INSTITUT FÜR DEUTSCHE SPRACHE

OPAL

Online publizierte Arbeiten zur Linguistik

0/2005

Nina Berend / Stefan Kleiner / Ralf Knöbl

Sprachliche Variabilität des Deutschen und ihre Erfassung mit Methoden der automatischen Spracherkennung

OPAL – Online publizierte Arbeiten zur Linguistik
Herausgegeben vom Institut für Deutsche Sprache

Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
opal@ids-mannheim.de

Technische Redaktion: Norbert Volz

Nina Berend/Stefan Kleiner/Ralf Knöbl

Sprachliche Variabilität des Deutschen und ihre Erfassung mit Methoden der automatischen Spracherkennung*

1. Zur Einführung

Ein wichtiges Merkmal der deutschen Sprache ist ihre Einheit in der schriftlichen und ihre Differenziertheit und Variabilität in der gesprochenen Form. Die sprachliche Variabilität im Bereich des Mündlichen betrifft nicht nur Dialekte oder Regiolekte, sondern sie ist auch für das gesprochene informelle und sogar für das formelle Standarddeutsch in verschiedenem Ausmaß typisch. Die Vielfalt und Heterogenität des gesprochenen Deutsch hängt mit Besonderheiten der historischen Entwicklung des Deutschen als Standardsprache zusammen und ist als Folge des spezifischen Standardisierungsprozesses auch in der Gegenwart noch präsent und relevant. Die Untersuchung und angemessene Dokumentation der sprachlichen Variation in der gesprochenen Standardsprache stellt jedoch in der germanistischen Soziolinguistik bis in die jüngste Zeit hinein ein Forschungsdesiderat dar. Am Institut für Deutsche Sprache (Mannheim) wird gegenwärtig ein neues Projekt konzipiert, das dieses Desiderat aufarbeiten soll und das sich mit der Analyse und Dokumentation der sprachlichen Variation im Deutschen beschäftigen wird. Ausgehend vom Konzept der Pluriarealität des Deutschen soll der regionale Aspekt der Variation im Zentrum stehen. Untersucht und dokumentiert wird schwerpunktmäßig die Variation auf lautlicher Ebene, aber auch die morphologische, syntaktische und pragmatische Variation soll in Zukunft untersucht und dokumentiert werden. Zielvarietäten sind die gesprochene formelle und die informelle nicht-private Standardsprache bzw. die regionalen Umgangssprachen. Eines der Ergebnisse des Projekts ist eine korpusbasierte Datenbank von sprachlichen Varianten des Deutschen, mit Angaben zu Standard- und Regiolekteigenschaften, stilistischen und sprechsprachlichen Charakterisierungen, zu Textsorten bzw. Kommunikationstypen und Formalität bzw. Informalität des Sprechens.

Die Datenbank soll für verschiedene Zwecke und verschiedene Nutzergruppen eingesetzt werden. Einer der wichtigsten Anwendungsbereiche der Datenbank wird der Unterricht Deutsch als Fremdsprache sein. Deutschlernende im Ausland beschäftigten sich bisher fast ausschließlich mit der invarianten formalen deutschen Standardsprache und bekommen so gut wie keine Vorstellung von der Vielfalt der gesprochenen Alltagssprache. Das hat für Studierende weit reichende Folgen, denn sie haben bisher kaum eine Möglichkeit, die kolloquiale Form des Deutschen im Unterricht kennen zu lernen, wie Barbour und Stevenson feststellen: „Das Deutsche ist wahrscheinlich die vielgestaltigste Sprache Europas, weshalb nicht wenige seiner ausländischen Studenten bei ihrem ersten Besuch eines sog. deutschsprachigen Landes verblüfft sind, dass die ihnen dort begegnende Sprache wenig Ähnlichkeit mit jener hat, die sie aus ihren Schulen und Universitäten kennen.“ (Barbour/Stevenson 1998, S. 2). Ähnlich äußert sich auch Durrell (1995), indem er die Problematik direkt anspricht: „Dem armen Aus-

* Der vorliegende Aufsatz basiert auf einem Vortrag, der auf dem 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Marburg, 5.-8. März 2003, gehalten wurde.

länder, der mühsam Kenntnisse in der formalen Hochsprache erworben hat, bleibt nur übrig, diese im Lande selbst zu verlernen, um effektiv im Alltag kommunizieren zu können.“ (S. 426). Die künftige Datenbank soll ausländischen Studierenden für diese Zwecke zur Verfügung gestellt werden und helfen, das tatsächlich gesprochene, variantenreiche Alltagsdeutsch zumindest ansatzweise kennen zu lernen. Die Anwendungsmöglichkeiten der Datenbank gehen aber weit darüber hinaus. Im Endausbau kann sie z.B. von Sprachwissenschaftlern, besonders von Dialektologen, Regionalsprachforschern und Variationslinguisten als Quelle für linguistische Untersuchungen genutzt werden. Sie kann behilflich sein beim Verfassen von Wörterbüchern, die der sprachlichen Wirklichkeit näher kommen wollen als das bisher oft der Fall ist. Sie wird z.B. Informationen darüber geben, wie die Struktur der sprachlichen Variation im gegenwärtig gesprochenen Standarddeutsch ist, welche Varianten für das gesprochene Alltagsdeutsch typisch sind und wie deren regionale und stilistische Verbreitung heute ist.

