

UNIVERSITÄT STUTTGART
INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG
AZENBERGSTRASSE 12
D 70174 STUTTGART

AUSSPRACHEREGELN FÜR DAS BULGARISCHE
IM SPRACHSYNTHESESYSTEM FESTIVAL

Stoyka Dachenska

Studienarbeit Nr.: 105

Betreuer: Antje Schweitzer
Prüfer: Prof. Dr. Grzegorz Dogil

Beginn: 28.07.2010
Ende: 07.01.2011

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.

Stuttgart, den 7. Januar 2011

Stoyka Dachenska

На майка и татко

Inhaltsverzeichnis

1. Einleitung	5
1.1. Ausgangssituation und Zielsetzung.....	5
1.2. Aufbau der Arbeit.....	6
2. Grundlagen	7
2.1. Aufgaben eines Sprachsynthesystems.....	7
2.2. Anwendungen eines Sprachsynthesystems.....	7
2.3. Architektur eines Sprachsynthesystems.....	8
2.4. Phonetische Transkription mit SAMPA.....	10
3. Eigenschaften der bulgarischen Sprache in Bezug auf Sprachsynthesysteme	11
3.1. Das Schriftsystem der bulgarischen Sprache	11
3.1.1. Das bulgarische Alphabet.....	12
3.1.2. Orthographie der bulgarischen Sprache.....	13
3.1.2.1. Groß- und Kleinschreibung.....	13
3.1.2.2. Abkürzungen und Akronyme.....	14
3.1.2.3. Doppelte Konsonanten und Vokale.....	15
3.1.2.4. Fremdwörter.....	16
3.1.2.5. Zahlen und Zahlwörter.....	16
3.1.2.6. Satzzeichen.....	17
3.2. Das Phonemsystem der bulgarischen Sprache	17
3.2.1. Graphem-zu-Phonem Beziehungen.....	18
3.2.2. Vokale.....	20
3.2.3. Konsonanten.....	22
3.3. Wortbetonung der bulgarischen Sprache	25
4. Ausspracheregeln für das Bulgarische im TTS-System Festival	26
4.1. Überblick	27
4.2. Erstellung einer Aussprachekomponente in Festival	28
4.2.1. Phonesets.....	29
4.2.2. Lexikon.....	32
4.2.2.1. Lexikoneinträge.....	32
4.2.2.2. Lexikonlookup.....	33
4.2.3. Letter-To-Sound-Regeln.....	33
5. Evaluierung	39
5.1. Methode	39
5.2. Auswertung	42
6. Zusammenfassung und Ausblick	43
Literaturverzeichnis.....	45
Abkürzungsverzeichnis.....	47
Anhang A.....	48
Anhang B.....	50
Anhang C.....	54

1. Einleitung

1.1. Ausgangssituation und Zielsetzung

Sprache ist das meist verbreitete Kommunikationsmittel. Kommunikation definieren Shanon und Weaver als die Gesamtheit aller Vorgänge, mit denen ein Lebewesen oder eine Maschine ein anderes Lebewesen oder eine Maschine beeinflusst [Shanon / Weaver, 1949]. Parallel zur Entwicklung von Computersystemen entstand der verstärkte Bedarf an Interaktion zwischen Mensch und Maschine. In den letzten Jahren entwickelten sich unterschiedliche Strategien, um den Informationsaustausch zwischen den beiden Akteuren zu modellieren. Eine Strategie ist die mündliche Kommunikation über ein Sprachdialogsystem. Eine Komponente eines Sprachdialogsystems ist, neben der Spracherkennung und der Sprachanalyse, die Sprachsynthese, d.h. die künstliche Erzeugung natürlicher Sprache.

Für einige Sprachen existieren bereits funktionsfähige und anwendbare Sprachsynthesysteme. Zum Beispiel sind im Verbmobil-Projekt¹ Fragmente des Deutschen, Englischen und Japanischen implementiert [Wahlster, 2000]. Mit der stärkeren Verbreitung von Computersystemen steigt der Bedarf auch weitere Sprachen zu synthetisieren und neue Sprachen für die vorhandene Sprachtechnologie zu entwickeln. Dies ist auch für das Bulgarische der Fall. In dieser Arbeit sollen Ausspracheregeln für das Bulgarische entwickelt und implementiert werden. Ein Vorteil des Bulgarischen besteht darin, dass die Abbildung von der Orthografie auf eine Folge von Lauten hier relativ einfach erscheint.

Ziel dieser Arbeit ist die Erstellung einer Sprachkomponente im Sprachsynthesystem Festival für das Bulgarische. Hierzu ist die Erarbeitung von Ausspracheregeln durch eine detaillierte Analyse der bulgarischen Phonologie notwendig. Darauf aufbauend soll ein vollständiger Satz von Ausspracheregeln für das Bulgarische entwickelt und für das frei verfügbare Sprachsynthesystem Festival [Black / Taylor / Caley, 2001] adaptiert werden. Zur Erstellung der Regeln sollen Beschreibungen der bulgarischen Grammatik und Phonologie herangezogen werden, zum Beispiel [BAN, 2005], [Radeva, 2003], [Stoyanov, 1999], [Tilkov, 1998].

¹ URL (20.11.2010): <http://verbmobil.dfki.de/>

Basis für die Evaluierung dieser Ausspracheregeln sind zufällig ausgewählte Sätze von unterschiedlichen Internetquellen, welche manuell phonetisch transkribiert wurden.

1.2. Aufbau der Arbeit

Die vorliegende Arbeit besteht aus sechs Kapiteln und wird wie folgt strukturiert.

Kapitel 2 befasst sich mit den theoretischen Ansätzen dieser Arbeit. Insbesondere werden die grundlegenden Begriffe der Sprachsynthese definiert, vorhandene Anwendungen vorgestellt und das Grundgerüst eines solchen Systems erläutert.

In Kapitel 3 wird ein Fragment der bulgarischen Grammatik im Bezug auf Sprachsynthesysteme analysiert. Dabei wird im Besonderen die Korrelation von Orthographie und Phonologie diskutiert. Im Detail wird hierbei eine umfassende Abbildung der Rechtschreibung auf der Aussprache erarbeitet. Diese beinhaltet eine ganzheitliche Betrachtung für das native Bulgarisch und eine partielle Analyse für Fremdwörter.

Die im Rahmen dieser Arbeit entwickelten Ausspracheregeln werden in Kapitel 4 nach phonologischen Merkmalen kategorisiert und erläutert.

Kapitel 5 befasst sich mit der Implementierung der Ausspracheregeln im Sprachsynthesystem Festival. Eine Stimme steht für diese Arbeit nicht zur Verfügung, schon allein deshalb kann keine vollständige Implementierung für die bulgarische Sprache vorgenommen werden. Abgesehen davon wäre auch die Entwicklung sprachspezifischer Module für bulgarische Prosodie und damit verbunden, die die Einbindung einer einfachen syntaktischen Analyse im Rahmen einer Studienarbeit zu aufwändig. Die Miteinbindung einer Stimme ist deshalb außerhalb des Fokus dieser Arbeit.

Kapitel 6 stellt die Ergebnisse dar und gibt einen Ausblick auf weitere Arbeiten in diesem Bereich.

2. Grundlagen

Dieses Kapitel befasst sich mit den theoretischen Aspekten von Sprachsynthesystemen im Allgemeinen. Im Folgenden werden deren Aufgaben, Anwendungsbereiche sowie die Struktur dieser Systeme dargestellt.

2.1. Aufgaben eines Sprachsynthesystems

Als Sprachsynthese wird die künstliche Erzeugung natürlicher Sprache mit Hilfe eines Computers zur Interaktion zwischen Mensch und Maschine bezeichnet. Als Eingabe bei den meisten Sprachsynthesystemen wird geschriebene Sprache verwendet. Als Ausgabe wird eine verständliche und natürlich klingende Sprache gefordert [Dutoit, 1997]. Diese Umsetzung der geschriebenen Sprache in Lautsprache steht in Analogie zu einer Person, die vorliest. Die Aufgabe der Sprachsynthese soll damit nichts Geringeres leisten als diese Person [Pfister / Kaufmann, 2008].

Bisherige Erfahrungen mit Sprachsynthesystemen zeigen, dass die Qualität der Sprachausgabe einen entscheidenden Einfluss auf die Akzeptanz der Anwendung, worin die Sprachsynthese eingebettet ist, hat. Daraus ergibt sich die Notwendigkeit, die Natürlichkeit der Stimme anzustreben [Taylor, 2007].

2.2. Anwendungen eines Sprachsynthesystems

Durch die zunehmende Computerisierung unseres Alltags wuchsen die Bedürfnisse der Menschen zur Bedienung der Maschinen mittels natürlicher Sprache. Bereits Ende der 70er Jahre erschien das erste vollständige Sprachsynthesystem [Klatt, 1987]. Sprachsynthesysteme werden auch als Text-To-Speech-Systeme (TTS) bezeichnet. In vielen Bereichen des modernen Lebens haben Sprachsynthesysteme bereits Einzug gehalten. Im Folgenden werden einige Beispiele kategorisiert nach sprachsynthese- und interaktionsfokussierten Anwendungen dargestellt:

1. Sprachsynthesefokussierte Anwendungen:

- Sprachsynthesysteme unterstützen behinderte Menschen bei der Bewältigung von Aufgaben im Alltag. Durch diese Systeme erhalten zum Beispiel seh- und

mobilitätsbehinderte Personen Zugang zu geschriebenen Informationen durch Sprachausgabe.

- Sprachsynthese kann auch beim Fremdspracherwerb behilflich sein. Sie kann das Erlernen der Aussprache von Wörtern unterstützen.

2. Interaktionsfokussierte Anwendungen:

- Sprachsynthesysteme eignen sich für die Interaktion zwischen Mensch und Maschine z.B. in Fahrzeugen, um eine Ablenkung des Fahrers vom Verkehrsgeschehen durch manuelle Bedienung zu vermeiden.
- In der Telekommunikations-Branche werden Sprachsynthesysteme zum Informationsaustausch bei Auskunft- und Reservierungssystemen zur Reduktion des Personalaufwands eingesetzt.

2.3. Architektur eines Sprachsynthesystems

In diesem Abschnitt wird die typische Architektur eines Sprachsynthesystems dargestellt. Nach Schweitzer [Schweitzer, 2008] besteht ein solches System aus folgenden Hauptkomponenten: linguistische Analyse, Prosodiesteuerung und akustische Synthese. Abbildung 2.1. stellt diese graphisch dar.

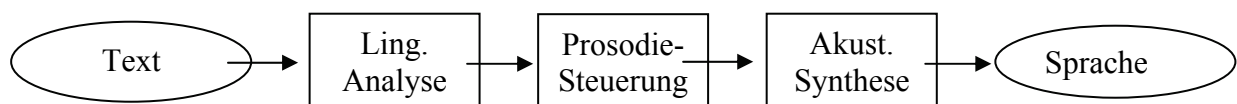


Abbildung 2.1. Hauptkomponente eines Sprachsynthesystems [Schweitzer, 2008]

Im Folgenden werden die Komponenten und Verarbeitungsschritte auf Basis der Systemarchitektur detailliert beschrieben.

1. Linguistische Analyse

Nach Schweitzer [Schweitzer, 2008] besteht die linguistische Analyse eines derzeitigen typischen Sprachsynthesystems aus Textvorverarbeitung, lexikalisch-morphologischer Analyse, syntaktischer Analyse und prosodischer Analyse.

- Textvorverarbeitung: die Textvorverarbeitung erfolgt an erster Stelle der Sprachsynthese. Als ihre Aufgabe zählt die Normalisierung des Eingabetextes. Darunter sind die folgenden Aufgaben zu verstehen: Entfernen von

Metainformation (z.B. überflüssige HTML-Tags, Formatierung, etc.), Erkennung von Satzenden, Tokenisierung, Tokentyp-Erkennung wie Abkürzungen, Währungen, Telefonnummern, Uhrzeiten, Daten, E-Mail-Adressen [Schweitzer, 2008].

- Lexikalisch-morphologische Analyse: nach der Textnormalisierung folgt die lexikalisch-morphologische Analyse. Durch die lexikalisch-morphologische Analyse werden Informationen über die Wörter gewonnen. An dieser Stelle werden Bestimmungen der Wortklassen und der Aussprache wie z.B. Phonemfolge, Silbentrennung, Betonung gemacht [Schweitzer, 2008]. Als Standardlösung dazu werden in den meisten Sprachsynthesysteme Aussprachewörterbücher verwendet.
- Syntaktische Analyse: während der syntaktischen Analyse wird die Umgebung eines Tokens² analysiert, d.h. es wird die syntaktische Struktur bestimmt [Schweitzer, 2008].
- Prosodische Analyse: nach Abschluss der syntaktischen Analyse werden die Phrasengrenzen und Akzente festgelegt.

Als Ergebnis von der linguistischen Analyse entsteht die linguistisch phonologische Struktur, die als Input für den nächsten Modul übergeben wird [Schweitzer, 2008].

2. Prosodiesteuerung

Die Prosodie wird nach der Bestimmung der Aussprache von den Wörtern generiert. Der Grad der Natürlichkeit eines Sprachsynthesystems ist abhängig von folgenden prosodischen Faktoren:

- Dauermodellierung: hierzu gehört die Bestimmung der Lautdauer für jeden einzelnen Laut, wobei Faktoren wie z.B. Lautidentität, Akzentstatus des zugehörigen Wortes, Position des Segments in der Phrase, benachbarte Segmente, Position des Segments in der Silbe, Einfluss darauf haben [Schweitzer, 2008].
- Intonationsmodellierung: dazu zählt die Bestimmung der konkreten F0-Kontur. [Schweitzer, 2008].
- Amplitudenmodellierung: hierzu gehört die Modellierung der Amplitude.

² URL (20.11.2010): <http://kitt.cl.uzh.ch/kitt/clglossar/index.php/Token>

Als Resultat der Prosodiegenerierung entstehen Phoneme, inklusiv gewünschter Dauer, Grundfrequenz und Amplitude. Diese dienen als Input für die akustische Analyse [Schweitzer, 2008].

3. Akustische Synthese

Das Modul der akustischen Synthese erzeugt aus der resultierenden linguistischen Repräsentation das digitale Sprachsignal, welches dann mittels eines Digital/Analog-Wandlers auf einem Lautsprecher ausgegeben werden kann [Breitenbücher, 1997]. Die unterschiedlichen Ansätze der Sprachsynthese werden im Rahmen dieser Arbeit nicht weiterverfolgt, da der Fokus dieser Arbeit auf der Erstellung von Ausspracheregeln für das Bulgarische liegt.

2.4. Phonetische Transkription mit SAMPA

In den folgenden Kapiteln wird ein zusammenfassender Überblick über die phonetische bzw. phonologische Transkription der bulgarischen Sprache gegeben. Es gibt zwei verschiedene Transkriptionssysteme: IPA³ und SAMPA⁴. Das internationale phonetische Alphabet (IPA) stellt den Standard in der phonetischen Transkription dar. Da SAMPA ein System zur phonetischen bzw. phonologischen Transkription bietet, das nur ASCII Zeichen verwendet und so als Standard im elektronischen Datenaustausch überall eingesetzt kann, wo aus technischen Gründen nur ASCII zur Verfügung steht, wird in dieser Arbeit das Transkriptionssystem SAMPA verwendet und im folgenden Abschnitt ausführlich beschrieben.

SAMPA (Speech Assessment Methods Phonetic Alphabet) ist ein maschinenlesbares phonetisches Alphabet. Es wurde in den Jahren 1987 bis 1989 von einer internationalen Gruppe von Phonetikern für die Sprachen der Europäischen Gemeinschaft entwickelt. Im Jahre 1996 wurde es für die bulgarische Sprache erweitert⁵.

Die SAMPA-Symbole für die Vokale des Bulgarischen entsprechen den IPA Symbolen nur teilweise. Für die Vokalphoneme [i, a, u] werden in den beiden Systemen dieselben Zeichen

³ URL (08.11.2010): <http://www.langsci.ucl.ac.uk/ipa/>

⁴ URL (08.11.2010): <http://www.phon.ucl.ac.uk/home/sampa>

⁵ URL (08.11.2010): <http://www.phon.ucl.ac.uk/home/sampa/bulgar.htm>

verwendet. Für den Rest, die Phoneme [ə, ɔ, e], existieren andere Symbole, die sich von die SAMPA Symbole unterscheiden.

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

Die Affrikate werden wie folgt dargestellt: [dz] für „дз“, [ts] für „ц“, [dZ] für „дж“ und [tS] für „ч“. Die palatalisierten Affrikaten werden als [dzʲ] und [tsʲ] wiedergegeben.

Die restlichen in IPA vom ASCII Zeichensatz abweichenden Symbole werden so dargestellt: [S] für „ш“ und [Z] für „ж“.

3. Eigenschaften der bulgarischen Sprache in Bezug auf Sprachsynthesysteme

In diesem Kapitel werden die Eigenschaften der bulgarischen Sprache, die für Sprachsynthesysteme relevant sind, dargelegt. Das Kapitel ist in zwei Unterkapitel geteilt. In dem ersten Unterkapitel erfolgt eine ausführliche Darstellung des Schriftsystems der bulgarischen Sprache. Dabei wird in Details auf die Geschichte des bulgarischen Alphabets eingegangen. Danach werden die Sprachreformen der bulgarischen Sprache erwähnt. Zunächst wird das Alphabet dargestellt, das heutzutage verwendet wird. Schließlich werden einzelne orthographische Phänomene der bulgarischen Sprache beschrieben, die eine Bedeutung für Sprachsynthesysteme haben. Das zweite Unterkapitel stellt das Phonemsystem der bulgarischen Sprache dar. Hier wird über Beziehungen zwischen Graphemen und Phonemen in der bulgarischen Sprache diskutiert. Ausschließend werden die einzelnen Komponenten des Phonemsystems der bulgarischen Sprache, die Vokale und die Konsonanten, detailliert beschrieben.

