Die Eignung unterschiedlicher deutscher Transkriptionssystemarchitekturen für zukünftige Anforderungen.

Andreas Mengel, Technische Universität Berlin, Institut für Kommunikationswissenschaft Katrin Rosenke, Technische Universität Berlin, Institut für Fernmeldetechnik

Zusammenfassung

Da für das Deutsche bereits verschiedene Datenbasen für die Transkription von Wörtern vorhanden sind, stellt sich die Frage des Nutzens der Weiter- und Neuentwicklung von Transkriptionssystemen. Deshalb wird zunächst betrachtet, welche Wörter in einem speziellen Anwendungsfall von vorhandenen Datenbasen abgedeckt werden. Ausgehend von den Wörtern, die typischerweise nicht in vorhandenen Datenbasen auftauchen, wird ein Testdatensatz für vier verschiedene Transkriptionssysteme erstellt. Bei den getesteten Systemen handelt es sich um ein regelbasiertes System, ein morphologiebasiertes System und zwei selbstlernende Systeme, eines davon ist ein neuronales Netz und das andere basiert auf statistischen Ansätzen. Die Fehler dieser Systeme werden analysiert, sie erlauben Rückschlüsse auf die Eignung der Systeme für zukünftige Anwendungen.

1 Einleitung

Für die Zwecke der Sprachsynthese und die damit verbundene automatische Umsetzung von Orthographie in Lautsignale ist eine dritte Ebene - die der Repräsentation durch Lautzeichen - vonnöten. In den letzten Jahren wurden unterschiedlichste Verfahren auch für das Deutsche für die Transkription von orthographischem Text vorgestellt. Die jeweils an diese Systeme gestellten Anforderungen und unterstützten Transkriptionstiefen sowie deren Qualität variieren dabei zum Teil beträchtlich voneinander. Weiterhin ist der Anwendungsbereich und die Testumgebung für die beschriebenen Systeme stets begrenzt, so daß die Qualitätsmessungen der Systeme mehr über den Grad der Anpassung an den verwendeten Wortschatz und den jeweiligen Transkriptionsstandard aussagen, als über deren tatsächliche Verwendungsqualität. Schließlich muß man fragen, inwiefern fehleranfällige Transkriptionssysteme statischen Datenbasen überlegen sind. Deshalb soll im vorliegenden Artikel die Leistungsfähigkeit verschiedener Transkriptionssysteme für zu erwartendes neues Datenmaterial untersucht werden.

2 Vorhandene Datenbasen

Für das Deutsche sind u.a. folgende Transkriptionsdatenbasen verfügbar: Die CELEX Lexical Database [BAA93], die 352.834 Einträge für deutsche Wörter umfaßt, und den deutschsprachigen Teil des ONOMASTICA-Projektes [ONO95], der 947.316 Namen und ihre Transkriptionen verzeichnet. Im folgenden Text wird der Begriff *Eintrag* für in der Form verschiedene Wörter, der Begriff *Nennung* für

Um zu untersuchen, wieviele Wörter von geschriebenen Texten durch diese Datenbasen abgedeckt werden, wurde eine auf CD-ROM verfügbare Tageszeitung [TAZ94] herangezogen. Das ausgewählte Korpus umfaßt 29.426.805 Nennungen, die sich auf 780.738 verschiedene Einträge verteilen. Von diesen konnten allerdings nur 210.808 Einträge in der CELEX und der ONOMASTICA-Datenbasis gefunden werden, das sind 27% aller Einträge. Andererseits decken diese 27% der Einträge bereits 92,4% aller im Text vorkommenden Nennungen ab.

Wenn 92,4% der auftretenden Einträge einer Textdatenbasis wie einer Tageszeitung weder in der CELEX noch in der ONOMASTICA-Datenbasis zu finden sind, stellt sich die Frage, um was für eine Art von Wörter es sich dabei handelt.

Mit Hilfe eines morphologischen Analyseprogramms [HAA95] wurde die Wortklassenzugehörigkeit von nicht in den in den Datenbasen vorhandenen Einträgen (569.930) festgestellt (Tabelle 1).

