COMPUTATIONAL ANALYSIS OF INTERFERENCE PHENOMENA ON THE LEXICAL LEVEL(*)

W. Skalmowski and M. Van Overbeke

0. Summary

This contribution presents the results of comparison of Dutch texts written by bilinguals[1] (speaking French and Dutch), with Dutch texts regarded as STANDARD WRITTEN DUTCH. The attention was focussed on French loan-words appearing in both types of texts and the differences in their use. Certain generalizations as to the mechanisms of interference are suggested.

1. Materials

The materials used for the present contribution belong to two groups :

- *group A :* texts written by francophones with ca. 6 years of Dutch training. These texts represent what we call *Francophone Written Dutch* (below FWD).
- *group B :* Texts from recent contemporary Dutch literature by both Dutch and Flemish authors. They will here represent *Standard Written Dutch* (SWD).

(*) We are greatly indebted for the assistance of our colleages Mr.L.DE BUSSCHERE, who prepared all computer programs needed in this investigation, Mr.R.EECKHOUT, who helped us with many suggestions as to the possibilities of information processing techniques and with critical remarks concerning the linguistic aspects of our problem, and - last but not least - the Direction of the MATHEMATICAL CENTRE of the University of Louvain, who put at our disposal the IBM-360 computer.

The texts of group A were written by 400 francophone 18 year-old pupils in the highest classes at the 61 private secundary schools in Brussels and its suburbs. This sample represents one fifth of the total population. From every pupil we obtained two Dutch compositions, one of them a piece of homework written in November 1967, another an examination composition from December of the same year. The reasons for this choice are evident, since the pupils can call in their parents' and their dictionaries' assistance in the first situation but not in the second.

From every composition the first 125 words were put on punch-cards together with coded information as to their source. In this way a corpus of ca. 100,000 words was compiled. In order to allow for comparison of relative parameters such as word-spread, vocabulary-growth etc., it was later divided into two parts each containing ca. 50,000 words (parts 1 and 2 below). The texts of group B, i.e. the SWD, were obtained by putting together extracts from literary work by 10 contemporary authors. This anthology gave us a corpus of some 10,000 words.

The first part of group A reflects ca. 50 different subject-matters, whereas the SWD-anthology reflects only 10 subject-matters or "themes". So the disproportion of corpora is outweighed by a *themes/tokens* ratio which is 1/10 in both corpora. In order to estimate the influence of subject-matter on word-choice and especially on the rate of vocabulary-growth, a comparison was made between the 10-author-corpus and a fragment of ca. 10,000 words from one single author. The results show that the vocabulary-growth remains almost unchanged, i.e. that the diversity of subject-matters does not substantially influence the numerical values of growth rate (*fig.1*). In accordance with this result, we suggested that each of these texts (groups A and B) be regarded as written by one single person.