3.1. Das Schriftsystem der bulgarischen Sprache

Die bulgarische Sprache gehört zu der südslawischen Sprachgruppe der indogermanischen Familie. Innerhalb dieser Sprachfamilie wird die bulgarische Sprache der östlichen Gruppe

der südslawischen Sprachen zugeordnet [Radeva, 2003]. Bulgarisch ist die Amtssprache der Republik Bulgarien und wird von etwa 8 Millionen Menschen gesprochen.

3.1.1. Das bulgarische Alphabet

Das bulgarische Alphabet entstand in den Zügen der Christianisierung des bulgarischen Volkes im Jahre 863. Der Fürst Rostislav (846-870) hatte das Byzantinische Reich gebeten, Geistliche zu entsenden, die das Christentum in slawischer Sprache verbreiten. Daraufhin wurden zwei Mönche, Konstantin-Kyryll, der Philosoph, und sein Bruder Method in die bulgarischen Gebiete entsandt. Sie leiteten den ersten Prototyp des bulgarischen Alphabets, genannt „Glagolica“, aus dem Griechischen ab.

Der Nachfolger der beiden Mönche, ihr Schüler Kliment Ochridski, entwickelte die Schrift weiter, indem er griechische Großbuchstaben mit glagolitischen Zeichen verbandt. Diese Schrift wurde nach dem Slawenlehrer „Kyrillica“ benannt [Radeva, 2003].

Die erste Sprachreform wurde zu Beginn des 19. Jh. vorgenommen und damit entfielen ungefähr zehn veraltete Buchstaben und Buchstabenverbindungen. Die zweite Reform wurde nach der Befreiung von der osmanischen Fremdherrschaft durchgeführt, wobei vor allem in der Wortfolge und dem Satzbau Veränderungen vorgenommen wurden. Die dritte Reform geschah nach 1945. Dabei wurden weitere Buchstaben gestrichen [Radeva, 2003].

Das moderne bulgarische Alphabet besteht aus 30 Buchstaben. Einen Überblick gibt Abbildung 3.1. Diese Buchstaben existieren in zwei Varianten: Schreibschrift und Druckschrift. Aus graphologischer Sicht sind die Buchstaben in zwei Arten unterteilt: Groß – und Kleinbuchstaben. Ausnahme hier ist der Buchstabe „б“, welcher nur kleingeschrieben wird [BAN, 2005].

А а	Б б	В в	Г г	Д д	Е е	Ж ж	З з	И и	Й й
К к	Л л	М м	Н н	О о	П п	Р р	С с	Т т	У у
Ф ф	Х х	Ц ц	Ч ч	Ш ш	Щ щ	Ъ ъ	Ь ь	Ю ю	Я я

Abbildung 3.1. Überblick über das bulgarische Alphabet

3.1.2. Orthographie der bulgarischen Sprache

Wie bereits in Kapitel 2 beschrieben, verwendet ein Sprachsynthesystem geschriebenen Text als Input. Deshalb ist es erforderlich, die Rechtschreibung der zu verarbeitenden Sprache darzustellen. Im folgenden Abschnitt werden spezielle Fälle der Orthographie der bulgarischen Sprache aufgeführt, die im Zusammenhang mit der Sprachsynthese stehen. Zunächst werden Verwendungsregeln von Groß- und Kleinbuchstaben in der bulgarischen Sprache aufgelistet. Ausschließend werden die Abkürzungsregeln in der bulgarischen Sprache erläutert. Daraufhin wird über die Bildung von Akronymen berichtet. Wann und an welcher Stelle doppelte Konsonanten und Vokalen entstehen können, und wie sie zu behandeln sind, wird drauffolgend diskutiert. Im Abschluss wird über die Regeln bei der Übernahme von Fremdwörtern in der bulgarischen Sprache berichtet. Schließlich werden Zahlen und Zahlwörter und Punctuation detailliert diskutiert.

3.1.2.1. Groß- und Kleinschreibung

Im Gegensatz zum Deutschen werden im Bulgarischen satzinterne Nomina nicht groß geschrieben. In der bulgarischen Sprache wird die Großschreibung in unterschiedlichen Fällen benutzt. Diese können in vier Hauptkategorien unterteilt werden:

- Die erste Kategorie stellt die Großschreibung am Anfang des Satzes und zu Beginn der direkten Rede (nach dem Doppelpunkt) dar [BAN, 2005].
- Zu der zweiten Kategorie werden alle Eigennamen gezählt, wobei sich mehrere Unterkategorien finden lassen wie z.B.: Namen von Personen, Familiennamen und Spitznamen („Иван“ [Ivan], „Йорданка“ [Jordanka], „Иван Иванов“ [Ivan Ivanov], „Йорданка Йорданова“ [Jordanka Jordanova], „Пешо Черния“ [pešo tSernija]); Namen von geografischen Eigenschaften wie Länder, Städte, Landschaften, Meere, Seen, Flüsse, Berge, Straßen; In die Worte Ost, West, Nord, Süd, wenn in Bezug auf das Land, Länder, Nationen eingesetzt, wie z.B. „Дивия Запад“ [divija zapad] *dt. wilder Westen*; Namen von Himmelskörpern als astronomische Begriffe; Namen von Betrieben, Unternehmen, Institutionen, Firmen; auch historische Ereignisse werden groß geschrieben [BAN, 2005].

- Zu der dritten Kategorie gehören die Pronomen in ihrer Höflichkeitsform. Titel und Wörter zur Kennzeichnung von Höflichkeiten sind auch mit einem Großbuchstaben zu versehen [BAN, 2005].
- Alle anderen Wörter werden im Bulgarischen klein geschrieben.

3.1.2.2. Abkürzungen und Akronyme

Orthographische Abkürzungen sind die verkürzten Darstellungsformen eines Wortes oder einer Wortgruppe. Diese werden aus Vereinfachungs- und Rationalisierungsgründen verwendet. In der Regel werden oft gebrauchte und allgemein bekannte Wörter oder Wortgruppen abgekürzt. Die abgekürzten Wörter werden beim Lesen ausgesprochen. Hier scheint es nur eine Regel zu geben: abgekürzt wird nur an einem Konsonanten gefolgt von einem Vokal, und die Abkürzung wird durch einen Punkt gekennzeichnet. Im Bulgarischen existiert eine Reihe von Möglichkeiten, um eine Abkürzung darzustellen:

- Die erste Möglichkeit betrifft die Eigennamen. Sie werden mit einem Großbuchstaben geschrieben und auf die erste oder auf die darauffolgenden abgekürzt, z.B. П. К. Яворов [p@ k@ javorov], Хр. Ботев [h@ r@ botev] [Tilkov, 1998].
- Die zweite Möglichkeit betrifft Gattungsnamen. Sie werden mit Kleinbuchstaben bei der Abkürzung geschrieben. Diese können in drei Haupttypen unterteilt werden. Der erste Typ beschreibt Abkürzungen nur mit dem Anfangsbuchstabe eines Wortes, z.B. „г“ [g@] – „година“ [godina] dt. *Jahr*. Der zweite Typ stellt Abkürzungen mit Anfangs- und Endbuchstaben eines Wortes dar, und dazwischen werden sie durch einen Bindestrich oder Schreckstrich getrennt, z.B. „г-н“ [g-n] – „господин“ [gospodin] dt. *Herr*, „д-р“ [d-r] – „доктор“ [doktor] dt. *Arzt* [Tilkov, 1998].
- Zum dritten Typ gehören Abkürzungen mit gemischtem Charakter wie z.B. „срв.“ [srv.] – „сравни“ [sravni] dt. *vergleiche*, wo nur die Anfangsbuchstaben mit einigen internen Konsonanten gekennzeichnet werden. Bei Abkürzungen für physikalische Maßeinheiten wird kein Punkt gesetzt, z.B. „л“ [l] – „литър“ [lit@r] dt. *der Liter* [Vatov, 1995].

Akronyme sind Textbestandteile, die sich aus den jeweiligen Anfangsbuchstaben von Mehrwortausdrücken, für welche sie stehen, zusammensetzen. Sie werden in Großbuchstaben geschrieben. Im Bulgarischen sind ein-, zwei- oder mehrsilbige Akronyme zu beobachten.

Wenn die abgekürzten Formen nicht wie eine Silbe ausgesprochen werden, bilden sie mit Hilfe der Vokalphoneme [e] oder [i] eine Silbe und ermöglichen somit die Aussprache, siehe Abbildung 3.2. Akronyme werden beim Lesen nicht in jedem Fall aufgelöst. Es gibt zwei Vorgehensweisen beim Aussprechen der Akronyme: entweder werden die Akronyme buchstabiert, oder sie werden wie ganz normale Wörter behandelt [Tilkov, 1998].

Akronym	Bedeutung	Aussprache	Übersetzung
БАН	Българска академия на науките	[ban]	Bulgarische Akademie der Wissenschaft
ЦК	Централен комитет	[tse ka]	Zentralkomitee
БНБ	Българска народна банка	[be ne be]	Bulgarische National Bank

Abbildung 3.2. Ausgewählte Akronyme

3.1.2.3. Doppelte Konsonanten und Vokale

Um die Beschreibung des orthographischen Systems des Bulgarischen zu vervollständigen, sollten doppelten Konsonanten- und Vokalgrapheme erklärt werden. Doppelte Konsonantengrapheme erscheinen im Wort an der Morphemgrenze, z.B. wird aus der Präposition „от“ [ot] *dt. von* und dem Adverb „там“ [tam] *dt. dort* das Adverb „оттам“ [ottam] *dt. von dort* gebildet. Diese Verdoppelung des Konsonanten [t] wirkt sich artikulatorisch durch ein längeres Halten des Verschlusses aus. Doppelte Vokalgrapheme stehen tatsächlich für zwei Vokale, von denen einer die Betonung trägt. Am häufigsten ist die Verbindung: [uu] und [ee], die jeweils zwei aufeinander folgende [ii] und [ee] repräsentieren, z.B. „той пее“ [toj`pee] *Dt. er singt* [Radeva, 2003]. Die doppelten Vokalgrapheme aa [aa] und yy [uu] repräsentieren Vokale mit doppelter Länge, z.B. „вакуум“ [ˈvakuum] *Dt. Vakuum* [Radeva, 2003].

Eine Besonderheit von doppelten Vokalgraphemen oo [OO] ist, dass es in bestimmten Wörtern fremden Ursprungs wie ein einfaches [O] artikuliert wird, z.B. „кьопоолу“ [ˈk’OpOlu] *Dt. Auberginenpaste*, und in anderen Fällen, in denen die Verdoppelung durch Zusammentreffen an Morphemgrenzen entsteht, deutlich als doppeltes [OO], z.B. „гръмоотвод“ [gr@mOOtvod] *Dt. Blitzableiter* [Radeva, 2003].

3.1.2.4. Fremdwörter

Fremdwörter stellen für die Formalisierung von Ausspracheregeln eine zusätzliche Schwierigkeit dar, da sie in der bulgarischen Sprache mit ihrer ursprünglichen Schreibweise übernommen werden und dadurch von den sprachspezifischen Ausspracheregeln abweichen. In der modernen bulgarischen Sprache werden viele Fremdwörter verwendet. Die neu adaptierten Wörter sind meist Benennungen von neuen Maschinen, Begriffen und Technologien oder Eigennamen. Hierzu wird die folgende Regel festgelegt: alle Fremdwörter, die in das Bulgarische übernommen werden, behalten ihre ursprüngliche Aussprache bei und werden in das bulgarische Alphabet nach den aktuellen Rechtschreibregeln transliteriert. Diese Grundregel kann nicht immer erfüllt werden. Die Ursache dafür ist, dass die Zusammensetzung des bulgarischen phonetischen Systems und die Mittel, über die das bulgarische Alphabet verfügt, in manchen Fällen nicht ausreichend sind z.B. „Волтер“ [volter] *Fr. Voltaire*, „Париж“ [pariZ] *Fr. Paris* [Stoyanov, 1999]. Die zwei Beispiele sind auch nicht die einzigen Abweichungen von der Grundregel. Da die Aussprache dieser Wörter sich nach der Orthographie der ursprünglichen Sprache richtet, wäre es sehr verwirrend, sie in das Regelset der bulgarischen Sprache aufzunehmen. Deshalb wird in dieser Arbeit die Aussprache solcher Wörter mit Hilfe eines Aussprachelexikons bestimmt.

3.1.2.5. Zahlen und Zahlwörter

Die Zahlwörter spielen in einem Sprachsynthesystem eine wesentliche Rolle, sobald diese im Text nicht ausgeschrieben sind, sondern durch Ziffern dargestellt werden. In einem solchen Fall muss die Aussprache der Ziffernsymbole bestimmt werden, was nicht durch dieselben Regeln erfolgen kann wie bei den Buchstaben eines Textes. Der folgende Abschnitt soll diese Problematik verdeutlichen.

Arabische Ziffern stehen entweder für Kardinalzahlen (z.B. 10 Äpfel) oder für Ordnungszahlen „1-и опит“ [p@rvi opit] *Dt. erster Versuch*. Die Ordnungszahlen im Bulgarischen haben zwei unterschiedliche Schreibvarianten. z.B. „1. опит“ [p@rvi opit] *Dt. erster Versuch*, aber auch die folgende Schreibvariante „1-и опит“ [p@rvi opit] steht für den gleichen Ausdruck. Genau dasselbe ist bei „1 опит“ [edin opit] *Dt. ein Versuch* und „1-н опит“ [edin opit] zu beobachten [Vatov, 1995].

Römische Ziffern werden als Ordnungszahlen verwendet und sind am häufigsten in Kalenderdaten zu finden. Nach den römischen Ziffern wird ein Punkt geschrieben, wie zum Beispiel 10.X.2010, was dem Ausdruck 10. Oktober 2010 entspricht. In allen anderen Fällen bei der Verwendung von römischen Ziffern sind keine Besonderheiten zu beobachten [Vatov, 1995].

3.1.2.6. Satzzeichen

In der geschriebenen Sprache werden außer Buchstaben auch eine Reihe von Satz- und Sonderzeichen verwendet, mit denen ein Sprachsynthesystem umgehen soll. Ein großer Teil der Satzzeichen haben rein „graphische“ Funktion. Sie werden verwendet, um Wörter, Phrasen, Zahlen und Buchstaben voneinander zu trennen. In dem folgenden Absatz werden die meist gebrauchten Satzzeichen mit ihrer Bedeutung aufgelistet.

1. Punkt: ein Punkt wird verwendet bei der Abkürzung von Wörtern, Namen und Ausdrücken. Beim Nummerieren mit arabischen oder römischen Ziffern wird ebenfalls ein Punkt gesetzt. Bei der Kennzeichnung von Kalenderdaten wird ebenfalls ein Punkt verwendet. Ein Punkt wird nach einer arabischen Zahl gesetzt, um zu signalisieren, dass ein Grundzahlwort eine Ordinalzahl und keine Kardinalzahl ist z.B. „1. клас“ – [p@rvi klas] *Dt. erste Klasse*, aber „1 клас“ – [edin klas] *Dt. eine Klasse* [BAN, 2005].

2. Fragezeichen: dieses Zeichen wird am Ende eines Satzes oder innerhalb eines Satzes geschrieben, um Zweifel, Erstaunen, Unverständnis oder eine Frage zu zeigen [BAN, 2005].

3.2. Das Phonemsystem der bulgarischen Sprache

Im vorherigen Abschnitt wurde die Orthographie der bulgarischen Sprache mit ihren Besonderheiten, die für die Sprachsynthese von Bedeutung sind, vorgestellt. Da für die Ausgabe eines Sprachsynthesystems das Lautinventar der zu synthetisierenden Sprache grundlegend ist, wird in diesem Abschnitt das Phonemsystem der bulgarischen Sprache ausführlich beschrieben.

Das Phonemsystem der bulgarischen Sprache besteht aus 45 Phonemen. Sechs davon sind Vokale und die restlichen 39 bilden die Gruppe der Konsonanten. Im folgenden Abschnitt

wird jedes Gruppenmitglied mit seinen Charakteristika ausführlich dargestellt. Die Regeln, die für eine einheitliche Schreibweise der Wörter in einer Sprache dienen, bilden das Rechtsschreibungssystem dieser Sprache.

Viele Rechtsschreibungssysteme verändern sich, indem sie sich von einigen veralteten Formen und Elementen befreien und sich an neue, aktuelle Formen anpassen. Die Rechtsschreibungsprinzipien, nach denen sich die meisten Rechtsschreibungssysteme richten, sind phonetische, morphologische und historische Prinzipien.

- Phonetisches Prinzip: definiert die Rechtsschreibung der Wörter durch die bereits festgelegte Aussprache. Die Schreibweise entspricht genau der Aussprache. Ein Beispiel dafür in der bulgarischen Sprache ist, als Ergebnis der Auswirkung des phonetischen Prinzip, das Ausfallen von [ə] „ъ“ und [er malək] „ь“ am Ende eines Wortes [BAN, 2005].
- Morphologisches Prinzip: welches das Hauptprinzip der bulgarischen Rechtsschreibung darstellt, verlangt die gleiche Schreibweise unabhängig von der Aussprache. Dies bedeutet z.B., dass die Auslautverhärtung der stimmhaften Konsonanten in der geschriebenen Form nicht wiedergespiegelt wird, z.B. es wird „por“ [rok] *Dt. Horn*, „porче“ [roktse] *Dt. Hörnchen* geschrieben [BAN, 2005].
- Historisches Prinzip: in der Geschichte der bulgarischen Rechtsschreibung wird ebenfalls die Anwendung des historischen Prinzips betrachtet. Zum Beispiel wurden aus rein historischen Gründen während einer langen Periode die Grapheme „ъ“ [ə] und „ь“ [er malək] traditionell weiter verwendet, ohne einen Aussprachewert zu besitzen. Heutzutage wird das historische Rechtsschreibungsprinzip in der Schreibweise von Wörtern wie „евтин“ [eftin] *Dt. günstig* und „втори“ [ftori] *Dt. zweite* betrachtet [BAN, 2005].