Die Datenbank wird auf den Ergebnissen der Analyse einschlägiger umfangreicher Korpora des gesprochenen Deutsch basieren. Um jedoch große Korpora analysieren zu können, ist es notwendig, automatische Analyseverfahren der Variation zu entwickeln. Mit traditionellen manuellen Methoden kann der Aufbau einer korpusbasierten Datenbank kaum verwirklicht werden. Dem eigentlichen Variationsprojekt wurde daher eine kleine Pilotstudie vorgeschaltet, die die Möglichkeiten der automatischen Analyse prüfen sollte. Dabei wurde der Frage nachgegangen, ob es möglich ist, regionale Varianten des Deutschen mit Verfahren der automatischen Spracherkennung zu untersuchen, d.h., ob es möglich ist, eine verlässliche Transkription der regionalen Varianten automatisch herzustellen. Diese Pilotstudie zur automatischen Transkription stützte sich auf das im IDS bereits vorhandene System SPRAT (Speech Recognition and Alignment Tool), das zum Alignieren (Text-Ton-Synchronisation) verwendet wird. Im Rahmen der Pilotstudie wurde dieses System modifiziert und in einer Reihe von Tests dessen automatische Transkription evaluiert (vgl. Abschnitt 3). Das Ziel des vorliegenden Beitrags ist es, die Ergebnisse dieser Pilotstudie vorzustellen. Zunächst aber soll ein kurzer Exkurs verdeutlichen, um welches System es sich beim IDS-Aligner SPRAT handelt.

2. Alignment – ein Exkurs

Das im IDS vorhandene Alignmentsystem SPRAT[1] ist keine Spracherkennung im eigentlichen Sinn, sondern eine Methode der Text-Ton-Synchronisation, die auf Prinzipien der Spracherkennung beruht. (Eine ausführliche Beschreibung findet sich in Schmidt/Neumann 1999, Schmidt 2000, Schmidt 2001 und im Internet unter `www.ids-mannheim.de/prag/alignment.html`). Als Ergebnis dieses Alignments wird der schriftliche Text über die Zuordnung von Zeitmarken mit dem entsprechenden Tonsignal verbunden. Dadurch ist der Zugriff aus dem Text auf jede beliebige Stelle bzw. das gewünschte Wort in der Tondatei möglich.

Die gemeinsame Grundlage des Alignments und der Spracherkennung im eigentlichen Sinn ist die mathematische Methode der Hidden-Markov-Modelle („stochastische Automaten“,

[1] SPRAT wurde im Rahmen des Projekts SERGES (Schriftliche Erfassung gesprochener Sprache) entwickelt und in einem Kooperationsprojekt mit dem IMS (= Institut für maschinelle Sprachverarbeitung) Stuttgart weiterentwickelt.

kurz HMM). HMMs modellieren Einzellaute bzw. Wörter als Folgen von 'Zuständen', die basalen Lauteigenschaften entsprechen.[2] Die Zustände sind durch Übergänge miteinander verbunden, die entweder auf einen anderen Zustand oder auf denselben Zustand verweisen. Die verbindenden Übergänge sind gewichtet mit Übergangswahrscheinlichkeitswerten; daneben werden in jedem Zustand Symbole („Observationen") ausgegeben. Die Zuweisung eines Symbols zu einem Zustand erfolgt mittels eines Emissionswahrscheinlichkeitswerts. Diese Wahrscheinlichkeitswerte werden in einem Training aus den Daten berechnet.[3] Die Trainingsdaten müssen dafür exakt segmentiert und klassifiziert sein. Beim Training wird für jeden Laut, der Teil des Phoninventars sein soll, ein HMM angelegt, das aus einer vorher festgelegten Anzahl von Zuständen besteht, und das anhand von möglichst vielen und heterogenen Realisationen des Lauts trainiert wird. Dadurch entsteht ein Prototyp, der möglichst alle akustischen Eigenschaften aller Vorkommen des betreffenden Lauttyps repräsentiert.

Der IDS-Aligner SPRAT besteht aus folgenden Komponenten:

1. Ein Inventar aus 46 Phonmodellen (HMMs). Diese Phonmodelle wurden in bereits trainiertem Zustand in das Tool SPRAT integriert und sind die Grundlage der Spracherkennung des Systems. Dieses Inventar der Phonmodelle ist an der Standardsprache orientiert, d.h., es wird zur Erkennung von standardsprachlichen (im Folgenden ssprl.) Äußerungen verwendet. Mit diesen Modellen können auch bestimmte regionalsprachliche Lautungen erkannt werden, nämlich solche, die mit ssprl. phonetisch identisch sind, aber eine andere Distribution aufweisen (vgl. Tabelle 1).
2. Ein Regelsatz mit ssprl. Graphem-Phonem-Korrespondenzen. Über diesen Regelsatz kann das System SPRAT seine Phonmodelle mit den tatsächlichen Lautungen (d.h. mit dem Tonsignal) vergleichen und so den entsprechenden Schreibungen im orthografischen Transkript zuordnen. Dieser Regelsatz wurde für die Tests der Variation verändert. Den bereits vorhandenen ssprl. Regeln wurden neue, regionale Regeln hinzugefügt. Dadurch sollte die Leistungsfähigkeit des Systems in Bezug auf Erkennung der regionalen Variation überprüft werden.
3. Ein Inventar mit HMMs (sog. Ganzwortmodelle) für die nichtsprachlichen Phänomene *Lachen*, *Husten*, *Störung*, *Atmen* und *Klatschen* und die Hesitations- bzw. Rezeptionssignale *Mhm*, *äh*, *ähm* und *m*. Die Modelle für die nichtsprachlichen Phänomene sowie für die Hesitationsphänomene sind in Kooperation mit dem IMS entwickelt und trainiert worden.[4]
4. Ein Verfahren zur Behandlung von Simultanpassagen.[5]

[2] Standard-HMM-Systeme wie SPRAT verwenden zur Klassifizierung artikulatorisch-akustischer Eigenschaften spektrale Eigenschaften des Sprachsignals (Kurzzeitspektren von 10-20ms Dauer), wohingegen sog. hybride Systeme auf neuronale Netze (ANNs) zurückgreifen (vgl. Hosom 2000).

[3] Vgl. Bodmer/Fach/Schmidt/Schütte (2001); vgl. auch Anm. 8.

[4] Dazu Fach (2001), S. 66-91; Projektberichte im Internet: `http://www.ims.uni-stuttgart.de/phonetik/projekte/IDS/activities.html` (Stand: März 2005).

[5] Vgl. Schmidt (2002).

3. Tests zur Variation

Um die Erkennungsleistung des IDS-Aligners SPRAT bei sprachlicher Variation auf Lautebene zu überprüfen, wurden zahlreiche Tests durchgeführt, von denen einige im Folgenden kurz beschrieben werden sollen.

