„Zentrum“ und „Peripherie“. Zur Bewertung der phonotaktischen Wortstruktur.

Sebastian Kempgen, Münster

0.

In korpusorientierten phonologischen Untersuchungen, beispielsweise zur Distribution (Phonotaktik), finden sich oft Bemerkungen derart, daß man Fremdwörter, Lehnwörter, onomatopoietische Wörter, Interjektionen, Eigennamen, also solche Einheiten, von denen man annimmt, daß sie sich in irgendeinerweise nicht voll in das System der Sprache (gemeint ist hier die Ausdrucksebene) eingegliedert hätten oder besonderen Regeln gehorchten, von vorneherein aus der Untersuchung ausgeschlossen habe (vgl. z.B. BLUHME 1971, 6 oder PANFILOV 1973, 7). Auf die Subjektivität eines solchen Vorgehens hat LEHFELDT (1971, 220f.) hingewiesen. Es ist also darauf zu achten, die Zirkelhaftigkeit der Argumentation zu vermeiden: Ob ein Wort in phonotaktischer Hinsicht zu einem bestimmten Zeitpunkt für eine Sprache *untypisch* ist oder nicht, sollte doch nicht *a priori* nach Gutdünken festgesetzt werden, sondern kann allenfalls *Ergebnis* einer entsprechenden Untersuchung sein. Der vorliegende Artikel ist einem Vorschlag zur Lösung der angedeuteten Problematik gewidmet. Wir sind der Ansicht, daß hier nur die Methoden der quantitativen Linguistik ein geeignetes Instrumentarium darstellen, da wir davon ausgehen, daß die Beurteilung, ob ein Wort in das phonotaktische System eingegliedert ist, nicht einer einfachen kategorischen Feststellung (ja/nein) unterliegt, daß vielmehr verschiedene *Grade der Eingliederung* anzunehmen sind. Ganz ähnliche Einsichten finden sich bei ALTMANN/LEHFELDT (1976). Es bleibt also die Aufgabe, diese Eigenschaft meßbar zu machen.

1.

Ausgehend von der Forderung, daß, will man eine Abgrenzung in zentrale und periphere *phonotaktische* Einheiten vornehmen, dann hierzu nur *phonotaktische*, nicht aber morphologische, semantische, historische o.ä. Kriterien heranzuziehen sind, schlagen wir vor, den folgenden Weg zu wählen. Als die Einheit, die wir untersuchen wollen, wählen wir das phonologische Wort (vgl. PULGRAM 1970, LEHFELDT 1971). Im übrigen kann das vorgeschlagene Verfahren sinngemäß auch auf andere Größen angewandt werden.

1.1

Nach dem bei LEHFELDT (1971) dargelegten Verfahren werden alle am Wortanfang bzw. -ende vorkommenden Konsonantenverbindungen einem statistischen Test unterworfen. Zunächst werden aus einem genügend umfangreichen Korpus die absoluten Häufigkeiten der einzelnen Verbindungen festgestellt. (Der Terminus ‚Konsonant' ist hier rein funktional zu verstehen als ein Phonem, das nicht den Kern einer Silbe bilden kann. Kroatisch /krk/ hat also genau wie /rab/ am Wortanfang wie am Wortende je einen Konsonanten.) Daraufhin wird für jede Verbindung ihr mathematischer Erwartungswert berechnet. Nach dem Verhältnis des tatsächlichen Wertes zum theoretisch berechneten Wert sowie – falls erforderlich – nach dem Ausgang eines statistischen Tests der Abweichung des tatsächlichen Wertes vom Erwartungswert auf ihre Signifikanz hin werden alle existierenden Verbindungen in zwei Klassen eingeteilt, in *marginale* (m) und *regelmäßige* (r). Alle nichtvorkommenden Verbindungen sind selbstverständlich ebenfalls marginal.

