Intonation, Ordnungseffekt und das Paradigma der kategorialen Wahrnehmung

Lieselotte Schiefer, Anton Batliner

Angaben zur Veröffentlichung / Publication details:

Schiefer, Lieselotte, and Anton Batliner. 1988. "Intonation, Ordnungseffekt und das Paradigma der kategorialen Wahrnehmung." In *Intonationsforschungen*, edited by Hans Altmann, 273–92. Tübingen: Niemeyer.

INTONATION, ORDNUNGSEFFEKT UND DAS PARADIGMA DER KATEGORIALEN WAHRNEHMUNG

Lieselotte Schiefer und Anton Batliner (München)

1. EINLEITUNG

Etablierte, genau beschriebene Modelle haben idealerweise zwei positive Eigenschaften: (1) Wenn man innerhalb des Modells definierte Begriffe verwendet, weiß jeder, wovon man spricht. (2) Eine strikte Definition ist auch Voraussetzung für die Falsifizierbarkeit des Modells bzw. einzelner Teilannahmen. Oft gibt es allerdings mehrere, sich ergänzende, aber auch miteinander konkurrierende Definitionen und Kriterien - so auch beim Paradigma der Kategorialen Wahrnehmung. Dieses Paradigma ist als Instrument gedacht, mit dem man kategoriale Wahrnehmung von nicht-kategorialer, d.h. kontinuierlicher Wahrnehmung abgrenzen kann. Es wurde entwickelt auf segmenteller Materialbasis bei der Wahrnehmung von Plosiven (vgl. Libermann et al. 1957) und mit der Zeit erweitert auf z.B. Vokale und nicht-sprachliche Stimuli. Wir legen im folgenden die wohl bekannteste Spielart des Modells, die der Haskins-Laboratories, zugrunde. Eine ausführliche Diskussion würde den Rahmen dieser Arbeit übersteigen; sie findet sich auch schon im Sammelreferat von Repp (1984), auf das wir uns in diesem Beitrag öfter beziehen werden.

2. DAS PARADIGMA DER KATEGORIALEN WAHRNEHMUNG

Ein Experiment im Paradigma der Kategorialen Wahrnehmung besteht typischerweise aus einem Identifikations- und einem Diskriminationstest (von nun an 'IT' und 'DT'). Im IT werden äquidistante Stimuli eines physikalischen Kontinuums, das zwei (gelegentlich auch mehrere) Kategorien enthält, in randomisierter Folge Hörern dargeboten, die diese einer der vorgegebenen Kategorien zuordnen müssen (*forced choice*). Beim DT werden benachbarte oder auch weiter entfernte Stimuli, die zu Paaren (AX-Test) zusammengestellt sind, auf Gleichheit beurteilt. Es werden dabei drei Stimulusabfolgen (AB, BA, und AA bzw. BB) getestet. Die erhaltene Diskriminationskurve setzt sich aus den richtigen Antworten (*hits*) zu allen Stimulusfolgen zusammen. In den meisten Untersu-

chungen wurde statt des AX-Tests der ABX-Test verwendet, in dem Stimulus-Tripel dargeboten werden, bei denen die beiden ersten Stimuli stets verschieden sind, Stimulus X dagegen Stimulus A oder B sein kann. Da in unseren Experimenten ausschließlich der AX-Test Verwendung fand, gehen wir auf den ABX-Test nicht ausführlicher ein. Wir entschieden uns für den AX-Test, da die Beanspruchung des Gedächtnisses (*memory load*) und damit die Wahrscheinlichkeit, daß die dargebotenen Stimuli zuerst klassifiziert und dann miteinander verglichen werden, beim Vergleich von nur zwei Stimuli (AX) geringer ist als beim Vergleich von drei Stimuli (ABX). Die Hypothese, daß der Mensch kategorial **wahrnimmt** und nicht nur **klassifiziert** (wie der Name sagt: 'Kategoriale Wahrnehmung', nicht 'Kategoriale Klassifikation'), wird also mit dem AX-Test einer strengeren Prüfung unterzogen (vgl. Repp 1984:266).

Kategoriale Wahrnehmung in dem strikten Design des Paradigmas ist dann anzunehmen, wenn die folgenden vier Kriterien (Repp 1984: 253) erfüllt sind:

> 1. Labeling probabilities change abruptly somewhere along the continuum; in other words, the identification functions have a rather steep slope. The point of maximum slope is the **category boundary** (equivalently defined as the point at which responses in two adjacent categories are equiprobable).
> 2. Discrimination functions show a peak at the category boundary; that is, stimuli are more easily discriminated when they fall on opposite sides of the boundary than when they fall on the same side.
> 3. Discrimination performance within each category is at or near chance level.
> 4. Discrimination functions are perfectly predictable from the labeling probabilities (using one of the simple formulae provided by the Haskins model [...]. This implies that (a) the discrimination peak is in exactly the right place and of the right height, and (b) the labeling probabilities are appropriate; that is, they apply independently of the context in which they were observed.

Das 1. Kriterium (*steepness of labeling functions*) ist schwach (vgl. Repp 1984:253), da die Steilheit der Identifikationsfunktion nicht unwesentlich von den Abständen der Stimuli zueinander (*spacing* bzw. Schrittgröβe) abhängt. Das 2. Kriterium fordert einen Gipfel in der Diskriminationskurve im Bereich der Kategoriengrenze, an der Stelle der maximalen perzeptiven Ungleichheit benachbarter Stimuli (*phoneme boundary effect*). Dieser Aspekt wird von Repp (1984, 253) als wesentlicher als die Kurvensteilheit beurteilt. Daβ dieses Kriterium jedoch trotz gegebener Steilheit der IT-Kurve nicht immer erfüllt sein muβ, wird im folgenden noch gezeigt werden. Das 3. Kriterium impliziert, daβ Stimuli, die der gleichen Kategorie angehören, nicht oder nur zufällig (*chance level*) diskriminiert werden können. Und das 4. Kriterium schließlich setzt voraus, daβ

die Diskriminationsleistung eindeutig aus der Identifikationsleistung voraussagbar ist z.B. nach der Haskins-Formel

$$pcor = .5 + [(p_1 - p_2)^2/2].$$

pcor steht für die Wahrscheinlichkeit einer korrekten Diskrimination, p_1 für die Wahrscheinlichkeit, daß A einer der Kategorien und p_2 für die Wahrscheinlichkeit, daß B der gleichen Kategorie zugeordnet wird (vgl. auch Cutting 1982). Die berechnete (*predicted*) kann mit der erhaltenen (*obtained*) Kurve unter Verwendung eines statistischen Verfahrens (z.B. Varianzanalyse, s. Repp et al. 1979) auf die Güte der Anpassung (*goodness of fit*) verglichen werden. Somit scheint einzig dieses Kriterium einer mathematisch-statistischen Falsifizierung zugänglich.

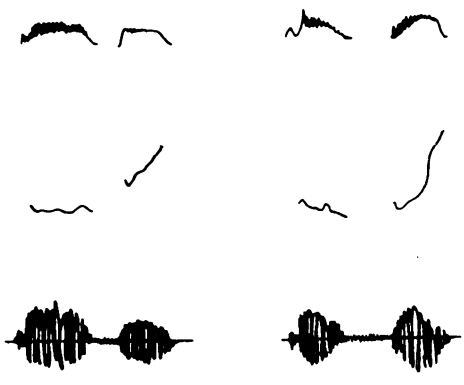
3. ZWEI EXPERIMENTE ZUR AKZENTWAHRNEHMUNG

Testaufbau

Material: Einer der Autoren (A.B.) produzierte im Tonstudio des Instituts für Phonetik in München mehrmals den Fragesatz *Kommen Sie?* in zwei Varianten, wobei der Akzent zum einen auf *Kommen*, zum anderen auf *Sie* lag (Telefunken M15, 19 m/sec.) Als Eckstimuli für die Generierung eines Kontinuums wurden diejenigen Realisationen ausgewählt, die von mehreren kompetenten Beurteilern als die natürlichsten eingestuft wurden und deren intonatorische Parameter die gewünschten Ausprägungen hatten (auditiv überprüft und kontrolliert mit Hilfe von Fo-Meter und Intensity-Meter). Fig.1 zeigt schematisierte Mingogramme dieser beiden Eckstimuli mit dem Zeitsignal, dem Grundfrequenz- (Fo-) Verlauf sowie der Intensität.