Im Mittelpunkt der durchgeführten Testreihe stand das Regelset des Alignmentsystems. In einem ersten Schritt wurde jeweils nur eine einzige Regel geändert, um die Erkennung konkreter Variationsphänomene zu überprüfen. Im zweiten Schritt wurden dann mehrere Regeln gleichzeitig geändert, um festzustellen, wie sich die Regeln aufeinander auswirken und wie sich die Leistung des Systems unter diesen Umständen verändert. Außerdem wurden unterschiedliche Konstellationen der Modelle für nonverbale Phänomene sowie für Hesitations- und Rezeptionssignale erprobt, um deren Einfluss auf die Variantenerkennung festzustellen.

Die empirische Basis für diese Tests waren verschiedene Korpora des gesprochenen Deutsch, die am Institut für Deutsche Sprache vorliegen. In erster Linie wurde dazu das „König"-Korpus herangezogen, weil es die entsprechende für die Tests notwendige regionale und stilistische Variation aufweist. Es besteht aus Aufnahmen, die als Grundlage für den *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland* (König 1989) dienten und die neben Vorlesesprache auch Spontansprache enthalten. Alle im Folgenden aufgeführten Beispiele stammen aus diesem Korpus. Außerdem wurde auch an Aufnahmen des Pfeffer-Korpus getestet (vgl. Pfeffer/Lohnes (Hg.) 1982).

Die ausgewählten Variationsphänomene, deren automatisches Erkennen getestet worden ist, sind in Tabelle 1 dargestellt. Sie sind im deutschen Sprachgebiet großräumig verbreitet und sind zum Teil für gelesene, meistens aber für frei gesprochene Sprache typisch. Es handelt sich dabei einerseits um typisch regionale Aussprachevarianten wie z.B. das stimmlose [s], das systematisch und nahezu sprecherunabhängig im oberdeutschen Sprachraum verwendet wird. Ein ähnliches Merkmal stellt die regionale Frikativierung von [pf] dar, die im norddeutschen Raum verbreitet ist. Andere Variationsphänomene wie z.B. die *t*-Tilgung sind wohl weniger als regional gebundene, sondern eher als Merkmale der stilistischen Variation zu sehen, die mit regionalen Mitteln realisiert wird. – Die Frage der genauen Einordnung und Beschreibung soll hier nicht weiter thematisiert werden, da sie im Zusammenhang des vorliegenden Beitrags eher zweitrangig ist.

Tabelle 1: Untersuchte Variationsphänomene[6]

ssprl. Lautung	**regionale Realisierung**	**Variationsphänomen**	**Beispiele**
[z]	[s]	Stimmtonbeteiligung bei <s>	***S**ohn*, *ver**s**ucht*
[C]	[S]	Koronalisierung bei <-ich(-)>	*i**ch***, *si**ch**er*
[C]	[g]	Frikativ/Plosiv bei <-ig(-)>	*wicht**ig***, *Schwier**ig**keiten*
[s]	[S]	Palatalisierung bei <-st(-)>/ <-sp(-)>	*fe**st***, *Polizi**st,** Re**sp**ekt*
[@]	Ø	*e*-Apokope	*müd**e***, *Roll**e***, (*ich*) *mach**e***
[n]	[@]	*n*-Apokope	*sage**n***, *Garte**n***
[g@]/[b@]	[g]/[b]	*e*-Synkope in <ge->/<be->	*G**e**fühl*, *g**e**nau*, *b**e**sonders*
[t]	Ø	*t*-Tilgung wortfinal	*is**t**, nich**t***
[pf]	[f]	Frikativierung von <pf->	***Pf**and*, ***Pf**erd*

4. Testergebnisse

An den Ergebnissen einiger durchgeführter Tests sollen in diesem Abschnitt die spezifischen Probleme und Phänomene erläutert werden, die bei der automatischen Erkennung von regionalen Varianten aufgefallen sind. Wie zuverlässig sind also die lautschriftlichen Transkriptionen von SPRAT bei bestimmten Variationsphänomenen? Anders formuliert: Wie gut decken sich die automatischen Transkriptionen mit der ohrenphonetischen, manuellen Transkription der Autoren dieses Beitrags? Damit diese manuelle Transkription als Bezugssystem und Vergleichsmaßstab dienen kann, wird im Folgenden implizit vorausgesetzt, dass das manuell Transkribierte tatsächlich die lautliche Realität erfasst. Auch wenn das natürlich nicht in jedem Fall zutrifft und bei gemeinsam abgehörten Passagen durchaus manchmal Uneinigkeit über Details des zu Transkribierenden auftauchten, bestehen wohl keine Einwände dagegen, die manuell erstellte Transkription als Referenzsystem zur Bewertung der maschinellen Transkriptionsleistung heranzuziehen.[7] Wenn im Folgenden von standardsprachlichen (ssprl.) Realisierungen die Rede ist, dann ist damit die Orthoepie, wie sie in den Aussprachewörterbüchern (Siebs 1969, Duden 2000, Krech et al. (Hg.) 1982) zu finden ist, gemeint. Die minimalen Diskrepanzen zwischen diesen orthoepischen Vorschriften sind für die durchgeführten Tests nicht von Belang.

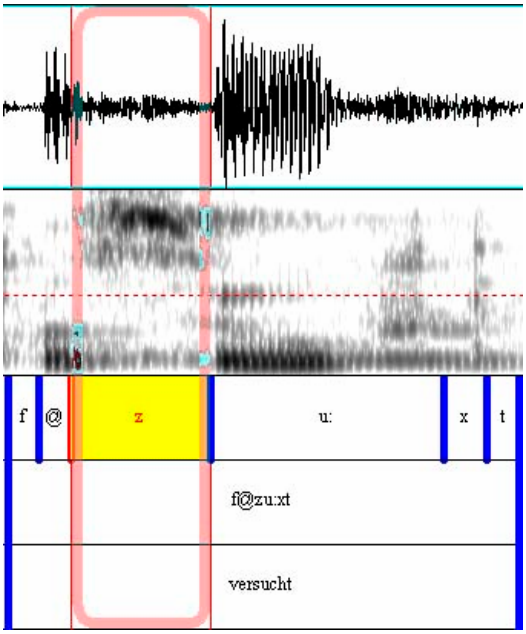
[6] Als Lautschrift wird in diesem Beitrag SAMPA verwendet, ein auf ASCII-Code basierendes, daher maschinenlesbares phonetisches Alphabet.

