

## Codierung natürlicher Sprache auf morphologischer Ebene

Sebastian Kempgen

Für einen Code sind mehrere Eigenschaften charakteristisch: die Anzahl der verwendeten Symbole (das „Alphabet“), die Regeln der Kombination dieser Elemente und außerdem die Länge solcher Kombinationen („Codewörter“). Hierbei können Codes mit *konstanter* und solche mit *variabler* Codewortlänge unterschieden werden. Die binäre Codierung des Alphabets natürlicher Sprachen auf Computern ist ein Beispiel für einen Code mit fester Wortlänge: jeden Buchstaben ersetzt man durch eine achtstellige Kombination aus 0/1-Zeichen („Bits“). Die natürlichen Sprachen, soweit sie in Buchstaben geschrieben werden, sind dagegen Codes mit *variabler* Wortlänge, da es ja lange und kurze Wörter gibt. Dies ist ein Weg, die Redundanz zu begrenzen und Ökonomie zu verwirklichen. Wenn ein Code künstlich konstruiert wird, so ist es u.a. Ziel dieser Konstruktion, den Zeitaufwand für die Übermittlung einer Nachricht möglichst gering zu halten. Ein in diesem Sinne optimaler Code muß offenbar „den häufigsten Symbolen die kürzesten Codewörter und, umgekehrt, den seltensten die längsten Codewörter zuschreiben“ (PADUČEVA 1961, 112). Es ist bekannt, daß natürliche Sprachen wie das Deutsche oder das Russische dieser Forderung im Prinzip genügen, und es kann sogar angenommen werden, daß die Informationsübertragung im Nervensystem des Menschen nach diesem Prinzip angelegt ist (vgl. die Hinweise bei PADUČEVA 1961, 113). Ein Blick in ein Häufigkeitswörterbuch wie ZASORINA (1977) zeigt sofort, daß häufige Wörter (wie z.B. die sogenannten „Funktionswörter“) besonders kurz sind. Die umfangreiche Literatur zur Darstellung der Länge eines Wortes als Funktion seiner Häufigkeit geht dabei auf die klassischen Arbeiten von G.K. ZIPF (1949) und sein „principle of least effort“ zurück.

Diese Gedankengänge wollen wir nun auf die morphologische Ebene übertragen. Was für Wörter insgesamt gilt, kann, muß aber nicht auch für Morpheme gelten. Wir wollen deshalb prüfen, ob dies der Fall ist. Bekanntlich werden ja nicht nur Wörter, sondern auch die einzelnen Wortformen eines und desselben Wortes aus verschiedenen Gründen mit stark unterschiedlicher Häufigkeit gebraucht. Folge dieser Tatsache ist es z.B., daß bei älteren Sprachstufen wie dem Kirchenslawischen u.U. nicht einmal alle Formen eines paradigmatischen Musterwortes tatsächlich auch belegt sind. Aber auch in der russischen Gegenwartssprache existieren bestimmte Wortformen einzelner Lexeme eher „virtuell“ denn tatsächlich (vgl. etwa die Bemerkungen bei ISAČENKO 1975, 406 zum Infinitiv der Iterativa, die vorwiegend im Präteritum gebraucht werden).

Wenn nun die Morpheme nach dem Prinzip eines optimalen Codes konstruiert wären, so müßte demnach erwartet werden, daß dem häufigsten Inhalt der kürzeste Ausdruck, dem seltensten Inhalt dagegen der längste Ausdruck entspricht. Da es aufgrund anderer Prinzipien (wie der Fehlererkennung und der Fehlerkorrektur, also der Sicherung gegen Störungen in der Übertragung) nicht anzunehmen ist, daß eine natürliche Sprache einzig auf minimalen Zeitaufwand hin optimiert ist, so wollen wir uns nicht mit einer einfachen ja/nein-Entscheidung begnügen; es soll vielmehr der Grad der Annäherung an die genannte Eigenschaft eines optimalen Codes bestimmt werden. Ein solcher Wert kann dann auch zum Sprachvergleich benutzt werden.

Konkret wollen wir die Frage, ob Morpheme im Russischen nach dem „Prinzip der kleinsten Anstrengung“ konstruiert sind oder nicht, an dem einfachen Beispiel der Präsensformen überprüfen.