3.2.1. Graphem-zu-Phonem Beziehungen

Die Erstellung des bulgarischen Alphabets baut auf das Grundprinzip auf, dass jedem Graphem nur ein Phonem zugeordnet werden kann. Das erleichtert das Schreiben von Ausspracheregeln, so dass bei Sprachen wie dem Bulgarischen fast alles über solche Regeln abgedeckt werden kann. Im Gegensatz zu Deutsch oder Englisch, für die ein Lexikon benötigt wird. Die Korrelation zwischen Graphemen und Phonemen bildet das Fundament des bulgarischen Rechtsschreibungssystems. Das phonetische und das graphemische System einer

Sprache stimmen in den meisten Fällen nicht überein, was einige Unregelmäßigkeiten verursacht. Die bulgarische Sprache verfügt über 30 Grapheme und 45 dazugehörige Phoneme (sechs Vokalphonemen und 39 Konsonantenphonemen). Diese große Differenz zwischen den zwei Systemen beruht auf der Tatsache, dass es im Bulgarischen fast zu jedem harten (nicht palatalisierten) Konsonanten eine weiche (palatalisierte) Variante gibt. Deshalb werden die Zeichen von den nicht palatalisieren Konsonanten übernommen und die Palatalität wird mit einem Apostroph markiert. Auf diese Weise ergibt sich der Unterschied zwischen der Anzahl der Graphemen und der Anzahl der Phoneme im Bulgarischen [Tilkov, 1998]. Diese Differenz stellt allerdings kein Hindernis für die genaue Abbildung der Phoneme des Bulgarischen dar. Um die Übersichtlichkeit zwischen Phoneme und Grapheme in der bulgarischen Sprache zu erhöhen, werden diese Beziehungen in mehrere Gruppen kategorisiert. Diese Relationen zwischen Phoneme und Grapheme in Bulgarisch werden in vier Abbildungen dargestellt.

Graphem	о	у	ъ	е	и	й	ж	ч	ш	х
Phonem	[O]	[u]	[@]	[e]	[i]	[j]	[Z]	[tS]	[S]	[h]

Abbildung 3.3. Konstellation ein Graphem entspricht genau ein Phonem

Wie in Abbildung 3.3. dargestellt, entsprechen die Phoneme [O, u, @, e, i, j, Z, tS, S, x] immer demselben Graphem, unabhängig davon ob sie sich in Anlaut, Inlaut oder Auslaut Position befinden.

Graphem	а	б	в	г	д	з	к	л	м	н	п	р	с	т	ф	ц
Phonem	[a], [@]	[b], [b']	[v], [v']	[g], [g']	[d], [d']	[z], [z']	[k], [k']	[l], [l']	[m], [m']	[n], [n']	[p], [p']	[r], [r']	[s], [s']	[t], [t']	[f], [f']	[ts], [ts']

Abbildung 3.4. Konstellation ein Graphem entspricht zwei Phoneme

Abbildung 3.4. stellt die Grapheme dar, die mehr als ein Phonem als Entsprechung haben. Diese Konstellation behandelt die größte Anzahl an Graphemen, nämlich „а, б, в, г, д, з, к, л, м, н, п, р, с, т, ф, ц“ [a, b, v, g, d, z, k, l, m, n, p, r, s, t, f, ts].

Graphem	щ	я	ю
Phonem	[St]	[ja]	[ju]

Abbildung 3.5. Konstellation ein Graphem entspricht Verbindung von zwei Phoneme

Abbildung 3.5. beschreibt Grapheme, die einer Verbindung von zwei Phonemen entsprechen. Dazu gehören insgesamt drei Buchstaben /щ, я, ю/ [St, ja, ju]. Die Grapheme /я/ [ja] und /ю/ [ju] haben die zusätzliche Funktion, den vorhergehende Konsonanten zu palatalisieren. Das Graphem /щ/ [St] dient nur zum Ausdruck zweier aufeinander folgender Laute, und zwar der Konsonanten /ш/ [S] und /т/ [t], z.B. „поща“ [poSta] *Dt. Post*.

Graphem	дж	дз
Phonem	[dZ]	[dz]

Abbildung 3.6. Konstellation zwei Grapheme entsprechen einem Phonem

Die Buchstaben /дж/ [dZ] und /дз/ [dz], die die vierte Konstellation bilden, entsprechen jeweils einem Phonem, z.B. „джоб“ [dZob] *Dt. Tasche*, „дзън“ [dz@n] *Dt. kling-kling* [Stoyanov, 1999]. In SAMPA werden sie allerdings mit zwei Zeichen dargestellt.

Die Grapheme „ь“ [er mal@k] ist in keine Tabelle aufgenommen, weil es keinen phonetischen Wert hat und nur zur Kennzeichnung palatalisierte Konsonanten vor /o/ verwendet wird, z.B. „бельо“ [bel'O] *Dt. Wäsche*. Deshalb kann es nie am Anfang eines Wortes stehen und existiert nur als Kleinbuchstabe [BAN, 2005].

3.2.2. Vokale

Das Vokalinventar des Bulgarischen umfasst sechs Phoneme: [i, e, @, a, O, u] /и, е, ъ, а, о, у/. Das bulgarische Vokalsystem wird in Abhängigkeit von der Betonung aufgebaut. In betonter Position gibt es sechs vokalischen Phoneme [i, e, a, @, O, u] /и, е, ъ, а, о, у/ und in akzentloser Position gibt es nur drei vokalische Phoneme [a, u, i] /а, у, и/. Unter Akzent haben die Vokale eine mittlere Länge, und in nicht akzentuierter Position werden sie kurz ausgesprochen [Radeva, 2003].

		Artikulationsstelle				
		vorn	mitte	hinten		
Zungenlage	hoch	i		u	eng	Gaumen/Zunge Abstand
	mitte	e	@	O		
	niedrig		a		weit	
		nicht labial		labial		
		Lippenrundung				

Abbildung 3.7. Vokale der bulgarischen Sprache [Radeva, 2003]

Wie Abbildung 3.7. zeigt, lassen sich die bulgarische Vokale durch die folgenden distinktive Merkmale identifizieren:

- Artikulationsstellen: [vorn, mitte, hinten]. Die zwei Vokalphoneme [e] /e/ und [i] /ɪ/ werden im vorderen Teil des Mundraumes artikuliert und die Zunge wird nach vorne gebracht, deshalb werden sie als vordere Vokale bezeichnet. Als mittlere Vokale gelten [a] /a/ und [ɔ] /ɔ/. Die Zunge bleibt zentral. Bei den hinteren Vokalen wie [O] /o/ und [u] /y/ wird die Zunge nach hinten genommen. Die mittleren und hinteren Vokale werden auch „harte“ Vokale genannt, weil sie keinen Einfluss auf die Palatalisierung der Konsonanten haben.
- Zungenlage: [niedrig, mitte, hoch]. Die bulgarische Vokale lassen sich in Abhängigkeit von der Zungenlage in drei Gruppen unterteilen: Vokale mit niedriger Zungenlage, Vokale mit mittlerer Zungenlage und Vokale mit hoher Zungenlage. Zur ersten Gruppe gehört der Vokal [a] /a/, weil er mit der niedrigsten Position der Zunge artikuliert wird. Die Vokale [e] /e/, [O] /o/ und [ɔ] /ɔ/ werden mit mittlerer Zungenposition ausgesprochen, somit bilden sie die zweite Gruppe. Die Vokale [i] /ɪ/ und [u] /y/ werden auch „hohe“ Vokale genannt. Sie werden mit der höchsten Anhebung der Zunge ausgesprochen.
- Abstand zwischen Gaumen und Zunge: offene und geschlossene Vokale [weit, eng]. Bei der Artikulation der Vokale [a] /a/, [e] /e/, [ɔ] /ɔ/ und [O] /o/ ist die Entfernung zwischen der höchsten Stelle der Zunge und dem Gaumen sehr weit, deshalb werden die Vokale auch „weite“ Vokale genannt. Dieser Abstand ist bei den Vokalen [i] /ɪ/ und [u] /y/ sehr eng, deshalb die Bezeichnung „enge“ Vokale.
- Lippenrundung: [labial, nicht labial]. Bei der Aussprache der Vokale [a] /a/, [e] /e/, [i] /ɪ/ und [ɔ] /ɔ/ sind die Lippen nicht oder ganz leicht beteiligt, deshalb werden sie „nicht gerundete“ Vokale genannt. Im Gegensatz dazu sind bei der Artikulation der Vokale [O] /o/ und [u] /y/ die Lippen beteiligt, deshalb die Bezeichnung „gerundete“ Vokale.

Die sechs Vokalphoneme der Bulgarischen werden mit acht Graphemen dargestellt. Die Zuordnung von Phonemen zu Graphemen wird in die folgende Tabelle abgebildet:

Grapheme	Phoneme	Beispiele	Deutsch
а	[a] und [@]	Ангел [angel] , чета [tSet@]	der Engel, lese
е	[e]	Ела [ela]	die Tanne
и	[i]	Игла [igla]	die Nadel
о	[O]	Обица [Obitsa]	der Ohring
у	[u]	Ухо [uxo]	das Ohr
ъ	[@]	Ъгъл [@g@l]	der Winkel

Abbildung 3.8. Graphem- zu- Phonem Beziehungen bei der Vokalen

Die Grapheme /ю/ [ju] und /я/ [ja] sind keine selbständige Laute. Wenn sie sich nach einem Konsonanten oder [a] und [@] befinden, werden diese palatalisiert [Stoyanov, 1999].

In der bulgarischen Sprache existieren keine Nasalvokale, nur orale Vokale. Vokale können in bestimmten Fällen nasalisiert werden. Das ist jedoch kein distinktives Merkmal und daher phonologisch nicht von Bedeutung [Tilkov, 1998].

3.2.3. Konsonanten

Die bulgarische Sprache verfügt über 39 Konsonantenphoneme: [p, b, t, d, k, g, ts, tS, dZ, f, v, s, z, S, Z, x, m, n, l, r, j, p', b', t', d', k', g', ts', dz', f', v', s', z', x', m', n', l', r'], und sie werden durch 21 Grapheme dargestellt: /п, б, т, д, к, г, ц, ч, ф, в, с, з, ш, ж, х, м, н, л, р, ѝ/ (siehe Abb. 3.9).

			Artikulationsstelle											
			labial				alveolar				post-alveolar		velar	
			bilabial		labiodental		alveodental		alveolar					
			sh	sl	sh	sl	sh	sl	sh	sl	sh	sl	sh	sl
Artikulationsart	Plosiv	hart	b	p			d	t					g	k
		weich	b'	p'			d'	t'					g'	k'
	Fikativ	hart			v	f	z	s			Z	S		x
		weich			v'	f'	z'	s'						x'
	Affrika t	hart					dz	ts			dZ	tS		
		weich					dz'	ts'						
	Lateral	hart					l							
		weich					l'							
	Vibrant	hart							r					
		weich							r'					
	Nasal	hart	m						n					
		weich	m'						n'					
	Glide										j			

Abbildung 3.9. Konsonanten des Bulgarischen [Tilkov, 1998]

Die zwei Hauptcharakteristika jedes Konsonanten sind seine Artikulationsstelle und seine Artikulationsart. Im Folgenden werden die distinktiven Merkmale der bulgarischen Konsonanten einzeln betrachtet.

- Artikulationsstelle: [bilabial, labiodental, alveodental, alveolar, postalveolar, velar]. In Abhängigkeit von der Artikulationsstelle lassen sich die bulgarischen Konsonanten in vier Hauptkategorien unterteilen: labial, alveolar, postalveolar und velar. Die erste Gruppe, die labialen Konsonanten wird in zwei weitere Untergruppen geteilt: bilabiale Konsonanten und labiodentale Konsonanten. Labiale Konsonanten werden mit Hilfe der Unterlippe ausgesprochen. Bei den bilabialen Konsonanten beteiligt sich auch die Oberlippe bei der Aussprache. Auf diese Art und Weise entstehen die Phoneme [p, p', b, b', m, m']. Die labiodentalen Konsonanten entstehen indem die untere Lippe die oberen Zähne berührt. Dazu gehören die Phoneme [f, f', v, v'].

Die zweite Gruppe wird ebenso in zwei Untergruppen geteilt: alveodentale und alveolare Konsonanten. Zu der Untergruppe der alveodentalen Konsonanten gehören die folgenden Phonemen: [d, d', t, t', z, z', s, s', dz, dz', l, l', ts, ts']. Zu der zweiten Untergruppe, der Gruppe der alveolare Konsonanten zählen die Phoneme: [r, r', n, n']. Zu der dritten Gruppe, der Gruppe der postalveolaren Konsonanten, gehören die Phoneme [S, Z, tS, dZ].

Die vierte Gruppe bildet die Gruppe der velaren Konsonanten. Bei der Aussprache von velaren Konsonanten sind der weichen Gaumen und der hintere Teil der Zunge beteiligt. Auf diese Weise entstehen die Phoneme [k, k', g, g', x, x'].

- Artikulationsart: [Plosive, Frikative, Affrikate, Lateral, Vibrant, Nasal, Glide]. Die Plosive werden durch einen Verschluss im Vokaltrakt gebildet. Dieser Verschluss kann entstehen in zwei Situationen: entweder durch Ober- und Unterlippe oder durch Zunge und Gaumen. Als plosive Konsonanten im Bulgarischen zählen die folgenden Phoneme: [p, p', b, b', t, t', d, d', k, k', g, g'] [Tilkov, 1998]. Die Konsonantenphoneme [g] und [k] werden als positionelle Variante vor [e] und [i] leicht palatalisiert und sie ähneln somit ihren palatalen Entsprechungen [Radeva, 2003]. Bei der Artikulation der frikativen Konsonanten wird eine Engstelle gebildet, die die ausströmende Luft verwirbelt und einen Reibelaut erzeugt. Zu den Frikativen gehören die folgenden Phoneme: [f, f', v, v', s, s', z, z', S, Z, x, x'] [Tilkov, 1998]. Hier muss beachtet werden, das [s] in vokalischer Umgebung und vor Lateralen, Sonoranten und dem vibrantischen Konsonanten [r] auch am Wortanfang immer

stimmlos ist. Das Konsonantenphonem [z] ist in der gleichen Position immer stimmhaft. Die Aussprache des nicht palatalen Phonems [x] wird nicht von seiner Umgebung beeinflusst. Palatales [xʲ] kommt nur in Fremdwörtern vor [Radeva, 2003]. Die dritte Hauptgruppe, die Gruppe der Affrikaten beinhaltet die Konsonanten, die die akustisch-artikulatorischen Merkmale der plosiven und der frikativen Konsonanten des Bulgarischen in sich verbinden. Zu den Affrikaten des Bulgarischen werden die folgenden Phoneme gezählt: [ts, tsʲ, dz, dzʲ, tS, dZ] [Tilkov, 1998]. Das laterale /l/ bzw. /lʲ/: das nicht palatale Phonem [l] tritt in zwei positionellen Varianten auf. Vor den hinteren Vokalen, Konsonanten und am Wortende handelt es sich um ein „hartes“ [l]. Vor den Vokalphoneme [i] und [e] ähnelt die Artikulation dem mittleren [l]. Das palatale [lʲ] hat Charakteristika, die für alle palatalen Konsonanten gelten [Radeva, 2003]. Das vibrantischen [r] wird unabhängig von seiner Position immer als [r] artikuliert, auch am Wortende [Radeva, 2003]. Bei den nasalen Konsonantenphoneme [m] und [n] bzw. [mʲ] und [nʲ] ist zu beachten, dass der Konsonant [n] vor [g] und [k] obligatorisch eine velare positionelle Variante, wie im Deutschen [ŋ] vor [k] z.B. in „Bank“ hat [Radeva, 2003]. Das Konsonantenphonem [j] wird vor [a] und [ɔ] mit Hilfe des Buchstabens „я“ [ja], vor dem Phonem [o] wird durch [jo] und vor [u] durch „ю“ [ju] symbolisiert [Radeva, 2003].

Eine weitere Besonderheit des bulgarischen Konsonantensystems ist der Unterschied zwischen palatalisierten und nicht palatalisierten Konsonanten, die auch weiche und harte Konsonanten genannt werden. Palatale Konsonanten werden hier mit dem Zeichen [ʲ] markiert und sie unterscheiden sich von ihren nicht palatalen Entsprechungen durch eine bei ihrer Artikulation obligatorische Hebung des Zungenrückens zum harten Gaumen. Dadurch erhalten die Konsonanten einen „hellen Klang“. Zum Beispiel „ключ“ [klʲutS] *Dt. Schlüssel* [Radeva, 2003]. Alle bulgarische Konsonanten, außer den Konsonantenphonemen [S, Z, tS, dZ, j], können palatalisiert oder nicht palatalisiert auftreten [Tilkov, 1998]. Das distinktive Merkmal Palatalität hat nur vor hinteren Vokalen eine phonologische Funktion. Palatalisierte Konsonanten kommen vor anderen Konsonanten oder am Wortende nicht vor [Radeva, 2003].

Eine wichtige Rolle in dem bulgarischen Konsonantensystem spielt der Unterschied zwischen stimmhaften und stimmlosen Konsonanten. Auf Grund dieser Opposition entstehen im Bulgarischen die folgenden Paare:

stimmlos	p	p'	f	f'	t	t'	s	s'	ts	ts'	S	tS	k	k'	h	h'
stimmhaft	b	b'	v	v'	d	d'	z	z'	dz	dz'	Z	dZ	g	g'		

Abbildung 3.10. Stimmlosigkeit/Stimmhaftigkeit

Im Zusammenhang mit dem Merkmal Stimmhaftigkeit ist das phonetische Phänomen Stimmassimilation sehr wichtig. Dieses kommt an verschiedenen Stellen vor:

Stimmassimilation innerhalb der Wortgrenzen: die Regel lautet, dass innerhalb eines Wortes stimmhafte vor stimmlosen Konsonanten stimmlos und stimmlose vor stimmhaften stimmhaft artikuliert werden z.B. „градски“ [gratski] *Dt. städtisch*. Diese Regel gilt nur für die folgenden Konsonantenphoneme: [b, b', v, v', d, d', z, z', dz, dz', Z, dZ, g, g'] [Radeva, 2003].