Fig.1: Eckstimuli (*Kommen Sie*):

links: Akzent auf *kommen*
rechts: Akzent auf *Sie*
oben: Intensität
Mitte: Fo (Grundfrequenz)
unten: Zeitsignal



Mit den auf einer PDP11/50 digitalisierten und segmentierten Eckstimuli wurde ein 10-stufiges Kontinuum unter Verwendung eines Programmpakets zur Interpolation zwischen natürlich produzierten Sprachsignalen erzeugt. (Dieses Paket ist genauer beschrieben in Simon 1983; zu den in der Zwischenzeit vorgenommenen Verbesserungen vgl. Hadersbeck 1987). Mit dem Verfahren können die Parameter Fo-Verlauf, Intensitätsverlauf, spektraler Energieverlauf, segmentale und damit auch suprasegmentale Zeitstruktur jeweils isoliert oder, wie in diesem Fall, gemeinsam behandelt werden. Für die Experimente ergaben sich also neben den resynthetisierten Eckstimuli acht synthetisierte, bezüglich der perzeptiv relevanten Parameter in ihrer physikalischen Ausprägung äquidistante Zwischenstufen.

Versuchspersonen waren acht Studenten der Linguistik bzw. der Phonetik, die freiwillig an den Experimenten teilnahmen.

Design: Es wurde ein IT und darauffolgend, mit einer Woche Abstand, ein DT durchgeführt. Beim IT wurden die zehn Testitems je zehnmal mit einem zeitlichen Abstand von 3.5 sec zwischen den einzelnen Items dargeboten. Beim DT wurden die Items im Einerschritt (1/2, 2/3, ... 2/1, 3/2 ...) gepaart und die drei unterschiedlichen Anordnungen AA, BB, AB bzw. BA mit einem Abstand von 3.5 sec zwischen den Paaren und 500 msec innerhalb der Paare je fünfmal randomisiert dargeboten. Die Versuchspersonen saßen dabei im Sprachlabor des Instituts für Phonetik vor einem Abstimmkästchen und hörten die Stimuli über die Raumlautsprecher. Genau nach jedem Stimulus(-paar) wurde die Abstimmung freigegeben; dies wurde durch eine Lampe an jedem Kästchen angezeigt. Die Versuchspersonen hatten dann drei Sekunden Zeit, ihre Antwort zu überlegen und die entprechende Taste zu drücken. Die Antworten wurden auf einer PDP11/03 gesammelt und zur weiteren Verarbeitung aufbereitet. Beim IT lautete die Instruktion: "Bitte drücken Sie die linke Taste, wenn der Akzent Ihrer Ansicht nach auf *kommen*, oder die rechte, wenn er auf *Sie* liegt." Beim DT lautete die Instruktion: "Bitte entscheiden Sie, ob die beiden Stimuli innerhalb eines Paares in allen Belangen gleich klingen oder nicht, und drücken Sie dann die linke Taste für 'gleich' oder die rechte für 'verschieden'."

Statistische Auswertung: Es wurden zweifaktorielle Varianzanalysen gerechnet. Im Rahmen unserer Argumentation ist allerdings eine adäquate Prüfstatistik problematisch, da wir verschiedene Konstellationen mit dem gleichen Datenmaterial berechnen müssen. Es ist aber auch nicht unser Ziel, mit prüfstatistisch signifikanten Ergebnissen zu argumentieren. Wir wollen nur illustrieren, welche Ergebnisse man erhält, wenn man den eigenen Entscheidungskriterien unterschiedliche Berechnungen zugrunde legt. D.h. wir tun so, **als ob** wir jeweils nur eine Berechnung durchführen würden. Wir geben deshalb normalerweise nur die Wahrscheinlichkeitswerte *p* an und sprechen nicht davon, daß das Ergebnis signifikant ist, sondern daß es signifikant sein **würde** - bei Zugrundelegung jeweils nur einer bestimmten Datenkonstellation.

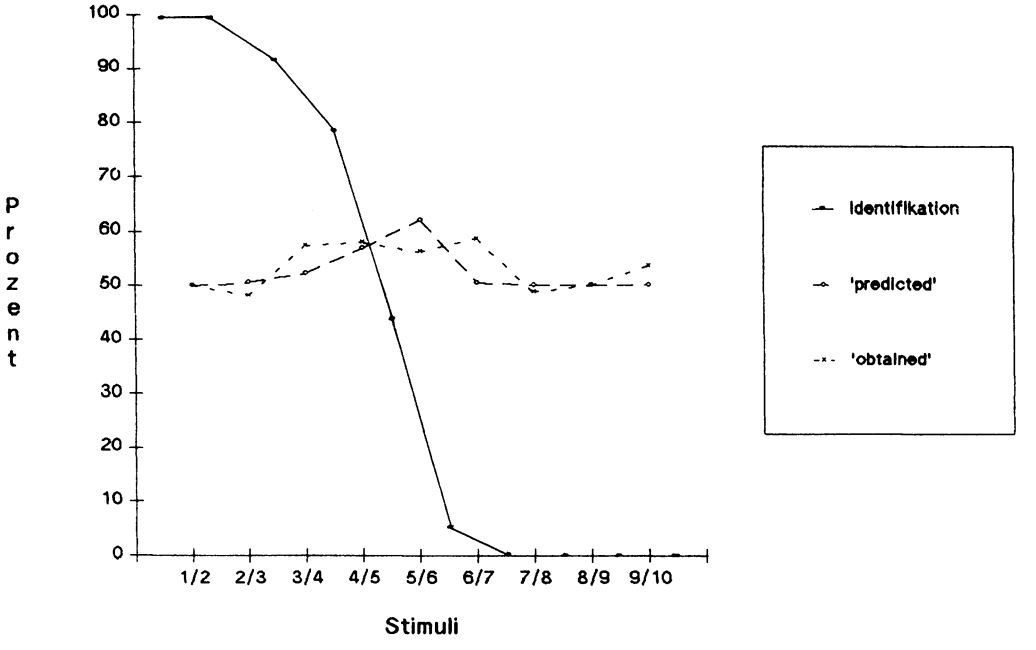
Ergebnisse

Die Versuchspersonen unterteilten im IT das Kontinuum in zwei Kategorien (Fig.2). Dabei wurde in den ersten vier Stimuli der Akzent auf *kommen*, in den letzten fünf Stimuli auf *Sie* wahrgenommen. Stimulus fünf konnte keiner Kategorie eindeutig zugeordnet werden.

Die aus den Ergebnissen des IT errechnete *predicted*-Kurve (ebenfalls Fig.2) weist einen niedrigen Diskriminationsgipfel bei Stimulus-Paar 5/6 auf. Die *obtained*-Kurve zeichnet sich durch ein niedriges Plateau von Paar 3/4 bis Paar 6/7 aus, das seinen niedrigsten Wert bei Paar 5/6 (sic!) hat, also genau an derjenigen Stelle, an der die *predicted*-Kurve ihren Gipfel aufweist. Trotz unterschiedlicher Form wären beide Kurven jedoch nicht signifikant verschieden ($p<.5$).