Die Präsensformen drücken die sechs möglichen Kombinationen der grammatischen Kategorien Person (1., 2., 3.) und der grammatischen Kategorie Numerus (Sg., Pl.) aus. Die Flexionsendungen sind dabei synchron betrachtet eines von drei Mitteln zum Ausdruck dieses Inhaltes: morphologische Alternationen und Akzentparadigmen kommen oft, aber nicht immer, hinzu.

Zunächst brauchen wir Daten über die allgemeine Häufigkeit der sechs Präsensformen. Solche Angaben lassen sich aus dem Wörterbuch von ŠTEINFELDT (1963, 141–167) gewinnen, da die Autorin im Wörterverzeichnis zu jedem Verb genau angibt, mit welcher Häufigkeit seine einzelnen Formen in der zugrundegelegten Stichprobe aufgetreten sind. Eine Summierung aller dieser Einzelwerte, bis auf die Häufigkeit von *est'*, führt zu folgendem Ergebnis:

	1.Ps.	2.Ps.	3.Ps.	Σ
Sg.	2124	1144	7855	11123
Pl.	1256	800	3261	5317
Σ	3380	1944	11116	16440

Wenn man die Werte in den Zellen dieser Tabelle auf die Gesamtsumme bezieht, so ergeben sich die folgenden prozentualen Anteile:

	1.Ps.	2.Ps.	3.Ps.	Σ
Sg.	12,92%	6,96%	47,78%	67,66%
Pl.	7,64%	4,86%	19,84%	32,34%
Σ	20,56%	11,82%	67,62%	100,00%

Ordnet man die Inhalte nach der Häufigkeit, mit der sie verwendet werden, so ergeben sich aus den genannten Daten die folgende *Ränge*: 1) 3.P.Sg., 2) 3.Ps.Pl., 3) 1.Ps.Sg., 4) 1.Ps.Pl., 5) 2.Ps.Sg., 6) 2.Ps.Pl. Es ist ohne Zweifel auffällig, daß die Reihenfolge ganz klar lautet: 3. Person – 1. Person – 2. Person, oder, in einer etwas anderen Terminologie, Referent – Sprecher – Hörer. Bei einer Optimierung des Russischen auf morphologischer Ebene nach der Länge müßte dem häufigsten Inhalt, nämlich der „3.Ps.Sg.“, der kürzeste Ausdruck entsprechen usw. entsprechend der Rangreihenfolge.

Betrachten wir die tatsächliche Länge der Präsensmorpheme im Russischen. In Buchstaben gezählt, ergibt sich folgendes:

1.Ps.Sg.: -у/-ю:	1	1.Ps.Pl.: -ем/-им:	2
2.Ps.Sg.: -ешь/-ишь:	3	2.Ps.Pl.: -ете/-ите:	3
3.Ps.Sg.: -ет/-ит:	2	3.Ps.Pl.: -ут/-ят:	2

oder, geordnet:	1)	1.Ps.Sg.
	2) -4)	3.Ps.Sg., 1.Ps.Pl., 3.Ps.Pl.
	5) -6)	2.Ps.Sg., 2.Ps.Pl.

Augenfällig im Sinne einer Bestätigung unserer Hypothese ist, daß die 2.Ps.Sg. und die 2.Ps.Pl. in beiden Bereichen – Häufigkeit auf der Inhaltsseite und Länge der Ausdrucksseite – die letzten Plätze einnehmen, so daß folglich auch die übrigen Elemente mindestens ähnlich verteilt sein müssen.

