QUANTITATIVE LINGUISTICS

Vol. 26

GLOTTOMETRIKA 7

edited

by

U. Rothe

Studienverlag Dr. N. Brockmeyer

Bochum 1984

QUANTITATIVE LINGUISTICS

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Glottometrika. – Bochum: Studienverlag Brockmeyer

7. ed. by U. Rothe. – 1984.
(Quantitative linguistics; Vol. 26)
ISBN 3-88339-423-8

NE: Rothe, Ursula [Hrsg.]; GT

ISBN 3-88339-423-8

Querenburger Höhe 281, 4630 Bochum 1
Druck: Thiebes GmbH & Co. Kommanditgesellschaft Hagen

CONTENTS

The Distribution of the Distance between Like Elements in a Sequence II

Peter Zörnig, Hagen

0. Introduction

In a previous study (ZÖRNIG, 1984) we derived the distribution of the distance between like elements in a sequence of n elements out of the set {1, ..., p}. To this end we considered all sequences of length n, where the element of r-th kind occurred exactly k_r times ($r = 1, \ldots, p$; $k_1 + \ldots + k_p = n$).

Now the distribution of the distance depends on the way of counting the distances. In the quoted study we counted the distances between a l l like elements taken two at a time, i.e. in a sequence

$$\ldots\ a_i\ \ldots\ a_j\ \ldots\ a_k\ \ldots$$

($a_i = a_j = a_k = r$; $a_\nu \neq r$ for $i < \nu < k$, $\nu \neq j$) we took into consideration not only the distances c(i,j) and c(j,k) but also c(i,k) (cf. section 1). In the present study we shall derive the distribution of the distances between adjacent elements only, i.e. in the above sequence c(i,k) will not be taken into account.

1. DEFINITIONS

As in ZÖRNIG (1984) we define $\mathbb{F}$: $\mathbb{F}_{k_1, \ldots, k_p}$ as the set of all finite sequences consisting of n elements of {1, ..., p}, where the element r occurs exactly k_r times ($r = 1, \ldots, p$ and $k_1, \ldots, k_p$ are natural numbers with $k_1 + \ldots + k_p = n$). Let $F = (a_1, \ldots, a_n)$ be a sequence out of $\mathbb{F}$. For any μ, ν with $1 \leq \mu < \nu \leq n$ the quantity $c(\mu,\nu) := \nu - \mu - 1$ is called

the distance between a_μ and a_ν. Now for any $F \in |F$ the quantity $c_i^{(r)}(F)$ represents the frequency of occurrence of the distance i between two a d j a c e n t elements of the r-th kind in a sequence F, i.e. $c_i^{(r)}(F)$ is the number of all pairs of indices (μ,ν) with the properties

$$a_\mu = a_\nu = r$$
$$a_\rho \neq r \qquad \text{for all } \rho \text{ with } \mu < \rho < \nu$$

and $\quad c(\mu,\nu) = i$

$(i = 0, \ldots, n-2;\ r = 1, \ldots, p;\ 1 \leq \mu < \nu \leq n)$.

For example (1, 3, 2, 3, 2, 3, 2, 1) is a sequence out of $|F_{2,3,3}$. Here we have

$$c_1^{(2)} = c_1^{(3)} = 2; \quad c_6^{(1)} = 1.$$

The number $c_i(F)$ is defined by

$$c_i(F) := \sum_{r=1}^{p} c_i^{(r)}(F) \ . \tag{1}$$

Let us finally introduce the sums

$$C_i^{(r)} := C_i^{(r)}(k_1, \ldots, k_p) := \sum_{F \in |F} c_i^{(r)}(F) \tag{2}$$

and

$$C_i := C_i(k_1, \ldots, k_p) := \sum_{r=1}^{p} C_i^{(r)}(k_1, \ldots, k_p) \tag{3}$$

From (1), (2), (3) also follows

$$C_i = \sum_{F \in |F} c_i(F). \tag{4}$$

In order to simplify some presentations of the following sections we define for non-negative integers n,i:

$$n_{(i)} := \begin{cases} \dfrac{n!}{(n-i)!} & \text{for } i \leq n \\ 0 & \text{for } i > n \ . \end{cases}$$

2. DERIVATIONS OF FORMULAE FOR $C_i^{(r)}$ AND C_i

The quantities $C_i^{(r)}$ and C_i from (2) and (3) can be represented as follows

Theorem 1: It holds

(a) $$C_i^{(r)} = \frac{(n-1-i)!}{k_1! \ldots k_p!}\, k_r(k_r-1)(n-k_r)_{(i)} \tag{1}$$

$(r = 1, \ldots, p)$,

(b) $$C_i = \frac{(n-1-i)!}{k_1! \ldots k_p!} \sum_{r=1}^{p} k_r(k_r-1)(n-k_r)_{(i)} \ .$$

Proof: For $1 \leq \mu < \nu \leq n$, $1 \leq r \leq p$ and $F = (a_1, \ldots, a_n) \in |F$ we define the auxiliary quantity

$$\beta_{\mu,\nu}^{(r)}(F) := \begin{cases} 1 & \text{for } a_\mu = a_\nu = r \text{ and } a_\rho \neq r \text{ for } \mu < \rho < \nu \\ 0 & \text{otherwise} \ . \end{cases}$$

Now we evidently have

$$c_i^{(r)}(F) = \sum_{\mu=1}^{n-1-i} \beta_{\mu,\mu+i+1}^{(r)}(F) \ .$$

Equation (2) yields

$$C_i^{(r)} = \sum_{F \in |F} c_i^{(r)}(F) = \sum_{F \in |F} \sum_{\mu=1}^{n-1-i} \beta_{\mu,\mu+i+1}^{(r)}(F) \ .$$

Exchanging the order of summation we obtain

$$C_i^{(r)} = \sum_{\mu=1}^{n-1-i} \sum_{F \in |F} \beta_{\mu,\mu+i+1}^{(r)}(F) \ . \tag{5}$$

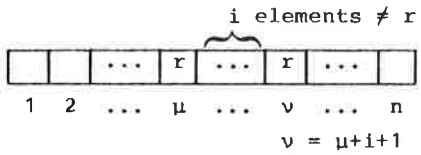
Now we have $\beta^{(r)}_{\mu,\mu+i+1}(F) = 1$ if in a sequence F there is an element of the r-th kind in positions μ and $\mu+i+1$ and no element of the r-th kind between these positions. Otherwise $\beta^{(r)}_{\mu,\mu+i+1}(F) = 0$. The sum

$$S := \sum_{F \in \mathbb{F}} \beta^{(r)}_{\mu,\mu+i+1}(F)$$

in (5) is therefore the number of all sequences $F = (a_1, \ldots, a_n) \in \mathbb{F}$ with the properties

$$a_\mu = a_{\mu+i+1} = r$$

and $\quad a_\rho \neq r \qquad$ for $\mu < \rho < \mu+i+1$.



Then we have

$$S = \binom{n-i-2}{k_r-2} \binom{n-k_r}{k_1, \ldots, k_{r-1}, k_{r+1}, \ldots, k_p}$$

because there are

$\binom{n-i-2}{k_r-2}$ ways to place the k_r-2 free elements of kind r on positions $1, \ldots, \mu-1, \mu+i+2, \ldots, n$ and

$\binom{n-k_r}{k_1, \ldots, k_{r-1}, k_{r+1}, \ldots, k_p}$ ways to place the elements of kinds $\neq r$ on the remaining positions. (S does not depend on μ). Inserting the above representation of S in (5) yields

$$C_i^{(r)} = \sum_{\mu=1}^{n-i-1} S = (n-i-1)S =$$

$$= \frac{(n-i-1)!}{k_1! \ldots k_p!} k_r (k_r-1) (n-k_r)_{(i)} ,$$

which proves (a). Part (b) follows immediately from (a) and definition (3).

For the special case when $p = 2$ we have (assume $k_1 \geq 2$ and therefore $k_2 \leq n-2$):

$$C_i^{(1)} = \frac{(n-i-1)!}{k_1! k_2!} k_1 (k_1-1) (k_2)_{(i)} =$$

$$= \frac{1}{(k_1-2)!} (n-1-i) \ldots (k_2-i-1)$$

and the representation of $C_i^{(2)}$ is analogous,

$$C_i = C_i^{(1)} + C_i^{(2)} .$$

Example: For sequences of three elements of kind 1 and four elements of kind 2 we obtain from the above formulae

$$C_i^{(1)} = (6-i)(5-i)$$

$$C_i^{(2)} = \frac{1}{2} (6-i)(5-i)(4-i) ,$$

hence

$$C_i = \frac{1}{2} (6-i)^2 (5-i) .$$

Finally we note that the numbers C_0 coincide with the numbers D_0 in ZÖRNIG (1984).

3. THE DISTRIBUTION OF THE DISTANCE

Let $F \in \mathbb{F}$ be a randomly chosen sequence from $\mathbb{F}$. Further let a_μ and a_ν be two randomly chosen adjacent identical elements of F (this means that a_μ and a_ν are of the same kind and all elements between a_μ and a_ν are different from a_μ and a_ν).

Let i be the distance between a_μ and a_ν. In this section we derive formulae for probability p_i, mean E(I) and variance V(I) of the random-variable I. The results are summarized by

Theorem 2: It holds for $p < n$ and $0 \le i \le n-2$:

(a) $$p_i = \frac{1}{(n-p)\,n_{(i+1)}} \sum_{r=1}^{p} k_r(k_r-1)(n-k_r)_{(i)} \qquad 2)$$

(b) $$E(I) = \frac{1}{n-p} \sum_{r=1}^{p} \frac{k_r-1}{k_r+1}(n-k_r)$$

(c) $$V(I) = n^2 - \frac{n+1}{n-p} \sum_{r=1}^{p} (3n + (n-1)k_r) \frac{k_r(k_r-1)}{(k_r+2)(k_r+1)} - \left(\frac{1}{n-p} \sum_{r=1}^{p} \frac{k_r-1}{k_r+1}(n-k_r) \right)^2 .$$

Proof: (a) Evidently p_i is given by

$$p_i = \frac{C_i}{\sum_{j=0}^{n-2} C_j} . \qquad (6)$$

Now the denominator of (6) can be simplified as follows: with respect to (4) we obtain

$$\sum_{j=0}^{n-2} C_j = \sum_{j=0}^{n-2} \sum_{F\in\mathbb{F}} c_j(F) = \sum_{F\in\mathbb{F}} \sum_{j=0}^{n-2} c_j(F) , \qquad (7)$$

where for every sequence F the sum $\sum_{j=0}^{n-2} c_j(F)$ is the number of all distances between adjacent like elements. So we have

$$\sum_{j=0}^{n-2} c_j(F) = (k_1-1) + \ldots + (k_p-1) = n-p \qquad (8)$$

for every sequence F.

Inserting (8) in (7) leads to

$$\sum_{j=0}^{n-2} C_j = \sum_{F\in\mathbb{F}} (n-p) = (n-p) \binom{n}{k_1 \ldots k_p} . \qquad 3) \qquad (9)$$

Inserting in (6) by equation (b) of Theorem 1 and equation (9) leads to

$$p_i = \frac{(n-i-1)!}{(n-p)\,n!} \sum_{r=1}^{p} k_r(k_r-1)(n-k_r)_{(i)}$$

which proves part (a).

(b) In order to prove equations (b) and (c) of Theorem 2 we use the following combinatorial equations:

$$\sum_{j=k}^{n} \binom{j}{k} = \binom{n+1}{k+1} \qquad 4) \qquad (10)$$

$$\sum_{j=k}^{n} j\binom{j}{k} = (k+1)\binom{n+2}{k+2} - \binom{n+1}{k+1} \qquad (11)$$

$$\sum_{j=k}^{n} j^2\binom{j}{k} = (k+2)(k+1)\binom{n+3}{k+3} - 3(k+1)\binom{n+2}{k+2} + \binom{n+1}{k+1} . \qquad (12)$$

From equation (a) follows

$$E(I) = \sum_{i=0}^{n-2} i p_i =$$

$$= \frac{1}{n-p} \sum_{r=1}^{p} \frac{1}{n!} k_r(k_r-1) \sum_{i=0}^{n-k_r} i(n-i-1)!(n-k_r)_{(i)} . \qquad (13)$$

For the last sum we obtain by the substitution j: = n-i-1 of indices:

$$S: = \sum_{i=0}^{n-k_r} i(n-1-i)!(n-k_r)_{(i)} = \sum_{j=k_r-1}^{n-1} (n-j-1)j!(n-k_r)_{(n-1-j)} =$$

$$= (k_r-1)!(n-k_r)!\left((n-1)\sum_{j=k_r-1}^{n-1}\binom{j}{k_r-1} - \sum_{j=k_r-1}^{n-1} j\binom{j}{k_r-1}\right).$$

With equation (10) and (11) this leads to

$$S = (k_r-1)!(n-k_r)!\left((n-1)\binom{n}{k_r} - k_r\binom{n+1}{k_r+1} + \binom{n}{k_r}\right) =$$

$$= n!\,\frac{n-k_r}{k_r(k_r+1)} \quad . \tag{14}$$

By inserting (14) into (13) we obtain

$$E(I) = \frac{1}{n-p}\sum_{r=1}^{p}\frac{k_r-1}{k_r+1}\,(n-k_r)\ ,$$

which proves part (b).

(c) Quite analogically to (13) and (14) we derive

$$E(I^2) = \sum_{i=0}^{n-2} i^2 p_i =$$

$$= \frac{1}{n-p}\sum_{r=1}^{p}\frac{1}{n!}\,k_r(k_r-1)\sum_{i=0}^{n-k_r} i^2(n-1-i)!\,(n-k_r)_{(i)} \tag{15}$$

with $S' = \sum_{i=0}^{n-k_r} i^2(n-1-i)!\,(n-k_r)_{(i)} =$

$$= \sum_{j=k_r-1}^{n-1}(n-1-j)^2 j!\binom{n-k_r}{j-(k_r-1)} =$$

$$= (k_r-1)!(n-k_r)!\left((n-1)^2\sum_{j=k_r-1}^{n-1}\binom{j}{k_r-1} -\right.$$

$$\left. - 2(n-1)\sum_{j=k_r-1}^{n-1} j\binom{j}{k_r-1} + \sum_{j=k_r-1}^{n-1} j^2\binom{j}{k_r-1}\right).$$

Using equations (10), (11) and (12) we obtain

$$S' = (k_r-1)!(n-k_r)!\left((n-1)^2\binom{n}{k_r} - 2(n-1)(k_r\binom{n+1}{k_r+1} -\right.$$

$$\left. - \binom{n}{k_r}) + (k_r+1)k_r\binom{n+2}{k_r+2} - 3k_r\binom{n+1}{k_r+1} + \binom{n}{k_r}\right) =$$

$$= n!\left(\frac{n^2}{k_r} - (n-1)\,\frac{3n+(n-1)k_r}{(k_r+2)(k_r+1)}\right). \tag{16}$$

Inserting (16) into (15) yields

$$E(I^2) = \frac{1}{n-p}\sum_{r=1}^{p}\left(n^2(k_r-1) - (n+1)k_r(k_r-1)\frac{3n+(n-1)k_r}{(k_r+2)(k_r+1)}\right) =$$

$$= n^2 - \frac{n+1}{n-p}\sum_{r=1}^{p}(3n+(n-1)k_r)\,\frac{k_r(k_r-1)}{(k_r+2)(k_r+1)} \quad .$$

Since $V(I) = E(I^2) - E^2(I)$, this finally proves part (c).

Example: For the special case, when

$k_1 = 3$, $k_2 = 14$, $k_3 = 21$, $k_4 = 10$, $k_5 = 2$; $p = 5$, $n = 50$
(cf. section 4) we have

$$E(I) = \frac{1}{45}\left(\frac{2}{4}\cdot 47 + \frac{13}{15}\cdot 36 + \frac{20}{22}\cdot 29 + \frac{9}{11}\cdot 40 +\right.$$

$$\left. + \frac{1}{3}\cdot 48\right) = 2.884 \quad .$$

Analogically we compute

$$V(I) = 21.717$$

with equation (c) in Theorem 2.

4. AN EXAMPLE OF APPLICATION

Let us use the example in ZÖRNIG (1984) for an application of the distribution derived in section 3. We considered the verses 81 to 130 in Vergils Aeneis from which we obtained the sequence

32122 10232 11221 31211 32132 04122 34212 13123

03212 22322 (17)

where the i-the element of the sequence (17) is the number of dactyls in verse (80+i) (i = 1, ..., 50). The sequence in (17) is an element of $\mathbb{F}_{k_1, \ldots, k_p}$, where

$k_1 = 3$, $k_2 = 14$, $k_3 = 21$, $k_4 = 10$, $k_5 = 2$; $p = 5$, $n = 50$.

In this work we want to test whether the distances between adjacent like elements in (17) are randomly distributed. Let $\hat{c}_i$ be the number of observed i-distances between adjacent like elements in (17) and $\bar{c}_i$ the theoretically expected number of such i-distances.

We obtain the theoretical $\bar{c}_i$ by a random choice of a sequence F out of $\mathbb{F}_{k_1, \ldots, k_p}$. Here we expect the number

$$\bar{c}_i = E(c_i) = \frac{1}{\binom{n}{k_1, \ldots, k_p}} \sum_{F \in \mathbb{F}} c_i(F) \qquad (18)$$

of i-distances. With respect to Theorem 1(a) equation (18) yields

$$\bar{c}_i = E(c_i) = \frac{C_i}{\binom{n}{k_1, \ldots, k_p}} =$$

$$= \frac{1}{n_{(i+1)}} \sum_{r=1}^{p} k_r (k_r - 1) (n - k_r)_{(i)} \, . \qquad (19)$$

Theorem 2(a) also yields

$$\bar{c}_i = (n-p) p_i \, , \qquad 5)$$

this means that $\bar{c}_i$ is also the expectation of i-distances by (n-p) trials of the double experiment mentioned in section 3. Therefore we take this double experiment as model to compute our theoretical $\bar{c}_i$ (cf. section 4 in ZÖRNIG (1984)).

For our example we obtain from (19):

$$\bar{c}_i = \frac{1}{50_{(i+1)}} (47_{(i)} \cdot 6 + 36_{(i)} \cdot 182 + 29_{(i)} \cdot 420 + 40_{(i)} \cdot 90 + 48_{(i)} \cdot 2) \, .$$

The observed and computed values are presented in Table 1 (cf. Fig. 1).

A chi-square test yields

$$X_2^2 = 7.45 \, , \qquad P = 0.024$$

The classes 6-13 and 14-48 were pooled; we have 8 classes and 5 parameters so that there are 2 degrees of freedom. The test shows that the empirical distribution of distances significantly diverges from the computed one and signalizes a specified tendency which can be observed especially at the distance i = 0.

5. CONCLUSIONS

The theory presented here is an alternative to the theory of runs. It is especially suitable for solutions of different problems in linguistics where the possibility of building runs of like elements is strongly restricted but tendencies to place like elements at special distances are quite usual. The method enables us to find hidden repetitive regularities at all levels of a text (phonological, morphological, semantic etc.).

REMARKS

1 For fixed numbers n, p, $k_1, \ldots, k_p$ the function $C_i^{(r)}$ is a polynomial in i of degree $k_r - 1$. Therefore C_i is a polynomial of degree k-1, where k is defined by $k := \max \{k_r \mid r = 1, \ldots, p\}$.

2 From (6) and (9) follows that p_i is also a polynomial in i of degree k-1 for fixed numbers n, p, $k_1, \ldots, k_p$ (cf. remark 1).

3 Analogically we derive the equation
$\sum_{j=0}^{n-2} C_j^{(r)} = (k_r-1)\binom{n}{k_1,\ldots,k_p}$, which is equivalent to (10).
Obviously (9) also follows from the equation above.

4 Equation (10) is a well known combinatorial equation.
Representations of $\sum_{j=k}^{n} j^q\binom{j}{k}$ (q = 1, 2, 3, ...) can be derived recursively.

5 Since $n-c_0(F)$ is the number of runs in the sequence F, the expected number of runs by a random choice of F out of $|F_{k_1,\ldots,k_p}$ is $n-\bar{c}_0$.

REFERENCE

ZÖRNIG, P., The distribution of the distance between like ele-
1984 ments in a sequence I. Boy, J., Köhler, R. (Ed.),
Glottometrika 6. Bochum, Brockmeyer, 1984, S. 1

Table 1: Observed and computed numbers of i-distances

i	$\hat{c}_i$	$\bar{c}_i$
0	8	14
1	10	9.269
2	9	6.193
3	5	4.185
4	4	2.868
5	4	1.996
6	2	1.414
7	1	1.019
8		0.748
9		0.560
10		0.426
11		0.331
12		0.261
13		0.210
14	1	0.173
15		0.144
16		0.123
17		0.106
18	1	0.093
19		0.083
20		0.075
21		0.068
22		0.062
23		0.057
24		0.053
25		0.049
26		0.045
27		0.042
28		0.039
29		0.036
30		0.033
31		0.030
32		0.028
33		0.025
34		0.023
35		0.021
36		0.019
37		0.017
38		0.015
39		0.013
40		0.011
41		0.009
42		0.008
43		0.006
44		0.005
45		0.004
46		0.003
47		0.002
48		0.001

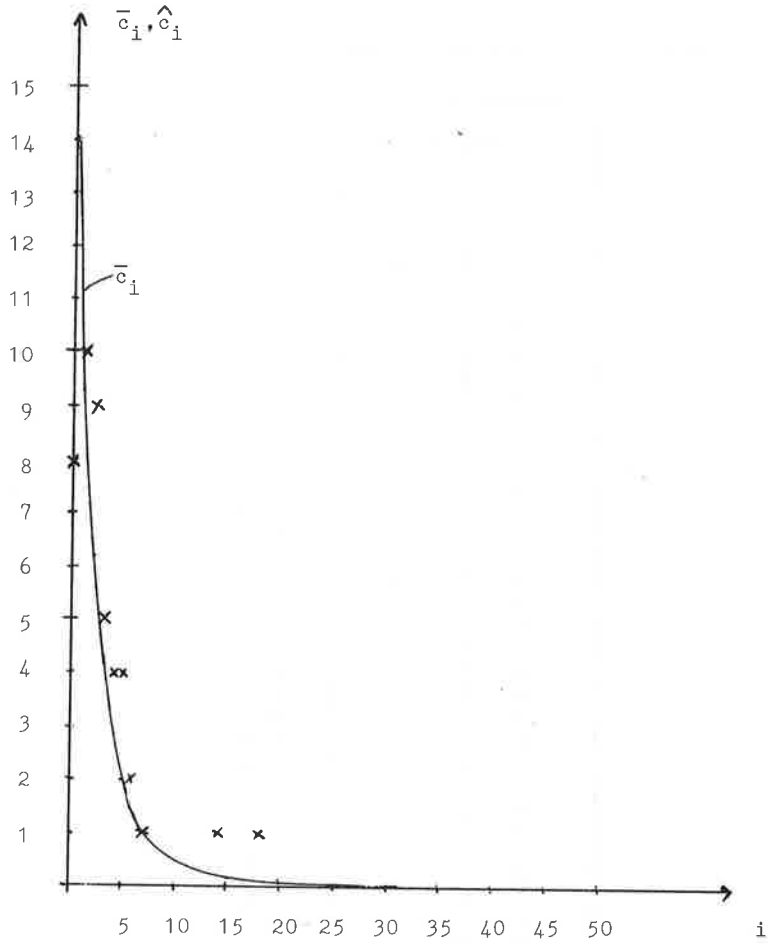


Fig.1. Observed and computed numbers of i distances

(The observed numbers $\hat{c}_i$ are marked with ×, the computed values $\bar{c}_i$ are given by the curve)

Die Erkennbarkeit von abgekürzten Wörtern - Experimentelle Untersuchungen und mathematische Modelle

D. Rumpel, D. Goldenberg, W. Boucsein

1. Einleitung

Bei der Benutzung von alphanumerischen Bezeichnungen in Rechner-Eingaben und -Speicherinhalten sollen Wörter möglichst stark abgekürzt werden, um sich bündig über Tastaturen eingeben zu lassen und wenig Speicherplatz zu verbrauchen. Andererseits sollen die Wörter aber nicht derart stark abgekürzt werden, daß der Sinnbezug zum vollen Wort verloren geht und die Abkürzungen als selbständiger Code erlernt werden müssen. Läßt sich hierbei ein brauchbarer Kompromiß bzw. ein Optimum finden? Wie sehen die dabei zu beachtenden Abhängigkeiten in der Erkennung eines Wortes aus seiner Abkürzung aus? Diese Fragestellung führte zu den im folgenden beschriebenen Ergebnissen, die an der Universität Duisburg in interfakultativer Kooperation der Autoren aus Elektrotechnik und Psychologie erarbeitet wurden.

Mit einer Probandengruppe wurden über ca. 2 Jahre Untersuchungen durchgeführt, deren statistische Auswertung das Datenmaterial für die quantitative Beantwortung folgender Fragen lieferte:

Prof.Dr.-Ing. D.Rumpel ist Ordinarius des Fachgebietes "Elektrische Anlagen und Netze" an der Universität Duisburg Gesamthochschule.
Dipl.-Ing. D. Goldenberg ist Assistentin im Fachgebiet.
Prof.Dr.rer.nat. W. Boucsein war zur Zeit der Arbeiten an der Universität Duisburg Gesamthochschule im Fach Psychologie tätig. Er ist jetzt Ordinarius des Fachgebietes für Physiologische Psychologie an der Universität Wuppertal Gesamthochschule.

- Wie ist der Verlauf der Erkennbarkeit (Wahrscheinlichkeit des Erratens abgekürzt angebotener Zielwörter) in Abhängigkeit von der Zielwortlänge und der gewählten Abkürzungslänge?
- Wie verändert sich die Erkennbarkeit, wenn den Probanden über die Zielwörter oder die Abkürzungen zusätzliche Instruktionen gegeben werden?
- Wie verändert sich die Erkennbarkeit, wenn die Abkürzungen in einem Kontext angeboten werden, aus dem die Probanden zusätzliche Information über die Zielwörter gewinnen können?

Bei dem Versuch, die gefundenen Verläufe mathematisch zu approximieren und zu beschreiben, entstanden Modelle, die plausible Abhängigkeiten aufweisen und informationstheoretisch interpretierbar sind.

2. Methodik der experimentellen Untersuchungen

2.1 Probanden

An den Untersuchungen nahmen insgesamt 15 Pbn teil, von denen 11 männlich und 4 weiblich waren. Es handelte sich um 6 Studenten, 4 Personen mit abgeschlossenem Hochschulstudium und 5 Angehörige verschiedener Berufe. Sie waren zwischen 20 und 50 Jahre alt. Die Pbn wurden von der Versuchsleiterin aus dem elektrotechnischen Institut bzw. aus ihrem Bekanntenkreis rekrutiert. Für eine Einzeluntersuchung standen jeweils 6 bis 8 Mitglieder der Probandengruppe zur Verfügung.

2.2 Versuchsplan

Die in der Einleitung genannten Fragen führten zur Festlegung von 2 Gruppen mit jeweils 3 Versuchsreihen, die ihrerseits 3 bis 12 Einzeluntersuchungen umfaßten. In jeder dieser Untersuchungen wurde ein Arbeitsbogen benutzt, der 50 bis 200 auf einheitliche Länge abgekürzte Zielwörter variabler Länge enthielt. Die Pbn wurden aufgefordert, das Zielwort zu erraten und in den Bogen einzutragen. Die Untersuchungsgruppen und -reihen unterschieden sich in den experimentellen Bedingungen, die vor der einzelnen Untersuchung für die Pbn festgelegt wurden:

A Gruppe "Einzelwörter"

Der Arbeitsbogen enthielt die Abkürzungen unverbundener Grundformen von Zielwörtern. Die 3 Versuchsreihen dieser Gruppe variieren die Vorkenntnisse der Pbn bezüglich der Zielwörter und Abkürzungen:

A1 Versuchsreihe "Allgemeine Wortkenntnis"

Den Pbn wurde mitgeteilt, daß es sich um Abkürzungen unverbundener Einzelwörter handelt, bei kurzen Wörtern evtl. auch um das vollausgeschriebene Wort. Im übrigen erhielten sie keine weiteren Instruktionen und hatten die Zielwörter aufgrund ihrer allgemeinen Kenntnis der deutschen Sprache zu erraten.

A2 Versuchsreihe "Spezielle Wortkenntnis"

Wie A1, jedoch wurden den Pbn unmittelbar vor dem Versuch die Zielwörter in gegenüber dem Arbeitsbogen geänderter Reihenfolge diktiert und sie erhielten 10 min Zeit, sie zu memorieren.

A3 Versuchsreihe "Spezielle Wort- und Abkürzungskenntnis"

Wie A2, jedoch wurden mit den Zielwörtern auch die zugehörigen gewählten Abkürzungen diktiert und 10 min memoriert.

B Gruppe "Kontext"

Die Arbeitsbögen enthielten einen laufenden Text aus auf einheitliche Länge abgekürzten Wörtern einschließlich Interpunktion. Die Pbn wurden instruiert, daß die Folge von Abkürzungen einen laufenden Text wiedergab. Die 3 Versuchsreihen dieser Gruppe variieren den Schwierigkeitsgrad des Originaltextes:

B1 Versuchsreihe "Bismarck"

Den Arbeitsbögen lagen Abschnitte aus Bismarck: "Gedanken und Erinnerungen" zugrunde.

B2 Versuchsreihe "Spiegel"

Den Arbeitsbögen lagen Textabschnitte aus Berichten des Nachrichtenmagazins "Der Spiegel" zugrunde.

B3 Versuchsreihe "Literarisch"

Den Arbeitsbögen lagen Abschnitte literarischer Texte (E. Hemingway in deutscher Übersetzung, St. Zweig, W. Bergengruen) zugrunde.

Tabelle 1 zeigt die zeitliche Abfolge der Einzelversuche bzw. der Sitzungen, in denen sie durchgeführt wurden. Zu Beginn des Versuchs lag die hier geschilderte Strategie in den Grundzügen fest, jedoch bestand noch keine Vorstellung über die zu erwartenden quantitativen Ergebnisse oder gar ein mathematisches Modell, das Vorhersagen erlaubte. Aus Tabelle 1 läßt sich erkennen, wie nach Vorliegen der ersten Ergebnisse die zu untersuchenden Parameterfelder erweitert wurden und z.T. Versuche mit gleichen Parametern, jedoch anderen Arbeitsbögen wiederholt wurden, um eine größere Datenbasis zu bekommen und so die Streuung der Ergebnispunkte zu vermindern. Letzteres gilt besonders für die Versuchsreihe "Allgemeine Wortkenntnis", die als Vergleichsbasis für die übrigen Versuchsreihen diente.

2.3 Erstellung der Arbeitsbögen

2.3.1 Versuchsgruppe A (Einzelwörter)

Bei der Gruppe A, deren Arbeitsbögen unverbundene Wörter enthielten, wurden die Zielwörter durch zufälliges Aufschlagen des Duden bzw. des deutschen Teils eines Langenscheidt-Wörterbuchs Deutsch-Englisch und Aufsuchen des nächsten Wortes der gewünschten Länge ermittelt. Es wurde darauf geachtet, daß kein Zielwort in mehreren Bögen vorkam.
Eine Ausnahme von dieser Regel mußte bei den selten vorkommenden 2- und 3-buchstabigen Wörtern gemacht werden; diese Wörter wurden gesammelt und nur darauf geachtet, daß in zeitlich aufeinanderfolgend behandelten Arbeitsbögen keine Wiederholungen vorkamen.

Um im Abkürzungsverfahren Univozität zu gewährleisten, wurden alle Abkürzungen von einer Person, der Versuchsleiterin, intuitiv-optimal vorgenommen; d.h. der erste Buchstabe der Abkürzung war mit dem des Zielwortes identisch, die übrigen wurden nach subjektiven Auffälligkeiten - wie Silbenanfänge oder Buchstabensignifikanz - aus dem Zielwort hinzugewählt, bis die für den betreffenden Bogen vorgesehene Abkürzungslänge erreicht

Tab. 1. Zeitplan der Untersuchungen

N = Anzahl der Probanden
n = Abkürzungslänge des Arbeitsbogens

Sitzung	VERSUCHSGRUPPE A						VERSUCHSGRUPPE B					
	A 1		A 2		A 3		B 1		B 2		B 3	
Datum	N	n =	N	n =	N	n =	N	n =	N	n =	N	n =
Juni 1980									7	5,6	7	2,3,4
Juli 1980									7	2,3,4		
3.12.80	7	2,3,4,5,6										
21. 5.81			8	2,3,4,5,6								
25. 5.81					6	2,3,4,5,6						
8. 7.81	6	7										
21. 7.81	7	8										
23. 7.81	6	8										
22.10.81	8	7,8,9,10										
3.11.81	8	12	8	7,8,9,10	7	5						
15. 6.83							7	2,3,4,5,6				
29. 6.83							7	4,6				
29. 8.83							10	7				

war. Zielwörter mit einer Länge gleich oder kleiner als die vorgesehene Abkürzungslänge wurden nicht abgekürzt und erschienen voll im Bogen.

Die Arbeitsbögen enthielten 50 bis 100 Abkürzungen von Zielwörtern zwischen 2 und 18 Buchstaben Länge so, daß jede Zielwortlänge mit 3 bis 6 Abkürzungen belegt war. Die Belegung mit nur 3 Abkürzungen pro Zielwortlänge führte zu starker Streuung in den Ergebnispunkten, was die oben erwähnten Versuchswiederholungen erforderlich machte. Als aus den vorliegenden Ergebnissen klar wurde, daß für Zielwortlängen, die wesentlich (2 bis 3 Buchstaben) kürzer sind als die vorgesehene Abkürzungslänge, eine praktisch sichere Erkennung vorausgesetzt werden darf, wurden diese Zielwortlängenbereiche zugunsten einer dichteren Belegung der größeren Zielwortlängen in den später erstellten Arbeitsbögen ausgelassen.

Die zu den einzelnen Zielwortlängen gehörigen Abkürzungen wurden möglichst unsystematisch auf dem Bogen verteilt. Abb. 1 zeigt ausschnittsweise einen Arbeitsbogen im Grundzustand und im vom Pbn ausgefüllten Zustand. Die in der ausgefüllten Version eingetragenen Bruchzahlen (z.B. 1/6) sind Auswertungsvermerke.

2.3.2 Versuchsgruppe B (Kontext)

Beim Erstellen der Arbeitsbögen für die Gruppe B (Kontext) wurde nach gleicher Methode ein laufender Text abgekürzt und in den Arbeitsbogen zeilenweise mit eingefügter Interpunktion eingetragen. Dabei waren Zielwortfolge und -auswahl durch den Text starr vorgegeben, wobei Wortwiederholungen mit gleicher Abkürzung vorkamen und vor allem Wörter großer Länge wesentlich seltener auftraten als solche mittlerer und kurzer Länge. Um eine einigermaßen ausreichende Belegung bis zu Wortlängen von 18 Buchstaben zu erreichen, mußten in den Versuchsreihen "Bismarck" und "Spiegel" Bögen von 200 bis 300 Zielwörtern erstellt werden. In der Gruppe "Literarisch" war keine ausreichende Belegung großer Wortlängen mehr zu erzielen.

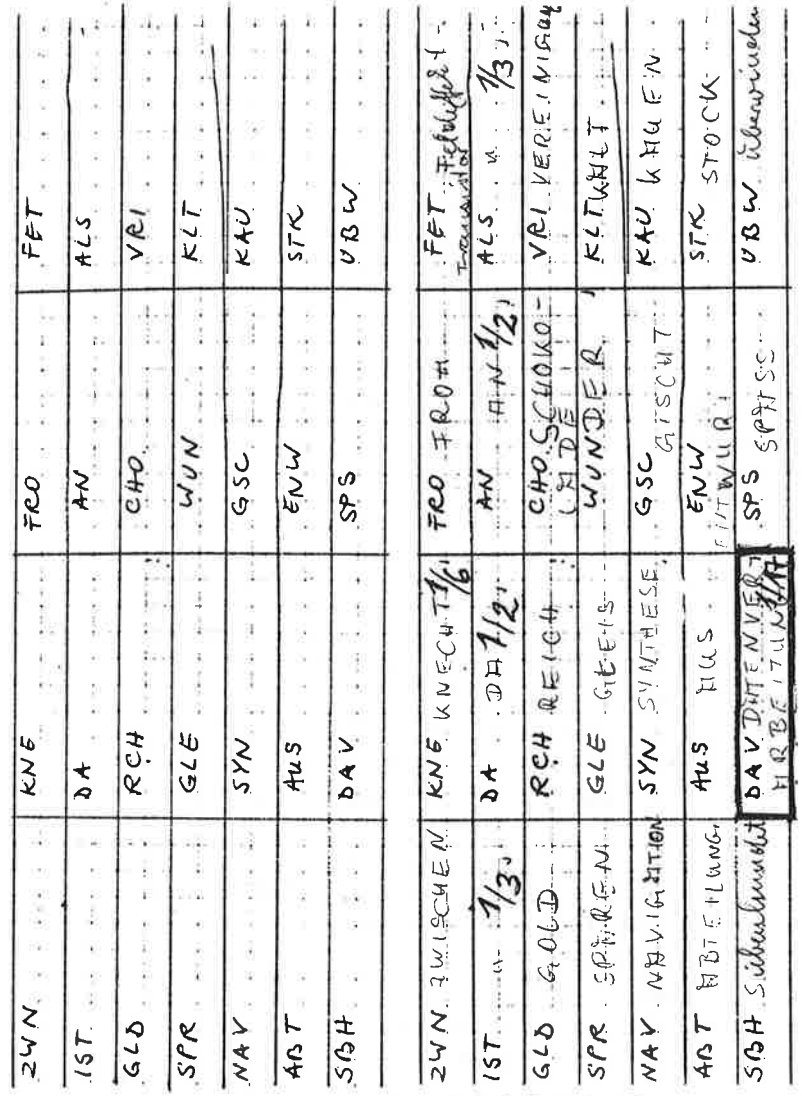


Abb.1 Arbeitsbogen, oben in unausgefülltem, unten in ausgefülltem Zustand (Ausschnitt)

2.4 Durchführung der Untersuchungen

Wie in Tab.1 aufgeführt, wurden jeweils 1 bis 6 aufeinanderfolgende Versuche auf einer Sitzung durchgeführt, an der jeweils 6 bis 8 Personen aus der Probandengruppe teilnahmen. Versuchsleiterin war die Koautorin Goldenberg.

Die Pbn wurden entsprechend der Versuchsreihe, an der sie teilnahmen (s. 2.2),instruiert bzw. vorbereitet und erhielten anschließend einen vorbereiteten Arbeitsbogen zum Ausfüllen vorgelegt. Die Pbn arbeiteten ohne Zeitdruck, da für jeden Versuch 25 min angesetzt waren. Diese Zeit wurde oft nicht ausgenutzt, wenn die Pbn einheitlich der Meinung waren "es fiele ihnen nichts mehr ein". Insofern können leistungsmotivationale und emotionale Einflüsse als verhältnismäßig gering angesehen werden. Bezüglich der experimentellen Bedingungen und der Versuchsreihen handelt es sich um einen Plan mit Meßwiederholungen, da immer wieder auf den gleichen Pool von 15 Pbn zurückgegriffen wurde.

2.5 Auswertung

Im Versuch wird der Pb aufgefordert, zu jeder Abkürzung ein mögliches Zielwort zu erraten und einzutragen. Dabei können 3 Fälle auftreten:

- das eingetragene Wort entspricht dem Zielwort (Treffer)
- das eingetragene Wort entspricht nicht dem Zielwort (Fehler)
- es wird kein zur Abkürzung passendes Wort gefunden (Schweigen)

Im vorliegenden Aufsatz werden nur die Treffer behandelt und nicht zwischen "Fehler" und "Schweigen" differenziert. Als Treffer wird in der Versuchsgruppe A das richtig und vollständig geratene Zielwort gezählt, in der Gruppe B (Kontext) wird zusätzlich verlangt, daß der Pb das Zielwort auch in der richtigen Flexion einträgt. Umlaute wurden in den Abkürzungen und Zielwörtern als 2 Buchstaben geschrieben und gewertet.

Die "Erkennbarkeit" E als Maß für den Erfolg der Rekonstruktion des Zielwortes aus seiner Abkürzung ist definiert als Summe über alle Pbn der richtig geratenen Zielwörter, dividiert durch die Summe über alle Pbn der angebotenen Abkürzungen, d. h. als Trefferrate oder Trefferwahrscheinlichkeit. Aus den Testdaten wurde die Erkennbarkeit E für jede Abkürzungslänge "n" in Abhängigkeit von der Zielwortlänge "l" ausgewertet und die entstehende Punkteschar in ein Raster E_n über l eingetragen. Die Punkteschar wurde durch eine Kurve approximiert. (Abb. 2 zeigt ein Beispiel für n = 5, l = 2 bis 18.)