Wortgrenzen-überschreitende Stimmassimilation: eine Besonderheit stellt das Verhalten der Präposition „в“ [v] bzw. „във“ [v@v] *Dt. in* dar. Vor Vokalen und den Konsonanten /v/, /r/, /l/, /m/, /n/ verliert das /v/ seine Stimmhaftigkeit z.B. „в око̀то“ [f okoto] *Dt. im Auge*, „във вла̀ка“ [v@f vlaka] *Dt. in den Zug* [Radeva, 2003].

Verlust der Stimmhaftigkeit im absoluten Auslaut: hier gilt die Regel, dass im absoluten Auslaut die Artikulation stimmhafter Konsonanten nicht möglich ist z.B. „град“ [grat] *Dt. die Stadt*, „градът“ [grad@t] *Dt. diese Stadt*, „град Берлин“ [grad berlin] *Dt. die Stadt Berlin* [Radeva, 2003].

3.3. Wortbetonung der bulgarischen Sprache

Die Betonung im Bulgarischen ist nicht an eine bestimmte Silbenposition im Wort gebunden und wird somit als frei bezeichnet. Zum Beispiel fällt die Betonung in Wörter wie „водопад“ [vo#⁶do#pad'] *Dt. Wasserfall* und „глава“ [gla#va'] *Dt. Kopf* auf die letzte Silbe, auf die vorletzte Silbe fällt sie in Wörtern wie „година“ [go#di'#na] *Dt. Jahr*. Die bulgarische Betonung ist nicht nur frei, sondern auch beweglich. Das bedeutet, sie kann ihre Position in verschiedenen morphologischen Formen ein- und desselben Worts ändern, z.B. „име“ [i'#me] *Dt. Name* und „имена“ [i#me#na'] *Dt. Namen* [Stoyanov, 1999].

⁶ Das Symbol # kennzeichnet in Folgende eine Silbengrenze

Im Bulgarischen gibt es eine Reihe von Wörtern bzw. Wortformen, die klitisch sind, d.h., dass solche Wörter bzw. Wortformen keinen eigenen Akzent tragen und sich mit dem folgenden (Enklitika) oder dem vorangehenden (Proklitika) akzentuierten Wort zu einer Akzenteinheit vereinen. Zu Proklitika gehören die Präpositionen und Konjunktionen wie „ako“ [ako] dt. *wenn*, „да“ [da] dt. *ja*, „или“ [ili] dt. *oder*, „но“ [no] dt. *aber*, die Negationspartikel „не“ [ne], die Partikel zur Bildung des indikativischen Futurs „ще“ [Ste]. Enklitika sind die sog. Kurzformen der Personalpronomina, die Kurzformen des Reflexivpronomens, die indikativischen Präsensformen von „съм“ [s@m] in periphrastischen Formen des Verbparadigma, Anredepartikeln „бе“ [be], „ма“ [ma], „ле“ [le] und die Fragepartikel „ли“ [li] [Radeva, 2003]. Die Tatsache, dass die Wortbetonung im Bulgarischen nicht gekennzeichnet wird, stellt eine große Herausforderung für Sprachsynthesysteme dar. Die Unregelmäßigkeiten bei dem Auftreten der Wortbetonung erschweren zusätzlich die Modellierung dieser Komponente in einem Sprachsynthesystem. Da die Weiterverfolgung des Problems den Rahmen dieser Arbeit sprengen würde, wird die Wortbetonung hier nicht weiter behandelt. Bei der Auswertung der Daten wird die Wortbetonung auch nicht berücksichtigt. Alle Wörter, die ohne die Markierung der Wortbetonung falsch transkribiert werden, werden in Lexikon eingetragen.

4. Ausspracheregeln für das Bulgarische im TTS-System Festival

Auf Basis der in Kapitel 3 vorgestellten Orthographie und Grammatik der bulgarischen Sprache werden in diesem Kapitel die Ausspracheregeln für das Bulgarische im Sprachsynthesystem Festival definiert. Dazu wird der Formalismus des Festival Sprachsynthesystems verwendet. Zunächst wird ein kürzerer Überblick über das Festival Sprachsynthesystem gegeben. Einzelne Module des Systems wie die Letter-to-Sound-Regeln oder der Lexikonlookup, die von großer Bedeutung für die Lösung dieser Aufgabe sind, werden detailliert beschrieben. Die im Rahmen dieser Arbeit entwickelten Skripte werden ausführlich diskutiert. Die Informationen in diesem Kapitel basieren auf der Festival Systemdokumentation⁷ [Black / Taylor / Caley, 2001].

⁷ URL (11.11.10): http://www.festvox.org/docs/manual-1.4.2/festival_toc.html

4.1. Überblick

Festival ist ein frei verfügbares mehrsprachiges System, entwickelt an der Universität von Edinburgh, das einen allgemeinen Rahmen für den Aufbau von Sprachsynthesystemen für beliebige Sprachen bietet.

Das System ist in C++ geschrieben und benutzt die Edinburgh Speech-Tools für Low-Level-Architektur. Es verfügt über einen Scheme (SIOD) basierten Kommando-Interpreter, wodurch die Einbindung neuer Sprachen ohne Modifikation des zugrundeliegenden C++ Programmcode erfolgen kann.

Festival als ein Sprachsynthesystem deckt drei unterschiedliche Anwendererebenen ab. Erstens diejenigen, die eine hochwertige Aussprache von beliebigem Text mit einem Minimum an Aufwand wollen. Zweitens, die Sprachentwickler. Die dritte Ebene ist die Entwicklung und Erprobung neuer Synthesemethoden.

Festival ist speziell für das einfache und effiziente Hinzufügen neuer Module konstruiert, so dass die Entwicklung leicht wird. Das ist ein Grund, warum die entwickelten Ausspracheregeln für das Bulgarische im Festival Formalismus verfasst wurden.

Das Sprachsynthesystem Festival verfügt über eine eigene interne Datenstruktur, die Utterance-Structure genannt und im Deutschen mit Äußerungsstruktur übersetzt wird. Der Aufbau der Äußerungsstruktur erfolgt, indem verschiedene Module durchlaufen werden, die der Äußerungsstruktur nach und nach neue Informationen hinzufügen. Welches Modul durchlaufen wird, hängt von der Wahl der Stimme und des Äußerungstyps ab. In IMS-Festival ruft der Äußerungstyp Text zurzeit folgende Module der Reihe nach auf: Initialize, Text, Token-Pos, Token, POS, Phrasify, Word, Pauses, Intonation, PostLex, Duration, Int-Targets, Wave-Synth. In den folgenden Abschnitten wird ein allgemeiner Überblick über die Festival Module im Text-Modus gegeben. Zu manchen Modulen wird eine ausführliche Beschreibung folgen und andere werden nur kurz erwähnt.

- Mit dem Aufruf des Moduls *Initialize* werden alle nötigen Relationen aus der Eingabe geladen und alle bestehenden Relationen gelöscht.
- Bei Aufruf des Moduls *Text* wird die Tokenisierung durchgeführt. Bei der Tokenisierung wird Interpunktion (" ' . , : ; ! ? ()[]) vor und nach einem Token abgetrennt und als Feature des Tokens gespeichert.

- Im Modul *Token-Pos* findet die Tokentyperkennung statt. Hier wird jedem Token einen Typ zugeordnet.
- Bei Aufruf des Moduls *Token* wird die Token-Wort-Konvertierung ausgeführt. An dieser Stelle wird die Expansion von Abkürzungen, E-Mail-Adressen, Geldbeträgen, Verhältniszahlen, Rechnungen (+ - * :), Telefonnummern, gemischten Token, römischen Zahlen, Sonderzeichen, Maßeinheiten, Abkürzungen, Jahreszahlen, Datumsangaben, Kardinalzahlen, Ordinalzahlen und rationalen Zahlen vorgenommen.
- Im Modul *POS* wird jedem Wort ein Part-Of-Speech-Tag zugewiesen.
- Bei Aufruf des Moduls *Phrasify* erfolgt die Bestimmung der Phrasengrenzen.
- Im Modul *Word* findet der Lexikonnachschlag statt. Die Relationen *SylStructure*, *Segment* und *Syllable* werden hier erzeugt. Eine Lexikon-Einheit in Festival besteht aus drei Teilen: Addendalexikon (siehe Kapitel 4.2.2), kompilierten Vollformlexikon und Ausspracheregeln. Beim Lexikonlookup wird zuerst im Addendalexikon nachgeschlagen, wenn der Eintrag dort nicht gefunden wird, wird in einem großen kompilierten Vollformlexikon nachgeschlagen. Falls auch hier kein Eintrag existiert, werden Ausspracheregeln auf das zu sprechende Wort angewendet.
- Das Modul *Pauses* fügt die Phrasengrenzen Pausen ein, die durch das Modul *Phrasify* bestimmt wurden.
- Im Modul *Intonation* findet die Intonationsgenerierung statt.
- Bei Aufruf des Moduls *PostLex* werden die postlexikalischen Regeln ausgeführt. Durch sie wird die Koartikulation modelliert, welche die Natürlichkeit der Synthese steigert.
- Im Modul *Duration* geschieht die Bestimmung der Lautdauer.
- Das Modul *IntTargets* bestimmt die Grundfrequenzkontur.
- Im Modul *Wave-Synth* wird das Sprachsignal generiert.

4.2. Erstellung einer Aussprachekomponente in Festival

Dieser Abschnitt wird die Definition einer Aussprachekomponente in Festival darstellen. Es werden nur die Komponenten beschrieben, die für die Erstellung einer Aussprachekomponente in Festival von Bedeutung sind.

Die allgemeine Methode für die Definition einer neuen Stimme [Black / Lenzo, 2007] ist es, die Parameter für die verschiedenen Komponenten zu definieren wie z.B. Phoneteset, Lexikon, LTS, Intonation etc. Da die Miteinbindung einer Stimme außerhalb des Fokus dieser Arbeit

ist, wird hier nur auf die Erstellung derjenigen Komponenten begrenzt, die für die Aussprache in Festival eine Rolle spielen, nämlich: Phonesets, Lexikon und Letter-To-Sound-Regeln.

4.2.1. Phonesets

Um eine neue Aussprachekomponente in Festival zu definieren, ist zuerst ein neues Phoneset erforderlich. Das Phoneset ist der Grundbaustein einer Stimme, und die meisten anderen Teile wie das Lexikon oder die Letter-To-Sound-Regeln sind im Hinblick auf dieses Set definiert, deshalb wird die Erstellung des Sets als erstes erfolgen [Black / Lenzo, 2007].

Das Lautinventar der bulgarischen Sprache wurde bereits im Kapitel 3.2. dargestellt. Hier wird nun die Scheme-basierte Notation, die von Festival verwendet wird, vorgestellt und kurz erläutert.

Phonesets sind Mengen bestehend aus Lautsymbolen, ihren distinktiven Merkmalen und deren möglichen Werten. Die Grundform eines Phonesets hat die folgende Form:

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

- NAME steht für eine eindeutige Bezeichnung für das betreffende Phoneset.
- FEATUREDEFS ist eine Liste der Merkmale, die in der jeweiligen Sprache distinktiv sind, sowie ihre möglichen Werte.
- PHONEDEFS stellt eine Liste der eigentlichen Laute der Sprache inklusive der Merkmale und Werte dar.

NAME	FEATUREDEFS	
phoneset bulgarian	Merkmale	Mögliche Werte
	Vokal oder Konsonant	VC + -
	Vokallänge (kurz, lang, Diphthonge, SCHWA)	VL s l d a 0
	Vokalhöhe (hoch, mitte, niedrig)	VH 1 2 3 -
	Artikulationsstelle (vorne, mitte, hinten)	VF 1 2 3 -
	Labial/Nicht labial	VLR + -
	Gaumen/Zunge Abstand (weit, eng)	VOC + -
	Artikulationsart (Plosiv, Frikativ, Affrikat, Nasal, Lateral, Vibrant, Glide)	CT p f a n l v g 0
	Artikulationsstelle (bilabial, labiodental, alveodental, alveolar, postalveolar, velar)	CP bl ld ad al pa vl 0
	Stimmhaftigkeit/Stimmlosigkeit (stimmhaft, stimmlos)	CVOX + -

Abbildung 4.1. Phoneset für Bulgarisch

Abbildung 4.1. zeigt distinktiven Merkmale, die bei der Definition des bulgarischen Phonesets im Rahmen dieser Arbeit berücksichtigt wurden. Die Abbildung hat zwei Hauptspalten: NAME und FEATUREDEFS. In der ersten Spalte steht der Name des betreffenden Phonesets und in diesem Fall heißt es „phoneset bulgarian“. Die zweite Hauptspalte stellt eine Aufzählung der distinktiven Merkmale der bulgarischen Sprache und ihrer Werte dar. Sie wird in zwei Unterspalten Merkmale und mögliche Werte geteilt. In der Unterspalte Merkmale sind die Merkmale des Lautinventars der bulgarischen Sprache aufgelistet. Als Erstes werden die Merkmale aller bulgarischen Vokale aufgezählt. Und in den darauffolgenden Zeilen erfolgt die Definition der Konsonanten der bulgarischen Sprache. In der Unterspalte mögliche Werte werden ihre möglichen Werte aufgezählt.

Für die formale Definition der distinktiven Merkmale im Bulgarischen werden diverse Zeichen eingeführt. Wie in Abbildung 4.1. dargestellt, wird

- das Merkmal Vokal oder Konsonant (VC) als erstes definiert. Somit erhalten die Vokale [i] und [e] ein (+), das für das Merkmal „Vokal“ steht und wohingegen die Konsonanten an dieser Stelle ein Minus (-) erhalten werden, was dem Merkmal „Konsonant“ entspricht.
- Das nächste Merkmal in der Liste ist die Vokallänge, wofür die Abkürzung VL steht, werden die folgenden Zeichen verwendet: „s“ steht für „short“ oder kurz ausgesprochene Vokale, „l“ entspricht „long“ oder lang ausgesprochene Vokale, „d“ kennzeichnet alle Doppellaute und mit „a“ wird SCHWA markiert.
- Die Vokalhöhe wird mit VH abgekürzt. Hier steht 1 für hohe Vokale, 3 für tiefe Vokale.
- Für die Zungenlage wird die Abkürzung VF verwendet, was für Vowel Frontness oder der vertikalen Position den Vokalen steht (1 steht hier für vordere Vokale, 2 für zentrale und 3 für hintere Vokale).
- Das nächste Merkmal, das die Lippenrundung von Vokale kennzeichnet und wofür die Abkürzung VLR steht, 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.
- Das Merkmal Gaumen-Zunge Abstand wird mit VOC abgekürzt und hat die Werte (+) für weite und (-) für enge Vokale. Darauffolgend erhalten zum Beispiel die Vokale [i] und [u] den Wert (-) und [e] und [O] den Wert (+).

- Die nächsten beiden Merkmale CT und CP in der Liste bezeichnen die Artikulationsstelle und die Artikulationsart von Konsonanten, weshalb bei den Vokalen hier jeweils 0 steht.
- Das letzte Merkmal, das mit CVOX abgekürzt ist, steht für Stimmhaftigkeit oder Stimmlosigkeit bei Konsonanten. Mit dem Wert Plus werden alle stimmhaften Konsonanten gekennzeichnet und ein Minus steht für die stimmlosen.

Um ein Phonetset vollständig zu beschreiben, ist die Festlegung des PHONEDEFS Parameters notwendig, welche das gesamte phonetische Inventar der bulgarischen Sprache mit seinen Unterscheidungsmerkmalen und deren Werten auflistet. Um einen groben Eindruck der eigentlichen distinktiven Merkmale und ihrer Werte im Bulgarischen zu erhalten, wird ein Ausschnitt aus der Definition des Phonetsets für Bulgarisch in Abbildung 4.2. dargestellt.

PHONEDEFS									
Phonem	Merkmale + Werte								
	VC	VL	VH	VF	VLR	VOC	CT	CP	CVOX
i	+	s	1	1	-	-	0	0	-
e	+	s	2	1	-	+	0	0	-
@	+	s	2	2	-	+	0	0	-
O	+	s	2	3	+	+	0	0	-
p	-	0	-	-	-	-	s	b	-
b	-	0	-	-	-	-	s	b	+
f	-	0	-	-	-	-	f	l	-
v	-	0	-	-	-	-	f	l	+
S	-	0	-	-	-	-	f	p	+
Z	-	0	-	-	-	-	f	p	+

Abbildung 4.2. Ausschnitt aus der Definition des Phonetsets für Bulgarisch

Wie in Abbildung 4.2. dargestellt, besitzt jedes Phonem eine Reihe von Unterscheidungsmerkmalen und Werten. In der Spalte Phonem sind die Laute der bulgarischen Sprache aufgelistet. Das erste Zeichen in der Abbildung steht für den zu definierenden Laut, im oberen Beispiel steht also „i“ für den Vokal [i], „e“ für den Vokal [e], „p“ für den Konsonant [p] und „b“ für den Konsonant [b]. Die zweite Spalte Merkmale+Werte ist in neun Unterspalten aufgeteilt. Jede Unterspalte stellt ein Unterscheidungsmerkmal dar, das bereits in FEATUREDEFS definiert ist, und seinen Wert dazu.

Das vollständige Phonetset für das Bulgarische ist in Anhang A zu finden.

4.2.2. Lexikon

Ein Lexikon in Festival ist ein Subsystem, das die Aussprache für die Wörter liefert. Es kann aus drei verschiedenen Teilen bestehen: ein Addenda, in der Regel von Hand hinzugefügte Wörter, um nachträglich noch Lexikoneinträge hinzufügen zu können; ein kompiliertes Lexikon, in der Regel groß (ca. 10.000 Wörter), und eine Methode für den Umgang mit Wörtern, die nicht in der Addenda oder den Lexikon zu finden sind. Da bei der Sprachsynthese immer mit unbekanntem Wörtern gerechnet werden muss, wird eine solche Methode benötigt. Dies wird durch Letter-To-Sound-Regeln erreicht. Im Folgenden werden alle drei Teile beschrieben.