1.2

Mit diesem statistischen Kriterium und unter Anwendung der von PULGRAM (1970) ausgearbeiteten und von LEHFELDT (1971) verbessert übernommenen Regeln ist eine eindeutige *Silbentrennung* möglich, eine Voraussetzung für unser weiteres Vorgehen. Es wird nun für ein beliebiges Wort i die Silbentrennung durchgeführt, wobei der Status (*m* oder *r*) der auftretenden Silbenonsets bzw. -codas notiert wird.Wir bilden dann einen Index aus der Zahl der marginalen zur Gesamtzahl der in i auftretenden Onsets bzw. Codas. Formal können wir das so ausdrücken:

$$F(i) = \frac{n_m(i)}{n_m(i) + n_r(i)}$$

Den erhaltenen Wert wollen wir den *Grad der Fremdheit von i* im phonotaktischen System der jeweiligen Sprache nennen. Der Grad der Fremdheit ist komplementär zum *Grad der Verankerung:*

$$V(i) = \frac{n_r(i)}{n_m(i) + n_r(i)} = 1 - F(i)$$

$n_m(i) + n_r(i)$ ist natürlich leicht zu berechnen als das Zweifache der Zahl der Silben von i (da ja jede Silbe stets einen Onset und eine Coda hat). Die Zahl der Silben wiederum ist (im Russischen z.B.) gleich der Zahl der

Vokale. Wie man leicht einsehen kann, liegen alle möglichen Werte unseres Indexes im Einheitsintervall, d.h. $F(i) \in < 0;1 >$.

Tabelle 1 zeigt einige (die wohl häufigsten) Werte von F(i).

Tabelle 1. Einige Werte von F(i)

	n_m									
Silben	0	1	2	3	4	5	6	7	8	9
1	0,0	0,50	1	--	--	--	--	--	--	--
2	0,0	0,25	0,5	0,75	1	--	--	--	--	--
3	0,0	0,1667	0,333	0,5	0,6667	0,833	1	--	--	--
4	0,0	0,125	0,25	0,375	0,5	0,625	0,75	0,875	1	--
5	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
6	0,0	0,0833	0,1667	0,25	0,3333	0,4167	0,5	0,5833	0,6667	0,75
7	0,0	0,0714	0,1429	0,2143	0,2857	0,3571	0,4286	0,5	0,5714	0,6429
8	0,0	0,0625	0,125	0,1875	0,25	0,3125	0,375	0,4375	0,5	0,5625
9	0,0	0,0556	0,1111	0,1667	0,2222	0,2778	0,3333	0,3889	0,4444	0,5
10	0,0	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45
11	0,0	0,0454	0,0909	0,1364	0,1818	0,2273	0,2727	0,3182	0,3636	0,4091
12	0,0	0,0417	0,0833	0,125	0,1667	0,2077	0,25	0,2917	0,3333	0,375
13	0,0	0,0385	0,0769	0,1154	0,1538	0,1923	0,2308	0,2692	0,3077	0,3462
14	0,0	0,0357	0,0714	0,1071	0,1429	0,1786	0,2143	0,25	0,2857	0,3214
15	0,0	0,0333	0,0667	0,1	0,1333	0,1667	0,2	0,2333	0,2667	0,3

1.3

Wir wollen das Verfahren an einigen Beispielen für das Russische erläutern.

1.3.1

vremja, /vr'ém'a/, ist zweisilbig: /vr'é-m'a/, hat also zwei Silbenonsets und zwei Codas. Eine Berechnung des Status russischer Konsonantenverbindungen aufgrund der bei BALDWIN (1969) angegebenen Häufigkeiten nach dem in 1.1 geschilderten Verfahren hat ergeben, daß /vr'/ am Wortanfang marginal ist. Der Onset /m'/ ist regelmäßig. Regelmäßig sind auch die beiden „Codas", da im Russischen der vokalische Silbenauslaut in jedem Falle regelmäßig ist. Wir kommen also zu folgenden Werten:

n_m (/vr'é-m'a/) = 1 und n_r (/vr'é-m'a/) = 3.

Daraus bilden wir den Quotienten

F (/vr'ém'a) = 1/(1 + 3) = 0,25.

Der Grad der Fremdheit von *vremja* im phonotaktischen System des Russischen ist also 0,25, während der Grad der Verankerung oder Eingliederung gleich 0,75 ist.

1.3.2

Das einsilbige *ritm*, /r'itm/, hat einen regelmäßigen Onset und eine marginale Coda. Für /r'itm/ kommen wir also zu folgendem Indexwert:

F(/r'itm/) = 1/(1 + 1) = 0,5.

Wir können also sagen, daß der Grad der Fremdheit gleich dem Grad der Verankerung ist; weiter ist *ritm* dem Russischen doppelt so fremd wie *vremja* (nur in phonotaktischer Hinsicht!).

Eine Generierung einsilbiger Wörter i, für die gilt, daß F(i) = 1, d.h. solcher Wörter, deren Silbenstruktur völlig marginal wäre, ergibt z.B. folgendes:

*vrelt', /vrelt'/,
*gvart', /gvart'/,
*ptorm, /ptorm/,
*vkretm, /fkr'etm/.