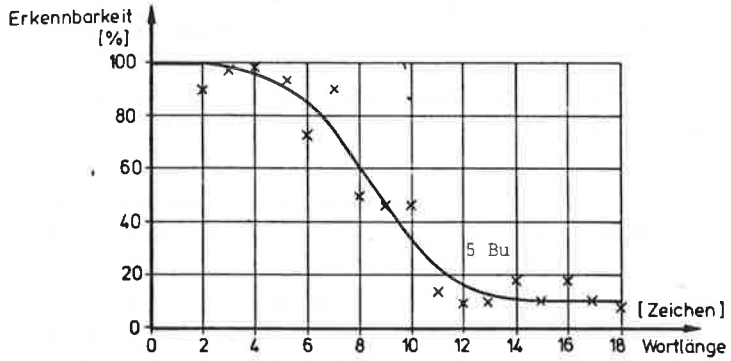


Abb. 2 Approximation der Ergebnisse durch eine Kurve (Beispiel Versuchsreihe A1, n = 5).

Auf diese Weise entstand für jeden Einzelversuch (bzw. bei Versuchswiederholungen: aus allen Versuchen mit gleichen Parametern) eine Kuve E_n über l. In der Gruppe A (Einzelwörter) konnten bei der Approximation alle Punkte der Schar mit gleichem Gewicht angesetzt werden. Bei der Gruppe B (Kontext) wurden die Punkte mit der zugrundeliegenden Abkürzungs- bzw. Zielwortzahl gewichtet, da die Anzahl der im Text verwendeten Wörter gleicher Länge stark schwankte und gegen große Wortlängen hin ausdünnte.

3. Experimentelle Ergebnisse

3.1 Ergebnisdiagramme

Die nach 2.5 ermittelten Kurven $E_n(\ell)$ wurden ohne weiteren Ausgleich in ein für jede Versuchsreihe gemeinsames Diagramm übertragen. Auf diese Weise entstanden die Kurvenscharen $E(n,\ell)$, die für die Versuchsgruppe A (Einzelwörter) in der Diagrammreihe auf der linken Seite von Abb.3, für die Versuchsgruppe B (Kontext) auf der linken Seite der Abb.4 wiedergegeben sind.

Aus diesen Diagrammen können bereits die folgenden Eigenheiten abgelesen werden:

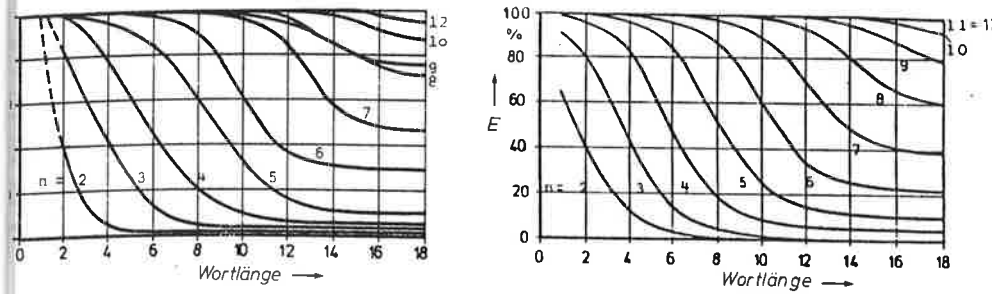
Versuchsgruppe A (Einzelwörter)

- Die Erkennbarkeit geht nicht gegen Null, auch wenn die Zielwortlänge die Abkürzungslänge wesentlich übersteigt. Sie konvergiert vielmehr gegen einen konstanten, von der Wortlänge unabhängigen Wert.
- Wenn die Zielwortlänge gleich oder kleiner als die Abkürzungslänge ist - d.h. das Zielwort voll im Arbeitsbogen steht - ist die Erkennbarkeit noch nicht notwendig 100 %. Dies zeigt sich insbesondere bei den Kurven kleiner Abkürzungslänge n.
- Die in den Versuchsreihen A2 und A3 gegebenen zusätzlichen Informationen führen im Vergleich mit A1 erwartungsgemäß zu einer wesentlichen Erhöhung der Erkennbarkeit, und zwar sowohl im gekrümmten als im geraden Teil der Kurvenschar.

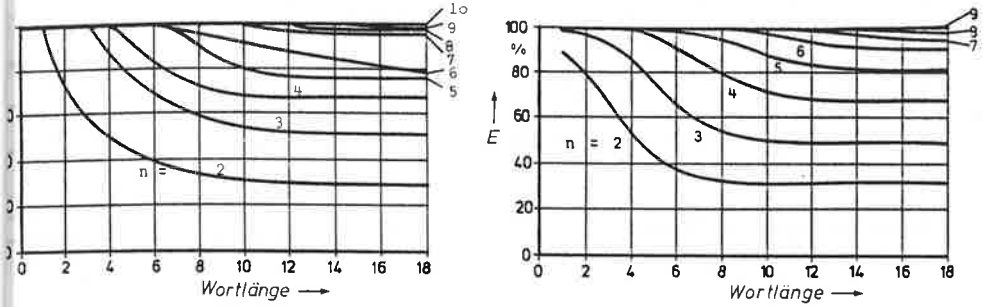
Versuchsgruppe B (Kontext)

- Gegenüber der Versuchsreihe A1 (Allgemeine Wortkenntnis) ist die Erkennbarkeit der Wörter merklich erhöht, jedoch auf grundsätzlich andere Weise als bei den Versuchsreihen A2 und A3. Der Abfall der Erkennbarkeit zu größeren Zielwortlängen hin verläuft wesentlich flacher. Die Konvergenz auf einen konstanten, wortlängenunabhängigen Wert ist angedeutet, wird jedoch im Rahmen der untersuchten Zielwortlängen meist nicht mehr erreicht.

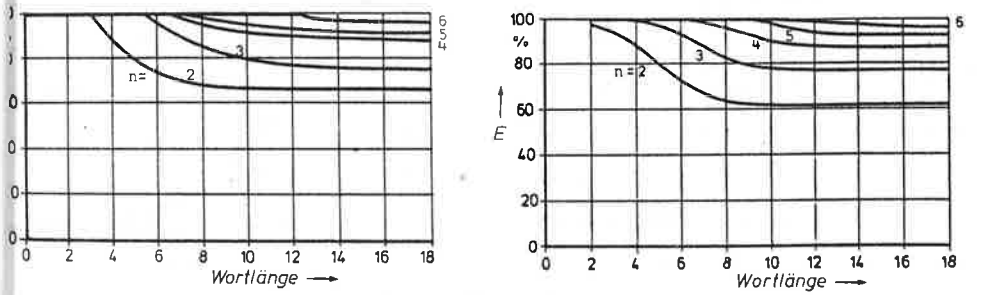
a) Experimentelle Werte b) Gerechnete Werte



A1 Allgemeine Wortkenntnis



A2 Spezielle Wortkenntnis

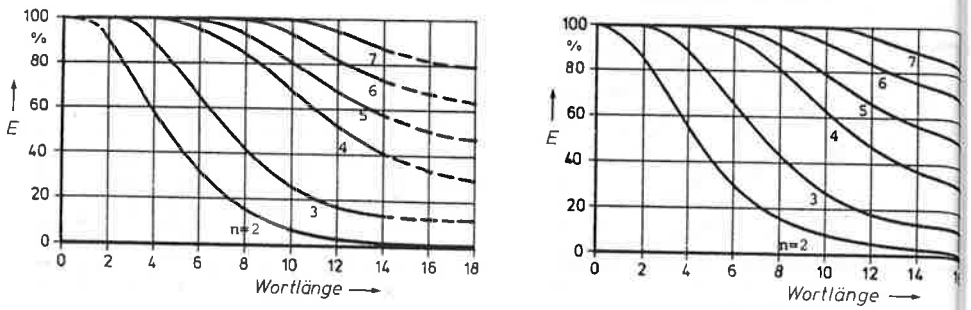


A3 Spezielle Wort- u.Abkürzungskenntnis

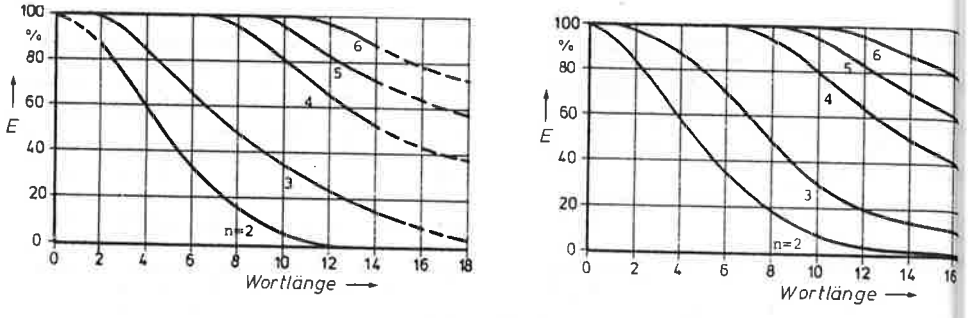
Abb.3 Versuchsgruppe A (Einzelwörter). Gegenüberstellung der experimentell gewonnenen Ergebniskurven (a) und der mit dem Modell synthetisierten Kurven (b).

a) Experimentelle Werte

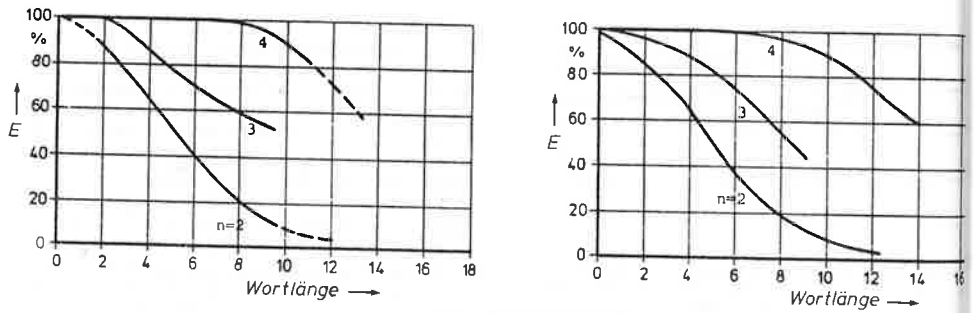
b) Gerechnete Werte



B1 Bismarck



B2 Spiegel



B3 Literarisch

Abb.4 Versuchsgruppe B (Kontext). Gegenüberstellung der experimentell gewonnenen Ergebniskurven (a) und der mit dem Modell synthetisierten Kurven (b).

- Die Kurvenscharen verhalten sich mit zunehmender Abkürzungslänge n diskontinuierlich. Zwischen n = 2 und n = 4 - besonders zwischen n = 3 und 4 - verbessert sich die Erkennbarkeit sprunghaft. Anschließend zeigt sich für n ≥ 4 ein kleineres, aber etwa konstantes Inkrement. Diese Erscheinung kann mit der subjektiven Feststellung der Versuchspersonen korreliert werden, daß ab n = 4 der "Sinn des Kontextes" klar wird.
- Die subjektiv beurteilte Textschwierigkeit fällt in den Versuchsreihen von B1 nach B3 erheblich ab. Diese Eigenschaft hat wenig Einfluß auf die Lage der Kurven für n = 2 und 3, aber die Kurven für n ≥ 4 reagieren mit einer von B1 nach B3 fortschreitenden Verschiebung nach oben, bei der der Abstand zwischen den Kurven n ≥ 4 etwa erhalten bleibt.

Das Verhalten der Kurven unter Kontexteinfluß gab Anlaß zu der folgenden Arbeitshypothese, die in der folgenden Modellbildung noch eine Rolle spielt: In den Versuchsreihen A2 und A3 wurde den Pbn zusätzliche Information geboten, die die Worterkennbarkeit steigerte. Analog hierzu wird in den Versuchsreihen B1 bis B3 durch das Kontextverständnis der Wortinterpretation zusätzliche Information zugeführt.

Hypothese: Das Kontextverständnis nimmt zunächst mit steigender Abkürzungslänge n zu, erreicht dann aber bei n = 4 einen Sättigungswert ("Der Sinn des Kontextes wird klar"). Der weitere Anstieg der Worterkennbarkeit über n = 4 resultiert nicht mehr aus steigendem Kontextverständnis.

3.2 Streuung der Ergebnisse

Die Ergebnispunkte, aus denen die Kurven entwickelt wurden, wiesen eine erhebliche Streuung auf (vgl. Abb.2). Als Ursache kommen hauptsächlich 3 Einflüsse infrage:

a) Die statistische Streuung infolge der relativ kleinen (6-8 Pbn) Probandengruppe pro Untersuchung und der geringen pro Ergebnispunkt untersuchten Wortzahl (3-6).
b) Schwankungen im mittleren Schwierigkeitsgrad der Zielwörter, die je einem Ergebnispunkt zugrunde lagen.

c) Schwankungen in der Zusammensetzung der Probandengruppe von Versuch zu Versuch, die die mittlere Qualifikation der Gruppe verschieben, sowie Schwankungen in der mittleren Tagesform der Probandengruppe.

Von diesen Einflüssen ist nur a) berechenbar, b) führt zu vergrößerter Streuung der Ergebnispunkte eines Versuchs, während c) eine systematische Verschiebung der Ergebnispunkte eines Versuches und damit eine Verschiebung der Ergebniskurve relativ zueinander bewirkt. Letzterer Einfluß ist in den Ergebniskurven von Abb. 3 und 4 noch enthalten und wird erst in der Approximation der Kurvenscharen durch das mathematische Modell (vgl. folgende Abschnitte) ausgemittelt.

Um alle drei Einflüsse a) bis c) zu erfassen, wurden die Abweichungen der Ergebnispunkte von über das Modell synthetisierten Kurven ausgewertet und dem aus Wort- und Probandenzahl berechneten erwarteten Streubereich gegenübergestellt. Abb. 5 enthält als Beispiel die Gegenüberstellung für die Versuchsreihe A1. Es zeigte sich, daß die Streuung im wesentlichen durch die statistische Streuung nach a) bestimmt ist.

4. Mathematische Analyse

4.1 Versuchsgruppe A (Einzelwörter)

Zunächst konnten nur die Ergebnisse der Gruppe A durch ein mathematisches Modell beschrieben werden. Die später durchgeführte Analyse der Kontextergebnisse fußt sehr stark auf diesem Modell und seiner Interpretation. Deshalb soll im folgenden das Modell der Erkennung von Einzelwörtern für sich dargestellt werden.

4.1.1 Mathematisches Modell der Erkennung von Einzelwörtern

Die Werte $E(n, l)$ der Versuchsgruppe A können in zwei Anteile zerlegt werden:

- einen Anteil, der nur von der Abkürzungslänge n, aber nicht von der Wortlänge l abhängig ist. Er wurde mit "Grundeffekt" benannt und wird mit "G" indiziert.
- einen zweiten Anteil, der von E = 100 % rasch auf E = 0 ab-

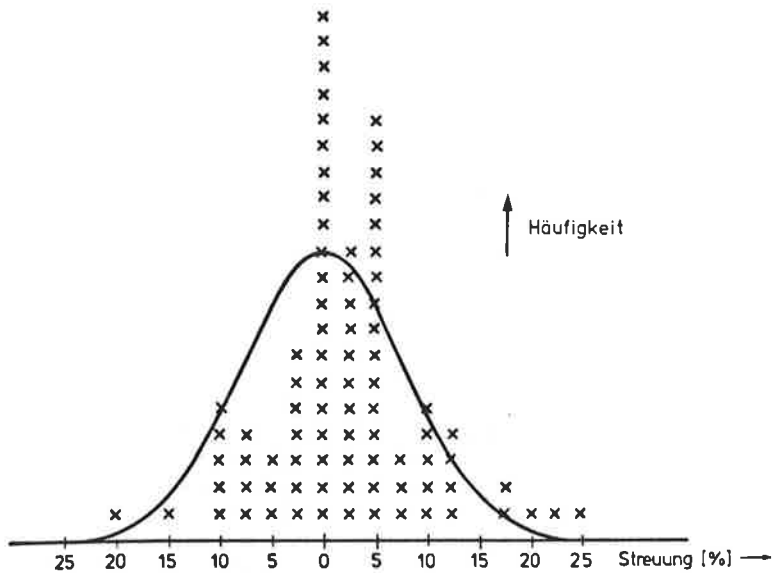


Abb. 5 Versuchsreihe A1. Streuung der experimentellen Ergebnispunkte um die Modellkurven (x) gegenüber der aus Probandenzahl und Wortzahl errechneten, mit σ = 7,5 Prozentpunkten erwarteten Streuung (—).

nimmt, wenn die Wortlänge die Abkürzungslänge übersteigt. Er wurde mit "Proximitätseffekt" benannt und wird mit "p" indiziert.

Die zwei Anteile setzen sich nach

$$E = E_G + E_P - E_G \cdot E_P \qquad (1)$$

zur gesamten Erkennbarkeit zusammen.

$E_G(n)$ kann aus den Versuchsergebnissen als der konstante Wert entnommen werden, gegen den die Kurven $E(n = \text{const}, l)$ mit steigendem l konvergieren. Damit läßt sich auch $E_P(n,l)$ entsprechend Gl. (1) aus den Kurven separieren.

Abb. 6 (links) zeigt als Beispiel die Kurven für die beiden Einzeleffekte, wie sie aus den Erkennbarkeitskurven der Versuchsreihe A1 (s. Abb. 3 links oben) separiert wurden. Die E_P-Kurven für die großen Abkürzungslängen n dürfen dabei nicht zu ernst genommen werden, da sie aus kleinen Differenzen gestreckt sind, wobei die in diesem Bereich ohnehin relativ große Streuung der Originalkurven stark vergrößert wird.

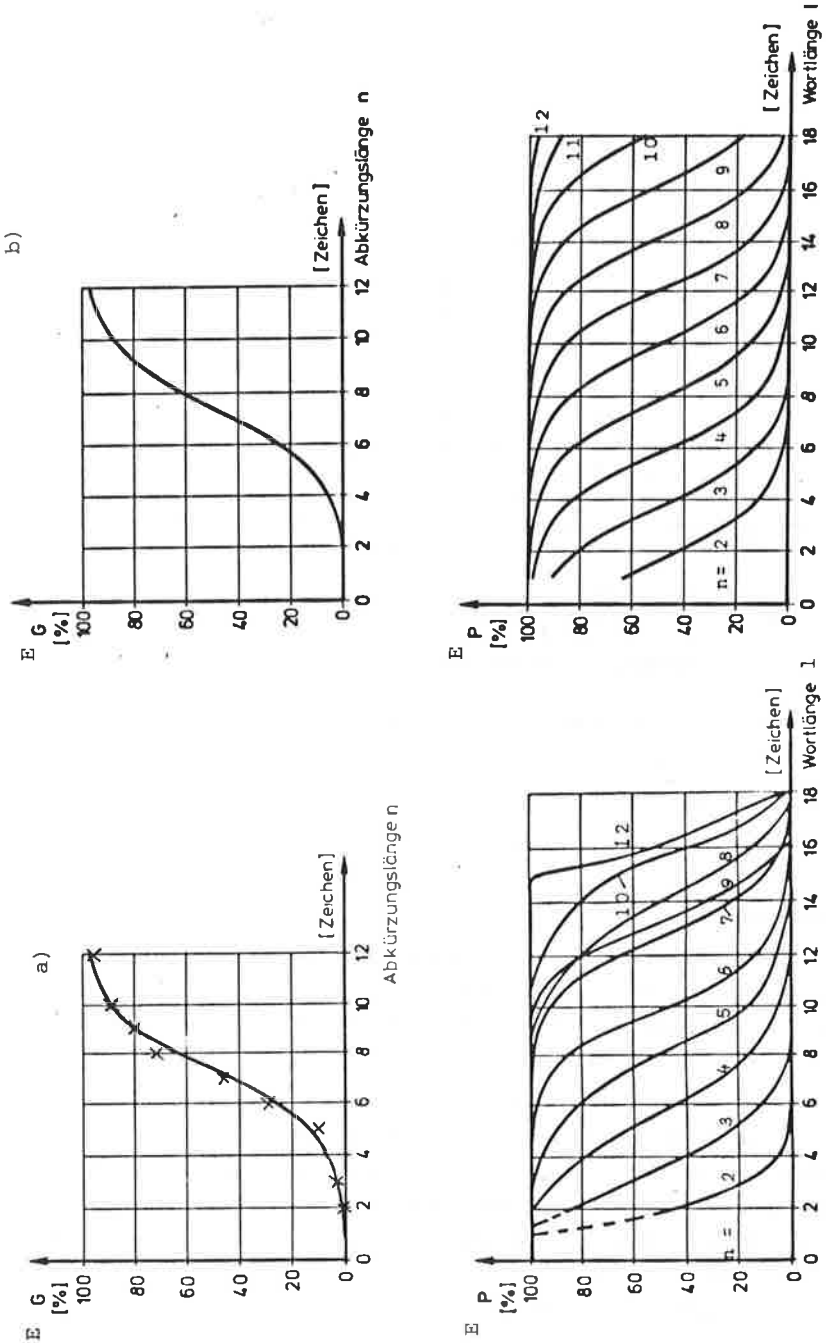


Abb.6 Versuchsreihe A1 (Allgem.Wortkenntnis). Gegenüberstellung der aus den Versuchsergebnissen separierten Kurven für die Einzeleffekte G und P (a) und der entsprechenden mit dem Modell

Durch Versuch und Fehlschlag wurde zunächst für E_G die brauchbare Approximation

$$E_G(n) = \frac{1}{1+2^{(A - K_1' \cdot n)}} \qquad (2)$$

gefunden, d.h. eine Transformation, über die sich die Grundeffekt-Kurve $E_G(n)$ in die lineare Funktion $A - K_1' \cdot n$ verwandeln läßt. (Die Basis 2 im Nenner von Gl.(2) wurde willkürlich gewählt, um den Exponenten auf das Maß "bit" zu bringen.) Geraume Zeit später stellte sich heraus, daß die gleiche Transformation auch in der Lage ist, die Kurvenschar $E_P(n,l)$ approximativ in eine Schar äquidistanter Geraden zu verwandeln. Man kann also zusammenfassend schreiben:

$$E_{G,P} = \frac{1}{1+2^{-\Delta i_{G,P}}} \qquad (3)$$

mit

$$\Delta i_G(n) = - A + K_1' \cdot n \qquad (4)$$

$$\Delta i_P(n,l) = - B - K_1 \cdot l + K_2 \cdot n \qquad (5)$$

Danach verblieb die Aufgabe, die Koeffizienten in Gln.(4) und (5) so zu bestimmen, daß sich nach Transformation entsprechend Gl.(3) und Überlagerung nach Gl.(1) die beste Näherung an die Kurvenscharen der Versuchsreihen A1, A2 und A3 ergab. Hierzu wurden die Versuchsergebnisse entsprechend Gl.(1) in Grund- und Proximitätsanteil separiert, die Einzelwerte nach Gl.(3) transformiert und die so ermittelten Punktescharen durch Geraden approximiert. Tab.2 enthält die so gefundenen Werte für die Koeffizienten. Abb.6 (rechte Seite) zeigt am Beispiel der Versuchsreihe A1 die mit diesen Koeffizienten nach Gln.(5) bis (3) und (1) synthetisierten Kurven der Einzeleffekte, Abb.3 (rechte Seite) die synthetisierten Gesamtkurven für A1 bis A3.

Tab. 2. Koeffizienten in Δi_G, Δi_P für Versuchsgruppe A

Versuchsreihe	Δi_G		Δi_P		
	-A	K_1'	-B	K_1	K_2
A1 ALLGEMEINE WORT-KENNTNIS	- 8,5	1,14	- 3	1,2	2,5
A2 SPEZIELLE WORT-KENNTNIS	- 3,3	1,1	- 1,29	1,21	2,5
A3 SPEZIELLE WORT- U. ABKÜRZUNGS-KENNTNIS	-1,26	0,985	+ 1,54	1,33	2,53

4.1.2 Interpretation des Modells

Erwähnenswert ist zunächst die Tatsache, daß ein einziges mathematisches Modell in der Lage ist, die Kurvenscharen aller drei Versuchsreihen A1 bis A3 angemessen zu approximieren.

Die lineare Gleichung (4) beschreibt im Bitmaß eine Informationsdifferenz zwischen einem bestehenden Anfangs-Informationsdefizit (-A) und einer durch die Abkürzung gelieferten Information $K_1'n$, bei der jeder Buchstabe mit K_1' bit bewertet wird. Gl.(5) beschreibt einen ähnlichen Vorgang, jedoch wird hier das Informationsdefizit in Abhängigkeit von der Zielwortlänge l mit K_1 bit/Buchstabe vergrößert, und die angebotene Abkürzung erhält über K_2 eine gegenüber dem Grundeffekt vergrößerte Informationswertigkeit pro Buchstabe.
Tabelle 2 zeigt, daß sich die Buchstabenwertigkeiten K_1', K_1 und K_2 trotz geänderter Versuchsbedingungen in A1 bis A3 praktisch invariant verhalten, während sich die Anfangs-Informationsdefizite A und B von A1 nach A3 fortschreitend stark verkleinern - in Einklang mit den Versuchsbedingungen A1 bis A3, die ja sukzessive den Pbn immer mehr Vorinformation zuführten.

Ein weiterer interessanter Aspekt steckt in Gl.(1). Sie hat die Form einer bekannten statistischen Formel, die die Gesamtwahrscheinlichkeit E eines Ereignisses beschreibt, das von zwei unabhängig arbeitenden Prozessen (mit den zugeordneten Wahrscheinlichkeiten E_P und E_G) ausgelöst wird. Das heißt, daß die Erkennbarkeit E zumindest näherungsweise aufgefaßt werden kann als die Überlagerung der Ergebnisse zweier unabhängig voneinander ablaufenden Erkennungsprozesse nach P und G.
Die Transformation nach Gl.(3) hat die Form einer Fermi-Dirac-Verteilung. Sie besagt, daß eine Informationsdifferenz $\Delta i = 0$ erst zu einer Erkennbarkeit von 50 % führt. Völlige Erkennbarkeit kann erst durch positive Informationsdifferenzen von 4 bis 6 bit, d.h. durch Redundanz, angenähert werden. Umgekehrt ergeben aber auch negative Informationsdifferenzen noch merkliche Erkennbarkeitswerte.
Die gefundenen Buchstabenwertigkeiten, die Transformation und die Überlagerung nach Gl.(1) werden später noch ausführlicher diskutiert.

4.2 Versuchsgruppe B (Kontext)

Die ersten Versuche, ein mathematisches Modell für die Erkennung von Wörtern im Kontext zu erstellen, gingen von der Vorstellung aus, daß auch der Kontexteinfluß die Buchstabenwertigkeiten konstant läßt und nur die vorbestehenden Informationsdefizite variiert. Es wurde daher probiert, die Kurvenscharen der Versuchsreihen B1 bis B3 mit dem Modell für unverbundene Wörter durch Variation lediglich der Koeffizienten A und B zu beschreiben. Das Ergebnis waren nicht-monotone Verläufe der B-Koeffizienten, die sowohl von l als auch von n abhängig waren; also ein völlig unplausibles Resultat. Immerhin konnte aus den Verläufen geschlossen werden, daß weitere - von den bisher gefundenen abweichende - Buchstabenwertigkeiten im Spiele waren. Im folgenden sind die Schritte beschrieben, die schließlich zu einer brauchbaren und plausiblen Approximation führten.

4.2.1 Mathematisches Modell der Erkennung abgekürzter Kontextwörter

Der zweite Versuch, die Kurvenscharen der Versuchsreihen B1, B2 und B3 (Kontext) durch ein einheitliches Modell zu approximieren, ging von folgenden Annahmen aus, die aus den Ergebnissen und Interpretationen des Modells für Einzelwörter abgeleitet wurden:

- Buchstabenwertigkeiten waren dort als Invarianten in "Effekten" vorgekommen. Demgemäß wurde angenommen, daß die neuen Buchstabenwertigkeiten auf die Existenz eines weiteren Effektes hindeuteten. Dieser wurde "K" für "Kontexteffekt" benannt.
- Sowohl der G- als auch der P-Effekt konnten durch eine lineare Gleichung in n, oder l und n beschrieben werden, die durch eine auf beide Effekte zutreffende Transformation die zugehörige Erkennbarkeit lieferte. Die gleiche mathematische Form mit der gleichen Transformation wurde daher auch für den "K"-Effekt angesetzt.
- Die Überlagerung von E_P und E_G nach Gl.(1) zur Gesamterkennbarkeit E wurde im vorstehenden als statistische Überlagerung zweier unabhängig wirkenden Effekte gedeutet. Diese Interpretation erlaubt die Ausweitung der Gleichung auf beliebig viele Effekte. Deshalb wurde unabhängige Überlagerung auch für den K-Effekt angenommen.

Aufgrund dieser Annahmen entsteht ein um einen Effekt erweitertes Modell der Form:

$$E = E_P+E_G+E_K-E_PE_G-E_PE_K-E_GE_K+E_PE_GE_K \qquad (6)$$

$$E_{G,P,K} = \frac{1}{1+2^{-\Delta i_{G,P,K}}} \qquad (7)$$

$$\Delta i_G = -A + K_1' n \qquad (8)$$

$$\Delta i_P = -B - K_1 \cdot l + K_2 \cdot n \qquad (9)$$

$$\Delta i_K = C - K_3 \cdot l + K_3' \cdot n \qquad (10)$$

Für die Approximation der Kontextkurvenscharen wurden die Buchstabenwertigkeiten $K_1' = 1{,}14$ $K_1 = 1{,}2$ $K_2 = 2{,}5$ bit/Buchstaben ungeändert aus der Versuchsreihe A1 übernommen, so daß nur noch die Buchstabengewichte K_3 und K_3' sowie die Informationsdefizite A, B und C zu bestimmen sind. Weitere einschränkende Bedingungen waren:

- Die Buchstabenwertigkeiten K_3 und K_3' sollen für alle 3 Versuchsreihen B1 bis B3 invariant sein, d.h. dasselbe Verhalten aufweisen, das sich für K_1', K_1 und K_2 in den Versuchsreihen A2 bis A3 gezeigt hatte.
- Innerhalb einer Versuchsreihe B1 bis B3 sollen die Informationsdefizite A, B und C nur von der Abkürzungslänge n, aber nicht von der Zielwortlänge l abhängen. D.h., es wurde angenommen, daß das Kontextverständnis - als Quelle der für die Wortinterpretation zusätzlich verfügbaren Information - nur eine Funktion des Textes und der gewählten Abkürzungslänge ist.
- Innerhalb einer Versuchsreihe B1 bis B3 sollen die Informationsdefizite A, B und C monoton einem Sättigungswert zustreben, der bei Abkürzungslängen über n = 4 erreicht wird. Diese Annahme entspricht der aus dem Verhalten der Kontextkurven abgeleiteten Arbeitshypothese (s. 2.6).

Unter diesen Randbedingungen wurden die erforderlichen Buchstabenwertigkeiten des K-Effektes bestimmt zu

$K_3 = 0{,}6$ $\qquad$ $K_3' = 0{,}7$ bit/Buchstabe

Die Verläufe der bei der Approximation gefundenen Werte für die A, B, C-Koeffizienten zeigt Abb.7; die mit diesen Werten nach Gln.(10) bis (6) synthetisierten Kurvenscharen sind in Abb.4 (rechte Seite) den experimentellen Kurven gegenübergestellt.

Zur Approximation wurde ein Rechner benutzt, auf dem das Modell nach Gl.(6) bis (10) abgesetzt war. Durch systematische Variation der noch zu bestimmenden Konstanten ließ sich so das beste Approximationsergebnis für die Einzelkurven aufsuchen.

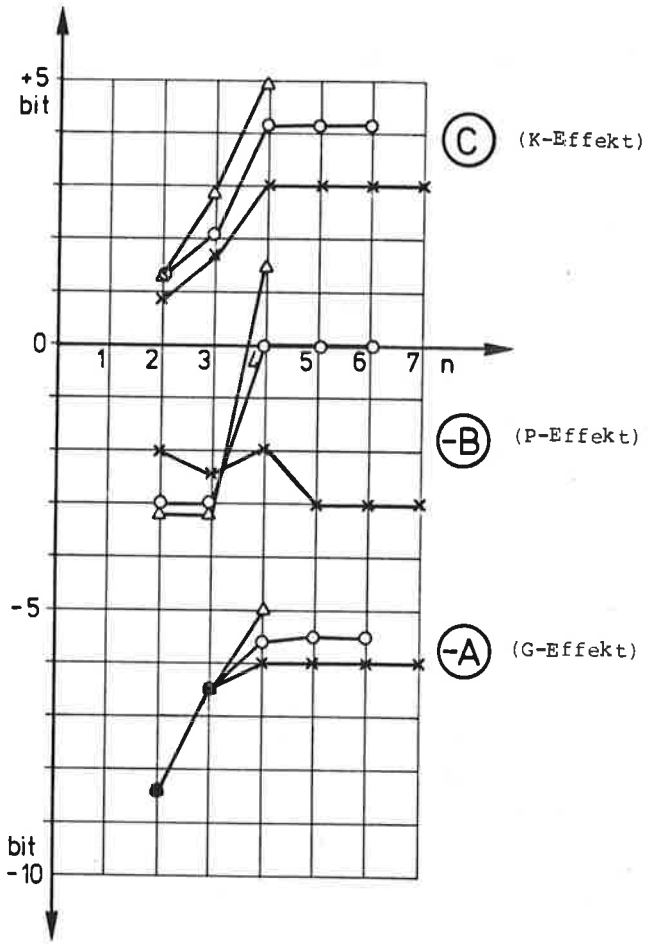


Abb.7 Versuchsgruppe B. Verlauf der Informationsdefizite bzw. -boni A,B,C in Abhängigkeit von der Abkürzungslänge und dem Schwierigkeitsgrad des Textes

x——x B1 (BISMARCK)

o——o B2 (SPIEGEL)

Δ——Δ B3 (LITERARISCH)

Alternativ zu dem 3-Effekt-Modell wurde zeitweise auch versucht, die Kontext-Kurven allein mit den beiden Effekten G und K zu approximieren, was erheblich schlechtere Resultate lieferte.

4.2.2 Interpretation des Modells

Das mathematische Modell unter Einschluß des Kontexteffektes ist zunächst als Zangengeburt aus den Ergebnissen und Interpretationen des Modells für die Erkennbarkeit von Einzelwörtern entstanden. Es erhält jedoch nachträglich seine Rechtfertigung daraus, daß es die Kurvenscharen aller 3 Versuchsreihen B1 bis B3 gut approximieren kann, und vor allem daraus, daß es für die Buchstabenwertigkeiten und Verläufe der Informations-Defizite bei den verschiedenen Textvarianten plausible Werte liefert.

Die Plausibilität der Buchstabenwertigkeit kann erst unter Beachtung der im folgenden Abschnitt 5. abgeleiteten Ergebnisse gezeigt werden.

Aus den Verläufen für A, B, C nach Abb.7 können folgende Eigenheiten abgelesen werden:

- Bei der Abkürzungslänge n = 2 beginnen A und B als Informationsdefizit, C als kleiner Informationsbonus.
- Mit steigender Abkürzungslänge n steigen auch die Werte von A, B und C im allgemeinen an, über n = 4 laufen sie auf einen Sättigungswert ein. Das beweist, daß die als Postulat in die Approximationsbedingungen eingeführte Arbeitshypothese eine vernünftige Approximation der Kurvenscharen zumindest nicht ausschließt.
- Die subjektiv beurteilte Textschwierigkeit nimmt von B1 (Bismarck) zu B3 (Literarisch) ab. Entsprechend zeigen die gesättigten Werte von A,B, C zunehmend höhere Informationsgewinne (bzw. einen größeren Abbau des Informationsdefizits). D.h. daß die Worterkennung aus leichteren Kontexten wesentlich mehr Zusatzinformation empfängt als aus schwierigen Kontexten.

- Die höheren Informationsgewinne aus den leichteren Kontexten zeigen sich insbesondere im P- und im K-Effekt (Werte B und C). Im G-Effekt (Wert A) ist der Einfluß gering.

5. Diskussion der gefundenen Buchstabenwertigkeiten

5.1 Vergleich mit den mittleren Buchstabenwertigkeiten der deutschen Sprache

Der mittlere Informationsgehalt eines Buchstabens ist stark davon abhängig, ob man einen isolierten Buchstaben, einen Buchstaben in einer Silbe, in einem Wort oder in einem Text betrachtet. Je größer der Komplex ist, innerhalb dessen die Ermittlung geschieht, desto kleiner fällt der mittlere Informationsgehalt des Buchstabens aus; vergleichbar etwa mit dem "Ahnenschwund", der um so stärker ins Gewicht fällt, je größer die betrachtete Personengruppe aus einer Population ist.
Der mittlere Informationsgehalt eines Buchstabens der deutschen Sprache beträgt (Küpfmüller 1954, Zemanek 1959, Steinbuch 1965, Fucks 1970):

Für einen isolierten Buchstaben: 4,3 (Anfangs-) bis 4,1 bit/Buchstabe

Für einen Buchstaben im isolierten Wort: ≈ 2 bit/Buchstabe
Für einen Buchstaben im Kontext: 1 bis 1,3 bit/Buchstabe
(Silben wurden bei der vorliegenden Untersuchung nicht betrachtet.)

Die Skala der Buchstabenwertigkeiten, die im vorliegenden Papier gefunden wurde, sieht dagegen wie folgt aus:

Koeffizient von l:

$K_1 = 1{,}2..1{,}3$
$K_3 = 0{,}6$

Koeffizient von n:

$K_2 = 2{,}5$ bit/Buchstabe
$K_1' = 1..1{,}14$ bit/Buchstabe
$K_3' = 0{,}7$ bit/Buchstabe

Dabei fällt auf, daß beide Skalen drei Schritte umfassen und die Schrittweite in beiden Fällen etwa dem Faktor 2 entspricht. Die beiden Skalen sind jedoch im Mittel um den Faktor T = 1,75 gegeneinander versetzt.

Tab. 3. Koeffizienten in Δi_G, Δi_P und Δi_K nach Berücksichtigung des Faktors T

Versuchsreihe	Δi_G		Δi_P			Δi_K		
	- A	K_1'	- B	K_1	K_2	C	K_3	K_3'
A1 ALLGEMEINE WORTKENNTNIS	-14,9	2	-5,25	2,1	4,37			
A2 SPEZIELLE WORTKENNTNIS	- 5,8	1,92	-2,25	2,12	4,37			
A3 SPEZIELLE WORT-U.ABKÜRZUNGSKENNTNIS	- 2,2	1,72	+2,7	2,33	4,43			
B1 BISMARCK								
n=2	-14,7	2	-3,5	2,1	4,37	1,4	1,05	1,22
3	-11,3		-4,0			3		
4	-10,5		-3,5			5,3		
5	-10,5		-5,3			5,3		
6	-10,5		-5,3			5,3		
7	-10,5		-5,3			5,3		
B2 SPIEGEL								
n=2	-14,7	2	-5,3	2,1	4,37	2,3	1,05	1,22
3	-11,3		-5,3			3,8		
4	- 9,5		0			7,35		
5	- 9,6		0			7,35		
6	- 9,6		0			7,35		
B3 LITERARISCH								
n=2	-14,7	2	-5,6	2,1	4,37	2,3	1,05	1,22
3	-11,3		-5,6			4,9		
4	- 8,75		+2,6			8,8		

Der Grundeffekt berücksichtigt die Zielwortlänge nicht und erweist sich dadurch als ein Effekt, der Wörter als Ganzes zählt. Setzt man dies voraus, dann sollte sein Anfangs-Informationsdefizit dem Entscheidungsgehalt des Wortschatzes entsprechen, in welchem das Zielwort gesucht wird. Nun nimmt nach Multiplikation mit 1,75 der Koeffizient A in der Testreihe A1 (Allgemeine Wortkenntnis) Werte von ca. 15 bit an. Gleiches gilt für die Werte von A in den Testreihen B1, B2, B3 für die Abkürzungslänge n = 2, bei der anscheinend noch keine erhebliche Zusatzinformation aus dem Kontext gezogen wird. Die Zahl $2^{15} \approx 33000$ entspricht tatsächlich etwa dem passiven Wortschatz eines Erwachsenen oder der Wortmenge (40-60000) der Quellen, aus denen die Zielwörter bezogen wurden.