Haben also die bei der Generierung der Test-Stimuli vorgenommenen Manipulationen der akustischen Parameter zu einer kategorialen Wahrnehmung des Akzents geführt oder muß man einen kontinuierlichen Übergang von einer Kategorie in die andere annehmen? Zur Beantwortung dieser Frage seien die in Teil 2 diskutierten Kriterien auf die Test-Ergebnisse angewandt. Kriterium 1 (*labeling*

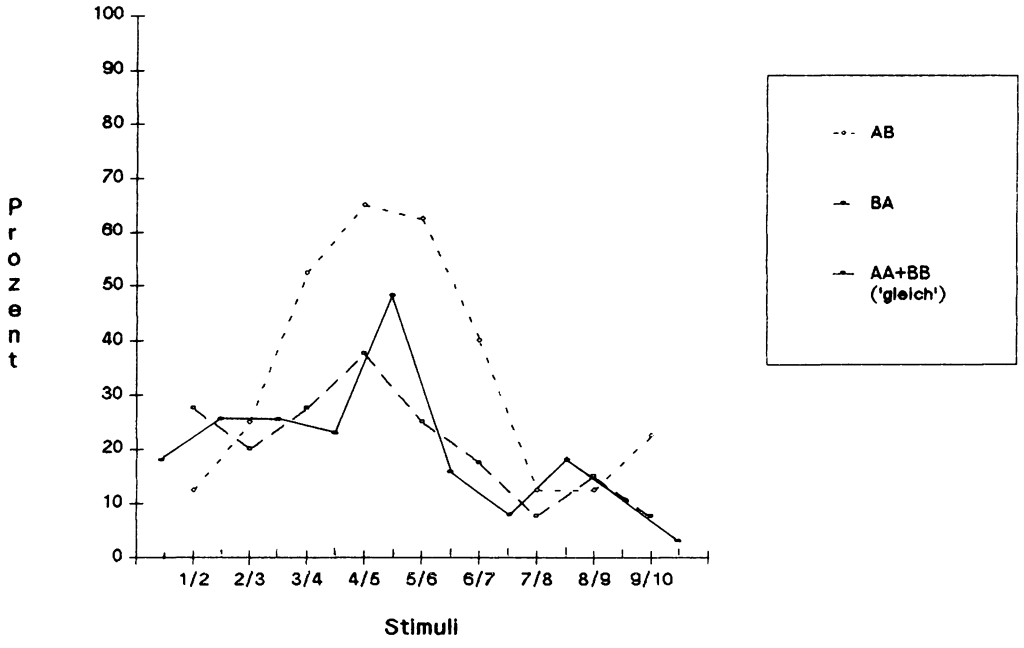
Fig.2: Identifikation und Diskrimination ('predicted' vs. 'obtained')



probabilities change abruptly) setzt einen abrupten Übergang von der einen zur anderen Kategorie im IT voraus. Dieses Kriterium scheint in unserem Fall nicht voll erfüllt, da Stimulus 5 keiner Kategorie eindeutig zugeordnet wurde. (Das mag mit der Schrittgröße *(spacing)* im Kontinuum zu tun haben; vgl. dazu unten Teil 4.) Kriterium 2 (*discrimination functions show a peak at the category boundary*) ist in keiner Weise erfüllt, da kein eindeutiger Diskriminationsgipfel existiert und darüberhinaus sogar eine Erniedrigung der Kurve bei Stimulus-Paar 5/6 vorliegt. Dagegen scheint das 3. Kriterium (*discrimination performance within each category is at or near chance level*) voll erfüllt. Die Ergebnisse entsprechen jedoch ebenfalls nicht Kriterium 4, da die Diskriminationsfunktion nicht aus dem IT voraussagbar ist (daran ändert auch die Tatsache nichts, daß die zu vergleichenden Kurven statistisch nicht signifikant verschieden wären).

Da drei der vier Kriterien nicht (voll) erfüllt sind, sprechen die Ergebnisse gegen eine kategoriale Wahrnehmung. Bedeutet dies jedoch, daß kontinuierliche Wahrnehmung anzunehmen ist? Bevor ein solches Urteil gefällt werden kann, müssen u.E. die Ergebnisse für die unterschiedlichen Darbietungsformen (AB, BA, AA bzw. BB) einzeln untersucht werden. Diese sind in Fig.3 getrennt dargestellt.

Fig.3: Diskrimination (AB, BA, 'gleich')



Drei wesentliche Ergebnisse sind zu diskutieren:

(1) Die 'verschieden'-Paare in Anordnung AB werden häufiger als verschieden perzipiert als die entsprechenden Paare in der Anordnung BA ($p<.001$). Beide Kurven zeigen einen Diskriminationsgipfel bei Paar 4/5; allerdings ist bei der AB-Kurve der Wert für das benachbarte Paar 5/6 fast genauso hoch. Das Phänomen der von der Anordnung der Stimuli innerhalb der Test-Paare abhängigen unterschiedlichen Beurteilung ansonsten gleicher Paare bezeichnen wir als **Ordnungseffekt**.

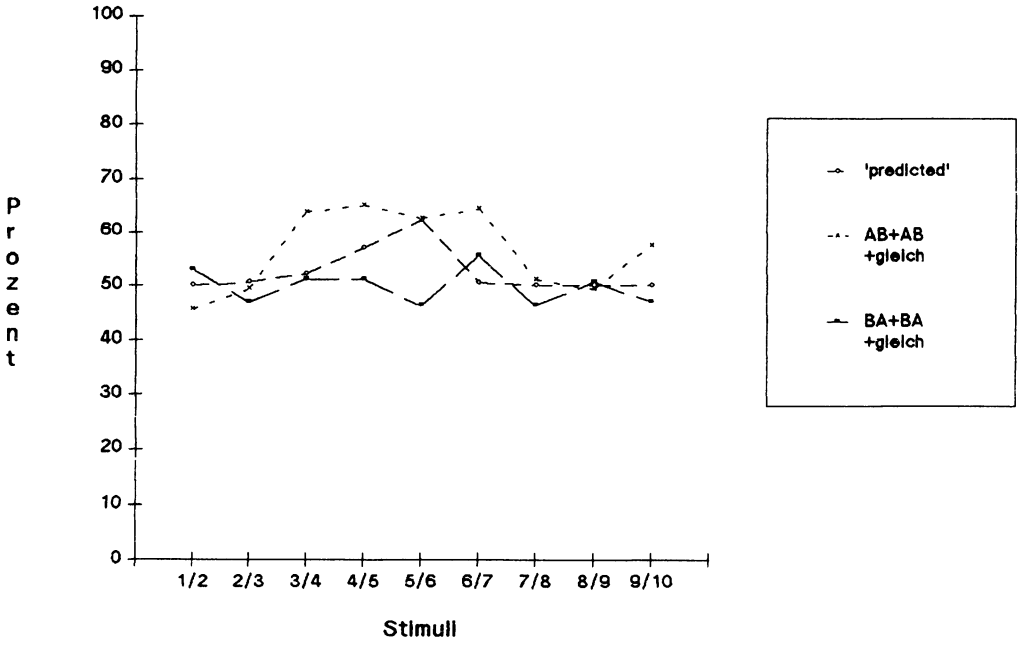
(2) Paar 1/2 wird in der Anordnung BA besser diskriminiert als in der Anordnung AB, während die Ergebnisse für Paar 9/10 in der AB-Darbietung besser sind. Beide Paare enthalten einen unmanipulierten (Stimulus 1 resp. 10) und einen manipulierten Stimulus (Stimulus 2 resp. 9). (Das erklärt auch die besseren Ergebnisse für die Eckpaare in allen berechneten Diskriminationskurven. Wir vermuten, daß bedingt durch ein Artefakt der Stimulusgenerierung die Intensität dieser resynthetisierten natürlichen Ausgangsstimuli im Verhältnis zu den synthetisierten Zwischenstufen zu hoch ist. Dies führt dann zu einer unverhältnismäßig guten Diskriminierbarkeit der Paare, bei denen der (prominente) Eckstimulus an zweiter Stelle steht. Der Fehler wurde inzwischen behoben, vgl. Hadersbeck 1987.)