4.2.2.1. Lexikoneinträge

Lexikoneinträge bestehen aus drei Teilen, das Wort in ihrer orthographischen Form, seine Wortart (das POS-Tag oder Part-Of-Speech Tag) und die Transkription des Wortes einschließlich der Silbenstruktur und des Wortakzentes. Der Wortakzent wird in Festival generell durch 0 für unbetonte und 1 für betonte Silben gekennzeichnet.

Bestandteile eines Lexikoneintrags		
orthographische Form des Wortes	Part-Of-Speech-Tag des Wortes	Transkription des Wortes Silbenstr. + Wortakzent
игра <i>dt. das Spiel</i>	n	((i) 1) ((gra) 0)
игра <i>dt. spielen</i>	v	((i) 1) ((gra) 0)
коса <i>dt. die Haare</i>	n	((ko) 0) ((sa) 1)
кожа <i>dt. die Haut</i>	n	((ko) 0) ((za) 1)
вълна <i>dt. die Wolle</i>	n	((v@l) 0) ((na) 1)
гама <i>dt. die Farbkonstellation</i>	n	((ga) 0) ((ma) 1)
риза <i>dt. der Hemd</i>	n	((ri) 0) ((za) 1)

Abbildung. 4.3. Bestandteile eines Lexikoneintrags

Wie in Abbildung 4.3. dargestellt, können zwei lexikalische Einträge dem gleichen Wort entsprechen, aber die Differenzierung zwischen den beiden wird durch ihren Part-Of-Speech-Tag erfolgen. Wie mehrere lexikalische Einträge mit gleicher orthographischer Form von dem System abgearbeitet werden, wird im Folgenden beschrieben.

4.2.2.2. Lexikonlookup

Bei der Suche nach einem Wort, entweder über die C++-, oder die Lisp-Schnittstelle, ist das Wort durch seine orthographische Form und sein Part-Of-Speech-Tag zu identifizieren.

Beim Lexikonlookup wird zuerst in der Addenda nachgeschlagen. Wenn eine entsprechende Übereinstimmung (lexikalisches Wort plus POS-Tag) vorliegt, wird sie zurückgegeben. Wenn keine Übereinstimmung in der Addenda gefunden wird, wird das kompilierte Lexikon, falls vorhanden, geprüft. Falls der Eintrag nicht im kompilierten Lexikon gefunden wird, wird das Wort an eine bereits definierte Methode übergeben, die sich um die Bearbeitung unbekannter Wörter kümmert. In der Regel wird diese Aufgabe von Letter-To-Sound-Regeln übernommen [Black / Taylor / Caley, 2001].

4.2.3. Letter-To-Sound-Regeln

Jedes Lexikon kann bestimmen, welche Maßnahmen ergriffen werden sollen, wenn ein Wort nicht in der Addenda oder in dem kompilierten Lexikon gefunden wird. Es gibt eine Reihe von Optionen. Eine davon sind die Letter-To-Sound-Regeln. Diese Methode verwendet extern erstellte Letter-To-Sound-Regeln.

Die manuelle Erstellung der Letter-To-Sound-Regeln ist zeitaufwändig, aber bei Sprachen wie Bulgarisch, wo eine relativ eindeutige Beziehung zwischen der Orthographie und der Aussprache eines Wortes besteht, bieten sich handgeschriebene Regeln an. Festival verfügt ebenfalls über eine alternative Methode, die Letter-To-Sound-Regeln automatisch zu erstellen. Dafür muss ein umfangreiches elektronisches Aussprachelexikon der zu bearbeitenden Sprache vorliegen. Für die bulgarische Sprache ist dies nicht der Fall. Aufgrund dessen wurde in dieser Arbeit eine manuelle Erstellung der Letter-To-Sound-Regeln gewählt.

In diesem Unterkapitel wird die Erstellung und die Adaption von solchen Regeln in Festival dargestellt. Ihre Grundform und ihre Notation in Festival wird ausführlich beschrieben. Einige konkrete Ausspracheregeln für das Bulgarische im Festival-Formalismus werden diskutiert. Der grundlegenden Mechanismus der Letter-To-Sound-Regeln ist einfach und gleichzeitig mächtig genug, um die Konstruktion von komplexen Regeln zu unterstützen. Die Grundform der Regel ist wie folgt:

(LEFTCONTEXT [ITEMS] RIGHTCONTEXT = NEWITEMS)

Dabei steht LEFTCONTEXT für den linken Graphemkontext. Analog dazu steht RIGHTCONTEXT für den rechten Graphemkontext. ITEMS beinhaltet die Grapheme, die transkribiert werden sollen und NEWITEMS entspricht dem Phonem, durch welches sie ersetzt werden. Die Ergebnisse der Regeln, also die NEWITEMS Phoneme, stehen im Fortgang der Prozessierung nicht mehr zur Transkription zur Verfügung.

Die Regeln werden von links nach rechts angewendet. Die seltensten und spezifischsten Regeln werden am Anfang und die Standardregeln am Ende des Regelsets notiert. Zum Beispiel wird im Bulgarischen das Graphem „л“ im Kontext von [ju] oder [ja] als palatalisiertes [l'] ausgesprochen, ansonsten als nicht palatalisiertes [l]. Somit erhält die palatalisierte [l'] ein höheres Ranking im Regelset. Um beide Kontexte für die palatalisierte Version abzudecken, wurde die Variable P definiert, die die Menge der Grapheme „ю“ und „я“ umfasst. Damit kann die Aussprache von „л“ durch zwei Regeln beschrieben werden, nämlich ([л] P = l') und ([л] = l). Da die bulgarische Sprache über zahlreiche Konsonanten verfügt, die im Kontext von [ju] und [ja] palatalisiert werden, wird somit die Regelanzahl deutlich reduziert. Sowohl der linke wie auch der rechte Kontext können Variablen enthalten, die für eine Menge von Graphemen stehen.

Die Menge aller Vokale der bulgarischen Sprache wird mit einer Variablen definiert. Dafür steht die Variable V. Mit der Definition dieser Variable wurde die Übersichtlichkeit und die Kompaktheit des Regelsets erhöht. Das lässt sich an dem folgenden Beispiel illustrieren: im Bulgarischen wird das Graphem „я“ am Anfang eines Wortes oder vor eine Vokal als [ja] ausgesprochen. Wenn sich das Graphem „я“ aber am Ende eines Wortes befindet, dann wird es mit [@] ausgesprochen. Die Aussprachevarianten des Graphems „я“ [ja] werden mit den folgenden Regeln festgelegt: (# [я] = j a) und ([я] V = j a) und ([я] # = @), wobei das Symbol # zur Kennzeichnung von Wortgrenzen dient.

Zwei andere Variablen CVOICED und CVOICELESS werden eingeführt um die Menge aller stimmhaften und stimmlosen Konsonanten abzudecken. Mit der Variable CVOICED sind die folgenden stimmhaften Konsonantenphoneme definiert: [b, v, d, z, dz, Z, dZ, g]. Im Gegensatz dazu sind die stimmlosen Konsonanten im Bulgarischen mit der Variable CVOICELESS definiert. Dazu gehören die folgenden Konsonantenphonem: [p, f, t, s, ts, S,

tS, k, h]. Mit der Definition dieser Variablen wird die Anzahl der Regeln deutlich reduziert. Somit werden Regeln wie ([B] k = f k) und ([B] c = f s) und ([B] T = f t) mit nur einer Regel dargestellt. In diesem Fall über die Folgende: ([B] CVOICELESS = f). Regeln wie ([c ɔ] = z b) und ([c ɗ] = z d) sind mit der Einführung der CVOICED Variable überflüssig, sie werden mit der Regel ([c] CVOICED = z) abgedeckt.

Der einzige Vokal⁸ der bulgarischen Sprache, der mehrere Aussprachevarianten besitzt, ist das Vokalgraphem „a“. Im Bulgarischen wird das Graphem „a“ am Ende eines Wortes als [@] ausgesprochen, ansonsten als [a]. Eine Ausnahme machen hier Wörter bei denen die Wortbetonung auf das Phonem [a] fällt und sich das Phonem [a] gleichzeitig im Auslaut befindet. In solchen Fällen wird das Graphem „a“ als [a] transkribiert. Da in dieser Arbeit die Wortbetonung nicht berücksichtigt wird, wird diese Eigenschaft der bulgarischen Sprache nicht modelliert. Die betroffenen Wörter werden in die Addenda eingetragen. Um die Aussprachevariante des Vokalgraphems „a“ in den restlichen Fällen abzubilden, werden die folgenden Regeln entwickelt: ([a] # = @) und ([a] = a) .

Wie bereits im Kapitel 3.2.2. beschrieben, spielt die Opposition Stimmhaftigkeit/Stimmlosigkeit eine wichtige Rolle im bulgarischen Konsonantensystem. Zum Beispiel wird im Bulgarischen das Graphem „б“ [b] am Ende eines Wortes in Folge von Auslautverhärtung in einer stimmlosen Variante „п“ [p] ausgesprochen. Dasselbe gilt für die Grapheme „в“ [v], „г“ [g], „д“ [d], „ж“ [Z], „дж“ [dZ], „з“ [z], „дз“ [dz]. Für diese Opposition werden die folgenden Regeln festgelegt: ([ɔ] # = p), ([B] # = f), ([T] # = k), ([ɗ] # = t), ([ʒ] # = S), ([ɗʒ] # = tS), ([ɜ] # = s), ([ɗɜ] # = ts) .

Wenn das Graphem „т“ [t] sich zwischen den zwei Konsonanten „с“ [s] und „м“ [m] befindet, wird es nicht ausgesprochen. Diese Besonderheit der bulgarischen Sprache wird mit dem folgenden Regel abgebildet: ([c T M] = s m) .

Die Hauptaufgabe der Letter-To-Sound-Regeln in Festival ist, jeden beliebigen Text zu transkribieren, d.h. eine Graphemfolge wird in eine Phonemfolge umgewandelt. In dem

⁸ In der Literatur finden sich unterschiedliche Auffassungen im Bezug auf den Vokalstatus von [ju] und [ja]. Hier wird nach [BAN, 2005], d.h. [ju] und [ja] sind Kombinationen aus Semivokal und Vokal, vorgegangen.

folgenden Abschnitt wird das Regelset fürs Bulgarische darstellen. Um einen besseren Überblick zu behalten, wird das Regelset in mehrere Untersets geteilt. Da die Regeln-Ranking in Festival von Bedeutung ist, sieht das tatsächliche Regelset anderes aus. Hier werden die Regeln gruppenweise dargestellt um den Leser ein besseren Überblick zu verschaffen. Als Folge dessen sind fünf Untersets entstanden, die im Folgenden einzeln beschrieben werden.

- ([o] = 0)
- ([y] = u)
- ([ъ] = @)
- ([e] = e)
- ([и] = i)
- ([ѝ] = j)
- ([ж] = Z)
- ([ч] = tS)
- ([ш] = S)
- ([x] = x)

Die erste Gruppe von handgeschriebenen Regeln beinhaltet alle Grapheme der bulgarischen Sprache, die, mit kleinen Ausnahmen, immer nur einem Phonem entsprechen. Unabhängig von ihrer Position innerhalb eines Wortes: (Anlaut, Inlaut oder Auslaut) werden diese zehn Grapheme mit Ausnahme des Graphems „ж“ [Z], das in Auslautposition mit ihrer stimmlose Variante ausgesprochen wird, immer mit demselben Phonem ausgesprochen. Deshalb wird zur ursprüngliche Unterset noch ein zusätzliche Regeln hinzugefügt: ([ж] # = S) .

Das Graphem „ъ“ kommt nie im Auslaut vor. Die restlichen neun Grapheme treten in Anlaut, Inlaut oder Auslaut auf.

- ([а] # = @) ([а] = @)
- ([б] P = b') ([б] = b)
- ([ф] ю = f') ([ф] = f)
- ([ц] я = ts') ([ц] = ts)

Zur zweiten Gruppe werden fast alle Konsonanten und ein Vokal gezählt. Hierzu gehören alle Grapheme, die zwei verschiedene Phoneme repräsentieren. Hier sind exemplarisch einige Regeln dieser Gruppe aufgelistet, analog funktionieren die Grapheme „в“ [v], „г“ [g], „д“ [d], „з“ [z], „к“ [k], „л“ [l], „м“ [m], „н“ [n], „п“ [p], „р“ [r], „с“ [s] und „т“ [t]. Um eine kompaktere Darstellung der Regeln in dieser Gruppe zu erreichen, wurde wie bereits oben beschrieben die Variable P definiert. Unter dieser Variablen sind die Grapheme ю [ju] und я [ja] gespeichert. Manche Konsonantengrapheme wie б [b] werden, wenn gefolgt von [ju] oder

[ja], palatalisiert. Zum Beispiel „бял“ [b'al] *dt. weiss*, „бюст“ [b'ust] *dt. Brust*. Etwas anders verhält es sich mit den Konsonantengraphemen ф [f] und ц [ts]; sie treten in Verbindung nur mit einer der beide palatalisierungsfähigen Lauten auf. Laut des PONS Wörterbuches fürs Bulgarische tritt das Graphem ф [f] nie gefolgt von я [ja] auf. Nach [f] kann nur ю [ju] folgen. Im Gegenteil dazu tritt das Graphem ц [ts] nur gefolgt von я [ja] auf. Da die Variable P die beiden palatalisierungsfähige Laute darstellt, wird bei der Erstellung von diesen zwei Regeln weg gelassen.

- ([щ] = s t)
- ([я] = j a)
- ([ю] = j u)

Das dritte Regelnunterset beinhaltet die Grapheme, die einer Kombination von zwei Phonemen entsprechen. Die oben dargestellten Grapheme щ [St], я [ja] und ю [ju] sind eine Zusammensetzung von jeweils zwei unterschiedlichen Phonemen. Das Graphem щ [St] zum Beispiel wird durch zwei auch selbständig vorkommenden Graphemen ш [S] und т [t] dargestellt. Beispiel: „поща“ [poSta] *dt. Post*. Hier wird das Verhalten der Phoneme [ju] und [ja] als selbständige Phoneme und nicht als palatalisierungsfähige Laute modelliert. Deshalb werden die folgenden Regeln hinzugefügt: ([я] V = j a) und ([ю] V = j u), wobei mit der variable V alle Vokale im Bulgarischen definiert sind.

- (п [ь] O = p' O)
- (л [ь] O = l' O)
- (к [ь] O = k' O)

Zum vierten Unterset gehört eine Reihe von Regeln, die das Verhalten des Graphems „б“ darstellen. Dieses Graphem hat keinen phonemischen Wert und wird immer nur vor das Vokalphonem [O] geschrieben, um den vorhergehenden Konsonanten zu palatalisieren. Zum Beispiel: „кьополю“ [k'opOlu] *dt. Auberginenpaste*.

- ([д ж] = d Z)
- ([д з] = d z)

Die fünfte Gruppe von handgeschriebenen Regeln beinhaltet die Regeln, die die Aussprache der zwei Grapheme дж [dZ] und дз [dz] beschreiben wird. z.B. „джоб“ [dZob] *dt. die Tasche*, „дзън“ [dz@n] *dt. kling-kling*. Dazu gehören auch Regeln, die das Verhalten von дж [dZ] und дз [dz] in Auslaut darstellen. Das wird mit die folgenden Regeln definiert: ([д ж] # = t S) und ([д з] # = t s).

- ([т т] = t t)
 ([e e] = e e)
 ([у у] = u u)
 ([о о] = o o)
 ([о о] = o)

Das sechste Regelunterset stellt das Verhalten aller doppelten Konsonanten und Vokale in der bulgarischen Sprache vor. Die Besonderheit hier ist die Artikulation des doppelten Vokalgraphems oo [OO]. In fremden Wörtern, wie in dem Wort „кьополу“ [ˈkʰopolu] dt. *Auberginenpaste*, wird das Vokalgraphem oo [OO] als einfaches [O] und in andere Fälle wie in dem Wort „гръмоотвод“ [gr@mOOtvot] dt. *Blitzableiter* als doppeltes [OO] artikuliert. Da keine Unterscheidungsmerkmale zwischen nativen und fremden Wörter bekannt sind, wird die Regel ([oo] = oo) ein höheres Ranking bekommen.

Die letzte Gruppe ist eine Mischgruppe. Diese Gruppe beschreibt das Verhalten der Präpositionen „във“ [v@v] dt. *in* und „със“ [s@s] dt. *mit*. Dazu gehören die folgenden Regeln:

- ([в ъ в] # в = v @ f)
- ([с ъ с] # с = s @ s)

Aus der im Kapitel drei dargestellten Grammatik der bulgarischen Sprache wurden in Kapitel vier eine Reihe von Letter-To-Sound Regeln manuell erstellt. Dabei wurden keine Regeln für die Expansion von Ziffern, Abkürzungen oder Akronyme geschrieben. Für die Syllabifizierung wurden auch keine Regeln definiert. Es wurden die englischen Syllabifizierungsregeln übernommen, aber da es dadurch große Abweichung gab, wurde die Syllabifizierung zum Schluss doch herausgenommen. Es gibt aber auch morphologische Eigenschaften der bulgarischen Sprache, die im Rahmen dieser Arbeit mit einfachen Transkribierungsregeln nicht modelliert werden können. Es gibt zum Beispiel die Regel, dass bei Verben der zweiten Konjugation in 1. Person Singular und 3. Person Plural im Präsens die Weichheit für die richtige Aussprache von Bedeutung ist, z.B. „мисля“ [misl’@] dt. *denken* oder „работят“ [rabot’@t] dt. *arbeiten*. Diese Fälle werden über das Lexikon abgehandelt und somit für eine richtige Aussprache gesorgt.

Die vollständige LTS Regelnset ist in Anhang B zu finden.

Was hier erwähnt werden soll, ist dass in der Implementierung der Ausspracheregeln *weiche Konsonanten* über *ein nachfolgendes „j“ kodiert wurden* und nicht wie bereits in Kapitel 3.2. und laut SAMPA zu erwarten ist über das Diakritikum '.

5. Evaluierung

Kapitel 5 fasst die Ergebnisse der Implementierung der Ausspracheregeln im Sprachsynthesystem Festival zusammen. Hier wird die Testmethode vorgestellt und ein Verfahren zur Ermittlung der Fehlerrate der neuen Ausspracheregeln beschrieben. Die zwei Scheme-Dateien und das Testkorpus, die im Rahmen dieser Arbeit entstanden sind, werden auch detailliert diskutiert.

