einsilbige und zweisilbige Aussprache von *ije*

2.4.1 Häufige Zusatzbuchstaben

Wie im Abschnitt 2.1 dargestellt wurde gibt es in der kroatischen Schrift drei verschiedene Diakritika, den Strich (-), den Akut (´) und den Hatschek (ˇ). Diese verbinden sich mit den Grundbuchstaben (Basisgraphemen) *c*, *d*, *s* und *z* zu den Graphemen *č*, *ć*, *đ*, *š* und *ž*.

Um einigen der im vorhergehenden Abschnitt angesprochenen Probleme zu begegnen sind zusätzliche Diakritika und damit auch zusätzliche Buchstaben in Gebrauch, die auch in einem Eingabetext für ein TTS-System auftreten könnten.

Eine häufig auftretende Schwierigkeit in der Schrift ist der Genitiv. Er ist im Kroatischen, wie in vielen slavischen Sprachen, ein häufig gebrauchter Fall. In der Deklination kann es aber wie in Abschnitt 3.2 noch genauer beschrieben werden wird zu Akzentänderungen kommen, welche dann den einzigen Unterschied zu anderen Formen darstellen. So kann es aufgrund der nicht schriftlich widergegebenen Akzentuierung zu Homographen kommen. Einige Beispiele hierfür sind in Abbildung 2.5 dargestellt. Die einzelnen Wörter sind mit Akzentzeichen versehen, wie sie in der kroatischen Literatur verwendet werden. Diese diakritischen Zeichen gehören aber nicht zur Standardorthographie — vgl. Abschnitt 2.9.3.

⁵„Beim Lesen gibt es keine Schwierigkeiten, da es in der Schrift keine Gruppe *lj* gibt, welche man als *l-j* lesen müsste, und *n-j* und *d-ž* finden sich nur an Kompositionsgrenzen...“ (Übersetzung DD)

		deutsch
dāna (Gsg)	dānā (Gpl)	Tag
jābuka (Nsg)	jābūkā (Gpl)	Apfel
knjīga (Nsg)	knjīgā (Gpl)	Buch
kōža (Nsg)	kōžā (Gpl)	Haut
kŕvi (Npl)	kŕvī (Gpl)	Blut
ljúbavi (Npl)	ljúbavī (Gpl)	Liebe
nōga (Nsg)	nōgā (Gpl)	Bein, Fuß
pūtnīka (Gsg)	pūtnīkā (Gpl)	Reisender
rūka (Nsg)	rūkā (Gpl)	Hand
sŕca (Gsg)	sŕcā (Gpl)	Herz
žēna (Nsg)	žēnā (Gpl)	Frau

Abbildung 2.5: „Akzenthomographen“

Wie die Beispiele in Abbildung 2.5 verdeutlichen gibt es in der Deklination Formen, die sich nur durch die Akzente oder auch nur durch die Länge unterscheiden. So muß man als Leser bei einem Wort wie *dana* den Kontext kennen um zu wissen, ob es sich um einen Tag oder mehrere handelt. Aufgrund des Wortes allein ist dies in der Standardorthographie nicht erkennbar. Um das Lesen von Texten zu erleichtern und Ambiguitäten zu vermeiden werden daher gelegentlich die Buchstaben *â* und *î* mit Zirkumflex verwendet. Diese Buchstaben werden aber nicht als eigenständige Buchstaben betrachtet und zum Alphabet gezählt. Selbst in der linguistischen Literatur werden sie für gewöhnlich ohne besondere zusätzliche Erklärung stillschweigend verwendet. Der Zirkumflex gehört nicht zur Standardorthographie des Kroatischen und wird auch nicht in allen möglichen Fällen verwendet. Einige Beispiele für die Häufigkeit der Verwendung im Genitiv Plural dieses Zeichens aus dem bereits erwähnten Testkorpus (Anhang D) sind in Abbildung 2.6 dargestellt.

Freq.	Token	
73	narodâ	dt. Nation, Volk
62	nagradâ	dt. Auszeichnung
53	zakonâ	dt. Gesetz
49	hrvatâ	dt. Kroatie
43	vjernikâ	dt. Gläubiger
33	vremenâ	dt. Zeit
29	tvrtkâ	dt. Firma
26	riječî	dt. Wort
26	zvjezdâ	dt. Stern
23	jezikâ	dt. Sprache
19	ljudî	dt. Leute, Menschen
11	obiteljî	dt. Familie

Abbildung 2.6: Kennzeichnung der Genitivlänge durch den Zirkumflex

Neben der Verwendung von *â* und *î* als orthographische Zeichen zur Markierung des Genitivs, wie in der Orthographie (Babić et. al. 1996) beschrieben, wird das *ˆ* Zeichen auch noch in anderen Fällen verwendet. Außer der Genitivlänge gibt es noch weitere Fälle in denen es aufgrund verschiedener Akzente zu Ambiguitäten kommt (vgl. Abschnitt 2.8). Der bei weitem häufigste Fall dieser Art ist das Wort *sam*. Genau genommen handelt es sich um zwei verschiedene Worte: das unbetonte *sam* stellt die erste Person Singular des Verbs *biti* (dt.

sein) dar und das mit einem langen fallenden Akzent betonte *sam* (mit Akzentzeichen: *sâm*) bedeutet selbst, allein. Wie bereits erwähnt, werden die Akzente in der Schrift normalerweise aber nicht wiedergegeben. Um diese Wörter dennoch unterscheiden zu können wird das zweite gelegentlich *sâm* geschrieben. Dies liegt wohl daran, daß *â* in einfachen 1-Byte Zeichensätzen wie z.B. ISO-8859-2 oder Windows Codepage 1250 enthalten ist. Das Zeichen *â* (mit einem „runden Zirkumflex“) hingegen ist nur in Unicode definiert und wird daher wohl nicht (oder nur selten) verwendet (siehe Abschnitt 1.2). In meinem Testkorpus kommt das Wort (bzw. die Zeichenkette) *sâm* immerhin 2597 mal vor. Ähnlich sieht es mit *samo* aus. Hier gibt es auch zwei Bedeutungen. Mit einem langen fallenden Akzent (*sâmo*) stellt es die neutrum Form von *sam* (dt. selbst/allein) dar. Mit einem kurzen fallenden Akzent (*sămo*) stellt es das Partikel *nur/bloß* dar. Daher wird auch in diesem Fall das Wort mit dem langen Akzent analog zu *sâm* oft *sâmo* geschrieben (75 Vorkommen im Testkorpus). Interessanterweise kommen auch andere Formen von *sam* im Testkorpus vor, die feminine Form *sâma* z.B. insgesamt 53 mal, obwohl diese Formen eindeutig sind und ohne weiteres ohne Zirkumflex geschrieben werden könnten.

Eine andere sehr häufige Verwendung des Zirkumflexes ist das Wort *kôd* (dt. Code) — das mit einem langen fallenden Akzent betont wird und mit den literarischen Akzentzeichen korrekterweise *kôd* geschrieben werden müßte (vgl. Abschnitt 2.9.3). Ohne Akzentzeichen geschrieben ist es von der Präposition *kod* (dt. an/bei), mit einem kurzen fallenden Akzent, nicht zu unterscheiden. Die Form *kod* erscheint 254508 mal, die Form *kôd* 155 mal im Testkorpus.

Diese Beispiele zeigen, daß Vokale mit Zirkumflex vorkommen können, und auch wichtige Informationen für die Textanalyse liefern. Es ist also sinnvoll, diese Fälle in einem TTS-System zu behandeln, wenn z.B. Texte aus dem Internet synthetisiert werden sollen.

2.4.2 Sonstige Zusatzbuchstaben

Außer diesen Zeichen zur Markierung der Vokallänge oder des (fallenden langen) Wortakzents gibt es noch weitere Buchstaben mit denen ein TTS-System unter Umständen umgehen können muß. Zunächst einmal könnte man sich Unicode Zeichen als Eingabe vorstellen.

Im Unicode Standard sind im Abschnitt *Latin Extended-B*⁶ gleich zwei Bereiche für speziell kroatische Bedürfnisse definiert: 1. „Croatian digraphs matching Serbian Cyrillic letters“ und 2. „Additions for Slovenian and Croatian“. Die an dieser Stelle definierten Buchstaben gehen über die gewöhnlichen Zeichen des kroatischen Alphabets hinaus könnten aber in allen in Unicode kodierten elektronischen Texten verwendet werden.

Unter „Croatian digraphs...“ sind die kroatischen Digraphen *dž*, *lj*, *nj* usw. als einzelne Zeichen definiert. Damit ist es theoretisch möglich alle Wörter, welche einen oder mehrere der drei Digraphen enthalten, auf eindeutige Weise zu kodieren. So ist es möglich den Unicode Zeichen 01C6 (*dž*), 01C9 (*lj*) und 01CC (*nj*) eindeutig die Phoneme /dʒ/, /lʲ/ und /nj/ entsprechend zuzuordnen. Den Zeichenfolgen 0064 + 017E (*d + ž*), 006C + 006A (*l + j*) und 006E + 006A (*n + j*) hingegen, könnten in einem so kodierten Text immer eindeutig die Phonemfolgen /d ʒ/, /l j/ und /n j/ entsprechend zugewiesen werden.

Unter „Additions for Slovenian and Croatian“ sind die Vokale und das *r* mit speziellen Dia-

⁶Internet: <http://www.unicode.org/charts/PDF/U0180.pdf>

kritika, dem doppelten Gravis (˘) und dem invertierten Brevis (˘), enthalten. Diese dienen der Markierung des Akzentes (wie in Abschnitt 3.2 erklärt wird). Diese zusätzliche Definition war notwendig, da solche Diakritika sonst in keinen anderen bereits in Unicode definierten Alphabeten anderer Sprachen verwendet werden.

Zu diesen hier genannten zusätzlichen Unicode Zeichen ist aber zu sagen, daß sie im Grunde überhaupt nicht verwendet werden. Zumindest kann gesagt werden, daß sie in Internet-Texten nicht vorkommen, da in meinem bereits viel zitierten Testkorpus kein einziges vorkommen gefunden wurde und auch eine Suche z.B. mit Google nach Wörtern mit diesen Zeichen keine Treffer liefert(e). Eine Kodierung in HTML wäre allerdings überhaupt kein Problem und moderne Internet-Browser sind auch in der Lage diese Zeichen darzustellen. Ob sich diese Zeichen jemals durchsetzen werden bleibt abzuwarten. In jedem Fall sollten sie beim Entwurf eines TTS-Systems aber berücksichtigt und ihre Verwendung für den gegebenen Einsatzbereich des Systems untersucht werden.

Neben diesen speziellen Zeichen für die Digraphen und die akzentuierten Vokale und das *r* werden vor allem in der linguistischen Literatur noch weitere Buchstaben verwendet. Dies sind die Zeichen, die für die phonologische Transkription verwendet werden. Diese sind zum größten Teil auch in Unicode definiert und könnten daher auch in verschiedenen Quellen wie E-Mails, Newsgroup-Beiträgen oder Internetseiten vorkommen. Eine genaue Auflistung dieser Zeichen ist in Abschnitt 2.9 zu finden.

Auch für die meisten dieser Buchstaben gilt, daß sie in „gewöhnlichen“ Alltagstexten eigentlich nicht vorkommen. Eine besondere Behandlung könnte aber, je nach Systemanforderung, notwendig sein.

2.5 Allgemeiner Schreibstil

Im vorangehenden Abschnitt wurde die Orthographie genauer vorgestellt und ihre für eine TTS Synthese wichtigen Aspekte. In diesem Abschnitt sollen weitere Erscheinungen in kroatischen Texten vorgestellt werden, die bei der Entwicklung von TTS-Systemen spezieller Aufmerksamkeit bedürfen.

2.5.1 „ASCII-Schreibweise“

In Newsgroups, Internetdiskussionsforen oder in E-Mails ist es weitverbreitet die diakritischen Zeichen einfach wegzulassen und nur die Grundbuchstaben zu schreiben (*c* statt *č* usw.). Dies hatte früher vor allem technische Gründe, so daß man sich im elektronischen Schriftverkehr mit dem einfachen ASCII-Zeichensatz begnügte oder begnügen musste. Die mit Diakritika versehenen Buchstaben werden nicht umgeschrieben, anders als z.B. im Deutschen wo *ö* zu *oe* wird.

Der einzige Buchstabe, der umgeschrieben wird ist *d*, der zu *dj* wird. Dies hat historische Gründe, da dieser Buchstabe bis vor einiger Zeit generell als Digraph geschrieben wurde.

Häufig ist diese Schreibweise ohne Diakritika auf kroatischen Internetseiten aus dem Ausland oder auch auf privaten Seiten zu finden.

Sprechern des Kroatischen bereitet so eine Schreibweise im Grunde keine Probleme. Computerprogramme sind aber leider nicht so flexibel. Soll ein System zum Vorlesen von E-Mails entwickelt werden muß ein Mechanismus eingebaut werden, um mit dieser vereinfachten Schreibweise umgehen zu können.

Die Anbringung von Diakritika ließe sich z.B. mit einem Modul zur Textnormalisierung in der Vorverarbeitung realisieren. Die zweite Möglichkeit besteht darin den Text so unverändert zu verarbeiten. Dafür müssen das Lexikon und die Ausspracheregeln entsprechend entworfen werden. In vielen Fällen wäre eine Bestimmung der korrekten Aussprache schon anhand eines Lexikons möglich. Zeichenketten wie etwa „ce“, „jos“, „moze“ oder „vec“ können eindeutig den Wörtern „će“, „još“, „može“, und „već“ zugeordnet werden⁷.

Leider kann es durch diese Schreibweise auch zu Homographen kommen (siehe Abschnitt 2.8). Der String „sto“ kann für *sto* (dt. hundert) oder *što* (dt. was) stehen; „vas“ für *vas* (Personalpronomen, 2. Person pl GA) oder *vaš* (Possesivpronomen, 2. Person pl m). Wie man sieht werden in solchen Fällen kompliziertere Analysen nötig, als ein einfacher Lexikonvergleich.

2.6 Fremdwörter

Fremdwörter können für die Formalisierung von Ausspracheregeln eine zusätzliche Schwierigkeit darstellen, da sie in einigen Sprachen mit ihrer ursprünglichen Schreibweise übernommen werden und dadurch von den sprachtypischen Ausspracheregeln abweichen können. Dies trifft teilweise auch für das Kroatische zu.

Bei der Aufnahme von Fremdwörtern ins Kroatische können zwei unterschiedliche Phänomene unterschieden werden: Zum einen ist das die Schreibung von Fremdwörtern, die keine Eigennamen sind, und zum anderen die Schreibung von fremden Eigennamen.

2.6.1 Fremdwörter, die keine Eigennamen sind

Fremdwörter, die keine Eigennamen darstellen, werden im Kroatischen nach ihrer Aussprache in der Sprache übernommen, aus der sie entlehnt wurden. Geschrieben werden sie dabei in der Regel nach den Regeln der kroatischen Orthographie. Werden Wörter mit Lauten übernommen, die es im kroatischen Lautinventar nicht gibt, so werden diese fremden Laute an die kroatischen angepaßt (vgl. Abschnitt 3.1).

Einige Beispiele für solche (teilweise nur regional verwendeten) Fremdwörter sind z.B. aus dem Deutschen *regal*, *šminka*, *šport*, *špek*, *šlag*, *tipfeler*, oder *vaga*. Aus dem Englischen stammen Wörter wie *fer*, *kompjutor*, *imidž*, *turizam*, *trener* oder *softver* und aus dem Französischen z.B. *feljton*, *kreten*, *kupon*, *plaža*, *žanr* oder *žeton*.

Was die Text-to-Speech Synthese angeht stellen solche Wörter keine zusätzlichen Probleme dar, da sie im Kroatischen so weit integriert sind, daß sie sich mit den allgemeinen Ausspracheregeln behandeln lassen. Sie werden nach den allgemeinen Regeln der kroatischen

⁷Diese vier Wortformen stellen die vier häufigsten Formen ohne Diakritika im Testkorpus dar. Die Häufigkeiten sind dabei wie folgt: ce 98.221, jos 62.516, moze 45.615 und vec 45.464 — die dazugehörigen orthographisch korrekten Formen: će 1.868.593, još 705.162, već 521.052 und može 419.735.

Orthographie geschrieben, was im Umkehrschluß auch bedeutet, daß sie wie andere kroatische Wörter ausgesprochen werden.

Es gibt allerdings auch Fälle, in denen ein Wort nicht an das Kroatische angepaßt wird. So ist z.B. der oben bereits erwähnte Ausdruck *softver* auch als *software* in kroatischen Texten zu finden — im Testkorpus (vgl. Anhang D) tritt die englische Schreibweise sogar in etwa doppelt so häufig auf, wie die kroatisierte: 4143 mal *software* und 1753 mal *softver*. Beim Wort *fer* verhält es sich wieder anders. Hier sieht die Verteilung wie folgt aus: 8282 mal *fer* und 2656 mal tritt das Wort in der Form *fair* auf. Die ursprüngliche Schreibweise wird auch oft als Stilmittel verwendet um den Fremdwortcharakter oder andere Assoziationen hervorzuheben (Babić et. al. 1996, S. 57). Als Beispiel sei das Wort *cool* genannt. Es erscheint im Testkorpus 6590 mal in der Schreibweise *cool* und 1260 mal in der Schreibweise *kul*.

Andere Ausdrücke, die zu internationalem kulturellem oder technischem Fachvokabular gezählt werden können, werden häufig oder auch fast ausschließlich in ihrer ursprünglichen Schreibweise verwendet. Einige Beispiele hierfür sind *adagio*, *copyright*, *interview*, *mail*, *show*, *web*, usw.

Die Aussprache dieser Wörter könnte in einem TTS-System mit Hilfe eines Aussprachelexikons bestimmt werden. Für solche Fremdwörter allgemeine Regeln zu formulieren scheint mir wenig aussichtsreich zu sein, da die Aussprache sich zum einen nach den Aussprache- bzw. Rechtschreibregeln der ursprünglichen Fremdsprache richtet (z.B. deutsch, englisch usw.) und zum anderen richtet sich die Aussprache nach dem kroatischen Lautinventar und der Phonologie des Kroatischen.

2.6.2 Fremde Eigennamen

Bei fremden Eigennamen sieht die Praxis im Kroatischen anders aus. Diese werden in ihrer ursprünglichen Schreibweise übernommen, sofern sie aus einer Sprache stammen, die mit einem lateinischen Alphabet geschrieben wird. Bei Sprachen mit anderen Schriften wird für gewöhnlich die dort übliche romanisierte Form übernommen. Für die Aussprache gelten gewöhnlich aber die gleichen Regeln, wie für die erste Gruppe der Fremdwörter, welche keine Eigennamen darstellen: Sie werden mit dem kroatischen Lautinventar gemäß ihrer ursprünglichen Lautung ausgesprochen.

Bei näherer Betrachtung fällt aber auf, daß bei der Übernahme von fremden Eigennamen einige nichtkroatische Zeichen häufiger nicht „korrekt“ übernommen werden. Die korrekte Form *Wojtyła* z.B. kommt seltener in kroatischen Texten vor als die nach der ursprünglich polnischen Schreibweise eigentlich falsche Form *Wojtyla*. Ebenso verhält es sich mit dem Wort *Solidarność*, das häufiger in der kroatischen Form *Solidarnošć* zu finden ist. *Erdoğan* ist häufiger in der Schreibweise *Erdogan*, *Antonín Dvořák* in der Schreibweise *Antonin Dvorak* zu finden — aber auch in Kombinationen von kroatischer und ursprünglicher Schreibweise wie z.B. *Antonín Dvorák*, *Antonin Dvořák* usw.

Im Gegensatz dazu kommen z.B. die deutschen Umlaute in Wörtern wie *Köln*, *Schröder*, *Schäuble*, *Zürich* oder *München* nur sehr selten in den Formen *Koln*, *Schroder*, *Schauble*, *Zurich* oder *Munchen* vor (für letzteres kann auch die kroatisierte Form *Minchen* bzw. *Minhen* vorkommen). Ob dies an der stärkeren Verbreitung einiger Begriffe in den nationalen und internationalen Medien liegt, oder daran, daß einige Buchstaben eher übernommen werden

als andere, müßte genauer untersucht werden.

Das Kroatische ist eine flektierende Sprache mit einer umfangreichen nominalen Flexion, wobei unterschiedliche Formen für Numerus (Singular und Plural), Genus (Maskulin, Feminin und Neutrum) und Kasus (Nominativ, Genitiv, Dativ, Akkusativ, Instrumental, Lokativ und Vokativ) existieren. Dies betrifft auch die fremdsprachigen Eigennamen. So werden z.B. die Wortendungen *-a* und *-o*, wenn sie unbetont sind, wie kroatische Suffixe behandelt und die entsprechenden Wörter in die entsprechende kroatische Deklinationsgruppe eingereiht. Die entsprechenden Endungen werden wie im Kroatischen üblich dann natürlich auch in der Schrift widergegeben. Aus dem Wort *Chicago* im Nominativ wird im Genitiv *Chicaga* und im Lokativ *Chicagu*, *Calcutta* im Nominativ wird zu *Calcutte* im Genitiv und *Calcutti* im Lokativ (entsprechendes gilt jeweils auch für die übrigen vier Fälle). Stehen *-a* oder *-o* am Wortende in einer betonten Silbe⁸ werden sie nicht wie kroatische Suffixe behandelt sondern bleiben in den flektierten Formen stehen (Babić et. al. 1996, S. 59). (*Victor*) *Hugo* z.B. wird zu *Hugoa*, *Hugou* usw, und *Rousseau* zu *Rousseaua* und *Rousseauu* usw. Gelegentlich werden Flexionsendungen auch durch einen Bindestrich abgetrennt: *Hugo-a* bzw. *Rousseau-a*.

Aus fremdsprachigen Eigennamen können auch durch morphologische Prozesse neue Wörter wie z.B. Adjektive abgeleitet werden. Diese Wörter werden nach den gewöhnlichen kroatischen Regeln der Morphologie gebildet, aber im Stamm oft unverändert geschrieben. Einige Beispiele hierfür sind die Adjektive *bohumsko* < *Bochum*, *kölnsko* < *Köln*, *münchensko* < *München*. Kommen über das reine Anhängen des Morphems *-ski*⁹ noch weitere phonologische Prozesse hinzu entstehen Formen wie z.B. *leipziški* < *Leipzig*, *newyorški* < *New York*, *rostočki* < *Rostock* usw. Diese flektierten Wortformen werden aber auch in einer kroatisierten Schreibweise verwendet. Abbildung 2.7 stellt dies anhand einiger Beispiele mit der jeweiligen Häufigkeit¹⁰ dar.

	Freq.		Freq.
bohumski	4	bohumski	4
kölnske	18	kelnske	4
leipziški	37	la.jpciški	9
luxembourška	24	luksemburška	50
münchenski	174	minhenski	54
newyorški	505	njujorški	391
schengenske	316	šengenske	34
welški	11	velški	125

Abbildung 2.7: Flektierte Wortformen in unterschiedlichen Schreibweisen

Fazit

Zur Bestimmung der korrekten Aussprache in einem Text-to-Speech System ist, wie man aus diesen Beispielen sehen kann, ein Aussprachelexikon für die nach „nichtkroatischen“ Regeln geschriebenen Wörter notwendig. Erschwerend kommt hinzu, daß das Kroatische über

⁸Vgl. dazu Abschnitt 3.2: nach der traditionellen Beschreibung des kroatischen Wortakzentes wird die letzte Silbe in mehrsilbigen Wörtern niemals betont!

⁹Die Suffixe *-ski*, *-ska* oder *-sko* stehen jeweils für eine maskuline, feminine oder neutrale Form.

¹⁰Vgl. Anhang D (Testkorpus)

eine umfangreiche Flexion verfügt und diese, auch in Zusammenhang mit phonologischen Prozessen, zu einer großen Zahl an Wortformen führt.