In der Versuchsreihe A2 (spezielle Wortkenntnis) wurde der relevante Wortschatz durch Memorieren auf 50 Zielwörter eingegrenzt, was einem Entscheidungsgehalt von ld(50) = 5,64 bit entspricht. Der in der Versuchsreihe A2 für das Anfangs-Informationsdefizit A nach Multiplikaton mit T gefundene Wert lautet 5,8 bit.

Wenn man die in diesem Abschnitt getroffene Interpretation der Buchstabenwertigkeiten akzeptiert, so folgt daraus, daß sich die Modelle noch verfeinern lassen durch genauere Erfassung der Informationswertigkeiten der Abkürzungsbuchstaben, speziell auch des ersten Buchstabens und seiner Sonderbehandlung im Abkürzungsverfahren.

6. Allgemeine Diskussion

Die hier vorgelegten Untersuchungen wurden in der Absicht begonnen, die Einflußgrößen bei der Interpretation abgekürzter Wörter durch den Menschen quantitativ zu erforschen und evtl. auf phänomenologischer Basis berechenbar zu machen. Die schließlich entstandenen Modelle tragen jedoch bereits funktionelle Züge und lassen sich informationstheoretisch interpretieren, so daß der Eindruck einer abgerundeten oder zumindest auf dieser Basis abrundbaren Theorie entsteht. Der Schein trügt. Die Ergebnisse werfen mehr Fragen auf als sie beantworten.

5.2 Konsequenz für die Transformation

Die Vermutung liegt nahe, daß es sich in beiden Fällen um die gleiche Skala handelt und daß bei der Modellbildung noch der Faktor T zu berücksichtigen ist. Dies würde bedeuten, daß alle bisher gefundenen Buchstabenwertigkeiten und A, B, C-Werte mit 1,75 zu multiplizieren sind und daß die Transformation die Form annimmt:

$$E_{G,P,K} = \frac{1}{1+2^{\frac{-\Delta i_{G,P,K}}{T}}} \qquad (11)$$

Die mit 1,75 multiplizierten Koeffizienten der Modelle sind in Tabelle 3 aufgeführt.
Die Multiplikation bewirkt, daß die Koeffizienten K_3 und K'_3 im Kontexteffekt Werte annehmen, die der mittleren Buchstabenwertigkeit im Kontext entsprechen. Im Grundeffekt, der Wörter als Ganzes zählt, nimmt $K'_1 = 1,7..2$ die Wertigkeit eines Buchstabens im Wort an. Im Proximitätseffekt wird die Zielwortlänge mit $K_1 = 2,1..2,3$ bewertet, was ebenfalls noch rund der mittleren Buchstabenwertigkeit im Wort entspricht; in der Abkürzung wird mit $K_2 = 4,4$ der Buchstabe deutlich höher bewertet als der mittlere Informationsgehalt eines isolierten Buchstabens (der recht genau bestimmt werden kann). Letztere Anhebung hat wohl ihre reale Ursache darin, daß durch Auswahl signifikanter Buchstaben der Informationsgehalt der Abkürzung tatsächlich über das Mittel angehoben wird.

Daß die Buchstabenwertigkeiten durch die Multiplikation mit T in einen Bereich gerückt werden, in dem sie interpretierbar werden und mit Werten korrespondieren, die bereits in der Literatur (s.o.) vorliegen, macht die Existenz eines Faktors T in der Transformation plausibel, obwohl er für die Kurvenapproximation irrelevant ist. Bemerkenswerter und beweiskräftiger scheinen folgende Korrespondenzen zu sein:

Die wesentlichste Frage dabei ist, wieso sich überhaupt aus der alleinigen Betrachtung der Erkennbarkeit bzw. der Treffer von Zielwörtern abgeschlossene mathematische Modelle mit guten Approximationseigenschaften gewinnen lassen, ohne Berücksichtigung des in Wirklichkeit dreiwertigen - nicht zweiwertig komplementären - Schemas "Treffer, Fehler, Schweigen" (s.2.5). Wenn den gefundenen "Effekten" tatsächlich unabhängig ablaufende Erkennungsprozesse zugrundeliegen, dann ist es keineswegs selbstverständlich, daß sie sich in ihrer Trefferrate ergänzen, ohne sich gegenseitig durch ihre Fehlerrate zuzudecken.
Weitere Fragen betreffen die Kompaktheit und Einheitlichkeit der beiden Modelle für den Worterkennungsprozeß aus der Abkürzung, obwohl dieser sicher aus mehreren trennbaren Komponenten besteht. Der Versuchsablauf kann z.B. gegliedert werden in die beiden Komponenten:

a) Der Pb produziert aus der Abkürzung (subjektiv) ein mögliches Zielwort.
b) Es wird (objektiv) festgestellt, ob das mögliche Zielwort dem ursprünglich abgekürzten Zielwort entspricht.

Eine andere Komponentengliederung des Erkennungsvorgangs könnte erfolgen nach

c) den anteiligen Leistungen des Pbn
d) den anteiligen Leistungen des Sprachaufbaus.

Die Grenzlinien dieser Teilung sind nicht identisch mit der Teilung nach a) und b), da der Sprachaufbau sowohl die objektive Sprachstatistik als auch das vom Pb bei der Wortproduktion subjektiv benutzte Sprachmodell bestimmt. Die Möglichkeit, einheitliche Erkennungsmodelle - ohne Berücksichtigung dieser Komponenten - zu formulieren, könnte ein Hinweis darauf sein, daß die Komponenten zueinander optimal eingestellt sind.

Die Transformation entsprechend einer Fermi-Dirac-Statistik, welche in den hier vorgelegten Modellen die Informationsdifferenz mit der Erkennbarkeit koppelt, scheint eine recht zentrale Rolle zu spielen. Es entsteht die Frage, ob sie einer der oben genannten Komponenten des Erkennungsvorgangs zugeordnet werden kann oder ob sie eine Funktion des Gesamtprozesses ist.

Interessant ist in dieser Beziehung der in Herdan 1962 zu findende Abschnitt "Chance, the ever-present alternative" (pp.179 bis 213), in dem er feststellt, daß sich bestimmte Wortverteilungen in laufenden Texten als Bose-Einstein-Statistik deuten lassen. Er erwähnt an einer Stelle ausdrücklich: "That we have only considered Bose-Einstein-Statistics... does not mean, that Fermi-Dirac statistics have no place in vocabulary partition". Die Verfasser legen Wert auf die Feststellung, daß ihnen Herdans Arbeit erst bekannt wurde, nachdem die Transformation (3) bzw. (11) aus den Versuchsergebnissen ermittelt und als Fermi-Dirac-Verteilung interpretiert worden war. Diese Information ist daher nicht a priori in die Modellbildung eingeflossen.

Schließlich sei nochmals darauf hingewiesen, daß die hier vorgelegten Ergebnisse an erwachsenen Pbn ohne Stressituation z.B. durch Zeitdruck und ohne besondere motivationale Anreize gewonnen wurden. Durch geeignete Variation situativer Bedingungen ließen sich noch weitere Einblicke in die zugrundeliegenden Mechanismen gewinnen.

7. Danksagung

Die hier beschriebenen Arbeiten wurden anteilig durch Forschungsmittel der Universität Duisburg Gesamthochschule gefördert. Die Verfasser haben weiterhin U. Raatz und Ch. Klein-Braley für wertvolle Hinweise und weiterführende Diskussionen zu danken.

Literatur

FUCKS,W. (1970) Gibt es mathematische Gesetze in der Sprache? Kybernetik Umschau (1970), S.197-207

GOLDENBERG,D., RUMPEL,D. (1983a) A Quantitative Analysis of the Recognition of Abbreviated Words by Man. Intern.Classification 10 (1983), No.2, S.84-86

GOLDENBERG,D., RUMPEL,D. (1983b) Recognition of Abbreviated Context-Words by Man. Intern.Classification 10 (1983), No.3, S.143-146

HERDAN,G.(1962) The Calculus of Linguistic Observations. Monton & Co. S'Gravenhage 1962

KÜPFMÜLLER,K. (1954) Die Entropie der deutschen Sprache. Fernmeldetechn.7 (1954), S.265-272

SCHMITZ,U. Vorbemerkungen zur Linguistik der Abkürzungen. (Prol. Ling. Abk.) in: Sprache, Diskurs und Text/ Akten des 17. Linguistischen Kolloquiums, Brüssel, 1982, Bd.1, 10-17.

STEINBUCH,K. (1965) Automat und Mensch. 3rd ed. Berlin: Springer 1965

ZEMANEK,H. (1959) Elementare Informationstheorie. Oldenbourg Verlag München und Wien

---

Anhang:

Die Fermi-(Dirac-)Statistik (oder -Verteilung) ist in den Lehrbüchern der Physik und der statistischen Thermodynamik zu finden, z.B. in der Form

$$f_{(\varepsilon)} = \frac{1}{1 + e^{\frac{\varepsilon - \mu}{T}}}$$

geschrieben. Darin stellt e die Basis des natürlichen Logarithmus, ε die Energie des betrachteten Orbitals, μ ein Maß für die Partikelkonzentration und T die absolute Temperatur dar. Die Fermi-Dirac-Statistik gilt für die Partikelklasse der Fermionen (z.B. Elektronen) und gibt an, mit welcher Wahrscheinlichkeit das betrachtete Orbital durch eine Partikel besetzt ist.

Semantic Diversification of Hungarian Verbal Prefixes

III. "föl-", "el-", "be-"

E. Beöthy, Amsterdam
G. Altmann, Bochum

1. In our previous papers (Beöthy, Altmann 1984 a, b) we proceeded from the fact that some classes of frequently used morphemes are subject to a semantic diversification process, i. e. they assume new meanings, combine them with already existing ones and become semantically "diffuse". This diversification process is nevertheless unique and can be expressed in mathematical terms. As has already become clear, this diversification is a result of the operation of "Zipfian forces" used by the speaker and the hearer in order to make communication effective. The speaker has the tendency to invest a word (a meaningful unit) with as many meanings as possible, while the hearer tries to invest it with a single meaning. These trends follow from Zipf's principle of least effort (Zipf 1972). If these assumptions are correct then it is possible to derive the rank-frequency distribution of the meanings from a stochastic birth process (Poisson process) in which the proportionality factor is not constant but a function of the rank attained and comprises both the forces mentioned. The solution of the equation leads to the displaced negative binomial distribution and was corroborated both by the data of "meg-" and "ki-".

It can be shown that the negative binomial distribution fits the trend with the other prefixes adequately as well. All occurrences of these prefixes in J.örkény's novel "A rózsa kiállitás" were ascertained and examined semantically. We obtained the results presented in tables 1, 2 and 3. Since at the same time we also analyzed the translations of the verbs with these prefixes into Dutch, only those occurrences are displayed that were really translated (two cases of "föl-", one case with "el-" and two cases with "be-" are missing).

Table 1: Frequencies of particular meanings of "föl-"

1.	Completed action + figurative meaning	11
2.	Completed action + new meaning	7
3.	New meaning	7
4.	Completed action + one-time action	6
5.	Completed action	5
6.	Figurative meaning + new meaning	4
7.	Direction	3
8.	Completed action + direction	3
9.	Completed action + its result	3
10.	Completed action + modified meaning	3

Table 2: Frequencies of particular meanings of "el-"

1.	Completed action	83
2.	Completed action + new sense	9
3.	Completed action + direction	3
4.	Lasting effect	2
5.	New meaning	2
6.	Direction	1
7.	Completed action + figurative meaning	1
8.	Figurative meaning + new meaning	1
9.	Completed action + its result	1

Table 3: Frequencies of particular meanings of "be-"

1.	Completed action + new meaning	20
2.	Completed action	11
3.	Completed action + direction	10
4.	Figurative meaning	7
5.	Direction	5
6.	Completed action + its result	3
7.	Result of the action	3
8.	Completed action + modified meaning	3
9.	Completed action + figurative meaning	2
10.	Figurative meaning + instantaneous action	1
11.	Modified meaning	1
12.	Direction + new meaning	1
13.	New meaning	1

All these distributions follow the negative binomial distribution

$$P_x = \binom{r + x - 2}{x - 1} p^r q^{x-1} \qquad x = 1,2,... \qquad (1)$$

as shown in table 4. Here we estimated the parameters from the mean and the first frequency class of the distribution (cf. Beöthy, Altmann 1984 a).

It is evident that the results are extremely satisfactory though it would perhaps be possible to obtain still better estimations.

Table 4: Observed and computed frequencies of the meanings of the three prefixes

	föl-		el-		be-	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	11	11.00	83	83.00	20	20.00
2	7	9.45	9	8.93	11	13.06
3	7	7.53	3	4.05	10	9.22
4	6	5.85	2	2.30	7	6.67
5	5	4.48	2	1.44	5	4.48
6	4	3.41	1	0.95	3	3.60
7	3	2.58	1	0.65	3	2.67
8	3	1.94	1	0.45	3	1.99
9	3	1.46	1	1.23	2	1.48
10	3	4.30	-	-	1	1.11
11	-	-			1	0.83
12					1	0.62
13					1	1.87
	p = 0.2642 r = 1.1671 X_7^2 = 3.50 P = 0.83		p = 0.2015 r = 0.1348 X_4^2 = 0.67 P = 0.95		p = 0.2414 r = 0.8610 X_9^2 = 1.88 P = 0.99	

2. Let us now consider the Dutch translation means for the given meanings of the Hungarian prefixes. We obtained the results presented in tables 5, 6 and 7.

Table 5: Frequencies of Dutch translation means for "el-"

Meaning	1 Paraphr.	2 ver-	3 No pref.	4 werd	5 Other verb	6 af-	7 uit-	8 los-	9 onder-	10 ver-	11 aan-	12 ont-	13 om-	14 op-	15 weg-	Σ
1	30	12	15	8	7	3	2	1	-	-	1	1	1	1	1	83
2	3	4	1	-	-	-	-	1	-	-	-	-	-	-	-	9
3	1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	3
4	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	2
5	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	2
6	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
7	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
8	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	1
9	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1
Σ	38	19	17	8	7	3	2	2	1	1	1	1	1	1	1	103

Table 6: Frequencies of Dutch translation means for "föl-"

Meaning	1 Paraphr.	2 op-	3 No pref.	4 uit-	5 be-	6 ver-	7 ab-	8 vor-	9 voor-	10 ont-	11 aan-	12 re-	Σ
1	2	3	2	2	1	-	1	-	-	-	-	-	11
2	2	3	-	-	-	1	-	1	-	-	-	-	7
3	1	2	2	2	-	-	-	-	-	-	-	-	7
4	5	-	1	-	-	-	-	-	-	-	-	-	6
5	3	-	-	-	-	1	-	-	1	-	-	-	5
6	-	1	1	-	-	-	-	-	-	1	1	-	4
7	2	-	-	1	-	-	-	-	-	-	-	-	3
8	1	1	1	-	-	-	-	-	-	-	-	-	3
9	-	-	2	-	-	-	-	-	-	-	-	1	3
10	-	2	-	-	1	-	-	-	-	-	-	-	3
Σ	16	12	9	5	2	2	1	1	1	1	1	1	52

Table 7: Frequencies of Dutch translation means for "be-"

Meaning	1 Paraphr.	2 in-	3 Other verb	4 No prefix	5 binnen-	6 aan-	7 voor-	8 ver-	9 toe-	10 af-	11 terug-	12 mee-	Σ
1	4	3	7	1	2	-	2	1	-	-	-	-	20
2	2	-	-	5	1	1	-	-	1	1	-	-	11
3	2	2	-	1	2	-	-	-	1	-	1	1	10
4	1	5	1	-	-	-	-	-	-	-	-	-	7
5	1	1	1	2	-	-	-	-	-	-	-	-	5
6	1	-	-	1	-	1	-	-	-	-	-	-	3
7	-	1	-	-	-	1	-	1	-	-	-	-	3
8	-	-	2	-	-	1	-	-	-	-	-	-	3
9	1	-	1	-	-	-	-	-	-	-	-	-	3
10	1	-	-	-	-	-	-	-	-	-	-	-	1
11	1	-	-	-	-	-	-	-	-	-	-	-	1
12	1	-	-	-	-	-	-	-	-	-	-	-	1
13	-	1	-	-	-	-	-	-	-	-	-	-	1
Σ	15	13	12	10	5	4	2	2	2	1	1	1	68

On the basis of an urn model it has been shown that the marginal rank-frequency distribution of Dutch translation means follows the negative binomial distribution as presented in Table 8 as well. All results corroborate our diversification model.

Now, if both marginal distributions are negative binomial then the distribution in the table must be negative binomial as well.

Though the numbers in the tables are too small for reasonable testing we can at least use them to demonstrate the model. The bivariate displaced negative binomial distribution has the probability function

$$P_{x,y} = \frac{(r+x+y-3)!}{(x-1)!\,(y-1)!\,(r-1)!}\, p^r q_1^{x-1} q_2^{y-1} \qquad x,y = 1,2,\ldots \qquad (11)$$

where $0 < p < 1$; $p+q_1+q_2=1$: A quite easy estimation is that using frequency classes $P_{1,1}$ $P_{1,2}$ $P_{2,1}$ (for other estimations see Bates, Neyman 1952; Sibuya, Yoshimura, Shimizu 1964). Using the fact that

$$P_{1,1} = p^r$$

$$P_{1,2} = rp^r q_2$$

$$P_{2,1} = rp^r q_1$$

we compute p iteratively from

$$\hat{p}_{i+1} = \hat{p}_i - \frac{\hat{p}_i + \frac{(\hat{P}_{1,2}+\hat{P}_{2,1})\ln \hat{p}_i - 1}{\hat{P}_{1,1}\ln \hat{P}_{1,1}}}{\frac{1}{\hat{p}_i}\cdot\frac{\hat{P}_{1,2}+\hat{P}_{2,1}}{\hat{P}_{1,1}\ln \hat{P}_{1,1}} + 1} \qquad (12)$$

where $\hat{P}_{1,1} = n_{11}/n$ etc. and

$$\hat{q}_2 = \frac{\hat{P}_{1,2}(1-\hat{p})}{\hat{P}_{1,2}+\hat{P}_{2,1}}$$

$$\hat{q}_1 = 1 - \hat{p} - \hat{q}_2 \qquad (13)$$

$$\hat{r} = \frac{\ln \hat{P}_{1,1}}{\ln \hat{p}}$$

Table 8: Rank-frequency distribution of Dutch translation means

el-			föl-		ke-	
y	n_y	NP_y	n_y	NP_y	n_y	NP_y
1	38	38.00	16	16.00	15	15.00
2	19	20.46	12	11.28	13	14.19
3	17	13.19	9	7.80	12	11.26
4	8	8.97	5	5.36	10	8.36
5	7	6.27	2	3.67	5	5.99
6	3	4.44	2	2.51	4	4.21
7	2	3.18	1	1.71	2	2.91
8	2	2.29	1	1.17	2	1.99
9	1	1.66	1	0.80	2	1.35
10	1	1.21	1	0.54	1	0.91
11	1	0.88	1	0.37	1	0.61
12	1	0.65	1	0.79	1	1.21
13	1	0.47				
14	1	0.35				
15	1	0.98				
p=0.2486 r=0.7165 χ^2_9 =3.57 P=0.94			p=0.3214 r=1.0385 χ^2_7 =2.38 P=0.94		p=0.3595 r=1.4773 χ^2_8 =1.43 P=0.99	

The recursion formulas for the computation of probabilities are

$$P_{1,1} = p^r$$

$$P_{x,y+1} = \frac{(r+x+y-2)\ q_2 P_{x,y}}{y} \qquad (14)$$

$$P_{x+1,y} = \frac{(r+x+y-2)\ q_1 P_{x,y}}{x} \qquad (15)$$

In order to use a chi square test for fitting, one must pool a number of cells; but that cannot be done systematically here.

3. The tables can be further analyzed and interpreted by means of information theoretical measures such as were introduced in the previous papers.
We use the following symbols:

X - meanings of Hungarian prefixes
Y - Dutch translation means
n_{ij} - frequencies of x_i translated as y_j (numbers in the tables)
n_i - marginal sums (last columns of the tables)
n_j - marginal sums (last rows of the tables)
n - total sum $= \sum_{ij} n_{ij}$

$p_{ij} = n_{ij}/n$; $p_i = n_i/n$; $p_j = n_j/n$

and define the following measures:

(i) $H(X) = -\sum_i p_i \mathrm{ld}\ p_i$ - meaning uncertainty of a prefix; a measure of diversification

(ii) $H(Y) = -\sum_j p_j \mathrm{ld}\ p_j$ - Dutch translation uncertainty

(iii) $H(X,Y) = -\sum_i \sum_j p_{ij} \mathrm{ld}\ p_{ij}$ - uncertainty of the system as a whole

(iv) $H(Y|X) = -\sum_i \sum_j p_{ij} \mathrm{ld}(p_{ij}/p_i)$ - conditional translation uncertainty given the prefix meaning

(v) $H(X|Y) = -\sum_i \sum_j p_{ij} \mathrm{ld}(p_{ij}/p_j)$ - conditional uncertainty of the prefix meaning given the translation

(vi) $I(X;Y) = \sum_i \sum_j p_{ij} \mathrm{ld}\left[p_{ij}/(p_i p_j)\right]$ - information transmitted; mutual information

The interpretation of these measures can be found in Beöthy, Altmann (1984a). The results of computation are presented in Table 9.

Table 9: Information theoretical measures

Measure	el-	föl-	be-
H(X)	1.1874	3.1717	3.0693
H(Y)	2.7833	2.7935	3.0204
H(X,Y)	3.6380	4.8182	5.0341
H(X|Y)	0.8547	2.0247	2.0137
H(Y|X)	2.4507	1.6466	1.9649
I(X;Y)	0.3326	1.1470	1.0556

In order to make these measures comparable we relate them to their maximum values attainable in the given context and define

$$H_{rel}(X) = \frac{H(X)}{\mathrm{ld}\ k}$$ k being the number of meanings of a prefix (x_{max}) (2)

$$H_{rel}(Y) = \frac{H(Y)}{\mathrm{ld}\ m}$$ m being the number of translation means used (y_{max}) (3)

$$H_{rel}(X,Y) = \frac{H(X,Y)}{H(X)+H(Y)} \qquad (4)$$

$$H_{rel}(X|Y) = \frac{H(X|Y)}{H(X)} \qquad (5)$$

$$H_{rel}(Y|X) = \frac{H(Y|X)}{H(Y)} \qquad (6)$$

$$I_{rel}(X;Y) = \sqrt{\frac{2I(X;Y)}{H(X)+H(Y)}} = \sqrt{2\left[1 - \frac{H(X,Y)}{H(X)+H(Y)}\right]} \qquad (7)$$

The last measure was introduced by Astola, Virtanen (1983) and called the entropy correlation coefficient.

The results of computations for all five of the examined Hungarian verbal prefixes as translated into Dutch are presented in Table 10.

Table 10: Relative information theoretical measures

Measure	meg-	ki-	el-	föl-	be-
$H_{rel}(X)$	0.4614	0.9171	0.3746	0.9548	0.8294
$H_{rel}(Y)$	0.7375	0.7526	0.7124	0.7792	0.8425
$H_{rel}(X,Y)$	0.9289	0.8313	0.9162	0.8077	0.8267
$H_{rel}(X\|Y)$	0.7887	0.7044	0.7198	0.6384	0.6561
$H_{rel}(Y\|X)$	0.8928	0.6069	0.8805	0.5894	0.6505
$I_{rel}(X;Y)$	0.3772	0.5809	0.4093	0.6201	0.5888

$H_{rel}(X)$ is the measure of semantic diversification. The greater its value the greater the homogeneity of the frequencies of individual meanings i.e. the more effaced is the original meaning. Usually the homogeneity is tested by means of a chi square test or the equivalent minimum discrimination information statistics

$$2I = 2\sum_{i=1}^{k} n_i \ln \frac{n_i}{E(n_i)} \qquad (8)$$

Since $E(n_i) = n/k$ it can be seen that

$$2I = \frac{2n}{\mathrm{ld}\ e}\left[H_o - H(X)\right] \qquad (9)$$

where $H_o = \mathrm{ld}\ k$, which is distributed approximately as a χ^2 with k-1 degrees of freedom so that the distribution of H(X) can easily be gained from the χ^2-distribution by means of the transformation

A further measure of semantic diversification was proposed in the form of the relative repeat rate

$$R_{rel} = \frac{1 - R}{1 - 1/k} \qquad (10)$$

where $$R = \sum_i p_i^2 = \frac{1}{n^2}\sum n_i^2$$

yielding for the prefixes the results presented in table 11.

Table 11: Relative repeat rates

	R	R_{rel}
föl-	0.1228	0.9747
ki-	0.1290	0.9581
be-	0.1579	0.9123
meg-	0.5810	0.4714
el-	0.6590	0.3836

The greater R_{rel} (= the smaller R) the greater the diversification. Again, it is possible to derive the moments of R using the multinomial distribution or, alternatively, to compute R directly from the negative binomial distribution.

References

Astola, J., Virtanen, I., A measure of overall statistical dependence based on the entropy concept. Proceedings of the University of Vaasa Nr. 91, 1983

Bates, G.E., Neyman, J., Contributions to the theory of accident proneness. University of California Publications in Statistics 1, 1952, 215-275

Beöthy, E., Altmann, G., Semantic diversification of Hungarian verbal prefixes. I."meg-", Nyelvtudományi közlemények (in press 1984 a)

Beöthy, E., Altmann, G., Semantic diversification of Hungarian verbal prefixes. II. "ki-, Finnish-Ugrische Mitteilungen (in press 1984 b)

Sibuya, M., Yoshimura, I., Shimizu, R., Negative multinomial distribution. Annals of the Institute of Statistical Mathematics Tokyo 16, 1964, 409-426

Zipf, G.K., Human behavior and the principle of least effort. New York, Hafner 1972 (2)

Einige formale Charakteristiken in Gesprächen mit mehreren Sprechern

R. Drewek, Zürich

This paper proposes a formal description of group conversation based on sets of recurrent patterns in turntaking. Every speaker of such a group has his own profile of communicative behaviour, which can be expressed in quantities of verbal participation and which might be indicative for the underlying social processes. A computer program helps to display and count those properties of a dialogue, for which an operational definition can be found. A brief sketch of ideas at the end shows the way how formal (and content oriented) approaches could enlarge the scope of conversational analysis.

1. Vorbemerkungen

Das subjektiv erfahrene und in introspektiver Rückschau rekonstruierte Gespräch bietet sich, vor allem im unmittelbaren *Danach* oft als verwirrende Vielfalt der Erinnerung an: wir wissen vielleicht

- mit wem wir uns über dieses oder jenes gestritten haben,
- daß man sich nicht verständlich machen konnte,
- wer wieder einmal zu lange auf die anderen eingeredet hat,
- daß man diesmal garnicht richtig zu Wort gekommen ist.

Vielfach bleibt ein summativer Eindruck *es sei wieder viel* gelaufen oder eben auch *eigentlich nichts*.

2. Theorieskizze

Der wissenschaftlichen Analyse erschließt sich diese kurze Vorstellung kommunikativer Phänomene auf vielerlei Ebenen; häufig wird folgende Einteilung (WATZLAWICK, 1974) verwendet:

A Inhaltsebene <-> B Beziehungsebene

Ein Linguist wird insbesondere die Ebene des (kommunizierten) Codes C im weitesten Sinn betonen, weil eine Betrachtung von A ihn an den sozio-psychologischen Rand seines Paradigmas (KUHN, 1974) und B ihn

sogar an Welt und Wirklichkeit stoßen würden. So wird er eingedenk seines Vorwissens um A und B die Strukturen *in* (?) C hermeneutisch herbeideuten und zu Theorien, wie z.B. einer Sprechakttheorie, verdichten. Diese Sachlage möchten wir in einem ersten Postulat verarbeiten:

> I. Im Code eines abgeschlossenen Gesprächs zeigen sich nicht nur Strukturen der verwendeten Sprache, er trägt zugleich Information über den abgelaufenen sozialen Prozess und die darin verknüpften psychischen Vorgänge.

Dazu sind einige erläuternde Sätze nötig:

I.1 Ein Gespräch sei abgeschlossen, wenn die räumliche und/oder zeitliche Koinzidenz aller beteiligten Personen aufgelöst ist

I.2 Der Code eines abgeschlossenen Gesprächs besteht aus der Gesamtheit der durch einen Beobachter (dies können mechanische Beobachter, z.B. Audio- und Videogeräte, aber auch teilnehmende menschliche Beobachter sein) in sprachlichen Symbolen notierbaren Geschehnisse. Codierbar sind folgende Dimensionen:

- die verbale und paraverbale Dimension:
  Wörter, Sprecherwechsel, Betonungen, Sätze, Äußerungen
- die nonverbale Dimension:
  Blickkontakte, Sitzpositionen, Gesten und Mimik
- die Ereignis- bzw. Handlungsdimension:
  Störungen durch Herabfallen eines Leuchters, Eintreten einer unbekannten Person, klimatische Veränderungen etc.
- die physikalische Dimension:
  die ablaufende Zeit, die Ausstattung des Gesprächsortes

I.3 Bei der Notierung eines Gesprächs entstehen verschiedene Verzerrungen mit unterschiedlichen Folgen für eine rationale Rekonstruktion.

I.3.1 Die Mannigfaltigkeit eines Systems von Notationssymbolen bestimmt die Selektion der Phänomene des realen Gesprächs (reduction bias).

I.3.2 Die Kompliziertheit der Analyse wächst mit der Komplexität der verwendeten Notation (artefactual bias).

I.3.3 Der Zustand des Beobachters und seine individuellen Wahrnehmungsgewohnheiten überführen die Totalität des realen Gesprächs in die Selektivität seines Protokolls (personal bias)

I.4 Die Zeit eines Gesprächs läßt sich in allen unter I.2 genannten Dimensionen ausdrücken, als physikalische Zeit in Sekunden, als Textzeit in der Anzahl abgelaufener verbaler Einheiten (Wörter, Äußerungen).

I.4.1 Mit der Verkleinerung des Zeitrasters wächst der Aufwand des aufzuschreibenden Materials, eine Vergröberung des Rasters hat den entgegengesetzten Effekt.

I.4.2 Die individuell im Gespräch von den Beteiligten empfundene Zeit kann, muß sich aber nicht im Zeitraster widerspiegeln.

I.5 Generell können zwei Typen der Notation nach dem Verhältnis zur Zeitdimension des realen Gesprächs unterschieden werden:
- die entzerrte lineare Kodierung (Nacheinander)
- die flächig parallelisierte Kodierung (Miteinander)

I.5.1 Die Entzerrung besteht darin, daß das ursprünglich simultan ablaufende Geschehen in eine Abfolge z.B. von Sprechern und ihren Äußerungen transformiert wird. (siehe Abbildung 1) Eine gewisse Inkonsequenz dieses Vorgehens liegt in der oft verwendeten geklammerten Markierung paralleler (Sprech-) Vorgänge.

```
Sprecher 1 :   ich bin dumm.
Sprecher 2 :   mmh.              ((es donnert))
      .            .
      .            .
```

Abbildung 1: lineare Notation eines Gesprächs

I.5.2 Die parallele Kodierung versucht vor allem kookkurente Vorgänge sichtbar festzuhalten, unterteilt aber das Geschehen nach dem Kanalmodell. (siehe Abbildung 2)

	Sprecher 1		Sprecher 2		Sprecher 3		
	Kanal 123456	Text	Kanal 123456	Text	Kanal 123456	Text	Ereignisse
Zeit							
0.0							
0.2		ich					
0.4	r	bin					
0.6	r	dumm	k	mmh.			es
0.8			k				donnert
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Abbildung 2: Partiturnotation eines Gesprächs

(Hier sind auf einem Zeitraster von 0,8 sec drei Sprecher nicht nur mit ihrem Text - Sprecher 3 schweigt - notiert, sondern parallel dazu pro Sprecher sechs Kanäle. Einem Kanal wären bestimmte Dimensionen

des nonverbalen Verhaltens (Gestik, Mimik, Körperhaltung usf.) zugeteilt und Vorkommnisse, wie das Runzeln der Stirn (Kanal 4: 'r') oder Kopfschütteln (Kanal 6: 'k') mit Symbolen eingetragen.)

I.5.3 Die Notierung eines Gesprächs kann auch als ein Übergang vom analogen Geschehen in eine Abfolge von digitalen Symbolen für Ereignisse (mit bestimmbarer Bedeutung) aufgefaßt werden

I.6 Gespräche sind zielgerichtet. Nach einem Gespräch haben sich die Beziehungen der Partner verändert: ihr (Alltags-) Wissen hat sich erweitert und ihre Einstellungen zueinander sind anders geworden.

I.6.1 Die Intentionalität von Gesprächen gibt ihnen den Charakter einer Handlung im sozialen Raum.

Aus dem Gesagten läßt sich ein zweites, eher forschungsorientiert programmatisches Postulat zusammenstellen:

> II. Die linguistische Gesprächsanalyse kann nur dann sinnvoll - und somit legitim - sein, wenn sie jeden Zug des Gesprächs im Gefüge der sozialen Positionen, der psychischen Zustände und des unter diesen Randbedingungen verwendeten Codes in seiner Intentionalität hinreichend evident erklären kann.

Auch das zweite Postulat ist durch einige Sätze zu erläutern:

II.1 Aus der sprachlichen Grammatik müßte eine Handlungsgrammatik hervorgehen und umgekehrt.
Der Fragesatz *ist* manchmal eine Frage von jemandem an jemanden unter bestimmten Bedingungen...

II.2 Das Postulat II gefährdet das system-linguistische Paradigma.

II.3 Die Reichweite einer linguistischen Analyse ist jedoch durch den Einfluß der Verzerrungen (vgl. I.3) eingeschränkt.

II.4 Einen Zug eines Gesprächs erklären heißt, ihn im Lichte möglicher Abhängigkeit text-innewohnender und kontextueller Variablen darstellen.

3. Erste Schritte

Nach diesem eher skizzenhaften Theorierahmen folgt der Schritt zur Empirisierung, die Einengung auf das Machbare. Ausgangspunkt bleibt die Frage, wie kann der Prozeß des Gesprächs, notiert nach möglichst

unaufwendigen Konventionen, in seiner Analyse erste Hinweise auf die zugrundeliegenden sozialen Prozesse abgeben. Welche sprachlichen Mechanismen verweisen auf die Gesprächsgruppenstruktur? Wie können Indikatoren hierfür operationalisiert werden?

Die älteste und bekannteste Form, Gespräch und Handlung zu notieren ist in der literarischen Gattung des Dramas tradiert. Auch die Gedanken der Philosophie Platos und seiner Schüler wurden in der Notierung von Gesprächen (Dialogen) überliefert. Sicher liegt der Zweck der literarischen Gattung Drama nicht gerade darin, Sprach- und Kommunikationsforschern als Material zu dienen. Dennoch bietet die Aufschreibkonvention dieser Gattung eine Form, wir haben sie unter I.5 als entzerrt lineare charakterisiert, welche die Bedingung der Einfachheit einer Notation sicher erfüllt. Unter der von Arno Holz geprägten programmatischen Wendung "Kunst = Natur - X" (HOLZ, 1891) könnte vor allem das Drama im Naturalismus unseren Anforderungen sehr nahe kommen.

Nehmen wir einmal an, ein Gespräch sei von einem Tonband in der Form eines Dramas aufgeschrieben worden. Um zu einer groben Analyse der Gesprächsstruktur zu gelangen, setzen wir zunächst folgende operationale Definitionen:

D 1 *Äußerung* ist der notierte Text eines Sprechers, der bis zu der Markierung des Sprechens einer anderen Person reicht.

D 2 Ein *Monolog* ist die Äußerung eines Sprechers, die in ihrer Ausdehnung einen Grenzwert überschreitet, der bezogen auf das gesamte Gespräch post hoc empirisch bestimmt wurde.

D 3 *Dialog* ist die ununterbrochene Abfolge der Äußerungen zweier Sprecher.

D 4 *n-log* ist eine kontinuierliche Äußerungsfolge von n Sprechern in einem rekursiven Raster von mindestens (2 * n- 1) Äußerungen. Die Sprecherfolge kann dabei innerhalb von n aufeinanderfolgenden Äußerungen frei permutieren.

D 5 *Sequenz* S(i) ist eine ununterbrochene, als n-log analysierte Äußerungskette.

D 6 *Ordnung*: ein n-log hat die Ordnung n, also die Anzahl in ihm konstant beteiligter Sprecher.

D 7 *Symmetrie* ist die Eigenschaft eines n-logs, wenn alle beteiligten Sprecher dieselbe Anzahl von Äußerungen haben.

D 8 *Overlap* wird die Verschachtelung zweier n-loge mit n > 1 genannt, bei der Sprecher im letzten Raster des vorhergehenden n-logs ins erste Raster des folgenden überwechseln.

D 9 Als *Adressat* einer Äußerung A(i) wird nur der Sprecher des jeweils unmittelbar folgenden Redebeitrags A(i+1) bezeichnet.

D 10 Eine *Phase* ist jede Abfolge von Sequenzen S(i)... S(i+j), in welcher mindestens ein Sprecher konstant an jeder Sequenz beteiligt ist.

D 11 Den *Kern* einer Phase bilden alle durchgehend an ihr beteiligten Sprecher.

4. Kommentar zu den Definitionen

Unter dem Gebot der Operationalisierung und der Einfachheit stützen sich alle Definitionen nur auf unmittelbar aus der Notierung des Gesprächs hervorgehenden Kriterien:

⟶ den gesprochenen Text, ⟶ seine Länge,
⟶ die Sprechermarkierungen, ⟶ ihre Abfolge.

Damit werden qualitativ alle konstitutiven Elemente der linear entzerrten Codierung ausgenutzt, mit anderen Worten: die "Voranalyse" dessen, der das Gespräch notiert hat, wird als (hermeneutische) Leistung miteinbezogen, um strukturelle Information zu gewinnen. Die inhaltliche Seite des Textes wird bewußt ausgeklammert; in einer parallelen Untersuchung wäre sie das Ziel inhaltsanalytischer Forschungsstrategien, welche die hier behandelte formale Organisation notwendig ergänzen muß. Denkbar ist zum Beispiel eine Analyse der referentiell-thematischen Verknüpfung von Äußerungen, die eine andere linguistische Ebene des Gesprächs aufdeckt und damit auch zu anderen Definitionen der oben genannten Begriffe führt.

Zu den einzelnen Definitionen:

In D 1 steht der Begriff *Text* im engeren Sinne sicher einmal für die verbale Dimension. Nicht auszuschließen wäre der Einbezug des Nonverbalen als 'Text', ebenso wie der paraverbalen Phänomene.
Die Abfolge der Sprechermarkierung wird in D 2 bis D 4 als ein mengentheoretisch überprüfbares Ordnungsgefüge interpretiert. Dabei wird die definitorische Strenge in zwei Punkten durchbrochen:

- Der *Monolog* (1-log, Sequenz 1. Ordnung) muß die Länge der Äußerung relativ zu den sonstigen Äußerungen desselben Sprechers heranziehen. Dies hat den Vorteil, daß monologische Partien, die in n-loge mit n> 1 sozusagen eingebettet sind, erkennbar werden. Es hat den Nachteil, daß eine *Monologschwelle* empirisch bestimmt werden muß, von der ab jeweils das Prädikat Monolog plausibel erscheint.

- Beim 0-log (Nullsequenz) handelt es sich nicht um ein Gespräch, das mangels Sprechern nicht stattgefunden hat, sondern um eine Restmenge von Äußerungsfolgen, die per definitionem keine Ordnung n> 1 aufweisen können, - meist infolge des Fehlens eines rekursiven Rasters von n> 1 Äußerungen. Einbettung von Monologen, Symmetrie D 7 und Overlap D 8 sind im gewissen Sinne Artefakte der formalen Analyse, deren Interpretabilität als Indikator der Gesprächsstruktur fraglich ist.

Die Definition des Adressaten D 9 stellt in ihrer - wenn auch verblüffenden Simplizität - einen aus theoretischer Sicht problematischen Fall dar: der Sprachgebrauch von 'Adressat' suggeriert eine inhaltliche Bestimmung der Äußerung eines Sprechers, sozusagen eine kommunikative Richtung. Dies kann aber in diesem formalen Ansatz nicht hinreichend überprüft werden. Andererseits ist es paradox anzunehmen, daß eine Folgeäußerung nicht in irgendeiner Weise mit der vorhergehenden verknüpft ist. Das Raster unserer Analyse fokussiert darum nicht die Art dieser Verknüpfung, sondern benutzt lediglich deren (unterstelltes) Vorhandensein. Adressaten sind nun mitunter nicht nur einzelne Gesprächspartner, sondern oft auch weitere Zuhörer, manchmal ein ganzes Publikum. Dieser Aspekt wird vorläufig ausgeklammert.