[7] Lautschriftliche Transkription ist bekanntlich nie völlig objektivierbar. Differenzen bei der Transkription desselben Sprachsignals ergeben sich nicht nur zwischen zwei verschiedenen Transkribierenden. Auch bei ein und demselben Transkriptor kommt es bei mit zeitlichem Abstand durchgeführtem mehrmaligem Abhören desselben Sprachsignals jeweils zu im Detail abweichenden Transkriptionen (vgl. Almeida/Braun 1982, König 1988).

4.1 *s*-Stimmtonbeteiligung

Getestet wurde hier, ob der *s*-Laut in den Positionen, in denen er standardsprachlich stimmhaft ist, auch bei regional stimmloser Aussprache korrekt als [s] automatisch transkribiert wird. Dazu mussten die bisherigen standardsprachlichen Regeln, die in diesen Positionen nur stimmhaftes [z] vorsahen, um die Variante [s] erweitert werden, d.h., das System hatte jetzt bei der Transkription in diesen Positionen die freie Wahl zwischen [s] und [z].

Abb. 1: Stimmhafte [z]-Transkription von stimmlosem [s] (Aufnahme Offenburg)



Anmerkung:
In dieser und in den folgenden Abbildungen handelt es sich um Bildschirmfotos des Phonetik-Programms *Praat*, mit dem die Analyse durchgeführt wurde. In den Bildschirmausschnitten werden oben ggf. das Oszillogramm, in der Mitte das entsprechende Breitbandsonagramm und unterhalb drei Transkriptspuren dargestellt. Die unterste repräsentiert das orthografische Transkript, also das in den Aligner eingespeiste schriftliche Material. Die oberste stellt das für die Tests besonders relevante automatische Transkriptionsergebnis mit den einzelnen phonetischen Segmenten dar. Die mittlere Spur entspricht der oberen. Hier wurden allerdings nur solche Segmentgrenzen beibehalten, die gleichzeitig auch Lexemgrenzen bilden. Das sind diejenigen Informationen, die als zeitliche Anker für das wortweise Alignment benötigt werden.

Das Beispiel in Abb. 1 zeigt ein automatisch transkribiertes, angeblich stimmhaftes [z] in *versucht*, das tatsächlich stimmlos realisiert wird. Insgesamt ist das Ergebnis der automatischen Transkription für stimmhaftes [z] und stimmloses [s] sehr schlecht – wie folgender Tabelle für die Aufnahmen zu entnehmen ist.

Tabelle 2: Transkription für stimmhaftes [z] und stimmloses [s]

	ssprl. realisiert [z]	**richtig transkribiert**	**regional realisiert [s]**	**richtig transkribiert**
Offenburg	19	19 (100%)	19	keine
Karlsruhe	3	3 (100%)	83	3 (4%)

Unterschiede zwischen [s] und [z] werden praktisch nicht erkannt, die Entscheidung fällt vielmehr auch bei stimmlos realisiertem [s] fast immer zugunsten der Transkription [z]. Als Ursache für diese Präferenz ist neben der großen akustischen Nähe der beiden Laute zu vermuten, dass beim ursprünglichen Training der Phonmodelle für stimmhaftes [z] und stimmloses [s] eventuell nicht genügend zwischen den beiden *s*-Lauten differenziert wurde. Ein Neutraining selbst durchzuführen war aufgrund des großen zeitlichen Aufwands im Rahmen dieser Pilotstudie jedoch nicht möglich.[8]

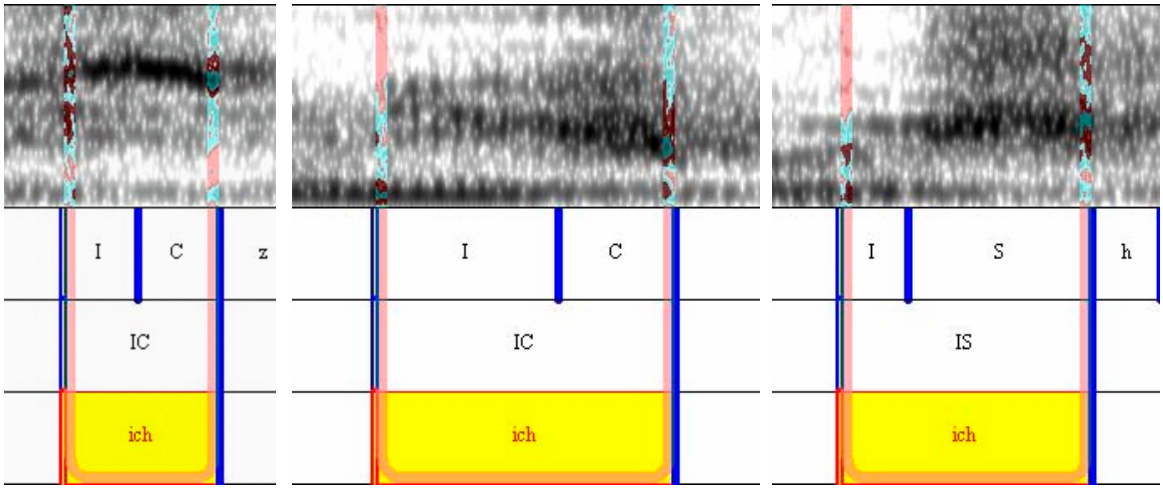
4.2 Koronalisierung

Für diesen Test wurde der orthografische Kontext <-ich(-)> wie in *ich*, *mich*, *sicher* usw. in der Aufnahme Wittlich analysiert, mit dem Zweck, herauszufinden, ob das als Koronalisierung bekannte Phänomen (vgl. Herrgen 1986) vom automatischen Transkriptionssystem korrekt erkannt wird. In die entsprechende ssprl. Regel, die für orthografisches <ch> nach hellen Vokalen als feste Transkription [C] (den *ich*-Laut) vorsah, wurde dazu zusätzlich die Möglichkeit eingebaut, an dieser Stelle auch [S] zu transkribieren.