2.7 Zahlen und Zahlwörter

Zahlen können Text-to-Speech Systemen Schwierigkeiten bereiten, wenn die dazugehörigen Zahlwörter (Numeralien) verschiedene Formen haben können. Im Kroatischen lassen sich die Zahlwörter zum größten Teil den veränderlichen Wortarten zuordnen, d.h. sie können dekliniert werden und dadurch verschiedene Formen haben.

2.7.1 Kardinalzahlen

Die Zahlwörter für Kardinalzahlen von 1 bis 4 haben verschiedene Wortformen, die durch den syntaktischen Kontext bestimmt werden, in dem sie stehen. Dies spielt in einem TTS-System eine Rolle, wenn Zahlen im Text nicht ausgeschrieben sind, sondern durch Ziffern dargestellt werden. In so einem Fall muß die Aussprache der Ziffernsymbole bestimmt werden, was nicht durch die selben Regeln erfolgen kann, wie bei den Buchstaben eines Textes. Der folgende Abschnitt soll dies ein wenig anhand der zu berücksichtigenden Regeln verdeutlichen.

Das Wort *jedan* (dt. eins) wird wie ein Adjektiv nach Genus, Kasus und Numerus dekliniert. Einige Beispiele hierfür:

- *jedan dan* (Nm sg, dt. ein Tag)
- *jedno pitanje* (Nn sg, dt. eine Frage)
- *jedne žene* (Gf sg, dt. eine Frau)
- *jednoj ženi* (Df sg)
- *jednu ženu* (Af sg)

Das Wort *dva* (dt. zwei) wird nur nach Kasus und Genus dekliniert, wobei die Formen für Maskulinum und Neutrum identisch sind. Ebenso werden auch die Wörter *oba* und *obadva* (dt. beide) dekliniert.

Das Wort *tri* (dt. drei) und das Wort *četiri* (dt. vier) sind noch einmal einfacher, da diese beiden nur nach dem Kasus dekliniert werden, wobei sie (wie auch schon das Wort *dva*) jeweils lediglich drei verschiedene Formen für die sieben Fälle besitzen.

Die Kardinalzahlen ab fünf stellen unveränderliche Wörter dar. Für die Zahlen von 11 bis 19 gibt es eigene Wörter im Kroatischen. Ab zwanzig werden die Zahlen durch Zusammensetzung gebildet und im Kroatischen „von links nach rechts gelesen“, ganz im Gegensatz zum Deutschen, wo z.B. 23 für *drei-und-zwanzig* steht. Die einzelnen Elemente von zusammengesetzten Zahlen können durch ein optionales *i* (dt. und) miteinander verbunden werden. Normalerweise wird dieses *i* aber nur vor der letzten Stelle eingefügt: 23 > *dvadeset (i) tri*, 532 > *petsto trideset (i) dva*, 1371 > *tisuću tristo sedamdeset (i) jedan* usw.

Was für die allein stehenden Wörter *jedan*, *dva*, *tri* und *četiri* gilt, gilt auch für damit zusammengesetzte Zahlen bzw. Zahlwörter. Diese Wörter werden auch als der letzte Teil in zusammengesetzten Zahlen wie z.B. in 21, 32, 43, 54, 181, 1001 usw. dekliniert wie oben beschrieben.

Zur Abtrennung von Dezimalstellen (bei der Schreibweise mit Ziffern) wird wie auch im Deutschen üblicherweise ein Komma verwendet. Zur übersichtlicheren Darstellung werden bei großen Zahlen die tausender Stellen durch einen Punkt oder ein Leerzeichen abgetrennt.

2.7.2 Weitere veränderliche Zahlwörter

Neben den oben bereits erwähnten veränderlichen Zahlwörtern zählen auch alle *Ordinalzahlen* (Ordnungszahlen) zu den veränderlichen Zahlwörtern, da sie wie Adjektive dekliniert werden. Wie bei zusammengesetzten Kardinalzahlen gilt auch bei den Ordinalzahlen, daß nur das letzte Element dekliniert wird. Die vorhergehenden Elemente werden in diesem Fall aus Kardinalzahlen gebildet. Geschrieben werden Ordinalzahlen wie im Deutschen entweder mit den entsprechenden Zahlwörtern oder mit Ziffern gefolgt von einem Punkt. Die Zeichenfolge 23 wird also als Kardinalzahl *dvadeset (i) tri* gelesen und die Zeichenfolge 23. je nach Kontext etwa als *dvadeset (i) treći* (NVm sg), *dvadeset (i) trećoj* (DLf sg) oder *dvadeset (i) trećim* (DLI pl).

Die Wörter *stotina* (dt. 100), *tisuća* (dt. 1000), *milijun*, *milijarda* und *bilijarda* stellen Substantive dar und werden wie feminine Nomen dekliniert.

Außerdem muß an dieser Stelle auch noch erwähnt werden, daß Jahreszahlen im Kroatischen durch Ordinalzahlen angegeben werden. Dadurch könnte man annehmen, daß Ordinalzahlen im Kroatischen, oder zumindest in kroatischen Texten, häufiger vorkommen, als in anderen Sprachen, wie z.B. im Deutschen oder Englischen. Untersuchungen zu diesem Thema liegen mir allerdings nicht vor.

Fazit

Wie das obige Beispiel mit der Ordinalzahl 23. verdeutlicht hat muß zur Bestimmung der korrekten Aussprache der entsprechenden Zeichenkette in einem Text auch der morphologische bzw. syntaktische Kontext bekannt sein. Je nach dem, um welche Zahl es sich handelt, müssen verschiedene Informationen aus dem geschriebenen Text ermittelt werden — im Fall von *jedan* wie bereits gesagt Numerus, Genus und Kasus. Hinzu kommt das Problem, daß Ordinalzahlen erst einmal von Kardinalzahlen unterschieden werden müssen, indem bestimmt wird, ob der Punkt einen Satz beendet, oder eben für eine Ordinalzahl steht.

2.8 Homonyme

Was allgemein schlicht mit dem Begriff *Homonym* bezeichnet wird kann in zwei Gruppen geteilt werden: *Homophone* und *Homographen*.

Eine (nach eigenen Angaben) erste Arbeit zur systematischen Darstellung der Homonyme im Kroatischen und Serbischen hat ANTON KNEŽEVIĆ (Knežević 1970) vorgestellt. Eine derartige Liste ließe sich heute mit computerlinguistischen Methoden ohne Probleme aus einem Aussprachelexikon extrahieren. Leider konnte ich für diese Arbeit auf kein solches Aussprachelexikon zurückgreifen.

Homophone

Homophone sind verschiedene Wörter die sich anhand ihrer Lautung nicht unterscheiden lassen. Verschieden können sie sein im Hinblick auf ihre Semantik oder ihre syntaktische Funktion. Da sich Homophone der Definition nach in der Lautung nicht unterscheiden stellen sie für ein Text-to-Speech System auch kein Problem dar. Da das TTS System gewöhnlicherweise als reiner Vorleseautomat verstanden wird muß es (in diesem Fall) keine semantische oder andersartige Disambiguierung durchführen. Die Wörter können also in diesem Fall ungeachtet ihrer Ambiguität synthetisiert werden.

Beispiele für Homophone im Deutschen wären etwa die Wörter *Bank*, *Bund/bunt*, *Lehre/Leere*, *Mutter*, *Stadt/statt*, *Wahl/Wal* usw. Einige Beispiele für kroatische Homophone sind *biti* (dt. schlagen / sein), *dan* (dt. Tag / geben, Part. pass.), *dužiti* (dt. verlängern / schulden), *gol* (dt. Tor / nackt), *grad* (dt. Stadt / Grad), *granica* (dt. Grenze / Zweig, dem.), *rujan* (dt. September / rötlich). Diese Wortpaare werden nicht nur alle gleich geschrieben, sondern tragen jeweils auch die gleichen Wortakzente, sind also echte Homophone.

Homographen

*Homographen*¹¹ sind in der Schrift völlig gleiche Wörter, die aber auch unterschiedliche Lautung haben können. Eine Schwierigkeit für die Bestimmung der korrekten Aussprache stellt die Tatsache dar, daß weder die Akzentstelle noch die Akzentart im Kroatischen in der Schrift gekennzeichnet werden.

So können im Kroatischen Homographen vor allem immer dann auftreten, wenn sich in der Flexion die Vokallänge oder die Akzentuierung ändert. Da sich in diesen Fällen auch keine Disambiguierung anhand der Wortart vornehmen läßt, müssen über das POS-Tagging hinausgehende Analysen vorgenommen werden. Wie bereits erwähnt kommt es zu solchen Änderungen besonders oft bei Nomina im Genitiv. Im Abschnitt 2.4.1 wurden spezielle Schreibweisen zur Unterscheidung von unterschiedlichen Vokallängen und Akzentuierungen vorgestellt. Da diese Schreibweise nicht zur Standardorthographie gehört und nur gelegentlich verwendet wird, sind all die angeführten Beispiele in der normalen Schreibweise ohne zusätzliche Diakritika als Homographen anzusehen.

Die in der Abbildung 2.5 aufgezählten Beispiele zeigen Akzentunterschiede, anhand der sich verschiedene Wortformen unterscheiden lassen. Es gibt allerdings auch Wörter, die sich lediglich in der Akzentuierung unterscheiden, aber nicht aus dem selben Wort durch Flexion oder andere morphologische Prozesse entstanden sind. Einige Beispiele für solche Homographen sind die in Abbildung 2.8 dargestellt. Die zweite und dritte Spalte gibt dabei jeweils die unterschiedlichen Aussprachen in IPA Notation an (vgl. Abschnitt 2.9).

¹¹Knežević spricht von „Homogrammen“

bit	/bît/ (dt. Bit)	/bî:t/ (dt. Wesen,Kern)
divan	/dīva:n/ (dt. Diwan)	/dī:van/ (dt. herrlich)
čelo	/tsê:lɔ/ (dt. Cello)	/tsě:lɔ/ (dt. Stirn)
jak	/jâk/ (dt. Yak)	/jâ:k/ (dt. stark)
kupiti	/kûpiti/ (dt. sammeln)	/kû:piti/ (dt. kaufen)
luk	/lûk/ (dt. Zwiebel)	/lû:k/ (dt. Bogen)
malina	/malina/ (dt. kleine Anzahl)	/mâlina/ (dt. Himbeere)
skup	/skûp/ (dt. Ansammlung)	/skû:p/ (dt. teuer)

Abbildung 2.8: Einige Homographen

Unechte Homographen

Ein zusätzliches Problem für TTS Systeme ergibt sich aus der Schreibweise ohne Diakritika (wie in Abschnitt 2.5.1 beschrieben). Diese „unechten Homographen“ müssen erkannt und zusätzlich zu den üblichen Analyseprozessen behandelt werden. Jedes der mit Diakritika versehenen Grapheme des Kroatischen kann davon betroffen sein. Wie „s“ in so einer „ASCII-Schreibweise“ für orthographisches *s* oder *š* stehen kann habe ich schon anhand zweier Beispiele im oben erwähnten Abschnitt 2.5.1 gezeigt. Hier nun noch die anderen Buchstaben mit Diakritika:

Die Zeichenkette „znaci“ kann für *znaci* (dt. Zeichen, N pl) oder *znači* (dt. bedeuten, 3. Person sg), „cesta“ für *cesta* (dt. Straße, N sg oder G pl) oder *česta* (dt. häufig, NAV pl) und analog dazu kann auch „cesto“ für *cesto* (dt. Straße, V¹² sg) oder *često* (dt. häufig, NAV sg) stehen.

Die Zeichenkette „reci“ kann für orthographisches *reci* (dt. sagen, imperativ; oder Linie / Zeile N pl) oder *reći* (dt. sagen, infinitiv) stehen; „kuca“ für *kuca* (dt. klopfen, 3. Person sg) oder *kuća* (dt. Haus, N sg oder G pl).

Die Zeichenkette „mladih“ kann für orthographisches *mladih* (dt. jung, G pl mn) oder *mladih* (dt. jung, komparativ G pl mn) stehen, „medu“ für *medu* (dt. Honig, DL sg) oder *među* (dt. Präp. zwischen, unter); „leda“ für *leda* (dt. Eis, G sg oder pl) oder *leđa* (dt. Rücken).

Und schließlich kann noch „veze“ für orthographisches *veze* (dt. Beziehung / Verbindung, G sg oder NAV pl) oder *veže* (dt. (ver-) binden, 3. Person sg) stehen, „zelje“ für *zelje* (dt. Kraut, NAV sg) oder *želje* (dt. Wunsch, G sg oder NAV pl) und „brzi“ für *brzi* (dt. schnell, m) oder *brži* (dt. schnell, m komparativ).

Wie man an diesen wenigen Beispielen sehen kann gibt es zahlreiche solche Homographen, wenn der Eingabetext in einer Schreibweise ohne Diakritika verfaßt ist. In Fällen wie bei *cesta*, *kuća* oder *želje* kommt zum Problem der „unechten Homographen“, wie ich es hier bezeichnet habe, noch das Problem der Akzentunterscheidung in den verschiedenen Kasus hinzu — es handelt sich hier also sozusagen um ein doppeltes Homographenproblem.

Was nach diesen Fällen noch an Homographen übrig bleibt, ist z.B. die Menge der Adverbien, die in der Regel die kurze Form des Adjektivs im Neutrum haben. Diese Unterscheidung dürfte für ein TTS-System aber keine Rolle spielen.

¹²Zugegeben: der Vokativ von „Straße“ ist eher unwahrscheinlich.

2.9 Phonetische und Phonologische Transkriptionen

Nachdem nun schon im gesamten Text phonetische Transkription verwendet wurde, soll hier noch einmal ein zusammenfassender Überblick über die phonetische bzw. phonologische Transkription des Kroatischen gegeben werden. Wie in diesem Abschnitt gezeigt werden soll, gibt es verschiedene Systeme, die in der Transkription verwendet werden.

2.9.1 IPA

Das *internationale phonetische Alphabet* stellt sozusagen den Standard in der phonetischen Transkription dar. Aus diesem Grund habe ich mir auch erlaubt, die phonetische Transkription mit IPA Symbolen ohne ausführliche Erklärung im Text zu verwenden.

Im Zusammenhang mit dem Lautinventar der kroatischen Sprache werden in Abschnitt 3.1 IPA Symbole verwendet. Dort wird auch der Abschnitt über die Transkription des Kroatischen aus dem IPA Handbuch angesprochen (IPA 1999; IPA Homepage). Die dort vorgestellte Transkription der kroatischen Laute ist in Abbildung 3.1 auf Seite 35 zusammengefaßt dargestellt.

Die verschiedenen Wortakzente des Kroatischen werden in IPA durch einen Zirkumflex $\hat{}$ über dem betreffenden Laut für einen fallenden Ton und durch einen Hatschek $\tilde{}$ für einen steigenden Ton gekennzeichnet. Die Länge wird durch das Symbol $:$ nach dem betreffenden Laut markiert. Der lange fallende Akzent in *med* (dt. Honig) z.B. wird dann als $/m\hat{e}:d/$ transkribiert, der kurze fallende in *mama* als $/m\hat{a}ma/$, der lange steigende in *zima* (dt. Winter) als $/zi:m\tilde{a}/$ und der kurze steigende Akzent in *noga* (dt. Bein / Fuß) als $/n\tilde{o}ga/$.

2.9.2 SAMPA

JURAJ BAKRAN und DAMIR HORGA haben eine SAMPA Transkription für das Kroatische vorgestellt (vgl. SAMPA Homepage). Die Grundidee von SAMPA war, ein System zur phonetischen / phonologischen Transkription zu bieten, das nur ASCII Zeichen verwendet und so als Standard im elektronischen Datenaustausch überall dort dienen kann, wo aus technischen Gründen nur ASCII zur Verfügung steht — oder damals nur zur Verfügung stand (in E-Mails oder Newsgroups z.B.). Auch heute wird SAMPA noch eingesetzt.

Die Symbole für die Vokale des Kroatischen entsprechen den IPA Symbolen: $i e a o u$. Das silbische r wird nicht gesondert transkribiert und wie auch das nicht-silbische als r geschrieben.

Die Symbole für die Plosive entsprechen ebenfalls den IPA Symbolen: $p b t d k g$. Ebenso wurden alle anderen Symbole der IPA übernommen, die in ASCII darstellbar sind: $f s z x m n l$ und j .

Die Affrikate werden wie folgt dargestellt: ts für $/ts/$, tS für $/tʃ/$ und dZ für $/dʒ/$. Die palatalisierten Affrikate werden als tS' für $/tʃ'/$ und dZ' für $/dʒ'/$ geschrieben. Die übrigen in IPA vom ASCII Zeichensatz abweichenden Symbole werden so dargestellt: S für $/ʃ/$, Z für

/ʒ/, J für /j/, L für /ɫ/ und v\ für v. Für letzteres wurde in X-SAMPA¹³ auch das Symbol P vorgeschlagen.

Bemerkenswert ist noch die Tatsache, daß orthographisches *j* mit *j* und *h* mit *x* transkribiert werden. Bei den Vokalsymbolen fällt auch auf, daß wie bei der IPA Transkription die Symbole *e* und *o* verwendet werden. In dieser Arbeit verwende ich aus Gründen der Vergleichbarkeit und schnelleren phonetischen Einordnung der Lautsymbole, insbesondere für den deutschen Leser, die Schreibweise *E* und *O*.

Die vier Wortakzente des Kroatischen werden in der SAMPA Transkription von BAKRAN und HORGGA mit den Symbolen <F> und <R> für fallenden und steigenden (engl. rising) Ton geschrieben. Die Länge wird durch einen Doppelpunkt : hinter dem betreffenden Laut gekennzeichnet und die betonte Silbe mit einem vorangestellten ". Die Beispiele für die vier Akzente und die Markierung der unbetonten Länge sehen wie folgt aus:

long fall	pas	"pa:s	<F>	belt
short fall	pas	"pas	<F>	dog
long rise	ruka	"ru:ka	<R>	hand
short rise	noga	"noga	<R>	leg
short unstr.	noga	"noga	<R>	leg
long unstr.	noga	"no:ga:	<R>	legs (gen.)

Aus: <http://www.phon.ucl.ac.uk/home/sampa/croatian.htm> (Bakran 1996b)

2.9.3 Slavistik und Kroatistik

Die nahezu phonologische Orthographie des Kroatischen macht es möglich eine spezielle Form der Transkription zu verwenden, wie sie in der Slavistik und in der Kroatistik üblich ist.

In linguistischen Arbeiten und Lehr- oder Wörterbüchern wird gewöhnlich diese slavistische Schreibweise zur phonetischen und phonologischen Transkription verwendet. Aufgrund der eindeutigen Zuordnung von Graphemen und Phonemen (siehe Abschnitt 2.2) werden alle einfachen, d.h. nicht zusammengesetzten Grapheme auch zur Transkription verwendet. Eine Ausnahme stellt dabei das Graphem *đ* für das Phonem /ɕ/ dar. Es wird traditionellerweise mit /ʒ/ transkribiert. Das silbische *r* wird durch /r̥/ (z.B. in (Babić et. al. 1996)), /r̄/ oder auch durch /r/ gekennzeichnet. In der engeren phonetischen Transkription wird *j* durch /j̣/ oder /j̥/ transkribiert, wo dies der phonetischen Realisierung entspricht und *h* durch /x/.

Die durch Digraphen repräsentierten Phoneme werden durch einfache Symbole dargestellt: /ʒ/ für *dž*, /lj/ für *lj* und /nj/ für *nj*. Gelegentlich ist die Schreibweise /g/ für *dž* zu finden, z.B. in den Publikationen der HAZU¹⁴.

Die Vokale werden mit den orthographischen Graphemen dargestellt. Der als *ije* geschriebene Diphthong wird durch /je/ dargestellt. Ebenso werden nicht-silbische Vokale mit einem daruntergestellten ˘ gekennzeichnet.

¹³Internet: <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>

¹⁴Hrvatska akademija znanosti i umjetnosti — dt. Kroatische Akademie der Wissenschaften und Künste

Zusätzliche Allophone, die keine Entsprechung im Alphabet haben werden ebenfalls durch in der Slavistik übliche Buchstaben wiedergegeben: [ś] für [ɕ], [ž] für [ʒ], [ɣ] für [ɣ], [F] oder [F] für [v] und [ʒ] für [ʒ] (die zweite Schreibweise stellt jeweils die entsprechenden IPA Symbole dar). Genau wie in IPA wird das Symbol [ŋ] verwendet.

Der Wortakzent wird mit den in Abbildung 2.9 dargestellten diakritischen Akzentzeichen gekennzeichnet. Die Länge in unbetonten Silben wird durch ein Makron $\bar{}$ über den betreffenden Vokalen gekennzeichnet.

	steigend	fallend
lang	dugoulazni [ː]	dugosilazni [ː]
kurz	kratkoulazni [˘]	kratkosilazni [˘]

Abbildung 2.9: Die kroatischen Akzente

Die verschiedenen Transkriptionen

Die verschiedenen verwendeten Transkriptionen sollen anhand einer Gegenüberstellung beispielhaft verdeutlicht werden (Abbildung 2.10). Der Text und die IPA Transkription sind, modifiziert, aus dem IPA Handbuch (IPA 1999) entnommen. In der SAMPA Transkription sind die Töne der Wortakzente mit ˈ für den steigenden und ˘ für den fallenden Ton vor dem betreffenden Laut markiert. Diese Symbole orientieren sich am Vorschlag der SAMPROSA¹⁵ Schreibweise. Die betonte Silbe muß aufgrund dieser Annotation auch nicht mit " markiert werden. Dadurch wird die hier dargestellte SAMPA Transkription insgesamt kürzer und übersichtlicher.

Orthographisch	Sjeverni ledeni vjetar i sunce su se prepirali o svojoj snazi. Stoga odluče da onome od njih pripadne pobjeda koji svuče čovjeka putnika. Vjetar započe snažno puhati, a budući da je čovjek čvrsto držao odjeću, navali on još jače.
Slavistisch	// sjěvĕrnĭ lĕdenĭ vjĕtar i sŭnce su se prĕpirali o svòjoj snázi // stòga òdlučĕ da ònome òd nĭch pripadne pòbjeda kòjĭ svŭče čòvjeka pŭtnika // vjĕtar zàpoče snážno pŭhati / a bŭdŭćĭ da je čòvjek čvrsto dŕžao òdjeću / nàvali òn još jàčĕ //
IPA	sjêvɛ:ɾni: lĕdeni: vjĕtar i sŭntɛ su se prĕpirali ɔ svòjɔj snâ:zi stòga òdluʧɛ: da ònome òd nix pripadne pòbjeda kòji svŭʧɛ ʧòvjeka pŭ:tnika vjĕtar zàpɔʧɛ snâzno pŭxati a bŭdu:ʧi da je ʧòvjek ʧvɔʃ:stɔ dŕʒao òdjetɛu nàvali ò:n jòʃ jâʧɛ:
SAMPA	sj`Ev`E:ɾni: l`EdEni: v\j`Etar is`u:ntsEsusE pr`Epirali Osv`OjOj sn`a:zi st`Oga `OdlutSE: da`OnOmE `OdJi:x pr`ipadnE p`ObjEda k`Oji: sv`u:tSE: tS`Ov\jEka p`u:tnika v\j`Etar z`apOtSE sn`a:ZnO p`u:xati ab`udu:tS`i dajEtS`OvjEk tSv`\`r:stO d`rZaO `OdjEtS`u n`avali `O:n j`OS j`atSE:

Abbildung 2.10: Die verschiedenen Transkriptionen

¹⁵Internet: <http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm>

Kapitel 3

Grammatik

Wieviel Information über Grammatik benötigt ein TTS-System? Dies ist die zentrale Frage, die für die Betrachtung der kroatischen Grammatik in dieser Arbeit bestimmend sein soll.