Sequenz D 5 ist lediglich ein paraphrasierender Begriff zu n-log.

Die Analyse in Phasen dient der Sichtbarmachung größerer Zusammenhänge als derjenigen der Sequenzen. Der (Sprecher-)Kern gibt dabei vielleicht erste Hinweise für eine Interpretation des sozialen Gefüges.

5. Algorithmus

Die Anwendung der oben definierten Begriffe wurde in Form eines Computerprogramms namens DMA realisiert. Es soll transkribierte Gespräche in ihrer Codierung durcharbeiten und die intendierten Struk-

turen sichtbar machen. DMA ist Teil der LDVLIB-Software, eines Programmpaketes für die computergestützte Textanalyse, das vom Autor am Rechenzentrum der Universität Zürich implementiert wurde. Im Rahmen des LDVLIB-Systems wurden auch die Textcodierungsregeln für Texte gesprochener Sprache festgelegt, um eine Normierung in Bezug auf maschinelle Lesbarkeit zu erreichen (DREWEK, 1975).

Wenn Teile einer Theorie zu einem computer-ausführbaren Programm geronnen sind, ist es wiederum möglich, experimentell Daten zu gewinnen, die Gesprächsanalysen im Rahmen der eingangs vorgestellten Theoreme ermöglichen. Doch zunächst zur Funktionsweise des Algorithmus selbst.

DMA läuft in drei Stufen ab:

Ablaufanalyse > Sequenzanalyse > Phasenanalyse

Die *Ablaufanalyse* berechnet die Länge jeder Äußerung (in Anzahl der Wörter) und notiert die beteiligten Sprecher. Es werden ein Ablaufdiagramm (siehe Abb.3), eine statistische Sprecherdeskription (siehe Abb.4) und eine Matrix der Sprecherfolge (siehe Abb.5) ausgedruckt.

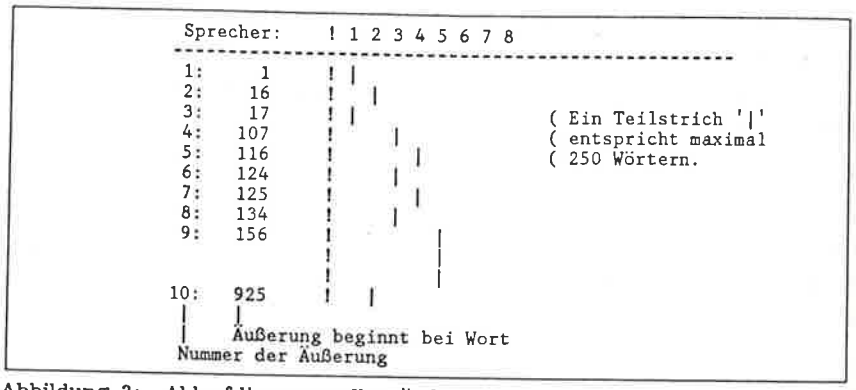


Abbildung 3: Ablaufdiagramm aller Äußerungen

Die erste Äußerung ist 15 Wörter lang, sie wurde von Sprecher-1 getätigt (die Sprecher werden automatisch in der Reihenfolge ihres

Auftretens numeriert). Die zweite Äußerung (Sprecher-2) besteht aus Wort 16 des Textes. (Mit der fortlaufenden Wortnummer können Äußerungsgrenzen im Textausdruck des LDVLIB-Programms IOZ relokalisiert werden). Ein Signaturstrich "|" steht hier für Längen bis zu 250 Wörtern, jedoch kann dieser Maßstab verändert werden. Die 9. Äußerung, die bei Wort 156 beginnt, ist etwas länger und daher durch drei Signaturen gekennzeichnet. Auf diese Art werden monologverdächtige Passagen schon im Ablauf sichtbar.

Äußerungen von oder an den Sprecher	Relativer Anteil an allen Äußerungen des Textes	Anzahl der vom Sprecher geäußerten Wörter	Anteil des Sprechers am Gesamttext	Mittlere Äußerungslänge	Standardabweichung	Name
59	23.37	260	126.0	44.1	50.28	Betty Bossi
16	6.32	20	0.2	1.3	0.45	Janine D.
53	20.95	1047	10.2	19.8	26.38	R. Schlawinsky
15	5.93	755	7.5	50.3	151.30	E. Steinberger

Abbildung 4: Sprecherkennwerte

Die für jeden einzelnen Sprecher berechneten Werte können der Typisierung der Gesprächsteilnehmer dienen, z.B. in Lang- und Kurzsprecher eingeschätzt nach ihrer mittleren Äußerungslänge; oder in Haupt- und Nebenfiguren je nach Anzahl ihrer Äußerungen und ihrem Textanteil am gesamten gesprochenen Text.

	Nachfolger: SP(n) S1	S2	S3	S4	n
Vorgänger: SP(v) S1	-	22	11	17	50
	-	44.0	22.0	34.0	
S2	57	-	9	3	69
	82.6	-	13.0	4.3	
S3	14	7	-	0	21
	66.7	33.3	-	0.0	
S4	89	9	0	-	98
	90.8	9.2	0.0	-	

absolute Werte : Anzahl Äußerungen,
Prozentwerte : Anteil aller Äußerungen im Text

Abbildung 5: Matrix der Sprecherabfolge MSF(v,n)

Diese Matrix ist so zu lesen, daß MSF(v,n) Prozent der Äußerungen von Sprecher SP(v) (in Zeile v) von Äußerungen des Sprechers SP(n) (in Spalte n) gefolgt wurden. Oder: der Adressat nach D 9 von Sprecher v war in MSF(v,n) Prozent aller Fälle der Sprecher SP(n).

Bei der *Sequenzanalyse* wird der Ablauf des Gesprächs mit einem geeigneten Algorithmus nach rekurrenten Mustern der Sprecherfolge abgesucht. Angenommen, die Abfolge der Äußerungen dreier Sprecher A, B und C sei wie folgt:

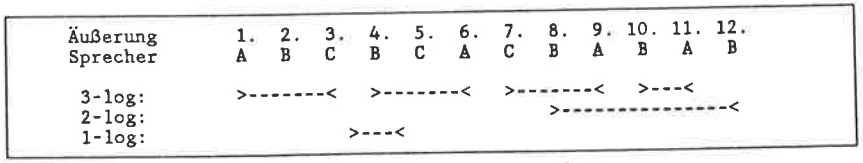


Abbildung 6: Klassifikation mehrdeutiger Sprecherfolge

Nach der Definition eines n-logs D 4 können hier drei Sequenzen unterschiedlicher Ordnung analysiert werden:
Die ersten zwölf Äußerungen zeigen variierende Muster der Abfolge aller drei Sprecher, das letzte 3er-Raster ist allerdings unvollständig, weil der Sprecher C fehlt. Hier liegt eine typisch asymmetrische Sequenz 3. Ordnung vor. Ab der achten Äußerung ist ein

rekurrentes Raster mit den zwei Sprechern A und B erkennbar, das aber in die vorangehende Folge 'eingehängt' erscheint. Dies ist ein Beispiel für einen Overlap (Definition D 8) zwischen einem Dialog und einem 3-log, der im Gespräch selbst interessante Interaktionsstrukturen zeigen dürfte (Hypothese), weil es sich um einen Übergang von einer komplexeren zu einer einfacheren Struktur handelt.

Angenommen, die vierte Äußerung der obigen Folge habe eine m-mal über dem Durchschnitt der Äußerungslängen des Sprechers B liegende Ausdehnung, dann wäre das Ergebnis der Sequenzanalyse ein eingebetteter Monolog (Sequenz 1. Ordnung). "m" ist hierbei die sogenannte Monologschwelle, die empirisch für den jeweiligen Dialog bestimmt wird.

Die Ergebnisse der Sequenzanalyse werden in einer Tabelle (siehe Abb.7) in dem Text-Sequenz-Muster (siehe Abb. 8) dargestellt.

Für jede analysierte Sequenz ist angegeben, welchen Ordnungsgrad sie hat, ob sie symmetrisch (Wert 1) ist oder nicht (Wert O), bei welchem Wort im Text sie beginnt und endet. Dann folgen drei Längenangaben, einmal in absoluter Wortzahl, dann aber relativ zur Gesamtzahl der Wörter im Text und schließlich gemessen in Anzahl der Äußerungen. Kommt ein Overlap vor, so wird seine Länge ebenfalls in Anzahl der Wörter ausgedruckt. Am Ende der Tabelle findet man die Codenummern der an der Sequenz beteiligten Sprecher. Eingebettete Monologe (vgl. Sequenz 6) werden mit einem Stern '*' bei der relativen Länge gekennzeichnet.

Nummer der Sequenz	Ordnung (n-log)	Symmetrie	Start Wort	Äußerung	Sequenzlänge in Worten	in Prozent des Textes	in Anzahl Äußerungen	Überlappung	Beteiligte Sprecher
1	0	0	1	1	107	1.07	3	0	01 02
2	2	0	108	4	49	0.49	5	0	03 04
3	0	0	157	9	88	0.88	5	0	05 01 03 04
4	2	0	245	14	59	0.35	5	0	03 05
5	2	0	278	18	734	7.32	11	26	03 04
6	1	1	304	19	596	5.96*	1	596	04
7	2	0	1012	29	228	2.28	9	0	01 04
8	0	0	1240	38	53	0.53	1	0	03

Abbildung 7: Sequenzanalyse

Das Text-Sequenz-Muster bietet eine Synopsis des gesamten Gesprächs, dessen gesamte Länge (in Wörtern) als 100% dargestellt wird. Die parallel verlaufenden Ketten von Pluszeichen ('+') repräsentieren die analysierten Sequenzen n-ter Ordnung mit n> 1; Punkte erscheinen als Signatur für Nullfolgen. Dieses Muster kann Ausgangspunkt für

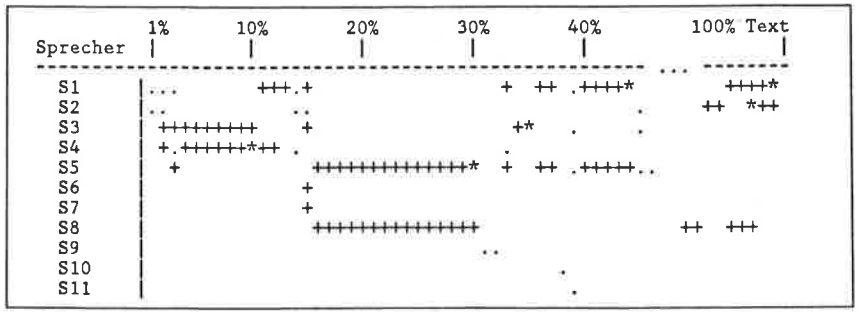


Abbildung 8: Text-Sequenz-Muster

einen typologischen Vergleich dieses Gesprächs mit anderen sein. Dabei kann das Eintreten verschiedener Sprecher zu verschiedenen Zeitpunkten, die Verteilung stabiler (analysierter) Sequenzen und instabiler (Nullfolgen) erste Hinweise für Interaktionsmuster geben.

Nebenher ergibt die Sequenzanalyse Aufschlüsse über die statistische Verteilung der Seqenzen. Aus Abbildung 9 ist ersichtlich, daß vor allem Nullfolgen und dialogische Partien dominieren. Der ähnlich große Anteil von Monologen ist von der Höhe des jeweils gesetzten Schwellwerts abhängig.

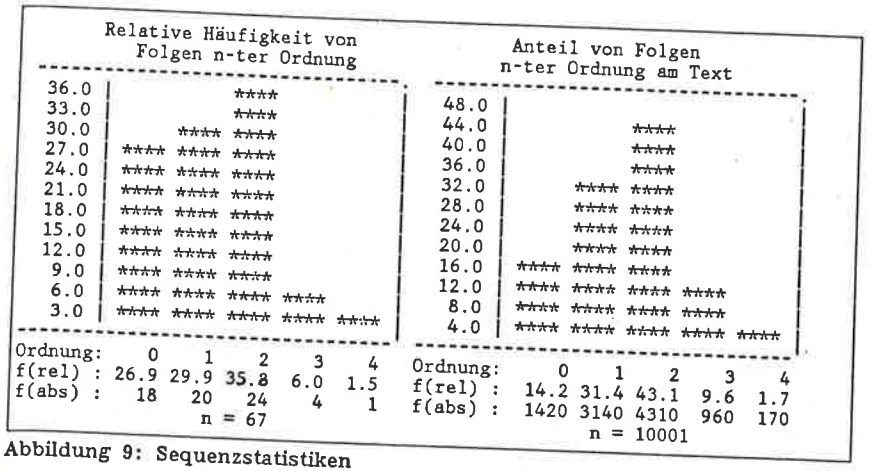


Abbildung 9: Sequenzstatistiken

Die vorläufig letzte Programmstufe ist die *Phasenanalyse*. Sinn der vorhergehenden Resultate war es, eher die Struktur des Gesprächsablaufs zu erhellen. Die Phasenanalyse soll hingegen Aufschluß über die Rolle der Sprecher geben, so wie sie sich aus den vorhergehenden Schritten ableiten lassen. Dabei geht man von dem Gedanken aus, daß ein aktiver Sprecher einen *Kanal* herstellt, - der informationstechnische Begriff 'Kanal' wäre unter interaktionsbezogenem Blickwinkel eher in *Konnex* umzumünzen, den er in unterschiedlichem Maße benutzt. Ein Sprecher, der während des gesamten Gesprächs einen Konnex zu beliebigen anderen Sprechern aufrechterhält, konstituiert sicher eine andere soziale Rolle, als jemand, der sich mit einer einzelnen Äußerung begnügt.

Das so entworfene Bild kann täuschen: es berücksichtigt nur den Umfang der kommunikativen Verbindungen, klammert aber soziale Faktoren (Status, Prestige, etc.) aus. Eine Prüfungssituation, in welcher der Kandidat dreißig Minuten redet und der statushöhere Dozent am Schluß nur *durchgefallen* sagt, mag das vorab Gesagte relativieren. Uns geht es jedoch nicht um eine direkte Ableitbarkeit von sozialen Faktoren aus der Phasenanalyse. Ihre Einschränkung liegt in ihrem Charakter als Indikator: sie kann zugrundeliegende soziale Strukturen (Rollen) zeigen, sollte aber bei weiteren Forschungsoperationen durch

andere Variablen gestützt werden.

Oft wird die Struktur einer Gruppe durch ein sogenanntes *Soziogramm* dargestellt. In der Phasenanalyse wird diese Darstellungsart auf die Gesprächssituation übertragen: im Zentrum stehen Sprecher, die als Kern definiert wurden (D 11). Angelagert werden all diejenigen Sprecher, die mit dem Kern mindestens einen Konnex in einer Sequenz aufweisen konnte. Die Nähe oder Ferne dieser Sprecher zum Kern wird durch ihre Konnexdichte gewichtet. Dazu kann die Anzahl unmittelbar mit dem Sprecherkern abgewickelter Äußerungen oder auch deren Gesamtumfang als Maß definiert werden.

Alle Daten hierzu sind bereits in der Sequenzanalyse berechnet. In der gegenwärtigen Programmversion von DMA ist aus darstellungstechnischen Gründen diese Stufe noch nicht programmiert. Wir stellen uns das Ergebnis der Phasenanalyse wie folgt vor:

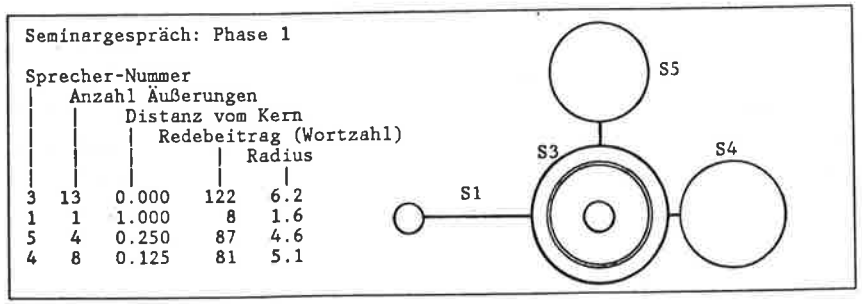


Abbildung 10: Phasenanalyse

6. Vorausblick

Die formale Seite von Gesprächen kann in der oben beschriebenen Weise nun mit Unterstützung eines Computerprogramms berechnet werden, weil die Definitionen aus unserer Theorieskizze operational tragbar waren. Der logisch folgende Schritt wird nun sein, möglichst viel Material auf diese Art zu untersuchen, um herauszufinden, ob der Beschreibungsalgorithmus Anforderungen in zwei Punkten, - wovon der eine eher linguistisch, der andere eher sozialpsychologischer Natur ist, standhält.

1. Kann gezeigt werden, daß verschiedene Typen von Gespräch (Interview, Verhör, Gerichtsverhandlung, Psychoanalyse, Seminar, Stammtischrunde, Dramen, Talkshow, Debatte...) signifikante Unterschiede in ihrer Struktur aufweisen, die durch den Algorithmus hinreichend beschrieben werden ?

2. Wie sind die gewonnenen sprachlichen Indikatoren in der sozialen Dimension des Gesprächs sowohl theoretisch, wie auch empirisch validierbar ?

Ein letzter Aspekt betrifft die Methode selbst. Nach einer Überprüfung am Material von Gesprächen selbst fragt es sich, wieweit und mit welchem (vertretbaren) Aufwand an Operationalisierung der bestehende Algorithmus ausgebaut werden kann. Hier muß auch diskutiert werden, welche Unterstützung der Computer bei der inhaltlichen Erschließung eines Gesprächs leisten könnte, - vielleicht im Rahmen der *content analysis* oder mit Methoden der *artificial intelligence*. Es existieren Programme (LDVLIB: KWI, VPS), mit denen die Vokabulare einzelner Sprecher in beliebigen Textteilen greifbar und auswertbar sind. So kann zum Beispiel das gemeinsame Vokabular zweier Sprecher bestimmt werden oder ihr Vokabularzuwachs in bestimmten Gesprächsperioden. Jedes dieser Vokabulare kann wiederum mit thematisch organisierten Wortlisten abgeglichen und statistisch beschrieben werden (LDVLIB: IHA). Eine solche inhaltsanalytische Auswertung der Themenkohärenz, wie sie die hier angedeutete formale Analyse sinnvoll ergänzen kann, wurde von Traue (TRAUE, 1978) an therapeutischen Dialogen vorgenommen. Ausgehend von den DMA hier zugrundegelegten Operationalisierungen wurde von Bodmer (BODMER, 1982) eine Dissertation vorgelegt, die mit einem zusätzlichen LDVLIB-Programm die Ausgewogenheit der Redebeiträge am Beispiel dramatischer Dialoge untersucht und Methoden der Lesbarkeitsmessung (FLESH-Formel) auf die gesprochene Sprache anwendet. Dort sind ebenfalls weitere Kennwerte zum Sprecherverhalten und der Beziehung der Sprecher untereinander diskutiert.

Eine umfangreiche Basis empirisch gewonnener Daten aus Gesprächen müßte bereit stehen, die eine Interpretation und Analyse im Sinne

der anfangs eingeführten Postulate sinnvoll und möglich machen kann. Gespräche bilden den Knotenpunkt sozialen Geschehens und entscheiden oft über Wohl und Wehe der Beteiligten: eine tiefe Kenntnis ihrer Struktur und ihres Funktionierens bietet Handhabe zu Mißbrauch, aber auch zu Verbesserung und Einsicht.

LITERATUR

BODMER, W., Quantifizierbare Aspekte von Dialogen - ihre Operationalisierung und deren Interpretation auf werk- und textsortenspezifischer Ebene, (Dissertation Uni) Zürich, 1982

DREWEK, R., LDVLIB, Dokumentation der Programme für computergestützte Textanalysen, (Rechenzentrum Uni) Zürich, 1975

HOLZ, A., Die Kunst, ihr Wesen und ihre Geschichte, 1891

KUHN, TH., Die Struktur wissenschaftlicher Revolutionen, Frankfurt 1967

TRAUE, H.C., Die Interaktionsanalyse: ein algorithmisches Verfahren zur Untersuchung verbaler Interaktion, (Dissertation Uni) Ulm, 1978

WATZLAWICK, P., BEAVIN, J.H., JACKSON, D.D., Menschliche Kommunikation. Formen, Störungen, Paradoxien, Bern, 1974

ZUR THEORIE DER KLUMPUNG VON TEXTENTITÄTEN

U. Strauß, Bochum
Ch. Sappok, Bochum
H. J. Diller, Bochum
G. Altmann, Bochum

1. Die Fragestellung, ob poetisch relevante Entitäten die Tendenz aufweisen, in bestimmten Abständen voneinander wiederholt aufzutreten, ist eines der ältesten Probleme der Poetik.

Die markantesten Beispiele sind Reim, Alliteration, Assonanz, Parallelismus und Referenz. Seit langem versucht man auch für Fälle, die nicht so regulär sind wie zum Beispiel der Reim, Wahrscheinlichkeitsmodelle aufzustellen oder eine hypothetisch angenommene Tendenz statistisch zu überprüfen. Ein Versuch, die Tendenz zur Wiederholung von Entitäten in kurzem Abstand nacheinander psychologisch zu untermauern, wurde von Skinner unternommen. Nach seiner Ansicht gilt: "the appearance of a sound in speech raises the probability of occurrence of that sound for some time thereafter" (Skinner 1939: 186). Da diese Tendenz unbewußt ist, muß sie nicht nur für Laute, sondern für beliebige sprachliche Entitäten gelten. Ferner muß es die Möglichkeit geben, diese Erscheinung theoretisch abzuleiten. Bisher wurden nur einzelne Hypothesen, insbesondere im phonischen Bereich, statistisch getestet (vgl. Skinner 1939, 1941, Sebeok und Zeps 1959, Knauer 1965, Altmann 1963, 1968, Hřebíček 1965; Herdan 1962: 79-85). In der vorliegenden Arbeit versuchen wir, ein allgemeines Modell für die Wiederholungstendenzen aufzustellen und an den Wiederholungen metrischer Muster im Hexameter zu überprüfen.

2. Wir betrachten ein Gedicht als ein bereits gegebenes Ganzes und untersuchen den Abstand zwischen zwei gleichen Einheiten A. Diese Einheiten können zum Beispiel phonischer, grammatischer, semantischer oder metrischer Natur sein. Der Abstand wird gemessen als die Anzahl aller von A verschiedenen Entitäten des gleichen Typs (A), die zwischen zwei Einheiten A stehen. Bildlich kann man sich also die

zwei Einheiten A als eine Urne vorstellen, in die man Kugeln (= andere Einheiten) plaziert.

Unser Problem, die Frage nach der Wahrscheinlichkeit eines Abstandes der Länge X zwischen zwei Einheiten A, läßt sich somit auf ein Urnenmodell überführen.

Nehmen wir an, daß zwischen dem ersten und dem zuletzt beobachteten k-ten Erscheinen der Einheit A genau r andere Einheiten $\overline{A}$ liegen. Die Vorkommen von A bilden k-1 Urnen (weiter k-1 = n), die anderen Einheiten ($\overline{A}$) entsprechen r Kugeln. Die Fragestellung lautet, wie die Wahrscheinlichkeit ist, daß es beim zufälligen Plazieren von r Kugeln in n Urnen genau n_o leere Urnen, n_1 Urnen mit jeweils einer Kugel, n_2 Kugeln mit jeweils 2 Kugeln, ..., sowie n_r Urnen mit jeweils r Kugeln gibt, wobei gilt:

$$n_o + n_1 + \ldots + n_r = n \qquad (1)$$

$$n_1 + 2n_2 + \ldots + rn_r = r \quad .$$

Die Zahl der Möglichkeiten, n Urnen in Gruppen von n_o, n_1 ..., n_r aufzuteilen, ist

$$\binom{n}{n_o}\binom{n-n_o}{n_1}\binom{n-n_o-n_1}{n_2} \ldots \binom{n-n_o-n_1-\ldots-n_{r-1}}{n_r} = \frac{n!}{n_o!n_1! \ldots n_r!} . \qquad (2)$$

Gleichzeitig ist aber die Zahl der Möglichkeiten, r Kugeln so aufzuteilen, daß in alle n_i Urnen genau i Kugeln kommen, gleich

$$\frac{r!}{(0!)^{n_o}\,(1!)^{n_1} \ldots (i!)^{n_i} \ldots (r!)^{n_r}} . \qquad (3)$$

Da die Zahl aller Möglichkeiten, r Kugeln in n Urnen zu plazieren gleich n^r ist, so erhält man durch Multiplikation von (2) und (3) und nach Division durch n^r

$$P\,(n_o, n_1, \ldots n_r) = \frac{n!\ r!}{n^r \prod_{i=1}^{r} n_i!\ \prod_{i=2}^{r} (i!)^{n_i}} \qquad (4)$$

Dies ist anzusehen als die Wahrscheinlichkeit, daß beim zufälligen Plazieren von r Kugeln in n Urnen genau n_o Urnen leer sind, n_1 Urnen jeweils eine Kugel enthalten usw. In der Sprache der Linguistik bedeutet diese Aussage, daß zwischen n_o Einheiten A der Abstand O besteht, zwischen n_1 Einheiten A genau jeweils 1 andere Einheit $\overline{A}$ steht usw.
Um nun die Verteilung der Abstände abzuleiten, suchen wir die erwartete Anzahl der Urnen n_i (i = 0, 1, ..., r). Es gilt:

$$E\,(n_i) = \sum \ldots \sum n_i \ \frac{n!\ r!}{n^r \prod_{i=1}^{r} n_i!\ \prod_{i=2}^{r} (i!)^{n_i}} \qquad (5)$$

$$= \frac{n\ r!}{n^r i!\ (r-i)!}\,(n-1)^{r-i} = n\binom{r}{i}\left(\frac{1}{n}\right)^{i}\left(1-\frac{1}{n}\right)^{r-i} .$$

Daraus folgt:

$$E\,(n_o) = n\left(1 - \frac{1}{n}\right)^{r}$$

$$E\,(n_1) = r\left(1 - \frac{1}{n}\right)^{r-1} \qquad (6)$$

$$E\,(n_2) = \frac{r(r-1)}{2!}\left(\frac{1}{n}\right)^{1}\left(1 - \frac{1}{n}\right)^{r-2}$$

.....

Man kann zur Bildung der $E\,(n_i)$ auch die Rekursionsformel

$$E\,(n_{i+1}) = \frac{r-i}{i+1}\,\frac{1}{n-1}\,E(n_i) \qquad (7)$$

verwenden.

Mit Hilfe dieses Modells läßt sich sehr einfach testen, zum Beispiel mit einem Chiquadrat-Test, ob die Verteilung empirischer Abstände von den theoretischen Abständen (6) abweicht. Dabei bleibt die Richtung der Abweichung unberücksichtigt. Für den Test, ob eine Tendenz zur Klumpung gleicher Entitäten besteht, reicht es gewöhnlich, die Zahl der Null-Abstände zu untersuchen. Hierbei ergibt sich aus (4):

$$V(n_o) = n(n-1)\left(1 - \frac{2}{n}\right)^r + n\left(1 - \frac{1}{n}\right)^r - n^2\left(1 - \frac{1}{n}\right)^{2r} . \qquad (8)$$

Bei großem n benutzen wir das Kriterium

$$\frac{n_o - E(n_o)}{\sqrt{V(n_o)}} = z , \qquad (9)$$

wobei z eine Normalvariable mit den Parametern (0,1) ist (vgl. David 1950). Ist Skinners Hypothese valide, so müßten beide Tests bei allen poetischen Entitäten signifikante Resultate aufweisen: die Plazierung gleicher Entitäten im Text ist nicht zufällig, sondern folgt einem (bisher unbekanntem) Gesetz. Folgt eine Einheit diesem Modell, so kann man annehmen, daß sie rein zufällig im Text verteilt ist.

Das besprochene Modell ist etwas "statisch" und setzt voraus, daß der Text fertig vorliegt. Dies ist keineswegs nachteilig, da Texte nach Fertigstellung noch korrigiert werden.

Als Überprüfungsbeispiel wählen wir die Verteilung der Abstände zwischen den rhythmischen Mustern des Typs DSSS (D = Daktylus, S = Spondeus) in Bridges "Poems in Classical Prosody, Epistle II: "To a Socialist in London", geschrieben in Hexametern. Wie üblich haben wir nur die ersten vier Versfüße verzeichnet (die letzten zwei sind fixiert). Die ersten 30 Verse dieses Gedichts sind wie folgt:

1. DSSS	11. SDSS	21. DDDS
2. SDSS	12. DSDS	22. DSDD
3. SDSS	13. SDSS	23. DDSS
4. DDSS	14. SSSS	24. SDSS
5. SDSS	15. DSSD	25. DSDS
6. DDSS	16. SDDS	26. DSDS
7. DSDD	17. SDSS	27. SSDS
8. SDSS	18. DSSS	28. DDSS
9. DSSS	19. DSSD	29. DSSS
10. SSSS	20. DDDS	30. DSSS

Das Muster DSSS wiederholt sich hier in den Abständen 7, 7, 10 und 0. In den ersten 300 Versen des Hexameters wurden 65 DSSS-Muster gefunden, somit ist n = 64 (Zahl der Urnen) und r = 300-65 = 255 (Zahl der Kugeln).

Die Erwartung der einzelnen Abstände ist gemäß (6):

$$E(n_o) = 64\left(1 - \frac{1}{64}\right)^{235} = 1.58$$

$$E(n_1) = \frac{235 - 0}{0 + 1}\left(\frac{1}{63}\right) 1.58 = 5.90$$

$$E(n_2) = \frac{235 - 1}{1 + 1}\left(\frac{1}{63}\right) 5.90 = 10,95 .$$

Die beobachteten und die erwarteten Werte sind in Tabelle 1 aufgeführt.

Tab. 1: Verteilung der Abstände zwischen den Wiederholungen des Musters DSSS in Bridges

Abstand	Beobachtet	Erwartet
i	n_i	$E(n_i)$
0	17	1.58
1	13	5.90
2	4	10.95
3	4	13.50
4	6	12.43
5	3	9.12
6	6	5.55
7	2	2.88
8	2	1.30
9	1	0.52
10	2	0.19
11	2	0.06
13	1	0.02
33	1	
	64	

Die Abweichung zwischen dem Modell und der Realität ist bereits optisch so groß, daß man die reine Zufälligkeit in den Daten ablehnen kann.

Der Chiquadrat-Test ergibt $X^2_6 = 200.72$, sowie bei der Zusammenfassung der ersten zwei Klassen $X^2_5 = 109.48$. Der theoretische Wert ist $X^2_{0.05(5)} = 11.07$. Man sieht, daß bei den ersten zwei Abständen (i = 0,1) eine starke Klumpung vorliegt. Der Test nach (9) unter Verwendung von (6) und (8) ergibt einen sehr hoch signifikanten z-Wert:

$$\frac{17 - 1.58}{\sqrt{64(63)(1 - \frac{2}{64})^{235} + 64(1 - \frac{1}{64})^{235} - 64^2(1 - \frac{1}{64})^{2(235)}}}$$

$$= \frac{15.42}{1.1834} = 13.03$$

Dieses Resultat zeigt, daß bei den Wiederholungen von identischen Einheiten ein unterbewußter Mechanismus die Abstände steuert. Zu seiner Verständigung werden wir zwei weitere Modelle vorstellen.

3. Es besteht die Möglichkeit, das Gedicht nicht in Form von Urnen, sondern als eine Kette von Elementen zu betrachten. Die Skinnersche Hypothese sagt, daß zwischen den Wiederholungen einer Einheit und daher auch zwischen dem Erscheinen der gegebenen und aller anderen Einheiten eine gewisse Abhängigkeit besteht. Man kann sich deshalb fragen, ob eine Folge von Einheiten A und $\overline{A}$ eine Markov-Kette bildet. Eine recht ausführliche Behandlung solcher Ketten für linguistische Zwecke wurde von Brainerd (1978) durchgeführt. Wir werden uns daher an seine Darstellungsweise halten.

Man betrachtet eine Folge von Elementen A und $\overline{A}$ als eine Markov-Kette mit zwei Zuständen, wobei der Zustand A (in unserem Beispiel das Muster DSSS) als 1 und der Zustand $\overline{A}$ als 0 bezeichnet werden kann. Die ersten 30 Verse im Gedicht von Bridges kann man schreiben als

100000001000000001000000000011.

Wenn es irgendwelche Konfigurationen (Teilketten) von Mustern gibt, so muß ein Muster aufgrund der Kenntnis seiner Vorgänger voraussagbar sein, das heißt, man muß die Wahrscheinlichkeit seines Vorkommens hinter den bekannten Vorgängern berechnen können. Es handelt sich hierbei um die bedingte Wahrscheinlichkeit, daß in der Position n das gegebene Muster M erscheint, wenn in den ersten n-1 Positionen bereits bekannte Muster stehen:

$$P\ (M_n | M_1 M_2 M_3 \ \ldots\ M_{n-1}) \qquad (10)$$

Hat man es nur mit zwei Mustern zu tun, die man als 1 und 0 bezeichnet, dann bedeutet $M_n = 1$ die Tatsache, daß sich der hexametererzeugende Prozess beim n-ten Schritt (bei der Erzeugung des n-ten Verses) im Zustand 1 befindet. Bezeichnet man den Zustand allgemein als x (x = 0,1) so kann man (10) als

$$P\ (M_n = x_n \,|\, M_1 = x_1;\ M_2 = x_2\ \ldots\ M_{n-1} = x_{n-1}) \qquad (11)$$

schreiben. Wenn die Musterfolge voneinander völlig unabhängig sind, dann reduziert sich (11) auf

$$P\ (M_n = x_n). \qquad (12)$$

Ein Prozess der Art (12) wird als Markov-Kette nullter Ordnung bezeichnet. Wenn das Erscheinen eines Musters lediglich von dem unmittelbar vor ihm erzeugten Muster abhängt, denn reduziert sich (11) auf

$$P\ (M_n = x_n \,|\, M_{n-1} = x_{n-1}). \qquad (13)$$

Ein Prozess dieser Art stellt die eigentliche Markov-Kette oder die Markov-Kette erster Ordnung dar.

Die Ordnung der Markov-Kette hängt davon ab, wie viele Vorgänger das Erscheinen eines Musters bedingen. So stellt

$$P(M_n = x_n \,|\, M_{n-2} = x_{n-2},\ M_{n-1} = x_{n-1}) \qquad (14)$$

eine Markov-Kette zweiter Ordnung,

$$P(M_n=x_n | M_{n-3}=x_{n-3}, M_{n-2}=x_{n-2}, M_{n-1}=x_{n-1}) \qquad (15)$$

zum Beispiel eine Markov-Kette dritter Ordnung dar.
Wenn wir nur ein bestimmtes Muster betrachten und mit 1, sowie alle anderen Muster mit 0 bezeichnen, dann stellen die 0 Folgen eine Lücke (Abstand) zwischen zwei Vorkommmen des Musters 1 dar. Die Länge dieser Lücke kann als die Zahl der Nullen betrachtet werden. Sie stellt eine Zufallsvariable dar, die wir mit Y bezeichnen. Die Wahrscheinlichkeitsverteilung von Y läßt sich aus der Markov-Kette ableiten und an die Daten anpassen.

a) Markov-Kette nullter Ordnung

Wenn die Kette nullter Ordnung ist, dann sind die einzelnen Ereignisse voneinander unabhängig. In dem Falle kann man die Wahrscheinlichkeit einer Sequenz von Mustern als

$$P(M_1)P(M_2)\ldots P(M_n) \qquad (16)$$

schreiben. Wenn nur zwei Zustände vorhanden sind, erhält man für den Abstand zwischen zwei Mustern 1, wobei das erste Muster 1 die Wahrscheinlichkeit 1 hat (also gegeben ist):

$$1P(M_1=0)P(M_2=0)\ldots P(M_k=0)P(M_{k+1}=1)$$

oder verkürzt

$$1P(0)\ldots P(0)P(1) = P(0)^k P(1). \qquad (17)$$

Da die Länge der Lücke zwischen zwei Mustern 1 als die Zufallsvariable Y betrachtet wird, so kann man schreiben:

$$P(Y=k) = P(1)P(0)^k; \quad k = 0,1,\ldots \qquad (18)$$

Da wir nur zwei Zustände haben, ist

$$P(0) = 1-P(1),$$

so daß man schließlich erhält:

$$P(Y=k) = P(1)\left[1-P(1)\right]^k, \quad k = 0,1,\ldots \qquad (19)$$

Dies ist die Wahrscheinlichkeitsfunktion der geometrischen Verteilung, mit deren Hilfe die Verteilung der Lücken in der Linguistik modelliert wurde. $P(1)$ ist der Parameter dieser Verteilung und kann aus den Daten geschätzt werden. Die einfachste Schätzung ist:

$$\hat{P}(1) = f_o/N \qquad (20)$$

wobei f_o/N die relative Häufigkeit der nullten Klasse ist.

Eine weitere Schätzung ergibt sich aus dem Mittelwert. Wegen

$$\mu_1' = \sum_{k=0}^{\infty} k\, P(1)\left[1-P(1)\right]^k = \frac{1-P(1)}{P(1)} \qquad (21)$$

erhält man

$$\hat{P}(1) = \frac{1}{1+\bar{y}} \qquad (22)$$

Hierbei ist $\bar{y}$ der Mittelwert der empirischen Verteilung. Für unser Beispiel in Tab. 1 erhalten wir gemäß (20):

$$\hat{P}(1) = 17/64 = 0.2656$$

und gemäß (22) wegen $\bar{y} = 235/64 = 3.6719$

$$\hat{P}(1) = \frac{1}{1+3.6719} = 0.2140 .$$

Die theoretischen Häufigkeiten, berechnet mit Hilfe dieser Schätzungen, sind in der Tabelle 2 aufgeführt.

Die einzelnen Häufigkeiten kann man rekursiv mit Hilfe der Formel

$$NP_{k+1} = \left[1-P(1)\right] NP_k \qquad (23)$$

berechnen. Überprüft man die Anpassung mit Hilfe eines Chiquadrat-Tests, so ergibt sich im ersten Fall (dritte Spalte von Tab. 2) $\chi^2_{10} = 14.08$ mit $P = 0.17$ und im zweiten Fall (vierte Spalte) $\chi^2_{11} = 8.52$ mit $P = .67$.

Tab. 2: Verteilung der Abstände zwischen den Wiederholungen von DSSS in Bridges

Abstand Y	Beobachtet f_y	Berechnet NP_y	
		$\hat{P}(1) = 0.2656$	$\hat{P}(1) = 0.2140$
0	17	17.00	13.70
1	13	12.48	10.77
2	4	9.17	8.46
3	4	6.73	6.65
4	6	4.94	5.23
5	3	3.63	4.11
6	6	2.67	3.23
7	2	1.96	2.54
8	2	1.44	2.00
9	1	1.06	1.57
10	2	0.78	1.23
11	2	0.57	0.97
12	0	0.42	0.76
13	1	0.31	0.60
33	1	0.84	2.18

Bei beiden Schätzungen ist die Markov-Kette nullter Ordnung, also die geometrische Verteilung, ein annehmbares Modell für die Verteilung der Abstände zwischen den Wiederholungen. Dies bedeutet aber gleichzeitig, daß es Wiederholungen gibt, die keine Abhängigkeit aufweisen, so daß sie der Skinnerschen Hypothese eigentlich nicht entsprechen, obwohl in Absatz 2 dieser Untersuchung eine Klumpungstendenz nachgewiesen wurde.

Wir überprüfen deshalb, ob die Markov-Kette erster Ordnung eine signifikante Verbesserung bringt.

b) Markov-Kette erster Ordnung

Bei der Markov-Kette erster Ordnung (Formel (13)) bedeutet der Abstand 0 einen Übergang von einem Zustand 1 zu einem gleichen Zustand 1:

$$P(Y=0) = P(M_n=1 \mid M_{n-1}=1) = P(1|1).$$

So haben wir zum Beispiel in der Kette

$$\underset{\sqcup}{1}0000 \ldots 00\underset{\sqcup}{1}$$

einen Übergang von 1 auf 0 mit der Wahrscheinlichkeit $P(M_1=0|M_0=1) = P(0|1)$; $(k-1)$-Übergänge von 0 auf 0 mit $P(0|0)^{k-1}$ und einen Übergang von 0 auf 1 mit $P(1|0)$.