(3) Bei den Ergebnissen für die 'gleich'-Paare (AA bzw. BB) weist Stimulus-Paar 5/5 einen eindeutigen Gipfel auf (nahe 50%). Dieses Ergebnis ist besonders interessant, da der Gipfel im Bereich der Kategorien-Grenze liegt, nämlich bei demjenigen Stimulus, der im IT keiner Kategorie eindeutig zugeordnet werden konnte. Vergleichbare Ergebnisse wurden von uns auch bei DTs zur Plosivwahrnehmung (Schiefer, bisher unveröffentlicht) gefunden und deuten darauf hin, daß ein systematischer Faktor dafür verantwortlich sein muß. Eine mögliche Erklärung wäre, daß der Hörer auch die im DT zum Vergleich dargebotenen Stimuli kategorisiert (klassifiziert), wobei zunächst der 1. Stimulus einer der vorgegebenen Kategorien zugeordnet wird. Der 2. Stimulus wird nicht auf seine akustische Struktur, sondern lediglich daraufhin überprüft, ob er der gleichen Kategorie angehört, die man dem 1. Stimulus zugewiesen hat. Ist dies der Fall, wird mit 'gleich' geantwortet, andernfalls mit 'verschieden'. Handelt es sich um einen kategoriell nicht eindeutig identifizierbaren Stimulus (in unserem Fall Stimulus 5), so wird zwar beim 1. Stimulus der gleiche Prozeß der Klassifizierung wie bei einem eindeutigen Stimulus durchlaufen, d.h. er wird einer der Kategorien zugeordnet, etwa 'Akzent auf *kommen*'. Der 2. Stimulus wird nun daraufhin

überprüft, ob der Akzent **eindeutig** auf *kommen* liegt. Da der Stimulus 'unentscheidbar' ist, ist dies nicht der Fall und es wird mit 'verschieden' geantwortet.

Wieweit beeinflussen nun diese Ergebnisse die Gesamt-Diskriminationsleistung (*obtained*-Kurve)? Es ist eindeutig, daß die Abflachung der *predicted*-Kurve im Bereich der Kategoriengrenze (die Paare 4/5 und 5/6) durch die *false alarms*, d.h. durch falsche 'verschieden'-Antworten auf 'gleich'-Paare bewirkt wird, da an beiden Paaren Stimulus 5 beteiligt ist. Basiert die Berechnung der Diskriminationskurve dagegen nicht auf AB **und** BA, sondern **nur** auf AB (Formel: *(AB+AB+AA+BB)/4*; vgl. Fig.4), so ist, ebenso wie bei der *obtained*-Kurve, kein Diskriminationsgipfel vorhanden, sondern die Ergebnisse für die Paare 3/4, 4/5, 5/6 und 6/7 sind fast identisch. Die Kurve liegt grundsätzlich **über** der *predicted*-Kurve ($p<.004$). Wird dagegen die Diskriminationskurve aufgrund der Ergebnisse für die BA-Darbietungsrichtung berechnet, so erhält man eine Kurve, die grundsätzlich **unter** der *predicted*-Kurve liegt ($p<.09$). Ein eindeutiger Gipfel fehlt auch hier; die beste Diskriminationsleistung erzielt Paar 6/7. Bei keiner der durchgeführten Berechnungen tritt ein Diskriminationsgipfel an der erwarteten Stelle auf.

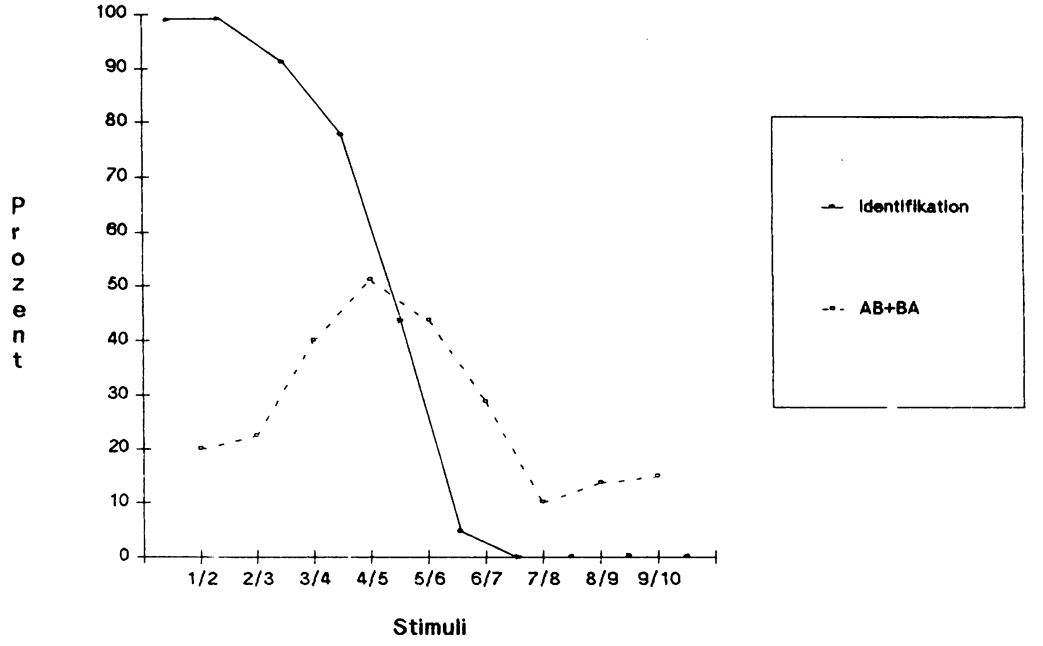
Fig.4: Diskrimination ('predicted', AB+AB+'gleich', BA+BA+'gleich')



Zusammenfassend kann für jede Art der Berechnung festgehalten werden: (1) Die *false alarms* für die 'gleich'-Paare führen zu einer Abflachung der Diskriminationskurve. (2) Die Einbeziehung der BA-Paare führt zu einer Erniedrigung, die der AB-Paare zu einer Erhöhung der Diskriminationskurve.

Es ist schließlich noch zu fragen, ob der Diskriminationsgipfel, der sich bei Addition der Antworten für die 'verschieden'-Paare (AB und BA) ergibt (gemittelte Kurve), wenn nicht in der Höhe so doch in der Position mit den Ergebnissen aus dem IT-Test übereinstimmt. Dies ist in Fig.5 dargestellt. Die Kurve zeigt einen Gipfel bei Paar 4/5 und entspricht damit den Ergebnissen aus dem IT, die einen Schnittpunkt der IT-Kurve mit der 50%-Linie zwischen Stimulus 4 und 5 aufweisen. (Man beachte, daß bei der Berechnung der *predicted*-Kurve der Schnittpunkt zwischen der IT-Kurve und der 50%-Linie nicht berücksichtigt wird, sondern ausschließlich die rechnerische Differenz zwischen den Antworthäufigkeiten für die einzelnen Stimuli maßgebend ist.)