Auf Aspekte der kroatischen Sprache, welche für eine Text-to-Speech Sprachsynthese nicht von Bedeutung sind werde ich in diesem kurzen Überblick zur Grammatik nicht eingehen. Dazu gehört z.B. der typisch slavische Verbalaspekt, der perfektive (vollendete) und imperfektive (unvollendete) Verbformen unterscheidet. Dies stellt einerseits zwar ein nicht nur für linguisten interessantes Thema dar, spielt andererseits aber keine Rolle für TTS-Systeme. Solche Verbaspektpaare sind durch diverse Affixe oder unterschiedliche Stammformen unterscheidbar¹ und können mit den Ausspracheregeln behandelt werden, ohne daß ein TTS-System den Verbalaspekt berücksichtigen müsste.

Es sollen nun also nur diejenigen Aspekte der kroatischen Grammatik behandelt werden, die für die Entwicklung von Ausspracheregeln eines TTS-Systems von Bedeutung sind.

3.1 Lautinventar

Grundlegend für die Ausgabe eines TTS-System ist wohl unbestreitbar das Lautinventar der zu synthetisierenden Sprache. Denn es sind die einzelnen Laute einer Sprache aus denen sich die gesprochenen Worte zusammensetzen. Aus diesem Grund soll das kroatische Lautinventar auch als erstes vorgestellt werden.

Lautinventar in der „IPA Darstellung“

Abbildung 3.1 zeigt das Lautinventar des Kroatischen² nach der Darstellung von LANDAU, LONČARIĆ, HORGA und ŠKARIĆ im IPA Handbuch (IPA 1999). Die Symbole für die Phoneme (d.h. die bedeutungsunterscheidenden Laute der kroatischen Sprache) sind **fett** dargestellt

¹Z.B. *učiti* (ipf) / *naučiti* (pf) (dt. lernen) oder *skakati* (ipf) / *skočiti* (pf) (dt. springen)

²Die betreffenden Laute werden in der tabellarischen Darstellungen und dem Vokaldreieck mit IPA Symbolen dargestellt. Die Transkription kroatischer Texte wird in Abschnitt 2.9 beschrieben.

und die zusätzlichen Allophone (die von den Phonemen lautlich abweichenden Realisierungen) sind in normaler Schriftstärke verzeichnet. In Abbildung 3.2 sind die Vokale des Kroatischen dargestellt. Zusätzlich zu den Vokalphonemen ist auch das Schwa eingezeichnet, das kein Phonem darstellt. In der Abbildung zu sehen ist auch der Diphthong /ie/ der nach der Beschreibung im IPA Handbuch bei /i/ beginnt und der Position von /e/ endet. Nach dieser Darstellung verfügt das Kroatische also über ein Inventar von 35 Konsonanten und 7 Vokalen.

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Stop								ʔ
Plosiv	p b		t d				k g	
Affrikat			ts ɖ		tʃ ɟʃ	tɕ ɟɕ		
Nasal	m	ɱ		n		ɲ	ŋ	
Frikativ		f v	s z		ʃ ʒ		x ɣ	h
Sibilant						ç ʒ		
Trill				r				
Approximant		ʋ				j	w	
Lateralapproximant				l		ʎ		

Abbildung 3.1: Die kroatischen Konsonanten nach (IPA 1999)

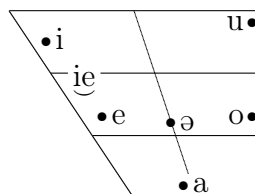


Abbildung 3.2: Die kroatischen Vokale nach (IPA 1999)

Lautinventar nach BARIĆ

Die Grammatik von BARIĆ et. al. (Barić et. al. 1995) dagegen nennt nur 33 Konsonanten und 8 Vokale. Dieser Unterschied bei den Konsonanten kommt dadurch zustande, daß die Allophone [ɱ] und [w] sowie der glotale Stop [ʔ] nicht, und statt dessen das silbische /r/ separat aufgeführt werden. Bei der Beschreibung der Vokale wird außerdem noch das nicht-silbische [i] als typische Realisierung von /j/ der Beschreibung hinzugefügt.

Lautinventar nach ŠKARIĆ

IVO ŠKARIĆ schreibt in seiner Phonetik des Kroatischen (Škarić 1991, seite 338):

„U hrvatskom standardnom jeziku ima 30 fonema i još dva koja su slobodna stilistička: /ě/ i /r/.“³

³„In der kroatischen Standardsprache gibt es 30 Phoneme und noch zwei, die frei stilistisch sind: /ě/ und

Zusätzlich zu den erwähnten 30 Phonemen beschreibt ŠKARIĆ noch 12 Allophone und kommt somit auf ein Lautinventar von 36 Konsonanten und 7 Vokalen. In die Liste der Konsonanten wird (im Vergleich zur IPA Beschreibung) zusätzlich der Halbvokal [j] eingereicht. Bei den Vokalen entspricht die Beschreibung dem des IPA Handbuchs.

In einem Artikel aus dem Jahre 2001 hingegen über die kroatische Rechtschreibung (und einen Vorschlag zur Verbesserung derselben) schreibt ŠKARIĆ:

„Na dobro utemeljenim razlozima stoji tvrdnja da u suvremenom [...] svehrvatskom jeziku [...] općem (tj. praktičkom standardu, a ne u propisanom "klassesichnome") postoji 29 (dvadesetdevet) fonema (a ne tradicionalno izbrojenih 30 niti 31 uz dodatak /ie/). Tih se 29 fonema dobije kad se od poznatih 30 oduzmu po jedna zvučna i bezvučna afrikata (tj. /dž/ i /ć/) te kad se popisu pridoda slogovni /r/.“⁴ (Škarić 2001b)

Lautinventar nach BROZOVIĆ

DALIBOR BROZOVIĆ beschreibt in seiner Phonologie des Kroatischen (Brozović 1991) 25 konsonantische Phoneme und 7 silbenbildende Phoneme (Selbstlaute) und listet dazu 597 Minimalpaare auf um somit sein Inventar an Phonemen zu begründen. Zu den Phonemen kommen noch 26 Allophone. Das gesamte Inventar ist in einer IPA-ähnlichen Tabelle in Abbildung 3.3 zusammengefaßt. Diese basiert auf einer Tabelle aus (Brozović 1991, Seite 404).

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Post-palatal	Velar	Glottal
Nichtkontinuierl.	p b		t d	ts dz	ʈ ɖ	ʧ ʤ	k^j g^j k^w g^w	ʔ
Kontinuierlich		f v	s z		ʃ ʒ	ç ʒ	x^j x^w ɣ	h ɦ
Nasal	m	ɱ	²n ²n^j		²ɲ	²ɳ	ŋ	
Vibrant				r				
Lateral			l^j l l^{wj}		ʎ			
Sonant	w	ʋ			j			
nichtsilb. Vokal	u̯				i̯			

Abbildung 3.3: Die kroatischen Konsonanten nach (Brozović 1991)

/r/“ (Übersetzung D. Duran)

⁴„Die Behauptung ruht auf gut begründeten Grundlagen, daß es in der gegenwärtigen [...] gesamt-kroatischen Gemeinsprache [...] (d.h. im praktischen Standard, und nicht im vorgeschriebenen "klassischen") 29 (neunundzwanzig) Phoneme gibt (und nicht die traditionell gezählten 30 oder 31 mit dem Zusatz /ie/). Diese 29 Phoneme erhält man, wenn man von den bekannten 30 je einen stimmhaften und einen stimmlosen Affrikat entfernt (d.h. /dž/ und /ć/) und wenn man der Beschreibung das silbische /r/ hinzufügt.“ (Auslassungen und Übersetzung DD) Dieser Artikel ist in der von Škarić vorgeschlagenen Orthographie geschrieben.

Wie in Abbildung 3.1 sind auch hier die Phoneme **fett** und die nicht-phonemischen Allophone in normaler Schriftstärke dargestellt. Zusätzlich ist in der Zeile „Nasal“ bei einigen Lauten eine 2 vorangestellt. Diese kennzeichnet laut BROZOVIĆ jeweils zwei Varianten desselben Lautes: eine kontinuierliche („kontinuiran izgovor“ — izgovor = Aussprache) und eine nicht-kontinuierliche („nekontinuiran izgovor“). Das ²n in der Spalte „Postpalatal“ stellt eine palatale Variante von /n/ dar, wie sie in Wörtern wie *opančar*, *kandža*, *branša* oder *inžinjer* vorkommt. Zu beachten ist, daß BROZOVIĆ diesen Laut eindeutig vom post-alveolaren Nasal /ɲ/ unterscheidet (Er nennt diese Laute „prednjotvrdonepčani (prednjopalatalni)“ = vor-palatal). In der Zeile „Lateral“ ist unter Dental l^w zu finden. Die beiden hochgestellten IPA Diakritika für „Labialisierung“ und „Velarisierung“ sollen bei diesem Laut eine labio-velarisierte („labiovelarizirani“) Variante kennzeichnen. Außerdem sind hier noch die nicht silbenbildenden Vokale $\underset{\cdot}{u}$ und $\underset{\cdot}{i}$ zu finden, von welchen letzteres in keiner der drei vorgenannten Beschreibungen separat aufgeführt wird.

Lautinventar nach BAKRAN

JURAJ BAKRAN unterscheidet in seiner akustischen Beschreibung des Kroatischen (Bakran 1996a) fünf Vokale und zusätzlich noch ein nichtphonemisches neutrales Schwa. Hinzu kommt ein glottaler Stop, der im Kroatischen keine phonologische Funktion erfüllt, aber zwischen zwei Vokalen auftreten kann. Zu den Konsonanten zählt BAKRAN:

- sechs Plosive (/p/ /t/ /k/ /b/ /d/ /g/)⁵
- vier stimmlose und zwei stimmhafte Frikative (/s/ /ʃ/ /f/ /x/ und /z/ /ʒ/)
- drei stimmlose und zwei stimmhafte Affrikate (/ts/ /tʃ/ /tɕ/ und /dʒ/ /dʒ/)
- drei Nasalphoneme (/m/ /n/ /ɲ/) und ein velares Allophon ([ŋ])
- zwei Laterale (/l/ /ʎ/)
- das Phonem /r/
- das Phonem /j/
- das Phonem /v/

Da es sich bei oben zitiertem Werk um eine akustische (spektrographische) Beschreibung des Kroatischen handelt, werden Allophone als verschiedene Realisierungen der oben genannten Phoneme insbesondere in unterschiedlichen Kontexten beschrieben. So werden Kontakte zwischen Plosiven und Nasalen als nasale Plosive beschrieben (Bakran 1996a, Seite 179), wie z.B. in Wörtern wie *dno*, *Etna*, *otmica* usw. Diese Plosiv-Nasal Folgen können auch explizit mit einer Plosion im oralen Raum artikuliert werden. Derartige Artikulationen klingen allerdings nicht sehr natürlich, wie BAKRAN schreibt:

„Ovakav se izgovor doživljava previše artikuliranim“⁶

⁵Bakran verwendet ausschließlich Standardorthographie und keine Transkription in dieser Arbeit (Bakran 1996a). Siehe Abschnitt 2.2 für die Graphem-Phonem-Beziehungen und Abschnitt 2.9 für die Transkription des Kroatischen.

⁶„So eine Aussprache wird als überartikuliert wahrgenommen“ (Übersetzung DD)

Das Phonem /r/ kann im Kroatischen einen Silbengipfel darstellen. Zwischen zwei Vokalen ist /r/ nie silbisch. In wortinitialer Position wird es von einem kurzen vokalischen Schwa-Element eingeleitet, in wortfinaler Position vor einer Pause und vor Konsonanten von einem kurzen Schwa-Element gefolgt. Trotz der beiden Funktionen als silbischer und nicht-silbischer Laut wird /r/ (im Gegensatz zu einigen anderen Autoren) von BAKRAN nur als ein Phonem beschrieben.

3.1.1 Das Lautinventar für diese Arbeit

Wie man aus diesen oben genannten Beispielen sieht ist es keine triviale Aufgabe, das Lautinventar einer Sprache zu beschreiben. Die Beschreibung wird immer von vielen Faktoren beeinflusst. Dazu gehört die Frage nach den Lauten einer Sprache, den zu unterscheidenden Allophonen, dem Unterschied zwischen Affrikaten und Lautfolgen und vor allem auch die Frage welche Laute als Phoneme zu bezeichnen sind. ZRINKA JELASKA gibt in ihrem Buch (Jelaska 2004) einen aktuellen Überblick über die Arbeiten auf dem Gebiet der phonologischen und phonetischen Beschreibung des Kroatischen und ihrer Schwierigkeiten und unterschiedlichen Lösungen.

Das Lautinventar des Kroatischen soll für die Zwecke dieser Arbeit nun wie folgt definiert sein:

- Das Vokalinventar umfaßt fünf Monophthonge /a/ /ɛ/ /i/ /ɔ/ /u/ und einen Diphthong /ie/ (Dieser Diphthong entspricht dem „langen Jat“ — vergleiche Abschnitt 1.3.1. Vergleiche auch Abschnitt 2.9 zur Verwendung der Transkriptionssymbole).

Zu diesen sechs Vokalphonemen kommen noch ein neutraler Zentralvokal [ə] sowie die nicht-silbischen Allophone [j] und [ɥ] hinzu. Jeder der Monophthonge kann sowohl lang als auch kurz auftreten. Trotz dieser Unterscheidung handelt es sich allerdings nicht um unterschiedliche Phoneme (wie im Deutschen etwa, wo man allgemein kurze und lange Vokale unterscheidet). Vielmehr kann die Länge selbst als Phonem aufgefaßt werden (da sie bedeutungsunterscheidend sein kann). (Siehe 3.2 Wortakzent).

Insgesamt entspricht das Vokalinventar also der Darstellung in Abbildung 3.2 — abgesehen von der Tatsache, daß ich für diese Arbeit ε anstelle von e und ɔ anstelle von o schreiben werde. Dies liegt einerseits daran, daß diese beiden kroatischen Vokale (in ihren Formantwerten) tiefer liegen als die Kardinalvokale e und o und diese Schreibweise damit zu Mißverständnissen führen könnte. Andererseits dürften die Symbole ε und ɔ vor allem dem deutschen Leser auch ohne Angabe der Formantwerte der entsprechenden Vokale die phonetische Natur verdeutlichen.

- Das silbische /r/ soll zusammen mit dem nicht-silbischen /r/ zu einem Phonem zusammengefaßt und unter die Konsonanten eingeordnet werden. Diese Zusammenfassung zu einem Phonem läßt sich dadurch begründen, daß /r/ und /r̥/ fast komplementär verteilt sind. Obwohl /r̥/ in vielen Arbeiten mit den Vokalen zusammengefaßt wird, bestehen im Gegensatz zu den Vokalen viele Einschränkungen für das Vorkommen des /r̥/. Es kann z.B. nur einmal in einem Wort auftreten, es kann nicht in Flexions- und Derivationsuffixen oder Pro- und Enklitika vorkommen und es steht meist in akzentuierten Silben. Jeder der Vokale kann vor und hinter jedem Konsonanten stehen, was für /r̥/ nicht gilt. Als Aussprache des silbischen /r/ wird die Variante ohne Schwa als Standard festgelegt.

- Die Menge der Verschlusslaute (Stops und Plosive) umfaßt die stimmlosen Phoneme /p/ /t/ /k/ und die stimmhaften /b/ /d/ /g/. Hinzu kommen die nasalen Plosive [dⁿ] und [tⁿ] und die lateralen [d^l] und [t^l] als kontextuelle Allophone.
- Die Menge der Frikative umfaßt die stimmlosen Laute /s/ /ʃ/ [ç] /f/ /x/ und die stimmhaften /z/ /ʒ/ [ʒ] [v] [ʎ].
- Die Affrikate sollen als eigene Laute (und nicht zusammengesetzt) behandelt werden. Ihre Menge umfaßt die stimmlosen Affrikate /ts/ /tʃ/ /tʃ/ sowie die stimmhaften [dʒ] /dʒ/ /dʒ/.
- Die Nasale sind /m/ /n/ /ɲ/ und [ŋ].
- Die Menge der Laterale umfaßt nur die beiden Phoneme /l/ und /ʎ/.
- Die Approximanten enthalten nur /v/ [w] und /j/.

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Stop/Plosiv	p b		t ⁿ d ⁿ t d t ^l d ^l				k g	
Frikativ		f v	s z		ʃ ʒ	ç z	x ɣ	
Affrikat			ts dʒ		tʃ dʒ	tʃ dʒ		
Nasal	m			n		ɲ	ŋ	
Trill				r				
Lateral				l		ʎ		
Approximant		u				j	w	

Abbildung 3.4: Die kroatischen Konsonanten für diese Arbeit

Das Vokalinventar macht somit 6 Phoneme und 2 zusätzliche Allophone aus und das Konsonanteninventar umfaßt 26 Phoneme und 11 zusätzliche Allophone (siehe Abbildung 3.4).

3.2 Wortakzent

Kroatisch wird zu den Sprachen mit polytonischem, freiem lexikalischem Wortakzent gezählt. Der Akzent ist sowohl lexikalisch als auch morphologisch bedingt. Innerhalb eines Wortes kann jede Silbe bis auf die letzte in mehrsilbigen Wörtern betont sein. Silben (betonte wie auch nach-akzentige unbetonte) können sowohl lang als auch kurz sein. Traditionellerweise werden zwei Tonverläufe unterschieden: ein fallender und ein steigender. Sie werden in der Literatur meist als Tonverläufe der Grundfrequenz der Stimme beschrieben. Diese beiden prosodischen Merkmale (Länge und Tonverlauf) werden in der klassischen Beschreibung des kroatischen Wortakzents zu einem Vierakzentsystem zusammengefaßt. Abbildung 2.9 auf Seite 33 zeigt diese vier lexikalischen Akzente mit den in der Literatur verwendeten Bezeichnungen und Symbolen (vgl. Abschnitt 2.9.3 zur kroatischen Transkription).

Der Akzent kann sich in der Flexion verschieben, d.h. auf eine andere Silbe wandern, oder ändern, d.h. von einem steigenden zu einem fallenden Akzent werden oder umgekehrt, oder

beides. Es gibt Proklitika, d.h. unbetonte Worte, die sich vor einem akzentuierten Wort mit diesem zusammen zu einem phonologischen Wort verbinden, die den (fallenden) Akzent auf sich ziehen können. Hier einige Beispiele:

- Veränderung des Akzentes:
mèdvjed (Nsg) — mèdvjēdā (Gpl)
sèla (Npl) — sèla (Gsg) — sèlā (Gpl) (dt. Dorf)
slīkē (Gsg) — slīkā (Gpl) (dt. Bild)
- Verschiebung des Akzentes:
grād (dt. Stadt) — ù grād (u - Präp. in)
vòdu (Asg dt. Wasser) — pò yodu (po - Präp. hier: nach)
Wie man sieht, kann sich der Akzent bei der Verschiebung auch verändern (das Makron, oder Längestrich, kennzeichnet unbetonte lange Silben — vgl. Abschnitt 2.4.1).

Die traditionellen Bezeichnungen für die verschiedenen Akzente (lang fallend, kurz steigend usw.) sollen in dieser Arbeit nur als Namen für die Akzente dienen. Über die tatsächliche phonetische Natur der Akzente kann ich im Rahmen dieser Arbeit keine Analysen anstellen. Daß die traditionellen Bezeichnungen (aus akustischer oder phonologischer Sicht) nicht unbedingt die exaktesten darstellen faßt z.B. ZVONIMIR JUNKOVIĆ in seinem Nachwort zu (Garde 1993) wie folgt zusammen:

„U našem jeziku jeka se očituje neposredno ispred naglašenog sloga: u riječima kao *žèn'a* ili *su:d'a* naglasak je na posljednjem samoglasniku, a jeka na slogu ispred naglaska. Kad se naglasak ostvaruje na početnom slogu, ili jedinom, jeke naravno, nema: *s'u:da*, *s'u:d*, *pr'aga*, *pr'ag*. Iz navedenih se primjera vidi da osobine koje se tradicionalno zovu silaznim naglascima stvarno obilježuju mjesto naglaska, a osobine što se nazivaju ulaznim naglascima zapravo su naglasne jeke, koje najavljuju naglasak; na taj način Garde objašnjava zašto se ulazni naglasci ne mogu očitovati u sklopu jednosložnih riječi ni na posljednjem slogu višesložnih.“⁷

Auch experimentalphonetische Untersuchungen beschreiben den steigenden Akzent als zweisilbigen Akzent, bei dem die Silbe unmittelbar nach der Akzentstelle einen hohen Ton trägt und mindestens ebenso hohe Intensität hat wie die vorhergehende. Die traditionellen Bezeichnungen und Schreibweisen werden aber auch in solchen Arbeiten beibehalten.

Wie bereits in Abschnitt 1.3.1 beschrieben wurde, basiert die kroatische Standardsprache auf dem neuštokavischen Dialekt und hat von diesem dessen Vierakzentsystem als Aussprachennorm übernommen. Die in der Literatur gegebenen Beschreibungen zum kroatischen Wortakzent beziehen sich daher oft auf eine „klassische Form“, womit unter Umständen nur eine östliche Variante der neuštokavischen Akzentuierung gemeint ist.

⁷„In unserer Sprache offenbart sich das Echo [= Betonungsecho, DD] unmittelbar vor der betonten Silbe: in Wörtern wie *žèn'a* oder *su:d'a* liegt die Betonung auf dem letzten Vokal, und das Echo auf der Silbe vor der Betonung. Wenn die Betonung auf der ersten, oder einzigen Silbe realisiert wird, gibt es natürlich auch kein Echo: *s'u:da*, *s'u:d*, *pr'aga*, *pr'ag*. Aus den angeführten Beispielen sieht man, daß die Charakteristika, welche traditionellerweise als fallende Akzente bezeichnet werden, in Wahrheit die Akzentstelle markieren, und diejenigen Charakteristika, die als steigende Akzente bezeichnet werden sind im Grunde genommen Betonungsechos, welche den Akzent ankündigen. Auf diese Weise erklärt Garde, warum steigende Akzente nicht in Verbindung mit einsilbigen Wörtern oder auf der letzten Silbe mehrsilbiger Wörter realisiert werden können.“ (Übersetzung DD)

So schreibt z.B. JOSIP MATEŠIĆ in der Einleitung seiner Arbeit zum „serbokroatischen“ Wortakzent (Matešić 1970):

„Die vorliegende Arbeit behandelt das Akzentsystem der modernen serbokroatischen Schriftsprache aus synchroner Sicht. Dieses System beruht auf den prosodischen Eigenschaften der neuštokavischen Mundarten, genauer jenes Typs dieser Mundarten, der in das Wörterbuch von Vuk St. Karadžić aufgenommen wurde...“

Und weiter:

„Außerdem gibt es die sog. „kategorialen“ Dubletten, die sich aus der Verschiedenheit des westlichen jekavischen (Kroatien) und des östlichen ekavischen (Serbien) Typs der Schriftsprache ergeben...“

Diese Arbeit von MATEŠIĆ ist also eine, die nicht nur einige Jahrzehnte alt ist, und somit möglicherweise keine genaue Beschreibung des zeitgenössischen kroatischen Akzentsystems mehr sein kann, sondern es ist auch noch eine Arbeit, die eine „serbokroatische Schriftsprache“ beschreibt und dabei eher serbische prosodische Merkmale berücksichtigt. Ob solche Arbeiten überhaupt geeignet sind als Grundlage für eine Beschreibung des kroatischen Akzentsystems und letzten Endes für eine Implementierung des modernen Kroatisch in einem TTS-System ist fraglich und müßte in jedem Fall zunächst genauer untersucht werden. Leider sind die meisten mir vorliegenden Arbeiten zu diesem Thema, was die Entstehungszeit und den Untersuchungsgegenstand angeht, mit dieser Arbeit von MATEŠIĆ vergleichbar und damit ähnlich zu bewerten.