Demnach ist

$$P(Y=k) = P_k = \begin{cases} P(1|1) & \text{für } k=0 \\ P(0|1)P(0|0)^{k-1}P(1|0) & \text{für } k=1,2\ldots \end{cases} \quad (24)$$

Dies ist eine modifizierte geometrische Verteilung (vgl. Johnson und Kotz 1969: 204 ff). Für die Anpassung brauchen wir nur die Schätzung von zwei Parametern, nämlich $P(1|1)$ und $P(1|0)$, weil

$$P(0|0) = 1 - P(1|0) \quad (25)$$

und

$$P(0|1) = 1 - P(1|1). \quad (26)$$

Da $P(1|1)$ die Wahrscheinlichkeit des Abstands 0 bedeutet, schätzen wir sie aus der relativen Häufigkeit der nullten Klasse:

$$\hat{P}(1|1) = f_o/N. \quad (27)$$

Weiter ist wegen (25)

$$P(1|0)P(0|1)\sum_{k=1}^{\infty} kP(0|0)^{k-1} = \frac{P(0|1)}{P(1|0)}, \tag{28}$$

so daß gilt:

$$P(1|0) = \frac{P(0|1)}{\mu_1'}. \tag{29}$$

Da $P(0|1)$ die Wahrscheinlichkeit eines Abstands größer als 0 bedeutet, also $P(0|1) = P(Y \geq 1) = 1 - P(Y = 0)$, folgt daraus:

$$\hat{P}(0|1) = 1-f_o/N. \tag{30}$$

Nimmt man als Schätzung für μ_1' den Durchschnitt $\bar{y}$ und setzt beides in (29) ein, so erhält man:

$$\hat{P}(1|0) = \frac{1 - f_o/N}{\bar{y}}. \tag{31}$$

Eine andere Möglichkeit der Schätzung ergbit sich mit Hilfe der Häufigkeiten der nullten und der ersten Klassen. Wir schätzen $P(1|1)$ wie oben in (27). Da ferner gilt:

$$\hat{P}(Y = 1) = \hat{P}(0|1)\ \hat{P}(1|0) = f_1/N,$$

erhalten wir wegen (30):

$$\hat{P}(1|0) = \frac{f_1/N}{1-f_o/N}. \tag{32}$$

In unserem Beispiel haben wir laut (27)

$$\hat{P}(1|1) = f_o/N = \frac{17}{64} = 0.2656.$$

Hieraus folgt gemäß (30):

$$\hat{P}(0|1) = 1 - 0.2656 = 0.7344.$$

Wegen $\bar{y} = 3.6718$ bekommen wir mit (31):

$$\hat{P}(1|0) = \frac{1 - 0.2656}{3.6719} = 0.2000$$

und

$$P(0|0) = 1 - 0.2000 = 0.8000.$$

Nach (32) erhalten wir:

$$\hat{P}(1|0) = \frac{13/64}{1-0.2656} = 0.2766$$

und

$$\hat{P}(0|0) = 1-0.2766 = 0.7234,$$

Die mit diesen Schätzungen berechneten Werte sind in Tabelle 3 angegeben. Im ersten Fall (dritte Spalte) ist $X^2_{11}=9.77$ mit $P=0.55$, im zweiten Fall (vierte Spalte) ist $X^2_8=15.52$ mit $P=0.05$.

Ob die Verbesserung der Anpassung beim Übergang von der Markov-Kette nullter Ordnung zur Markov-Kette erster Ordnung signifikant ist, überprüfen wir mit Hilfe des Likelihood-ratio-Kriteriums. Wir bilden das Likelihood-ratio, in dem wir die Maximum-Likelihood-Funktion der Markov-Kette erster Ordnung in den Zähler und die der Markov-Kette nullter Ordnung in den Nenner setzen.

Für die erste Ordnung der Markov-Kette gilt:

$$L_1 = P(1|1)^{f_o} \prod_{k=1}^{n} \left[P(1|0)P(0|1)\ P(0|0)^{k-1}\right]^{f_k} \tag{33}$$

und für die nullte Ordnung der Markov-Kette:

$$L_o = \prod_{k=0}^{n} \left[P(1)\ P(0)^k\right]^{f_k}. \tag{34}$$

Tab. 3: Anpassung der Markov-Kette erster Ordnung

Abstand Y	Beobachtet f_y	Berechnet mit $\hat{P}(1/1)=0.2656$ $\hat{P}(1/0)=0.2000$	Berechnet mit $\hat{P}(1/1)=0.2656$ $\hat{P}(1/0)=0.2766$
0	17	17.00	17.00
1	13	9.40	13.00
2	4	7.52	9.40
3	4	6.02	6.80
4	6	4.81	4.92
5	3	3.85	3.56
6	6	3.08	2.58
7	2	2.46	1.86
8	2	1.97	1.35
9	1	1.58	0.97
10	2	1.26	0.71
11	2	1.01	0.51
12	0	0.81	0.37
13	1	0.64	0.27
33	1	2.59	0.70

Hieraus ergibt sich:

$$\lambda = \frac{L_1}{L_0} = \left[\frac{\hat{P}(1|1)}{\hat{P}(1)}\right]^{f_0} \left[\frac{\hat{P}(0|1)\hat{P}(1|0)}{\hat{P}(1)\ \hat{P}(0|0)}\right]^{N-f_0} \left[\frac{\hat{P}(0|0)}{\hat{P}(0)}\right]^{N\bar{y}}. \qquad (35)$$

In diese Formel setzt man die Maximum-Likelihood-Schätzungen der einzelnen Parameter ein, die sich aus (22), (27) und (31) ergeben. Mit (35) erhalten wir

$$\lambda = \left[\frac{0.2656}{0.2140}\right]^{17} \left[\frac{0.7344(0.2)}{0.2140(0.8)}\right]^{64-17} \left[\frac{0.8}{0.786}\right]^{64(3.6719} = 1.86 .$$

Da $2 \ln \lambda$ ungefähr wie ein χ^2 mit 1 Freiheitsgrad verteilt ist und daher $2 \ln (1.86) = 1.24$ nicht signifikant ist ($P = 0.27$), sehen wir, daß der Übergang zur ersten Ordnung keine wesentliche Verbesserung bringt.

Die Tatsache, daß die Markov-Kette nullter Ordnung für die Anpassung hinreichend ist, hat für das Modell die ungünstige Folge, daß man hier kaum von Abhängigkeiten, wie sie die Skinnersche Hypothese ansetzt, sprechen kann. Die Abstände folgen der geometrischen Verteilung, wobei der sie erzeugende Mechanismus unklar bleibt. Auf der anderen Seite läßt sich zeigen, daß die Markov-Kette nullter Ordnung nicht immer ausreicht. Der Rückgriff auf Markov-Ketten höherer Ordnungen führt in vielen Fällen zur Verbesserung der Anpassung dadurch, daß die geometrische Verteilung immer mehr modifiziert wird. Dies führt zu einer Vermehrung der Parameter, die sich jedoch nicht alle als "Klumpungsparameter" interpretieren lassen.

Durch Modifizierung wird das Modell zerstückelt, es kann aber kaum angenommen werden, daß im Texterzeuger nebeneinander Mechanismen vorhanden sind, die separat 0-Abstände, 1-Abstände und weitere Abstände erzeugen. Hinzu kommt der Umstand, daß die nichtmodifizierte geometrische Verteilung monoton fallend ist, während dies in der Empirie nicht immer der Fall ist. Nicht monoton fallende Verteilungen wie man sie oft in der Empirie beobachten kann, lassen sich mit der geometrischen Verteilung nur durch Modifizierung erfassen.

Auf der anderen Seite erlaubt die Markov-Kette, den Abhängigkeitsgrad festzustellen. Dadurch bietet sie eine neue Methode zur Untersuchung der Eigenschaften der Wiederholungen. Es ist gut möglich, daß phonische Elemente ganz andere Ketten bilden als metrische, semantische und andere Entitäten. Dies bedarf aber einer umfangreichen Untersuchung auf allen Ebenen der Sprache (für grammatische Entitäten vgl. Brainerd 1976; für phonologische Köhler 1980).

Brainerd kommt in seiner Untersuchung bis zu Markov-Ketten dritter Ordnung. Es läßt sich nicht bestreiten, daß im Text solche Abhängigkeiten vorkommen. Es besteht auch kein Zweifel daran, daß man mit Markov-Ketten große Teile der Syntax modellieren

kann - jedoch nicht in dem Sinn, den Miller und Chomsky (1963) abgelehnt haben, sondern ungefähr in der Art, wie man Computer-Musik modelliert.

4. Diese geschilderten Umstände zwingen dazu, von anderen Annahmen auszugehen und nach einem "kompakteren" Modell zu suchen, in dem die "Klumpungshypothese" in einer expliziten Form erscheint.

Wie wir in Absatz 2 gesehen haben, ist die Verteilung der $\bar{A}$-Muster zwischen den A-Mustern nicht ganz analog der zufälligen Verteilung von Kugeln in Urnen. Sie folgt irgendeiner "Klumpungstendenz" oder "Klumpungsabsicht", die man auch mit Markov-Ketten nicht erfassen kann. Es besteht aber eine Tendenz, möglichst viele "Urnen" leer zu lassen und die Kugeln in einigen wenigen Urnen zu häufen. Wir wollten versuchen, diese Tendenz mit Hilfe des Poisson-Prozesses abzuleiten.

Betrachten wir die A-Einheiten für das ganze Gedicht als gegeben und stellen wir uns vor, daß zur Zeit t = 0 keine $\bar{A}$-Einheiten zwischen ihnen liegen (alle Zwischenräume sind leer). Wir bezeichnen die Wahrscheinlichkeit, daß zur Zeit t in einem Zwischenraum genau x $\bar{A}$-Einheiten vorhanden sind, als $P_x(t)$. Wir nehmen an, daß im Zeitintervall (t, t+dt) eine $\bar{A}$-Einheit in einem Zwischenraum, in dem schon x $\bar{A}$-Einheiten vorhanden sind mit der Wahrscheinlichkeit $f_x(t)dt$ plaziert wird. Der Proportionalitätsfaktor ist also nicht konstant, sondern eine Funktion, die es uns ermöglicht, die Klumpungstendenz zu erfassen. Die Wahrscheinlichkeit, daß im Intervall (t, t+dt) mehr als eine $\bar{A}$-Einheit in den Zwischenraum plaziert wird, ist vernachlässigbar klein. Die Wahrscheinlichkeit, daß keine $\bar{A}$-Einheit hereinkommt, ist somit $1-f_x(t)dt$. Die Unabhängigkeit der Ereignisse in den einzelnen Zeitintervallen wird wie üblich vorausgesetzt.

Setzen wir diese Wahrscheinlichkeit zusammen, so erhalten wir

$$P_o(t+dt) = P_o(t)\left[1-f_o(t)dt\right]. \qquad (36)$$

Die Wahrscheinlichkeit also, daß zum Zeitpunkt t+dt keine $\bar{A}$-Einheit in einem Zwischenraum liegt, setzt sich zusammen aus dem Produkt von $P_o(t)$ (keine $\bar{A}$-Einheit zur Zeit t) und von $1-f_o(t)dt$

(kein Zuwachs in dt.) Ferner gilt:

$$P_x(t+dt) = P_x(t)\left[1-f_x(t)\right]dt + P_{x-1}(t)f_{x-1}(t)dt, \qquad (37)$$

d. h. entweder gab es zur Zeit t bereits x $\bar{A}$-Einheiten und keinen Zuwachs in dt oder es gab x-1 $\bar{A}$-Einheiten und einen Zuwachs mit der Wahrscheinlichkeit $f_{x-1}(t)dt$.

Diese beiden Gleichungen können wir im Grenzübergang für x = 1,2,... schreiben als:

$$\lim_{dt\to o} \frac{P_o(t+dt)-P_o(t)}{dt} = \frac{dP_o(t)}{dt} = -P_o(t)f_o(t) \qquad (38\ a)$$

$$\lim_{dt\to o} \frac{P_x(t+dt)-P_x(t)}{dt} = \frac{dP_x(t)}{dt} = P_x(t)f_x(t)+P_{x-1}(t)f_{x-1}(t)$$

$$\text{für } x = 1,2,3... \qquad (38\ b)$$

Da zur Zeit t=0 keine $\bar{A}$-Einheiten dazwischen liegen, haben wir die Anfangsbedingungen $P_o(0) = 1$ und $P_x(0) = 0$ für x = 1,2... zu berücksichtigen.

Diese Gleichungen lassen sich z. B. mit Hilfe von erzeugenden Funktionen lösen. Wir multiplizieren (38 a) mit s^o (also mit 1) und jede Gleichung in (38 b) mit s^x (x = 1,2...). Dann erhält man:

$$\frac{dP_o(t)s^o}{dt} = -P_o(t)f_o(t)s^o$$

$$\frac{dP_1(t)s^1}{dt} = -P_1(t)f_1(t)s^1 + P_o(t)f_o(t)s^1$$

$$\frac{dP_2(t)s^2}{dt} = -P_2(t)f_2(t)s^2 + P_1(t)f_1(t)s^2 .$$

.....

Addiert man beide Seiten, so erhält man:

$$\frac{d}{dt}\sum_{x=0}^{\infty} P_x(t)s^x = -\sum_{x=0}^{\infty} P_2(t)f_x(t)s^x + s\sum_{x=0}^{\infty} P_x(t)f_x(t)s^x$$

$$= (s-1)\sum_{x=0}^{\infty} P_x(t)f_x(t)s^x \quad . \tag{39}$$

Hier ist $\sum_{x=0}^{\infty} P_x(t)s^x$ die wahrscheinlichkeitserzeugende Funktion, die wir als $G(t,s)$ bezeichnen. Die Funktion auf der rechten Seite in (39) kann man als:

$$\sum_{x=0}^{\infty} P_x(t)f_x(t)s^x = F(t,s) \tag{40}$$

bezeichnen. Die Lösung von (39) hängt davon ab, wie man $f_x(t)$ bestimmt. Gemäß der Skinnerschen Hypothese und der obigen Überlegungen, die zu diesem Modell führten, ist $f_x(t)$ der Proportionalitätsfaktor. Wir wählen ihn so, daß er nicht von der Zeit abhängt, sondern allein davon, wie viele $\bar{A}$-Einheiten bereits im Zwischenraum liegen.

Da eine Tendenz bestehen soll, die A-Einheiten eng nebeneinander zu plazieren, neigen leere Zwischenräume dazu, neue $\bar{A}$-Einheiten abzuweisen. Dies führt automatisch dazu, daß ein Zwischenraum umso mehr weitere $\bar{A}$-Einheiten "anzieht", je mehr $\bar{A}$-Einheiten in diesem Zwischenraum liegen.

Die $\bar{A}$-Einheiten häufen sich also in einigen wenigen Zwischenräumen. Mit anderen Worten: es gibt viele Zwischenräume mit wenigen $\bar{A}$-Einheiten und wenige Zwischenräume mit vielen $\bar{A}$-Einheiten.

Ansatzweise nehmen wir an, daß die "Anziehung" weiterer $\bar{A}$-Einheiten in einem Zwischenraum eine lineare Funktion der Anzahl der

hier bereits vorhandenen $\bar{A}$-Einheiten ist:

$$f_x(t) = a + bx \quad . \tag{41}$$

Setzen wir (41) in (40) ein, so erhalten wir:

$$F(t,s) = \sum_{x=0}^{\infty} P_x(t)(a+bx)s^x = a\sum_{x=0}^{\infty} P_x(t)s^x + b\sum_{x=0}^{\infty} xP_x(t)s^x \quad . \tag{42}$$

Da $\sum_{x=0}^{\infty} xP_x(t)s^x = s\frac{\partial}{\partial s}\sum_{x=0}^{\infty} P_x(t)s^x$ ist,

gilt:

$$F(t,s) = aG(t,s) + bs\frac{\partial}{\partial s}G(t,s) \quad . \tag{43}$$

Setzt man (43) in (39) ein, so erhält man:

$$\frac{\partial}{\partial t}G(t,s) = (s-1)\left[aG(t,s) + bs\frac{\partial}{\partial s}G(t,s)\right] . \tag{44}$$

Die Lösung dieser partiellen Differntialgleichung ergibt sich als:

$$G(t,s) = \left[e^{bt} - (e^{bt}-1)s\right]^{-a/b} \tag{45}$$

oder, wenn man $e^{bt} - 1 = P$, $e^{bt} \triangleq Q$ sowie $a/b = k$ setzt, als:

$$G(s) = (Q-Ps)^{-k} . \tag{46}$$

Dies ist die wahrscheinlichkeitserzeugende Funktion der negativen Binomialverteilung, also (vgl. Johnson, Kotz 1969: 125):

$$P_x = \binom{k+x-1}{x} \left(\frac{P}{Q}\right)^x Q^{-k}, \quad x = 0,1,2,... \tag{47}$$

Es ist selbstverständlich möglich, statt (41) andere "Klumpungsfunktionen" zu verwenden und mit ihnen (39) lösen, was dann zu anderen Wahrscheinlichkeitsverteilungen führen kann. Bei diesem Verfahren bekommen wir nicht die Ordnung der Abhängigkeit (die, wie gezeigt, bei einer tatsächlichen vorhandenen Klumpung ganz irreführend sogar 0 sein kann), sondern irgendwelche Klumpungskonstanten in einer Wahrscheinlichkeitsverteilung, die für die Abstandsbildung bei den Wiederholungen zuständig ist. Die Zeitvariable, die wir zur Ableitung des Poisson-Prozesses gebraucht haben, ist nur eine Hilfsvariable, mit der die Erzeugung des Textes veranschaulicht wird. Sie kann zum Schluß gleich 1 gesetzt werden.

Die Tatsache, daß wir die negative Binomialverteilung erhalten haben, ist in der Hinsicht erfreulich, daß die geometrische Verteilung eben ein Spezialfall der negativen Binomialverteilung ist (mit k = 1 in (47)).

Ein weiterer Umstand ist noch zu bedenken. Die negative Binomialverteilung hätten wir auch aus einem Wartezeitmodell bekommen.

Nehmen wir nämlich an, daß die A-Elemente mit Wahrscheinlichkeit P vorkommen, so ist die Wahrscheinlichkeit, daß bis zum Erscheinen des k-ten A-Elements im Text genau $\overline{A}$-Elemente vorkommen werden gleich

$$P_x = \binom{k+x+1}{x} p^k q^x, \quad x=0,1,2,... \tag{48}$$

Diese Formel ist mit (47) identisch, wenn man q=P/Q und p=1/Q setzt. Diese Ableitung hat aber den Nachteil, daß hier der "Klumpungsmechanismus" nicht zutage tritt. Andererseits hat sie aber den Vorteil, daß der Parameter p eine eindeutige Interpretation als die Wahrscheinlichkeit der A-Einheit besitzt. Wir können also p aus der relativen Häufigkeit von A im Text abschätzen:

$$\hat{p} = f_A/n, \tag{49}$$

woraus sich

$$\hat{Q} = 1/\hat{p} \text{ und } \hat{P} = \hat{Q}-1 \tag{50}$$

ergibt.

Wegen

$$P_o = Q^{-k}$$

erhält man dann die Schätzung des "Klumpungsparamters" k als

$$\hat{k} = \frac{-\ln f_o/N}{\ln \hat{Q}} \tag{51}$$

In unserem Beispiel gibt es 65 DSSS-Muster in 300 Versen, also ist

$$\hat{p} = 65/300 = 0.2167$$
$$\hat{Q} = 4.6154$$
$$\hat{P} = 3.6154$$

und

$$\hat{k} = \frac{-\ln (17/64)}{\ln 4.6154} = 0.8668.$$

Eine andere Schätzungsmöglichkeit ergibt sich aus

$$\hat{P} = \frac{s^2 - \bar{y}}{\bar{y}} \quad \text{und} \quad \hat{k} = \frac{\bar{y}^2}{s^2 - \bar{y}} \tag{52}$$

Aus den Daten folgt

$$s^2 = 25.3351$$

$$\bar{y} = 3.6719$$

und daraus

$$\hat{P} = \frac{25.3351 - 3.6719}{3.6719} = 5.8997$$

$$\hat{k} = \frac{3.6719^2}{25.3351 - 3.6719} = 0.6224 \quad .$$

Die dritte Möglichkeit, P iterativ zu berechnen, folgt aus:

$$\frac{\hat{P}}{\ln (1+\hat{P})} = \frac{\bar{y}}{-\ln (f_o/N)} \quad . \qquad (53)$$

In unserem Fall also:

$$\frac{\hat{P}}{\ln (1+P)} = \frac{3.6719}{-\ln (17/64)} = 2.7698 \, .$$

Dies ergibt P = 4.9305.

Die Schätzung von k kann auch hier laut (51) oder (52) erfolgen. Aus (51) bekommen wir $\hat{k}$= 0.8309 und aus (52) erhalten wir $\hat{k}$=0.6224 wie oben.

Zwei weitere, etwas langwierigere Methoden findet man in Johnson und Kotz (1969: 131-15). Die einzelnen erwarteten

Häufigkeiten bekommen wir durch

$$NP_o = NQ^{-k} = N(1+P)^{-k} \qquad (54)$$

und die weiteren rekursiv als

$$NP_{y+1} = \frac{k + y}{y + 1} \frac{P}{Q} NP_y \quad . \qquad (55)$$

Die erwarteten Häufigkeiten sind in der Tabelle 4 aufgeführt. Wie man sieht, ist die Anpassung bei jeder der vier Schätzungen sehr gut. Durch geeignete Gruppierung lassen sich die X^2-Werte noch stark reduzieren.

Tab. 4: Anpassung der negativen Binomialverteilung

y	f_y	NP			
		$\hat{P}$=3.6154 $\hat{k}$=0.8668	$\hat{P}$=5.8997 $\hat{k}$=0.6224	$\hat{P}$=4.9305 $\hat{k}$=0.8309	$\hat{P}$=4.9305 $\hat{k}$=0.6224
0	17	17.00	19.23	14.58	21.14
1	13	11.54	10.24	10.17	10.94
2	4	8.44	7.10	7.67	7.38
3	4	6.32	5.31	6.01	5.36
4	6	4.78	4.11	4.79	4.04
5	3	3.65	3.25	3.85	3.10
6	6	2.79	2.60	3.11	2.42
7	2	2.15	2.11	2.52	1.90
8	2	1.65	1.72	2.05	1.51
9	1	1.28	1.41	1.67	1.20
10	2	0.99	1.16	1.37	0.96
11	2	0.76	0.95	1.12	0.77
12	0	0.59	0.79	0.92	0.62
13	1	0.46	0.66	0.75	0.50
33	1	1.60	3.37	3.51	2.18
	X^2	10.75	9.85	10.21	13.17
	FG	10	10	11	10
	P	0.38	0.45	0.51	0.21

Zum Vergleich bringen wir noch die Verteilung der Abstände zwischen den Vorkommen des Musters SDSS bei Bridges sowie DDSD und SDSD in polnischen Hexametern bei Wallenrod.

In beiden Fällen geben wir nur eine Anpassung mit Hilfe der drei Verteilungen, Markov-Kette 0-ter Ordnung (geometrische Verteilung), Markov-Kette erster Ordnung (modifizierte geometrische Verteilung) sowie negative Binomialverteilung an. Wir benutzten hier nur die einfachste Schätzung der Parameter, die sich mit der Maximum-Likelihood oder Minimum-Chiquadrat Methode usw. noch verbessern läßt.

Tab. 5: Anpassung an die Abstände zwischen SDSS bei Bridges

y	f_y	Berechnet		
		0.Ordnung $\hat{P}(1)=0.1987$	1.Ordnung $\hat{P}(1/1)=0.1695$ $\hat{P}(1/0)=0.2059$	neg. Binomial $\hat{P}=3.3318$ $\hat{k}=1.2107$
0	10	11.72	10.00	10.00
1	12	9.39	10.01	9.31
2	7	7.53	8.01	7.92
3	6	6.03	6.36	6.52
4	8	4.83	5.05	5.28
5	1	3.87	4.01	4.23
6	2	3.10	3.19	3.37
7	3	2.49	2.53	2.67
8	4	1.99	2.01	2.11
9	0	1.60	1.60	1.66
10	1	1.28	1.27	1.30
11	1	1.03	1.01	1.02
12	1			
19	2	4.13	3.88	3.62
21	1			
	X^2	9.42	8.69	8.85
	FG	11	11	10
	P	0.58	0.65	0.55

Tab. 6: Anpassung an die Abstände zwischen DDSD bei Wallenrod

y	f_y	Berechnet		
		0.Ordnung $\hat{P}(1)=0.2211$	1.Ordnung $\hat{P}(1/1)=0.2951$ $\hat{P}(1/0)=0.2000$	neg. Binomial $\hat{P}=3.4601$ $\hat{k}=1.0186$
0	18	13.49	18.00	13.30
1	6	10.51	8.60	10.51
2	7	8.18	6.88	8.23
3	9	6.37	5.50	6.42
4	5	4.96	4.40	5.01
5	2	3.87	3.52	3.90
6	3	3.01	2.82	3.03
7	1	2.35	2.25	2.36
8	0	1.83	1.80	1.84
9	0	1.42	1.44	1.43
10	4	1.11	1.15	1.11
11	2	0.86	0.92	0.86
12	3	0.67	0.74	0.67
13	0	0.52	0.59	0.52
14	0	0.41	0.47	0.40
15	1	1.44	1.92	1.40
	X^2	9.17	6.85	9.42
	FG	9	8	8
	P	0.42	0.55	0.31

Tab. 7: Anpassung an die Abstände zwischen SDSD bei Wallenrod

y	f_y	Berechnet		
		0.Ordnung $\hat{P}=(1)=0.6048$	1.Ordnung $\hat{P}=(1/1)=0.6048$ $\hat{P}=(1/0)=0.5739$	neg. Binomial $\hat{P}=0.7829$ $\hat{k}=0.8796$
0	101	101.00	101.00	100.42
1	38	39.92	37.88	38.79
2	15	15.77	16.14	16.01
3	8	6.23	6.88	6.75
4	4	2.46	2.93	2.87
7	1	1.62	2.17	2.16
	x^2	1.83	1.28	1.38
	FG	4	3	3
	P	0.77	0.73	0.71

5. Wir kommen zu folgendem Ergebnis:
Die Markov-Kette nullter Ordnung liefert in allen unseren Fällen eine hinreichende Anpassung. Dies zeugt aber von der Nichtexistenz irgendwelcher Abhängigkeiten und spricht gegen die Skinnersche Hypothese. Die A-Einheiten werden demnach im Prozess der Texterzeugung zufällig plaziert. Diese Zufälligkeit ist nicht identisch mit der Plazierung in Urnen, da es sich hier offensichtlich um Sequenzen handelt.

Die Markov-Kette erster Ordnung weist eine etwas bessere Anpassung auf, die aber zu Lasten einer Modifzierung und Zerstückelung des Modelles geht. Nichtsdestoweniger bringt sie keine signifikante Verbesserung. Für die vier präsentierten Beispiele bekommen wir die folgenden Werte von 2 ln 2: Bridges DSSS: 1.24, Bridges SDSS: 0.405, Wallenrod DDSD: 2.38, Wallenrod SDSD: 1.66. Falls die Kette nullter Ordnung nicht zutrifft, so kann man mit Markov-Ketten die Abhängigkeitsordnung feststellen. Die

sequentielle Abhängigkeit ist aber so allgemein, daß sie die Klumpung nicht erklärt. Auch wenn im Text die Übergänge vom Zustand 1 zum Zustand 1 vermieden würden, könnte man eine echte Markov-Kette erhalten.

Die Ableitung aus dem Poisson-Prozess mit einer Hypothese über Klumpung liefert dagegen sowohl eine gute Anpassung als auch eine fruchtbare Gesetzeshypothese im Skinnerschen Sinne. Die Fruchtbarkeit besteht darin, daß man in diesem Modell beliebig andere "Klumpungsfunktionen" $f_x(t)$ ansetzen und überprüfen kann, daß ferner die Parameter gut interpretierbar, wenn auch vorläufig theoretisch nicht ableitbar sind.

Eine wichtige Tatsache ist noch festzustellen. Ein Text galt lange als eine mit Hilfe von Regeln erzeugte Entität, in der nur der Stil für Variabilität sorgte. Die Tatsache, daß man seine Eigenschaften mit stochastischen Prozessen modellieren kann, eröffnet der Forschung neue Möglichkeiten.

Eine andere, kombinatorische Art der Modellierung bringt in einigen Arbeiten Zörnig (1984 und in diesem Band).

LITERATUR

Altmann, G., Phonic structure of Malay pantun.
Archiv orientalni 31, 1963, 274-286

Altmann, G., Some phonic features of Malay shaer.
Asian and African Studies 4, 1968, 9-16

Brainerd, B., On the Markov nature of the text.
Linguistics 176, 1976, 5-30

David, F.N., Two combinatorial tests of whether a sample has come from a given population.
Biometrika 37, 1950, 97-110

Herdan, G., The calculus of linguistic observations.
's-Gravenhage 1962

Hřebíček, L., Euphony in Abay Kunanbayev's Poetry.
African Studies 1, 1965, 123-130

Johnson,N.L., Kotz,S., Discrete Distributions.
New York, Houghton Mifflin 1969

Knauer, K., Die Analyse von Feinstrukturen im sprachlichen Zeitkunstwerk. In: Kreuzer, H., Gunzenhäuser, R. (Hrgs.), Mathematik und Dichtung, München 1965, 193-210

Köhler, R., Folgen von distinktiven Merkmalen als Markov-Ketten.
Bochum 1980 (Diss.)

Miller, G.A., Chomsky, N., Finitary models of language users.
In: Luce, R.D., Bush,R.D., Galanter,R.R. (Hrgs.), Handbook of mathematical psychology. New York, Wiley 1963, 419-491

Sebeok,T.A., Zeps, V.J., On non-random distribution of initial phonemes in Cheremis verse. Lingua 8,1959,370-384

Skinner, B.F. The alliteration in Shakespeares's sonnets:
A study in literary behavior. Psychological Record 3, 1939

Skinner, B.F., A quantitative estimate of certain types of sound patterning in poetry. The American Journal of Psychology 46, 1941, 64-79

Zörning, P., The distribution of the distance between like elements in a sequence I. In : Boy, J. Köhler, R. (eds.) Glottometrika 6, Bochum, Brockmeyer 1984, 1-15

FRENCH INFLUENCE ON VIETNAMESE ENGLISH

An Experimental Investigation of the Effects of French Transfer on the Orthographic Recognition and Production of the English Lexicon by Vietnamese Speakers

Jan M. Ulijn, Eindhoven
Susan J. Wolfe, Santa Cruz
Adele Donn, San Francisco

Vietnamese and other Indo-Chinese immigrants to the United States often know some French before they try to acquire English. This case of trilingualism is considered psycholinguistically and linguistically, and some hypotheses are formulated about lexical transfer processes in unrelated/related language settings. The hypotheses are experimentally tested by havin g 88 Vietnamese subjects, recent immigrants to the United States with minimal English knowledge, read Vietnamese sentences and demonstrate English and French knowledge of a target word within each sentence through a translation task. Data show that for cases of French knowledge, transfer effects of lexical items seem to overrule the effect of word knowledge in that subjects demonstrated better English knowledge of English-French cognates than formal contrasts, and better English knowledge of formal contrasts than misleading cognates. For cases of no French knowledge, however, a significantly different pattern was found in that subjects identified the target words as well as the French knowledge group and then they even performed significantly better on the formal contrasts than those with French knowledge. No difference was found between recognition and production. Conclusions are drawn regarding the lexical transfer process and possible implications of strategies for teaching English to native Vietnamese speakers, both for people with and without French knowledge, are considered.

Jan Ulijn teaches psycholinguistics and technical communication theory at Eindhoven University of Technology, Department of Applied Linguistics, in The Netherlands.
Susan Wolfe, Program in Experimental Psychology, University of California, Santa Cruz, is currently a Human Factors Engineer, investigating human/computer interaction issues for General Electric Company in the United States.
Adele Donn, Department of English, San Francisco State University, is currently teaching ESL at San Francisco Community College in the United States.

VIETNAMESE IMMIGRANTS TO THE U.S. AND THEIR FRENCH BIAS: HOW TO TEACH THEM ENGLISH

In the past 15 years, several hypotheses have been formulated about foreign language acqusition processes in relation to the mother tongue: (1) the contrastive analysis hypotheses (CAH) in its strong or weak variations about the role of the mother tongue in foreign language use (contrasts with the native language on a more linguistic basis should predict, or at least explain, psycholinguistic difficulties in learning a foreign language); (2) the interlanguage hypothesis (ILH) which proposes that a L2 learner builds up a separate linguistic system (interlanguage) that is affected by both L1 and L2 elements, but differs from those in systematic ways reflected by specific errors; and (3) the identity hypothesis (IDH) which suggests that L2 acquisition is identical (L1 = L2) or similar (L1$\approx$L2) to L1(cf. Bausch & Kasper, 1979, McLaughlin, 1982). These hypotheses have been mostly examined in rather exploratory studies on the syntactic level (CAH, IDH and ILH) and on the morphological level (ILH and IDH), and very little attention has been paid to the lexical level (that is , lexicon in the sense of content words, and less in the sense of function words which operate mostly on the syntactic level). Within the framework of the interlanguage hypothesis, there is an exception in the experiments by Levenston and Blum (1977) and Blum and Levenston (1983) on the lexical simplification of English acquired by native Hebrew speakers.

In recent books on second language acquisition (Hatch, 1978; Krashen, 1982; and Pugh & Ulijn, 1984) and second language reading (Mackay et al., 1979), some attention is paid to the lexicon, but hardly any to lexical contrasts with the first language (not to be confused with the use of the first language through translation). In none of these publications is there mention of cases where three or more languages are involved (as

is the case of Vietnamese refugees coming to the U.S. with some knowledge of another language, such as French).

Only in the last few years has there been an increasing interest in the teaching of the English lexicon. This is reflected in contributions to the TESOL Quarterly and Proceedings of TESOL Conventions (Martin, 1976; Richards, 1976; Judd, 1978; and Rivero &Best, 1978). In none of these cases, however, is there focus on lexical contrasts with other languages. There are a few recent exceptions, however, particularly for a related language pair like English-French, which is of specific relevance to Indo-Chinese immigrants. In 1976, Hammer and Monod published an English-French cognate dictionary (cognate meaning words with the same or similar form in two languages with the same meaning, such as the French and English word orange). In 1979, Hammer also contributed an extensive literature survey on cognates and misleading cognates (same or similar form with different meaning, like (French car = English bus) $\neq$ (English car = French voiture): Her study, dating back to 1920, deals basically with English, but also with English-Spanish and English-German. She pointed out a void in the literature, however, dealing with misleading cognates as they relate to negative transfer. Recent attempts to teach misleading cognates in the ESL classroom are from Wilcox and Lehman (1980), for English-Spanish and English-French speakers.

As far as we know, until now, there have been no studies investigating what happens with immigrants have a native language completely unrelated to English (i.e. Arabic, Japanese, Chinese, Vietnamese) and already know another European language related to English (i.e., French, German or Spanish). Considering the fact that a massive number of Indo-Chinese immigrants have come to the U.S. (300,000 in January 1980, and more have continued to arrive), this question is important for developing strategies for teaching English to such people.

The Case of the Vietnamese: What Exactly is Their French Bias?

There were two important waves of Vietnamese immigration to the U.S. and other countries: one that started in 1975 with the reuniting of Vietnam under one communist regime, and one that started in 1979 with the expulsion of ethnic Chinese from Vietnam. There existed an educational difference between the two groups, with the education level of the first being, on the average, higher than that of the second. Consequently, the first group was exposed to more French than the second. Table 1 gives an indication of the structure of the foreign language instruction in the Vietnamese education system. University education is excluded, since this represents a very restricted population.

Table 1

Foreign Language (French and English) Education in Vietnam

School Period	Language Instruction in Primary School (PS) and Secondary School (SS)
Before 1945	PS and SS: curriculum in French SS: English language teaching included
1945 - 54	PS: curriculum in Vietnamese with French language taught 10 hours/week for 5 years SS: curriculum in French with French language taught 24 hours/week for 7 years and English language taught 3 hours/week for the last 3 years
1954 - 75	PS: no foreign language instruction SS: English or French taught 7 hours/week for first 4 years and both French and English taught 4 hours/week during last 3 years
1975 - 79	PS: no foreign language instruction SS: English or French taught 3 hours/week

(Information collected from Vietnamese teachers of foreign languages and immigrants to the U.S.)

From 1862 to 1945, the Vietnamese education system was heavily influenced by French cultural domination (Duong, 1978). Chinese or French language and culture was emphasized, rather than Vietnamese. In addition, literary education was stressed to the detriment of vocational and technical education, and memorization and imitation were encouraged, at the expense of initiatives in problem solving. Between 1954 and 1975, the country was split into a communist Northern part and an American dominated Southern part. We refer, here, to the situation in the Southern part. Immigrants report that there was only French, and no English, teaching in the North during this period. Finally, since 1954, all curricula are in Vietnamese. Reports about the situation since 1978 are somewhat confusing. There has been only one foreign language taught in some cases, with Russian, English, French, Chinese or Japanese being learned at home, with private tutors.

It is, therfore, a fair assumption that immigrants educated before 1954 learned a considerable amount of French. Those educated later probably have some knowledge of French as well as English (except those educated at private French institutions, who would have a far better knowledge of French). Refugees without secondary school education have no French knowledge. This category is highly representative of the second wave of immigration. Personnel at Vietnamese resettlement centers in the San Francisco Bay Area estimate that 10 % of the Vietnamese adult refugees (older than 18 years) have learned French. In absolute terms, this figure warrants paying particular attention to the French bias of Vietnamese immigrants in their acquisition of English.

This is of particular interest since English and French share 10,993 cognates which should be facilitating, but also 950 misleading cognates which could cause considerable interference (Hammer, 1979). It is interesting to note, therefore, that in their manual for Indo-Chinese refugee education, Grognet et al. (1976) do not list the French bias among the sources of difficulty or facilitating factors in learning English. While

pronunciation and grammar problems are discussed, vocabulary is not.

Further evidence about the extent to which French interferes or aids Vietnamese immigrants in acquiring English may show the way to more efficient teaching of such people. In addition, this knowledge may have implications for teaching English to other people with a good knowledge of the French language.

Psycholinguistic Aspects of the Problem: The Lexicon and Conceptual Systems in Bilingualism and Second Language Acquisition

As we wish to examine the role of the lexicon with respect to several languages in the human brain, theories on the specific nature of the lexicon in bilingualism, trilingualism, etc., are particularly relevant. We are deliberately leaving out theoretical proposals about the general lexicalization processes (see, for instance, the models of Morton, 1983; and Forster, 1976), but we will try to account for the effects of factors which play a role in relating words to concepts in reading, or relating concepts to words in writing. High frequency, recent usage, context, semantic apparency with other well known concepts, and conceptual lexical knowledge play roles in reading and writing. Specifically related to reading are homonymy and polysemy, and to writng, synonymy and phonological accessibility (Clark & Clark, 1977; and Schreuder & Levelt, 1978). Additionally, the notion of context, allowing the reader to infer the meaning of a given word, is important (see Mosenthal et al., 1978, for a proposal on a multicontextual model of word recognition.

All of these factors have an impact on the functioning of the lexicon and conceptual system in any language (L1, L2, L3 etc.).

This discussion will focus more on the links between several languages when they merge in the brain while reading and writing L2 or L3.

What exactly is bilingualism? Elsewhere (Ulijn, Wolfe, & Donn, 1981), we have defined the different types as compound, coordinate and subordinate and we have reviewed the available experimental evidence of how the conceptual system and the lexicons are organized in typical cases of bilingualism (see also Baetens-Beardsmore, 1982, and Grosjean, 1982, for more recent and extensive surveys). Here we simply suggest that two lexicons should be necessary, though they can be interdependent, and one single conceptual system should be sufficient to account for a brain in which two languages are operating. Separate lexicons for each langugage seem to be needed in order to avoid constantly mixing the languages. We disagree with Nas (1983) who states that compound bilinguals should have a common lexical store for two languages. His experimental data do not convincingly support the idea, since they are gathered out of sentence context (the same holds true for the works of Kerkman, 1981, and Kirsner et al., 1980). There is , however, plausible experimental evidence for the same conceptual system for all languages. Rosenzweig (1962) compared word-association responses in English, French, German, andItalian, and found that, across languages, similar associations tend to occur among words with similar meaning. Steinberg (1975) supplied evidence that speakers of Chinese, Finnish, Japanese and Slovenian share the same conceptual system for determining semantic sentence interpretations.