Zur Illustration der gerade bei diesem Beispiel besonders auffälligen Problematik wurde folgendes Beispiel gewählt: In dem Satz „Wenn *ich* so, wenn *ich* mitten in der Diskussion bin, so gewisse Redewendungen, hört man sofort wo *ich* herkomme“ tritt dreimal das Pronomen *ich* auf, es wird aber nicht in jedem Fall vom Sprecher gleich realisiert, sondern es wird in unterschiedlichem Maß koronalisiert. Es lässt sich sozusagen eine Zunahme der [S]-Haltigkeit feststellen, auch wenn die Unterschiede teilweise nicht besonders groß und nicht ganz leicht wahrzunehmen sind. Am Sonagramm sind die unterschiedlichen Verdichtungen im Frequenzspektrum bei den drei Frikativen aber deutlich zu erkennen (im Ausschnitt ist ein Spektrum von 0-8 kHz dargestellt). Wo der Mensch aber noch leicht Abstufungen wahrnehmen und diese eventuell auch lautschriftlich differenzieren kann, kann das System SPRAT nur eine Entweder-Oder-Entscheidung treffen, da es nur die trainierten Phone [C] und [S] als Alternativen hat. Zwischenwerte können nur einem der beiden Phonmodelle zugewiesen werden. Das sieht man auch an den automatischen Transkriptionen (vgl. Abb. 2).

[8] Um eine Minute an zuverlässigem Trainingsmaterial herzustellen, benötigen geübte Transkribierende nach Erfahrungswerten phonetischer Institute zwischen 5 und 10 Stunden Arbeitszeit. Es geht hier aber nicht nur um die phonetische Transkription, sondern die Laute müssen auch fast auf die Millisekunde genau segmentiert werden, was einen viel größeren zeitlichen Aufwand benötigt. Um den kompletten Phonsatz sinnvoll neu zu trainieren, wären mindestens zwei Stunden transkribiertes und segmentiertes Tonmaterial nötig. Zu diesem Zweck hätten also bis zu 1200 Stunden Arbeitszeit geopfert werden müssen.

Abb. 2: Transkription [C] und [S] für ssprl. <ich> (Aufnahme Wittlich)



Die Koronalisierungen des Sprechers aus Wittlich weichen zwar sichtbar (und hörbar) vom ssprl. [C] ab und ein menschlicher Transkribent würde sie deshalb wohl alle dem [S] zuschlagen, wenn nur die Wahl zwischen zwei Transkriptionsmöglichkeiten bestünde. Für die Maschine beginnt der Bereich, ab dem sie ein [S] transkribiert, aber offenbar erst viel näher am „echten“ [S]. Daher auch die geringe Trefferquote von nur 44%.

Tabelle 3: Transkription [C] und [S] für ssprl. <ich>

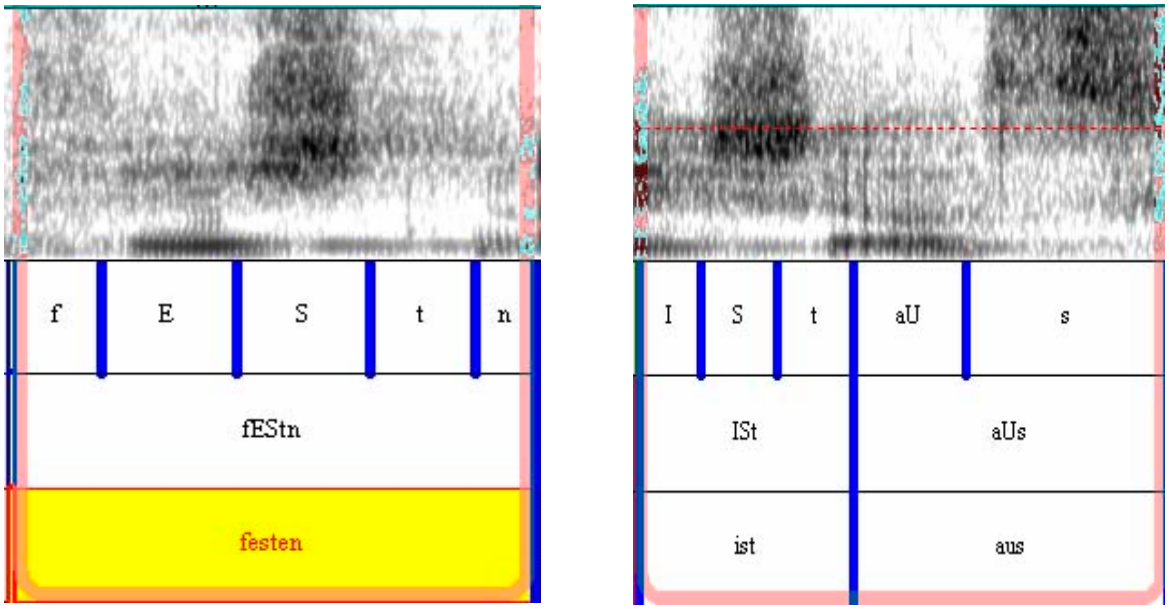
	ssprl. realisiert [C]	**richtig transkribiert**	**regional realisiert [S]**	**richtig transkribiert**
Wittlich	2	2 (100%)	48	21 (44%)

Vermutlich wäre es sinnvoll, hier ein neues Phon mit diesen „Zwischenwerten“ zu trainieren. So könnte durch ein um ein Glied erweitertes System der Phonmodelle auch die automatische Transkriptionsleistung verbessert werden. Es ist allerdings auch zu bedenken, ob ein zusätzliches Phon die Fehlerrate nicht noch weiter erhöhen würde, da das System sich dann statt zwischen zwei phonetisch einigermaßen verschiedenen Varianten zwischen drei sich phonetisch sehr nahestehenden Varianten entscheiden müsste.