Östliche und westliche Akzentuierung

Daß die östliche und die westliche Variante der neuštokavischen Akzentuierung sich unterscheiden und von Sprechern des Kroatischen auch als eher einem fremden bzw. dem eigenen Standard zugehörig wahrgenommen werden zeigen z.B. experimentelle Untersuchungen von IVO ŠKARIĆ (Škarić et. al. 1996; Škarić 2001a; Škarić & Lazić 2002).

„Zato je sasvim legitimno da se kodificira za te riječi i silazni naglasak na nepoččetnim slogovima ako suvremeni jezični osjećaj upravo tako traži.“⁸
(Škarić et. al. 1996, On Stressing Loanwords)

„The so called classical stress was judged as more typical for Serbian and the non-standard stress as more typical for Croatian.“
(Škarić 2001a, Distinctive Prosody)

Andere (oft zitierte) Arbeiten können als (alleinige) Grundlage für weitere Forschung und Entwicklung auf diesem Gebiet von vornherein ausgeschlossen werden — wenn man den

⁸„Daher ist es völlig legitim für diese Wörter [die in der Sprache neu sind und nur den gegenwärtigen nicht aber den historischen phonologischen Regeln unterliegen, (DD)] den fallenden Akzent auch auf nicht-initialen Silben zu kodifizieren, wenn das gegenwärtige Sprachgefühl dies genau so verlangt.“ (Übersetzung DD)

Anspruch erhebt die tatsächliche Akzentuierung des Kroatischen zu beschreiben. Hier spielt das in Abschnitt 1.1.1 gesagte eine Rolle: welchen Ansprüchen muß das zu entwerfende TTS-System genügen? Soll es die literarische Standardsprache nachbilden, oder den allgemein akzeptierten Standard der Alltagssprache — den *gefühlten* Standard?

Als Beispiele für aus dieser Sicht eher „ungeeignete“ Beschreibungen des kroatischen Akzentsystems seien hier folgende genannt:

- ILSE LEHISTE und PAVLE IVIĆ:

„This book deals mainly with the so-called Neoštokavian accentuation, that is, with the prosodic system of the Serbocroatian standard language and of somewhat more than half of the Serbocroatian dialects.“ (Lehiste & Ivić 1986)

„The main informant for the study was P.I., one of the authors. P.I. was born in 1924 in Belgrade. [...] His pronunciation is based on standard Serbocroatian as spoken by educated Vojvodinians, with certain regional characteristics. [...] The twelve additional informants speak the same dialect of modern standard Serbocroatian as the main informant, with only minor exceptions.“ (Lehiste & Ivić 1986, S. 35)

Bei dieser Beschreibung des Wortakzents ist auch nicht ganz klar, was man unter „Serbocroatian standard language“ verstehen sollte. Außerdem ist die Wahl der Sprecher auch nicht repräsentativ und aussagekräftig für das Kroatische.

- KARL-HEINZ POLLOK: „Der neuštokavische Akzent und die Struktur der Melodiegestalt der Rede“
Die frühe experimentalphonetische Untersuchung (Pollok 1964) des neuštokavischen Akzents kann als deskriptive Arbeit aufgrund der Sprecherauswahl für das Kroatische keine verlässlichen Informationen liefern. Die untersuchten Sprecher sprachen alle ekavisch, was wie bereits gesagt für das Kroatische nicht typisch ist. Außerdem kam nur einer der Sprecher aus Kroatien (was allerdings auch nicht automatisch bedeutet, dass die betreffende Person Kroatisch spricht).
- IRMGARD MAHNKEN: „Studien zur serbokroatischen Satzmelodie“
Diese Arbeit zur Beschreibung der „Satzmelodie“ beruht auf den Untersuchungen von POLLOK, bietet also auch keine Erkenntnisse für das Kroatische. (Mahnken 1964)
- PETER REHDER: „Beiträge zur Erforschung der serbokroatischen Prosodie“
Auch hier (Rehder 1968) handelt es sich auf Grund der Sprecherauswahl für die Untersuchung um eine wenig Aussagekräftige Quelle. Die untersuchten Sprecher waren alle wohnhaft in Belgrad, die meisten auch dort geboren.
- SVETLANA GODJEVAC
GODJEVAC hat ein ToBi⁹ System für „Serbo-Kroatisch“ vorgestellt (Godjevac 2001). Auch wenn die Autorin selbst nicht sagt in ihren Arbeiten das Kroatische zu beschreiben, bezieht sie es dennoch in ihren Begriff „Standard Serbo-Croatian“ mit ein.

⁹Internet: <http://www.ling.ohio-state.edu/~tobi>

„This work is a study of focus projection in Serbo-Croatian. Serbo-Croatian is a South-Slavic language spoken in the former Yugoslavia, now consisting of three separate countries: Croatia, Yugoslavia (Serbia and Montenegro), and Bosnia and Herzegovina.“ (Godjevac 2000)

„By the term Standard Serbo-Croatian, I here refer to the Eastern Standard Variant.“ (Godjevac 2001, S. 1)

Die Akzentuierung im Kroatischen stellt (nicht nur) für TTS-Systeme eine große Herausforderung dar. Die unterschiedlichen oder teilweise veralteten Beschreibungen in der Literatur erschweren die (schnelle) Entwicklung von prosodischen Regeln für Sprachsynthesysteme zusätzlich. Je nach Anforderung an das TTS-System müssen im Zweifelsfall experimentalphonetische Untersuchungen durchgeführt werden um zu einer zeitgemäßen und exakten Beschreibung des kroatischen Akzentsystems zu kommen. Eine Sprachsynthese, die in einer Weise spricht, die von der Zielgruppe nicht akzeptiert oder als ungewöhnlich empfunden wird, erscheint wenig sinnvoll.

Fazit

BAKRAN schreibt zum Wortakzent (Bakran 1996a):

„... osim toga realizacija normativnih četiriju različitih akcenata najčešće je toliko neutralizirana da je razlike među pojedinim tipovima akcenta moguće ilustrirati samo u hiperartikuliranom – izvještačenom izgovoru.“¹⁰

und:

„Istinit opis svih aspekata akustičke realizacije akcenta riječi u hrvatskom zahtjeva obimna istraživanja koja u hrvatskom još nisu provedena.“¹¹

Somit kann ich für diese Arbeit auf keine aktuelle und vollständige akustische Beschreibung des Wortakzentes im Kroatischen zurückgreifen und die Ausarbeitung einer solchen Darstellung würde den Rahmen dieser Arbeit bei weitem sprengen.

¹⁰„... außerdem ist die Realisierung der vier unterschiedlichen normativen Akzente meistens derart neutralisiert, daß es nur möglich ist die Unterschiede zwischen den einzelnen Akzenttypen in überartikulierter, gekünstelter Aussprache zu illustrieren.“ (Übersetzung DD)

¹¹„Eine wahrheitsgetreue Beschreibung aller Aspekte der akustischen Realisierung der Wortakzente im Kroatischen erfordert umfangreiche Untersuchungen, welche im Kroatischen noch nicht durchgeführt wurden.“ (Übersetzung DD)

Kapitel 4

TTS–Ausspracheregeln

In diesem Kapitel sollen Ausspracheregeln für das Kroatische auf Basis der in den Kapiteln „Die geschriebene kroatische Sprache“ und „Grammatik“ vorgestellten Gegebenheiten der kroatischen Sprache und Orthographie entwickelt werden. Dazu wird der Formalismus des Festival Sprachsynthese Systems verwendet (vgl. Black et. al. 1999; Black & Lenzo 2003).

4.1 Festival und Festvox

„... making it possible for anyone to build a new voice.“
<http://www.festvox.org>

Festival ist ein frei verfügbares multilinguales System zur Entwicklung von Text-to-Speech Sprachsynthesystemen, das an der Universität von Edinburgh entwickelt wurde. Das System steht unter einer der „MIT-Lizenz“ oder auch „X11-Lizenz“ ähnlichen Lizenz¹, welche sowohl kommerzielle wie auch nicht-kommerzielle Verwendung gestattet. Festival bietet ein volles Text-to-Speech System für (britisches und amerikanisches) Englisch und bietet auch eine Diphoninventar für (kastilisches) Spanisch. Darüber hinaus bietet Festival auch ein Grundgerüst zur Entwicklung von neuen Systemen für beliebige Sprachen.

Festival ist in C++ geschrieben und verwendet einen Scheme Kommandointerpreter, der auf dem SIOD² Scheme-Standard basiert (Black & Lenzo 2003, Kap. 24). Durch diesen integrierten Scheme Interpreter soll es möglich sein neue Sprache in Festival einzubinden ohne den zugrundeliegenden C++ Programmcode zu modifizieren. Der Quellcode ist offen. Somit ist Festival gut geeignet, um Text-to-Speech Systeme für beliebige Sprachen und Anwendungsgebiete zu entwickeln.

Festvox bietet eine Reihe von Tools und Anleitungen zur Entwicklung und Einbindung neuer Stimmen in Festival — was in Festival auch gleich zu setzen ist mit der Einbindung neuer

¹Der Copyright Hinweis von Festival ist unter <http://www.cstr.ed.ac.uk/projects/festival/freecopyright.html> zu finden und die allgemeine Vorlage der MIT-Lizenz kann z.B. unter <http://www.opensource.org/licenses/mit-license.php> eingesehen werden.

²siehe <http://people.delphiforums.com/gjc/siod.html>

Sprachen. Das Festvox Projekt soll die Entwicklung neuer synthetischer Stimmen systematisieren und vereinfachen.

Die freie Verfügbarkeit, die Modifizierbarkeit, sowie die Tatsache, daß Festival am INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG der Universität Stuttgart verwendet wird, hat mich zu der Entscheidung gebracht, die entwickelten Ausspracheregeln für das Kroatische im Festival Formalismus zu verfassen.

4.2 Lautinventar

In Abschnitt 3.1.1 wurde das Lautinventar des Kroatischen bereits dargestellt. Hier soll nun eine formale Darstellung in der von Festival verwendeten Notation erfolgen.

Die Definition des Lautinventars ist in Festival grundlegend für die Entwicklung einer neuen Sprache (bzw. Stimme). Definiert wird das Lautinventar im Scheme Format und als „*phoneset*“ bezeichnet.

Die Grundstruktur der Definition eines solchen *phonesets* sieht wie folgt aus:

```
(defPhoneSet
  NAME
  FEATUREDEFS
  PHONEDEFS)
```

NAME steht für eine eindeutige Bezeichnung für das betreffende *phoneset*, *FEATUREDEFS* ist eine Liste von Merkmalen und den dazugehörigen möglichen Werten und *PHONEDEFS* stellt die eigentliche Liste der einzelnen Phone (Laute) dar.

Die in einem *phoneset* definierte Liste der einzelnen Laute legt in Festival nicht nur das Lautinventar für die Synthese fest, sondern es liefert dem System auch das Zeicheninventar für die (interne) Transkription der eingegebenen orthographischen Texte. Wie in Abschnitt 2.9 bereits beschrieben, stellt die IPA Schreibweise in der Linguistik den Standard zur phonologischen und phonetischen Transkription dar. Die IPA Transkriptionen sind in Unicode mit den im Bereich 0250–036F definierten Zeichen³ darstellbar. Da es in Festival leider keine eingebaute Unterstützung für Unicode gibt muß für die Transkription der einzelnen Laute ein 1-Byte Zeichenkodierungssystem verwendet werden (oder Festival in entsprechender Weise angepasst werden). In so einem 1-Byte Zeichenkodierungssystem bietet sich SAMPA zur Darstellung der Laute an (vgl. Abschnitt 2.9.2).

4.2.1 Die Vokale

Für die 14 Vokallaute des Kroatischen (wie in Abschnitt 3.1.1 beschrieben) werden Einträge der folgenden Form in die Definition des *phonesets* aufgenommen:

³ 0250–20AF: IPA Extensions (<http://www.unicode.org/charts/PDF/U0250.pdf>), 02B0–02FF: Spacing Modifier letters (<http://www.unicode.org/charts/PDF/U02B0.pdf>) und 0300–036F: Combining Diacritical Marks (<http://www.unicode.org/charts/PDF/U0300.pdf>)

```

(a  + m 3 2 - 0 0 + + -)
(i  + m 1 1 - 0 0 + + -)
(u  + m 1 3 + 0 0 + + -)
...
(a: + m 3 2 - 0 0 + + +)
...
(iE + d 2 1 + 0 0 + + +)
(@  + s 2 2 - 0 0 + - 0)
(i_~ + m 1 1 - 0 0 + - 0)
(u_~ + m 1 3 + 0 0 + - 0)

```

Das erste Zeichen in der Liste steht für den zu definierenden Laut, im oberen Beispiel steht also **a** für den Vokal /a/, **a:** für das lange /a:/, **iE** für den Diphthong /iε/ usw. Die folgenden Elemente der Liste stellen die Werte der in **FEATUREDEFS** definierten Merkmale dar. In den oben gezeigten Beispielen der Vokale /i/ /a/ und /u/ steht + für das Merkmal „Vokal“ und hat den Wert Plus, wohingegen die Konsonanten an dieser Stelle ein Minus (-) erhalten werden. Das **m** bezeichnet einen Vokal als Monophthong (als einfachen Vokal also) und das **d** als Diphthong. Das nächste Merkmal in der Liste ist die Vokalhöhe (1 steht hier für hohe bzw. geschlossene Vokale, 3 für tiefe bzw. offene), gefolgt von der vertikalen Position (1 steht hier für vordere Vokale, 2 für zentrale und 3 für hintere Vokale). Das nächste Merkmal gibt an, ob der Laut gerundet (+) oder ungerundet (-) ist. Die Konsonanten werden an dieser Stelle eine 0 als Wert erhalten, da für sie das Merkmal Lippenrundung keine Rolle spielt. Die nächsten beiden Merkmale in der Liste bezeichnen die Artikulationsstelle und die Artikulationsart von Konsonanten, weshalb bei den Vokalen hier jeweils 0 steht. Die letzten drei Merkmale stehen für Stimmhaftigkeit, Silbizität (da im Kroatischen nicht nur Vokale sondern auch das /r/ einen Silbengipfel bilden kann) und die Länge (nur wichtig für silbische Laute, die anderen erhalten den Wert 0 an dieser Stelle).

Für die Transkription der nicht-silbischen Vokale [i] und [u] habe ich mich für die X-SAMPA⁴ Notation **i_~** und **u_~** entschieden.

Die Definition der Merkmale ist nicht durch das Festival System vorgegeben, sondern kann bzw. muß von den Autoren einer neuen Stimme/Sprache definiert werden. Da in Festival auch mehrere *phonesets* definiert werden können, dienen die in **FEATUREDEFS** definierten Merkmale auch zur Konvertierung von Lauten aus einem *phoneset* in Laute eines anderen.

4.2.2 Die Konsonanten

Die Liste der Konsonanten enthält 40 Elemente. Diese werden nach demselben Schema definiert, wie die Vokale auch. Einige Beispiele für die Konsonanten sind:

```

(p  - 0 - - 0 s bl - - 0)
(b  - 0 - - 0 s bl + - 0)
...
(t_n - 0 - - 0 s nr - - 0)
(t_l - 0 - - 0 s lr - - 0)

```

⁴siehe <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>

```

...
(S_j - 0 - - 0 f pl - - 0)
(Z_j - 0 - - 0 f pl + - 0)
...
(n_ = - 0 - - 0 n av + + -)

```

Die Symbole t_n und t_l stehen für die jeweiligen kontextuellen Allophone von /t/ — dem nasalen Plosiv [tⁿ] und dem lateralen [t^l]. An den nächsten beiden Beispielen sieht man die Definition der palatalisierten Allophone [ç] und [ž]. Die übliche SAMPA Schreibweise für diese Laute wäre S' und Z'. Da das Zeichen ' in Scheme aber eine spezielle Bedeutung hat verwende ich statt dessen die alternative Notation nach X-SAMPA mit _j. Das letzte der oben gezeigten Beispiele zeigt die Definition eines silbischen Konsonanten, hier die des silbischen /ŋ/.

Der vollständige Scheme-Code für das von mir entwickelte Lautinventar für das Kroatische ist in Anhang A zu finden.

4.3 Lexikon

Festival verwendet zur Bestimmung der Aussprache ein Aussprachelexikon. Dieses besteht aus drei Teilen. Der erste Teil ist in der Regel, d.h. bei Festival in Sprachen wie Englisch oder Deutsch, der wichtigste und besteht aus einem großen, kompilierten Aussprachelexikon. Ein solches Lexikon stand mir nicht zur Verfügung. Für Kroatisch wird ein großes Aussprachelexikon möglicherweise aber überhaupt nicht benötigt. Ob die Bestimmung des Wortakzentes und die Aussprache einiger (teilweise morphologisch bedingter) Ausnahmefälle am besten mit einem Aussprachelexikon gelöst werden kann, kann ich im Rahmen dieser Arbeit leider nicht klären.

Neben dem Aussprachelexikon kommen in Festival noch zwei weitere Module zum Einsatz: eine handgeschriebene Liste von zusätzlichen Lexikoneinträgen sowie einer Methode zur Bestimmung der Aussprache von Wörter, die weder im großen Lexikon noch in der zusätzlichen Liste enthalten sind. Dies wird durch Transkriptionsregeln erreicht.

Für diese Arbeit habe ich eine Reihe von Lexikoneinträgen von Hand erstellt. Hier können vor allem Ausnahmefälle behandelt werden. Dies können z.B. häufig verwendete Fremdwörter sein, die in ihrer Schreibweise nicht kroatisiert sind. Auch Wörter mit selten vorkommenden Lautfolgen können hier aufgelistet werden. Hier einige Beispiele für (handgeschriebene) Lexikoneinträge in Festival:

```

(lex.add.entry '("croatia" nil (((k r 0) 0) ((a) 2) ((ts i) 0) ((a) 0)) ))
...
(lex.add.entry '("adagio" nil (((a) 0)((d a) 2)((dZ 0) 0)) ))
(lex.add.entry '("allegro" nil (((a) 0) ((l E:) 1) ((g r 0) 0)) ))
...
(lex.add.entry '("reljef" nil (((r E l) 2) ((j E f) 0)) ))
...
(lex.add.entry '("copyright" nil (((k 0) 2) ((p i) 0) ((r a i^ t) 0)) ))

```

...

```
(lex.add.entry '("nautika" nil (((n a u_) 2) ((t i) 0) ((k a) 0)))
```

Die Lexikoneinträge enthalten, wie man oben sehen kann, das Wort in orthographischer Form, gefolgt von der Wortart (dem POS-tag⁵) und der Transkription des Wortes einschließlich der Silbenstruktur und des Wortakzentes. Der Wortakzent wird für Englisch durch 0 für unbetonte und 1 für betonte Silben gekennzeichnet. Um die kroatischen Wortakzente zu markieren habe ich eine 1 für den fallenden und eine 2 für den steigenden Akzent gewählt. Dies spielt für die Arbeit im Grunde aber keine Rolle, da keine weitere Behandlung des Wortakzentes erfolgen wird.

4.4 Transkriptionsregeln (Letter to sound rules)

In diesem Abschnitt sollen kurz die Notation und Funktionsweise von Transkriptionsregeln in Festival dargestellt werden. Im Anschluß daran erfolgt eine allgemeine Darstellung von möglichen Transkriptionen für jeden einzelnen Buchstaben gefolgt von einigen konkreten Ausspracheregeln im Festival-Formalismus.

Festival bietet die Möglichkeit Transkriptionsregeln automatisch zu erstellen. Zu diesem Zweck muß ein (sehr großes) elektronisches Aussprachelexikon der zu bearbeitenden Sprache vorliegen. Ein derartiges Aussprachelexikon des Kroatischen stand mir für diese Arbeit nicht zur Verfügung. Neben der automatischen Erstellung der Regeln ist es auch möglich, diese von Hand zu schreiben und vor allem für eine Sprache mit einer Orthographie, wie das Kroatische sie verwendet, bietet sich die Erstellung von Hand an. Dies kann unter Umständen je nach Sprache sogar einfacher sein, als eine automatische Erstellung.

„For well defined languages like Spanish and Croatian writing rules by hand can be more simple than training.“ (Black & Lenzo 2003, Kap. 7.4.)

Die Transkriptionsregeln (letter-to-sound rules, kurz: lts) in Festival werden nach folgender Grundform notiert: (LK [Z] RK = NZ). Dabei steht LK für den linken Kontext, RK für den rechten Kontext, Z für das zu transkribierende Zeichen und NZ für das Zeichen, durch welches Z ersetzt werden soll. Die Regeln werden jeweils von links nach rechts der Reihe nach angewendet und transkribieren nur die einzelnen (isolierten) Wörter. Regeln, die Effekte in einem Satz über Wortgrenzen hinweg beschreiben, werden als postlexikalische Regeln bezeichnet. Auch ist zu beachten, daß die Ergebnisse der Regelanwendungen, also die NZ Elemente, nicht wieder als Eingabe für die folgenden lts-Regeln zur Verfügung stehen.

Sowohl der linke wie auch der rechte Kontext kann Variablen enthalten, die für eine Menge von Lauten stehen. Durch einen Befehl wie (V i E a O u) kann z.B. eine Menge definiert werden, welche die angegebenen fünf Vokale enthält und durch V bezeichnet wird. Zusätzlich gibt es noch das Symbol # zur Kennzeichnung von Wortgrenzen. Beispielsweise ließe sich nun mit der Regel (# [y] V = j) festlegen, daß ein y am Wortanfang vor einem Vokal als [j] ausgesprochen werden soll.

⁵Da Festival keinen standardmäßigen part-of-speech Tagger für das Kroatische beinhaltet und mir für diese Arbeit auch keiner zur Verfügung stand, steht an der Stelle der Wortart in meinem Lexikon bei allen Einträgen nur nil.

4.4.1 Die möglichen Transkriptionen

Die lts-Regeln in Festival konvertieren jedes orthographische Zeichen (Buchstaben) in ein phonetisches, d.h. sie transkribieren einen Text. Um die von mir erstellten Regeln besser überblicken zu können möchte ich in diesem Abschnitt kurz die allgemeinen Möglichkeiten zur Transkription zu jedem Buchstaben aufzählen. Da Festival keine Unicode Unterstützung bietet, werde ich der Einfachheit halber auch für die Eingabe nur einen 1-Byte Zeichensatz annehmen, und damit nur in Unicode definierte Zeichen wie z.B. *â* oder *à* nicht berücksichtigen.

Die Zeichenfolgen *dž*, *nj* und *lj* stellen im Kroatischen zwar eigene Grapheme dar, werden aber dennoch mit zwei einzelnen Zeichen (Buchstaben) geschrieben. In der folgenden Beschreibung sollen diese und weitere relevante Buchstabenfolgen jeweils unter ihrem ersten Element eingeordnet werden. D.h. die Regel, daß *nj* mit J transkribiert werden kann, wird in zwei Regeln aufgeteilt, nämlich die, daß *n* zu J werden und *j* gelöscht werden kann — letzteres wird durch $j \mapsto \emptyset$ dargestellt.

- $a \mapsto [\mathbf{a}]$
 $e \mapsto [\mathbf{E}]$
 $i \mapsto [\mathbf{i}], [\mathbf{iE}]$
 $o \mapsto [\mathbf{O}]$
 $u \mapsto [\mathbf{u}], [\mathbf{u}^{\wedge}]$

Die Vokale können alle sowohl kurz als auch lang sein. Da die Länge im Kroatischen aber als eigenständiges Phonem angesehen werden kann (vgl. Abschnitt 3.1.1), wähle ich für die standardmäßige Transkription der Vokalbuchstaben die kurzen Vokale.

Für das *i* gibt es zusätzlich noch die Möglichkeit es als Diphthong zu transkribieren. Dies ist genau dann der Fall, wenn die Zeichenkette *ije* für den Diphthong /ie/ steht (vgl. Abschnitt 2.4).