Of course, a given language is an important means of expression of culture and, therfore, not all concepts are shared by all languages. This explains the possibility of semantic shifts in bilinguals, as shown by Ervin-Tripp (1961), in color-naming by Navajo-English bilinguals (as Navajo and English speakers use different color systems). Normally, humans will have many more

shared concepts than differing concepts. Therefore, how can we account for different response patterns in word association (cf. Lambert, Havelky & Corsby, 1972; and Heuer, 1973)? The number of entries in the lexicons and the conceptual system are unstable, as are their inner organization.
Words and concepts go in and out as a function of learning and forgetting. The manner of learning will probably determine how concepts and words are stored. This, in turn, will determine the way the concepts will be retrieved. Therefore, what can be retrieved from the conceptual system through Lexicon 1 (and vice-versa) may be different in organization from what Lexicon 2 allows to be retrieved from the same conceptual system. In some bilinguals, however, the organisation may be the same for both languages, as shown the free recall experiments of Young and Wavar (1968).

Which type of bilingual situation do we have when a Vietnamese adult, whose conceptual system is almost finalized , comes to the U.S.? In this case, which is the focus of the present experiment, English acquisition will reflect an English lexical re-labeling of known concepts by inner translation (a direct relationship to Lexicon 1 is established). Some Vietnamese concepts will be specific for that culture and nonexistent in English (semantic voids), and others will be acquired with the English label, although they may exist in the Vietnamese society (for instance, technical concepts learned in the U.S.). Figure 1 illustrates this relationship. This could be considered coordinate bilingualism. As long as English experience develops, the direct relation between Lexicon 2 and the conceptual system gets stronger, and the relation between Lexicon 1 and Lexicon 2 gets weaker. The detour via Lexicon 1 (translation) is no longer necessary. Consequently, if this transformation process is strong enough, coordinate bilingualism could turn into compound bilingualism.

Here, translation back and forth through Lexicon 1 provides a clear delineation of concepts, which is important for effective communication in the second language. This could be the reason learners systematically avoid words that have no equivalent in their native language (Blum & Levenston, 1983). It may also elucidate why teaching new words when explained by native language equivalents appears more effective than teaching by eyplaining in the target language.

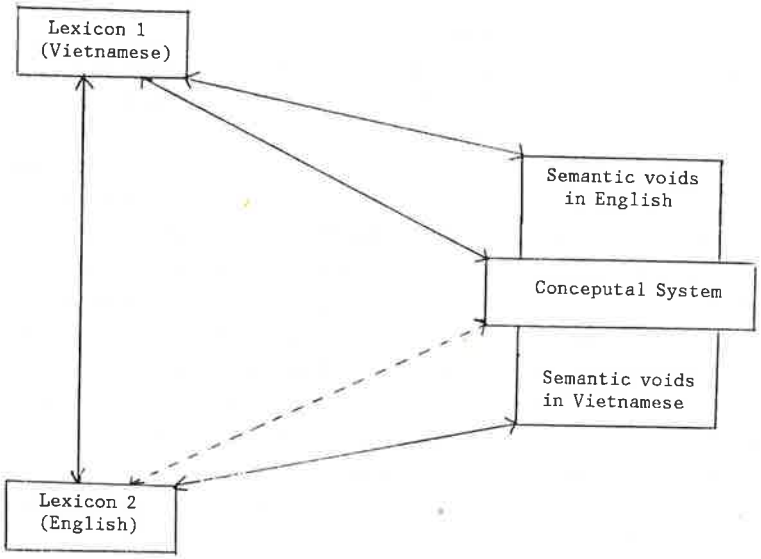


Fig. 1: Relationship between Lexicons and Conceptual Systems (with Semantic Voids)

Moreover, in the assessment of bilinguals, it makes sense to differentiate between perception (comprehension or recognition) and production to see whether the lexicalization procedures in reading are different from those of writing. It is also important to know what happens when more than two languages are involved, and to what extent the relations between the conceptual system and lexicons depend on the degree of similarity of the languages involved. A discussion of these issues follows.

Acquisition of Several Non-Native Languages and the Lexical Transfer Processes

What happens to the lexicons and the conceptual system when more than two languages co-occur? As with bilinguals, effective multilingualism depends on many factors such as simultaneous or successive acquisition and the age at which the languages are acquired. One of the main factors seems to be similarity of the acquired languages. Particularly in the case of related languages, the tendency to transfer from L1 to L2 or from L2 to L3 seems to be very strong: positive if the similarity is actual, negative if the similarity is misleading. The definition of transfer and interference varies a lot in the literature. For purposes of discussion, positive transfer leads to correct L2 and L3 use, while negative transfer or interference leads to incorrect use of L2 or L3 (for a terminological discussion, see Rattunde, 1971). Kellerman (1978) argues that the transfer of L1 items to L2 expressions is an active learning strategy depending on the learner's notion of the distance between L1 and L2. Some L2 items are more likely to be transferred than others to the extent that they are believed to be less native language specific. This is demonstrated in an experiment where Dutch subjects had to judge the acceptability of English or German (L2) polysemous lexical items (English break / German brechen / Dutch breken). The core meaning of these lexical items shared by L1 and L2 was more likely to be transferred than more idiomatic and figurative meanings.

In the case of three languages, there are at least four grossly simplified combinations of possible relatedness, as follows:

1. L1 -u- L2 -u- L3
2. L1 -r- L2 -r- L3
3. L1 -u- L2 -r- L3
4. L1 -r- L2 -u- L3

where u = unrelated and r = related. Two other combinations, in which even L1 could suffer, are instances of a backwash effect which we do not consider in this paper:

5. L1 -r- L3 -u- L2
6. L1 -u- L3 -r- L2

Very little literature is available about the lexical transfer processes involved with Combinations 1, 3, and 4, while there is some information regarding Combination 2. The learning of foreign languages in The Netherlands is a case in point, as nearly every secondary school student learns English, French andGerman almost simultaneously. Although it is hard to define L2, L3 and L4 in chronological order, incidences of interference are interesting. Ickenroth (1976) argues that when a particular L1 student (Dutch) does not know a L2 word (French), for instance, one of his "escape" routes is adapting a word from L1 or L3 (English), etc. He would provide the Dutch koper / English copper with a French pronunciation kopêr. Knibbeler (1977) surveyed French errors made by Dutch adults with some school knowledge of French, English and German. He found some lexical interference from English and not as much from German, which is much less related to French than English is . (Dutch speakers normally are more proficient in English and German than French, as these languages are more closely related to Dutch than is French. Furthermore, Dutch speakers usually have more exposure to English and German than to French). For the past ten years, the same lexical

transfer and interference has been noted in French spoken and written by Dutch engineering students in French courses at Eindhoven University of Technology (Ulijn, personal observations). On the basis of English compositions written by 187 bilingual students, LoCoco (1976) concludes that the less equally balanced bilingual learner tends to rely more on his stronger language, while the more equally balanced bilingual learner relies less on the previously acquired languages (and this reliance appears evenly distributed over German and Spanish). Since German shares the germanic lexical layer with English and Spanish shares the roman one, there is probably equal transfer from both.

What could be predicted, then, for a Vietnamese speaker who knows some French? This falls into our third condition of relatedness, as Vietnamese (L1) is unrelated to French (L2) which, in turn, is related to English (L3). Since unrelatedness implies no transfer, neither negative transfer nor interference can take place. The second language, French, has to be acquired as a distinct language different from the first language. In learning the third language, English, which is related to French in lexical form and unrelated to Vietnamese, the learner would not have any help or interference from Vietnamese, but could from French. The relation between the conceptual system and the lexicons is diagrammed in Figure 2 (for simplification, the semantic voids are omitted).

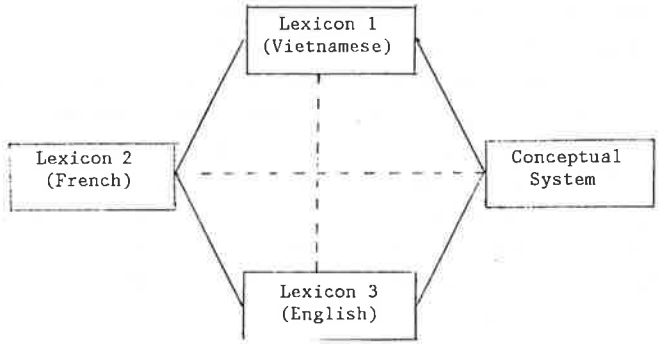


Fig. 2: Three Lexicons and Their Relationsship to the Conceptual System (with shortcuts expressed as dotted lines)

Lexicon 3 -- Conceptual System would be possible at latter stages of increasing compound trilingualism. This L2 - L 3 interference rather than L1 - L3 interference seems to correspond to anecdotal and experimental evidence on the phonological level and other levels (lexis included). A Zairese speaker with perfect knowledge of Tshiluba (L1) and French (L2) would speak English (L3) with a French accent (Ferguson, Note 1). Similar L2 - L3 interference, where L1 is related to both L2 and L3, has been experienced by one of the present authors (Donn), where she would transfer the French "r" (L2) and not the English "r" (L1) to Spanish (L3). Chumbow (1981) reports on three studies of trilingualism (we do not consider pidgin English as a separate language as he does). They show that there is much more transfer from L2 to L3 than from L1 to L3, when L2 and L3 are related languages (English and French) and when L1 is unrelated to either L2 or L3. Error analyses were performed in trilingual situations in West Cameroun (L1 = Ngemba, L2 = [pidgin] English and L3 = French) on in phonological level; in Western Nigeria (L1 = Yoruba, L2 = [pidgin] English and L3 = French) and in Cameroun (L1 = Bulu, L2 = French and L3 = English) on other levels (lexis included).

Linguistic Aspects of the Problem and Types of Lexical Contrasts between Vietnamese, French and English

In order to predict what happens when a Vietnamese speaker tries to transfer knowledge to English lexical recognition and production, we need a framework to adequately describe lexical contrasts. After having reviewed several techniques of constrastive analysis available in the literature, Ulijn (1978) proposed a schema

where both conceptual structure and linguistic form of words vary. A total of 12 types of contrasts can be reduced to three categories.

1. Cognates (CS): words which have the same or similar spellings (form) and the same meanings (concept) in two or more languages, regardless of their origins. For example, English table and French table, where L1 = L2 in both concept and form.

2. Formal Contrasts (FCs): words which have completely different spellings and the same meanings. For example, English window and French fenêtre, where L1 = L2 in concept but L1 ≠ L2 in form.

3. Faux-amis, Deceptive or Misleading Cognates (MCs): words which have the same or similar spellings but the possibility of a different meaning in two or more languages (that is, the same or similar forms are not exact translation equivalents in two languages and they represent two different concepts) (note 2). For example, English pain = French douleur while French pain = English bread. In this case, L1 ≠ L2 in concept but L1 = L2 in form. The misleadingsness, of course, is a hypothesis rather than an assumption. Note that the terms concept and meaning are interchangeable in this discussion. For further discussion on possible definitions, see Carroll (1964), Ulijn (1981) and the recent literature on linguistic semantics.

A brief discussion should be made of some of the details of the specific languages involved. Vietnamese is a tonal language spoken by 50 million people in Vietnam. The Northern dialect has six tones, and the Southern has five. Vietnamese is one of the rare languages in southeast Asia which uses a Roman letter alphabet (quôc ngù) developed by 16th century Spanish missionaries to replace native Chinese characters. The tones are denoted by diacritical marks. Phonologically, syntactically and lexically, Vietnamese is unrelated to both French and English. There are some loan words from French, because of a century of French domination, but their spelling is often drastically modified (in the Vietnamese-English dictionary of Nguyen Dinh Hoa, 1971, French loan

words are specifically noted). Before 1954, the year of the French withdrawal from Vietnam, there was also a French pidgin (Tây Bôi) which is no longer spoken (Phillips, 1975). Reinecke (1971) states that remarkably few Vietnamese words were used in this pidgin. French loan words and pidginized French words that are widely known by Vietnamese speakers are so modified from original French that the Vietnamese speakers have to master French as a completely different language from Vietnamese.

For historical reasons, French and English have a strong lexical relationship even though English is a Germanic language and French is a Romance language. As Hammer (1979) points out, the number of cognates and misleading cognates between French and English is three times more than between other related language pairs such as English/Spanish and English/German. Some of these occur in more complicated relationship (for example, English gas = French essence and gaz, while gas and gaz are cognates, gas and essence are misleading cognates). However, the three basic lexical relationships (Cs, FCs, MCs) seem to occur most frequently between French and English.

What are the origins of the misleading cognates in the borrowing process between two languages? They may reveal some general underlying cross-linguistic phenomena when either language is used in a multilingual setting.

At one time in the history of a language, a transfer takes place from one language to another as a cognate. These words, however, are often used in the new linguistic community for other specific needs without consulting the original linguistic community. This can be particularly harmful in the area of technical and scientific terminology, where misleading cognates between languages could create serious misunderstandings. After World War II, new technical terms were standardized in international committees to avoid divergent uses of such terms in different languages (Ulijn, 1979). Still, not all human activities are covered, as French éditeur = English publisher, but English editor = French rédacteur, just to mention one example.

A similar phenomenon can often be observed in the development of pidgins. Grass, in Neo-Melanesian pidgin, would be anything that grows outward from a surface in a blade-like shape, and this would replace other original L1 words (Hall, 1966). In this meaning, grass becomes a misleading cognate of the English word with the same spelling.

The creation of misleading cognates between languages seems to be a very natural process in bilingual development at the community level, as well as on an individual basis. When people acquire a new language after their native or other languages, they must be aware of these formal and conceptual similarities, particularly in cases where the languages are as similar as French and English are.

What is the relative difficulty of cognates, formal contrasts and misleading cognates between existing and new languages when acquiring a new language? Previously, we mentioned some evidence about general English-French interference. There is some specific experimental evidence pertaining to the distinctions between Cs, FCs, and MCs (albeit not in the Vietnamese-French-English situation). In the Shadok project (Ulijn, 1981), Dutch engineering students appeared not only to be hampered by misleading cognates with the L1 (Dutch), but also with their L2 (English), when reading L3 (French). Misleading cognates were significantly more difficult than Dutch-English-French cognates, such as in international technical terms. In continued word association tasks with cognates and formal contrasts, by French-English bilinguals, there were more similar response words to cognates than to formal contrasts (Taylor, 1976). Relationships within the conceptual system are, therefore, more obvious with cognates than with formal contrasts. In a study on lexical errors in English composition written by Finnish and Swedish secondary school students in Finland, both groups of students made a large number of errors on Swedish-English misleading cognates (Wikberg, 1979). Young Swedish-speaking Finns (L1 = Swedish), being part of the minority, know considerably more Finnish than the Finnish speakers (L1 = Finnish) know Swedish. This is due to both the bilingual environment and to school conditions. The

Finns (L2 - L3 transfer) made even more errors than the Swedes (L1 - L3 transfer) with 11.4 versus 8,4 % of the total number of errors for each population. Ringbom (1981) found that the same categories transferred more from Swedish to English than from Finnish to English (cf. Ulijn, forthcoming, for other trilingual studies). The situation of the Finns is highly comparable to the Vietnamese case, as Finnish (L1) is unrelated to Swedish (L2), which is, in turn, related to English (L3). Our attention should now focus on what predictions can be made, partly on the basis of these studies, about the transfer of French/English Cs, FCs, and MCs on English lexical recognition and production by Vietnamese speakers with knowledge of the particular French words.

HYPOTHESES ON THE RECOGNITION AND PRODUCTION OF ENGLISH LEXICON BY NATIVE VIETNAMESE SPEAKERS WHO HAVE KNOWLEDGE OF FRENCH

Since Vietnamese is unrelated to both French and English, we posit no transfer to these languages (an assumption which we have not tested in the present experiment). We, therefore, limit our hypotheses to the transfer effect of French/English Cs, FCs and MCs to English lexical recognition and production by Vietnamese who know French.

When a Vietnamese speaker attempts to understand, in some context, an English word that is a C of the French word, and the person knows this French equivalent, earlier experience with the relatedness of French and English will invite a guess of the meaning of the English word, and the guess will be correct. If the word is an FC of the French word, the Vietnamese has no particular strategy for guessing the correct meaning, and the accuracy of the guess is merely a function of the person's English word knowledge (no transfer is possible). In the case of MCs, the person encounters a word which appears to have the same or similar form of a familiar French word, and he is led on the wrong track, which leads to an incorrect notion in the conceptual system.

In English production, the opposite approach is taken. If a Vietnamese speaker does not know the particular English words, among other escape routes and communication strategies, he could make an attempt to transfer the French form to English. This will yield correct results where in the case of Cs, but incorrect in the cases of FCs and MCs. After some experience with English forms, the person is likely to notice the French form for FCs and will need to develop other means for producing such English words. As in the case of English recognition, MCs are the most problematic. The production of an English MC appears obvious, given knowledge of a French word with similar form. However, this leads to an incorrect response.

These three types of lexical relationships are diagrammed in Figure 3. In the case of Cs. the transfer is positive. In the case of MCs, it is negative. To clarify this last case, consider the following: If a Vietnamese speaker has to recognize the English car, the speaker will think that it is a cognate with the French car and will, in doing so, reach for an incorrect notion in the conceptual system. As English bus ≠ French voiture, an incorrect semantic interpretation will be made. If the same speaker has to produce English bus or coach and knows the French word car, this also will yield the expression of a concept other than what was intended (English car = French voiture ≠ French car).

Our hypothesis is that Vietnamese speakers who know the particular corresponding French words, produce significantly fewer errors in Cs than in FCs, and significantly fewer errors in FCs than in MCs, both in the recognition and production of the related English words. In cases where the corresponding French word is not known, the number of correct guesses in recognition and production will simply be a function of English word knowledge. Such words should manifest a different response pattern from those influenced by French knowledge, at least for those Vietnamese speakers: in the initial stages of English language acquisition. The following experiment assesses this hypothesis.

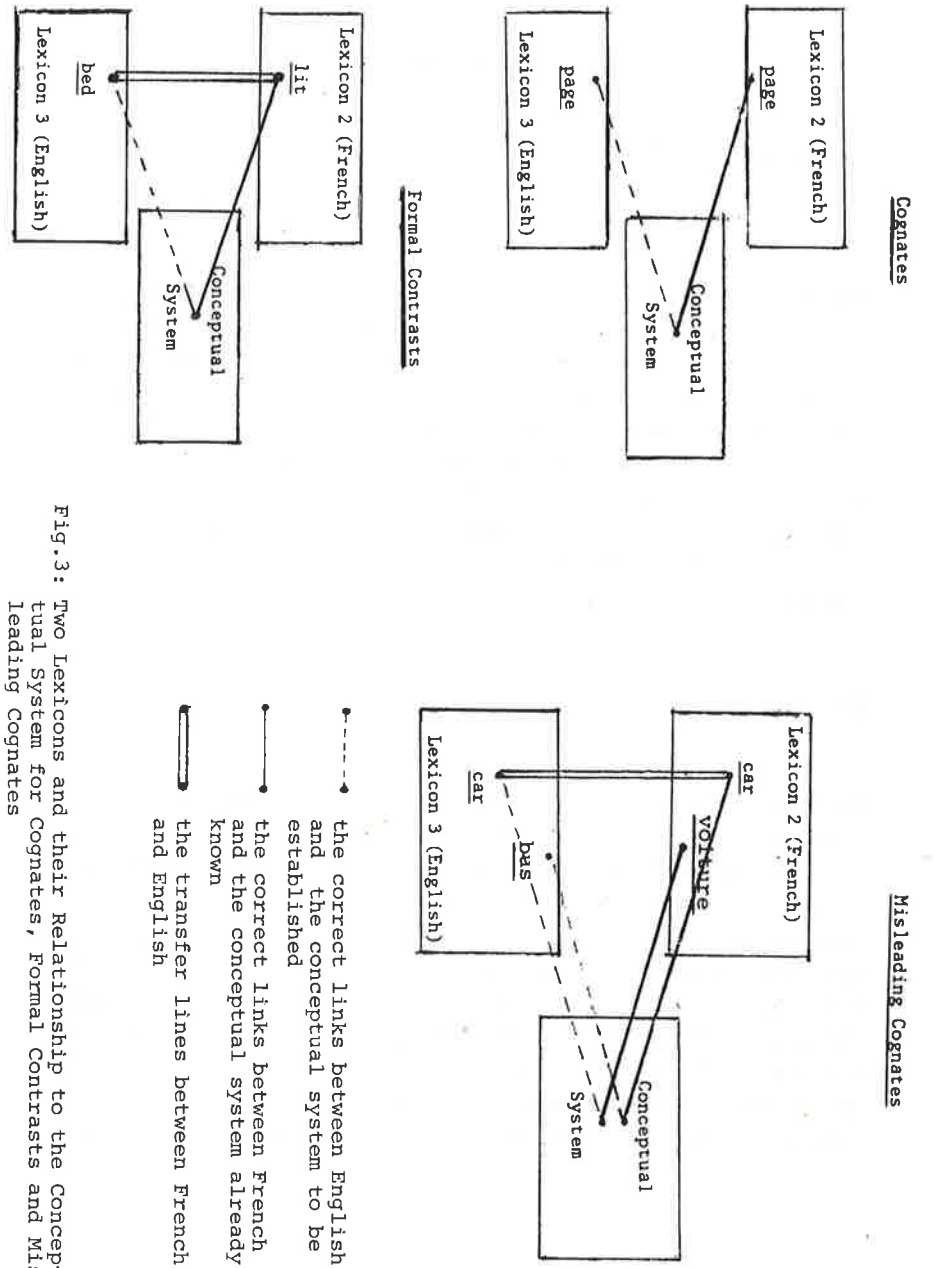


Fig.3: Two Lexicons and their Relationship to the Conceptual System for Cognates, Formal Contrasts and Misleading Cognates

METHOD

SUBJECTS

The subjects were 54 male and 34 female recent Vietnamese immigrants to the United States, with little English knowledge. Their average age was 30.06 years (with a range of 17 - 63 years) and they had been in the country for an average of 6.54 months. For those knowing French (about 20 %), the average length of French study was 6.35 years. Contact was made with the subjects via various Indo-Chinese resettlement programs and elementary English classes in the San Francisco Bay Area, and subjects were asked to participate by administrators in the various institutes. Subjects were randomly assigned to one of two conditions.

MATERIALS

Sixty nouns were selected (20 each of Cs, FCs and MCs) on the basis of frequency counts of French (Mackey et al., 1966; Gougenheim, 1967, and Juilland, 1970) and English (Carroll et al., 1971; Wei & Light, 1973, and Abbas, 1979) to achieve comparable frequencies for each of the three lists of Cs, FCs and MCs. Both a Vietnamese-English phrase book (Duong Thanh Bihn, 1975) and an English-Vietnamese phrase book (Nguyen Hy Quang, 1975) were consulted. Various sources were also consulted for the French/English cognates (Hammer, 1979), formal contrasts Mackey et al., 1966), and misleading cognates (Boillot, 1930; Anderson & Harmer, 1938; Seward, 1947, and Koessler, 1975).

A problem with the selection of misleading cognates, however, was noted in that there is almost always a slight overlap of meaning in the pairs of words (e.g., anniversaire - anniversary). Cases such as coin - coin, with a complete difference in meaning, are rare. Our previously stated definition of MCs, that the two words are not exact translation equivalents, holds. Preference was given to high frequency "survival" concepts which would be important to recent immigrants. Additionally, obvious semantic voids in either Vietnamese or English were excluded. These 60 nouns were eventually reduced to 30 (ten of each type of lexical contrast), and were each embedded into a sentence which allowed for a natural lexical recognition and production task, avoiding situations where context clearly gives away the meaning of the word. (See Appendix A for a list of the 30 nouns used).

The definitive test consisted of one of two ransomly selected orders of the 30 sentences (see Appendix B). Within each sentence, the target (underlined) word appeared in either Vietnamese or English. The target word in each sentence was to be translated into English and French when appearing in Vietnamese and into Vietnamese and French when appearing in English. It was randomly decided which of the cognate target words would appear in Vietnamese and which in English, and likewise with the formal contrasts and misleading cognates. Those target words which appeared in English in the first random order of the test appeared in Vietnamese in the other order, and vice versa. No target word appeared in both English and Vietnamese within the same test form.

Answer sheets were provided, with the column heads of "English", "Vietnamese" and "French" (written in Vietnamese). For each test item, the box corresponding to the inappropriate language was crossed out (e.g., when the target word appeared in English, the English box was unavailable). This insured both that the subjects indicated their responses in the appropriate column and that they understood in which language the target word appeared. In all cases the French column was left free, as the

subjects were always asked to indicate the French word, if it was known.

Subjects participated in the experiment in variously sized groups. Once assembled, subjects were randomly assigned to one of the two conditions, given a test booklet containing written instructions in Vietnamese, a biographical questionnaire, one of the two test forms and its corresponding answer sheet. Time was allowed for the subjects to read the instructions (explaining the translation task, filling in the "French" column whenever possible, and guessing when unsure, as well as stressing that accurate spelling was not crucial, emphasizing the need to work independently, and reassuring the subject that their results on the task would have no bearing on their ability to find job in the U.S., etc.). Additional time was allowed for questions, and a Vietnamese speaker was available throughout the session to answer any questions.

Subjects were given a maximum of one hour to work on the task. Upon completion, subjects returned the test booklet and answer sheet and were thanked for their participation.

DESIGN

A 3 X 2 factorial design with repeated measures was used, with three levels of target words (French-English cognates, formal contrasts or misleading cognates) and two levels of language presentation of target words (target words either in Vietnamese or English in each sentence, so that the task is either English recognition or production).

RESULTS

In order to analyze the results, an assumption was made by the authors. It was decided that English knowledge was demonstrated whenever the subject gave a correct response in English or Vietnamese. This is obviously the case when a subject gives the correct response in English (production), but it is also the case when a correct response is given in Vietnamese, as the subject has to have read an English target word in order to provide the Vietnamese (recognition). Table 2 provides an error analysis of each of the 30 words, broken down according to the accuracy of the English word given knowledge or no knowledge of the French equivalent.

The major analysis of the data, therefore, interprets the results by looking at the three types of lexical contrasts (cognates, formal contrasts and misleading cognates) over two levels of English knowledge (where the French equivalent was either known or not known). Percentages of the total number of correct responses for each cell are shown in Table 3. Figure 4 graphically represents the results, illustrating both the recognition and production aspects of the effect of French knowledge on learning English words.

There is a statistically significant effect of type of lexical contrast when French knowledge is demonstrated, with French-English cognates known at a higher rate than formal contrasts, which are known at a higher rate than misleading cognates (F = 29.19, df = 2/27, $p < .001$). A multiple comparison test shows that all levels of this factor are significantly different from each other (Newman-Keuls: r = 3, crit. diff. = 7.76; r = 2, crit. diff. = 6.43). Without French knowledge, a significant effect of the type of lexical contrast is observed (F = 9.38, df = 2/27, $p < .001$). This significant effect is attributed to

the high percentage of correct translations of formal contrasts, however, as evidenced by a multiple comparison test (Newman-Keuls; r = 3, crit. diff. = 7.76; r = 2, crit. diff. = 6.43). A check was made on the distinction between production (translating Vietnamese into English) and recognition (translating English into Vietnamese), and the data suggest no significant difference between the two (F = 1.42, df = 1/58, $p < .5$).

Table 2

Error Analysis for Cs, FCs and MCs (88 observations of each word)

	French Knowledge		Non-French Knowledge	
	correct	incorrect	correct	incorrect
Cs				
page	36	3	28	21
village	32	6	18	32
fruit	35	0	25	28
restaurant	37	0	28	23
table	39	1	42	6
age	36	3	34	15
cousin	24	7	22	35
million	31	3	26	28
police	37	1	43	7
prison	24	6	18	40
FCs				
bird	28	0	35	25
foot	32	6	18	32
street*	29	0	45	8
day	26	0	44	18
water	33	2	44	9
window	27	2	54	5
bed	22	3	45	18
pen*	16	2	52	12
door	25	1	53	9
wall	23	3	37	25
MCs				
car	31	0	55	2
cave	1	7	7	73
coin	9	4	24	51
fabric	8	1	14	65
crayon	16	3	28	41
patrons	13	4	19	52
lecture*	8	3	19	52
library	15	3	48	22
carts	9	2	29	48
anniversary	12	1	23	52

*number of observations = 82

Table 3. Percentage of Total Number of Correct Responses Given English Knowledge

	French Knowledge	Non-French Knowledge
Cognates	37.61	32.28
Formal Contrasts	28.67	51.72
Misleading Cognates	13.93	30.39

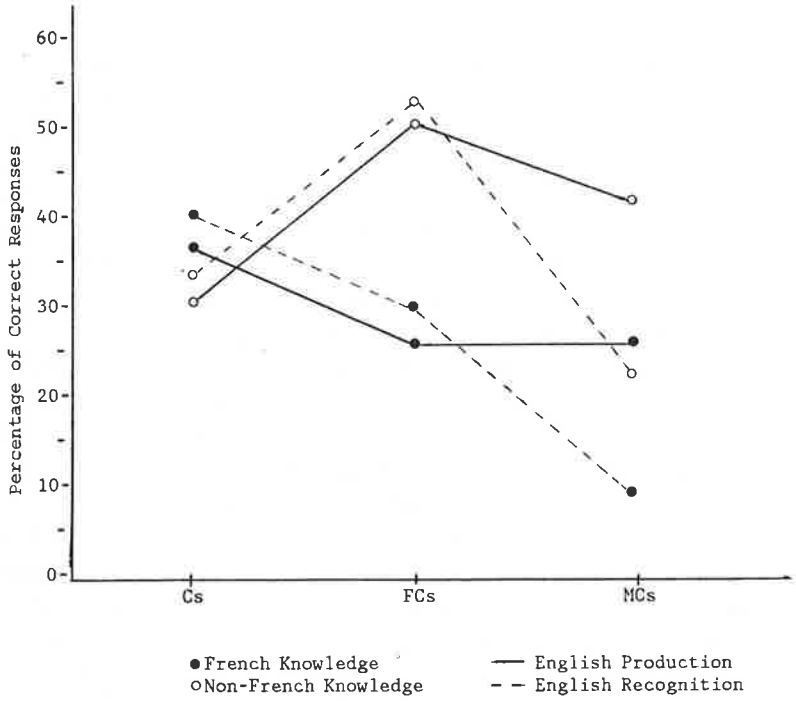


Fig. 4: The Effect of French Knowledge Across Lexical Contrasts on English Recognition and Production

DISCUSSION

The data appear to support the hypothesis that the transfer effect of items from French to English overrules word knowledge to some extent. Using 30 common words, the case where transfer effects are beneficial (Cs) produces the most correct translation. The case where transfer effects serve no useful purpose (FCs) produces fewer correct responses, and the case where transfer effects produce error (MCs) yields the worst performance.

The fact that significant results were also obtained for the non-French knowledge group does not weaken the theory. The multiple comparison test verifies that the results are statistically significant due to the FC group only. The high percentage of correct translations of FCs in the case of non-French knowledge could be due to the fact that theses words are more "basic", contrary to the frequency counts. However, as there is no systematic reason why FCs should be significantly easier than any others (see Appendix A for a list of the words), it is reasonable that chance is a factor in this case. Furthermore, as the results for French and non-French knowledge yield different trends, there is evidence that knowledge of French has a direct bearing on learning English. The purpose of this study is only to look at the effect of knowing a second language (L2) on learning a third (L3), when the mother tongue (L1) is unrelated to either L2 or L3, and L2 and L3 are linguistically related. Therefore, this research is only concerned with the fact that knowledge of the French word influences using (recognizing or producing) the English counterpart, and not how English words are used when the related French word is not in the person's vocabulary.

The fact that the transfer effect is observed both during recognition and production is important to our transfer hypothesis as well. The trends are the same for both tasks, and there is no statistically significant difference between the two, over the three types of lexical contrasts. We can, therefore, be fairly sure that the transfer effect is not merely an artifact of only one aspect of language use. The present results suggest that the transfer effect is present in both of the skills necessary for mastering a language -- recognition and production. However, one could say that even in the case of cognates, for those with French knowledge the percentage of incorrect responses is high (62.39 %). This emphasizes how little English knowledge the subjects still had after an average stay of 6.35 months in the U.S.

Additional observations of some of the responses, although not statistically analyzed, suggest the strength of the transfer effect in that the reverse effect is also noted. Vietnamese speakers knowing the English word tend to transfer this knowledge to French, if they think that these languages are similar for a particular word (e.g., indicating an English word wall and a French word walle, when the actual French translation is mur). Such findings are indicative of the effect of transfer of linguistically related languages.

CONCLUSIONS OF PSYCHOLINGUISTIC THEORY ON LEXICAL TRANSFER PROCESSES

Ulijn (1981) concludes that if a thorough conceptual analysis is of utmost importance in reading a foreign language, all types of lexical contrasts between foreign and native language will hamper, since content words are the main carriers of meaning in the text. In the Shadok project data, there was an indication

that L1/L2/L3 cognates were easier to recognize than L1/L2/L3 misleading cognates. This indication was observed for both Dutch (L1) and English (L2) by Dutch engineering students reading French. These indications are largely confirmed in the present study as effects in an L1/L2/L3 setting, where L1 = Vietnamese, L2 = French and L3 = English, in an English lexical recognition and production task for Vietnamese native speakers knowing the target words in French.

Furthermore, formal contrasts, probably the largest category of lexical contrast in any language pair, appear in the middle of the range of difficulty. Vietnamese, knowing French and beginning to read English, are helped by French/English cognates (positive transfer), "misled" by misleading cognates (negative transfer) and do not receive any help or interference from formal contrasts (no transfer). The successful recognition of this category is merely a function of word knowledge.

Thus, as far as lexical transfer processes are concerned, lexical recognition and production are quite similar. Relations between conceptual system and lexicon seem to be comparable in both cases, whereas in other subsystems of language process, certain activities are limited to production and others to recognition (e.g., the script recognizer and sentence/text parser in reading, and the sentence/text generator and script producer in writing, Kempen, 1976).

Although the present authors strongly believe that our data support our hypothesis, we realize that it is important not to overgeneralize the results. This study is limited to a beginning knowledge of English. Lexical transfer effects are minor, comparatively, in advanced stages of learning, as has been demonstrated with German (van Weeren, 1977) and English (Huisjes-Schreuder, 1978) acquired by Dutch students. In advanced stages of language learning, even misleading cognates produced relatively few errors.

Additionally, it should be remembered that the present study is limited to written language. Transfer effects in oral language are probably largely affected by the pronunciation of the particular words. Vietnamese, French, and English phonological systems are so vastly different that knowledge of French pronunciation does not even allow for correct recognition of French words produced as English (cognates). On the production side, French/English cognates pronounced in a Vietnamese (or possibly French) manner would be barely intelligible to a native English speaker.

Although we warn against overgeneralizing the present results beyond written language at the beginning stages of learning, we, nevertheless, believe that these lexical transfer effects are also valid for other related language pairs (e.g., English/German, English/Spanish, Dutch/German and Russian/Polish) although replication studies are obviously needed with adapted words and sentence stimuli. In analyzing other language pairs, it is important to define the lexical distance between the languages observed.

The present study researched only the three basic types of lexical contrasts. Relations between the conceptual system and lexicon are not of a one-to-one correspondence. Cs, FCs and MCs often occur in combinations as homonyms or synonyms in one language and not in the other. Further research along these lines, with additional words in each category and closer control of MCs would be challenging, as the lexical and conceptual variables involved in such settings are not easily manipulated experimentally.

Possible Implications for Teaching English to Native Vietnamese Speakers

The importance of cognate-based teaching of a foreign language related to the native language is largely demonstrated by Hammer

(1979). If English and French share nearly 11,000 words, the learner needs to acquire very few new words, at least in the written form without syntactic imbedding. In the beginning, at least, it is only necessary to learn how to apply knowledge of the person's first language to the second, and related language. The commonalities between French and English are also relevant for multilingual settings in which French or English is the second language and the other is the third.

In teaching English to immigrants, sometimes the native language is taken into account. However, for different reasons, immigrants may know several languages. If English is related to any of these languages, it makes sense to consider this factor and place the students in homogeneous ESL classes. African, Arab, Caribbean (Haitian) or Polynesian students might know French, and within the classroom setting, this lexical transfer effect could be used in the teaching of English.

These implications, however, apply only to learning to read or write in such a related language. In language pairs where the pronunciation differs so much as between English and French, cognates will simply not be easy to identify in their oral form. So, cognate-based teaching in listening and speaking a related language could suffer from this lack of phonetic similarity. Nevertheless, a reasonable access to the written form probably facilitates a positive transfer here, too.

Once validated, the exercise used in this experiment could serve as a placement test, in an adapted version, even for other (Indo-Chinese) immigrants. At the least, it could function as a tool to measure the French bias of Vietnamese speakers. Additionally, English/French lexical transfer seems to be relevant for multilingual settings such as the Dutch secondary school, where English and French are acquired simultaneously. Too little is presently known about this type of acquisition.

In conclusion, lexical transfer can be positive, however, it can also be negative. Too much cautioning against the traps of the misleading cognates could be counterproductive, with more errors being produced than avoided. It might be advisable to provide ESL or FL students with exercises based implicitly on lexical contrasts and not exaggerate explicit teaching of the "wolves in sheep's clothing" (MCs), particularly not in the beginning of the learning process. Whenever explicit teaching is needed, translation seems to be the only reliable way to elucidate the exact meaning of the CS, FCs and MCs shared by the two languages involved.

The results also show, however, that the subjects who did not know the French equivalent of the target word did as well (and in the case of FCs, significantly better) in identifying the target word as those who knew the French equivalent. Therefore, it would seem wise not to overemphasize the role of translation from an intermediate language closely related to English in attempting to teach English vocabulary.

ACKNOWLEDGMENTS

The Vietnamese project reported here was carried out while the principal investigator was visiting Stanford University (U.S.A.) through a grant from the Netherlands Organization for the Advancement of Pure Research (1979 - 80). He is grateful for comments he received during various presentations of the proposal at Stanford from the participants to the Reading Research Seminar (by Dr. R. Calfee), the Practicum in Consulting on Methodological Problems in Education Research (by Dr. R. Sitgreaves), and the Child Language Noon Seminar (Department of Linguistics). Additionally, gratitude must be paid to Dr. Robert Politzer, School of Education, Stanford University, for his helpful comments on an early draft of this paper.

Furthermore, the authors wish to acknowledge the invaluable help of Ms. Patricia Huong Nguyen, ESL teacher of the Indo-Chinese Training Program in San Francisco, who did the Vietnamese translation of the sentences, and Dr. Nguyen Van Canh, who provided important information about the Vietnamese language and education system. As Director of the Indo-Chinese Cultural and Social Center of San Mateo County, Dr. Van Canh also made his staff available for secretarial services. Additionally, we appreciate comments from Dr. R. Grotjahn (University of Bochum, FRG), Dr. G. Kempen (University of Nijmegen), Mr. M. J. Morse (San Francisco State University) and Mr. K. Wikberg (University of Tromsø, Norway). Finally, we would like to thank the staff members of the different Indo-Chinese Refugee Resettlement Centers in San Jose, San Francisco and Redwood Dity, California, and of the San Francisco Community College Centers, who helped us find the necessary subjects.

NOTES

1. Ferguson, C. Personal Communication, Academic year 1979 - 80
2. This is not to deny that translation-equivalent terms may differ in connotation cross-culturally, as has been shown in experiments with English-Japanese bilinguals by King (1980)

REFERENCES

ABBAS, M. The vocabulary of application forms. Reading Improvement, 1979, 16, 28-31.

ALBERT, M. L., & OBLER, L. K. The bilingual brain. Neuropsychological and neurolinguistic aspects of bilingualism. New York: Academic Press, 1978.

ANDERSON, J. G., & HARMER, L. C. Le mot juste, a dictionary of English and French homonyms. New York: Dutton, 1938.

BAETENS-BEARDSMORE, H. Bilingualism: Basic principles. Clevedon: Tieto, 1982.

BAUSCH, K. R., & KASPER, G. Der Zweitsprachenerwerb: Möglichkeiten und Grenzen der 'grosser' Hypothesen. Linguistische Berichte, 1979, 64, 3-35.

BLUM, S. & LEVENSTON, E. Universals of lexical simplification. In: C. Faerch & G. Kasper (Eds.), Strategies in Interlanguage Communication, London: Longman, 1983, 119-139.

BOILLOT, F. Le vrai ami du traducteur anglais-français et français-anglais. Presses Universitaires Françaises, 1930.

CARROLL, J. B. Words, meanings and concepts. Harvard Educational Review, 1964, 34 (2), 178-202.

CARROLL, J. B. et al. Word frequency book, the American heritage. New York: Houghton Mifflin, 1971.

CHUMBOW, B. S. The mother tongue hypothesis in a multilingual setting. In J. G. Savard & L. Laforge (Eds.), Proceedings of the 5th Congress of AILA, Quebec: Presses de l'Université Laval, 1981, 42-55.