Wortart	Einträge	Nennungen
Substantive	343.000	1.342.753
Verben	55.814	324.515
Adjektive	45.341	198.648
sonstige	180	10.444
gesamt	444.335	1.876.360
nicht analysierbar	125.595	324.440

Tabelle 1
Wortarten der nicht in den Transkriptionsdatenbasen gefundenen Einträge

125.595 Einträge konnten nicht automatisch identifiziert werden. Es liegt die Vermutung nahe, daß es sich zum großen Teil um Tippfehler, außergewöhnliche Wortkonstruktionen, nicht-native Wortneuschöpfungen und ausländische Namen handelt. Eine Analyse der Häufigkeiten dieser Einträge ergibt, daß es sich mit diesen 125.595 Einträgen um ein Aufkommen von 324.440 (1,1%) Nennungen handelt. Damit hat jeder dieser Einträge durchschnittlich 2,58 Nennungen, was im Gegensatz zu allen Einträgen, die durchschnittlich 37,69 Nennungen aufweisen, wenig ist.

Es wurde festgestellt, daß 343.000 von nicht in der CELEX vorhandenen Einträgen Substantive sind. Das entspricht einem Anteil von 52,13%; gemessen an den als Wörter klassifizierbaren Einträgen handelt es sich um 77,19%. Dies ist im Vergleich zu den Einträgen der CELEX ein sehr hoher Anteil, da hier die Substantive mit nur 22,49% vertreten sind (Tabelle 2).

Wortart	Einträge	Nennungen
Substantive	73.515	847.703
Verben	168.459	1.266.407
Adjektive	82.765	44.0821
sonstige	2.069	1.853.143
gesamt	326.808	4.408.074
nicht analysierbar	26.026	17.926

Tabelle 2
Wortarten in der CELEX

Beim überwiegenden Anteil von neuen bzw. nicht in den verfügbaren Datenbasen aufgeführten Wörtern handelt es sich also um Substantive. Eine weitere

Komposita - Wörter mit mindestens zwei lexikalischen Morphemen: *Hausboot, Industrialisierungskampagne, Bearbeitungsgebühren* - handelt. Betrachtet man auch die Adjektive und Verben unter diesem Gesichtspunkt, so zeigt sich, daß insgesamt 83,75% (372.126) der analysierbaren Wörter Komposita sind.

An dieser Stelle muß darauf hingewiesen werden, daß im Deutschen gerade von der Möglichkeit der Wortneuschöpfung durch Komposition und Derivation besonders reger Gebrauch gemacht wird, andererseits aber gerade nur diese Kategorie neuer Wörter mit dem erwähnten System identifiziert werden kann. Neuschöpfungen durch Übername aus anderen Sprachen oder ohne Zuhilfenahme morphologischer Mittel können mit diesem System definitionsgemäß nicht erkannt werden. Dennoch ist die große Anzahl der Substantive und des kompositorischen Prinzips bei den nicht in den Datenbasen zu findenen Wörtern unbestreitbar.

Ein weiterer Punkt verdient Aufmerksamkeit: In der CELEX sind Substantive gegenüber Verben und Adjektiven viel geringer vertreten als in der Zeitungsdatenbasis. Dies hat drei Gründe. Erstens ist die Anzahl der möglichen Formen pro Adjektiv oder Verb viel größer als die von Substantiven. Zweitens sind die in der CELEX aufgeführten Flexionen nicht nur aus Textanalysen und -sammlungen hervorgegangen, sondern auch synthetisch erzeugt worden, was nicht nur dazu führt das die Formen der Wörter komplett sind, sondern z.T. auch absurde Formen auftreten (*gespensterhaftester, niederregnetest*). Drittens finden sich abhängig von der Textsorte bestimmte Formen der Wortarten besonders häufig, die anderer besonders selten: Bei Verben beispielsweise wird in Texten wie Kommentaren, Beschreibungen und Nachrichten fast ausschließlich die dritte Person Singular, die dritte Person Plural und der Infinitiv benutzt, wobei die beiden letztgenannten darüberhinaus in ihrer Form identisch sind.