Für das *u* gibt es die Möglichkeit es als nicht-silbisches [u] zu transkribieren, wenn es in Wörtern wie *auto* oder *euro*. Hierbei handelt es sich aber nur um einige wenige ausnahmen, da zwischen zwei (in der Schrift) aufeinanderfolgenden Vokalen im Kroatischen sonst immer eine Silbengrenze liegt.

- $\hat{i} \mapsto [\mathbf{i:}]$
 $\hat{a} \mapsto [\mathbf{a:}]$
 $\hat{o} \mapsto [\mathbf{O:}]$

Wie in Abschnitt 2.4.1 gezeigt, werden auch diese Buchstaben in kroatischen Texten verwendet. Als Transkription kann hier jeweils ein langer Vokal angenommen werden. Die Zeichen *ê* *û* oder *ř* kommen in 1-Byte Zeichensätzen wie ISO 8859-2 oder CP1250 nicht vor und werden hier daher nicht berücksichtigt.

- $\acute{i} \mapsto [\mathbf{i:}]$
 $\acute{e} \mapsto [\mathbf{E:}]$
 $\acute{a} \mapsto [\mathbf{a:}]$
 $\acute{o} \mapsto [\mathbf{O:}]$
 $\acute{u} \mapsto [\mathbf{u:}]$

Diese Zeichen kommen hauptsächlich in fremdsprachigen Eigennamen wie z.B. *Rodríguez*, *José*, *László*, *Bulcsú* usw. vor. Sie könnten aber auch als kroatische Akzentzeichen verwendet werden, wie sie in Abschnitt 2.9.3 vorgestellt wurden (und einen langen

steigenden Akzent kennzeichnen). In der einen oder anderen Verwendung sollen die Buchstaben in jedem Fall als lange Vokale interpretiert werden.

- $b \mapsto [b], [p], \emptyset$
- $p \mapsto [p], [b], \emptyset$
- $d \mapsto [d], [t], [dz], [dZ], [ts], [tS], \emptyset$
- $t \mapsto [t], [d], [ts], \emptyset$
- $g \mapsto [g], [k]$
- $k \mapsto [k], [g]$
- $z \mapsto [z], [s], \emptyset$
- $s \mapsto [s], [z], \emptyset$
- $\acute{d} \mapsto [dZ_j], [tS_j]$
- $\acute{c} \mapsto [tS_j], [dZ_j]$
- $\check{c} \mapsto [tS], [dZ]$

Im Kroatischen gibt es die Effekte der Angleichung an die Stimmhaftigkeit und an die Artikulationsstelle. Dabei gilt, daß vor stimmhaften Lauten nur stimmhafte Allophone der Konsonanten stehen und entsprechend bei stimmlosen nur stimmlose. Die Laute /m/, /n/, /ɲ/, /l/, /v/, /j/ und /r/ verhalten sich dabei in Bezug auf die Stimmhaftigkeit neutral. Die Angleichung ist in der Orthographie in der Regel wiedergegeben. Es gibt allerdings einige Ausnahmen wozu einige Komposita und Fremdwörter zählen. Auch wird die Angleichung an einigen Morphemgrenzen nicht schriftlich wiedergegeben. Außerdem werden zwei gleiche aufeinanderfolgende Konsonanten niemals doppelt ausgesprochen, sondern zu einem Konsonanten zusammengefaßt. Insbesondere wird in der gesprochenen Sprache die Phonemfolge /ts/ gewöhnlich zur Affrikate /ts/ zusammengefaßt und /t/ vor den Affrikaten /ts/, /tʃ/, /tʃ/ oder zwischen /s/, /z/, /ʃ/, /ʒ/ und Konsonanten (außer /r/ und /v/) gelöscht.

Das Wort *podčiniti* (mit dem Präfix *pod-*, dt. unterwerfen, biegen) wird z.B. als [pɔtʃinɪti] ausgesprochen, *gradski* (mit dem Suffix *-ski*, dt. städtisch) als [gratski], und *postdiplomski* wird zu [pɔzdiploḡmski]⁶, *azbestni* wird zu [azbɛsni] usw.

Für diese und alle anderen Fälle müssen Ausspracheregeln erstellt werden. In Festival heißt das lts-Regeln für Effekte innerhalb der Wortgrenzen und postlexikalische Regeln für Effekte über solche hinweg.

- $f \mapsto [f], [v]$
- $h \mapsto [x], [G]$
- $c \mapsto [ts], [dz]$

Die Phoneme /f/, /x/ und /ts/ werden durch die Grapheme *f*, *h* und *c* repräsentiert. Auch diese Laute sind von der Angleichung der Stimmhaftigkeit betroffen. Im Gegensatz zu den oben aufgezählten Beispielen gibt es für die stimmhaften Allophone [v], [ɣ] und [dz] der Phoneme /f/, /x/ und /ts/ keine entsprechenden Buchstaben. Dies liegt vielleicht daran, daß der Effekt in Zusammenhang mit diesen Lauten nur über Wortgrenzen hinweg auftritt⁷ (und somit in Festival mit postlexikalischen Regeln behandelt werden muß).

⁶[std] > [sθd] > [zd]

⁷Das einzige „kroatische“ Wort, in welchem *f* vor einem stimmhaften Konsonanten steht, scheint meinem Testkorpus zufolge *Afganistan* zu sein. Mit *h* vor einem stimmhaften Konsonanten gibt es nur *Vrhbosna* (alter Ort und Erzbistum von Sarajevo), und das ist ein Kompositum aus *vrh* (dt. Gipfel/Spitze) und *Bosna*. Als Wörter mit einem *c* vor einem stimmhaften Konsonanten kommen nur *cd* und seine flektierten oder abgeleiteten Formen wie z.B. *cdovima* (dt. „CD“ pl I) vor — und das ist kein gewöhnliches Wort sondern ein Akronym (vgl. Abschnitt 2.3.2).

- š ↦ [S], [S_j], [Z], ∅
 ž ↦ [Z], [Z_j], [S], ∅
 ś ↦ [S_j]
 ź ↦ [Z_j]

Die Konsonanten /ʃ/ und /ʒ/ sind auch von der Angleichung der Stimmhaftigkeit betroffen. Hinzu kommt die Angleichung der Artikulationsstelle. Steht /ʃ/ vor einem /tʃ/ wird es auch palatalisiert und als [ç] ausgesprochen und /ʒ/ vor /dʒ/ wird zu [z]. Das Wort *lišće* (dt. Laub) wird [liɕtʃe] ausgesprochen und *grožđe* (dt. Trauben) [grɔzdʒe]. Für die Allophone [ç] und [z] gibt es keine speziellen Buchstaben im Kroatischen, in spezieller linguistischer Literatur werden aber *ś* und *ź* verwendet. Diese Zeichen sind auch in ISO 8859-2 enthalten und werden hier daher auch behandelt.

- n ↦ [n], [m], [N], [J], [n_=]
 m ↦ [m]
 l ↦ [l], [L], [l_=]

Das *n* wird vor /p/ und /b/ an die Artikulationsstelle angeglichen und zu /m/, wie in *jedanput* (dt. einmal) > [jedamput]. Vor /k/ und /g/ wird es zu [ŋ].

Die Zeichen *n* und *l* stellen auch jeweils den ersten Teil der Grapheme *nj* und *lj* dar und können daher in der Transkription zu /ɲ/ und /ʎ/ werden. Dies kann natürlich nur vor *j* geschehen, das in diesem Fall lediglich ein orthographisches Zeichen darstellt und gelöscht wird. Da es aber auch Buchstabenfolgen *nj* gibt, die getrennte Laute darstellen, müssen die Regeln diese Fälle entsprechend berücksichtigen (vgl. Abschnitt 2.4). In einigen seltenen Fällen können /n/ und /l/ auch silbisch sein.

- j ↦ [j], [i_~], ∅
 v ↦ [P], [w]

Das *j* wird generell als nicht-silbisches [j] ausgesprochen, lediglich am Wortanfang als [j]. Steht *v* vor *u*, so wird die Artikulationsstelle verschoben und der Laut als [w] ausgesprochen, sonst als [v].

- r ↦ [r], [r_=], [r_=:]

Das *r* bedarf besonderer Aufmerksamkeit. Es kann sowohl für das nicht-silbische wie für das silbische /r/ stehen, wobei letzteres auch noch lang oder kurz sein kann. Einige Grundregeln lauten, daß *r* am Wortanfang vor einem Konsonanten immer silbisch ist (z.B. *rzati*, dt. wiehern), oder zwischen zwei Konsonanten⁸ (z.B. *vrt*, dt. Garten) oder bei einigen Fremdwörtern am Wortende nach einem Konsonanten (z.B. *žanr*, dt. Genre).

Es gibt aber auch morphologische Regeln, die im Rahmen einfacher Transkriptionsregeln nicht erstellt werden können. So gibt es z.B. die Regel, daß *r* zwischen einem Konsonanten und *o*, welches aus einer morphologischen Alternation aus *l* entstanden ist, silbisch ist. Außerdem kann das *r* nach einem Präfix, das mit einem Vokal endet, silbisch sein, oder auch in Komposita wobei der erste Teil auf einen Vokal endet (vgl. Brozović 1991; Barić et. al. 1995).

⁸Dies gilt, wenn der Konsonant vor *r* nicht einer der folgenden ist: *j*, *r*, *l*, *lj*, *n*, *nj*, *ć*, *dž* oder *d*.

- $\ddot{a} \mapsto [E]$
- $\ddot{o} \mapsto [E]$
- $\ddot{u} \mapsto [i]$
- $q \mapsto [k], [k\ u], [k\ P]$
- $x \mapsto [k\ s]$
- $y \mapsto [i], [j]$
- $w \mapsto [P]$

Dies sind einige Beispiele für nicht-kroatische Buchstaben und den typischerweise dazugehörenden Lauten bzw. Lautfolgen.

4.4.2 Die letter-to-sound rules in Festival

Aus der oben dargestellten Zusammenfassung habe ich eine Reihe von Transkriptionsregeln für Festival erstellt. Die Regeln werden in Festival in einer Liste von einzelnen Regeln gespeichert und dabei so erstellt, daß die seltensten und spezifischsten Regeln als erstes notiert werden und die Standardregeln als letztes. Die sehr spezielle Regel, nach der das Präfix *cro* nach dem Lateinischen mit /k/ gesprochen wird, sollte also vor der allgemeinen Regel definiert werden, nach der *c* als /ts/ gelesen wird. Die Regel, daß *d* vor *c* gelöscht wird, muß vor der allgemeineren Regel kommen, daß *d* vor einem stimmlosen Konsonanten auch stimmlos wird. Im folgenden Beispiel sind diese Regeln dargestellt (NVOX bezeichnet eine Menge mit den stimmlosen Konsonanten):

```
( # [ c ] r o = k )
( [ d c ] = ts )
...
( [ d ] NVOX = t )
...
( [ c ] = ts )
( [ d ] = d )
```

Die vollständige Definition der lts-Regeln ist in Anhang C zu finden.

4.5 Test und Auswertung

Mit den oben dargestellten Regeln sind drei Scheme-Skripte für Festival entstanden: ein Lautinventar (phoneset, siehe Anhang A), eine Reihe von handgeschriebenen Lexikoneinträgen (lexicon addenda, siehe Anhang B) sowie allgemeine Transkriptionsregeln (letter-to-sound rules, siehe Anhang C).

Da diese drei Teile nicht ausreichen, um eine vollständige neue Sprache zu implementieren, kann die Korrektheit nur eingeschränkt überprüft werden. In Festival sind neben diesen drei Aspekten noch weitere Regeln zu definieren, dazu zählen unter anderem Postlexikalische Regeln, Regeln zur Tokenisierung (Expansion von Ziffern, Abkürzungen, Akronymen, Daten, E-Mail Adressen usw.), Syllabifizierung (Silbifizierung), Lautdauermodellierung, F0-Generierung, Wortarterkennung (part-of-speech tagging).

Um einen ersten Eindruck davon zu erhalten, wie „einfach“ die kroatische Orthographie tatsächlich ist und wie viel sich schon allein mit der Definition der oben genannten Ausspracheregeln erreichen läßt habe ich sie an ausgewählten journalistischen Texten getestet (statistisch ausgedrückt eine Teilerhebung mit bewußter Auswahl „typischer“ Fälle nach subjektiven Kriterien). Diese Texte setzten sich wie folgt zusammen: eine Meldung des Fernsehsenders *HRT*⁹ mit 160 Tokens (von Leerzeichen umschlossene Textelemente, ohne Satzzeichen), eine Kolumne der Zeitung *Novi list*¹⁰ mit 676 Tokens, ein Artikel der Zeitung *Slobodna Dalmacija*¹¹ mit 453 Tokens, zwei Artikel der Zeitung *Večernji list*¹² mit zusammen 531 Tokens, sowie einem Text aus dem IPA Handbuch mit 105 Tokens (IPA 1999, S. 69). Der letztgenannte Text wurde bewußt ausgewählt, da dieser im IPA Handbuch auch transkribiert ist und ich somit einen direkten Vergleich meiner automatischen Transkription mit der im IPA Handbuch dargestellten anstellen kann (vgl. Abschnitt 4.5.2).

Eine vollständige Text-to-Speech Synthese ist zu diesem Zeitpunkt natürlich noch nicht möglich. Da es sich beim Festival System um eine konkatenative Synthese handelt ist ein Audioinventar nötig, um das Sprachsignal letztendlich zu erzeugen. JURAJ BAKRAN hat ein kroatisches Diphoninventar für MBROLA¹³ erstellt (vgl. Bakran 1998). Dieses ließe sich, laut MBROLA Lizenzbestimmungen für nicht-kommerzielle und nicht-militärische Zwecke, für die Synthese in Festival integrieren. Im Rahmen dieser Arbeit habe ich allerdings auf diesen Schritt verzichtet; wodurch dann der eigentliche Zweck der Text-to-Speech Synthese, die automatische Erzeugung einer lautsprachlichen Äußerung, für das Kroatische leider nicht demonstriert werden kann.

4.5.1 Testvorgang und Auswertung

Die oben zitierten Texte wurden in Festival als ganze Texte „synthetisiert“ und die erzeugte Festival *utterance*-Struktur in eine Textdatei exportiert (vgl. Black et. al. 1999; Black & Lenzo 2003). Aus dieser Struktur wurde nur die Transkription anhand der Silbenstruktur extrahiert und von Hand ausgewertet. Der folgende Text (aus dem *Večernji list*) wurde dann z.B. so von Festival analysiert: „Zna se da je pušenje u trudnoći čimbenik rizika za normalan razvoj ploda ...“

```
...
156 id _162 ; name zE ; stress 1 ;
157 id _165 ; name ?En ; stress 1 ;
158 id _168 ; name ?a: ; stress 1 ;
159 id _170 ; name sE ; stress 0 ;
160 id _173 ; name da ; stress 0 ;
161 id _176 ; name jE ; stress 0 ;
162 id _179 ; name pu: ; stress 1 ;
163 id _182 ; name SE ; stress 0 ;
164 id _185 ; name JE ; stress 0 ;
```

⁹Quelle: <http://www.hrt.hr/auto/vijesti/KRV.html#1128424043>

¹⁰„Kolumne, komentari: Sanader ojačao u svijetu, a kod kuće oslabio“ vom 17.12.2005 (<http://www.novolist.hr>)

¹¹Quelle: <http://www.slobodnadalmacija.hr/20050702/novosti04.asp>

¹²Quelle: <http://www.vecernji-list.hr/newsroom/zanimljivosti/442115/index.do> (ohne Überschrift) und <http://www.vecernji-list.hr/newsroom/culture/441482/index.do> (ohne Überschrift)

¹³Internet: <http://tcts.fpms.ac.be/synthesis/mbrola.html>

```

165 id _188 ; name ?u: ; stress 1 ;
166 id _190 ; name tru:d ; stress 1 ;
167 id _195 ; name n0 ; stress 0 ;
168 id _198 ; name tS_ji: ; stress 0 ;
169 id _201 ; name tSi:m ; stress 1 ;
170 id _205 ; name bE ; stress 0 ;
171 id _208 ; name ni:k ; stress 0 ;
172 id _212 ; name ri: ; stress 1 ;
173 id _215 ; name zi: ; stress 0 ;
174 id _218 ; name ka ; stress 0 ;
175 id _221 ; name za ; stress 0 ;
176 id _224 ; name nOr ; stress 1 ;
177 id _228 ; name ma ; stress 0 ;
178 id _231 ; name lan ; stress 0 ;
179 id _235 ; name raz ; stress 1 ;
180 id _239 ; name p0 ; stress 0 ;
181 id _242 ; name i_^ ; stress 0 ;
182 id _244 ; name pl0 ; stress 1 ;
183 id _248 ; name da ; stress 0 ;
...

```

An diesem kurzen Ausschnitt erkennt man bereits einige Fehler, die bei dieser Vorgehensweise entstanden sind. Das erste Wort *zna* wurde z.B. buchstabiert und nicht als Wort erkannt (zE ?En ?a:). Dies liegt daran, daß auch andere Module bei der Synthetisierung verwendet wurden, welche ich nicht für das Kroatische definiert bzw. angepaßt habe. Man erkennt auch, daß hier (nach Regeln für das Deutsche) ein ? wohl für den glottalen Verschlusslaut eingefügt wurde. Dieser ist im Kroatischen nicht obligatorisch und könnte sogar als Fehler angesehen werden. Wendet man die Transkriptionsregeln aber explizit auf das einzelne Wort an, so erhält man eine korrekte Analyse:

```

festival> (lts.apply "zna" 'croatian_lit_dd)
(z n a)

```

Der Befehl `lts.apply` wendet die vorher definierten `lts`-Regeln mit der Bezeichnung `'croatian_lit_dd` auf die angegebene Zeichenkette an. Mit dem Befehl `lex.lookup` erhält man eine noch weitergehende Analyse einer gegebenen Zeichenkette. In diesem Fall würde die Analyse wie folgt aussehen:

```

festival> (lex.lookup "zna")
("zna" nil (((z n a) 1)))

```

Neben der reinen Transkription ist nun auch die Silbenstruktur zu erkennen. Sieht man sich den oben dargestellten Ausschnitt der Textanalyse an, fällt auch auf, daß die Silbenstruktur nicht in allen Fällen korrekt ist — dies war so aber auch zu erwarten, da die Regeln zur Syllabifizierung nicht an das Kroatische angepaßt wurden. Was vor allem zu Fehlerhaften Silbengrenzen führte waren die von mir definierten nicht-silbischen Vokale. Die standardmäßige Bestimmung der Silbengrenzen unbekannter, d.h. nicht im Lexikon enthaltener Wörter,

orientiert sich in Festival vor allem am Merkmal *vc*, das angibt ob es sich bei einem Laut um einen Vokal oder einen Konsonanten handelt. Auch eine Änderung des Lautinventars, so daß die nicht-silbischen Vokale in diesem Merkmal den Wert Minus erhalten haben, hat die fehlerhafte Bestimmung der Silbengrenzen nicht wesentlich verbessert. Dies bedeutet, ein speziell für das Kroatische entworfener Algorithmus muß an dieser Stelle unbedingt eingesetzt werden. Ein weiteres (bedeutendes) Problem ist die Bestimmung der akzentuierten Silbe, sowie der Lautdauer. Dies konnte im Rahmen dieser Arbeit aber nicht gelöst werden (siehe Abschnitt 3.2).

Weitere Fehler ergaben sich vor allem aus der fehlenden Behandlung von Ziffern, Abkürzungen/Akronymen und Fremdwörtern. Eine Zusammenfassung der enthaltenen Fehler der hier teilweise transkribierten insgesamt 1925 Tokens umfaßt 552 einzelne Fehler in der Transkription¹⁴. Davon entfallen allein 285 auf fehlerhaft gesetzte Silbengrenzen. Von den 267 verbleibenden Fehlern, entfallen 205 auf fälschlicherweise buchstabenweise transkribierte (d.h. buchstabierte) Wörter. Zieht man diese auch noch ab, so verbleiben 62 Fehler, die einer näheren Betrachtung bedürfen.

Insgesamt 17 Fehler entstanden durch Ziffern im Text, da ich keine Methode zum Umgang mit Ziffern und Zahlen definiert hatte. 7 Fehler entstanden durch die im Text enthaltenen Fremdwörter *croatian medical journala* (Zeitschriftentitel, mit kroatischer Deklinationsendung im letzten Wort), *performance*, *Haag*, *Haagom* (wieder mit kroatischer Deklinationsendung) sowie *Carla del Ponte*. Letzteres wurde als (tsar la dEl pOn tE) transkribiert, wobei der Nachname hier zufälligerweise korrekt gewesen wäre. 2 Fehler entstanden durch die Abkürzungen *tzv.* und *dr.* — beide könnten nur durch einen einfachen Lexikoneintrag ohne Kasus und Genus Information nicht korrekt expandiert werden, da sie je nach dem Kontext flektiert werden (z.B. *tzv.* > *takozvani* / *takozvane* / *takozvanoj* / *takozvanih* / *takozvanoga* usw, dt. sogenannt...). Zu bemerken ist auch noch, daß die fälschlicherweise angewandte Buchstabierung in vielen Fällen auch dazu geführt hat, daß enthaltene Akronyme zufälligerweise annähernd korrekt transkribiert wurden, wie z.B. *HDZ-a* > (xa dE zE ?a:) (mit Deklinationssuffix *-a*).

Von den Transkriptionsregeln wurden folgende nicht (immer) korrekt¹⁵ angewendet:

- Die Realisierung von /d/ + /n/ als [dⁿ]
In den Worten *svakodnevno*, *trodnevna* und *ugledni* wird die Zeichenfolge „dn“ nicht als *d_n* transkribiert, was im Grunde nicht unbedingt falsch wäre, aber eine falsch gesetzte Silbengrenze zur Folge hatte. Die Einzelanalyse (mit `lex.lookup`) für diese drei Wörter liefert folgendes Ergebnis:


```

("svakodnevno" nil (((s P a) 1) ((k 0 d) 0) ((n E P) 0) ((n 0) 0)))
("trodnevna" nil (((t r 0 d) 1) ((n E P) 0) ((n a) 0)))
("ugledni" nil (((u g) 1) ((l E d) 0) ((n i) 0)))
      
```
- Realisierung der Zeichenkette „ije“ als /iɛ/ oder /ije/
Von insgesamt 56 Vorkommen dieser Zeichenkette in den gewählten Texten, wurden 4 falsch transkribiert. Dies waren: *kvalitetnijeg*, *čijem*, *prijevremene* und *urijeme*. Die ersten drei sollten das zweisilbige /ije/ und das letzte den Diphthong /iɛ/ enthalten. Analysiert wurden sie allerdings genau andersherum, also z.B. wie folgt:

¹⁴nicht berücksichtigt sind hierbei fehlerhafte Aussprache von Satzzeichen, Vokallängen, das ? sowie Fehler in der Akzentstelle und der Akzentart

¹⁵„korrekt“ heißt in diesem Zusammenhang: „so wie erwartet“

```
("prijevremene" nil ((p r iE P) 1) ((r E) 0) ((m E) 0) ((n E) 0)))
("vrijeme" nil ((P r i) 1) ((j E) 0) ((m E) 0)))
```

- Ungenügende Lexikoneinträge

Zwei Fehler sind aufgrund des Lexikons aufgetreten. Das *s* stellt einen Homographen im Kroatischen dar. Es kann einerseits eine Präposition darstellen, andererseits aber auch ein Nomen (das „S“). Eine Unterscheidung wäre leicht anhand der Wortart (prep / nn) zu treffen. Da ich allerdings über keinen POS-Tagger zur Wortarterkennung für das Kroatische verfüge wurde zur Analyse der einzige vorhandene Lexikoneintrag verwendet und *s* zu [Es] transkribiert. Diese Entscheidung, das Nomen anstelle der Präposition im Lexikon zu verwenden hat sich als Fehler herausgestellt, da das *s* allein in den hier getesteten Texten 16 mal vorgekommen ist und damit auch 16 Fehler hervorgerufen hat.