CLARK, H. H. & CLARK, E. V. Psychology and language. New York: Harcourt & Brace Jovanich, 1977.

DUONG THANH BINH. Vietnamese-English phrase book with useful word list. Arlington, VA: Center of Applied Lingustics, 1975.

ERVIN-TRIPP, S. M. Semantic shift in bilingualism. American Journal of Psychology, 1961, 74, 233-241.

FORSTER, K. J. Accessing the mental lexicon. In F. J. Wales & E. Walker (Eds.) New Approaches to Language Mechanisms, Amsterdam: North-Holland, 1976, 257-287.

GOUGENHEIM, G. L'élaboration du français fondamental. Paris: Didier, 1967.

GROGNET, A. G. et al. A manual for Indochinese refugee education Arlington, VA: Center for Applied Linguistics, 1976.

GROSJEAN, F. Life with two languages: An introduction to bilingualism Cambridge, MA: Harvard University Press, 1982.

HALL, R. Pidgin and creole languages. Ithaca, NY: Cornell University Press, 1966.

HAMMER, P. What's the use of cognates? Unpublished manuscript, University of Alberta at Edmonton (Canada), 1979.

HAMMER, P., & MONOD, M. J. English-French cognate dictionars. University of Alberta at Edmonton (Canada), 1976.

HATCH, E. (Ed.). Second language acquisition, a book of readings. Rowley, MA: Newbury, 1978.

HEUER, H. Wortassoziationen in der Fremdsprachendidaktik. In W. Hullen (Ed.) Neusser Vorträge zur Fremdprachendidaktik, Berlin, 1973, 66-83.

HUISJES-SCHREUDER, E. C. M. Het effect van waarschuwen bij de presentatie van verwarrende woordparen. Levende Talen , 1978, 337, 613-617.

ICKENROTH, J. On the elusiveness of interlanguage. Institute of Applied Linguistics, Universitiy of Utrecht (The Netherlands), 1976

JORDENS, P. L. Contrastivité et transfer . In J. P. Menting & J. M. Ulijn (Eds.), La Linguistique Appliquée aux Pays-Bays Etudes de Lingustique Appliquée, Paris: Didier, 1979, 94-101.

JUDD, E. L. Vocabulary teaching and TESOL: A need for re-evaluation of existing assumptions. TESOL Quarterly, 1978, 12(1), 71-76.

JUILLAND; A. G. et al. Frequency dictionary of French words. The Hague: Mouton, 1970.

KELLERMAN, E. Giving learners a break: native language intuitions as a source of predictions about transferability. Working Papers in Bilingualism, 1978, 15, 60-92.

KEMPEN, G. De taalgebruiker in de mens. Groningen: Tjeenk Willink, 1976

KEMPEN, G. Sentence construction by a psychologically plausible formulator. In R. N. Campbell & P. T. Smith (Eds.), Recent Advances in the Psychology of Language-Formal and Experimental Approaches, New York: Plenum Press, 1978, 103-123.

KERKMAN, H. De organisatie von het lexicon bij bilingualen. Toegepaste Taalwetenschap in artikelen, 1981, 11, 190-196

KING, S. T. Cognitive correlates of culturally dissimilar word meaning in the two languages of the bilingual. Ph. D. dissertation, George Washington University, 1980.

KIRSNER, K. et al.; Bilingualism and lexical representation, Quarterly Journal of Experimental Psychology, 1980, 32, 585-594.

KNIBBELER, W. Frans van Nederlanders. Doctoral dissertation, University of Utrecht, The Hague: Staatsuitgeverij (The Netherlands), 1977.

KOESSLER, M. Les faux amis des vocabulaires anglais et américain. Paris: Vuibert, 1975.

KRASHEN, S. D. Principles and practice in second language acquisition. Oxford: Pergamon, 1982.

LAMBERT, W. E. , HAVELKY, J., & CROSBY, C. The influence of language acqusitions context on bilingualism. In W.E. Lambert, Language, Psychology and Culture, Palo Alto, CA: Stanford University Press, 1972, 51-62.

LEVENSTON, E. A., & BLUM, S. Aspects of lexical simplification in the speech and writing of advanced adult learners. In L. Roulet (Ed.), The Notions of Simplification, Interlanguages, and Pidgins and their Relation to Second Laguage Pedagogy, Geneve: Droz, 1977, 51-71.

LOCOCO, G. M. V. A cross-section study on L3 acquisition. Working Papers in Bilingualism, 1979, 9, 44-75.

MACKAY, R. et al. (Eds.). Reading in a second language. Rowley, MA: Newbury, 1979.

MACKEY, W. F. et al. Le vocabulaire disponible du français. Vol. II, Paris/Montreal: Didier, 1966.

MARTIN, A. V. Teaching academic vocabulary to foreign graduate students. TESOL Quarterly, 1976, 10(1), 91-98

MCLAUGHLIN, B. Second language learning and bilingualism in children and adults. In S. Rosenberg (Ed.), Handbook of Applied Psycholinguistics, Hillsdale, NJ: Erlbaum, 1982, 329-384.

MORTON, J. A. Le lexique interne. La Recherche, 1983, 143, 479-481.

MOSENTHAL, P., WALMSLEY, S. A., & ALLINGTON, R. L. Word recognition reconsidered toward a multi-context model. Visible Language, 1978, 12(4), 448-468.

NAS, G. L. J. Bilingual visual words recognition: a study of lexical access coding in Dutch-English bilinguals, Ph. D. dissertation, Nijmegen University, 1983.

NGUYEN, DINH HOA. Vietnamese-English student dictionary. Southern Illinois University Press, London and Amsterdam: Feffer & Simons, 1971.

NGUYEN, HY QUANG. English-Vietnamese phrase book with useful word list. Arlington VA: Center of Applied Linguistics, 1975.

PHILLIPS, J. Vietnamese contact with French: Acquisitional variation in an language contact situation. Ph. D. dissertation, University of Indiana, 1975.

POLITZER; R. L. Teaching French, an introduction to applied linguistics. Waltham MA: Blaisdell, 1965.

PUGH, A. K. & ULIJN, J. M. (Eds.) Reading for professional purposes: Studies in native and foreign languages. London, Heinemann, 1984

RATTUNDE, E. Transfer-Interferenz? Probleme der Begriffsdefinition bei der Fehleranalyse. Die Neueren Sprachen, 1971, 4-14.

REINECKE, J. Pidgin French in Vietnam. In D. Hymes (Ed.), Pidginization and Creolization of Languages, Cambridge Univesity Press, 1971, 47-56.

RICHARDS, J. C. The role of vocabulary teaching. TESOL Quarterly, 1976, 10(1). 77-90

RINGBOM, H. The influence of other languages on the vocabulary of foreign language learners. Paper presented at AILA congress, Lund, Abo Akademi, 1981.

RIVERO, G. A., & BEST, M. Strategies for solving lexical problems through discourse and context. In C. H. Blatchford & J. Schachter (Eds.) On TESOl '78: EFL Policies and Programm Practices, Washington, D. C.: TESOL, 1978, 191-198.

ROSENZWEIG, M. R. Comparisons among word-association responses in English, French, German and Italian. American Journal of Psychology, 1962, 74, 347-360.

SCHREUDER, R., & LEVELT, W. J. M. Psychologische theorieën over het lexicon. Forum der Letteren , 1978, 19(1).

SEWARD, R. D. Dictionary of French deceptive cognates. New York and San Francisco: Vanni, 1947.

STEINBERG, D. Semantic universals in sentence processing and interpretation. A study of Chinese, Finnish, Japanese and Slovenian speakers. Journal of Psycholinguistic Research 1976, 5, 169-193.

TAYLOR, J. Similarity between French and English words: a factor to be considered in bilingual beharvior. Journal of Psycholinguistic Research , 1976, 5, 85-94

ULIJN, J. M. French as a foreign language in engineering education - an investigation into reading comprehension. Ph. D. dissertation Nijmegen University 1978. (ERIC Document Reproduction Service No. ED 157 011)

ULIJN, J. M. Le régistre scientifique et technique et ses constantes et variantes supra-linguistiques. Fachsprache, International Journal of Languages for Speciall Purposes, Vienna; Braumüller, 1979.

ULIJN, J. M. Conceptual and syntactic strategies in reading a foreign language. In E. Hopkins & R. Grotjahn (Eds.), Studies in Language Teaching and Language Acquisition, Quantitative Linguistics, Vol, 9, Bochum (FRG): Brockmeyer 1981, 129-166

ULIJN, J. M., & KEMPEN, G. The role of the first language in second language reading comprehension: some experimental evidence. In G. Nickel (Ed.), Proceedings of the 4th International Congress of Applied Linguistics, Stuttgart: Hochschul-Verlag, 1976, 495-507

ULIJN, J. M., WOLFE, S. J., & DONN, A. The lexical transfer effect of French knowledge in the acquisition of English by native Vietnamese speakers. Internal report, Eindhoven University of Technology, 1981.

ULIJN, J. M. Cross-language transfer in reading vs. writing: the CAH revisited in a psycholinguistic perspective, forthcoming.

WEEREN, J. VAN Interferenz und Valenz. Doctoral dissertation, University of Leiden (The Netherlands).

WEI, M., & LIGHT, F. A newspaper's vocabulary: A raw frequency count of the words in the South China Morning Post. Chinese University of Hong Kong, 1973.

WIKBERG, K. Lexical errors in English, made by Finnish and Swedish senior secondary school students: A comparison- Paper contributed to the Interlanguage Symposium at Hanaholmen, Helsinki (Finland), 1979.

WILCOX, G., & LEHMAN, J. A wolf in sheep's clothing, Spanish/English and French/English deceptive cognates for the non-bilingual teacher. Paper contributed to the TESOL Convention, San Francisco, 1980.

YOUNG, R. K., & WAVAR, M. Retroactive inhibition with bilinguals. Journal of Experimental Psychology, 1968, 77, 108-115.

APPENDIX A

THE 30 ENGLISH NOUNS USED IN THE EXPERIMENT (FOLLOWED BY THE FRENCH WORD, IF ORTHOGRAPHICALLY DIFFERENT)

Cognates	Formal Contrasts	Misleading Cognates
age/âge	bed/lit	anniversary/anniversaire
cousin	bird/oiseau	car
fruit	day/jour	cart/carte
million/milion	door/porte	cave
page	foot/pied	coin
police	pen/plume	crayon
prison	street/rue	fabric/fabrique
restaurant	wall/mur	lecture
table	water/eau	library/librairie
village	window/fenêtre	patron

APPENDIX B

ENGLISH TRANSLATION OF THE SENTENCES USED IN THE EXPERIMENT

Cognates (Cs)

1. That shape can be found on each page.
2. The village that you are looking for is ten miles from here.
3. Cold fruit is always enjoyable during the hot summer.
4. This restaurant is too expensive for my budget.
5. She polished the table for more than an hour.
6. The official asked the refugee about his age.
7. His cousin is visiting from out of town.
8. The unempolyment rate will not drop below three million.
9. The police suggested that we leave.
10. The disturbance at the prison caused much alarm.

Formal Contrasts (FCs)

1. There is a beautiful bird over there.
2. His foot was badly injured in the accident.
3. The street was full of people in the evening.
4. At the end of the day, the students returned home.
5. You must have water in order to live.
6. There is only one window in the office.
7. The bed was hard and uncomfortable.
8. The pen fell out of his pocket onto the floor.
9. It was possible to find the door in the darkness.
10. A wall remained standing after the fire.

Misleading Cognates (MCs)

1. Traveling by car is very exciting.
2. The temperature in the cave was always cool.
3. He couldn't see the coin from where he was standing.
4. The fabric was designed by a highly paid consultant.
5. The picture was drawn with a crayon.
6. We have had many patrons this past year.
7. The lecture was very informative.
8. I found a purse in the library.
9. There is an exhibit of antique carts at the museum.
10. It rained on the day of her anniversary.

Kontentanalytische Untersuchungen zur inhaltlichen und formalen Komplexität von Texten

M. H. Schwibbe, Göttingen
K. Räder, Göttingen

Es wird eine Methode entwickelt und vorgestellt, die a) textsortenübergreifend und b) gemeinsam formale wie auch inhaltliche Textcharakteristika erfaßt. Während wir uns bei den formalen Aspekten an traditionelle Variablenkonzepte wie Subordinationsindex, Satzlänge und Frequenz von Konjunktionen halten, werden die inhaltlichen Aspekte durch die psycholinguistischen Konstrukte Abstraktheit, Dogmatismus, Redundanz und Wortgeläufigkeit parametrisiert. Eine Faktoranalyse zur Variablenreduktion über 744 Texte gruppiert die Variablen - wie erwartet - nach den Komplexitätsaspekten, die Diskriminanzanalyse der Faktorscores trennt u.a.: a) Zeitungen, Literatur, Wissenschaftstexte, Aufsätze, Briefe, b) Zeitungstexte aus Politik und Sport, c) Trivial- von ernster Literatur, d) Texte der FAZ von BILD, e) Aufsätze verschiedener Altersstufen. Inhaltlich hoch komplex sind politische und wissenschaftliche Texte, den Gegenpol bilden unter Aktivation verfaßte Briefe. Formal hoch komplex sind Abituraufsätze, einfach strukturiert Texte der Bildzeitung. Die Kombination der beiden Forschungsansätze zur Komplexitätserfassung ermöglicht eine differentielle Beschreibung von Texten unterschiedlicher motivationaler, kognitiver und entwicklungspsychologischer Provenience.

1. Problemstellung

Der heuristische Wert kontentanalytischer Untersuchungen zur Komplexität von Form und Inhalt von Textmaterialien wird zumeist durch die Beschränkung der einzelnen Studien auf jeweils nur einige Parametrisierungen dieses Konstrukts geschmälert. Daraus resultiert ein Mangel an konzeptübergreifenden Analysen, die die Variationen der Variablen in einen größeren Rahmen stellen und dadurch zur Theoriebildung über textuale Komplexität beitragen können. Die vorliegende Arbeit analysiert deshalb empirisch face-, kriteriums- und externvalide stilstatistische Indikatoren, die der textualen Komplexität affin konzipiert sind, auf ihre strukturellen Zusammenhänge.

Forschungshistorisch lassen sich zumeist zwei verschiedene Konzepte zur Bestimmung der Komplexität von Textmaterialien aufzeigen, die sich in ihren Parameterbildungen teilweise überlappen: die Readability-Forschung, die die Wirkung von Texten auf den Rezipienten untersucht, und sprachstatistischen Analysen, die aus Textmerkmalen Inferenzen hinsichtlich kognitiver Prozesse der Textverfasser zu erschließen suchen. Der ersterwähnte Forschungsansatz sucht nach generellen Merkmalen, die unmittelbaren Einfluß auf die Lesbarkeit und die Verständlichkeit von Texten haben (FLESH 1949, 1950). Dabei konnten mehrfach zwei Hauptprädiktoren zur Bestimmung der Textverständlichkeit nachgewiesen werden: Wortschwierigkeit - gemessen in der Wortlänge und der Auftretenshäufigkeit - und Satzschwierigkeit - gemessen an der Länge und Tiefe der Sätze. Neuere Untersuchungen weisen darüber hinaus auch der lexikalischen Redundanz Bedeutung für die Lesbarkeit zu (TAUBER et al., 1980, DICKES und STEINER, 1977). Aus diesen Untersuchungen wurden eine Fülle von Formeln zur Textlesbarkeit (siehe z.B. NGUYEN und HENKIN 1980) entwickelt, deren externe Validität allerdings durch die Heterogenität der Definitionen des Prädikators 'Lesbarkeit' (siehe dazu HOFER, 1976) eingeschränkt wird. Ein anderes Konzept zur Bestimmung der Komplexität von Texten wird von LANGER et al. (1974) verfolgt: Die Autoren extrahieren aus Skalierungen faktoranalytisch 4 Dimensionen, die mit Lesbarkeit im Zusammenhang stehen, wobei der Dimension 'Einfachheit versus Komplexheit' - wie z.B. TAUBER et al. (1980) zeigen - die größte Bedeutung zukommt: höchste Korrelationen zu Parametern des Cloze-procedure als Mass für die Verständlichkeit, Satzlänge und lexikalische Redundanz. Die Validität der letztgenannten Variable analysiert auch GAMP (1971) und betrachtet die Redundanz als Textqualität unter informationstheoretischen Gesichtspunkten: Texte mit niedriger Entropie bezeichnet der Autor als 'strukturiert', Texte mit hoher Entropie - z.B. mit großem Wortschatz oder mit großer Variation in der Wortkombination - als komplexer, weniger geordnet und unbestimmter.

Diesen und ähnlichen Textcharakteristika wird auch bei den Untersuchungen zu kognitiven Strukturen seit längerem nachgegangen: Die Grundlage dafür bieten Annahmen und Überlegungen der 'Allgemeinen Semantik' (KORSZYBSKY 1933, SANFORD 1942, HAYAKAWA 1949), daß sprachliches Handeln als Ausdruck dynamisch/kognitiver Prozesse auch Rück-

schlüsse auf die dahinter stehende Organisation von Denkstrukturen zuläßt. Dieser Autorengruppe - besonders SANFORD - gilt die Verwendung von 'allness terms' als ein Anzeichen von 'low order abstraction', ein Ansatz, den OSGOOD und WALKER (1959), ERTEL (1978) und SCHWIBBE et al. (1983) weiterentwickelt haben. Diese Variable wurde von letztgenannten Autoren zur Bestimmung kognitiv/sprachlicher Prägnanz und Glättung eingesetzt. Zur Parametrisierung des Gegenpols dieser Dimension des Denkens - der abstraktiven Verkürzung (i.S. von KLIX, 1976) - wurde im Rahmen der Readabilityforschung (dann jedoch von ihr ausgegliedert) (siehe dazu JENKINS und JONES (1951)) ein Verfahren zur Bestimmung von 'high order abstraction' - der'abstraktiven Verdichtung' (i.S. von KLIX) zur ökonomischen Verarbeitung großer Informationsmengen - entwickelt: das Abstraktheitssuffix-Verfahren. Für die deutsche Sprache haben GÜNTHER und GROEBEN (1978) sowie SCHWIBBE und RÄDER (1982) dieses Konzept validiert.

Die einem Individuum eigene Informationsmenge, die es zu verdichten gilt und die die Elemente der kognitiven Struktur bilden, mißt auf der sprachlichen Ebene die Wortredundanz (z.B. TTR: Type/Token-Ratio), die in einem mehrfach statistisch abgesicherten Verhältnis zur Intelligenz (CHOTLOS, 1944, WECHSLER, 1961), zur Kreativität (LEWANDOWSKY, 1980) und zur diachronischen Sprachentwicklung (DEUTSCH et al. 1964, SCHWIBBE, 1982) steht. Neben der Größe der Menge differenter Wortelemente kommt auch deren Geläufigkeit als Wortattribut Bedeutung für die Organisation kognitiver Strukturen zu. Die Verwendung eines eher ungewöhnlichen Sprachmaterials - z.B. über die Zugehörigkeit zu den häufigsten Wörtern einer Sprache gemessen - indiziert Niveau im sprachlich/kognitiven Handeln in weitester Hinsicht.

Ein dritter Forschungsansatz, der komplexitätsaffine Variablen untersucht, ist die Sprachtypologie. Hierbei werden unter anderem syntaktische Texteigenschaften wie Satzlänge, Satztiefe und Satzbreite (ALTMANN und LEHFELDT 1973, FUCHS 1968) als geeignete Diskriminatoren für Textsorten und -typen angesehen. Auf der Wortebene werden Variablen wie Phonementropie oder Wortlänge als morphologische Eigenschaften zur Typologisierung herangezogen (siehe z.B. GREENBERG 1961), Allerdings bestehen hinsichtlich der theoretischen Ansätze

und den daraus folgenden Zugehensweisen zum Auffinden und Bestätigen von Textgruppen (Diskriminanzanalyse vs. Clusteranalyse, Ausschöpfung des möglichen Variablen- und Typenraums) beträchtliche Diskrepanzen, sodaß das Konzept der Typologisierung bei den weiteren Analysen hier nur peripher verfolgt wird (1).

Den aus den dargestellten Forschungskonzepten resultierenden Beschreibungsaspekten - einfach vs. komplex, dogmatisch vs. undogmatisch, elaboriert vs. restringiert, abstrakt vs. konkret etc. - liegt u.E. ein genereller Faktor zugrunde, der mit 'Komplexität' bezeichnet werden kann. Es läßt sich sogar eine übergeordnete Operationalisierung von Komplexität finden, unter die die Indikativität der genannten Parameter subsummiert werden kann, der 'Komplexität' sensu BERLYNE (1974). Ausgehend von Befunden der Wahrnehmungspsychologie formuliert der Autor folgende - über Mengen und Mengenmaße definierte - Beziehungen: a) je mehr Elemente in einem Gebilde zusammengefaßt sind, desto komplexer ist das Gebilde. Auf der Wort- und Satzebene wurde dieses durch Längenmaße, auf der Satzebene auch durch die Tiefe der Satzschachtelung indiziert, b) je heterogener die Elemente, um so komplexer ist das Ganze. Das TTR oder ein verwandtes Maß parametrisiert damit Komplexität des zugrundeliegenden Textes, c) je regelmäßiger die Konturen eines Gebildes, um so einfacher ist das Gebilde. Die Regelmäßigkeit und Prägnanz sprachlicher Gebilde wird durch den Dogmatismusquotienten abgebildet, der - wie ERTEL (1981) ausgeführt - 'Struktur' -indizierenden Charakter hat, d) je mehr Relationen zwischen den Elementen eines Gebildes bestehen, umso komplexer ist das Gebilde. Diese Relationen können auf semantischer Ebene durch den Abstraktionsindex, auf der syntaktischen Ebene durch Konjunktionen gemessen werden, e) je höher die Ungewißheit über das Auftreten der Elemente in einem Gebilde, um so komplexer ist das Ganze. Diese Unsicherheit kann in Textmaterialien durch die Auftretenswahrscheinlichkeit der Wörter parametrisiert werden. Diese Beziehungen sowie die Parametrisierungsebenen und -methoden sind tabellarisch in der Abbildung (1) dargestellt.

Definition: Ein Gebilde ist umso komplexer, ...	Sprachliche Operationalisierung durch Kontentanalyt. Indikatoren	Hypothet. Korr. mit Komplexität
je mehr Elemente in dem Gebilde zusammengefasst sind.	Wortebene: Wortlänge Satzebene: Satzlänge, Satzschachtelung	Positiv Positiv
je heterogener die Elemente.	Textebene: Wortredundanz (zb. TTR)	Negativ
je unregelmäßiger die Konturen der Subgebilde.	Textebene: Prägnanzindizierender Dogmatismusquotient	Negativ
je mehr Relationen zwischen den Elementen bestehen.	Textebene: Abstraktheitsindex Frequenz von Konjunktionen	Positiv Positiv
je geringer die Auftretenswahrscheinlichkeit der Elemente.	Wortebene: Geläufigkeitsindex	Negativ

Abb. 1. Komplexität sensu Berlyne

Es lassen sich - wie ersichtlich - verschiedene Formen der Parametrisierung von Komplexität unter mengentheoretischen Beschreibungsgesichtspunkten konsequent zur Operationalisierung eines so definierten Konstruktes zusammenfassen. Dieses Konstrukt soll im folgenden auf seine interne und inhaltliche Validität untersucht werden.

2. Material und Methoden

Die Textgrundlage dieser Studie bildet eine Sammlung von 744 Einzeltexten der Textsorten 'Aufsätze', 'wissenschaftliche Literatur', 'Briefe', 'Belletristik', und 'Zeitungen'. Die einzelnen Textsorten sind jeweils in inhaltlich differente Untergruppen partitioniert. Für jeden dieser Texte werden folgende Variablen bestimmt: der Prägnanzindikator (Dogma-Anteil referierter Überzeugungen sensu ERTEL (1981), SCHWIBBE et al. (1983)), der YULE-Index (Redundanz auf der Wortebene), die Anzahl von Konjunktionen (auf die Textlänge - abzüglich der Formworte - relativiert), der Anteil der Wörter, die unter die häufigsten 1500 Wörter des MEIER-Lexikons (MEIER 1964) fallen, als Geläufigkeitsindex, die mittlere Wortlänge (in Graphemen), der mittlere Quotient aus der Anzahl der Sätze und der Anzahl der Nebensätze (mit -1 multipliziert (2)) als Indikator der Satzschachtelung, und der Abstraktheitsindex (nach dem Suffixverfahren - GÜNTHER und GROEBEN 1978, SCHWIBBE und RÄDER 1982). Zur Überprüfung der Dimensionalität der Indikatoren wird eine Faktoranalyse (principle components, iterative Kommunalitätenschätzung, Varimaxrotation) durchgeführt. Die Stabilität der Faktorstruktur wird über die getrennt für die 'odd' und 'even' Fälle berechnete Struktur und deren Ähnlichkeitsvergleich durch Zielrotation bestimmt (siehe dazu SIXTL, 1964). Mittels einer 'schrittweisen Regression' wird eine Optimierung der Vorhersage der Strukturdimensionen aus den einzelnen Variablen vorgenommen. Die Prüfung auf Unterschiedlichkeit der Messwertausprägungen in den einzelnen Textsorten und deren Untergruppen wird varianzanalytisch (einseitige Fragestellung) durchgeführt. Der Eta-Quadrat-Wert wird aus der Prüfstatistik F berechnet und gibt das Ausmaß erklärter Varianz der abhängigen Variablen wieder. Mittels

der Diskriminanzanalyse wird die diskriminative Potenz der Variablenkombination zur Trennung von Textsorten und Gruppen geprüft, die Effizienz wird ebenfalls über Eta-Quadrat - als Equivalent von WILK's Lambda - bestimmt.

3. Ergebnisse

3.1 Zur Dimensionalität der Indikatoren

Die Interkorrelationsmatrix der genannten Indikatoren wird faktorisiert, das Eigenwertkriterium (1.00) empfiehlt die Extraktion von 2 Faktoren. Auf dem ersten Faktor, der 36% der totalen Varianz erklärt, laden die einzelnen Variablen entsprechend den oben angestellten Überlegungen dem Vorzeichen entsprechend (siehe Abbildung 2). Die Stabilität dieser Struktur beträgt r = 0.94. Damit können die Annahmen einer grundlegenden Dimension hinter der Variation komplexitätsindizierender Variablen bestätigt werden. Eine Rotation der beiden Faktoren trennt die Variablen in die Gruppe der semantisch-inhaltlichen und die der syntaktisch-formalen Komplexität indizierenden Variablen. Dieser Befund stützt die Überlegungen zur Relevanz von Wort- und Satzkomplexität im Gesamtkonstrukt. Für jeden Text und jeden Faktor werden die Faktorscores bestimmt und in Z-Werte (Mittelwert: 100, Streuung: 10) umgerechnet. Da die Faktorstruktur orthogonal rotiert wurde, sind diese neuen Messwerte für Komplexität unkorreliert.

In diesem Zusammenhang wird die Frage nach einer möglichst ökonomischen, jedoch genauen Schätzung der Messwerte dieser Dimensionen an beliebigen Texten akut. Die Regressionsanalyse zeigt, daß die 'inhaltliche Komplexität' (=IK) durch die Wortlänge (in Graphemen) allein schon mit einer erklärten Varianz von 81% (r = 0.90) bestimmt werden kann. Wird in die Vorhersagegleichung noch der YULE-Index einbezogen, erhöht sich die Sicherheit auf 90% (r = 0.95). Die Vorhersagegleichung lautet:

IK = 51,77 + 9,00 x (Wortlänge) + 0,149 x (Yule-Index).

Die 'formale Komplexität' (=FK) kann durch die 'Satzschachtelung' allein mit einer Varianzaufklärung von 69%, durch die Anzahl der Konjunktionen zusätzlich - mit gleicher Sicherheit wie die erste

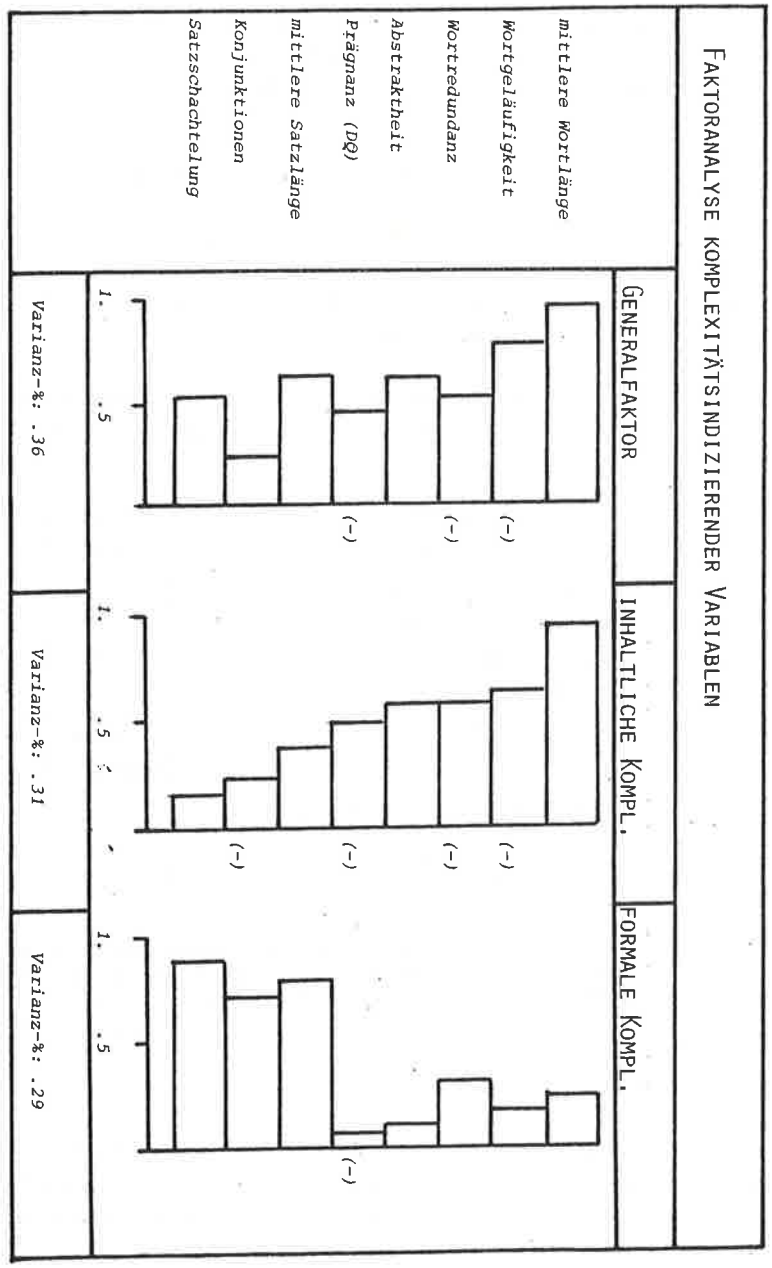


Abb. 2. Faktoranalyse komplexitätsindizierender Variablen

Dimension - mit einem r = 0.95 (erklärte Varianz: 90%) bestimmt werden. Die Vorhersagegleichung lautet:

FK = 93,46 + 8,34 x (Satzschachtelung) + 125,29 x (Konjunktionen).

3.2 Komplexitätsunterschiede zwischen Textsorten und deren Untergruppen

Die Mittelwerte der Komplexitätsaspekte für die einzelnen Textsorten sind in der Tabelle (1) dargestellt.

Tabelle 1: Standardwerte inhaltlicher und formaler Komplexität für die Gesamtstichprobe der Texte.

Gruppen	inhal. K.	form. K.	N
AUFSÄTZE	101.9	106.4	236
ZEITUNGEN	104.5	88.8	158
BRIEFE	84.4	101.8	91
LITERATUR	93.0	99.4	128
WISSENS.	107.9	101.1	139
F, df	228.24	126.67	4/739
p	.00	.00	
eta-Qua.	.55	.41	
eta-Qua. (Disk) : .75			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua. : Ausmaß erklärter Invarianz
WISSENS. : Wissenschaftliche Literatur

Als besonders inhaltlich komplex können danach wissenschaftliche Texte, Zeitungen und Aufsätze gelten, Briefe und Literatur sind danach inhaltlich eher einfach strukturiert. Auf der formalen

Ebene imponiert besonders das Niveau der Textsorte 'Aufsätze'. Zeitungen sind hingegen als formal besonders einfach konzipiert anzusehen. Die Diskriminationsfähigkeit der Dimensionen für diese Textsorten ist beträchtlich. Aufsätze, Briefe und Literatur sind formal komplexer als inhaltlich, Zeitungen und wissenschaftliche Literatur zeigen ein dazu inverses Bild.

3.2.1 Analyse der Aufsätze

Zunächst soll die Textsorte 'Aufsätze' untersucht werden. In dieser Textsorte liegt eine Untergruppe von 28 Einzeltexten zu verschiedenen Themen zur Rubrik 'Zeitlupe 20' der Wochenzeitschrift 'DIE ZEIT' von 14 bis 20 Jährigen vor (jeweils 2 von männlichen und zwei von weiblichen Verfassern eingesandte Textmengen pro Altersstufe). Die Mittelwerte für die IK und die FK sowie die Ergebnisse der Varianz- und Diskriminanzanalyse sind in der Tabelle (2) dargestellt.

Tabelle 2: Standardwerte inhaltlicher und formaler Komplexität für unterschiedliche Altersstufen (AUFSÄTZE)

Gruppen	inhal. K.	form. K.	N
14 Jahre	96.1	105.2	4
15 Jahre	100.0	100.2	4
16 Jahre	100.0	105.6	4
17 Jahre	103.9	106.4	4
18 Jahre	103.5	107.8	4
19 Jahre	107.5	108.9	4
20 Jahre	105.0	105.8	4
F, df	4.44	1.05	6/21
p	.00	.42	
eta-Qua.	.56	.23	
eta-Qua.(Disk) : . 68			

F, df : F-Wert der Varianzanalyse, Freiheitgrade.
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua. : Ausmaß erklärter Invarianz

Es bestätigen sich die Befunde und Annahmen der Entwicklungs- und Kognitionspsychologie über die diachronische Sprachentwicklung in einer altersbedingten Zunahme beider Komplexitätsaspekte. Die Effekte sind auf der inhaltlichen stärker als auf der formalen Ebene ausgeprägt.

Diese Textsorte kann weiterhin in Altersgruppen der Verfasser (durchschnittlich 15, 18 und 22 Jahre) eingeteilt werden. Bei den beiden ersten Altersgruppen handelt es sich um Schul- bzw. Abituraufsätze. Die Texte der durchschnittlich 22 Jährigen wurden von Studenten zu Themen der 'Zeitlupe 20' verfaßt. Die Ergebnisse weisen aus, daß auch in dieser Stichprobe eine diachronische Zunahme an IK zu verzeichnen ist. Auch hier sind (siehe Tabelle 3) die Effekte der Gruppierung auf die FK geringer:

Tabelle 3: Standardwerte inhaltlicher und formaler Komplexität für die Textsorte AUFSÄTZE

Gruppen	inhal. K.	form. K.	N
15 Jahre	96.4	104.5	58
18 Jahre	103.1	108.3	75
22 Jahre	104.9	105.9	75
F, df	33.34	8.70	2/205
p	.00	.00	
eta-Qua.	.25	.08	
eta-Qua. (Disk) : .33			

p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua.: Ausmaß erklärter Varianz

15 Jahre = Schüler (15 Jahre alt)
18 Jahre = Schüler (18 Jahre alt)
22 Jahre = Studenten (22 Jahre alt)

die höchste Komplexität dieser Ebene zeigen die Abituraufsätze, die augenscheinlich eine Ausbildungsstufe abbilden, die besonderen Wert auf formale Kriterien legt. Komplexitätsmindernd wirkt sich

demgegenüber die Zusammensetzung der Stichprobe der 22-jährigen aus, die auch Nicht-Abiturienten umfaßt. Dies macht eine Reduktion der Komplexität dieser Texte gegenüber denen der Abiturienten wahrscheinlich.

3.2.2 Analyse der wissenschaftlichen Literatur

Die wissenschaftliche Literatur ist in die Subgruppen 'strenge Wissenschaft' (Texte aus der Fachliteratur) und 'Populärwissenschaft' (Texte aus Büchern und Magazinen) unterteilt. Die Ergebnisse der teststatistischen Prüfung und die Mittelwerte sind in der Tabelle (4) dargestellt.

Tabelle 4: Standardwerte inhaltlicher und formaler Komplexität für die Textsorte WISSENSCHAFT

Gruppen	inhal. K.	form. K.	N
WISS.L.	110.9	102.7	51
POP. L.	106.2	100.3	88
F, df	22.54	3.76	1/137
p	.00	.05	
eta-Qua.	.14	.03	
eta-Qua. (Disk) : .15			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua.: Ausmaß erklärter Varianz

WISS.L. = wissenschaftliche Literatur
POP. L. = populärwissenschaftliche Literatur

Dabei imponiert besonders der Unterschied zwischen den Mittelwerten der IK. Der Unterschied in der FK erreicht hingegen gerade das Niveau statistischer Bedeutsamkeit. Insgesamt betrachtet sind wissenschaftliche Texte inhaltlich komplexer als formal.

Diese Textsorte läßt sich außerdem in Textgruppen naturwissenschaftlicher, sozialwissenschaftlicher und geisteswissenschaftlicher Herkunft unterscheiden. Das Entscheidungsrationale für diese Gruppierung bildet die ehemalige Einteilung der Studiengänge in Fakultäten der Universität Göttingen. Die Mittelwerte und die testtheoretischen Kennwerte sind für diese Untergruppen in der Tabelle (5) dargestellt.

Tabelle 5: Standardwerte inhaltlicher und formaler Komplexität für die Textsorte WISSENSCHAFT

Gruppen	inhal. K.	form. K.	N
NATURW.	106.8	96.9	58
SOZ.W.	110.0	103.9	40
GEIST.W.	107.2	104.5	41
F, df	4.90	23.54	2/136
p	.00	.00	
eta-Qua.	.07	.26	
eta-Qua. (Disk) : .30			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlickeit des F-Wertes
eta.-Qua.: Ausmaß erklärter Varianz

NATURW. = naturwissenschaftliche Texte
SOZ.W. = sozialwissenschaftliche Texte
GEIST.W.= geisteswissenschaftliche Texte

Als formal gering komplex stellen sich naturwissenschaftliche Texte dar. Als inhaltlich besonders komplex erscheinen sozialwissenschaftliche Texte. Die Dimension der FK erscheint wesentlich diskriminativer als die Dimension der IK.

3.2.3 Analyse der Zeitungstexte

Die Texte der Textsorte 'Zeitungen' sind in die Gruppen 'Bildzeitung', 'Neues Deutschland', 'Die Welt' und die 'Frankfurter Allgemeine Zeitung' partitioniert. Die Mittelwerte und Teststatistiken sind der Tabelle (6) zu entnehmen.

Tabelle 6: Standardwerte inhaltlicher und formaler Komplexität für die Textsorte ZEITUNG

Gruppen	inhal. K.	form. K.	N
BILD	97.7	74.6	33
ND	108.6	95.9	30
WELT	103.9	91.1	73
FAZ	110.9	92.8	22
F, df	44.49	124.91	3/154
p	.00	.00	
eta-Qua.	.46	.71	
eta-Qua. (Disk) : .76			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua.: Ausmaß erklärter Varianz

BILD = Bildzeitung
ND = Neues Deutschland
WELT = Die Welt
FAZ = Frankfurter Allgemeine Zeitung

Beide Komplexitätsdimensionen können zwischen den Untergruppen differenzieren, die der FK ist allerdings trennschärfer als die der IK. Als besonders einfach strukturiert in beider Hinsicht sind die Texte der Bildzeitung. Inhaltlich und formal komplex erscheinen innerhalb der Zeitungstexte die der FAZ und des ND. Die Diskriminationsfähigkeit der beiden Dimensionen zusammen ist mit einem Etaquadrat = 0.76 (R = 0.87) beträchtlich. Die IK der Texte ist durchgehend höher als die FK.

3.2.4 Analyse der Texte der 'FAZ' und des 'ND'

Die Texte des 'Neuen Deutschland' und der 'Welt' wurden bereits vom 'Institut für deutsche Sprache' in Bonn (siehe dazu HELLMANN, 1972) u.a. in die Gruppen 'Politik', 'Wirtschaft', 'Soziales', 'Sport' und 'Kunst' eingeteilt. Die Mittelwerte der Komplexitätsdimensionen für diese Gruppen sind in der Tabelle (7) dargestellt.