Die oben dargestellten gefundenen Unterschiede zwischen CELEX und Zeitungskorpus sind also gut erklärbar im Hinblick auf zu erwartende Verteilungen der Einträgsarten und zeigen deutlich die gute Verwendbarkeit des hier analysierten Korpus für die Untersuchung von für die Sprachsynthese in Frage kommende Texte. Nimmt man jetzt weiter an, daß es immer einen gewissen Anteil von Einträgen in Texten geben wird, der nicht in vorhandenen Datenbasen verzeichnet ist, und geht man davon aus, daß es sich bei den meisten von ihnen um kurzlebige Komposita handelt, wie es die oben angeführte Analyse nahelegt, so wird man diese Art von Einträgen mit Transkriptionssystemen bewältigen müssen.

3 Testdaten

5.000 verschiedene Komposita wurden ausgewählt. Ihren Aufbau zeigt Tabelle 3.

Anzahl	Silben	lexikalische Morpheme
2	686	4.763
3	1.463	236
4	1.477	1
5	851	-
6	375	-
7	116	-
8	26	-
9	6	-

Tabelle 3
Silben- und Morphemstruktur der Testdaten

Diese 5.000 Wörter wurden mit Angabe der Laute, Silbengrenzen, Haupt- und Nebenbetonungen transkribiert. Wörter, die fremdsprachliche Laute, die im deutschen Lautsystem nicht vorkommen, beinhalten, und solche, deren Buchstabe-zu-Laut-Beziehungen für das Deutsche besonders ungewöhnlich sind (*Bandleader* ['bE:nt-,li:-d6]), wurden nicht in das Testvokabular miteinbezogen.

In den Daten finden sich insgesamt 63.305 Buchstaben, 57.065 Laute, 19.248 Silben und 11.387 Betonungen.

4 Verwendete Systeme

4.1 Regelbasiertes System

Das System FON ist ein rein regelbasiertes System, das auf der Grundlage von kontextsensitiven Regeln, die eine feste Abfolge haben, die orthographischen Strings sukzessiv in phonetische umwandeln. Obwohl dieses System ohne Lexikon arbeitet, ist es dennoch möglich, die Regularitäten der deutschen Orthographie zu nutzen, um auch lexikalische Morphemgrenzen zu bestimmen.

4.2 Morphologiebasiertes System

Das hier benutzte morphologisch System ist eine Kombination verschiedener Ansätze und Softwaremodule. Zunächst wird mit dem bereits erwähnten Morphologiesystem [HAA95] eine morphologische Analyse durchgeführt. Die analysierten Morpheme werden dann in einem Morphemlexikon nach ihren transkriptorischen Entsprechungen durchsucht, welche dann zusammengesetzt werden. Diese Morphem-Transkriptionsketten werden schließlich für Aspekte der Silbentrennung und Betonung mit einem regelbasierten System weiterverarbeitet.

4.3 Neuronales System

Das erste System, das die Transkription anhand von Beispielen „lernt“, ist ein neuronales Netz (siehe auch ROS95). Es handelt sich dabei um ein dreistufiges Multilayer-Perceptron, das mit dem Backpropagation-Verfahren trainiert wurde. Am Eingang des Systems werden 9 Buchstaben des zu transkribierenden Wortes

die vier Buchstaben davor und dahinter dienen der Angabe des Kontexts. Durch eine Eingangskodierung werden diese Buchstaben in Eingangswerte für das MLP umgewandelt.

Das MLP hat 134 Eingangsneuronen eine verdeckte Schicht mit 90 Neuronen und 20 Ausgangsneuronen, die die phonetische Eigenschaften der zu ermittelnden Laute repräsentieren. Die zwei vom MLP zuletzt ermittelten Laute werden wieder auf den Eingang zurückgeführt. Das MLP wurde mit 40.000 dem hier benutzten Transkriptionsstandard angepaßten Wörtern der CELEX Datenbasis trainiert.

4.4 Statistisches System

SELEGRAPH ist ein selbstlernendes statistisches System [AND94], das anhand eines bestehenden Transkriptionslexikons alle Kontexte aller Buchstaben bis zur Länge von vier Buchstaben nach links und nach rechts statistisch auswertet und bei neuen Eingangsdaten die für den entsprechenden Kontext wahrscheinlichste Lautentsprechung ausgibt. Suprasegmentelle Informationen (Silbengrenzen und Betonungen) werden nach kombinatorischen Gesichtspunkten eingefügt.