Es wäre interessant zu untersuchen, ob sich die Hypothese, daß es im Russischen kein einsilbiges Wort mit zugleich marginalem Onset und marginaler Coda gibt, halten läßt. Andernfalls wäre für einsilbige Wörter des Russischen 0,5 der höchste tatsächlich beobachtbare Wert.

1.3.3

Wir haben uns oben gegen eine aprioristische Bewertung der Fremdheit ausgesprochen. Unser Vorgehen verlangt in dieser Hinsicht nur eine Beschränkung. Da wir die Silbenstruktur zur Grundlage unserer Messung gemacht haben, müssen alle Wörter, die auf diese Weise untersucht werden sollen, *silbenfähig* sein, was für das Russische nichts weiteres heißt, als daß sie mindestens einen Vokal enthalten müssen. Durch diese Vorschrift werden solche Einheiten wie *gm*, /gm/, oder *tss*, /ts/, aus dem in die Untersuchung eingehenden Korpus ausgeschlossen. Im Gegensatz zu den oben kritisierten ist dieses Kriterium jedoch ein formal bestimmbares, kein intuitives.

2.

Mit der Berechnung dieses Quotienten, die sehr einfach ist, kann das Problem aber noch nicht als gelöst betrachtet werden. (Da bisher keine geeigneten Daten vorliegen und im Rahmen dieser Arbeit auch nicht gewonnen werden können, müssen wir uns darauf beschränken, das Vorgehen im folgenden theoretisch dazustellen.) Ein wichtiger Schritt ist jedoch getan: Wir haben einen metrischen Begriff eingeführt, der zunächst einmal eine Präzisierung der Fragestellung erlaubt: von welchem Wert des Indexes F(i) an sind wir berechtigt, von einem *niedrigen,* einem *mittleren* oder einem *hohen* Grad der Ausprägung der gemessenen Eigenschaft zu sprechen? Ist der für /r'itm/ ermittelte Wert von 0,5 so hoch, daß dieses Wort eher der Peripherie als dem Zentrum im phonotaktischen System des Russischen zuzuordnen ist? Wir können uns nicht damit zufriedengeben, hier eine intuitive Beurteilung vorzunehmen. Mit dieser Fragestellung verbunden ist die Notwendigkeit, nicht bei der Berechnung numerischer Werte stehenzubleiben, sondern diese Werte als Grundlage für eine „qualitative" linguistische Interpretation zu nehmen. Wir wollen also jedes Wort i in eine von drei Klassen einordnen können (wobei die Wahl von gerade drei Klassen konventionell ist; das Verfahren kann im Prinzip auch zur Festsetzung der Grenzen einer beliebigen größeren Anzahl von Klassen angewendet werden).

2.1

Um hier weiterzukommen, wäre folgendermaßen zu verfahren (der Weg entspricht dem bei LEHFELDT 1975): nach dem in 1.2 geschilderten Vorschlag wird eine große Anzahl von Wörtern bearbeitet. Das Korpus muß dabei groß genug sein, um – im statistischen Sinne – signifikante Ergebnisse zu liefern. Auf diese Weise gewinnen wir Angaben über die Häufigkeit regelmäßiger bzw. marginaler Onsets und Codas (nach ihrem Umfang getrennt für null-, ein-, zwei- bis n-phonemige Verbindungen). Daraus ergibt sich, mit anderen Worten, die empirische Verteilung unserer Variablen F = „Zahl der marginalen Onsets/Codas zur Summe aller Onsets/Codas eines Wortes". Diese empirischen Verteilungen sind jetzt durch theoretische Verteilungen zu approximieren, wobei selbstverständlich die Güte der Näherung zu testen ist (dies sind wohl die schwierigsten Schritte des gesamten Verfahrens).

Ist dies geschafft, können wir weiter folgendes tun:

2.2

Wir berechnen aus der theoretischen Verteilung den *Erwartungswert* unserer Variablen F. Dann sind zwei Zahlen zu finden, die die untere bzw. die obere Grenze eines 95%-Konfidenzintervalls, das wir um den Erwartungswert legen, bilden. Anders ausgedrückt: wir suchen zwei Zahlen, t_1 und t_2, von denen gilt, daß die Wahrscheinlichkeit, mit der die Variable F gerade diese Zahlen erreicht und t_2 nach oben bzw. t_1 nach unten überschreitet, gleich 0,025 ist. Die Rangfolge dieser Werte läßt sich anschaulich so darstellen:

$F \in < 0, \ldots t_1, \ldots t_2, \ldots 1>$.