4.3 *s*-Palatalisierung im Kontext <-st(-)>

Das nächste Beispiel für eine linguistisch interessante Variable ist die besonders für das Südwestdeutsche typische *s*-Palatalisierung vor Plosiv im In- und Auslaut wie im hier aufgeführten Beispiel *festen* aus Kempten (vgl. Abb. 4 links).

Abb. 3: Transkription von [S] für ssprl. <-st(-)> (Aufnahme Kempten)



Zur Bewertung der automatischen Transkriptionsleistung in diesem Fall wurde folgendermaßen vorgegangen. Statt der für die Standardsprache notwendigen kontextabhängigen Regel, bei initialer <st>-Graphie [S], in allen anderen Positionen aber [s] zu transkribieren, wurde dem System die Freiheit gelassen, ohne Rücksicht auf die Position für jedes orthografische <s> vor <t> und <p> sowohl [s] als auch [S] als Alternativen zuzulassen.[9] Wie die folgende Tabelle zeigt, stimmt die automatische Transkription hier weit häufiger mit den manuell transkribierten Lautungen überein als das bei allen anderen getesteten Variablen der Fall ist.

Tabelle 4: Transkription von [S] für ssprl. <-st(-)>

	regional realisiert [S]	**richtig transkribiert**
Kempten	16	12 (75%)
Offenburg	19	16 (84%)
Pfeffer-Korpus	72	51 (71%)

Wie lassen sich diese überraschend guten Resultate deuten? Die Argumentation führt wieder auf die phonetisch-akustische Ähnlichkeit bzw. in diesem Fall Unähnlichkeit der beiden voneinander zu unterscheidenden Laute zurück. [s] und [S] sind ja auch für das menschliche Ohr sehr gut voneinander zu unterscheiden und die palatalisierte Variante ist sicher nicht zufällig ein fast schon stereotypes Kennzeichen von südwestdeutscher Regionalsprache. Es ist ja möglicherweise kein Zufall, dass gerade der Unterschied zwischen [s] und [S] durch die Erwähnung im Alten Testament (im Buch der Richter 12, 5,6) als *Schibboleth* zum Terminus für ein sprachliches Erkennungsmerkmal schlechthin geworden ist. Es sei daran erinnert, dass dort jene Dialektsprecher, die statt „***Sch****ibboleth*“ „***S****ibboleth*“ sagten, als nicht zum eigenen

[9] Mit dieser Wahlfreiheit hätte man umgekehrt natürlich auch testen können, wie gut die Ent-Palatalisierung von initialem <st->/<sp-> in Norddeutschland automatisch erkannt wird.

Stamm gehörig entlarvt wurden. Der auditiv wahrnehmbare deutliche Unterschied zwischen [s] und [S] lässt sich auch spektral gut sichtbar machen, d.h., auch SPRAT hat hier vergleichsweise wenig Schwierigkeiten, die beiden Laute auseinander zu halten. Das lässt sich am Beispiel rechts in Abb. 3 erkennen. Dort kommen [S] und [s] kurz hintereinander vor. Bei [S] sieht man im Sonagramm die Verdichtung in der Mitte, d.h. bei 3-4 kHz, bei [s] liegen die primär beteiligten Frequenzen dagegen am oberen Rand des dargestellten Spektrums, d.h. bei 6-8 kHz. Ein weiterer begünstigender Faktor ist, dass in diesem Fall im analysierten Material nur selten Zwischenwerte (wie ein dorsales [s´]) vorkommen, die eine automatische Erkennung hätten erschweren können (wie das bei der Koronalisierung der Fall war).

4.4 *e*-Synkope in <ge->

Untersucht wurde hier die Synkope von [@] im Präfix <ge-> (außer vor Plosiven), also in Fällen wie *Gfühl*, *gnau*, die in Abb. 4 als Beispiele gewählt wurden. Dazu wurde zunächst die Regel, die für <ge-> nur die vokalhaltige Transkription [g@] vorsah, um die vokallose Alternative [g] erweitert. Da der Plosiv des Präfixes bei ausgefallenem Vokal in aller Regel stimmlos realisiert und je nach Folgekonsonant zum Teil auch fortisiert wird, wurde für die Aufnahme Kempten ein weiterer Testlauf durchgeführt, in dem der Maschine neben ssprl. [g@] als regionale Alternative statt Lenis-[g] die Lautung Fortis-[k] als Transkriptionsmöglichkeit zur Wahl gestellt wurde. Insgesamt lässt sich an den Werten in der folgenden Tabelle erkennen, dass die automatische Transkription zur vokalhaltigen Variante tendiert.

Tabelle 5: *e*-Synkope in <ge->

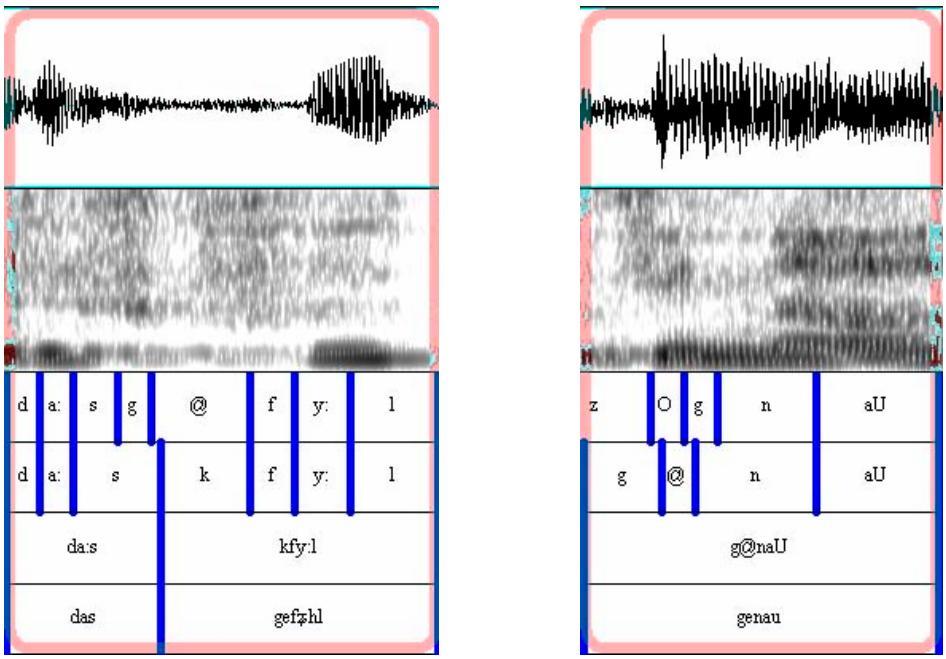
	ssprl. realisiert [g@]	**richtig transkribiert**	**regional realisiert [g/k]**	**richtig transkribiert**
Kempten [k]	4	3 (75%)	10	5 (50%)
Kempten [g]	4	2 (50%)	10	4 (40%)
Nordhorn [g]	23	19 (83%)	keine	keine
Offenburg [g]	9	8 (89%)	7	4 (57%)