Tabelle 7: Standardwerte inhaltlicher und formaler Komplexität innerhalb der Textsorte ZEITUNG

Gruppen	inhal. K.	form. K.	N
POLITIK	109.6	97.0	29
WIRT.	103.9	89.6	28
SOZIALES	104.4	92.0	18
SPORT	99.4	90.9	8
KUNST	100.2	92.2	13
F, df	12.94	8.09	4/91
p	.00	.00	
eta-Qua.	.36	.26	
eta-Qua.(Disk) : .48			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua.: Ausmaß erklärter Varianz

POLITIK = Rubrik Politik
WIRT. = Rubrik Wirtschaft
SOZIALES = Rubrik Soziales
SPORT = Rubrik Sport
KUNST = Rubrik Kunst

Inhaltlich komplex sind Texte der Politik, aber auch von Wirtschaft und Sozialem. Einfach strukturiert sind die Texte von Kunst und Sport. Gegenüber der IK fallen die Werte der FK der gesamten Zeitungsstichprobe deutlich ab und liegen unter dem Gesamtmittelwert von 100. Die Unterschiede in der Komplexität zwischen den Gruppen sind auf der inhaltlichen Ebene größer als auf der formalen.

3.2.5 Analyse der Textsorte 'Literatur'

Diese Textsorte ist in die Subgruppen 'ernste Literatur', 'Trivial-Literatur' und 'Lyrik' unterteilt. Die Ergebnisse der Mittelwertbildung und der varianz- und diskriminanzanalytischen Prüfung sind in der Tabelle (8) dargestellt.

Tabelle 8: Standardwerte inhaltlicher und formaler Komplexität für die Textsorte LITERATUR

Gruppen	inhal. K.	form. K.	N
ERNST L.	95.4	98.5	61
TRIV. L.	90.1	96.2	30
LYRIK	91.4	104.7	29
F, df	7.02	6.45	2/117
p	.00	.00	
eta-Qua.	.11	.09	
eta-Qua. (Disk) : .20			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua.: Ausmaß erklärter Varianz

ERNST L. = ernste Literatur
TRIV. L. = Trivialliteratur

Inhaltlich sind die Textgruppen sehr wenig komplex und liegen deutlich unter dem Mittelwert von 100, den höchsten Wert weist die ernste Literatur auf. Hinsichtlich der FK nimmt die Gruppe 'Lyrik' eine Sonderstellung ein: mit einem Komplexitätsscore von 104.7 liegt diese im Bereich der Gesamtstichprobe der Aufsätze und der geisteswissenschaftlichen Literatur. Die Inhomogenität der untersuchten Textgruppe schlägt sich in der geringen Diskriminierbarkeit durch die Komplexitätsdimensionen nieder. Die Untergruppen sind beide formal komplexer als inhaltlich.

3.2.6 Analyse der Briefstichprobe

Die Textsorte 'Brief' ist in die Gruppen 'normale Privatbriefe', 'todesnah verfaßte Texte' von Widerstandskämpfern und 'Suizidabschiedsbriefe' unterteilt. Mittelwerte und Teststatistiken sind in der Tabelle (9) zusammengestellt.

Tabelle 9: Standardwerte inhaltlicher und formaler Komplexität für die Textsorte BRIEF

Gruppen	inhal. K.	form. K	N
NBP	88.8	100.5	34
TVT	82.6	108.1	28
SUI	80.9	97.5	29
F, df	11.98	16.96	2/88
p	.00	.00	
eta-Qua.	.21	.28	
eta-Qua. (Disk) : .45			

F, df : F-Wert der Varianzanalyse, Freiheitgrade
p : Irrtumswahrscheinlichkeit des F-Werts
eta.-Qua.: Ausmaß erklärter Varianz

NPB = normale Privatbriefe
TVT = todesnah verfaßte Briefe
SUI = Suizidabschiedsbriefe

Wie ersichtlich weisen die unter Aktivation und Streß verfaßten Texte (SUI und TVT) auf der Skala der IK die geringsten Werte aller analysierten Textgruppen auf. Das formale Komplexitätsniveau der normalen Briefe liegt im mittleren Bereich, das der Suizid-Briefe ist demgegenüber reduziert. Hingegen zeichnen sich die todesnah verfaßten Texte durch ein hohes Niveau der FK aus, ein Befund, auf den in der Diskussion noch einzugehen sein wird. Die diskriminative Potenz der Dimensionen ist nahezu ausgeglichen. Die FK ist durchgehend höher als die IK.

4. Synoptische Betrachtung der Textsorten und -gruppen

Die Lage der einzelnen Textsorten und ihrer Untergruppen im zweidimensionalen Komplexitätsraum ist in der Abbildung (3) dargestellt. Inhaltlich hoch komplex sind demnach politische Zeitungstexte und wissenschaftliche Literatur, besonders in den Sozialwissenschaften. Den Gegenpol bilden Texte, die unter extremem Streß verfaßt worden sind: todesnah verfaßte Texte und Suizidabschiedsbriefe. Formal hoch komplex sind Abituraufsätze, einfach strukturiert Texte der Bildzeitung. Während die Quadranten I (geringe FK, hohe IK) und II (hohe FK und hohe IK) des Komplexitätsraums relativ dicht besetzt sind, sind die Quadranten III und IV (hohe FK und geringe FK bei geringer IK) so gut wie unbesetzt: d.h. die vorliegende Stichprobe enthält kaum Texte, die inhaltlich wenig komplex sind und auf der Ebene der FK die Extreme besetzen. Der IV. Quadrant kann - wenn wir nur die geschriebene Sprache zugrundelegen - eventuell durch Schulbuchtexte und sprachliche Materialien der unteren Altersstufen gefüllt werden. Ebenfalls kann hier an die Textsorte 'Comics' gedacht werden. Schwieriger wird es, Texte für den III. Quadranten (geringe IK, hohe FK) zu finden. Wir denken dabei an bestimmte Formen der Schizophasie, die sich unter anderem durch starke Redundierungen (siehe MITTENECKER 1953) und dogmatische Aussagen auszeichnen und, wie ein 'sprachlicher Rigiditätsindex' (SCHWIBBE et al. 1981) ausweist, ähnliche Werte wie die 'todesnahe Sprache' annehmen. Andererseits enkodieren schizophrene Verfasser formal hoch komplex: Satzlängen und Satzschachtelungen liegen deutlich oberhalb normaler Briefe und Aufsätze.

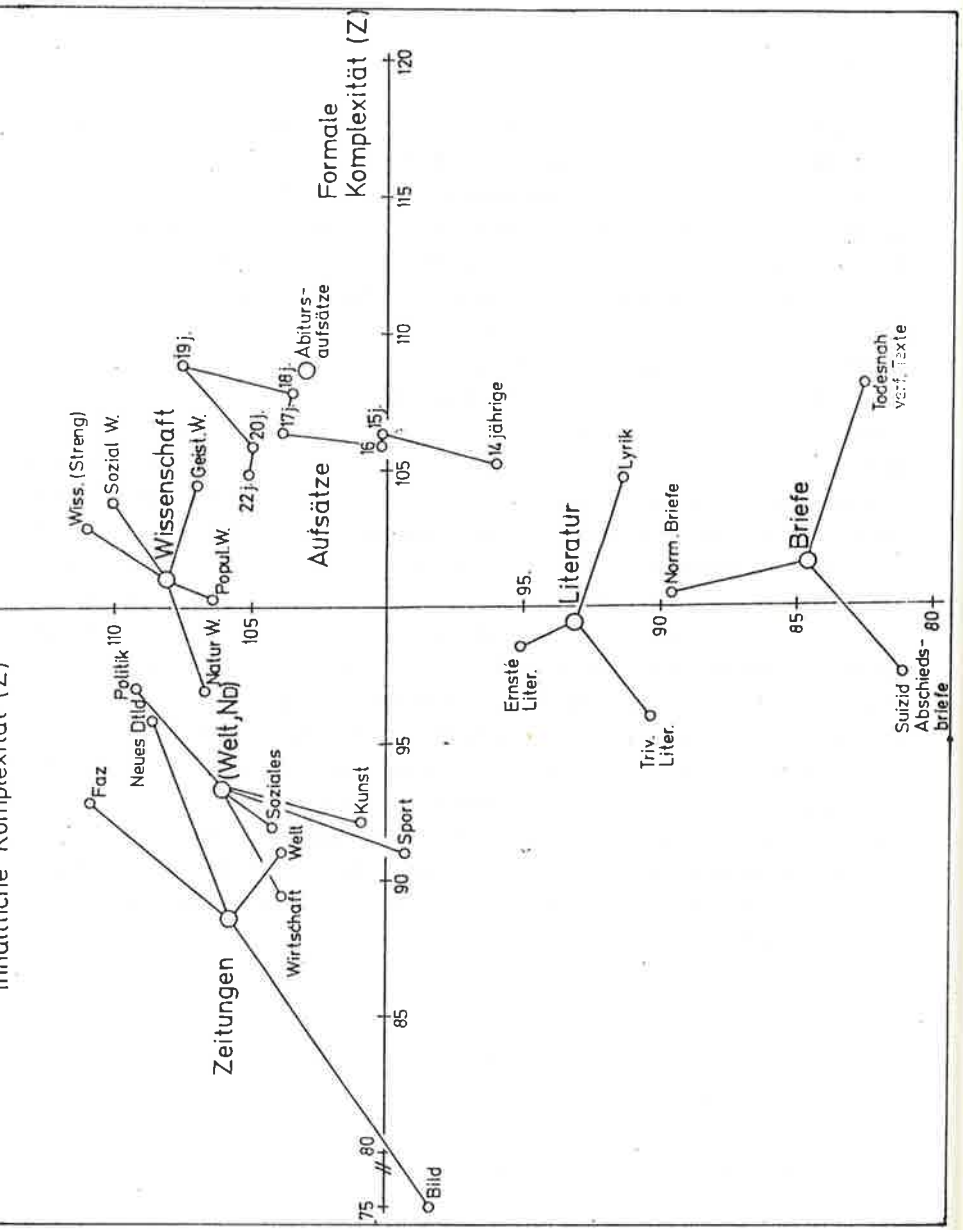


Abb. 3. Lage der Textsorten und ihrer Untergruppen im zweidimensionalen Komplexitätsraum

5. Zusammenfassende Wertung

Hinter der Variation von Variablen der Textkomplexität ist ein Faktor zu identifizieren, der als Abbildung dieses Konstrukts gelten kann. Die Pole dieser Dimension sind als 'prägnant, strukturiert' vs. 'entropiehaltig, komplex' zu bezeichnen. Die formal-syntaktische und die semantisch-inhaltliche Komplexität können als Konstituenten dieses Konstrukts bezeichnet werden. Damit bestätigt dieses Ergebnis die Befunde bisherigeer Forschungskonzepte zu Aspekten der Komplexität bis hin - wie HOFER (1976) ausführt - zu den 'Ideen und Worte zählenden' Talmudisten.

Die Mengentheorie liefert für die am Komplexitätkonstrukt beteiligten Variablen eine die Forschungsansätze übergreifende Beschreibung: Mittels der Mengenmaße und der Maße für die Elemente und ihre Relationen können die verschiedenen Parametrisierungsformen von Komplexität auf grundlegende kognitiv/sprachliche Komponenten der Informationsverarbeitung zurückgeführt werden. Unter kognitiv/sprachlichen Prozessen betrachtet zwingt eine große Informationsmenge, wie sie die Maße der sprachlichen Entropie/Redundanz anzeigen, zur ökonomischen Durchdringung des Informationsangebotes, zur abstraktiven Verdichtung und Vernetzung der Einzelelemente. Demgegenüber steht die von 'einstweiliger Irrelevanz' durch den dogmatisierenden Denkprozeß befreite und geglättete Information: redundant, strukturiert und mit erwartungssicheren Elementen. Die hohe Informationsmenge mit vielen distinkten Elementen, über deren Auftreten beim Rezipienten daher große Unsicherheit herrscht, wird in syntaktisch langen, tiefgeschachtelten und damit komplexen Strukturen vermittelt. Zur Überschaubarkeit der großen Informationsmenge werden die Einzelelemente in kurze/kleine Substrukturen verpackt, ein Ökonomieprinzip, das auch durch die sogn. 'Menzerath'sche Regel' beschrieben werden kann.

Die beiden Dimensionen der Komplexität erweisen sich als ausgezeichnete Diskriminatoren von Texten, die unter verschiedenen psychischen Bedingungen verfaßt wurden: Die Komplexität von Aufsätzen nimmt mit der diachronischen Sprachentwicklung zu, bei den inhaltlichen Aspekten stärker als bei den rein formalen. Nach dem schulischen Höhepunkt formaler Komplexität bei den Abituraufsätzen ist sogar

eine leichte Reduktion dieses Aspekts zu verzeichnen.

Streß reduziert die inhaltliche und formale Komplexität: Aktivatorisch bedingtes Sprachhandeln (hier operationalisiert durch die todesnahe Situation von Briefschreibern) ist durch Redundanz kognitiv-sprachlicher Operationen, dogmatischen Denkstil und Regression auf geläufiges Wortgut bestimmt. Hier ist allerdings noch auf eine Besonderheit hinzuweisen: Die Briefe, die von Widerstandskämpfern stammen, zeichnen sich gegenüber Normal- und Suizidabschiedsbriefen durch eine erhöhte formale Komplexität aus. Dieses mag einerseits darin begründet sein, daß die Verfasser dieser Briefe sich eher aus der Mittel- und Oberschicht rekrutierten, die bekanntermaßen einen eher elaborierten Kode schreiben. Andererseits entstammen diese Briefe einer Anthologie und sind daher sicherlich redaktionell überarbeitet bzw. selegiert.

Niveauvolles Sprachhandeln schlägt sich in komplexeren inhaltlichen und formalen Textstrukturen nieder. Ernste Literatur unterscheidet sich von Trivialliteratur, die BILD-Zeitung von anderen Zeitungen, wissenschaftliche von populärwissenschaftlicher Literatur, wobei allerdings - mit Ausnahme der Bild-Zeitung - die Unterschiede auf der 'formalen' Dimension geringer ausfallen als auf der 'inhaltlichen' Dimension.

Der praktische Nutzen dieser Studie besteht im Nachweis der ökonomischen Schätzbarkeit der Komplexitätsdimensionen durch einige wenige, einfach bestimmbare Parameter der geschriebenen Sprache. Der Validitätsbereich auch so einfach berechenbarer Indikatoren wie Wort- und Satzlängen konnte in dieser Studie über den Aspekt der Wort- und Satzschwierigkeit hinaus auf generelle Konzepte der organismischen Informationsverarbeitung ausgeweitet werden. Dieses bietet die Möglichkeit, strukturelle Aussagen über pädagogische Textmaterialien in ihrer Beziehung zur Sprachproduktion der jeweiligen Zielgruppe zu treffen, Veränderungen von Denk- und Sprachprozessen lassen sich in dieser Hinsicht konsistent beschreiben, Textvariabilitäten und -spezifitäten können anhand dieser Dimensionen aufgezeigt werden. Die Ortbestimmung der analysierten Texte auf den Dimensionen des hier vorgestellten Komplexitätsraums ermöglicht darüber hinaus eine Einordnung weiterer Materialien.

Literatur

ALTMANN, G., LEHFELDT, W., Allgemeine Sprachtypologie. München: Fink (UTB), 1973

BERLYNE, D.E., Konflikt, Erregung, Neugier. Stuttgart: Klett, 1974

CHOTLOS, J.W., A statistical and comparative analysis of individual written language samples. Psychol. Monogr. 56, 1944, 76-111

DEUTSCH, M.P., MALIVER, A., BROWN, D., CHERRY, E., Communication of information in the elementary school classroom. New York: Institute for developmental Studies, Department for Psychiatry, New York Medical College, 1964

DICKES, P., STEIWER, L., Ausarbeitung von Lesbarkeitsformeln für die deutsche Sprache. Z. Entw. Päd. Psychol. 9, 1977, 20-28

ERTEL, S., Liberale und autoritäre Denkstile. Ein sprachstatistisch-psychologischer Ansatz. In: v. THADDEN, A. (Ed.), Die Krise des Liberalismus zwischen den Weltkriegen. Göttingen: Vandenhoek und Ruprecht 1978, 234-255

ERTEL, S., Prägnanztendenzen in Wahrnehmung und Bewußtsein. Z. f. Semiotik, 3, 1981, 107-141

FLESH, R.A., A new readability yardstick. J. appl. Psychol. 32, 1948, 221-233

FLESH, R.A., Measuring the level of abstraction. J. appl. Psychol. 34, 1950, 384-390

FUCKS, W., Nach allen Regeln der Kunst. Stuttgart: DT. Verlagsanstalt, 1968

GAMP, R., Möglichkeiten einer informationstheoretischen Analyse der Werke Immanuel Kants. In: Autorenkollektiv: Untersuchungen zur Sprache Kants, Hamburg: Buske, (=IPK-Forschungsberichte, 26), 1971, S. 137-166

GREENBERG, J.H., Essays in Linguistics. Chicago: University Press, 1961

GUENTHER, U., GROEBEN, N., Abstraktheitssuffixverfahren. Vorschlag einer objektiven und ökonomischen Messung der Abstraktheit/Konkretheit von Texten. Z. exp. ang. Psychol., 25, 1978, 55-74

HAYAKAWA, S.I., Language in Thought and Action. New York: Wiley, 1949

HELLMANN, M., Untersuchungen an östlichen und westlichen Zeitungstexten. Einige Arbeiten der Außenstelle Bonn des Instituts für die deutsche Sprache. In: Literatur und Datenverarbeitung, SCHANZE, H. (Hrsg.), Tübingen, 1972, 66-70

HOFER, M., Textverständlichkeit: Zwischen Theorie und Praxeologie. Unterrichtswissenschaft 2, 1976, 143-150

JENKINS, J.J., JONES, R.L., Flesh's measuring the level of abstraction. J. appl. Psychol. 35, 1951, 68-74

JOHNSON, W., Language and speech hygiene. An application of general semantics. Chicago: Inst. Gener. Semantics, 1939, (=Gen. Seman. Monogr., 1)

KLIX, F., Information und Verhalten. Berlin (DDR): Verlag der Wissenschaften, 1976

KORZYBSKI, A., Science and sanity. An introduction to Non-Aristotalian systems in general semantics. Lancaster (Connecticut): Science Press, 1933

LANGER, I., SCHULZ von THUN, TAUSCH, R., Verständlichkeit in Schule, Verwaltung, Politik, Wissenschaft. München: Reinhardt, 1974

LEWANDOWSKY, T., Linguistisches Wörterbuch. Heidelberg: Quelle und Meyer (UTB), 1979/80 (=Uni-Tb., Bd. 200, 201, 300)

MEIER, H., Deutsche Sprachstatistik. Hildesheim: Olms, 1964

MITTENECKER, E., Perseveration und Persönlichkeit. Z. exp. ang. Psychol. 1, 1953, 5-31, 265-284

NGUYEN, L.T., HENKIN, A.B., A readability formula for Vietnamese. J. Reading 25, 1980, 216-223

OSGOOD, C.E., WALKER, E.G., Motivation and Language behavior: A content analysis of suicide notes. J. abnorm. soc. Psychol. 59, 1959, 58-67

SANFORD, F.H., Speech and personality. Psychol. Bull. 39, 1942, 811-845

SCHWIBBE, M.H., RÄDER, K., Über die Entwicklung eines testäquivalenten Verfahrens zur kontentanalytischen Abstraktheitsmessung. Z. exp. ang. Psychol. 24, 1982, 628-648

SCHWIBBE, M.H., RÄDER, K., SCHWIBBE, G., Rigidität, Perseveration und Sprache: Teil I. Medizin. Psychol., 1981, 207-219

SCHWIBBE, M.H., SCHWIBBE, G., RÄDER, K., HONG, S-K., Validierungsuntersuchungen zu den Dimensionen des kontentanalytisch definierten Dogmatismuskonstrukts. Z. exp. ang. Psychol., 1983, im Druck

SCHWIBBE, G., Intelligenz und Sprache: Zur Vorhersagbarkeit des intellektuellen Niveaus mittels kontentanalytischer Indikatoren. Bochum: Brockmeyer, 1984

SIXTL, F., Ein Verfahren zur Rotation von Faktorladungen nach einem vorgegebenen Kriterium. Arch. ges. Psychol. 116, 1964, 92-96

TAUBER, M., STOLL, F., DREWEK, R., Erfassen, Lesbarkeitsformeln und Textbeurteilung, verschiedene Dimensionen der Textverständlichkeit. Z. exp. ang. Psychol. 27, 1980, 135-146

WECHSLER, D., Die Messung der Intelligenz Erwachsener. Bern-Stuttgart: Huber, 1961

Vocabulaire et Stylistique 1. Théâtre et Dialogue, by Daniel Dugast. Travaux de linguistique quantitative, 8. Editions Slatkine, Geneva, 1979. 293 pages. SFr. 35.--.
Reviewed by S. M. Embleton, Toronto.

This volume has a lengthy subtitle ("Etudes de Lexicométrie Organisationelle sur les théâtres de Corneille, Racine et Giraudoux, sur des pièces de Corneille, Racine, Molière et Beaumarchais, sur un entretien entre Maurice Clavel et Philippe Sollers, précédées d'un historique des méthodes quantitatives en lexicologie et des fondements d'une explication nouvelle: UBER") which describes fairly accurately its content. In some ways it is more a collection of essays relating to the theme of theatre and dialogue, rather than a tightly structured and cohesive book. Each essay is in a sense "incomplete" (in that many directions for future research are only sketched or suggested), but yet they complement one another, leaving the reader with a satisfied feeling of completenesss and an impression of the field rather like that of a mosaic. The second volume of Vocabulaire et Stylistique (in preparation) will apparently be similar in structure, but centred around the novel.

The "Introduction" (pages 13-64) is divided into three sections. The first is a historical sketch of approaches to lexical richness/abundance (la richesse lexicale) covering the work of J.B. Estoup, G.K. Zipf, B. Mandelbrot, M. Petruszewycz, J.B. Carroll, J.W. Chotlos, W. Johnson, P. Guiraud, G. Herdan, Q. Rubet, C.E. Shannon, J. Tuldava et al, E. Brunet, and H.-D. Maas, and concluding with a review of the "UBER" theory (presenting the equation of the curve giving the number of "word" (more accurately vocables or "dictionary entries") in a text as a function of its length). The section explains and exemplifies the notion "chreode", originally borrowed from biology, which designates "un champ morphogénétique soumis á l'axe du temps" (essentially a developmental pathway). The third section deals with a topic often neglected in statistical applications in the humanities, the degree of confidence which one can attach to one's predictions and calculations.

The remainder of the book, entitled "Le Théâtre du Répertoire français", is divided into three major parts. The first (pages 65-116) of these is more general in nature, looking at the works of Corneille, Racine, and Giraudoux, and is itself divided into three subsections, treating in turn the chreode used by Giraudoux in his theatrical works (different from the one used in his novels), a study of genre and chreode in Corneille and Racine, and a study of vocabulary growth from play (i.e. chronologically) in Corneille and Racine. Giraudoux and Corneille are shown to tend towards a more "sober" style (i.e. use of a less rich vocabulary) with increasing age, where as the opposite is true of Racine. The many graphs throughout this part are a useful supplement to the charts and figures presented in the next. The second part of the book (pages 117-186) is more specific, dealing with individual plays of the authors. The first subsection deals with regular "slices" of the text and vocabulary growth in Corneille's Sertorius and Sophonisbe and in Racine's Andromaque, Mithridate, and Phèdre. The second subsection looks at individual roles as characterized by different lexical richnesses in Corneille's L'Illusion Comique, Molière's L'Avare, and Beaumarchais's Le Mariage de Figaro. The third subsection centres on vocabulary growth in a comedy of Racine (Les Plaideurs) and a drama of Hugo (Ruy Blas). Again, many graphs provide a welcome and useful visual supplement to the charts and textual discussion. The third major part of the book (pages 187-272) is a detailed study of vocabulary, chreodes and growth in a one-hour 1976 conversation (transcribed, comprising 5113 words) between Maurice Clavel and Philippe Sollers. The reader could be forgiven for at first wondering why an analysis of such a conversation would be appropriate in a book devoted to theatre, but one should not forget that the book is actually entitled theatre and dialogue.As Dugast Explains (page 189), "sous une forme plus simple que celle qui rassemble plusieurs locuteurs, le dialogue peut être le lieu de l'étude de l'échange verbal." Thus a detailed study of a dialogue in some sense gets at the root of what theatre is all about. The pages of charts and graphs easily outnumber the pages of textual discussion in this section, but the data and

illustration are necessary to back up the detailed in-depth analysis. The book ends with a brief conclusion (pages 273-279), aptly subtitled "Sur les perspectives ouvertes par la lexicométrie", followed by some notes and a discussion of sources.

I have few minor quarrels with certain aspects of the production of this book. Some pages seem to be in the wrong place, at least in my copy of the book. For example, there is a page entitled "Conclusion" (with no page number) erroneously inserted between pages 172 and 173; the graphs referred to on page 159 as "preceding" in fact follow, on pages 160 and 161. For this reason and for simple ease and accuracy of reference, it would have been helpful if all of the graphs could have had short captions. There are very few typographical errors (although there tend to be more later in the book and virtually none in the early part of the book), but some (e.g. the fact that chi is sometimes spelled with c and sometimes with k; a reference to "page 822" which should read "page 157") may belie the fact that the book (as mentioned above) is more a collection of related essays (one even carries a separate date) possibly written at different times or assembled from different sources.

On first glance, this book appears to be centred around mathematics and the statistical manipulation of data; there are pages upon pages of graphs, charts, formulae, derivations, and numbers. Despite all these, this is not what the book is really about; its heart and soul lie in Dugast's realization that the numbers are not an end in themselves, but rather a catalyst or tool, aiding one by giving a heightened perception of certain literary facts and allowing one a means of making certain impressions objective. Two quotes will illustrate this point: "C'est du moins ce que tenterait de montrer l'analyse quantitative. Elle met en évidence quelques faits exacts qui demandent interprétation. C'est en retournant au texte, seule source de l'intuition, que ces quelques faits stylistiques trouveront leur explication". (page 82); and "Un indice seul ne signifie rien; il indique, dans un contexte donné, que quelque chose se passe. ...Le style de l'auteur ... ne sera jamais enfermé dans une série de chiffres!" (page 115). The

linguist often gets sufficiently excited by his/her analysis in its own right that the object of such analysis, namely language itself, is forgotten; likewise, the pitfall of forgetting the text awaits the quantitative linguist. Dugast skillfully avoids such traps, although at one point he has to check himself consciously: "Mais ne sommes-nous pas alors en train de jouer davantage avec les nombres qu'avec les mots?" (page 35).

Dugast's book is a useful contribution to the ever-increasing literature in quantitative linguistics. One can look forward to the publication of the forthcoming parallel work on the novel, Vocabulaire et Stylistique 2: Le Roman.

Dynamisme du texte et stylostatistique: élaboration des index et de la concordance pour Alice's Adventures in Wonderland. Problèmes, méthodes, analyse statistique de quelques données, by Philippe Thoiron. Travaux de linguistique quantitative, 11. Editions Slatkine, Geneva, 1980. 691 pages. SFr. 110.--.
Reviewed by S. M. Embleton, Toronto.

This book, a revised version of the author's doctoral thesis, is mostly devoted to a detailed account of a construction of a word frequency index and a concordance for Lewis Carroll's Alice's Adventures in Wonderland. At the time of writing this review, the conception volume, the actual index and concordance by Philippe Thoiron and Alain Pave, has not yet been published; it will appear as volume 16 in the same series. Although the two books logically go together and perhaps should therefore be reviewed together, this is obiously impossible at the present moment. However, there is much of general value, not just specific to Alice's Adventures in Wonderland, to be gained from Thoiron's discussion of methodology, the various problems he encountered, and his solutions to these problems, and thus the book can be read with profit even without the companion volume.

After a brief introduction (1-25) to the purposes and goals of the book as a whole, Part One (26-321, divided into six chapters) is concerned with a variety of topics related to the construction of the concordance and index, including lemmatization and grammatical tagging of the text, the coding systems used, the description of the hardware and software of the computer available to Thoiron, and above all of the problems resulting and the decisions which had to be made. It would be impossible to discuss any of these meaningfully in the space available here, but there are perhaps two points which are worth making. First, the computer used (a CAE 510, with paper tape input and roughly 12K of memory in total) is really pathetically inadequate for the current undertaking (Alice's Adventures in Wonderland is over 27,000 words long); Thoiron calls it (almost lovingly it seems at times!) "un ordinateur de faible puissance". It is quite incredible though what Thoiron can achieve with such inferior equipment, proving perhaps that large research

grants of fancy equipment are no substitute for the scholar's ingenuity when it comes to achieving results. Second, Thoiron does not sweep anything under the carpet; all decisions to be made and problems encountered are fully discussed from all points of view. Thus any decision of solution is seen to result from careful reasoning. The reader may not agree with all the decisions, but that is not important. Thoiron's emphasis throughout is on methodology, which of course makes the book valuable reading for anyone about to embark on a similar enterprise.

Part Two (322-530, divided into three chapters) is less successful than Part One and considerably more difficult to read. The discussion of the concepts "norm" and "deviation" and their role in stylostatistics is valuable, but the unduly lengthy sections on perception and memory in Chapter 1, although interesting, seem rather out of place and irrelevant to the topic at hand. Chapter 2 looks in detail at questions related to the distribution of the definite article "the" and the extent to which it is a stylistic index. The related question of the distribution of nominals, in particular of definite nominals, is also examined. Chapter 3 returns to the concept of norm, this time viewing it as a "dynamic" norm rather than a "static" norm and thus able to change gradually and in a principled way throughout a literary work. Since style is typically defined as "deviation with respect to a norm", this change in the concept of norm is obviously of great importance. To illustrate, Thoiron turns again to the definite article, nominals, and definite nominals. The increasing density of the definite article throughout Alice's Adventures in Wonderland is the result of two factors (growth in the density of nominals and growth in definition) and can be modelled using linear regression. The culmination of the chapter is when Thoiron is able to show that Alice has passed from childhood (with mostly female contacts; even contacts with the father are mediated through the mother) to adulthood (where male contacts are much increased), as reflected in the increasing number of male participants in the action, most of whom have names of the type the + nominal (the March hare, the hatter, the knave, the white rabbit, the mock turtle, etc.). It is fascinating to see how a statistical analysis can sequentially generate such hypotheses --

'les anomalies de la distribution de THE conduisaient à une étude de la répartition des noms dont les résultats permettaient de mettre en évidence de nouveau la structure dichotomique du texte" (529). After this climax, the concluding chapter, the various appendices, and the bibliography (divided into six sections, according to topic) will definitely seem like a slow dénouement! The appendices themselves are close to a hundred pages long, and most readers would agree that many of the tables and graphs provided unnecessarily add to the bulk of an already weighty volume. At the very least, if it really was felt essential to include all of these tables and graphs, they should have been included in the text at the point where they were being discussed. As it stands now, the table or graph number is provided, but no page reference is given, making it even more awkward for the reader to flip back and forth.

Thoiron's work presented in this volume has been monumental, both in its scope as well as in its sheer size and inclusion of detail. Many readers, especially newcomers to the field, will be dissuaded from reading this work, because of its length (nearly 500 pages) and rather daunting appearance. There are places where Thoiron diverges from the topic or dwells overly long on some point, but in general the length is justified since it is a result of completeness and thoroughness. Part One of the book should be required reading for anyone thinking of constructing a concordance, particularly if their budget dictates less-than-ideal computing equipment. Chapter 3 of Part Two is the real "gem" though, in showing so elegantly what computer-aided stylistic analysis can achieve; it constitutes a perfect rebuttal to any literary scholar who persists in believing that computers have no place in stylistic studies.

CURRENT BIBLIOGRAPHY

ABBREVIATION

PSML Prague Studies in Mathematical Linguistics, 8, 1983

GENERAL

1. ALEKSEEV, P.M., GRIGOR'EVA, A.S.: K probleme stabil'nosti častot lingvističeskix edinic [On the problem of quantitative stability of linguistic units]. Inženernaja lingvistika i optimizacija prepodavanija inostrannyx jazykov v vuze. Leningrad 1983, 124-130.

2. KRÁLÍK, I.: Some notes on the frequency-rank relation. PSML, 67-80.

3. ORLOV, Ju.K., ČITAŠVILI, R.Ja.: O raspredelenii častotnogo spektra v malyx vyborkax iz raspredelenij s bol'šim čislom vozmožnyx sobytij [On the distribution of frequency spectrum in small samples from populations with a large number of possible events]. Soobščenija Akademii nauk Gruzinskoj SSR 108, No 2, 1982, 297-300.

4. dies.: Nekotorye problemy statističeskogo ocenivanija v otnositel'no malyx vyborkax [On some problems of statistical estimation in relatively small samples]. Soobščenija Akademii nauk Gruzinskoj SSR 108, No 3, 1982, 513-516.

5. dies.: Obobščennoe Z-raspredelenie, porožd ajuščee izvestnye 'rangovye raspredelenija' [Generalized Z-distribution generating the well-known 'rank distributions']. Soobščenija Akademii nauk Gruzinskoj SSR 110, No 2, 1983, 269-272.

6. dies.: O statističeskom smysle raspredelenija Cipfa [On the statistical significance of Zipf's law]. Soobščenija Akademii nauk Gruzinskoj SSR 109, No 3, 1983, 505-508.

7. POLINSKAJA, M.S.: Metod kvantifikacii svjazei meždu elementami jazykovoj struktury [A method of quantifying the relations between elements in the structure of language]. Učenye zapiski Tartuskogo gosudarstvennogo universiteta 658, 1983, 82-100.

PHONOLOGY

8. KAMIMURA, R., ODA, J.: A study on the optimal latent structure of language. Mathematical Linguistics 14, 1984, 206-221.

9. MAEKAWA, K.: A theory for vowel confusion: Based on the data from two Japanese dialects, Tsugaru and Izumo. Mathematical Linguistics 14, 1984, 149-162.

MORPHOLOGY

10. ISHII, M., NOMURA, M.: Word-formation of compounds in "Japanese scientific terms mechanical engineering", on the basis of classification of stems. Mathematical Linguistics 14, 1984, 163-175.

11. PUSZTAY, I.: Verbalpräfixe im Deutschen und Ungarischen. Ein Beitrag zum Thema: Ungarisch als Fremdsprache. Bachofer, W., Fischer, H. (Eds.) Ungarn - Deutschland. Studien zur Sprache, Kultur, Geographie und Geschichte. München: Trofemik, 1983, 63-76.

SEMANTICS

12. GOEKE, D., KORNELIUS, J.: Empirie in der Wortfeldanalyse: Quantitative Bestimmungen zu semantischen Ähnlichkeiten zwischen englischen Bewegungsverben. Anglistik und Englischunterricht 13, 1981, 133-146.

13. dies.: Wortfelder aus bemessenen Ordnungen. Trier: Wissenschaftlicher Verlag, 1984, 116 pp.

14. MORIMOTO, H.: An analysis of Chinese character (KANJI) by the semantic differential Method. Mathematical Linguistics 14, 1983, 129-137.

TEXT ANALYSIS

15. ALASANINA, G.G., ORLOV, Ju.K.: Opyt statističeskoj klassifikacii tekstov (na materiale sočinenij "Kartlis sxovreba") [An attempt at statistically classifying texts (based on material of the Old-Georgian historical work "Kartlis cxovreb")]. Izvestija Akademii nauk Gruzinskoj SSR. Serija istorii, arxeologii, ėtnografii i istorii iskusstva 1983, No 2, 57-79.

16. CONFORTIOVÁ, H.: On prepositions in non-fiction style. PSML, 31-42.

17. KLIMEŠ, L.: On some quantitative aspects of the sentences in Palacký's works. PMSL, 93-100.

18. KOROLEV, E.I., KORSAKOVA, I.I., SAFRANOVA, M.V.: Častota upotreblenija slov v tekste i ix leksičeskie xarakteristiki [The frequency of word usage in texts and their lexical characteristics]. Naučno-texničeskaja informacija, Ser. 2, 1984/2, 8-14.

19. KRÁMSKÝ, J.: A stylostatistical examination of conjunctions in modern English. PSML, 81-92.

20. LIŠKOVÁ, Z.: Intersentential connections in journalistic texts. PSML, 111-119.

21. LUDVÍKOVÁ, M.: Quantitative Aspects of verb categories (based on present-day non-fiction texts). PSML, 19-30.

22. MANASJAN, N.S.: Ob odnoj statističeskoj modeli upotreblenija terminov v tekste [A statistical model of terminological usage in texts]. Inženernaja lingvistika i op-

timizacija prepodavanija inostrannyx jazykov v vuze. Leningrad 1983, 131-136.

23. MATVEEVA, E.G., PANOVA, N.S., RYŽOVA, E.Ju., FEDORČUK, A.V.: Statističeskoe issledovanie sintaksičeskoj struktury naučno-texničeskix tekstov [Statistical research on the syntactical structure of scientific and technical texts]. Voprosy informacionnoj teorii i praktiki, No. 51, 1984, 70-80.

24. MIZUTANI, S.: Lexical analysis of Japanese popular songs 1929-1944. Mathematical Linguistics 14, 1984, 185-206.

25. NEBESKÁ, I.: Compound/complex sentences in non-fiction texts. PSML, 53-65.

26. ŠTĚPÁN, J.: On some aspects of syntactic complexity in fiction style. PSML, 101-110.

27. TĚŠITELOVÁ, M.: Some quantitative characteristics of non-fiction texts in present-day Czech. PSML, 9-18.

28. TEŠITELOVA, M. [TĚŠITELOVÁ, M.]: O tak nazyvaemom delovom, nexudožestvennom stile s kvantitativnoj točki zrenija [A quantitative approach to so called business style]. Učenye zapiski Tartuskogo gosudarstvennogo universiteta 658, 1983, 136-148.

29. UHLÍŘOVA, L.: Simple sentence structure from the quantitative point of view (based on present-day Czech non-fiction texts). PSML, 43-51.

Psycholinguistics

30. HONG, S-K.: Kognitive Komplexität und Dogmatismus - theoretischer und empirischer Zusammenhang. Diss. Göttingen 1982.

31. RÄDER, K., SCHWIBBE, M.H.: Nonlineare Beziehungsanalysen zum Polyanna-Prinzip in der Sprache. Psychologische Beiträge 24, 1982, 226-295.

32. dies.: Zur Bedeutung des "Polyanna-Prinzips" bei der Attribuierung von Familiennamen. Psychologie 30, 1983, 448-457.

33. SCHWIBBE, M.H., RÄDER, K., SCHWIBBE, G.: Rigidität und Sprache. Teil II: Empirische Untersuchungen zur Variation von Sprachparametern aus dem Bereich des Rigiditätskonstrukts. Medizinische Psychologie 8, 1982, 1-19.

34. SCHWIBBE, M.H., SCHWIBBE, G., RÄDER, K., SOO-KEE HONG: Untersuchungen zur Validierung der Dimensionen des kontentanalytisch fundierten Dogmatismus-Konstrukts. Zeitschrift für experimentelle und angewandte Psychologie 3, 1983, 639-654.

Historical Linguistics

35. DIETZE, J.: Frequenzstatistische Möglichkeiten zur Nutzung der linguistischen Datenverarbeitung in der Diachronie der russischen Sprache. ZPhSK 37, 1984, 355-360.

36. PHILLIPS, B.S.: Word frequency and the actuation of sound change. Language 60, 1984, 320-342.

Language Variation

37. GROTJAHN, R.: On the use of quantitative methods in the study of interlanguage. Applied Linguistics 3, 1983, 235-241.

38. THELANDER, M.: A qualitative approach to quantitative data of speech variation. Romaine, S. (Ed.): Sociolinguistic Variation in Speech Communities. London 1982, 65-83.

39. VEJLERT, A.A.: Lingvostatičeskie universalii kak predmet tipologičeskix sopostavlenij (k probleme metoda) [Linguistic and statistical universals as the object of typological comparison (methodological problems)]. Sopostavitel'nyj analiz grammatičeskoj i leksičeskoj semantiki.

Kujbyšev 1982, 3-12.

DIALECTOLOGY

40. GOEBL, H.: Parquet polygonal et treillis triangulaire: les deux versants de la dialectométrie interponctuelle. Revue de linguistique romane 47, 1983, 353-412.

41. KRISTENSEN, K., THELANDER, M.: On dialect levelling in Denmark and Sweden. Folia Linguistica 18, 1984, 223-246.

All contributions to the CURRENT BIBLIOGRAPHY should be sent to Prof. Dr. Werner Lehfeldt, Universität Konstanz, Fachgruppe Sprachwissenschaft, P.O.Box 5560, D 7750 Konstanz.