Fig.5: Identifikation und Diskrimination (AB+BA)



Kehren wir nun zur Frage nach der Kategorialität der Ergebnisse zurück, die bei Heranziehung des Paradigmas in seiner strikten Form sicher zu verneinen ist. Die Alternative dazu - kontinuierliche Wahrnehmung - ist allerdings bei genauer Betrachtung der Kurven ebenfalls unplausibel. Ein Weg aus dem Dilemma ist eine weniger strikte Anwendung des Paradigmas, wobei etwa lediglich die Form der IT-Kurve und die gemittelte DT-Kurve einer heuristischen Beurteilung unterzogen werden. Bei einem solchen Vorgehen wird man sicher eine kategoriale Wahrnehmung annehmen. Der Preis dafür ist das Fehlen einer exakten Definition und der objektiven Meßbarkeit von "Kategorialität". Wir werden auf diesen Punkt in der Diskussion zurückkommen.

4. DER ORDNUNGSEFFEKT

Für die Ergebnisse aus IT und DT sind zumindest zwei Faktoren mit 'verantwortlich', die primär nichts mit Kategorialität, sondern mit der Konstruktion der Teststimuli und ihrer Darbietung im Experiment zu tun haben. Der eine Faktor ist bekannt: Die Schrittgröße, d.h. die Dichte der Abfolge der Stimuli im physikalischen Kontinuum (*spacing*). Ein zu geringer Abstand der Stimuli zueinander führt im IT zu einem weniger abrupten Übergang von einer zur anderen Kategorie. Im DT resultiert daraus einerseits eine Abflachung der Diskriminationskurve bei den 'verschieden' Paaren und eine Erhöhung der *false alarms* an der Kategoriengrenze bei den 'gleich'-Paaren. Ein zu groß gewählter Schritt dagegen würde im IT zwar zu dem gewünschten abrupten Kategorienwechsel, im DT jedoch zu einer zu starken Erhöhung der Diskriminationskurve (*ceiling effect*) für die 'verschieden'-Paare führen, während mit einer Reduzierung der *false alarms* bei den 'gleich'-Paaren zu rechnen ist. Man muß also gegebenenfalls in Pilotexperimenten die adäquate Schrittgröße ermitteln, die dann im eigentlichen Experiment getestet werden kann.

Ein zweiter, die Ergebnisse maßgeblich beeinflussender Faktor, ist der Ordnungseffekt, der sich in einer (häufig signifikant) besseren Diskriminierbarkeit von Stimulus-Paaren in einer der beiden Darbietungsrichtungen AB oder BA äußert. Der Ordnungseffekt wurde bereits im 19. Jahrhundert durch Fechner entdeckt, der ihn den 'constanten Fehler' bei der Ermittlung des 'Masses der Unterschiedsempfindlichkeit' nannte (Fechner 1964: 90ff). In der Folgezeit, besonders durch die Arbeiten von Woodrow (1935, 1951), Woodworth (1950), Stott (1933, 1935) sowie Woodrow/Stott (1936) wurde der Ordnungseffekt (in der englischsprachigen Literatur als *time-order error* bezeichnet) in der experimen-

tellen Psychologie einer intensiven Erforschung unterzogen. Ein *time-order error* konnte z.B. nachgewiesen werden bei der Wahrnehmung von Dauer, Länge, Lautstärke, Helligkeit und Geschmack. Als Erklärung für das Phänomen wurde zunächst eine Verknüpfung mit Gedächtnis- und Perzeptionsprozessen angenommen, während spätere Untersuchungen einen *subject bias* dafür verantwortlich machen, sei es in Form eines *simple response bias* (Luce 1959) oder eines *criterion bias* (Wickelgren 1968). Eine endgültige Erklärung für das Phänomen steht jedoch nach wie vor noch aus.

In der Psychophysik wird der Ordnungseffekt normalerweise durch eine Mittelung der Ergebnisse neutralisiert. Die Phonetik verfährt ähnlich; dort wurde mit wenigen Ausnahmen (s. etwa Repp et al. 1979; Smith 1976; Rosen 1977; Chuang-Wang 1978a, 1978b) der Ordnungseffekt nicht zur Kenntnis genommen und unseres Wissens nicht hinsichtlich seiner Auswirkung auf Experimente im Paradigma der Kategorialen Wahrnehmung thematisiert. Die Ergebnisse aus dem oben beschriebenen Experiment lassen die Annahme zu, daß der Ordnungseffekt eine systematische 'Störvariable' im Paradigma ist, die nicht als Test-Artefakt erklärbar ist.

5. EXPERIMENTE ZUM 'DESIGN-EFFEKT'

Mit den im folgenden geschilderten Experimenten wurde die Hypothese getestet, daß eine gemischte (durchrandomisierte) und eine getrennte Darbietung, bei der entweder nur die Anordnung AB oder die Anordnung BA auftritt, gleichermaßen zu einem signifikanten Ordnungseffekt führen. Dieser Annahme nach kann der Ordnungseffekt zwar durchaus vom experimentellen Design beeinflußt werden, er ist aber nicht **ausschließlich** auf dieses Design zurückzuführen.

Testaufbau

Material: Ausgangsmaterial für die Stimuli war ein natürlich produziertes, monotones *ja* (Aufnahmebedingungen wie oben in Teil 3 beschrieben). Dieser kurze und durchgehend stimmhafte Ausgangsstimulus wurde gewählt, um eine exakte Manipulation der interessierenden Parameter zu gewährleisten. Der Stimulus wurde unter Verwendung eines Segmentationsprogramms (für Einzelheiten s. Batliner/Schiefer 1987) in einzelne Perioden segmentiert. Die Fo-Manipulation der sieben generierten Stimuli begann stets im quasi-stationären Teil des Vokals, dessen Fo bei 85 Hz lag. Der Fo-Offset war bei allen Stimuli konstant bei 114 Hz. Bei Stimulus 1 erstreckte sich der Fo-Anstieg über die letzten 4 Perioden (das entspricht einer Dauer von 39.3 msec); für die übrigen Stimuli wurde der Fo-Anstieg um jeweils zwei Perioden verlängert, so daß Stimulus 7 einen Anstieg

über 16 Perioden (entsprechend 159.5 msec) aufwies. Das auf diese Weise generierte Kontinuum basierte somit auf unterschiedlicher Dauer und Steilheit (*slope*) des Fo-Anstiegs, während der Fo-Offset, der Fo-Range und die mittlere Fo in allen Stimuli unverändert war. Dauer und Steilheit des Fo-Anstiegs waren äquidistant; die Daueruntersch iede zwischen den einzelnen Stimuli betrugen ca. 20 msec, die durch die pitchsynchrone Manipulation (periodenweise Segmentierung, kein fixes Zeitfenster) bedingten Schwankungen lagen im Bereich von +/- 1.5 msec. Außer mit dem beschriebenen Kontinuum wurde mit zwei weiteren Kontinua gearbeitet, die auf der Manipulation anderer Fo-Parameter basierten. Damit sollte zum einen die perzeptive Relevanz der unterschiedlichen Parameter wie Fo-Offset, Dauer und Steilheit des Fo-Anstiegs und mittlere Fo überprüft werden; zum anderen sollten diese Experimente ermitteln, welche Anordnung innerhalb der Stimulus-Paare bei den verschiedenen Manipulationen besser diskriminierbar ist. Da die Ergebnisse aus den drei Kontinua sich in den hier interessierenden Punkten nicht unterscheiden, beschränken wir uns auf die Darstellung der Ergebnisse eines Kontinuums. (Es wird im folgenden aus Platzgründen auch nicht auf alle damit durchgeführten Tests eingegangen, da eine Diskussion der ITs und damit der Frage, ob überhaupt und in welchem Ausmaß unterschiedliche Kategorien im Kontinuum enthalten sind, zu weit führen würde und in unserem Zusammenhang auch nicht interessiert. Über diese Ergebnisse werden wir an anderer Stelle berichten.)