Einen weiteren Fehler hat das Wort *jedanaesteročlanog* (dt. elfköpfig / elf Mitglieder zählend) hervorgerufen. Die Zahlen von 11 bis 19 werden abweichend von den Standardregeln ausgesprochen bzw. geschrieben. Für *jedanaest* (dt. elf) enthält das Lexikon folgenden Eintrag:

```
(lex.add.entry
  '("jedanaest" nil ((i_ E) 0) ((d a) 2) ((n a i_ s t) 0) ))
```

Das Wort *jedanaesteročlanog* allerdings ist nicht im Lexikon enthalten und wird folglich nach den Standardregeln mit (i_ E d a n a E s t E r 0 tS l a n 0 g) transkribiert.

- Probleme mit -VjV-

Das *j* zwischen zwei Vokalen stellt im Kroatischen generell ein Problem dar, da es entweder nur als stark reduziertes [j] bzw. nicht-silbisches [j̥] realisiert wird oder vollständig wegfällt. Insgesamt traten so 11 Analysen auf, die man als nicht korrekt bezeichnen kann. Die Wörter *muzejima*, *kojima* oder *svojim* wurden z.B. mit einem nicht-silbischen [j̥] transkribiert, wo keines sein sollte:

```
(m u z E i_ i m a)
(k 0 i_ i m a)
(s P 0 i_ i m)
```

In *prijavljujemo* hingegen wurde das (erste) *j* gelöscht, obwohl es ausgesprochen wird:

```
(p r i a P L u i_ E m 0)
```

4.5.2 Vergleich mit manueller Transkription

Um einen groben Eindruck über die Qualität der mit den in dieser Arbeit entwickelten Ausspracheregeln erreichten Transkription zu geben, möchte ich nun einen der transkribierten Texte in voller Länge darstellen und diesem die Transkription aus dem IPA Handbuch (IPA 1999, S. 69) gegenüberstellen. Die anhand meiner Regeln erstellte Transkription ist in der Spalte DD zu sehen. Von Festival gesetzte Akzente sind durch " gekennzeichnet, falsch gesetzte Silbengrenzen durch | und in Klammern gesetzte Transkriptionen sind nachträglich mit Hilfe von `lex.lookup` von Hand hinzugefügt, da an dieser Stelle buchstabenweise transkribiert wurde.

IPA	DD
sjêve:rn̩i: l̩ed̩eni: u̩jêtar i̩ s̩untse su se pr̩êpirali o su̩ôjɔj sn̩ã:zi	"si̩^ EPe:rn̩i: "lEdEni: "Pi̩^ Etar "?i: "su:ntsE (su) sE "prEpi:r̩ali: "?0 "sPOi̩^ 0 i̩^ ("snazi)
stôga ôdlut̩fe: da ôno:me ôd̩ni:x pri̩padne pôbjeda kôji sv̩ut̩fe t̩fôvjeka p̩r̩tnika	"stOga "?Odlu:tSE da "?OnOmE ?Od ("Jix) "pri:pad nE "pObi̩^ Eda ("kOi̩^ i) "sPu:tSE "tSOPi̩^ Eka "pu:t̩ni:ka
ujêtar zãpot̩fe sn̩ãzno p̩uxati a b̩udut̩ci da je t̩fôv̩jek t̩fu:sto d̩r̩zao ôd̩jet̩cu nãvali ô:n jôf jãt̩fe:	"Pi̩^ Etar "zapOtSE "snaZnO "pu:xati: "?a: "bu:du:tS_ji: da jE "tSOPi̩^ Ek "tSPr_=stO "dr_=ZaO "?Odi̩^ EtS_ju: "naPali: "?On ("i̩^ oS) ("i̩^ atSE)
t̩fôv̩jek pãk jôf jãt̩fe: ot̩ st̩udeni pr̩it̩snu:t nau̩r̩t̩fe nã sebe jôf v̩ife: ôd̩jet̩ce: dôk se ujêtar ne ũmori: i̩ pr̩êpusti ga tãda s̩ũ:nts̩u	"tSOPi̩^ Ek "pak ("i̩^ oS) ?Od "stu:dEni: "pri:t̩i:snu:t "naPu:tSE na "sEbE ("i̩^ oS) "Pi:SE "?Odi̩^ EtS_jE "dOk sE "Pi̩^ Etar nE "?u:mOri: "?i: "prEpu:sti: ga "tada ("suntsu)
ôno: u pot̩f̩et̩ku zãsija ũm̩jeren̩o	"?OnO "?u: "pOtSEtku: "zasi:a "?u:mi̩^ ErEnO
kãd je t̩fôv̩jek sk̩nuo suv̩i:fak ôd̩jet̩ce: pôv̩i:si ôno: jôf jãt̩fe: z̩êgu dôk se t̩fôv̩jek	kad jE "tSOPi̩^ Ek "ski:nu:O "su:Pi:Sak "?Odi̩^ EtS_jE "pOPi:si: "?OnO ("i̩^ oS) "?i̩^ atSE "ZEgu: "dOk sE "tSOPi̩^ Ek
u nemog̩t̩nosti da ôdoli s̩ũt̩feov̩j topl̩ni ne sv̩i: t̩fe:	"?u: "nEmOgu:tS_jnOsti: "da "?OdOli: "su:ntSEPO i̩^ "tOp li:ni: nE "sPu:tSE
i n̩ê pod̩ze: na k̩ũ:pape u r̩ij̩ê:ku tek̩ũt̩its̩u	"?i: nE "pOdZ_jE na "ku:paJE "?u: "riEku: "tEku:tS_ji:t̩su:
pr̩i:t̩fa pokãz̩uje: da je t̩f̩ê:sto uspj̩j̩n̩ije: uujerã:va:pe n̩êgoli nã:si:ãe	pri:tSa pOkazu:i̩^ E da jE tSEstO "?u:spi̩^ ESni:jE "?u:Pi̩^ EraPaJE nEgOli: nasi:LE

Eine zusätzliche Angabe der orthographischen Form ist hier nicht unbedingt notwendig. Es muß allerdings angemerkt werden, daß in der IPA Transkription die Zeichenfolge *ije* mit [ije] transkribiert wurde. Die Autoren schreiben dazu folgendes:

„ The diphthong /ie/ begins at the position of the /i/ monophthong and ends at the position of the monophthong /e/. It can also be pronounced [ije] but this still functions as a single syllable.“ (IPA 1999)

Ich habe mich, wie bereits gesagt, dazu entschlossen den Diphthong als /iE/ zu transkribieren. Daher ergibt sich der Unterschied in der Transkription von *rijeku* (dt. Fluß L): [rij̩ê:ku]

und [riEku:]. Außerdem verwenden die Autoren der IPA Transkription die Zeichen /e/ und /o/ wo ich mich für /E/ und /O/ entschieden habe.

Die in der rechten Spalte ohne Akzentzeichen versehenen Wörter sind in dieser Form aus dem Lexikon entnommen (als nicht-akzentuierte Klitika, siehe Anhang B). In der Phrase *od njih* ergibt sich hier ein Unterschied im Gegensatz zur IPA Version aufgrund eines postlexikalischen Effektes, der im Rahmen dieser Arbeit nicht implementiert wurde: Proklitika können in einer prosodischen Einheit den Wortakzent auf sich ziehen, so daß die normalerweise nicht-akzentuierten Proklitika den Akzent tragen. Die IPA Transkription lautet also [ôd̥ni:x], wohingegen der Akzent in meiner Version nicht verschoben wurde: [ʔOd̥ "Jix]. Das selbe gilt für [i nĕ podze:] im Gegensatz zu ["ʔi: nE "pOdZ_jE]. An [ot̥st̥udeni] im Gegensatz zu [ʔOd̥ "stu:dEni:] läßt sich der Effekt der Angleichung an die Stimmhaftigkeit beobachten, der in diesem Fall in den postlexikalische Regeln definiert werden müßte. Eindeutig falsch ist die Transkription von *koji* ["kOĩ_ˆ|i], sowie die automatisch (von nicht von mir angepaßten Modulen) gesetzten ?, die Längenzeichen (:) und in den meisten Fällen die Akzentuierung.

Schlußbetrachtungen

Das Ziel dieser Arbeit war die Erstellung von Ausspracheregeln in der Text-to-Speech Synthese für das Kroatische. Zu diesem Zweck wurde im ersten Kapitel zunächst eine allgemeine Einführung gegeben, die bereits auf einige spezielle Probleme einging. Zu erwähnen sind hier von vor allem die Probleme, welche sich aus den unterschiedlichen Zeichenkodierungssystemen ergeben können. Da das Kroatische, im Vergleich zum Englischen, zusätzliche Buchstaben verwendet, ist eine Kodierung der Texte in ASCII nicht (orthographisch korrekt) möglich. Dieses Problem wurde dann später bei der Implementierung der Ausspracheregeln für das Festival Sprachsynthesesystem bedeutend.

Da die Kroatische Sprache in Deutschland nicht so bekannt und nicht oft beschrieben ist wie viele andere europäische und nicht-europäische Sprachen wurde die Beschreibung des Kroatischen in dieser Arbeit entsprechend umfangreich behandelt. Die Beschreibung der einzelnen linguistischen Merkmale richtete sich dabei immer nach den Erfordernissen einer Text-to-Speech Synthese. Es stellte sich heraus, daß die Beschreibungen des Kroatischen in der Literatur oft nicht einheitlich sind. Eine für Linguisten besonders interessante und in der Literatur oft angesprochene Eigenheit des Kroatischen, wie auch der gesamten kroatisch-bosnisch-serbischen Sprachen, ist der polytonische Wortakzent. Dieser konnte von mir im Rahmen dieser Arbeit aber nicht behandelt werden. Obwohl der korrekte Wortakzent, wie auch die gesamte Prosodie, zur Synthetisierung einer natürlich klingenden Stimme unabdingbar ist. Der Grund für das Weglassen des Wortakzentes aus der Erstellung der Ausspracheregeln war der sehr große Umfang und die Komplexität dieses Themas. Im Rahmen dieser Studienarbeit wäre es nicht möglich gewesen eine ausreichend genaue Darstellung des Wortakzentes zu erreichen. Dies wurde ausführlich in Kapitel 3, in Abschnitt 3.2, dargestellt.

Von der kroatischen Orthographie wird in der Literatur oft als einer „phonologischen“ gesprochen. Dies wurde in Kapitel 2 ausführlich behandelt. Die Orthographie ist für ein Text-to-Speech System besonders wichtig, da es nach seiner Definition geschriebenen Text als Eingabedaten erhält. Wie sich in dieser Arbeit herausgestellt hat ist die kroatische Orthographie tatsächlich zu einem hohen Grad phonologisch, was sich in der hohen Präzision der Transkription unter Verwendung nur sehr weniger Regeln bemerkbar macht. Die in dieser Arbeit entwickelten Ausspracheregeln wurden im Festival Formalismus in ein Scheme-Skript gefaßt und so anhand einiger Texte getestet. Dabei zeigte sich, daß die transkribierten Laute fast alle korrekt waren. Mit nur etwa 50 falsch transkribierten Lauten in einem Text von 1925 Tokens wurde damit schon eine hohe Präzision erreicht. Natürlich handelt es sich hierbei nur um eine stichprobenartige Untersuchung (siehe Abschnitt 4.5). Was nicht von den Regeln abgedeckt wurde zeigte allerdings eine hohe Fehlerrate. Die Behandlung von Ziffern und Abkürzungen sowie die Disambiguierung von Homographen und die Syllabifizierung wurden nicht implementiert was zu vielen Fehlern führte. In dem getesteten Text traten z.B. 285

Fehler allein aufgrund falscher Silbengrenzen auf.

Im Bereich der Sprachsynthese ist für das Kroatische noch viel Arbeit nötig. Es gibt noch eine Reihe von offenen Fragen, wie diese Arbeit gezeigt hat. Dabei ist teilweise auch noch Grundlagenarbeit zu leisten, wie etwa bei der Beschreibung des Wortakzentes, oder der abschließenden Beschreibung des Gesamtkroatischen Lautinventars. Auch zeigte sich in dieser Arbeit, daß es an einigen (leicht verfügbaren) Ressourcen für die computerlinguistische Arbeit mangelt. Zu erwähnen wäre etwa ein umfangreiches maschinenlesbares Aussprachelexikon (falls es ein solches gibt ist mir dies — auch nach umfangreichen Recherchen — derzeit nicht bekannt). Neben dem Aussprachelexikon wäre auch ein Modul zur morphologischen Analyse sowie ein POS-Tagger äußerst hilfreich, da Kroatisch eine stark flektierende Sprache ist und ein Vollformenlexikon für solche Sprache nicht besonders praktisch erscheint. Zur weiteren Arbeit mit dem Festival System wäre die Integration eines Audioinventars zur Synthese sowie die Unterstützung von Unicode unbedingt notwendig.

Ich denke meine Arbeit hat gezeigt, was auf diesem Gebiet möglich ist.

Anhang A

Phonset

```
----- phone_set_croatian.scm -----
1 ;; Croatian PhoneSet
2 ;; (C) 2005 Daniel Duran - all rights reserved
3 ;; (ISO 8859-2 encoded)
4 ;; -----
5 ;; -----
6 ;; to use the syll feature change standard syllabification in:
7 ;; src/arch/festival/Phone.cc:ph_sonority()
8 ;; or define new lts_function with syllabification
9 (defPhoneSet
10  croatian_lit_dd
11  (
12   ;; Phone Features -----|
13   ;; (sketch)
14   ;; vowel or consonant
15   (vc + -)
16   ;; vowel class: monophthong diphthong semivowel
17   (vclass m d s 0)
18   ;; vowel height: high mid low
19   (vheight 1 2 3 -)
20   ;; vowel frontness: front mid back
21   (vfront 1 2 3 -)
22   ;; lip rounding
23   (vrnd + - 0)
24   ;; consonant type: stop fricative affricative nasal trill lateral approximant
25   (ctype s f a n t l x 0)
26   ;; place of articulation: bilabial labiodental dental alveolar post-alveolar
27   ;;                               palatal velar glottal
28   ;;                               dental+lateral-release dental+nasal-release
29   (cplace bl ld dt av pa pl vl gt lr nr 0)
30   ;; voice
31   (cvox + -)
32   ;; syllabicity
33   (syll + -)
34   ;; (syllable) length -- distinction needed for lexical encoding of word
35   ;; accents (?)
```

```

36 (long + - 0)
37 )
38 (
39 ;; Phone set members -----|
40 ;;
41 ;; * SAMPA = Symbols used in SAMPA transcription by Bakran and Horga
42 ;;           http://www.phon.ucl.ac.uk/home/sampa/croatian.htm
43 ;; .....|
44 ;;   V/C?
45 ;;   . VCLASS
46 ;;   . . VH
47 ;;   . . . VF
48 ;;   . . . . VRND
49 ;;   . . . . . CTYPE
50 ;;   . . . . . . CPLACE
51 ;;   . . . . . . . VOX
52 ;;   . . . . . . . . SYLL
53 ;;   . . . . . . . . . LONG
54 ;;   . . . . . . . . . .
55 (# - 0 - - - 0 0 - - 0) ; silence
56 ; syllabic vowels (short)
57 (a + m 3 2 - 0 0 + + -)
58 (E + m 2 1 - 0 0 + + -) ; SAMPA*: e
59 (i + m 1 1 - 0 0 + + -)
60 (O + m 2 3 + 0 0 + + -) ; SAMPA*: o
61 (u + m 1 3 + 0 0 + + -)
62 ; syllabic vowels (long)
63 (a: + m 3 2 - 0 0 + + +)
64 (E: + m 2 1 - 0 0 + + +) ; SAMPA*: e
65 (i: + m 1 1 - 0 0 + + +)
66 (O: + m 2 3 + 0 0 + + +) ; SAMPA*: o
67 (u: + m 1 3 + 0 0 + + +)
68 (iE + d 2 1 + 0 0 + + +)
69 ; non-syllabic vowels
70 (@ + s 2 2 - 0 0 + - 0)
71 (i_~ + m 1 1 - 0 0 + - 0)
72 (u_~ + m 1 3 + 0 0 + - 0)
73 ; consonants: stops / plosives
74 (p - 0 - - 0 s bl - - 0)
75 (b - 0 - - 0 s bl + - 0)
76 (t - 0 - - 0 s dt - - 0)
77 (t_n - 0 - - 0 s nr - - 0) ; nasal release
78 (t_l - 0 - - 0 s lr - - 0) ; lateral release
79 (d - 0 - - 0 s dt + - 0)
80 (d_n - 0 - - 0 s nr + - 0) ; nasal release
81 (d_l - 0 - - 0 s lr + - 0) ; lateral release
82 (k - 0 - - 0 s vl - - 0)
83 (g - 0 - - 0 s vl + - 0)
84 ;(? - 0 - - 0 s gt - - 0)
85 ; consonants: fricatives
86 (f - 0 - - 0 f ld - - 0)
87 (v - 0 - - 0 f ld + - 0)
88 (s - 0 - - 0 f dt - - 0)
89 (z - 0 - - 0 f dt + - 0)

```

```

90 (S - 0 - - 0 f pa - - 0)
91 (Z - 0 - - 0 f pa + - 0)
92 (S_j - 0 - - 0 f pl - - 0) ;
93 (Z_j - 0 - - 0 f pl + - 0)
94 (x - 0 - - 0 f vl - - 0)
95 (G - 0 - - 0 f vl + - 0)
96 ;(h - 0 - - 0 f gt - - 0)
97 ;(h\ - 0 - - 0 f gt + - 0) ; voiced glottal fricative?
98 ; consonants: affricatives
99 (ts - 0 - - 0 a dt - - 0)
100 (dz - 0 - - 0 a dt + - 0)
101 (tS - 0 - - 0 a pa - - 0)
102 (dZ - 0 - - 0 a pa + - 0)
103 (tS_j - 0 - - 0 a pl - - 0)
104 (dZ_j - 0 - - 0 a pl + - 0)
105 ; consonants: nasals
106 (m - 0 - - 0 n bl + - 0)
107 ;(F - 0 - - 0 n ld + - 0)
108 (n - 0 - - 0 n av + - 0)
109 (n_= - 0 - - 0 n av + + -) ; syllabic /n/
110 (J - 0 - - 0 n pl + - 0)
111 (N - 0 - - 0 n vl + - 0)
112 ; consonants: trills
113 (r - 0 - - 0 t av + - 0)
114 (r_= + 0 - - 0 t av + + -) ; syllabic /r/
115 (r_=: + 0 - - 0 t av + + +) ; syllabic /r:/
116 ; consonants: laterals
117 (l - 0 - - 0 l av + - 0)
118 (l_= - 0 - - 0 l av + + -) ; syllabic /l/
119 (L - 0 - - 0 l pl + - 0)
120 ; consonants: approximants
121 (w - 0 - - 0 x bl + - 0)
122 (P - 0 - - 0 x ld + - 0) ; SAMPA*: v\
123 (j - 0 - - 0 x pa + - 0)
124 )
125 )
126 (PhoneSet.silences '#)
127

```

phone_set_croatian.scm

Anhang B

Lexikon

```
lexicon_croatian.scm
1 ;; -----|
2 ;; -----|
3 ;; Croatian lexicon (addenda - incomplete) |
4 ;; |
5 ;; (C) 2005 Daniel Duran - all rights reserved |
6 ;; file name: lexicon_croatian.scm |
7 ;; (ISO 8859-2) |
8 ;; |
9 ;; == Lexical word accent == |
10 ;; falling accent: 1 - SAMPA: <F> |
11 ;; rising accent: 2 - SAMPA: <R> |
12 ;; source 1: Vladimir Anić & Ivo Goldstein "Rječnik stranih riječi" |
13 ;; source 2: Vladimir Anić "Rječnik hrvatskoga jezika" |
14 ;; source 3: Barić et al "Gramatika" |
15 |
16 (lex.add.entry '("EU" nil (((E) 0) ((u) 1)) )) ;?
17 ;; .....|
18 ;; orthographic exceptions:
19 |
20 (lex.add.entry '("croatia" nil (((k r 0) 0) ((a) 2) ((ts i) 0) ((a) 0)) )) ;1
21 (lex.add.entry '("croatije" nil (((k r 0) 0) ((a) 2) ((ts i) 0) ((i~ E) 0)) ))
22 (lex.add.entry '("croatiju" nil (((k r 0) 0) ((a) 2) ((ts i) 0) ((i~ u) 0)) ))
23 (lex.add.entry '("croatiji" nil (((k r 0) 0) ((a) 2) ((ts i) 0) ((i) 0)) ))
24 |
25 ;; numbers 11-19:
26 (lex.add.entry '("jedanaest" nil (((i~ E) 0) ((d a) 2) ((n a i~ s t) 0)) )) ;3
27 (lex.add.entry '("dvanaest" nil (((d v a:) 2) ((n a i~ s t) 0)) )) ;3 + 2
28 (lex.add.entry '("trinaest" nil (((t r i:) 2) ((n a i~ s t) 0)) )) ;3 + 2
29 (lex.add.entry '("četnaest" nil (((tS E) 0) ((t r=) 2) ((n a i~ s t) 0)) )) ;3 + 2
30 (lex.add.entry '("petnaest" nil (((p E t) 2) ((n a i~ s t) 0)) )) ;3 + 2
31 (lex.add.entry '("šesnaest" nil (((S E s) 2) ((n a i~ s t) 0)) )) ;3 + 2
32 (lex.add.entry '("sedamnaest" nil (((s E) 0) ((d a m) 2) ((n a i~ s t) 0)) )) ;3 + 2
33 (lex.add.entry '("osamnaest" nil (((0) 0) ((s a m) 2) ((n a i~ s t) 0)) )) ;3 + 2
34 (lex.add.entry '("devetnaest" nil (((d E) 0) ((P E t) 2) ((n a i~ s t) 0)) )) ;3 + 2
35
```

```

36 ;; classical music terminology (incomplete)
37 (lex.add.entry '("adagio" nil (((a) 0)((d a) 2)((dZ 0) 0)) )) ;1
38 (lex.add.entry '("allegro" nil (((a) 0) ((l E:) 1) ((g r 0) 0)) )) ;1
39 (lex.add.entry '("allegretto" nil (((a) 0) ((l E) 0) ((g r E:) 1) ((t 0) 0)) )) ;1
40 ;(lex.add.entry '("canto" nil ( ) )) ; "bel canto" --> homographs: [bElk"ant0] < bel canto
41 ;                                     ["bEl] (phys.)
42 (lex.add.entry '("furioso" nil (((f u) 0) ((r i) 0) ((0:) 1) ((z 0) 0)) )) ;1
43 (lex.add.entry '("intermezzo" nil (((i n) 0) ((t E r) 0) ((m E) 2) ((ts 0) 0)) )) ;1
44 (lex.add.entry '("staccato" nil (((s t a) 0) ((k a:) 1) ((t 0) 0)) )) ;1
45 (lex.add.entry '("vivace" nil ((v i) 0) ((v a:) 1) ((tS E) 0)) )) ;1 ; ...
46
47 ;; foreign words, cultural and technical terminology (incomplete)
48 (lex.add.entry '("atelje" nil (((a) 2) ((t E l) 0) ((j E) 0)) )) ;2
49 (lex.add.entry '("copyright" nil (((k 0) 2) ((p i) 0) ((r a i_ t) 0)) )) ;1
50 (lex.add.entry '("grandprix" nil (((g r a n) 0) ((p r i:) 1)) ))
51 ; standard form: "grand prix" --> homographs: [gra:n <F>] < lat. granum,
52 ;                                     [gran <F>] (gradn coeur),
53 ;                                     [gran] (grand prix, grand mal),
54 ;                                     [grE:nd <F>] (grand slam)
55 (lex.add.entry '("reljef" nil ((r E l) 2) ((j E f) 0)) )) ;2
56 (lex.add.entry '("restaurant" nil ((r E s) 0) ((t 0) 2) ((r a: n) 0)) )) ;1
57 ; ...
58
59 ;; foreign words with non-syllabic u (incomplete)
60 (lex.add.entry '("kauzalan" nil (((k a u_) 1) ((z a:) 0) ((l a n) 0)) )) ;1
61 (lex.add.entry '("kauzalni" nil (((k a u_) 1) ((z a: l) 0) ((n i:) 0)) )) ;1
62 (lex.add.entry '("kauzalno" nil (((k a u_) 1) ((z a: l) 0) ((n 0:) 0)) )) ;1
63 (lex.add.entry '("kauzalnost" nil (((k a u_) 0) ((z a: l) 2) ((n 0: s t) 0)) )) ;1
64 (lex.add.entry '("kauzalnosti" nil (((k a u_) 0) ((z a: l) 2) ((n 0: s) 0) ((t i) 0)) )) ;1
65 (lex.add.entry '("kauzalnošću" nil
66     ((k a u_) 0) ((z a: l) 2) ((n 0: S_j) 0) ((tS_j u) 0)) ))
67 (lex.add.entry '("kauzalitet" nil (((k a u_) 0) ((z a) 0) ((l i) 2) ((t E: t) 0))))
68 (lex.add.entry '("nautika" nil (((n a u_) 2) ((t i) 0) ((k a) 0)) )) ;1
69 (lex.add.entry '("nautike" nil (((n a u_) 2) ((t i) 0) ((k E) 0)) ))
70 (lex.add.entry '("nautiku" nil ( ((n a u_) 2) ((t i) 0) ((k u) 0) )) ) ; ...
71 ;; (family) names with gj --> [dZ'] (arhaic orthography) (sketch)
72 ;(lex.add.entry '("gjuro" nil ((dZ' u:) 1) ((r 0) 0)) )) ;?
73 ;; .....|
74 ;; "unchangable" words |
75 ;; isolated letters (nouns?)
76 (lex.add.entry '("a" nil ( ((a:) 1) )) ) ; POS: nm ?
77 (lex.add.entry '("b" nil ( ((b E) 1) )) )
78 (lex.add.entry '("c" nil ( ((ts E) 1) )) )
79 (lex.add.entry '("č" nil ( ((tS E) 1) )) )
80 (lex.add.entry '("ć" nil ( ((tS_j E) 1) )) )
81 (lex.add.entry '("d" nil ( ((d E) 1) )) )
82 (lex.add.entry '("dž" nil ( ((dZ E) 1) )) )
83 (lex.add.entry '("đ" nil ( ((dZ_j E) 1) )) )
84 (lex.add.entry '("e" nil ( ((E:) 1) )) )
85 (lex.add.entry '("f" nil ( ((E f) 1) )) )
86 (lex.add.entry '("g" nil ( ((g E) 1) )) )
87 (lex.add.entry '("h" nil ( ((x a) 1) )) )
88 (lex.add.entry '("i" nil ( ((i:) 1) )) )
89 (lex.add.entry '("j" nil ( ((j E) 1) )) )