Bevor das System die Zuordnung von Buchstaben- zu Lautketten der vorhandenen Daten auswerten kann, muß es in einer Vorverarbeitungsstufe (Alignment) feststellen, welcher Buchstabe eines Wortes welchem Laut in der zugehörigen Lautschrift entspricht. Eine ähnliche Vorverarbeitung wurde auch beim Training des neuronalen Systems benötig. Dort wurde das Alignment allerdings nachträglich so optimiert, daß keine Fehler mehr auftraten; d.h. daß diese Vorverarbeitung beim neuronalen System nicht in die Bewertung mit einfließt.

SELEGRAPH wurde mit den gleichen Daten trainiert wie das neuronale System. Bei beiden selbstlernenden Systemen wurde eine kurze Nachbearbeitung durchgeführt, die sicherstellte, daß jedes Wort genau eine Hauptbetonung erhielt.

5 Analyse der Fehler

5.1 Zählweise der Fehler

Um eine genauere Aussage über die Leistungsfähigkeit der untersuchten Systeme zu erhalten, wurden nicht nur die falsch transkribierten Wörter gezählt, sondern es wurde auch versucht, die Schwere und die Art der im speziellen aufgetretenen Fehler zu bewerten. Die Auswertung der Fehler erfolgte automatisch. Es wurden Lautfehler, Silbenfehler und Betonungsfehler unterschieden. Lautfehler wurden als leicht bezeichnet, wenn ähnliche Laute (siehe Tabelle 4) verwechselt wurden. Dazu zählt auch die Verwechslung langer und kurzer Vokale (z.B. a: $\leftrightarrow$ a).

Vokale						Konsonanten					
e	$\leftrightarrow$	E	i	$\leftrightarrow$	I	b	$\leftrightarrow$	p	d	$\leftrightarrow$	t
o	$\leftrightarrow$	O	u	$\leftrightarrow$	U	g	$\leftrightarrow$	k	z	$\leftrightarrow$	s
y	$\leftrightarrow$	Y	2	$\leftrightarrow$	9	n	$\leftrightarrow$	N	R	$\leftrightarrow$	6
@	$\leftrightarrow$	E	@	$\leftrightarrow$	e	f	$\leftrightarrow$	v	s	$\leftrightarrow$	S

Tabelle 4
leichte Lautfehler
in SAMPA-Notation

Für die Zählung der Lautfehler mußte zunächst festgestellt werden, ob das ausgewertete Transkriptionssystem einen zusätzlichen Laut eingefügt oder einen Laut weggelassen hatte. Wenn dies der Fall war, wurde die Stelle in der Lautschrift markiert. Probleme traten dann auf, wenn zwei oder mehrere Laute zuviel oder zuwenig vorhanden waren; diese Einträge mußten von Hand bearbeitet werden. Nachdem auf diese Art nun eine Zuordnung der Laute der falschen und richtigen Transkription erreicht worden war, erfolgte die Zählung der Fehler auf folgende Weise:

Die falsche und die richtige Transkription wurden lautweise verglichen. Wenn die Laute nicht übereinstimmten, wurde ein Lautfehler gezählt. Ein Betonungsfehler wurde gezählt, wenn an der entsprechenden Stelle in der Lautschrift eine Betonung (Haupt- oder Nebenbetonung) fehlte oder zuviel war oder wenn anstelle einer Hauptbetonung eine Nebenbetonung vorhanden war bzw. umgekehrt. D.h. wenn eine Betonung vom getesten System einen Laut früher oder später gesetzt worden war, wurde dies mit zwei Betonungsfehlern verrechnet. Das gleiche gilt für die Zählung der Silbenfehler. Die Zählung der Fehler soll kurz anhand von zwei Beispielen erläutert werden:

Wort	richtige Lautschrift	falsche Lautschrift
Beweislage	b@-'vaIs-,la:-g@	'be:-,vaIs-,la:-g@
Bildrand	'bIlt-,Rant	'bIl-,dRant

Das erste Wort *Beweislage* wurde vom System FON falsch transkribiert, es läßt sich leicht nachzählen, daß in diesem Fall ein leichter Lautfehler und zwei Betonungsfehler anzurechnen sind. Das zweite Beispiel zeigt die falsche Transkription des Wortes *Bildrand* des neuronalen Systems. Durch die Verschiebung der Silbengrenze entstehen hier zwei Silben- und zwei Betonungsfehler sowie ein leichter Lautfehler.