Design: Für die AX-DTs wurden die 7 Stimuli im 2er Schritt gepaart (1/3, 2/4, 3/5 ... 3/1, 4/2, 5/3 ...) und in randomisierter Folge je 5mal wiederholt. Das Interstimulus-Intervall betrug 500 msec, die Pause zwischen den Paaren 3.5 sec. Im 1. DT wurden alle Darbietungsfolgen (AB, BA, AA bzw. BB) randomisiert dargeboten. In den beiden weiteren DT wurden die Darbietungsrichtungen AB und BA jeweils mit den 'gleich'-Paaren getrennt randomisiert und getestet. Im 2. DT wurden zunächst die BA-Paare, dann die AB-Paare dargeboten, während im 3. DT die Reihenfolge umgekehrt war. Dies ergibt die folgenden Testkonfigurationen:

1. Diskriminationstest: AB, BA, AA/BB gemischt randomisiert
2. Diskriminationstest: 1.Teil: BA, AA/BB; 2.Teil: AB, AA/BB
3. Diskriminationstest: 1.Teil: AB, AA/BB; 2.Teil: BA, AA/BB

Die Darbietung der Tests erfolgte mit einer Revox 77 über Raumlautsprecher. Testpersonen waren die Teilnehmer (n = 10, 11, 15) eines Seminars. (Zur Instruktion und dem allgemeinen Prozedere vgl. oben Teil 3.) Da die Tests mit mindestens einer Woche Abstand durchgeführt wurden, betrachteten wir sie als 'unabhängig', also nicht als wiederholte Messungen an den gleichen Versuchspersonen. Für jeden Test rechneten wir zwei-faktorielle Varianzanalysen mit den Faktoren 'Paar' (1/3, 2/4, 3/5, 4/6, 5/7) und 'Darbietungsrichtung' (AB, BA).

Ergebnisse

Die Ergebnisse sind in Fig. 6-8 dargestellt. In allen DTs ist die Diskrimination der AB-Paare besser als die der BA-Paare (1.DT: $F(1,9)=79.83$; $p<.001$; 2.DT: $F(1,10)=24.35$; $p<.001$; 3.DT: $F(1,14)=75.77$; $p<.001$). Wir vergleichen die Tests zwar nicht prüfstatistisch untereinander, da dafür die Voraussetzungen fehlen, es läßt sich jedoch an Hand der Abbildungen feststellen, daß das unterschiedliche Design Einfluß auf die Ergebnisse hatte: Beginnt der Test mit der weniger prominenten Stimulus-Abfolge BA (Fig.7), so werden diese Paare besser diskriminiert als die gleiche Abfolge im durchrandomisierten Test (Fig.6); stehen dagegen die prominenten Paare AB an erster Stelle (Fig.8), so werden diese etwas schlechter diskriminiert als im durchrandomisierten Test (Fig.6). Wir haben es hier also mit einer gegenseitigen Beeinflußung der Darbietungsfolgen zu tun: die Anwesenheit der prominenten Paare AB im gleichen Test führt zu einer schlechteren Diskrimination der BA-Paare, während diese ihrerseits zu einer besseren Diskrimination der AB-Paare führen (*anchoring*; vgl. den folgenden Abschnitt). Die Ergebnisse aus den zweiten Hälften der nicht-durchrandomisierten

Fig.6: randomisiert

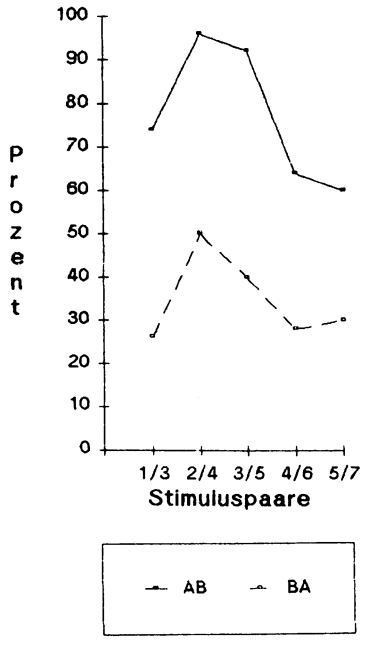


Fig.7: Teil1:BA, Teil2:AB

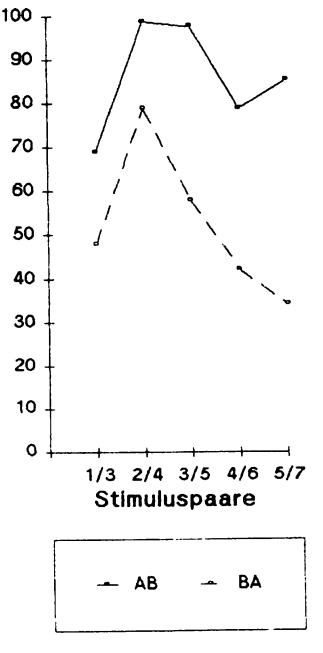
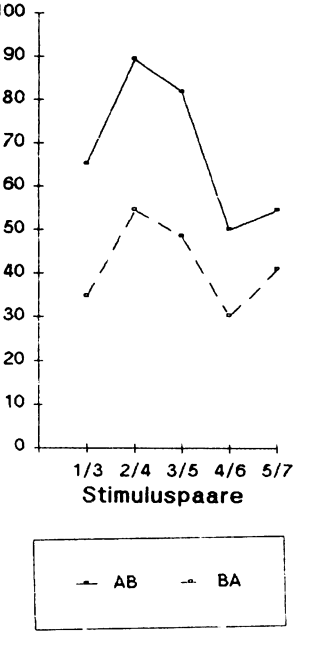


Fig.8: Teil1:AB, Teil2:BA



Tests unterscheiden sich nur geringfügig von den entsprechenden Verläufen des durchrandomisierten Tests; wir vermuten, daß dafür verschiedene, sich gegenseitig in ihrer Wirkung neutralisierende Faktoren verantwortlich sind.

Für eine Überprüfung dieser Ergebnisse wurde bei einem Fo-Kontinuum, das im Prinzip gleich aufgebaut war, aber eine andere Schrittgröße (30 msec statt 20 msec) hatte, das Testdesign geändert. Dabei prüften wir die Hypothese, daß trotz gemischter Darbietung von Paaren **unterschiedlicher** Schrittgröße (1/2, 2/3, 3/4 vs. 1/3, 2/4, 3/5 vs. 1/4, 2/5 usw.) auch bei den Paaren mit geringerem Unterschied (1er Schritt) ein signifikanter Ordnungseffekt auftritt. Diese Hypothese nimmt also an, daß sich ein Ankereffekt (*anchoring*), der bewirkt, daß die Versuchspersonen ihre Bewertung anhand der deutlichsten Unterschiede innerhalb eines Paares kalibirieren, nicht entscheidend auf den Ordnungseffekt auswirkt. Es wurden zwei DTs erstellt. Im 1. DT hatten die 'verschieden'-Paare die Schrittgröße 1, während im 2. DT alle möglichen Schrittgrößen (1er bis 6er Schritt) vorlagen. Die AB-Paare wurden gemäß unserer Annahme beim 1er und auch beim 2er Schritt signifikant besser als die BA-Paare diskriminiert ($p<.001$). Ab dem 3er Schritt stellte sich ein *ceiling effekt* ein: beide Anordnungen wurden sehr gut diskriminiert. Die Ergebnisse entsprechen ansonsten den oben diskutierten Experimenten.