Mit anderen Worten: erhalten wir einen Wert,der zwischen 0 und t_1 oder zwischen t_2 und 1 liegt, so ist die Wahrscheinlichkeit, daß ein solches Ergebnis durch das Spiel des „Zufalls“ allein erreicht wird, minimal. Wir haben dann Grund zu der Annahme, daß hier ein anderer, „struktureller“ Zug der Sprache deutlich wird. Jetzt können wir mit den folgenden Entscheidungsregeln zu der gewünschten Klassifikation kommen:

a) Ist F(i) kleiner oder gleich t_1, dann sprechen wir von einem *signifikant niedrigen Wert*;

b) Ist umgekehrt F(i) größer oder gleich t_2, dann nennen wir das Ergebnis *signifikant hoch*.

c) Ist schließlich F(i) gleichzeitig größer als t_1 und kleiner als t_2, dann sprechen wir von einem mittleren oder *normalen* Grad der Ausprägung der hier untersuchten Eigenschaft.

Für die Klasse, die sich durch die Anwendung der Regel b) ergibt, können wir auch die Bezeichnung *peripher* verwenden.

Für phonotaktische Untersuchungen könnte man z.B. jetzt sagen: Ich beziehe nur solche Objekte i in meine Untersuchung ein, von denen gilt, erstens, i hat mindestens einen Vokal, und zweitens i weist keinen signifikant hohen Grad an Fremdheit der Silbenstruktur zum betrachteten Zeitpunkt in der untersuchten Sprache auf. – Solche Abgrenzungskriterien wären nicht mehr intuitiv, sondern können jederzeit empirisch überprüft werden.

3.

Auch für sprachtypologische Zwecke läßt sich unser Index nutzbar machen. Dazu definieren wir den *Durchschnittswert der Fremdheit* der Wörter im phonotaktischen System der Sprache L als

$$\overline{F}_L = \frac{\sum \mathrm{F(i)}}{N_i}$$

wobei wir mit N_i die Zahl der untersuchten Wörter bezeichnen. Ebensogut kann man dem Sprachvergleich auch die *durchschnittliche Verankerung* zugrundelegen, die wieder zu F komplementär ist:

$$V_L = 1 - \overline{F}_L$$

4.

Es soll abschließend noch einmal ausdrücklich darauf hingewiesen werden, daß unser Vorgehen in mehrfacher Weise nicht statisch ist, sondern der Sprachveränderung Rechnung trägt: ändert sich die Frequenz einer Verbindung, so kann sich auch ihr Status ändern, was eine Neuberechnung zeigen würde. Dies hat u.U. zur Konsequenz, daß in zwischenvokalischen Konsonantenketten, in der die betreffende Verbindung auftritt, die Silbengrenze an einer anderen Stelle anzusetzen ist. Damit ändern sich aber auch die empirischen Verteilungen der regelmäßigen bzw. der marginalen Onsets bzw. Codas, woraufhin zu prüfen ist, ob die bisher angenommene theoretische Verteilung noch „gut genug“ ist. Muß eine neue Approximation gesucht werden, ergeben sich damit auch andere Klassengrenzen. Zusammen mit veränderten Indexwerten können wir in verschiedenen Entwicklungsstadien einer Sprache so zu ganz verschiedenen Ergebnissen kommen.

Literatur

ALTMANN, G., LEHFELDT, W. (1976): *Einführung in die quantitative Phonologie* (im Druck).

BLUHME, H. (1971): Notes on Polish Phonoactics, in: *Linguistics* 69, 5-23.

LEHFELDT, W. (1971): Ein Algorithmus zur automatischen Silbentrennung, in: *Phonetica* 24, 212-237.

PANFILOV, E.D. (1973): *Fonologičeskie slogi klassičeskoj latyni* (Issledovanie spiska). Čast' I. Odnosložnye slovoformy, Leningrad (Izdatel'stvo Leningradskogo universiteta).

PULGRAM, E. (1970): Syllable, Word, Nexus, Cursus (= *Janua Linguaram, Series Minor* Nr. 81), The Hague—Paris (Mouton).

LEHFELDT, W. (1975): Die Verteilung der Phonemanzahl in den natürlichen Sprachen, in: *Phonetica* 31, 274-287.

BALDWIN, J.R. (1969): Syllable division in Russian, in: *Z.f. Phonetik* 22, 211-217.