5.2 Ergebnisse

5.3 Anzahl der Fehler

In Tabelle 5 sind für alle getesten Systeme die Ergebnisse angegeben. In jeder Spalte ist das jeweils beste Ergebniss fett hervorgehoben. Die Laut-, Silben- und Betonungsfehler wurden jeweils auf die Anzahl der falschen Wörter bezogen.

System	regelbasiert	morphologie-basiert	neuronal	statistisch
falsche Wörter in %	69,04	**17,94**	61,72	83,64
nicht automatisch analysierbar	20	**10**	74	42
Lautfehler pro falsches Wort	1,79	1,31	1,61	**0,66**
Lautfehler pro leichte LF	4,80	3,81	4,24	**3,60**
Silbenfehler pro falsches Wort	0,98	**0,63**	2,70	0,95
Betonungsfehler pro falsches Wort	2,11	**1,30**	1,66	1,92
Summe der Fehler pro falsches Wort	4,89	**3,24**	4,22	5,28

Tabelle 5 Fehler der Systeme bezogen auf die Anzahl der falschen Wörter

Allgemein kann man sagen, daß das morphologiebasierte System die besten Ergebnisse liefert, es transkribiert nur 17,94% der Wörter falsch. Außerdem ist hier auch die Summe der Fehler pro falsch transkribiertes Wort am geringsten. Sechs der zehn Wörter, die von Hand nachbearbeitet werden mußten, waren unvollständig, das heißt, ein oder mehrere Morpheme wurden nicht im Morphemlexikon gefunden. Das Ergebnis dieses Systems ließe sich durch eine gezielte - hier wurde stets die erste Variante selektiert - Auswahl bei Wörtern, bei denen mehrere Zerlegungen möglich sind, noch verbessern.

Das statistische System liefert erstaunlicherweise die besten Ergebnisse bei den Lautfehlern. Am schlechtesten findet es die richtigen Silbengrenzen; hier würde eine bessere Nachbearbeitung aber einiges helfen. Bei diesem System muß auch erwähnt werden, daß allein schon 403 Wörter bei der Lernphase weggefallen sind, weil sie für das System in unbrauchbarer Weise aligned wurden. Wenn man diese Vorstufe verbessert, lassen sich sicher noch bessere Ergebnisse erzielen. Bei den Transkriptionen von SELEGRAPH fehlen häufig Laute; dies wird zwar nur als ein Fehler gerechnet, verstümmelt die Lautschrift aber recht stark. Gleiches gilt für Silben und Betonungen. Noch störender ist es, daß häufiger überflüssige Laute in der Lautschrift auftauchen.

Das neuronale System liefert bei Betonungen und Silbengrenzen recht gute Ergebnisse. Dies läßt sich zum Teil auf die Wirksamkeit der Rückkopplung zurückführen. Bei den Lauten jedoch wirkt sich die Rückkopplung teilweise störend aus, es kann dadurch zu Folgefehlern kommen.

5.4 Verteilung der Fehler auf die Wörter

Für die Fehleranalyse ist nicht nur die durchschnittliche Fehleranzahl pro falsch transkribiertem Wort interessant, sondern auch, wie sich die Fehler tatsächlich auf

der falschen Wörter nur leichte Lautfehler, Lautfehler allgemein, Silbenfehler oder Betonungsfehler bzw. bestimmte Fehlerkombinationen enthalten.