Wir wollen prüfen, ob dieser Augenschein einer präzisen Bewertung standhält. Dazu stellen wir die Ränge eines jeden Elementes in Bezug auf Häufigkeit und Länge zusammen, wobei Elementen gleichen Ranges der Durchschnitt der ihnen zukommenden Rangzahlen zugeschrieben wird:

	Rang Häufigkeit	Rang Länge	$D_i$	$D_i^2$
1.Ps.Sg.	3	1	2	4
2.Ps.Sg.	5	5,5	0,5	0,25
3.Ps.Sg.	1	3	2	4
1.Ps.Pl.	4	3	1	1
2.Ps.Pl.	6	5,5	0,5	0,25
3.Ps.Pl.	2	3	1	1
				$\Sigma$ 10,50

Rechts sind die einfachen und die quadrierten Differenzen aufgeführt, die wir zur Berechnung des SPEARMANSchen Rangkorrelationskoeffizienten benötigen. Er ist für den Fall, daß Elemente mit gleichen Rängen auftreten,

folgendermaßen definiert (vgl. ALTMANN/LEHFELDT 1980, 201; umgeformt bei SIEGEL 1976, 197):

$$r_s = \frac{K^3 - K - 6(\sum T_x + \sum T_y) - 6\sum D^2}{\sqrt{(K^3 - K - 12\sum T_x)(K^3 - K - 12\sum T_y)}}$$

Hierbei ist

$$T = \frac{t^3 - t}{12}.$$

$K$  steht für die Anzahl der Elemente,  $t$  für die Anzahl der Elemente mit jeweils gleichen Rängen. Wir erhalten für unseren Fall:

$$\begin{aligned} r_s &= \frac{6^3 - 6 - 6(0 + 2.5) - 6(10.5)}{\sqrt{(6^3 - 6 - 12(0))(6^3 - 6 - 12(2.5))}} \\ &= \frac{216 - 6 - 15 - 63}{\sqrt{(216 - 6 - 0)(216 - 6 - 30)}} \\ &= \frac{132}{194.4222} = 0.6789 \end{aligned}$$

$$\text{da } \sum T_y = \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} = 0.5 + 2 = 2.5.$$

Die Werte von  $r_s$  liegen im Intervall  $\langle -1; 1 \rangle$ ; da wir einen positiven Wert erhalten haben, deutet dies auf eine Tendenz zu gleichen Rängen in beiden Bereichen. Transformiert man den Bereich  $\langle -1; 1 \rangle$  durch die einfache Umrechnung  $(x+1)/2$  in das Einheitsintervall  $\langle 0; 1 \rangle$ , so erhalten wir den Wert  $(0,6789+1)/2 = 0.8395$ . Mit anderen Worten: Der Grad, mit dem die russischen Präsenzmorpheme der Zeitoptimierung entsprechen, beträgt rund 84%.

Dieser Wert läßt sich nicht nur für den Sprachvergleich verwenden, sondern kann auch einzelsprachlich interpretiert werden, indem wir ihn auf seine Signifikanz überprüfen. Die Überlegungen, die hierzu im einzelnen angestellt werden, kann man etwa bei SIEGEL (1976) nachlesen. Es reicht hier, darauf hinzuweisen, daß man die untere Schwelle für einen mit mindestens 95% Sicherheit vorhandenen signifikanten Zusammenhang zwischen den beiden Rangverteilungen einer Tabelle entnehmen kann (s. SIEGEL 1976, 270). Bei  $n = K = 6$  und  $\alpha = 0.05$  und einseitiger Fragestellung liest man dort 0.829 als Schwellenwert ab. Da unser  $r_s = 0.6789$  ist, also kleiner als der geforderte Wert, so können wir keine 95%ige Sicherheit für

die Annahme eines nicht-zufälligen Zusammenhanges behaupten. Immerhin können wir festhalten, daß das Ergebnis unserer Ausgangshypothese nicht widerspricht – es hätten sich ja auch negativ signifikante Werte ergeben können.

Beim Bau der Präsensmorpheme des Russischen spielt die Optimierung auf möglichst geringen Zeitaufwand hin also durchaus eine Rolle, doch ist dies offensichtlich nicht die einzige Zielstellung; andere Faktoren sind ebenfalls zu berücksichtigen. Hierbei ist vor allem an eine andere Eigenschaft von Codes zu denken, nämlich eine Sicherung gegen Störungen, die in der Redundanz zum Ausdruck kommt. Es werden solche Codes, die Fehler nur erkennen lassen, von solchen unterschieden, die zugleich auch eine Fehlerkorrektur zulassen. Auf diesen Aspekt müßte man gesondert eingehen.