```

```

90 (lex.add.entry '("k" nil ( ((k a) 1) )) )
91 (lex.add.entry '("l" nil ( ((E l) 1) )) )
92 (lex.add.entry '("lj" nil ( ((E L) 1) )) )
93 (lex.add.entry '("m" nil ( ((E m) 1) )) )
94 (lex.add.entry '("n" nil ( ((E n) 1) )) )
95 (lex.add.entry '("nj" nil ( ((E J) 1) )) )
96 (lex.add.entry '("o" nil ( ((O:) 1) )) )
97 (lex.add.entry '("p" nil ( ((p E) 1) )) )
98 (lex.add.entry '("q" nil ( ((k P E) 1) )) )
99 (lex.add.entry '("r" nil ( ((E r) 1) )) )
100 (lex.add.entry '("s" nil ( ((E s) 1) )) )
101 (lex.add.entry '("š" nil ( ((E S) 1) )) )
102 (lex.add.entry '("t" nil ( ((t E) 1) )) )
103 (lex.add.entry '("u" nil ( ((u:) 1) )) )
104 (lex.add.entry '("v" nil ( ((P E) 1) )) )
105 (lex.add.entry '("w" nil ( ((d u) 1) ((p l O) 0) ((P E) 1) )) )
106 (lex.add.entry '("x" nil ( ((i k s) 1) )) )
107 (lex.add.entry '("y" nil ( ((i) 1) ((p s i) 0) ((l O: n) 0) )) )
108 (lex.add.entry '("z" nil ( ((z E) 1) )) )
109 (lex.add.entry '("ž" nil ( ((Z E) 1) )) )
110
111 ;; adverbs
112 ;; adverbial pronouns (incomplete)
113 (lex.add.entry '("dokad" nil (((d O) 2) ((k u a d) 0)) ))
114 (lex.add.entry '("dokada" nil (((d O) 2) ((k a) 0) ((d a) 0)) ))
115 (lex.add.entry '("dokle" nil (((d O:) 2) ((k l E) 0)) ))
116 (lex.add.entry '("gdje" nil (((g d i_) E) 1)) ))
117 (lex.add.entry '("kad" nil ((k a d) 1)) ))
118 ;(lex.add.entry '("kada" nil ((k a) 2) ((d a) 0)) ) ; homograph! -> POS needed for resolving
119 (lex.add.entry '("kako" nil ((k a) 1) ((k O) 0)) ))
120 (lex.add.entry '("kamo" nil ((k a) 1) ((m O) 0)) ))
121 (lex.add.entry '("koliko" nil ((k O) 0) ((l i) 2) ((k O) 0)) ))
122 (lex.add.entry '("kud" nil ((k u d) 1)) ))
123 (lex.add.entry '("kuda" nil ((k u) 2) ((d a) 0)) ))
124 (lex.add.entry '("otkud" nil (((O t) 2) ((k u d) 0)) ))
125 (lex.add.entry '("otkuda" nil (((O t) 2) ((k u) 0) ((d a) 0)) ))
126 (lex.add.entry '("otkad" nil (((O t) 2) ((k a d) 0)) ))
127 (lex.add.entry '("otkada" nil (((O t) 2) ((k a) 0) ((d a) 0)) ))
128 (lex.add.entry '("odakle" nil (((O) 0) ((d a:) 2) ((k l E:) 0)) ))
129 (lex.add.entry '("tako" nil (((t a) 2) ((K O:) 0)) ))
130 (lex.add.entry '("vrlo" nil ((P r_=) 1) ((l O) 0)) ))
131 (lex.add.entry '("zato" nil ((z a) 2) ((t O) 0)) ))
132
133 ;; accented particles
134 (lex.add.entry '("zar" nil ((z a: r) 1)) ))
135 (lex.add.entry '("ta" nil ((t a) 1)) ))
136
137 ;; conjunctions (incomplete)
138 (lex.add.entry '("čim" nil (((tS i m) 1)) ))
139 (lex.add.entry '("budići" nil (((b u) 2) ((d u:) 0) ((tS_j i) 0)) ))
140 (lex.add.entry '("dakle" nil ((d a) 2) ((k l E) 0)) ))
141 (lex.add.entry '("iako" nil (((i) 1) ((a) 0) ((k O) 0)) ))
142 (lex.add.entry '("makar" nil ((m a) 2) ((k a r) 0)) ))
143 (lex.add.entry '("niti" nil (((n i) 1) ((t i) 0)) ))

```

```

144
145 ;; clitics (non accented words)
146 ;; 1. proclitics (can get accent from following accented words!)
147 ;; a. prepositions:
148 ;(lex.add.entry '("k" nil (((k) 0)) )) ; homograph! -> POS needed for resolving
149 ;(lex.add.entry '("o" nil (((0) 0)) )) ; homograph! -> POS needed for resolving
150 ;(lex.add.entry '("s" nil (((s) 0)) )) ; homograph! -> POS needed for resolving
151 ;(lex.add.entry '("u" nil (((u) 0)) )) ; homograph! -> POS needed for resolving
152 (lex.add.entry '("bez" nil (((b E z) 0)) ))
153 (lex.add.entry '("do" nil (((d 0) 0)) ))
154 (lex.add.entry '("iz" nil (((i z) 0)) ))
155 (lex.add.entry '("kod" nil (((k 0 d) 0)) )) ; homograph! -> POS needed for resolving
156 (lex.add.entry '("kroz" nil (((k r 0 z) 0)) ))
157 (lex.add.entry '("na" nil (((n a) 0)) ))
158 (lex.add.entry '("nad" nil (((n a d) 0)) ))
159 (lex.add.entry '("niz" nil (((n i z) 0)) ))
160 (lex.add.entry '("ob" nil (((0 b) 0)) ))
161 (lex.add.entry '("od" nil (((0 d) 0)) ))
162 (lex.add.entry '("po" nil (((p 0) 0)) ))
163 (lex.add.entry '("pod" nil (((p 0 d) 0)) )) ; homograph! -> POS needed for resolving
164 (lex.add.entry '("pri" nil (((p r i) 0)) ))
165 (lex.add.entry '("uz" nil (((u z) 0)) ))
166 (lex.add.entry '("za" nil (((z a) 0)) ))
167 (lex.add.entry '("zbog" nil (((z b 0 g) 0)) ))
168 (lex.add.entry '("pred" nil (((p r E d) 0)) ))
169 (lex.add.entry '("ispod" nil (((i s) 0) ((p 0 d) 0)) ))
170 (lex.add.entry '("iznad" nil (((i z) 0) ((n a d) 0)) ))
171 (lex.add.entry '("među" nil (((m E) 0) ((dZ_j u) 0)) ))
172 (lex.add.entry '("mimo" nil (((m i) 0) ((m 0) 0)) ))
173 ;(lex.add.entry '("nada" nil (((n a) 0) ((d a) 0)) )) ; homograph! -> POS needed for resolving
174 ;(lex.add.entry '("oko" nil (((0) 0) ((k 0) 0)) )) ; homograph! -> POS needed for resolving
175 (lex.add.entry '("prama" nil (((p r a) 0) ((m a) 0) ))
176 ;(lex.add.entry '("poda" nil (((p 0) 0) ((d a) 0)) )) ; homograph! -> POS needed for resolving
177 (lex.add.entry '("pokraj" nil (((p 0) 0) ((k r a i_) 0)) ))
178 (lex.add.entry '("preko" nil (((p r E) 0) ((k 0) 0)) ))
179 (lex.add.entry '("prema" nil (((p r E) 0) ((m a) 0)) ))
180 (lex.add.entry '("preko" nil (((p r E) 0) ((k 0) 0)) ))
181 (lex.add.entry '("spram" nil (((s p r a m) 0)) ))
182 (lex.add.entry '("između" nil (((i z) 0) ((m E) 0) ((dZ_j u) 0)) ))
183 (lex.add.entry '("umjesto" nil (((u) 0) ((m i_ E s) 0) ((t 0) 0)) ))
184 ;; b. conjungtions (incomplete)
185 ;(lex.add.entry '("a" nil () )) ; homograph! -> POS needed for resolving
186 ;(lex.add.entry '("i" nil () )) ; homograph! -> POS needed for resolving
187 (lex.add.entry '("da" nil (((d a) 0)) ))
188 (lex.add.entry '("kad" nil (((k a d) 0)) )) ; can be also accented (source: 3)
189 (lex.add.entry '("ma" nil (((m a) 0)) ))
190 (lex.add.entry '("ni" nil (((n i) 0)) ))
191 (lex.add.entry '("pa" nil (((p a) 0)) ))
192 (lex.add.entry '("što" nil (((S t 0) 0)) ))
193 ;; c. negation particle
194 (lex.add.entry '("ne" nil (((n E) 0)) )) ; homograph(?) -> POS needed for resolving
195
196 ;; 2. enclitics (unaccented words)
197 ;; a. personal pronouns (short forms)

```

```

198 (lex.add.entry '("ga" nil (((g a) 0)) ))
199 (lex.add.entry '("ih" nil (((i x) 0)) ))
200 (lex.add.entry '("im" nil (((i m) 0)) ))
201 (lex.add.entry '("je" nil (((j E) 0)) ))
202 (lex.add.entry '("joj" nil (((j 0 i_) 0)) )) ; homograph(?) -> POS needed for resolving
203 (lex.add.entry '("ju" nil (((j u) 0)) ))
204 (lex.add.entry '("me" nil (((m E) 0)) ))
205 (lex.add.entry '("mi" nil (((m i) 0)) ))
206 (lex.add.entry '("mu" nil (((m u) 0)) ))
207 (lex.add.entry '("nam" nil (((n a m) 0)) ))
208 (lex.add.entry '("nas" nil (((n a s) 0)) ))
209 ;(lex.add.entry '("nj" nil (((J) 0)) )) ; homograph -> POS needed for resolving
210 (lex.add.entry '("nju" nil (((J u) 0)) ))
211 (lex.add.entry '("te" nil (((t E) 0)) ))
212 (lex.add.entry '("vam" nil (((P a m) 0)) ))
213 (lex.add.entry '("vas" nil (((P a s) 0)) ))
214
215 ;; b. reflexive pronoun (short form)
216 (lex.add.entry '("se" nil (((s E) 0)) ))
217
218 ;; c. non-accented verbforms
219 ;; - biti (= be, present tense)
220 ;(lex.add.entry '("sam" nil (((s a m) 0)) )) ; homograph! -> POS needed for resolving
221 (lex.add.entry '("si" nil (((s i) 0)) ))
222 ;(lex.add.entry '("je" nil ( ) )) ; (homograph)
223 (lex.add.entry '("sma" nil (((s m a) 0)) ))
224 (lex.add.entry '("ste" nil (((s t E) 0)) ))
225 (lex.add.entry '("su" nil (((s u) 0)) ))
226 ;; - htjeti (= want, present tense)
227 (lex.add.entry '("ću" nil (((tS_j u) 0)) ))
228 (lex.add.entry '("češ" nil (((tS_j E S) 0)) ))
229 (lex.add.entry '("če" nil (((tS_j E) 0)) ))
230 (lex.add.entry '("čemo" nil (((tS_j E) 0) ((m 0) 0)) ))
231 (lex.add.entry '("čete" nil (((tS_j E) 0) ((t E) 0)) ))
232 ;; - biti (= be, "aorist" tense)
233 (lex.add.entry '("bih" nil (((b i x) 0)) )) ; homograph! -> POS needed for resolving
234 (lex.add.entry '("bi" nil (((b i) 0)) ))
235 (lex.add.entry '("bismo" nil (((b i) 0) ((s m 0) 0)) ))
236 (lex.add.entry '("biste" nil (((b i) 0) ((s t E) 0)) ))
237
238 ;; d. interrogative conjunction / particle
239 (lex.add.entry '("li" nil (((l i) 0)) ))
240

```

lexicon_croatian.scm

Anhang C

Letter-to-Sound Rules

```
----- letter_to_sound_croatian.scm -----
1 ;; -----|
2 ;; -----|
3 ;; Croatian letter-to-sound rules (incomplete) |
4 ;; (C) 2005 Daniel Duran - all rights reserved |
5 ;; (ISO 8859-2 encoded) |
6 ;; |
7 ;; |
8 ;; "Latin2" := {ISO 8859-2 | Windows CPs^250} |
9 (lts.ruleset |
10 ;; Name of rule set: |
11   croatian_lit_dd |
12 ( |
13 ;; Sets used in the rules: |
14 (V a e i o u) ; all vowel symbols |
15 (C b c d f g h j k l m n p r s t v z) |
16 (C1 b c č f g h k m p s š t v z ž) ; for syllabic /r/ in {C1+d}rC2 context |
17 (C2 b c č ć d f g h j k l m n p s š t z ž) ; - r v |
18 (S s š) ; for /t/ deletion in "FR t C2" context |
19 (Z z ž); |
20 (VOX a E i O u b d g j l m n v w y z) ; voiced |
21 (CVOX b d đ g z ž) ; voiced consonants |
22 (NVOX c č ć f h k p s š t) ; voiceless |
23 (PB p b) |
24 (KG k g) |
25 (VF v f) |
26 ) |
27 ( |
28 ;; Rules .....|
29 ;; latin orthography |
30 ( # [ c ] r o = k ) ; "CroNet" (Croatia -> lexicon) |
31 ;; .....|
32 ;; orthographic exceptions: |
33 ;; assimilation |
34 ( [ s t ] d = z ) ; "postdiplomski" |
35 ;; consonant neutralisation
```

```

36 ( [ d č ] = tS )
37 ( [ d ć ] = tS_j )
38 ( [ d c ] = ts )
39 ( [ d š ] t i n V # = tS )
40 ( [ d s ] k V # = ts ) ; "...d"+"-ski" "gradski, gradsko"
41 ( [ d s ] t v V # = ts ) ; "...d"+"-stvo" "gradski, gradsko"
42 ( # [ d n ] V = d_n ) ; "dnevnik, dno"
43 ( # [ d l ] V = d_l ) ; "dlake, dlan"
44 ( [ t č ] = tS )
45 ( [ t c ] = ts ) ; "otca"
46 ( [ t s ] k V C * V * # = ts ) ; "bratski"
47 ( [ t s ] t v V C * V * # = ts ) ; "hrvatstvo"
48 ;( # [ t n ] V = t_n ) ; ?
49 ( # [ t l ] V = t_l ) ; "tlo, tlak"
50 ( [ s t ] n V C * # = s ) ; "azbestni, protestni"
51 ( [ č s ] k i # = tS )
52 ( [ ć s ] k i # = tS_j )
53 ;; de-voicing
54 ( [ b ] NVOX = p )
55 ( [ d ] NVOX = t )
56 ( [ d ž ] NVOX = tS )
57 ( [ đ ] NVOX = tS_j )
58 ( [ g ] NVOX = k ) ; "gangster"
59 ( [ z ] NVOX = s )
60 ( [ ž ] NVOX = S )
61 ;; voicing
62 ( [ č ] CVOX = dZ )
63 ( [ ć ] CVOX = dZ_j )
64 ( [ c ] CVOX = dz )
65 ( [ f ] CVOX = v )
66 ( [ h ] CVOX = G )
67 ( [ k ] CVOX = g )
68 ( [ p ] CVOX = b )
69 ( [ š ] CVOX = Z )
70 ( [ s ] CVOX = z )
71 ( [ t ] CVOX = d )
72 ;; labialization
73 ( [ n ] PB = m )
74 ;; "long" Consonats
75 ( # p o [ d č ] = tS ) ; "pod-"+"č..."
76 ( # n a [ d c ] = ts ) ; "nad-"+"c..."
77 ;; place of articulation
78 ( # [ v ] u = w ) ; "vuk"
79 ( # [ v ] o = w )
80 ;( [ m ] VF = F ) ; tramvaj
81 ( [ n ] KG = N )
82 ( [ š ] ć = S_j )
83 ( [ ž ] đ = Z_j )
84 ;( # [ h ] C = h )
85 ;; .....|
86 ;; ambiguities
87 ;; syllabic /r/
88 ( # [ r ] C = r_ = ) ; "rzati"
89 ( C1 [ r ] C = r_ = ) ; "vrt"

```

```

90 ( d [ r ] C = r_ = ) ; "drvo"
91 ( C [ r ] # = r_ = ) ; only in foreign words "masakr, žanr"
92 ;; graphemes n+j --> /n j/
93 ( # i [ n j ] V = n j ) ; in foreign words: "injekcija, injunktiv"
94 ( # k o [ n j ] V C = n j ) ; "konjunktura, konjugacija"
95 ( # i z v a [ n j ] V C = n j ) ; "izvan-"+"j..." "izvanjezični"
96 ;; graphemes l+j --> /lj/ --> lexicon
97 ;; graphemes d+ž --> /dž/
98 ( # n a [ d ž ] V C = d Z ) ; "nad-"+"ž..." "nadživjeti"
99 ( # p o [ d ž ] V C = d Z ) ; "podžupan"
100 ( # p r e [ d ž ] V C = d Z ) ; "predživot"
101 ;; syllabic allophones
102 ( C [ l ] # = l_ = ) ; "bicikl, ansambl"
103 ( CVOX [ n ] # = n_ = ) ; "njutn, krafn"
104 ( NVOX [ n ] # = n_ = ) ; ...
105 ; not after l,r...
106 ;; grapheme j -> [j] or []
107 ( # [ j ] u = j ) ; "jug"
108 ( # [ j ] i = j )
109 ( [ i j ] a = i ) ; "rakija, dijaliza"
110 ;( [ i j ] e = i ) ; "orijent" ----- IE ??
111 ;( [ i j ] e # = i ) ; "reakcije"
112 ( [ i j ] i # = i ) ; "studiji"
113 ( [ u j ] i # = u ) ; "zuji"
114 ( [ i j ] u C * # = i ) ; "radijus"
115 ;; "ije" --> two syllables
116 ( # n [ i j e ] d = i j E ) ; "nijedno-"
117 ( C [ i j e ] # = i j E ) ; "prije, nije, županije, rakije"
118 ( C [ i j e ] h # = i j E ) ; arhaic forms
119 ( C [ i j e ] m V * # = i j E )
120 ;; non-syllabic u
121 ;( e [ u ] r o = u_ ^ ) ; "europa, neurologija"
122 ( # e [ u ] = u_ ^ ) ; "euforija, euharistija" -- put isolated "EU" in lexicon!
123 ( # a [ u ] = u_ ^ )
124 ;; .....|
125 ;; non Croatian graphemes (in context):
126 ;; q
127 ( V [ q u e ] # = k ) ; "Dominique, unique"
128 ( [ q ] u # = k ) ; "Iraqu"
129 ( [ q u ] = k P ) ; k v\
130 ( # [ q ] V = k ) ; "qaida"
131 ( [ q ] # = k )
132 ( [ q ] C = k )
133 ;; y
134 ( # [ y ] C = i )
135 ( C [ y ] # = i )
136 ( C [ y ] C = i )
137 ;; cz
138 ( [ c z ] = tS ) ; "cz" does never occur in croatian words, exeption: foreign names
139 ;; .....|
140 ;; standard Rules .....|
141 ;; Vowels
142 ( [ i j e ] = iE ) ; ???
143 ;; accented vowels --> word accent (in Croatian words)

```

```

144 ;( [ â ] = a: ) ; a:
145 ;( [ î ] = i: ) ; i:
146 ;( [ ô ] = O: ) ; O:
147 ;( [ â ] = a: ) ; non-Croatian graphemes
148 ;( [ é ] = E: ) ; ...
149 ;( [ í ] = i: ) ; ...
150 ;( [ ó ] = O: ) ; ...
151 ;( [ ú ] = u: ) ; ...
152 ( [ ä ] = E ) ; non-Croatian graphemes
153 ( [ ö ] = E ) ; ...
154 ( [ ü ] = i ) ; ...
155 ( [ a ] = a )
156 ( [ e ] = E ) ; SAMPA: e
157 ( [ i ] = i )
158 ( [ o ] = O ) ; SAMPA: o
159 ( [ u ] = u )
160 ;; Consonants
161 ( [ b ] = b )
162 ( [ č ] = tS )
163 ( [ ć ] = tS_j )
164 ( [ d ž ] = dZ )
165 ( [ đ ] = dZ_j )
166 ( [ d ] = d )
167 ( [ f ] = f )
168 ( [ g ] = g )
169 ( [ h ] = x )
170 ( [ k ] = k )
171 ( [ l j ] = L )
172 ( [ l ] = l )
173 ( [ m ] = m )
174 ( [ n j ] = J )
175 ( [ n ] = n )
176 ( [ j ] = i_~ )
177 ( [ p ] = p )
178 ( [ r ] = r )
179 ( [ š ] = S )
180 ( [ t ] = t )
181 ( [ v ] = P ) ; SAMPA: v\
182 ( [ ž ] = Z )
183 ( [ y ] = j ) ; non-Croatian grapheme
184 ( [ w ] = P ) ; non-Croatian grapheme
185 ;( [ ś ] = S_j ) ; non-standard grapheme
186 ;( [ ź ] = Z_j ) ; non-standard grapheme
187 ( [ s ] = s )
188 ( [ z ] = z )
189 ( [ c ] = ts )
190 ( [ x ] = k s ) ; non-Croatian grapheme
191 ( [ q ] = k P ) ; non-Croatian grapheme
192 ))

```

193

letter_to_sound_croatian.scn

Anhang D

Testkorpus

Für diese Arbeit habe ich mir einen eigenen kleinen Testkorpus erstellt, da mir kein Korpus des Kroatischen zur vollen Verfügung stand. Unter <http://www.hnk.ffzg.hr> sind Informationen zum kroatischen Nationalkorpus „Hrvatski nacionalni korpus“ (HNK) zu finden. Außerdem wird eine eingeschränkte Abfrage über das Internet angeboten. Zur Zeit, als ich häufiger Informationen aus dem Korpus benötigte war das Angebot aber immer wieder über einige Tage hinweg nicht verfügbar. Dies war (auch) ein Grund dafür einen eigenen Korpus für die Zwecke dieser Arbeit zu erstellen. Dies geschah (vor allem was die Programmierung der benötigten Werkzeuge angeht) in Zusammenarbeit mit MATEUSZ WIĄCEK, der eine Arbeit zum Polnischen verfaßt hat und zu diesem Zweck auch eine große Menge an Text benötigte.