Diese Tendenz zur vokalhaltigen Transkription ist vermutlich dadurch zu erklären, dass die Segmentgrenzen beim automatischen Segmentieren häufig nicht exakt getroffen werden. So ergeben sich leichte Verschiebungen bei der Text-Ton-Synchronisation und dann wird z.B. bei synkopierten Formen der folgende Tonsilbenvokal vom Aligner irrtümlich für den Vokal des Präfixes gehalten und als solcher transkribiert.

Der Vergleich beider Testläufe für die Aufnahme Kempten zeigt, dass sich Unterschiede ergeben, die sich auf die Verwendung des einen oder des anderen Phonmodells zurückführen lassen. So wird die Erkennungsleistung bei [k] als Alternative neben [g@] sowohl bei der ssprl. Aussprache mit Vokal im Präfix als auch bei der regionalsprachlich synkopierten um je einen Treffer verbessert. Diese verbesserten Ergebnisse finden sich bei den beiden Beispielen in Abb. 4 in der zweiten Transkriptspur. Beim linken Beispiel wurde zunächst [g@fy:l] transkribiert, danach richtig [kfy:l], beim rechten Beispiel wurde zunächst [gnau] transkribiert, danach richtig [g@nau]. An diesem Einzelbeispiel zeigt sich also, dass mit einer ent-

sprechend detaillierten Modellierung der tatsächlichen Aussprache durch Regeln die automatische Transkriptionsleistung durchaus verbessert werden kann.

Abb. 4: Transkription von ssprl. <ge-> mit zwei unterschiedlichen Regelsets (Aufnahme Kempten)



5. Probleme allgemeiner Art

An den gezeigten Beispielen wurden einige recht spezifische Problematiken an Einzelfällen demonstriert. In diesem Abschnitt soll noch auf einige generelle Faktoren eingegangen werden, die die automatische Transkription wesentlich beeinflussen.

Wie man sich vorstellen kann, hat vor allem die Qualität der jeweiligen Tonaufnahme einen entscheidenden Einfluss auf die automatische Transkription. Das betrifft insbesondere außersprachliche Störgeräusche verschiedenster Art, die kurzzeitig oder dauerhaft mit der Sprachaufnahme interferieren, aber natürlich auch die grundsätzliche technische Qualität der Aufnahme. Dann sind es vor allem die Eigenschaften der natürlich gesprochenen Sprache, die die Spracherkennung und damit die automatische Transkription erschweren. Darunter fallen Hesitations- und Rezeptionsphänomene ebenso wie sprechsprachliche Verschleifungen und Reduktionen, die sich vor allem in unbetonten Positionen ergeben. Schließlich beeinflussen auch nonverbale Phänomene wie Atmen, Husten usw. das Alignment, denn das System muss sie als solche erkennen und darf sie nicht fälschlich für Sprachsignale halten.

Diese genannten generellen Faktoren wirken sich insofern auf die automatische Transkription aus, als sie die korrekte Synchronisation von Text und Ton erschweren. Die automatische Variantenerkennung ist aber nur auf der Grundlage von gut synchronisierten (alignierten) Daten möglich. Darum sind z.B. die Alignmentergebnisse bei gelesener Sprache, bei der viele der oben genannten Faktoren wegfallen oder nur abgeschwächt auftreten, wesentlich besser

als z.B. bei einem Dialog. Im Dialog kommen als zusätzliche Hürde für das Alignment auch noch Simultanpassagen hinzu, in denen sich die Sprachsignale von zwei Gesprächspartnern überlagern. Bei den von uns bewerteten Transkriptionsergebnissen sind im Übrigen nur die Fälle berücksichtigt worden, bei denen die Text-Ton-Synchronisation prinzipiell in Ordnung war.

6. Fazit

Die beschriebenen Tests und Erfahrungen haben gezeigt, dass die automatische Transkription und damit Detektion regionaler Varianten beim gegenwärtigen Stand der Technik mit einem Spracherkennungssystem vom Typ wie es am IDS zur Verfügung steht, sehr fehleranfällig ist und nicht ohne weiteres als wissenschaftliches Analyseinstrument benutzt werden kann. Bei den meisten getesteten Variablen ist die Trefferquote zu gering. Von den vier oben beschriebenen Tests waren einzig bei [S] und [s] die Erkennungsraten akzeptabel. An mehreren Stellen des Systems könnte man durch zusätzlichen Arbeitsaufwand sicher Verbesserungen der Transkriptionsleistung herbeiführen, z.B.

a) mit dem Neutraining zumindest bestimmter bereits vorhandener Phone (wie zum Beispiel bei stimmhaftem [z] und stimmlosem [s]);

b) mit der Erweiterung des Phoninventars um Phone, die für bestimmte Regionalsprachen typisch sind (koronalisiertes [C], aber z.B. auch dunkles bairisches [A] usw.);

c) mit der weiteren Präzisierung und Erweiterung des Regelsatzes, auch um Regeln für sprechsprachliche Phänomene und um ein Ausnahmewörterbuch, in dem gerade hochfrequente Formen bestimmter Lexeme, die nicht sinnvoll über GPK-Regeln ableitbar sind, vorkommen (wie z.B. bei *nicht*, *nich*, *nüsch*, *ni*, *net*, *it* usw.).