5.1. Methode

Im Rahmen dieser Arbeit wurden Ausspracheregeln im Sprachsynthesystem Festival für das Bulgarische entwickelt. Als Folge dessen sind zwei Scheme-Dateien entstanden. Mit den entwickelten Scheme-Skripten kann die Korrektheit nur eingeschränkt überprüft werden. Um einen Eindruck von der Einfachheit der bulgarischen Orthographie zu erhalten und die Korrektheit der in Kapitel 4 dargestellten Ausspracheregeln zu überprüfen, werden zufällig ausgesuchte Texte als ganzer Text mit Festival transkribiert. Für diesen Zweck wurde einen Referenzkorpus erstellt, indem mehrere Texte zufällig ausgesucht und zusammengesetzt wurden. Die Texte decken unterschiedliche Bereiche der bulgarischen Sprache. Insgesamt besteht der Referenzkorpus aus 2377 Tokens. Zu dieser Zusammensetzung gehören die folgenden Texte: eine Kolumnen der Zeitung „Капитал“⁹ [ka pi tal] mit 465 Tokens, zwei Artikel der Zeitung „Стандарт“¹⁰ [stan dart] mit 235 Tokens, mehrere Artikeln der Zeitschrift „Ева“¹¹ [eva] mit 767 Tokens, drei Artikel der Zeitschrift „Блясък“¹² [bl'a s@k] mit zusammen 503 Tokens und drei weitere Artikel der Zeitschrift „Психология“¹³ [psi ho lo gija] mit insgesamt 397 Tokens. Als nächster Schritt wird das Letter-To-Sound-Regelnset für die bulgarische Sprache, das im Rahmen dieser Arbeit entstanden ist auf den Referenzkorpus¹⁴ laufen gelassen. Dafür wird ein Skript verwendet, das Text auf der Shell als Input nimmt

⁹ URL (20.11.2010): <http://www.capital.bg/>

¹⁰ URL (20.11.2010): <http://www.standartnews.com/>

¹¹ URL (20.11.2010): <http://www.eva.bg/>

¹² URL (20.11.2010): <http://www.bliasak.bg/>

¹³ URL (20.11.2010): <http://www.brain-and-mind.bg/>

¹⁴ Das Skript wurde von Antje Schweitzer zur Verfügung gestellt.

und die Transkription ausgibt. Die erzeugte Struktur wird in eine Textdatei exportiert. Daraus wird nur die Transkription extrahiert mit dem „Goldstandard“¹⁵ verglichen und mit Levenshtein-Editier Algorithmus ausgewertet. Levenshtein-Editier Algorithmus ist ein Maß für die Ähnlichkeit zweier Zeichenketten. Der Abstand ist die Anzahl der Deletionen, Insertionen oder Substitutionen die erforderlich sind die Quelle Zeichenkette in die Ziel Zeichenkette zu verwandeln. Und jesto größer der Abstand, desto unterschiedlicher die Zeichenketten.

Als Endergebnis entstand eine Text Datei (siehe Abbildung 5.1.), die aus vier Spalten besteht. In der ersten Spalte steht das ursprüngliche Wort, das mit Festival transkribiert wird. Die zweite Spalte beinhaltet die Transkription dieses Wortes, die mit Hilfe der in Kapitel vier dargestellten Letter-To-Sound-Regeln, erfolgte. Spalte drei stellt den manuell annotierten Goldstandard dar. In Spalte vier stehen die Ergebnisse, die mit dem Levenshtein-Editier Algorithmus berechnet wurden.

ursprüngliche Wort	LTS-Regelnset	Referenzkorpus	LS Algorithmus
Френският	frenskijat	frenskijat	0
Vogue	vOgue	vOk	3
отпразнува	Otpraznuv@	Otpraznuv@	0
си	si	si	0
годишнина	gOdiSnin@	gOdiSnin@	0
с	s	s	0
пищен	piSten	piSten	0
бал	bal	bal	0
с	s	s	0
маски	maski	maski	0
вдъхновен	vd@hnOven	vd@hnOven	0
от	Ot	Ot	0
филма	film@	filma	1
на	n@	n@	0
Стенли	stenli	stenli	0
Кубрик	kubrik	kubrik	0
Широко	SirOkO	SirOkO	0
затворени	zatvOreni	zatvOreni	0
очи	OtSi	OtSi	0
Естествено	estestvenO	estestvenO	0
едни	edni	edni	0
от	Ot	Ot	0
влиятелните	vlijatelnite	vlijatelnite	0

¹⁵ <http://www.merriampark.com/ldperl.htm>

ЛИЧНОСТИ	litSnOsti	litSnOsti	0
ОТ	Ot	Ot	0
МОДНИЯ	mOdnija	mOdnija	0
СВЯТ	sfjat	sfjat	0

Abbildung 5.1. Auszug

Wie in Abbildung 5.1. dargestellt, ist die vierte Spalte nicht immer 0. In Zeile zwei und in Zeile 13 sind Werte unterschiedlich von null berechnet worden. Die Zahl drei in Zeile zwei steht für die Anzahl von Fehler, die für dieses Wort berechnet sind. Dasselbe gilt für Zeile 13, wo nur ein Fehler registriert ist. Die beiden Fehler sind aus unterschiedlichem Grunde entstanden. Das Wort in Zeile zwei, das in dem Referenzkorpus mit ihrer ursprünglichen Schreibweise übernommen ist, ist ein Fremdwort. In dem Regelset der bulgarischen Sprache wurden keine Regeln definiert, die die Aussprache von Fremdwörtern bestimmen können. Die Aussprache solche Wörter erfolgt mit Hilfe eines Aussprachelexikons. Da das Wort „Vogue“ nicht im Lexikon ist, wird sie auch falsch transkribiert. Der Fehler in Zeile 13 ist auf Grund der fehlenden Wortbetonung entstanden. Da die Wortbetonung im Bulgarischen nicht gekennzeichnet wird (siehe Kapitel 3.3.), wird die Modellierung von Regeln, die diese Eigenschaft der bulgarischen Sprache betreffen im Rahmen dieser Arbeit nicht diskutiert und die daraus entstandenen Fehler werden bei der Auswertung nicht berücksichtigt.

Weitere Fehler ergaben sich bei der Anwendung einiger Transkriptionsregeln. Die erste Regel, die mehrmals eine nicht korrekte Transkription lieferte, ist die Regel:

([t] CVOICED = d)

Diese Regel besagt, dass immer, wenn nach dem Graphem [t] ein Graphem der Variable CVOICED folgt, wird ein [d] transkribiert. Mit der Variable CVOICED sind die folgenden Phoneme definiert: „б“ [b], „в“ [v], „д“ [d], „з“ [z], „дз“ [dz], „ж“ [Z], „дж“ [dZ], „г“ [g]. Wie sich festgestellt hat, liefert diese Regel bei bestimmten Konstellationen nicht das korrekte Ergebnis. Wenn zum Beispiel nach dem Graphem „т“ [t] das Phonem [v] folgt, wird das Graphem „т“ [t] als ihre stimmhafte Variante ausgesprochen und das ist nicht gewünscht. Deshalb wurde die folgende Regel (# [т в] = т в) hinzugefügt und die Regel ([t] CVOICED = d) ein niedrige Ranking zugewiesen.

Das Wort „така“ [ta ka] wurde auch falsch [ta k@] transkribiert. Eine falsche Transkription wird auch bei weiteren Wörtern folgen, bei denen die Wortbetonung auf die

letzte Silbe fällt. Da keine Regel in dem Fall modelliert werden kann, wird das Wort in dem Lexikon aufgenommen werden.

5.2. Auswertung

Es existieren unterschiedlichen Auswertungsverfahren, aber ein Maß der Evaluation ist die Wortfehlerrate, mit WER (engl. word error rate) abgekürzt. Sie wird meist bei der Evaluation von Spracherkennungssysteme verwendet. Zur Berechnung dieses Maßes werden Referenzsatz und Hypothese einander gegenüber gestellt und deren minimale Editierdistanz berechnet, die auch als Levenshtein-Editierdistanz¹⁶ bezeichnet wird. „Dazu werden Referenz und Hypothese so aufeinander abgebildet, dass die Summe aller Fehler minimiert wird“ [Carstensen, 2001]. Dabei werden die Fehler in drei Kategorien eingeteilt: Ersetzungen (ein korrektes Wort wurde durch ein falsches ersetzt), hier mit SUB (substitute) abgekürzt, Auslassungen (ein Wort wurde weggelassen), die englische Entsprechung ist DELETE, hier mit DEL abgekürzt und Einfügungen (ein zusätzliches Wort wurde eingefügt), wofür das englische Übersetzung INSERT, hier mit INS abgekürzt, übernommen [Pfister / Kaufmann, 2008]. Anhand dieser drei Fehlerklassen wird die Wortfehlerrate für eine Folge von n Wörter nach Carstensen [Carstensen, 2001] wie folgt definiert:

$$WER = 100 \times \frac{N_{SUB} + N_{INS} + N_{DEL}}{N}$$

wobei N die Gesamtzahl der Wörter des Referenzsatzes ist.

Die Wortfehlerrate kann auch auf die Evaluierung von LTS-Regeln übertragen werden, man berechnet dann analog die Phonemfehlerrate. Die Phonemfehlerrate wurde für das gesamte Korpus berechnet und die Wortfehlerrate nur für 1120 Tokens. Fehler ergaben sich vor allem aus der fehlenden Behandlung von Ziffern und Abkürzungen/Akronymen. Aber da hier keine Methode dafür definiert wurde, werden solche Fehler bei der Auswertung nicht berücksichtigt. Bei der Berechnung der Phonemfehlerrate ergaben sich die folgenden Ergebnisse: von insgesamt 2377 Tokens wurden nur 93 (ca. 4%) falsch transkribiert. Eine Zusammenfassung der enthaltenen Fehler sieht wie folgt aus: von die 93 (ca. 4%) falsch transkribierte Tokens entfallen allein 39 (ca. 1,6 %) auf Fremdwörter. Von den 54 (ca. 2,3 %)

¹⁶ URL (13.11.10): <http://www.levenshtein.de/>

verbleibenden Fehler, entfallen 48 (ca. 2%) auf die Wortbetonung (d.h. wenn die Wortbetonung auf die letzte Silbe fällt und das ist auch zufällig das Graphem „a“ [a], dann wird sie nicht als [a] transkribiert und nicht wie in dem LTS-Regelnsatz für Bulgarische bereits definiert als [a]). Wie bereits in Kapitel 3.3. dargestellt ist die Wortbetonung der bulgarischen Sprache ein Punkt worüber hier nicht diskutiert wird und es wurden auch keine Regeln zur Behandlung dieses Phänomens entwickelt. Die verbleibenden 6 Fehler entfallen auf die Morphologie.

Das zeigt, dass die handgeschriebenen Regeln für die bulgarische Sprache fast alle Laute korrekt darstellen. Natürlich handelt es sich hierbei nur um eine stichprobenartige Untersuchung. Die Behandlung von Ziffern und Abkürzungen sowie die Disambiguierung von Homographen die Syllabifizierung und die Wortbetonung wurden nicht implementiert was zu vielen Fehlern führte. Aber das Thema dieser Arbeit ist die Erstellung der Ausspracheregeln und die oben genannten Phänomene keine Einfluss darauf haben, kann ich sagen, dass diese mit kleinen Ausnahmen sehr gut gelungen ist.

6. Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war die Erstellung von Ausspracheregeln für das Bulgarische im Sprachsynthesystem Festival. Zu diesem Zweck wurde im dritten Kapitel zunächst eine allgemeine Einführung in der Orthographie der bulgarischen Sprache gegeben, die bereits auf einige spezielle Probleme einging. Da die Bulgarische Sprache nicht so bekannt, nicht oft beschrieben ist und nur von 8 Millionen Menschen gesprochen wird, wurden die Eigenschaften des Bulgarischen sehr umfangreich besprochen. Die Beschreibung der einzelnen linguistischen Merkmale richtete sich dabei immer nach den Erfordernissen eines Sprachsynthesystems. Darauf basierend wurde in Kapitel vier ein Set von Ausspracheregeln für das Bulgarische entwickelt und für das Festival adaptiert. Basis für die Evaluierung dieser Ausspracheregeln sind zufällig ausgewählte Sätze von unterschiedlichen Internetquellen, welche manuell phonetisch transkribiert wurden. In Kapitel fünf wurde die Vorgehensweise beschrieben und ein Auswertungsverfahren dargestellt. Hier wurden auch die Ergebnisse präsentiert, die noch einmal die Hypothese bestätigt haben, dass die bulgarische Orthographie sehr „synthese-freundlich“ ist, da die Abbildung von Orthographie auf Aussprache sehr regelmäßig ist. Es zeigte sich, dass die transkribierten Laute fast alle korrekt waren. Mit nur 93 Fehlern in einem Text von 2377 Tokens wurde eine hohe Präzision erreicht.

Im Bereich der Sprachsynthese ist für das Bulgarische noch viel Arbeit nötig. Zu erwähnen wäre etwa ein umfangreiches maschinenlesbares Aussprachelexikon. Zur weiteren Arbeit mit dem Festival System wäre die Integration eines Audioinventars zur Synthese sowie die Unterstützung von Unicode notwendig. Es gibt noch eine Reihe von Probleme bei die ist noch Grundarbeit zu leisten, wie zum Beispiel bei der wissenschaftlichen Beschreibung der Wortbetonung (falls es solche Studien gibt, ist mir nicht gelungen- trotz umfangreichen Recherchen – sie zu finden). Es zeigte sich auch, dass es an aktuellen Ressourcen für die Bearbeitung computerlinguistische Probleme mangelt.

Literaturverzeichnis

BAN (2005): Nov pravopisen rechnik na balgarskijat ezik. Sofia: BAN

Black, A. / Taylor, P. / Caley, R. (2001): *The Festival Speech Synthesis System*.
System documentation, Edition 1.4 for Festival Version 1.4.2. Stand 25th July 2001.
http://www.festvox.org/docs/manual-1.4.2/festival_toc.html

Black, A. / Lenzo, K. (2007): *Building Synthetic Voices*
http://festvox.org/festvox/festvox_toc.html

Breitenbücher, M. (1997): Textverarbeitung zur deutschen Version des Festival Text-To-Speech Synthese Systems. IMS Phonetik

Carstensen, K.-U. (2001): *Computerlinguistik und Sprachtechnologie: Eine Einführung*.
Berlin: Spektrum Akademischer Verlag

Dutoit, T. (1997): *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers

IPA Webseite: URL (16.11.2010): <http://www.langsci.ucl.ac.uk/ipa/>

Klatt, D. H. (1987): Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America* 82:737-793.

Pfister, B. / Kaufmann, T. (2008): *Sprachverarbeitung Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Berlin: Springer-Verlag

PONS Neues Universalwörterbuch Bulgarisch-Deutsch (2006). Stuttgart: Ernst Klett Sprachen GmbH

Radeva, V. (2003): *Bulgarische Grammatik: morphologisch-syntaktische Grundzüge*.
Hamburg: Buske

SAMPA Webseite: URL (16.11.2010): <http://www.phon.ucl.ac.uk/home/sampa/>

Simeonova, R. (1988): *Grundzüge einer kontrastiven Phonetik deutsch/bulgarisch*. Sofia: Nauka i Izkustvo

Schweitzer, A. (2008): Unterlagen des Seminars Sprachsynthese im Sommersemester 2008 am Institut für maschinelle Verarbeitung der Universität Stuttgart

Shanon / Weaver (1949): *The mathematical theory of communication*. Urbana: University of Illinois Press

Sproat, R. (1998): *Multilingual Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers

Stoyanov, S. (1999): *Gramatika na balgarskija knizhoven ezik: fonetika i morfologija*. Veliko Tarnovo: Abagar

Taylor, P. (2007): Text-to-Speech Synthesis. University of Cambridge. draft

Tilkov, D. (1998): *Gramatika na savremennija balgarski knizhoven ezik*. Sofia: Abagar Publishing

Unicode Webseite: URL (16.11.2010): <http://unicode.org/>

Vatov, V. (1995): *Phonetika i leksikologija na balgarskijat ezik*. Veliko Tarnovo: Abagar

Wahlster, W. (2000): *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer

Walter, H. (1987): *Lehrbuch der bulgarischen Sprache*. Leipzig: Enzyklopädie

Abkürzungsverzeichnis

Abb.	Abbildung
bzw.	beziehungsweise
CT	Artikulationsart / Consonant Type
CP	Artikulationsstelle / Consonant Place
CVOX	Stimmhaftigkeit - Stimmlosigkeit / Consonant Voicing
d.h.	das heißt
Dt.	Deutsch
DEL	Auslasung / Delete
etc	und so weiter
evtl.	eventuell
inkl.	inklusiv
IPA	International Phonetic Alphabet
IMS	Institute für Maschinelle Sprachverarbeitung
INS	Einfügung / Insert
LTS	Letter-To-Sound
SAMPA	Speech Assessment Methods Phonetic Alphabet
sog.	Sogenannt
SUB	Ersetzung / Substitute
TTS	Text-To-Speech
vgl.	vergleiche
VC	Vokal oder Konsonant / Vowel or Consonant
VL	Vokallänge / Vowel Length
VH	Vokalhöhe / Vowel Height
VF	Artikulationsstelle Vokale / Vowel Frontness
VOC	Gaumen-Zunge Abstand
VLR	Labial - Nicht Labial
WER	Wortfehlerrate / Word Error Rate
z.B.	zum Beispiel