System	regelbasiert	morphologie-basiert	neuronal	statistisch
leichte Lautfehler	2,00	3,59	2,40	0,14
Lautfehler	8,00	26,23	13,35	1,96
Silbenfehler	0,23	2,58	0,75	9,64
Betonungsfehler	10,46	24,44	11,63	2,25
Betonungs- und Silbenfehler	3,01	1,23	4,63	46,34
Silben- und Lautfehler	1,39	6,50	3,31	2,94
Betonungs- und Lautfehler	36,36	16,48	29,20	3,08
alle Fehlerarten	38,56	18,95	33,86	33,64

Tabelle 6 Verteilung der Fehlerarten, angegeben ist der Prozentsatz der falschen Wörter, die ausschließlich die genannte(n) Fehlerart(en) enthalten

Beim regelbasiertem System enthalten 38,56% der falschen Transkriptionen alle drei Fehlerarten, fast ebensogroß ist der Anteil der Einträge, die ausschließlich Betonungs- und Lautfehler enthalten. Den drittgrößten Anteil bilden mit 10,46% die Einträge, die nur Betonungsfehler aufweisen. Beim neuronalen System ist die Verteilung der Fehler recht ähnlich, nur ist hier der Anteil der falschen Transkriptionen, die allein Lautfehler enthalten (13,35%), etwas größer. Die Ähnlichkeit der beiden Systeme ist interessant, und deutet darauf hin, daß das neuronalen System Regelmäßigkeiten aus den Daten extrahiert und bei der Anwendung dieser „Regeln“ ähnliche Schwierigkeiten hat wie das regelbasierte System.

Bei über einem Viertel (26,23%) der falschen Trankriptionen des morphologiebasierten Systems treten auschließlich Lautfehler auf, ein knappes weiteres Viertel (24,44%) enthält nur Betonungsfehler. Bei diesem System ist auch der Anteil der Transkriptionen, die nur leichte Lautfehler enthalten, am höchsten. Dieses und die Tatsache, daß die Einträge mit gemischten Fehlern geringer sind, deutet darauf hin, daß die Fehler hier nicht so schwerwiegend sind wie bei anderen Systemen.

Beim statistischen System fällt auf, daß der Anteil der Einträge, die kombinierte Silben- und Betonungsfehler enthalten, überdurchschnittlich groß ist. Gleiches gilt für die Transkriptionen, die nur Silbenfehler enthalten. Man sieht daran wiederum, daß das System vor allem Schwierigkeiten mit der Bestimmung

6 Ausblick

Bei der Anwendung von vier verschiedenen Transkriptionssystemen auf die bei Neubildungen besonders häufig vorkommende Eintragsart Kompositum wurde festgestellt, daß gerade diese von allen Systemen ungenügend gut transkribiert wird. Dieses wiegt um so schwerer, als diese gerade durch ihre geringere Auftretenswahrscheinlichkeit nur schlecht durch den Kontext erschlossen werden können. Unbeantwortet ist jedoch die Frage, wie gravierend fehlerhafte Aussprache und Transkription für das Verständnis von Informationstexten generell ist.

Aus dieser exemplarischen Untersuchung folgt, daß es für den Entwurf und die Integration von Transkriptions- und Sprachsynthesesystemen besonders darauf ankommt, die im Anwendungsfall zu erwartenden Daten auf die folgenden Charakteristika hin zu analysieren: Die benötigte Qualität und Tiefe der Transkription, die Auftretenshäufigkeit der Daten und die vorhandenen Resourcen.

Literatur

[AND94] O. Anderson, P. Dalsgaard: *A Self-Learning Approach to Transcription of Danish Names*, ICSLP 94, 1627-1630.

[BAA93] R.H. Baayen, R. Piepenbrock, H. van Rijn: *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. 1993.

[HAA95] M. Haapalainen, A. Majorin: *GERTWOL und Morphologische Disambiguierung für das Deutsche*, NODALIDA-95.

[ONO95] ONOMASTICA: *Transcription Database for Proper Names of 11 European Languages*, Edinburgh: CCIR, University of Edinburgh.

[ROS95] K. Rosenke: *Verschiedene neuronale Strukturen für die Transkription von deutschen Wörtern*, Konferenz „Elektronische Sprachsignalverarbeitung“, 1995, in diesem Band

[TAZ94] taz 1994: taz CD-ROM. Berlin: die tageszeitung.