Diese Verbesserungen können der Text-Ton-Synchronisation bei regionalsprachlichen Aufnahmen insgesamt zu Gute kommen. Sie können so z.B. bei der Aufbereitung und Alignierung großer regionalsprachlicher Korpora, wie sie am IDS vorliegen, nach Möglichkeit Anwendung finden.

Abschließend lässt sich Folgendes festhalten: Zwar kann mit dem automatischen Transkriptionssystem keine fehlerfreie lautschriftliche Transkription hergestellt werden. Dieses System kann aber sehr wohl dafür verwendet werden, eine digitale Basistranskription zu erstellen, die dann auf eine relativ komfortable Weise manuell verbessert werden kann. Mit diesem quasi halbautomatischen Verfahren kann die Analyse der Variation in großen Korpora und die Erstellung einer Datenbank von Varianten – wie sie am Anfang des Beitrags beschrieben wurde – immerhin unterstützt und beschleunigt werden.

Literatur:

Almeida, Antonio/Braun, Angelika (1982): Probleme der phonetischen Transkription. In: Besch, Werner/Knoop, Ulrich/Putschke, Wolfgang/Wiegand, Herbert Ernst (Hg.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. 1. Halbbd. Berlin. S. 597-615.

Barbour, Stephen/Stevenson, Patrick (1998): Variation im Deutschen. Soziolinguistische Perspektiven. Berlin/New York.

Bodmer, Franck/Fach, Marcus L./Schmidt, Rudolf/Schütte, Wilfried (2002): Von der Tonbandaufnahme zur integrierten Text-Ton-Datenbank. Instrumente für die Arbeit mit Gesprächskorpora. In: Pusch, Claus D./Raible, Wolfgang (Hg.): Romanistische Korpuslingustik: Korpora und gesprochene Sprache. Romance Corpus Linguistics: Corpora and Spoken Language. (= Script Oralia 126). Tübingen, S. 209-243.

Duden (2000): Aussprachewörterbuch. Wörterbuch der deutschen Standardaussprache. 4. Aufl. Mannheim.

Durrell, Martin (1995): Sprachliche Variation als Kommunikationsbarriere. In: Popp, Heidrun (Hg.): Deutsch als Fremdsprache: An den Quellen eines Faches. Festschrift für Gerhard Helbig zum 65. Geburtstag. S. 417-428.

Fach, Marcus L. (2001): Automatische Segmentierung, Verwaltung und Abfrage von Korpora gesprochener Sprache. (= AIMS Working Papers. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung 7 (1)). Stuttgart. [Im Internet unter: `http://www.ims.uni-stuttgart.de/~fach/Diss.pdf` (Stand: März 2005)].

Herrgen, Joachim (1986): Koronalisierung und Hyperkorrektion. Das palatale Allophon des /CH/-Phonems und seine Variation im Westmitteldeutschen. Stuttgart.

Hosom, John-Paul (2000): Automatic time alignment of phonemes using acoustic-phonetic information. Internet: `http://www.ece.ogi.edu/~hosom/hosom_thesis.pdf` (Stand: März 2005).

König, Werner (1988): Zum Problem der engen phonetischen Transkription. In: Zeitschrift für Dialektologie und Linguistik 55, S. 155-178.

König, Werner (1989): Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland. 2 Bde. Ismaning.

Krech et al. (Hg.) (1982): Großes Wörterbuch der deutschen Aussprache. Hrsg. v. dem Kollektiv Eva-Maria Krech, Eduard Kurka, Helmut Stelzig, Eberhard Stock, Ursula Stötzer, Rudi Teske unt. Mitw. v. Kurt Jung-Alsen. Leipzig.

Pfeffer, Alan J./Lohnes, F. Walter (Hg.) (1984): Grunddeutsch. Texte zur gesprochenen deutschen Gegenwartssprache. 3 Bde. (= Phonai 28-30). Tübingen.

Schmidt, Rudolf (2000): Maschinelle Text-Ton-Synchronisation in Wissenschaft und Wirtschaft. In: Schmitz, Klaus-Dirk (Hg.): Sprachtechnologie für eine dynamische Wirtschaft im Medienzeitalter. Tagungsakten der XXVI. Jahrestagung der Internationalen Vereinigung Sprache und Wirtschaft e.V. 23.-25. November 2000. Fachhochschule Köln. Wien. S. 69-79.

Schmidt, Rudolf (2001): Instrumente zur Erstellung multimedialer Gesprächskorpora. In: Lobin, Henning (Hg.): Sprach- und Texttechnologie in digitalen Medien. Proceedings der GLDV-Frühjahrstagung 2001, 28.-30. März 2001. Justus Liebig-Universität Gießen. Norderstedt. S. 115-127.

Schmidt, Rudolf (2002): Automatic text-to-speech alignment: A method for handling disturbances and simultaneous utterances. In: Busemann, Stephan (Hg.): KONVENS 2002, 6. Konferenz zur Verarbeitung natürlicher Sprache. Proceedings. Saarbrücken, 30.09.-02.10.2002. Kaiserslautern/Saarbrücken.

Schmidt, Rudolf/Neumann, Robert (1999): Automatic Text-to-Speech-Alignment: Aspects of Robustification. In: Matousek, Václav/Mautner, Pavel/Ocelíková, Jana/Sojka, Petr (Hg.): Text, speech and dialogue. Second international workshop. Proceedings. TSD '99, Plzen, Czech Republic, September 13-17, 1999. (= Lecture notes in computer science 1692: Lecture notes in artificial intelligence). Berlin/Heidelberg/New York u.a. S. 72-76.

Siebs, Theodor (1969): Deutsche Aussprache: Reine und gemäßigte Hochlautung mit Aussprachewörterbuch. Hrsg. v. Helmut de Boor, Hugo Moser u. Christian Winkler. 19., umgearb. Aufl. Berlin.