Anhang A

Phonset

```
-----ims_bulgarian_phonset.scm-----
;;Bulgarian PhoneSet
;;Stoyka Dachenska 2010
;;Encoding: iso-8859-1

(defPhoneSet
  bulgarian_sampa
  (
    ;; Phone Features
    ;; vowel or consonant
    (vc + -)
    ;; vowel length: short long diphthong schwa
    (vl s l d a 0)
    ;; vowel height: high mid low
    (vh 1 2 3 -)
    ;; vowel frontness: front mid back
    (vf 1 2 3 -)
    ;; lip rounding
    (vlr + -)
    ;; vowel openness
    (voc + -)
    ;; consonant type: plosive fricative affricative nasal lateral vibrant glide
    (ct p f a n l v g 0)
    ;; place of articulation: bilabial/bl/ labio-dental/ld/ alveo-dental/ad/ alveolar/al/
    ;;                               post- alveolar/pa/ velar/vl/
    (cp bl ld ad al pa vl 0)
    ;; consonant voicing: voiced voiceless
    (cvox + -)
  )
  ;; Phone set members
  (
    (# - 0 - - - - 0 0 -)

    ;;Vowels
    ;; VC VL VH VF VLR VOC CT CP CVOX
    (i + s 1 1 - - 0 0 -)
    (e + s 2 1 - + 0 0 -)
    (a + s 3 2 - + 0 0 -)
    (@ + s 2 2 - + 0 0 -)
    (O + s 2 3 + + 0 0 -)
    (u + s 1 3 + - 0 0 -)
  )
)
```

::Consonants

:: VC VL VH VF VLR VOC CT CP CVOX

(p	-	0	-	-	-	-	p	bl	-)
(b	-	0	-	-	-	-	p	bl	+))
(t	-	0	-	-	-	-	p	ad	-)
(d	-	0	-	-	-	-	p	ad	+))
(k	-	0	-	-	-	-	p	vl	-)
(g	-	0	-	-	-	-	p	vl	+))
(f	-	0	-	-	-	-	f	ld	-)
(v	-	0	-	-	-	-	f	ld	+))
(s	-	0	-	-	-	-	f	ad	-)
(z	-	0	-	-	-	-	f	ad	+))
(S	-	0	-	-	-	-	f	pa	-)
(Z	-	0	-	-	-	-	f	pa	+))
(x	-	0	-	-	-	-	f	vl	-)
(ts	-	0	-	-	-	-	a	ad	-)
(dz	-	0	-	-	-	-	a	ad	+))
(tS	-	0	-	-	-	-	a	pa	+))
(dZ	-	0	-	-	-	-	a	pa	+))
(m	-	0	-	-	-	-	n	bl	-)
(n	-	0	-	-	-	-	n	al	-)
(l	-	0	-	-	-	-	l	ad	-)
(r	-	0	-	-	-	-	v	al	-)
(j	-	0	-	-	-	-	g	pa	-)

)

)

(PhoneSet.silences '(#))

-----ims_bulgarian_phones.scm-----

Anhang B

Letter-To-Sound-Regeln

-----ims_bulgarian_lts.scm-----

```
;;Letter-To-Sound Rules for Bulgarian
;;Stoyka Dachenska 2010
;;Encoding: iso-8859-1
```

```
(lts.ruleset
;; Name of rule set
bulgarian
```

```
;;Variables used in the LTS Rules for Bulgarian
(
(V a e и o y ъ ) ;; all vowels in Bulgarian
(CVOICED б в д з дз ж дж г ) ;; voiced consonants
(CVOICELESS п ф т с ц ш ч к х ) ;; voiceless consonants
(P я ю );; palatalised
)
```

```
;; Letter-To-Sound Rules for Bulgarian
```

```
(
([ o ] = O )
([ y ] = u )
([ ъ ] = @ )
([ e ] = e )
([ и ] = i )
([ ѝ ] = j )

([ ж ] # = S )
([ ж ] CVOICELESS = S )
([ ж ] = Z )

([ ч ] = tS)

([ ш ] = S )

([ x ] = h )

([ a ] # = @ )
```

([a] = a)

;; ([б] P = b')

([б я] = b j a)

([б ю] = b j u)

([б ч] = p t S)

([б с] = p s)

([б ш] = p S t)

([б] CVOICELESS = p)

([б] # = p)

([б] = b)

([вѣв] # v = v@f)

([в] # k = f)

([в] # п = f)

([в] # p = f)

([в] # л = f)

([в] # м = f)

([в] # н = f)

([в] # и = f)

([в] # т = f)

([в] # б = f)

([в] # с = f)

([в] # ч = f)

([в] # ш = f)

([в] # ц = f)

([в т] = f t)

([в к] = f k)

(# [в м] = f m)

([в я] = v j a)

([в] CVOICELESS = f)

([в] # = f)

([в] = v)

([г я] = g j a)

([г ю] = g j u)

;; ([г е] = g j e)

;; ([г и] = g j i)

([г] # = k)

([г] CVOICELESS = k)

([г] = g)

;; ([д] P = d')

([д я] = d j a)

([д ю] = d j u)

([д] # = t)

([д] CVOICELESS = t)

([д] = d)

;; ([з] P = z')

([з я] = z j a)

([з ю] = z j u)
([з] CVOICELESS = s)
([з] # = s)
([з] = z)

;;([к] P = k')
([к я] = k j a)
([к ю] = k j u)

;;([к] CVOICED = g)
([к] = k)

;;([л] P = l')
([л я] = l j a)
([л ю] = l j u)
;;([л] e = l')
;;([л] и = l')
([л] = l)

;;([м] P = m')
([м я] = m j a)
([м ю] = m j u)
([м] = m)

;;([н] P = n')
([н я] = n j a)
([н ю] = n j u)
;;([н] k = N)
;;([н] r = N)
([н] = n)

;;([п] P = p')
([п я] = p j a)
([п ю] = p j u)
([п б] = b)
([п] = p)

([р я] = r j a)
([р] = r)

;;([с] P = s')
([с] # б = z)
([с] # д = z)
([с я] = s j a)
([с ю] = s j u)
([с ъ с] # c = s@z)
;;([с] CVOICED = z)
([с в] = s f)
([с б] = z b)
([с д] = z d)

([c] = s)

;;([T] P = t')

([T я] = t j a)

([T ю] = t j u)

(# [T B] = t v)

;;([T] CVOICED = d)

([T] = t)

([ф ю] = f j u)

([ф] CVOICED = v)

([ф] = f)

([ц я] = t s j a)

([ц] CVOICED = d z)

([ц] = t s)

([ш] = S t)

(# [я] = j a)

([я] V = j a)

([я] = j a)

(# [ю] = j u)

([ю] V = j u)

([ю] = j u)

([д ж] # = t S)

([д ж] CVOICELESS = t S)

([д ж] = d Z)

([д з] # = t s)

([д з] CVOICELESS = t s)

([д з] = d z)

([г ь O] = g j O)

([к ь O] = k j O)

([л ь O] = l j O)

([с ь O] = s j O)

([ф ь O] = f j O)

([T T] = t t)

([H H] = n n)

([И И] = i i)

([e e] = e e)

([a a] = a a)

([y y] = u u)

([o o] = O)

([o o] = O O)

([c т м] = s m) ;; астма [asma] - [t] wird nicht ausgesprochen!!!
([тск] = t s k)

)
)

-----ims_bulgarian_lts.scm-----

Anhang C

Ergebnisse

сиси скитницата императрица

Елисавета Австрийска, по известна като императрица Сиси, е красива, аристократична, омъжена за владетеля на най могъщата империя и нещастна. За късмет по нейно време професията папараци не е съществувала и няма кадри, в които да изглежда разстроена или разчорлена след езда.

Животът на Сиси започва като приказка. Родена е със синя кръв. Баща ѝ херцог Максимилиан, който е от баварския владетелски род, има деца. Втората му дъщеря Сиси се появява на бял свят през навръх Рождество и с поникнало в устата зъбче. Неделя е денят на късметлиите, така че всичко вещае щастлива съдба за малката принцеса. В детството си тя не се отличава с особена хубост. Като гледа кръглото лице и прекалено свободните маниери на дъщеря си, майка ѝ херцогиня Людовика, въздиша тъжно. Не, Сиси не може да се мери с по голямата си сестра Хелена, която не само е с благородно бледи черти, но и впечатлява със своя ум и съдържаност.

sisi skitnitsat@ imperatrits@

elisavet@ afstrijsk@ pO izvestn@ katO imperatrits@ sisi e krasiv@ aristOktratitSn@
Om@Zen@ z@ vladetelja na naj mOg@Stat@ imperija i neStastn@ z@ k@smet pO nejnO
vreme prOfesijat@ paparatsi ne e s@Stestvuval@ i njam@ kadri v kOitO da izgleZd@
rasstrOen@ ili rastSOrlen@ slet ezda ZivOt@t n@ sisi zapOtSv@ katO prikask@ rOden@ e
s@s sinja kr@f baSta j hertsOk maksimilian kOjtO e Ot bavarskija vladetelski rOt ima detsa
ftOrat@ mu d@Sterja sisi se pOjavjav@ n@ bjal sfjat n@vr@h rOZdestvO i s pOniknalO v
ustat@ z@ptSe nedelja e denjat n@ k@smetliite taka tSe fsitSkO veStae Stastliv@ s@dba
z@ malkat printses@ f detstvOtO si tja ne se OtlitSav@ s OsOben@ hubOst katO gleda
kr@glOtO litse i prekalenO sfObOdnite manieri na d@Sterja si majk@ j hertsOginja
ljudOvik@ v@zdiS@ t@ZnO ne sisi ne mOZe d@ se meri s pO gOljamat@ si sestra helen@
kOjatO e ne samO s blagOrOdnO beli tSerti nO i fpetSatljav@ s@s sfOja um i zd@rZanOst

сиси	sisi	sisi 0
скитницата	skitnitsat@	skitnitsat@ 0
императрица	imperatrits@	imperatrits@ 0
Елисавета	elisavet@	elisavet@ 0
Австрийска	afstrijsk@	afstrijsk@ 0

по	pO	pO 0	
известна	izvestn@	izvestn@ 0	izvestn@ 0
като	katO	katO 0	
императрица	imperatrits@	imperatrits@ 0	imperatrits@ 0
Сиси	sisi	sisi 0	
е	е	е 0	
красива	krasiv@	krasiv@ 0	krasiv@ 0
аристократична	aristOkratitSn@	aristOkratitSn@ 0	aristOkratitSn@ 0
омъжена	Om@Zen@	Om@Zen@ 0	Om@Zen@ 0
за	z@	z@ 0	
владетеля	vladetelja	vladetelja 0	vladetelja 0
на	n@	n@ 0	
най	naj	naj 0	
могъщата	mOg@Stat@	mOg@Stat@ 0	mOg@Stat@ 0
империя	imperija	imperija 0	imperija 0
и	i	i 0	
нещастна	neStastn@	neStastn@ 0	neStastn@ 0
За	z@	z@ 0	
късмет	k@smet	k@smet 0	k@smet 0
по	pO	pO 0	
нейно	nejnO	nejnO 0	
време	vreme	vreme 0	
професията	prOfesijat@	prOfesijat@ 0	prOfesijat@ 0
папараци	pararatsi	pararatsi 0	pararatsi 0
не	ne	ne 0	
е	е	е 0	
съществувала	s@Stestvuval@	s@Stestvuval@ 0	s@Stestvuval@ 0
и	i	i 0	
няма	njam@	njam@ 0	
кадри	kadri	kadri 0	
в	f	f 0	
които	kOitO	kOitO 0	
да	d@	d@ 0	
изглежда	izgleZd@	izgleZd@ 0	izgleZd@ 0
разстроена	rasstrOen@	rasstrOen@ 0	rasstrOen@ 0
или	ili	ili 0	
разчорлена	rastSOrlen@	ractSOrlen@ 0	ractSOrlen@ 0
след	slet	slet 0	
езда	ezda	ezda 0	
Животът	ZivOt@t	ZivOt@t 0	ZivOt@t 0
на	n@	n@ 0	
Сиси	sisi	sisi 0	
започва	zapOtSv@	zapOtSv@ 0	zapOtSv@ 0
като	katO	katO 0	
приказка	prikask@	prikask@ 0	prikask@ 0
Родена	rOden@	rOden@ 0	rOden@ 0
е	е	е 0	
със	s@s	s@s 0	
синя	sinja	sinja 0	
кръв	kr@f	kr@f 0	
Баща	baSt@	baSta 1	

й	j	j 0	
херцог	hertsOk	hertsOk 0	
Максимилиан	maksimilian	maksimilian 0	
който	kOjtO	kOjtO 0	
е	e	e 0	
от	Ot	Ot 0	
баварския	bavarskija	bavarskija 0	
владетелски	vladetelski	vladetelski 0	
род	rOt	rOt 0	
има	im@	ima 1	
деца	detsa	detsa 0	
Втората	ftOrat@	ftOrat@ 0	
му	mu	mu 0	
дъщеря	d@Sterja	d@Sterja 0	
Сиси	sisi	sisi 0	
се	se	se 0	
появява	pOjavjav@	pOjavjav@ 0	
на	n@	n@ 0	
бял	bjal	bjal 0	
свят	sfjat	sfjat 0	
ппрез	pres	pres 0	
навръх	navr@h	navr@h 0	
Рождество	rOZdestvO	rOZdestvO 0	
и	i	i 0	
с	s	s 0	
поникнало	pOniknalO	pOniknalO 0	
в	f	f 0	
устата	ustat@	ustat@ 0	
зъбче	z@ptSe	z@ptSe 0	
Неделя	nedelja	nedelja 0	
е	e	e 0	
денят	denjat	denjat 0	
на	n@	n@ 0	
късметлиите	k@smetliite	k@smetliite 0	
така	taka	taka 0	
че	tSe	tSe 0	
всичко	fsitSkO	fsitSkO 0	
вещае	veStae	veStae 0	
щастлива	Stastliv@	Stastliv@ 0	
съдба	s@db@	s@dba 1	
за	z@	z@ 0	
малката	malkat@	malkat@ 0	
принцеса	printses@	printses@ 0	
В	f	f 0	
детството	detstvOtO	detstvOtO 0	
си	si	si 0	
тя	tja	tja 0	
не	ne	ne 0	
се	se	se 0	
отличава	OtlitSav@	OtlitSav@ 0	
с	s	s 0	

особена	OsOben@	OsOben@ 0
хубост	hubOst	hubOst 0
Като	katO	katO 0
гледа	gled@	gled@ 0
кръглото	kr@glOtO	kr@glOtO 0
лице	litse	litse 0
и	i	i 0
прекалено	prekalenO	prekalenO 0
свободните	sfObOdnite	sfObOdnite 0
маниери	manieri	manieri 0
на	n@	n@ 0
дъщеря	d@Sterja	d@Sterja 0
си	si	si 0
майка	majk@	majk@ 0
й	j	j 0
херцогиня	hertsOginja	hertsOginja 0
Людовика	ljudOvik@	ljudOvik@ 0
въздиша	v@zdiS@	v@zdiS@ 0
тъжно	t@ZnO	t@ZnO 0
Не	ne	ne 0
Сиси	sisi	sisi 0
не	ne	ne 0
може	mOZe	mOZe 0
да	d@	d@ 0
се	se	se 0
мери	meri	meri 0
с	s	s 0
по	pO	pO 0
голямата	gOljamat@	gOljamat@ 0
си	si	si 0
сестра	sestra	sestra 0
Хелена	helen@	helen@ 0
която	kOjatO	kOjatO 0
не	ne	ne 0
само	samO	samO 0
е	e	e 0
с	s	s 0
благородно	blagOrOdnO	blagOrOdnO 0
бледи	bledi	bledi 0
черти	tSerti	tSerti 0
но	nO	nO 0
и	i	i 0
впечатлява	fpetSatljav@	fpetSatljav@ 0
със	s@s	s@z 0
своя	sfOja	sfOja 0
ум	um	um 0
и	i	i 0
сдържаност	zd@rZanOst	zd@rZanOst 0

ани лейбовиц интимно

Гениалната фотографка на нашето време Ани Лейбовиц пазеше ревниво вратата към своя личен живот. Съвсем скоро в Националната портретна галерия в Лондон Лейбовиц счупи ключалката и пусна публиката в най интимните кътчета на душата си.

Легенди се носят за работата ѝ и нейната свръхвзискателност. В документалния филм за живота ѝ, направен от сестра ѝ Барбара и излъчен у нас по телевизия, главната редакторка на американския Vogue Анна Уинтур споделя Лейбовиц ни подлудява, разбира се. А бюджетът е понятие, което въобще не влиза в нейното съзнание.

Циниците биха казали, че тя подкрепя голямата фабрика на знаменитостите, истината е, че тя е техен коментатор. Някои от нейните портрети имат спираща дъха, нежна интимност.

В последните няколко години Лейбовиц минава изцяло към дигитална фотография. Обвиняват я, че си играе много с постпродукцията на фотографиите си. Според нея това твърдение е пресилено, въпреки признанието, че обича този тип нереална реалност.

Висока, с очила и маниер да командва непринудено, Лейбовиц почти винаги е облечена в черно и ръкомаха широко с дългите си ръце, смее се гърлено и умее да се самоиронизира.