Der (namenlose) Testkorpus besteht aus reinen Textdaten (ein Rohdatenkorpus). Er umfaßt 29 Dateien zu je 10 MB, insgesamt also ein Datenvolumen von 2.998.171.201 Bytes an Text. Das daraus erzeugte Lexikon enthält ungefähr 2.500.000 Wortformen in mehr als 40.000.000 Tokens.

Die Zusammensetzung des Testkorpuses ist in Tabelle D.1 zu sehen, ein Vergleich mit dem HNK zeigt daran anschließend Tabelle D.2. Dort sind die 100 Häufigsten Wortformen des Testkorpus und des HNK dargestellt (Quelle: http://www.hnk.ffzg.hr/razfrek_e.htm).

Quelle	Bytes	%	
<i>Umgangssprachliche Texte:</i>			
blog.hr	105.662.627	3,5242%	bis April 2005
newsgroups	155.779.991	5,1958%	Nov 2004 bis Mär 2005
<i>Umgangssprachliches gesamt</i>	261.442.618	8,7201%	
<i>Juristische Texte:</i>			
Ministerien	25.748.437	0,8588%	Ministarstvo financija, Ministarstvo znanosti, obrazovanja i športa, Ministarstvo europskih integracija (Feb, Mär 2005)
Narodne Novine	93.468.870	3,1175%	16.01.1990 bis 17.03.2005

Fortsetzung auf der nächsten Seite

<i>... Fortsetzung</i>			
Quelle	Bytes	%	
Parteiseiten	9.190.577	0,3065%	DC, HDZ, Hrvatska Narodna Stranka, Hrvatska Pučka Stranka, HSLS, HSS, LS, SDP, (alle Mär 2005)
predsjednik.hr	9.125.174	0,3043%	Mär 2005
sabor.hr	128.491.139	4,2857%	Feb 2005
<i>Juristische Texte gesamt</i>	266.024.197	8,8729%	
<i>Publizistik:</i>			
Städteseiten	43.567.837	1,4531%	Crikvenica, Đakovo, Gospić, Karlovac, Koprivnica, Ogulin, Osijek, Sitno, Slavonski Brod, Varaždin, Velika Gorica, Vinodol, Vukovar, Zagreb, (Apr, Mai 2005)
Dom i Svijet	17.780.144	0,5930%	Broj 220 (1998) bis 440 (16.02.2004)
Glas Podravine	13.250.598	0,4420%	Broj 37 (2000) bis 29 (25.07.2003.)
Gloria	13.868.945	0,4626%	Broj 249 (10. 1999) bis 539 (05. 2005)
HRT	49.263.426	1,6431%	HTV Sport bis Apr 2005 HRT Vijesti 01. 04. 1998 bis 17. 03. 2005
HrvatskoSlovo	27.378.940	0,9132%	Broj 216 (11.06.1999) bis 433 (08.08.2003)
Matura-hr.com	2.777.588	0,0926%	2002/03 und 2003/04 (Namen!)
Monitor	3.322.436	0,1108%	Mai 2005
Motostil	476.899	0,0142%	Mai 2005
Novi List	758.753.478	25,3072%	2.1.2002 bis 25.2.2005
RTL televizija	3.417.234	0,1140%	Mai 2005
Slobodna Dalmacija	703.209.675	23,4546%	03.06.1999 bis 12.05.2005
Večernji List	55.420.665	1,8485%	Jan bis Apr 2005 (teilweise)
Vjesnik	561.851.505	18,7398%	28.10.1998 bis 07.04.2005
Zarez	45.288.242	1,5105%	14.09.2000 bis 05.05.2005
Andere	21.680.236	0,7231%	Firmenseiten Journalistisches, Portale Musik
<i>Publizistik gesamt</i>	2.321.307.848	77,2441%	
<i>Wissenschaftstexte:</i>			
Wikipedia HR	14.174.554	0,4728%	stand 06.05.2005
Andere	30.204.810	1,0074%	Aufsätze, Arbeiten, Biographien, Handbücher, Literaturkritik, Referate, Schultexte

Fortsetzung auf der nächsten Seite

<i>... Fortsetzung</i>			
Quelle	Bytes	%	
<i>Wissenschaftstexte gesamt</i>	44.379.364	1,4802%	
<i>Fiktion:</i>			
Lyrik	7.469.223	0,2491%	kroat. Autoren
Prosa	17.688.216	0,5900%	kroat. Autoren und Übersetzungen
Liedtexte	2.681.629	0,0894%	HipHop, Pop, Schlager, Volkslieder
<i>Fiktion gesamt</i>	27.839.068	0,9285%	
<i>Religiöse Texte:</i>			
Biblija	4.271.681	0,1425%	
Glas Koncila	59.744.310	1,9927%	2002 bis 2005
Andere	13.911.972	0,4640	Katekizam, Papstbriefe, Publikationen
<i>Religiöse Texte gesamt</i>	77.927.963	2,5992%	

Tabelle D.1: Die Zusammensetzung des Testkorpus

#	HNK		DD	
1.	i	287853	i	12305564
2.	u	252533	je	10577859
3.	je	241718	u	10496853
4.	se	143249	se	5804783
5.	da	118606	na	5572219
6.	na	113855	da	4790410
7.	su	78945	za	4219477
8.	za	75651	su	3208619
9.	s	51981	a	3005294
10.	a	51365	od	2577501
11.	od	50944	o	2125180
12.	ne	46232	s	2107693
13.	to	41924	će	1849536
14.	koji	40958	ne	1829732
15.	o	40860	koji	1668418
16.	što	40859	to	1434668
17.	kao	33338	iz	1416521
18.	iz	30621	U	1358473
19.	će	29584	što	1324124
20.	bi	28607	bi	1289396
21.	sam	26298	nije	1108664
22.	nije	26128	d	1014612
23.	Kako	24925	te	1009961
24.	te	23734	do	983596

Fortsetzung auf der nächsten Seite

Tabelle D.2 – Fortsetzung				
#	HNK		DD	
25.	ili	21255	kao	960717
26.	ali	20019	kako	949945
27.	do	19133	ili	875890
28.	koje	18542	sam	825410
29.	sve	18158	koja	774313
30.	samo	17109	ali	752497
31.	koja	15752	vjesnik	735478
32.	jer	14914	koje	729585
33.	po	14513	Hrvatska	727237
34.	tako	14454	godine	722956
35.	biti	14190	dana	694065
36.	više	13925	sve	674964
37.	bio	13812	hr	668257
38.	još	13746	po	663377
39.	pa	13254	I	659050
40.	godine	13168	samo	656514
41.	već	12968	biti	650667
42.	bilo	12888	sa	630313
43.	kad	12769	jer	590433
44.	može	12673	još	560402
45.	Mi	12265	pa	544856
46.	smo	12139	više	534679
47.	sa	11859	Na	518749
48.	Prema	11674	on	503770
49.	ni	11460	mi	502547
50.	Hrvatske	11271	li	495768
51.	nakon	11268	smo	495411
52.	zbog	10993	već	482314
53.	li	10869	bilo	472311
54.	ga	10775	Zagreb	456431
55.	No	9886	nakon	451224
56.	Ako	9781	bio	448064
57.	nego	9586	prava	438888
58.	on	9253	zbog	436650
59.	uz	9143	ni	423384
60.	toga	8987	Hrvatske	408719
61.	prije	8437	kronika	408271
62.	bez	8411	tako	401540
63.	ima	8407	može	400746
64.	bila	8053	Vjesnik	390754
65.	svoje	7938	HDZ	388053
66.	Hrvatskoj	7767	prema	385710
67.	ih	7601	ima	382512
68.	dana	7593	ga	379960
69.	godina	7486	bez	372185
70.	tome	7466	Novi	366876

Fortsetzung auf der nächsten Seite

Tabelle D.2 – Fortsetzung				
#	HNK		DD	
71.	taj	7413	kuna	356716
72.	mu	7383	godina	354420
73.	Nisu	7305	kad	351221
74.	treba	7106	nego	347320
75.	vrlo	7019	uz	346641
76.	jedan	6902	Sport	345202
77.	oko	6897	ako	336836
78.	bih	6760	oko	325852
79.	dok	6747	prije	325342
80.	mogu	6727	nisu	322262
81.	vrijeme	6717	ih	313983
82.	Ja	6648	Crna	304812
83.	ljudi	6629	nam	299945
84.	kada	6485	Kultura	299070
85.	oni	6374	komentari	293686
86.	danas	6369	Kolumne	293280
87.	sada	6235	Sva	290201
88.	bili	6010	line	287093
89.	nas	5995	kada	286642
90.	tu	5952	pridržana	281841
91.	nema	5802	arhiva	281110
92.	između	5725	bila	280491
93.	ona	5565	posto	278978
94.	kod	5542	ja	275009
95.	ipak	5530	dok	274798
96.	gdje	5527	toga	273961
97.	kojima	5518	protiv	272945
98.	me	5482	jedan	272513
99.	način	5463	treba	272231
100.	Hrvatska	5447	svih	272018

Tabelle D.2: Die 100 häufigsten kroatischen Wortformen

Abkürzungsverzeichnis

Die folgenden Abkürzungen und Symbole sind im Text dieser Arbeit zu finden. Teilweise werden die Abkürzungen auch kombiniert, so steht beispielsweise „NAV“ für „Nominativ, Akkusativ, Vokativ“ oder „mn“ für „maskulin, neutrum“.

A	Akkusativ
D	Dativ
DD	Anmerkung des Autors, Daniel Duran
Freq.	Frequenz, Häufigkeit
G	Genitiv
Hex	Hexadezimal
IPA	International Phonetic Alphabet, International Phonetic Association
Kap.	Kapitel
L	Lokativ
N	Nominativ
Part.	Partizip
Präp.	Präposition
S.	Seite
SAMPA	Speech Assessment Methods Phonetic Alphabet
V	Vokativ, der Anredefall
bzw.	beziehungsweise
dem.	Deminutiv
dt.	deutsch, deutsche Übersetzung
et. al.	et alii, und andere
f	feminin
ipf.	imperfektiv, unvollendeter Verbalaspekt
m	maskulin
n	neutrum
pass.	passiv
pf.	perfektiv, vollendeter Verbalaspekt
pl	Plural, Mehrzahl
sg	Singular, Einzahl
usw.	und so weiter
vgl.	vergleiche ...!
z.B.	zum Beispiel

Abkürzungen (Fortsetzung)

∅	Nullphonem, leeres oder gelöscht Element
/.../	phonologische Transkription
[...]	phonetische Transkription
<	ist entstanden aus
>	wird zu

Literaturverzeichnis

- [Anić 2000] Anić, Vladimir: „Rječnik hrvatskoga jezika“. Novi Liber, Zagreb 2000.
(dt. *Wörterbuch der kroatischen Sprache*, DD)
- [Anić & Goldstein 2002] Anić, Vladimir & Goldstein, Ivo: „Rječnik stranih riječi“. Novi Liber, Zagreb 2002.
(dt. *Fremdwörterbuch*, DD)
- [Babić et. al. 1996] Babić, Stjepan; Finka, Božidar & Moguš, Milan: „Hrvatski pravopis“. Školska Knjiga, Zagreb 1996.
(dt. *Kroatische Rechtschreibung*, DD)
- [Bakran 1996a] Bakran, Juraj: „Zvučna slika hrvatskoga govora“. Ibis grafika, Zagreb 1996.
(dt. *Das akustische Bild der kroatischen (gesprochenen) Sprache*, DD)
- [Bakran 1996b] Bakran, Juraj & Horga, Damir: „SAMPA za Hrvatski“. In *Govor — Časopis za fonetiku*, 13 Nr. 1-2, Hrvatsko Filološko Društvo, Zagreb 1996.
(dt. *SAMPA für Kroatisch*, DD)
- [Bakran 1998] Bakran, Juraj & Lazić, Nikolaj: „Fonetski problemi difonske sinteze Hrvatskoga govora“. In *Govor — Časopis za fonetiku*, 15 Nr. 2, Hrvatsko Filološko Društvo, Zagreb 1998.
(dt. *Phonetische Probleme bei der Diphonsynthese der kroatischen (gesprochenen) Sprache*, DD)
- [Barić et. al. 1995] Barić, Eugenija; Lončarić, Mijo; Malić, Dragica; Pavešić, Slavko; Peti, Mirko; Zečević, Vesna & Znika, Marija: „Hrvatska Gramatika“. Školska Knjiga, Zagreb 1995.
(dt. *Kroatische Grammatik*, DD)
- [Black et. al. 1999] Black, Alan W.; Taylor, Paul & Caley, Richard: „The Festival Speech Synthesis System“. Edition 1.4. Dokumentation zu *Festival Version 1.4.0*. Internetveröffentlichung, 1999.
Internet: <http://www.cstr.ed.ac.uk/projects/festival/manual/>
- [Black & Lenzo 2003] Black, Alan W. & Lenzo, Kevin A.: „Building Synthetic Voices“. Dokumentation zu *FestVox 2.0 Edition, 2nd January 2003*. Internetveröffentlichung, 2003.
Internet: <http://festvox.org/bsv/>
(Titel früherer Versionen: *Building Voices in the Festival Speech Synthesis System*)

- [Brozović 1991] Brozović, Dalibor: „Fonologija hrvatskoga književnog jezika“. In Stjepan Babić et. al. *Povjesni pregled, glasovi i oblici hrvatskoga književnog jezika: nacrti za gramatiku*, Hrvatska akademija znanosti i umjetnosti, Zagreb 1991.
(dt. *Die Phonologie der kroatischen Literatursprache*, DD)
- [Delić & Perčinić] Perčinić, Mario & Delić, Vlado: „Razvoj sintetizatora govora za hrvatsko i srpsko govorno područje“. Savez slijepih Hrvatske, Zagreb & Fakultet tehničkih nauka, Novi Sad (2002?).
Internet: <http://www.ftn.ns.ac.yu/dogs/radovi/TTS1.pdf> und
<http://www.savez-slijepih.hr/hr/strucniradovi/referati/sintetizator/>
(dt. *Entwicklung von Sprachsynthesystemen für das kroatische und serbische Sprachgebiet*, DD)
- [Garde 1993] Garde, Paul. (Übersetzer: Dragutin Raguž): „Naglasak“. Školska Knjiga, Zagreb 1993. Original: *L'accent*, Press universitaires de France, Paris 1968.
- [Godjevac 2000] Godjevac, Svetlana: „Intonation, word order and focus projection in Serbo-Croatian“. Dissertation, The Ohio State University; Unveröffentlicht (?) 2000.
- [Godjevac 2001] Godjevac, Svetlana: „Serbo-Croatian ToBI (SC_ToBI)“. Unveröffentlicht (?) 2001.
Internet: <http://www.ling.ohio-state.edu/~tobi/>
(„Serbo-Croatian“ zur Zeit der Fertigstellung dieser Arbeit nicht mehr verfügbar)
- [IPA Homepage] The International Phonetic Association.
Internet: <http://www.arts.gla.ac.uk/IPA/>
- [IPA 1999] Landau, Ernestina; Lončarić, Mijo; Horga, Damir & Škarić, Ivo: „Croatian“. In *Handbook of the International Phonetic Association*, Cambridge University Press, 1999.
- [Jelaska 2004] Jelaska, Zrinka: „Fonološki opisi hrvatskoga jezika: Glasovi, slogovi, naglasci“. Hrvatska sveučilišna naklada, Zagreb 2004.
(dt. *Phonologische Beschreibungen der kroatischen Sprache: Laute, Silben, Akzente*, DD)
- [Knežević 1970] Knežević, Anton: „Homophone und Homogramme in der Schriftsprache der Kroaten und Serben“. Verlag Anton Hain, Meisenheim am Glan 1970.
- [Lehiste & Ivić 1986] Lehiste, Ilse & Ivić, Pavle: „Word and Sentence Prosody in Serbocroatian“. The MIT Press, Cambridge, London 1986.
- [Mahnken 1964] Mahnken, Irmgard: „Studien zur serbokroatischen Satzmelodie“. In *Opera Slavica*, Band III. Vandenhoeck & Ruprecht, Göttingen 1964.
- [Matešić 1970] Matešić, Josip: „Der Wortakzent in der serbokroatischen Schriftsprache“. Carl Winter Universitätsverlag, Heidelberg 1970.
- [Möbius 2001] Möbius, Bernd: „German and Multilingual Speech Synthesis“. Habilitationsschrift, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Universität Stuttgart), AIMS 7 (4), Stuttgart 2001.
- [Möbius 2002] Möbius, Bernd: „Sprachsynthese I“. Hauptseminar am Institut für Maschinelle Sprachverarbeitung, Stuttgart 2002/2003.
<http://www.ims.uni-stuttgart.de/lehre/teaching/2002-WS/Synthese1/>

- [Moguš 1991] Moguš, Milan: „Povjesni Pregled Hrvatskoga Književnog Jezika“. In Stjepan Babić et. al. *Povjesni pregled, glasovi i oblici hrvatskoga književnog jezika: nacrti za gramatiku*, Hrvatska akademija znanosti i umjetnosti, Zagreb 1991.
(dt. *Geschichtliche Übersicht der kroatischen Literatursprache*, DD)
- [Pollok 1964] Pollok, Karl-Heinz: „Der neuštokavische Akzent und die Struktur der Melodiegestalt der Rede“. In *Opera Slavica*, Band III. Vandenhoeck & Ruprecht, Göttingen 1964.
- [Raguž 1997] Raguž, Dragutin: „Praktična hrvatska gramatika“. Medicinska naklada, Zagreb 1997.
(dt. *Praktische kroatische Grammatik*, DD)
- [Rehder 1968] Rehder, Peter: „Beiträge zur Erforschung der serbokroatischen Prosodie“. Verlag Otto Sagner, München 1968.
- [SAMPA Homepage] SAMPA — computer readable phonetic alphabet.
Internet: <http://www.phon.ucl.ac.uk/home/sampa>
- [Škarić 1991] Škarić, Ivo: „Fonetika hrvatskoga književnog jezika“. In Stjepan Babić et. al. *Povjesni pregled, glasovi i oblici hrvatskoga književnog jezika: nacrti za gramatiku*, Hrvatska akademija znanosti i umjetnosti, Zagreb 1991.
(dt. *Phonetik der kroatischen Literatursprache*, DD)
- [Škarić et. al. 1996] Škarić, Ivo; Škavić, Đurđa & Varošaneć-Škarić, Gordana: „Kako se naglašavaju posuđenice“. In *Jezik — časopis za kulturu hrvatskoga književnog jezika*, 43 Nr. 4, Hrvatsko Filološko Društvo, Zagreb 1996.
(dt. *Wie Lehnwörter betont werden*, DD)
- [Škarić 2001a] Škarić, Ivo: „Razlikovna prozodija“. In *Jezik — časopis za kulturu hrvatskoga književnog jezika*, 48 Nr. 1, Hrvatsko Filološko Društvo, Zagreb 2001.
(dt. *Distinktive Prosodie*, DD)
- [Škarić 2001b] Škarić, Ivo: „Kakav pravopis (između fonetike i fonologije)“. In *Govor — Časopis za fonetiku*, 18 Nr. 1, Hrvatsko Filološko Društvo, Zagreb 2001.
(dt. *Was für eine Rechtschreibung (zwischen Phonetik und Phonologie)*, DD)
Internet: http://www.ffzg.hr/fonet/skaric/skaric-kakav_pravopis.pdf (mit einer Zusammenfassung in englischer Sprache)
- [Škarić & Lazić 2002] Škarić, Ivo & Lazić, Nikolaj: „Vrijednosni sudovi o hrvatskim naglascima“. In *Govor — Časopis za fonetiku*, 19 Nr. 1, Hrvatsko Filološko Društvo, Zagreb 2002.
(dt. *Beurteilungen der kroatischen Betonungen*, DD)
Internet: <http://www.ffzg.hr/fonet/govor/pdf/2002/govor-2002-19-1/en/skaric-lazic.pdf> (Zusammenfassung in englischer Sprache)
- [Unicode Standard 2005] Unicode Consortium „The Unicode Standard“. 2005.
Internet: <http://www.unicode.org/standard/standard.html>

Pronunciation Rules for Croatian Text-to-Speech Synthesis

Abstract

This student research paper describes the development of pronunciations rules for a Croatian text-to-speech system. These rules were developed within the framework of the *Festival Speech Synthesis System*.

The first chapter gives a brief overview on speech synthesis as a field of research in computational linguistics and some notes on practical applications. Different concepts of speech synthesis are described followed by a short overview of text-to-speech synthesis (TTS). This chapter discusses also the term *Croatian language* and gives a definition for the purpose of this paper.

The second chapter focuses on the written forms of the Croatian language. Special attention is given to those aspects of the written language that are of importance for the text-to-speech synthesis. The Croatian orthography is first described in general followed by some specific details and features of Croatian texts. Problems of higher relevance for TTS systems are described in depth — such as the special characters used in Croatian orthography, the disambiguation of homographs and the problems that (might) occur in the analysis process of digits, acronyms or abbreviation within written texts. This chapter gives also an overview of different systems of phonetic and phonologic transcriptions that are used in Croatia and Croatian linguistics. The IPA system is described as well as SAMPA and the Croatian method for phonological transcriptions which is typically used in slavistics and Croatian literature.

The third chapter describes the grammar of the Croatian language. The grammar is not represented in full detail but only those aspects are described which are of special importance to TTS systems. These aspects are phonetics, phonology and morphology. First the phoneme and phone inventory of the Croatian is described and different descriptions from the literature are represented. After that an overview of the lexical accent (word accent) is given. This is also described by different examples from the linguistic literature.

The fourth chapter provides the actual implementation of the pronunciation rules using the Festival Speech Synthesis System. After an short introduction to the Festival system, the developed rules are described in detail. Within the Festival formalism three modules were developed: (1) a phoneset providing the Croatian phone inventory, (2) general letter-to-sound rules (lts) used to transcribe unknown words and (3) a small hand written lexicon (addenda). These rules are then tested and a short analysis concludes this chapter.

Under the number five a short conclusion is given along with some open questions and perspectives for future work and development on Croatian speech synthesis.

The appendix contains the Festival Scheme-scripts with the developed rules as well as a short description of a 40 000 000 token *test corpus* of raw text data. This corpus was compiled to aid research within this work and to provide an additional source of information about the contemporary written Croatian language.