Die Ergebnisse dieser beiden Testserien sprechen dafür, den Ordnungseffekt nicht als simplen Designeffekt aufzufassen. (Damit ist natürlich nicht ausgeschlossen, daß er grundsätzlich ein experimenteller Effekt ist. Es ist noch nicht klar, welche Rolle er in natürlichsprachlicher Kommunikation spielt.) Weitere Experimente zu den folgenden zwei Fragen wurden mit dem gleichen Material durchgeführt und sollen an anderer Stelle dargestellt werden: (1) Handelt es sich beim Ordnungseffekt um ein systematisches Phänomen, d.h. lassen sich bei der Wahrnehmung von Tonhöhen Bedingungen für die prominenten Paare angeben? (2) Ist ein unterschiedlicher Ordnungseffekt festzustellen, wenn Tonhöhenunterschiede bei sprachlichem bzw. nicht-sprachlichem Material wahrgenommen werden? Ohne der Diskussion dieser Ergebnisse vorzugreifen, kann jedoch festgestellt werden, daß **immer** eine prominente (besser diskriminierbare) Anordnung AB einer nicht-prominenten Anordnung BA gegenübersteht. Als prominent erwiesen sich bisher Anordnungen, bei denen (a) der höhere Offset (s. Teil 3), (b) die längere Dauer des Fo-Anstiegs (wie gerade gezeigt) oder (c) die größere Fo-Veränderung (vgl. Batliner/Schiefer 1987) an zweiter Stelle stand.

6. DISKUSSION

Es konnte nachgewiesen werden, daß der Ordnungseffekt einen entscheidenden Einfluß auf die Berechnung der Kategorialität im strikten Paradigma ausübt. Wir wollen nun drei mögliche, unterschiedlich weitreichende Folgerungen aus diesem Befund diskutieren:

(1) Das strikte Paradigma der Kategorialen Wahrnehmung ist abzulehnen. – Man könnte dagegen einwenden, daß das Paradigma nicht für die Wahrnehmung von Tonverläufen, sondern für die Wahrnehmung von Segmenten, insbesondere Plosiven konzipiert wurde. Der Ordnungseffekt läßt sich aber auch im segmentellen Bereich nachweisen, vgl. Chuang/Wang 1978a. Wir fanden ihn auch bei VOT- und Amplitudenwahrnehmung (Schiefer, bisher unveröffentlicht). U.E. blieb der Ordnungseffekt im Rahmen des Paradigmas nicht deswegen unbeobachtet, weil es ihn nicht gibt, sondern weil er bei den üblichen Testdesigns, z.B. dem ABX-Test, vgl. oben Teil 2, nicht so deutlich sichtbar wurde.

(2) Der Ordnungseffekt wird in das Paradigma integriert. – Es gibt natürlich Wege, den Ordnungseffekt in das Paradigma zu integrieren, etwa indem man ihn wie die Schrittgröße einbaut – damit wird er zum heuristischen Problem. Wir können uns allerdings keine befriedigenden Kriterien vorstellen, anhand derer er in das **strikte** Paradigma integrierbar wäre. Man müßte also einen anderen Kriterienkatalog aufstellen. Es bleibt zu fragen, ob es sich dann noch um das gleiche Paradigma handelt. Damit kommen wir zur dritten, am weitesten reichenden Folgerung:

(3) Das globale Paradigma der Kategorialen Wahrnehmung ist grundsätzlich abzulehnen. – Egal, welche Kriterien ein solches Paradigma ansetzt: Mit dieser Folgerung wird in Frage gestellt, daß man mit einem **einzigen** Paradigma sinnvoll **alle** lautsprachlichen Phänomene bearbeiten kann. Prima vista mag der Ordnungseffekt ein zu bescheidenes Phänomen sein, um eine solche Schlußfolgerung motivieren zu können. Wir wollen deshalb einige weitere relevante Punkte zumindest anführen, auch wenn wir sie nicht im Detail diskutieren können:

Die Kategoriengrenze: Der Ordnungseffekt tangiert insbesondere die Kriterien, mit denen die Beschaffenheit der Kategoriengrenze bestimmt wird. Über dieses innerhalb des Paradigmas doch grundlegende Konzept schreibt Repp (1984:320):

> One true shortcoming of the categorical perception paradigm is that it has overemphasized the importance of the boundaries between phonetic categories. [...] The boundaries [...] are a mere epiphenomenon, apparent only in a particular experimental situation. [...] beyond the realm of artificial speech continua, the boundary concept has little to offer.

Nun ist die Kategoriengrenze unmittelbar oder mittelbar an allen der vier eingangs vorgestellten Kriterien beteiligt. Welchen Wert hat aber ein Paradigma, das sich wesentlich auf dieses 'Epiphänomen' stützt, wobei das Epiphänomen als solches schon durch einen anderen systematischen Faktor, nämlich den Ordnungseffekt - egal, ob es sich bei diesem um ein 'echtes' Phänomen oder auch nur um ein Epiphänomen handelt - in Frage gestellt ist?

Der Begriff 'Kategorialität': Ausgangspunkt für das Paradigma war ein experimentelles Ergebnis: synthetisierte Plosivkontinua wurden diskontinuierlich, also kategorial wahrgenommen (Liberman et al. 1957; zur Geschichte des Paradigmas vgl. wieder Repp 1984). Da dies anfangs nur bei solchen **sprachlichen** Phänomenen zu beobachten war, folgerte man daraus: 'speech is special', die Wahrnehmung von Sprache ist kategorial, nicht kontinuierlich strukturiert. Im Laufe der Jahre mußte die Anfangsaussage zwar modifiziert werden - Vokale werden wenig bis gar nicht kategorial wahrgenommen, es gibt eine kategoriale Wahrnehmung auch im nicht-sprachlichen Bereich - , der Grundgedanke blieb aber attraktiv: Es gibt einfach sehr viele Kategorien (mit distinktivem Zeichencharakter) in der Sprache, und das Paradigma schien ein geeignetes Instrument zu sein, mit dem sich solche Kategorien im phonetischen Bereich beschreiben lassen. Der Begriff 'Kategorialität' ist also grundlegend, aber zugleich schillernd. Es gibt ihn, strikt definiert, im Paradigma, es gibt ihn in allen möglichen Varianten in der Linguistik, und es gibt ihn im normalen Sprachgebrauch als die Eigenschaft von jemand oder von etwas, unterschiedlichen Gruppen anzugehören. In der Praxis kann das dazu führen, daß man den Begriff, der innerhalb des Modells definiert ist, in einer weniger strikten oder gar umgangssprachlichen Bedeutung verwendet. Diese Verwendung ist problematisch, wenn stillschweigend der weniger strikte Begriff angewandt, aber doch die mit dem strikten Modell verbundene theoretische Aussage damit gemeint ist. Eine solche Begriffscamouflage zeigt sich z.B. bei Lindsay/Ainsworth (1985): einerseits eine explizite Bezugnahme auf das Modell, andererseits eine unzureichende Anwendung der Kriterien. Im Einzelfall läßt sich das natürlich einfach als unzulässiges Vorgehen kritisieren. Insgesamt ist es aber ein Phänomen der normativen und zugleich verwirrenden Kraft des faktischen Wissenschaftsdiskurses: Der Begriff wird in so unterschiedlichen Bedeutungen verwendet, daß er

bedeutungsleer wird. Dies beruht u.E. letztlich auf der inhaltlich inadäquaten **globalen** Verwendung dieses Begriffs, wie wir im folgenden Punkt skizzieren wollen.