ани lejbOvits intimnO

genialnat@ fOtOgrafk@ n@ naSetO vreme ani lejbOvits pazeSe revnivO vratat@ k@m sfOja litSen ZivOt s@fsem skOrO f natsiOnalnat@ pOrtretn@ galerija f lOndOn lejbOvits stSupi kljutSalkat@ i pugn@ publikat@ f naj intimnite k@ttSet@ n@ duSat@ si

legendi se nOsjat z@ rabOtat@ j i nejnat@ sfr@hvziskatelnOst v dOkumentalnija film z@ ZivOt@ j napraven Ot sestra j barbar@ i izl@tSen u nas pO televizija glavnat@ redaktOrk@ n@ amerikanskija vOk ann@ uintur spOdelja lejbOvits ni pOdludjav@ razbir@ se a bjudZet@ e pOnjatie kOetO v@OpSte ne vliz@ f nejnOtO s@znanie

tsinitsite bih@ kazali tSe tja pOtkrepja gOljamat@ fabrik@ n@ znamenitOstite istinat@ e tSe tja e tehen kOmentatOr njakOi Ot nejnite pOrtreti имат spiraSt@ d@h@ neZn@ intimnOst

f pOslednite njakOlko gOdini lejbOvits minav@ istsjalO k@m digitaln@ fOtOgrafija Obvinjavat ja tSe si igrae mnOgO s pOstprOduksijat@ n@ fOtOgrafiite si spOret neja tOva tv@rdenie e presileno v@preki priznanietO tSe ObitS@ tOzi tip nerealn@ realnOst

visOk@ s OtSila i manier d@ kOmandv@ neprinudenO lejbOvits pOtSti vinagi e ObletSen@ f tSernO i r@kOmah@ SirOkO s d@lgite si r@tse smee se g@rlenO i umee d@ se samOirOnizir@

ани	ани	ани 0
лейбовиц	lejbOvits	lejbOvits 0
интимно	intimnO	intimnO 0
Гениалната	genialnat@	genialnat@ 0

фотографка	fOtOgrafk@	fOtOgrafk@ 0
на	n@	n@ 0
нашето	naSetO	naSetO 0
време	vreme	vreme 0
Ани	ani	ani 0
Лейбовиц	lejbOvits	lejbOvits 0
пазеше	pazeSe	pazeSe 0
ревниво	revnivO	revnivO 0
вратата	vratat@	vratat@ 0
към	k@m	k@m 0
своя	sfOja	sfOja 0
личен	litSen	litSen 0
живот	ZivOt	ZivOt 0
Съвсем	s@fsem	s@fsem 0
скоро	skOrO	skOrO 0
в	f	f 0
Националната	natsiOnalnat@	natsiOnalnat@ 0
портретна	pOrtretn@	pOrtretn@ 0
галерия	galerija	galerija 0
в	f	f 0
Лондон	lOndOn	lOndOn 0
Лейбовиц	lejbOvits	lejbOvits 0
счупи	stSupi	stSupi 0
ключалката	kljutSalkat@	kljutSalkat@ 0
и	i	i 0
пусна	pusn@	pusn@ 0
публиката	publikat@	publikat@ 0
в	f	f 0
най	naj	naj 0
интимните	intimnite	intimnite 0
кътчета	k@ttSet@	k@ttSet@ 0
на	n@	n@ 0
душата	duSat@	duSat@ 0
си	si	si 0
Легенди	legendi	legendi 0
се	se	se 0
носят	nOsjat	nOsjat 0
за	z@	z@ 0
работата	rabOtat@	rabOtat@ 0
й	j	j 0
и	i	i 0
нейната	nejnat@	nejnat@ 0
свърхвзискателност	sfr@hvziskatelnOst	sfr@hvziskatelnOst 0
В	f	f 0
документалния	dOkumentalnija	dOkumentalnija 0
филм	film	film 0
за	z@	z@ 0
живота	ZivOt@	ZivOt@ 0
й	j	j 0
направен	napraven	napraven 0
от	Ot	Ot 0

сестра	sestra	sestra 0	
й	j	j 0	
Барбара	barbar@	barbar@ 0	
и	i	i 0	
излъчен	izl@tSen	izl@tSen 0	
у	u	u 0	
нас	nas	nas 0	
по	pO	pO 0	
телевизия	televizija	televizija 0	
главната	glavnat@	glavnat@ 0	
редакторка	redaktOrk@	redaktOrk@ 0	
на	n@	n@ 0	
американския	amerikanskija	amerikanskija 0	
Vogue	vOgue	vOk 3	
Анна	ann@	ann@ 0	
Уинтур	uintur	uintur 0	
споделя	spOdelja	spOdelja 0	
Лейбовиц	lejbOvits	lejbOvits 0	
ни	ni	ni 0	
подлудява	pOdludjav@	pOdludjav@ 0	
разбира	razbir@	razbir@ 0	
се	se	se 0	
А	a:	a 1	
бюджетът	bjudZet@t	bjudZet@t 0	
е	e	e 0	
понятие	pOnjatie	pOnjatie 0	
което	kOetO	kOetO 0	
въобще	v@OpSte	v@OpSte 0	
не	ne	ne 0	
влиза	vliz@	vliz@ 0	
в	f	f 0	
нейното	nejnOtO	nejnOtO 0	
съзнание	s@znanie	s@znanie 0	
Циниците	tsinitsite	tsinitsite 0	
биха	bih@	bih@ 0	
казали	kazali	kazali 0	
че	tSe	tSe 0	
тя	tja	tja 0	
подкрепя	pOtkrepja	pOtkrepja 0	
голямата	gOljamat@	gOljamat@ 0	
фабрика	fabrik@	fabrik@ 0	
на	n@	n@ 0	
знаменитостите	znamenitOstite	znamenitOstite 0	
истината	istinat@	istinat@ 0	
е	e	e 0	
че	tSe	tSe 0	
тя	tja	tja 0	
е	e	e 0	
техен	tehen	tehen 0	
коментатор	kOmentatOr	kOmentatOr 0	
Някои	njakOi	njakOi 0	

от	Ot	Ot 0	
нейните	nejnite	nejnite 0	
портрети	pOrtreti	pOrtreti 0	
имат	imat	imat 0	
спираща	spiraSt@	spiraSt@ 0	
дъха	d@h@	d@h@ 0	
нежна	neZn@	neZn@ 0	
интимност	intimnOst	intimnOst 0	
В	f	f 0	
последните	pOslednite	pOslednite 0	
няколко	njakOlko	njakOlko 0	
години	gOdini	gOdini 0	
Лейбовиц	lejbOvits	lejbOvits 0	
минава	minav@	minav@ 0	
изцяло	istsjalO	istsjalO 0	
към	k@m	k@m 0	
дигитална	digitaln@	digitaln@ 0	
фотография	fOtOgrafija	fOtOgrafija 0	
Обвиняват	Obvinjavat	Obvinjavat 0	
я	ja	ja 0	
че	tSe	tSe 0	
си	si	si 0	
играе	играе	играе 0	
много	mnOgO	mnOgO 0	
с	s	s 0	
постпродукцията	pOstprOduksijat@	pOstprOduksijat@ 0	
на	n@	n@ 0	
фотографиите	fOtOgrafiite	fOtOgrafiite 0	
си	si	si 0	
Според	spOret	spOret 0	
нея	neja	neja 0	
това	tOva	tOva 0	
твърдение	tv@rdenie	tv@rdenie 0	
е	e	e 0	
пресилено	presilenO	presilenO 0	
въпреки	v@preki	v@preki 0	
признанието	priznanietO	priznanietO 0	
че	tSe	tSe 0	
обича	ObitS@	ObitS@ 0	
този	tOzi	tOzi 0	
тип	tip	tip 0	
нереална	nerealn@	nerealn@ 0	
реалност	realnOst	realnOst 0	
Висока	visOk@	visOk@ 0	
с	s	s 0	
очила	OtSila	OtSila 0	
и	i	i 0	
маниер	manier	manier 0	
да	d@	d@ 0	
командва	kOmandv@	kOmandv@ 0	
непринудено	neprinudenO	neprinudenO 0	

Лейбовиц	lejbOvits	lejbOvits 0
почти	pOtSti	pOtSti 0
винаги	vinagi	vinagi 0
е	е	е 0
облечена	ObletSen@	ObletSen@ 0
в	f	f 0
черно	tSernO	tSernO 0
и	i	i 0
ръкомаха	r@kOmah@	r@kOmah@ 0
широко	SirOkO	SirOkO 0
с	s	s 0
дългите	d@lgite	d@lgite 0
си	si	si 0
ръце	r@tse	r@tse 0
смее	smee	smee 0
се	se	se 0
гърлено	g@rlenO	g@rlenO 0
и	i	i 0
умее	umee	umee 0
да	d@	d@ 0
се	se	se 0
самоиронизира	samOirOnizir@	samOirOnizir@ 0

мадона

Тя пристига в Ню Йорк с култовата реплика, отправена към таксиметровия шофьор Откарайте ме в центъра на събитията. Мадона е истинското й име, не псевдоним, както мислят много хора, и освен това тя няма италиански корени. Майка й, също Мадона, е от френско канадски произход. Днес Мадона печели по милиона долара на година, което означава, че състоянието й е около милиона. И е все още в центъра на събитията.

madOn@

tja pristig@ s kultOvat@ replik@ Otpraven@ k@m taksimetrOvija SOfjOr Otkarajte me f tsent@r@ n@ s@bitijat@ madOn@ e istinskOtO j ime ne psevdOnim kaktO misljat mnOgO hOra i Osfen tOva tja njam@ italianski kOreni majk@ j s@StO madeOn@ e Ot frenskO kanatski prOishOt dnes madOn@ petSeli pO miliOn@ dOlar@ n@ gOdin@ kOetO OznatSav@ tSe s@stOjanietO j e OkOlO miliOn@ i e f se OSte f tsent@r@ n@ s@bitijat@

мадона	madOn@	madOn@ 0
Тя	tja	tja 0
пристига	pristig@	pristig@ 0
с	s	s 0
култовата	kultOvat@	kultOvat@ 0
реплика	replik@	replik@ 0
отправен	Otpraven@	Otpraven@ 0
към	k@m	k@m 0
таксиметровия	taksimetrOvija	taksimetrOvija 0

шофьор	SOfjOr	SOfjOr 0
Откарайте	Otkarajte	Otkarajte 0
ме	me	me 0
в	f	f 0
центъра	tsent@r@	tsent@r@ 0
на	n@	n@ 0
събитията	s@bitijat@	s@bitijat@ 0
Мадона	madOn@	madOn@ 0
е	e	e 0
истинското	istinskOtO	istinskOtO 0
й	j	j 0
име	ime	ime 0
не	ne	ne 0
псевдоним	psevdOnim	psevdOnim 0
както	kaktO	kaktO 0
мислят	misljat	misljat 0
много	mnOgO	mnOgO 0
хора	hOra	hOra 0
и	i	i 0
освен	Osfen	Osfen 0
това	tOva	tOva 0
тя	tja	tja 0
няма	njam@	njam@ 0
италиански	italianski	italianski 0
корени	kOreni	kOreni 0
Майка	majk@	majk@ 0
й	j	j 0
също	s@StO	s@StO 0
Мадона	madOn@	madeOn@ 0
е	e	e 0
от	Ot	Ot 0
френско	frenskO	frenskO 0
канадски	kanatski	kanatski 0
произход	prOishOt	prOishOt 0
Днес	dnes	dnes 0
Мадона	madOn@	madOn@ 0
печели	petSeli	petSeli 0
по	pO	pO 0
милиона	miliOn@	miliOn@ 0
долара	dOlar@	dOlar@ 0
на	n@	n@ 0
година	gOdin@	gOdin@ 0
което	kOetO	kOetO 0
означава	OznatSav@	OznatSav@ 0
че	tSe	tSe 0
състоянието	s@stOjanietO	s@stOjanietO 0
й	j	j 0
е	e	e 0
около	OkOIIO	OkOIIO 0
милиона	miliOn@	miliOn@ 0
и	i	i 0

е	е	е 0	
все	fse	fse 0	
още	OSte	OSte 0	
в	f	f 0	
центъра	tsent@r@	tsent@r@ 0	
на	n@	n@ 0	
събитията	s@bitijat@	s@bitijat@ 0	

Дрога с касова бележка

От лятото насам дизайнерските дроги вече са достъпни в България не само през интернет, но и в няколко специализирани магазина. Там се продават синтетични аналози на някои от най-популярните наркотици. Те имат същия ефект, но понеже химическият им състав е различен, не попадат в списъка на забранените вещества.

Това едва ли ще продължи още много време. Вероятно скоро и България ще се присъедини към другите европейски държави, които са забранили тези съставки. Тогава обаче неуморните химици в Китай, където основно се произвеждат тонове от новите дроги, сигурно ще са измислили друга формула. Състезанието по забрана и синтезиране на нови психоактивни вещества вероятно ще продължи още известно време, но истината е, че държавата никога няма да го спечели, ако продължава сегашната си политика за борба с наркотиците.

drOg@ s kasOv@ beleSk@

Ot ljatOtO nasam dizajnerskite drOgi vetSe s@ dOst@pni f b@lgarija ne samO pres internet nO i f njakOlKO spetsializirani magazin@ tam se prOdavat sintetitSni analOzi na njakOi Ot naj pOpuljarnite narkOtitsi te imat s@Stija efekt nO pOneZe himitSeskijat im s@staf e razlitSen ne pOpadat f spis@k@ n@ zabranenite veStestva

tOva edva li Ste prOd@lZi OSte mnOgO vreme verOjatnO skOrO i b@lgarija Ste se pris@edini k@m drugite evrOpejski d@rZavi kOitO s@ zabranili tezi s@stafki tOgav@ ObatSe neumOrnite himitsi f kitaj k@detO OsnOvnO se prOizveZdat tOnOve Ot nOvite drOgi sigurnO Ste s@ izmislili drug@ fOrmul@ s@stezanietO pO zabran@ i sintezirane n@ nOvi psihOaktivni veStestva verOjatnO Ste prOd@lZi OSte izvestnO vreme nO istinat@ e tSe d@rZavat@ nikOg@ njam@ d@ gO spetSeli akO prOd@lZav@ segaSnat@ si pOlitik@ z@ bOrba s narkOtitsite.

Дрога	drOg@	drOg@ 0	
с	s	s 0	
касова	kasOv@	kasOv@ 0	
бележка	beleSk@	beleSk@ 0	
От	Ot	Ot 0	
лятото	ljatOtO	ljatOtO 0	
насам	nasam	nasam 0	
дизайнерските	dizajnerskite	dizajnerskite 0	
дроги	drOgi	drOgi 0	
вече	vetSe	vetSe 0	

са	s@	s@ 0	
достъпни	dOst@pni	dOst@pni 0	
в	f	f 0	
България	b@lgarija	b@lgarija 0	
не	ne	ne 0	
само	samO	samO 0	
през	pres	pres 0	
интернет	internet	internet 0	
но	nO	nO 0	
и	i	i 0	
в	f	f 0	
няколко	njakOlkO	njakOlkO 0	
специализирани	spetsializirani	spetsializirani 0	
магазина	magazin@	magazin@ 0	
Там	tam	tam 0	
се	se	se 0	
продават	prOdat	prOdat 0	
синтетични	sintetitSni	sintetitSni 0	
аналози	analOzi	analOzi 0	
на	n@	n@ 0	
някои	njakOi	njakOi 0	
от	Ot	Ot 0	
най	naj	naj 0	
популярните	pOpuljarnite	pOpuljarnite 0	
наркотици	narkOtitsi	narkOtitsi 0	
Те	te	te 0	
имат	imat	imat 0	
същия	s@Stija	s@Stija 0	
ефект	efekt	efekt 0	
но	nO	nO 0	
понеже	pOneZe	pOneZe 0	
химическият	himitSeskijat	himitSeskijat 0	
им	im	im 0	
състав	s@staf	s@staf 0	
е	e	e	
различен	razlitSen	razlitSen 0	
не	ne	ne 0	
попадат	pOpadat	pOpadat 0	
в	f	f 0	
списъка	spis@k@	spis@k@ 0	
на	n@	n@ 0	
забранените	zabranenite	zabranenite 0	
вещества	veStestva	veStestva 0	
Това	tOva	tOva 0	
едва	edva	edva 0	
ли	li	li 0	
ще	Ste	Ste 0	
продължи	prOd@lZi	prOd@lZi 0	
още	OSte	OSte 0	
много	mnOgO	mnOgO 0	
време	vreme	vreme 0	

Вероятно	verOjatnO	verOjatnO 0
скоро	skOrO	skOrO 0
и	i	i 0
България	b@lgarija	b@lgarija 0
ще	Ste	Ste 0
се	se	se 0
присъедини	pris@edini	pris@edini 0
към	k@m	k@m 0
другите	drugite	drugite 0
европейски	evrOpejski	evrOpejski 0
държави	d@rZavi	d@rZavi 0
които	kOitO	kOitO 0
са	s@	s@ 0
забрали	zabranili	zabranili 0
тези	tezi	tezi 0
съставки	s@stafki	s@stafki 0
Това	tOgav@	tOgav@ 0
обаче	ObatSe	ObatSe 0
неуморните	neumOrnite	neumOrnite 0
химици	himitsi	himitsi 0
в	f	f 0
Китай	kitaj	kitaj 0
Където	k@detO	k@detO 0
основно	OsnOvnO	OsnOvnO 0
се	se	se 0
произвеждат	prOizveZdat	prOizveZdat 0
тонове	tOnOve	tOnOve 0
от	Ot	Ot 0
новите	nOvite	nOvite 0
дрого	drOgi	drOgi 0
сигурно	sigurnO	sigurnO 0
ще	Ste	Ste 0
са	s@	s@ 0
измислили	izmislili	izmislili 0
друга	drug@	drug@ 0
формула	fOrmul@	fOrmul@ 0
Състезанието	s@stezanietO	s@stezanietO 0
по	pO	pO 0
забрана	zabran@	zabran@ 0
и	i	i 0
синтезиране	sintezirane	sintezirane 0
на	n@	n@ 0
нови	nOvi	nOvi 0
психоактивни	psihOaktivni	psihOaktivni 0
вещества	veStestva	veStestva 0
вероятно	verOjatnO	verOjatnO 0
ще	Ste	Ste 0
продължи	prOd@lZi	prOd@lZi 0
още	OSte	OSte 0
известно	izvestnO	izvestnO 0
време	vreme	vreme 0

но	nO	nO 0	
истината	istinat@	istinat@ 0	istinat@ 0
е	e	e 0	
че	tSe	tSe 0	
държавата	d@rZavat@	d@rZavat@ 0	d@rZavat@ 0
никога	nikOg@	nikOg@ 0	nikOg@ 0
няма	njam@	njam@ 0	
да	d@	d@ 0	
го	gO	gO 0	
спечели	spetSeli	spetSeli 0	spetSeli 0
ako	akO	akO 0	
продължава	prOd@lZav@	prOd@lZav@ 0	prOd@lZav@ 0
сегашната	segaSnat@	segaSnat@ 0	segaSnat@ 0
си	si	si 0	
политика	pOlitik@	pOlitik@ 0	pOlitik@ 0
за	z@	z@ 0	
борба	bOrb@	bOrba 1	
с	s	s 0	
наркотиците	narkOtitsite	narkOtitsite 0	narkOtitsite 0