„Schuld“ an dem nicht vollständig positiven Ergebnis sind die Endung der 3.Ps.Sg, die für einen optimalen Code zu lang ist, und die Endung der 1.Ps.Sg., die für einen optimalen Code zu kurz ist.

Nun treten ja die Endungsmorpheme in der Kommunikation nicht für sich auf, sondern nur als Teile von Wortformen. Da aber innerhalb des Präsensparadigmas vor den Endungen durchaus verschieden lange Stammallomorphe (eines und desselben Lexems) auftreten können, so ist es notwendig, den Einfluß dieser unterschiedlichen Allomorphlängen zu berücksichtigen. Die isolierte Betrachtung der Endungen gilt jedoch zugleich für die Fälle, in denen die Stammallomorphe stets die gleiche Länge aufweisen. Das sind von den fünf Strukturtypen des russischen Verbs (vgl. KEMPGEN 1989, 145) vier. In einem – gar nicht so seltenen Fall – jedoch ist die Form der 1.Ps.Sg. tatsächlich um einen Buchstaben länger als die übrigen Formen, nämlich dann, wenn bei mehrsilbigen Verben auf *-буть*, *-нуть*, *-муть*, *-вуть*, *-футь* das sog. „epenthetische I“ eingeschoben wird. Mit anderen Worten: Für die Rolle der morphologischen Alternationen bedeutet dies, daß sie, sofern sie überhaupt auf diesen Bereich einen Einfluß ausüben, dann jedenfalls die Tendenz zur Zeitminimierung unterstützen. Diese Feststellung ist jedoch rein theoretischer Natur, da sie sich auf den systemischen Vergleich von Wortformen bezieht. Selbstverständlich bedeutet es in der praktischen Anwendung keine Zeitersparnis, wenn eine Form um einen Buchstaben verlängert wird, sofern dies nicht von einer Verkürzung einer häufigeren Form ausgeglichen wird.

Der vorliegende Beitrag konnte die aufgeworfene Frage nach der Zeitoptimierung des Russischen auf morphologischer Ebene nur an einem winzigen Ausschnitt aus der Flexionsmorphologie überprüfen. Die Ergebnisse ermuntern jedoch dazu, die Hypothese an weiteren Bereichen zu testen. Dazu sollten jedoch umfangreichere Angaben über die Häufigkeiten einzelner Wortformen vorliegen als dies bei STEINFELDT (1966) gegeben ist. Es ist ohne Zweifel ein großes Manko der russischen Lexikographie,

daß es bisher kein Häufigkeitswörterbuch vom Umfange von ZASORINA (1977) gibt, das die Lexemhäufigkeit nach einzelnen Wortformen aufschlüsselt. Ein solches Wörterbuch wäre für viele Fragestellungen, nicht nur für die hier behandelte, eine unverzichtbare Arbeitsgrundlage.

### *Literatur*

Altmann, G., Lehfeldt, W.

1980 Einführung in die quantitative Phonologie (*Quantitative Linguistics*, vol. 7). Bochum: Brockmeyer.

Isačenko, A.V.:

1975 Die russische Sprache der Gegenwart. Formenlehre. München. 3. Auflage.

Kempgen, S.:

1989 Grammatik der russischen Verben (*Slavistische Studienbücher, Neue Folge Bd. 3*). Wiesbaden: Otto Harrassowitz.

Padučeva, E.V.:

1961 Vozmožnosti izučenija jazyka metodami informacii. In: O.S. Achmanova, I.A. Mel'čuk, E.V. Padučeva, R.M. Frumkina, *O točnych metodach issledovanija jazyka*, Moskva, 98–149.

Siegel, S.:

1976 Nichtparametrische statistische Methoden. Mit einem Vorwort und Flußdiagramm zur Deutschen Ausgabe von W. Schüle. Frankfurt.

Šteinfeldt, E.:

1966 Häufigkeitwörterbuch der russischen Sprache. 2500 meistgebrauchte Wörter der modernen russischen Schriftsprache. Handbuch für Russischlehrer. Moskva.

Zasorina, L.N. (red.):

1974 Častotnyj slovar' russkogo jazyka. Okolo 40000 slov. Moskva.

Zipf, G.K.:

1965 Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology. NewYork—London (zuerst 1949).