Die Adäquatheit des Begriffs 'Kategorialität': Modelle haben die Tendenz, sich zu verselbstständigen. D.h. man bleibt zu sehr dem experimentellen Mikrokosmos verhaftet und vernachlässigt, daß das Modell ja nur ein Hilfsmittel sein soll, um die natürlichsprachliche Kommunikation besser verstehen zu können. (Vgl. auch das Schlußwort in Repp 1984:322: "It is to be expected [...] that the traditional methodology will eventually give way to new approaches that more directly address the important theoretical and practical problems raised by **communication in the real world**." [unsere Hervorhebung]) Nun ist die Modellierung der *communication in the real world* in Experimenten unterschiedlich lebensnah: Was den segmentalen Bereich betrifft, so ist die übliche Aufgabe, etwa innerhalb eines Kontinuums von [p] nach [b] zu differenzieren, per se unnatürlich. Der Mensch differenziert normalerweise keine Laute, sondern zumindest Wörter - etwa *Pein* vs. *Bein* - wenn nicht doch immer Satz- oder Textbedeutungen. Entscheidend bleibt aber, daß es in diesem Fall wirklich nur zwei Kategorien gibt, zwischen denen differenziert werden kann, eben Fortis und Lenis. Damit ist nicht gemeint, daß der Hörer keine Zwischenstufen zwischen [p] und [b] hören und ihnen keinen Sinn beimessen kann, etwa als Indikator der regionalen Herkunft des Sprechers. Solche Differenzierungen laufen aber auf einer anderen Ebene: sie haben **Anzeichen**-, aber keine **Zeichen**funktion. Bei der Intonation ist die Aufgabe, etwa zwischen Frage und Exklamativ zu unterscheiden, eher 'aus dem Leben gegriffen'. Man mag einwenden, daß diese Aufgabe viel komplexer ist als die Unterscheidung zwischen /p/ und /b/. Im Sinne eines *ecological approach*, also eines Ansatzes, der die Bedingungen der natürlichsprachlichen Kommunikation stärker ins Kalkül zieht, ist sie angemessener: Die Frage, ob es sich um ein [p] oder [b] handelt, beschäftigt den Phonetiker, aber nicht den 'normalen' Menschen. Die Frage, ob es sich um eine Frage oder einen Exklamativ handelt, hat dagegen jeden 'normalen' Menschen schon öfter beschäftigt. Allerdings sind die Kategorien Frage und Exklamativ weniger eindeutig unterschieden als etwa /p/ und /b/. Es gibt zwar auch eindeutige, prototypische Fragen versus eindeutige Nicht-Fragen (seien das nun Exklamative oder Aussagen; zu dieser Differenzierung vgl. Batliner 1988). Es gibt aber ebenfalls mehrere Abstufungen der Fragehaltigkeit, mit der eine Frage gestellt wird (vgl. dazu Batliner 1987). Wieder etwas anders liegt es bei der Akzentuierung, wie in unserem Testsatz *Kommen Sie.* Hier gibt es im Regelfall nur eine alternative Fokussierung - damit wird natürlich eine kategoriale Wahrnehmung begünstigt. Bei anderen Konstellationen ist aber auch eine

Doppelfokussierung möglich (wobei noch dahingestellt bleiben muß, ob sie sich auch in zwei gleichwertigen Akzentuierungen niederschlägt.)

Es gibt also ein ganzes Spektrum unterschiedlich strukturierter Kategorien ('echt' binäre mit scharfen Grenzen, binäre mit breiten Übergangsbereichen, durch den Kontext mehr oder weniger beeinflußbare, etc.), und es ist mehr als fraglich, ob sie sich gemeinsam mit ein und demselben Instrumentarium und ein und demselben Begriff 'Kategorialität' beschreiben lassen. Aus all diesen Gründen halten wir eine Abkehr vom strikten Paradigma besonders bei der Betrachtung von Intonationsphänomenen und eine vorsichtig dosierte Verwendung des Begriffs "kategorial" für wünschenswert.

LITERATUR:

Batliner, A. (1987): Kategorialität und Kontexteffekte bei Frage- und Exklamativmodus im Deutschen. Perzeptionsexperimente zur Rolle des Fo-Verlaufs. Ms.

Batliner, A. (1988): Der Exklamativ: Mehr als Aussage oder doch nur mehr oder weniger Aussage? Experimente zur Rolle von Höhe und Position des Fo-Gipfels. (In diesem Band).

Batliner, A. / Schiefer, L. (1987): Stimulus category, reaction time, and order effect - an experiment on pitch discrimination. Proceedings XIth ICPhS, Vol. 5: 46-49.

Chuang, C.-K. / Wang, W.S.-Y. (1978a): The time-order error in judgement of prosodic features: The pitch, the intensity, and the duration. J. Acoust. Soc. Am. 62: S48(A). (Zitiert nach Chuang/Wang 1978b).

Chuang, C.-K. / Wang, W.S.-Y. (1978b): Psychophysical pitch biases related to vowel quality, intensity difference, and sequential order. J. Acoust. Soc. Am. 64: 1004-1014.

Cutting, J. E. (1982): Plucks and bows are categorically perceived, sometimes. Perception and Psychophysics 31: 462-476.

Fechner, G. Th. (1964): Elemente der Psychophysik. Amsterdam, E.J. Bonset. (Nachdruck der Ausgabe Leipzig 1860).

Hadersbeck, M. (1987): A new program for manipulation of natural speech - interpolation between two natural utterances. Proceedings XIth ICPhS, Vol. 5: 35-38.

Liberman, A.M. / Harris, K.S. / Hoffman, H.S. / Griffith, B.C. (1957): The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54: 358-368.

Lindsay, D. / Ainsworth, W. A. (1985): Two models of nuclear intonation. Journal of Phonetics 13: 163–173.

Luce, R. D. (1959): Individual choice behavior. New York u.a., John Wiley & Sons.

Repp, B. H. (1984): Categorical perception: Issues, methods, findings. In: Lass, N. J. (Hg.): Speech and Language. Orlando u.a., Academic Press: 243–335.

Repp, B.H. / Healy, A.F. / Crowder, R.G. (1979): Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance 5: 129–145.

Rosen, S.M. (1977): The effect of fundamental frequency patterns on perceived duration. Speech Transmission Laboratory Quarterly Progress and Status Report 1: 17–30.

Simon, Th. (1983): Manipulation of natural speech signals according to the speech parameters of different speakers. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 17: 233–245.

Smith, M.R. (1976): An investigation of changes in categorical perception of vowels. Unpubl. Manuskript. Department of Linguistics, University of Connecticut. (Zitiert nach Repp et al. 1979).

Stott, L.H. (1933): The discrimination of short tonal duration. Ph.D. diss. University of Illionois Library. (Zitiert nach Woodrow 1951).

Stott, L.H. (1935): Time-order errors in the discrimination of short tonal durations. Journal of Experimental Psychology 18: 741–766.

Wickelgren, W.A. (1968): Unidimensional strength theory and component analysis of noise in absolute and comparative judgements. Journal of Mathematical Psychology 5: 102–122.

Woodrow, H. (1935): The effect of practice upon time-order errors in the comparison of temporal intervals. The Psychol. Rev. 42: 127–152.

Woodrow, H. (1951): Time perception. in: Stevens, S.S. (Hg.): Handbook of Experimental Psychology. New York u.a., John Wiley & Sons: 1224–1236.

Woodrow, H. – Stott, L.H. (1936): The effect of practice on positive time-order errors. Journal of Experimental Psychology 19: 694–705.

Woodworth, R.S. (1950²): Experimental psychology. New York, Holt, Rinehart, and Winston. (1. Auflage 1938).

Anmerkung: Die folgende umfangreiche Aufsatzsammlung erschien erst nach Fertigstellung dieses Aufsatzes und konnte deshalb nicht berücksichtigt werden:

Harnad, S. (Hg.) (1987): Categorical perception. The groundwork of cognition. Cambridge u.a., Cambridge University Press.