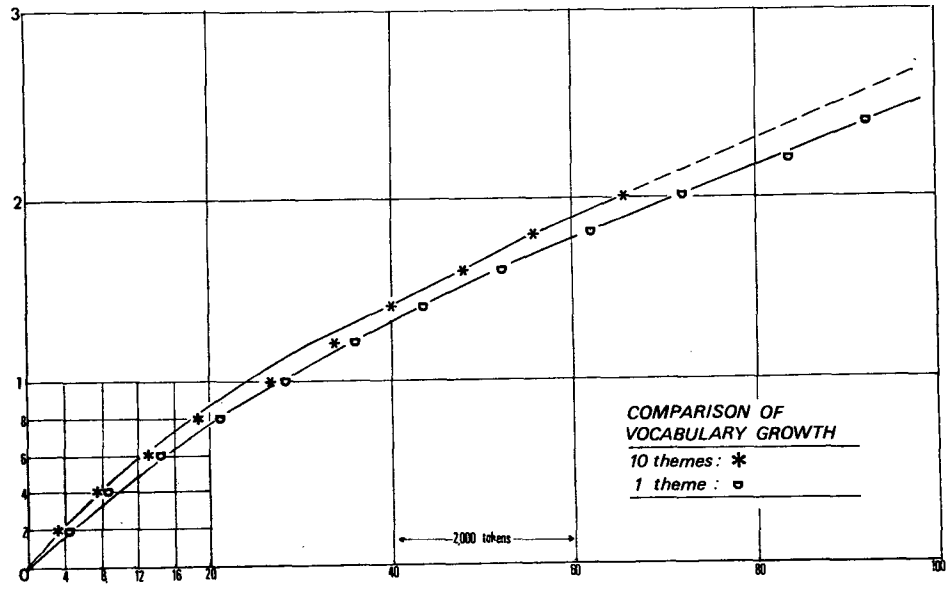

FIG. 1

2. *Lexical interference*

The main purpose of this contribution is to test and verify certain non-computational insights made about language interference in general. Dutch presents a very poignant example of this phenomenon since its vocabulary contains a very large number of French loan- and foreign words and there is still an "open door" allowing the intrusion of lexical gallicism in practically unlimited quantity. Thus the Dutch vocabulary holds a lot of parallel lexemes of both origins, e.g. *analyse/ ontleding, fenomeen/verschijnsel, deceptie/ontgoocheling* etc.

This situation strongly resembles that of English with its Anglo-Saxon and Romance words, although the semantic differentiation of such word-pairs seems to have progressed much more there. Whereas the native Dutch speaker plays both keys with an unbiased ease, for the Belgian francophone this ambiguous situation produces certain constraints and difficulties, which have visible effects on word-choice, growth rate of foreign words and vocabulary size in general.

For reasons of simplicity our investigation did not adopt the usual distinction between loan-words and foreign words since this is based on the different degrees of integration of foreign lexemes, measured by differences in pronunciation, social acceptability within the speaking community and certain prescriptive arrangements such as their inclusion in vocabularies and dictionaries, whose authority is generally accepted. As the aim of our investigation was to find ways of providing numerical values for interference phenomena, we proceeded in a purely descriptive way, using only etymological criteria to distinguish between original and foreign lexical elements. Thus we considered units containing either lexical or morphological elements, or both, as loan-words. So *bonjouren* with its French lexical element was entered, but also *trotseren* because of its French word-fromational part. Composita containing only one foreign element (e.g. *avondtoilet*) were treated as loan-words unless this element had already been entered as an autonomous word. No distinction was made between foreign words included in the Standard Dutch Vocabulary of van Dale [2], (e.g. *assaut*) and those which are not mentioned there (e.g. *auberge*), both examples occurring in our investigation materials. Since the computer program did not provide for lemma-like items, all different morphological forms and derivations of words were regarded as different types; thus *expressie, expressief, ex-*

pressionist etc. are counted as different items. Also for reasons of simplicity all non-French foreign words are relegated here to the category of pure Dutch items.

3. *Lexical interference and word-length*

As a first approximation test the percentage of foreign words in the vocabulary in both FWD- and SWD-texts was established. The results are as follows :

	TOKENS	TYPES	*log*TYPES	FOREIGN TYP.	*log*F.T.	% F.TYPES
FWD	47,307	5,653	0.8375	648	0.4954	11.85
SWD	10,358	2,616	0.8807	141	0.4285	5.38

The difference of foreign vocabulary ratio in both groups results in distributional differences of words of diverging letter-number. Though the overall word-length of tokens in both groups is nearly identical (4.51 for SWD and 4.61 for FWD) an application of the chi-square test proved the divergences of word distribution (words belonging to different word-classes) to be highly significant. The average word-length of types (M) is different in both groups :

	M	σ
FWD	7.85	2.97
SWD	7.03	2.72

As the pronunciation of French words is in most cases adapted to the Dutch ones (and this is reflected in the orthography), it was not plausible to suppose that this divergence was due noly to the proportional difference of foreign words. It was found that the divergence was partly due to the use of composita in FWD; their distribution differs considerably from SWD. This

is strikingly evident for word-length 10 (*fig.2*) The fact that FWD-authors would "switch in" this Dutch-formational device in cases where the Dutch native speaker does not, shows that francophones are "over-aware" of this means of translating the French genitive construction by a Dutch compositum (e.g. *pot de fleurs> bloempot*). This fact strengthens the assumption made in this paper, that the lexical level of language is very closely connected with higher (syntactic) levels, so that statistically statable facts may be explained only in connection with certain more general models of speech production.

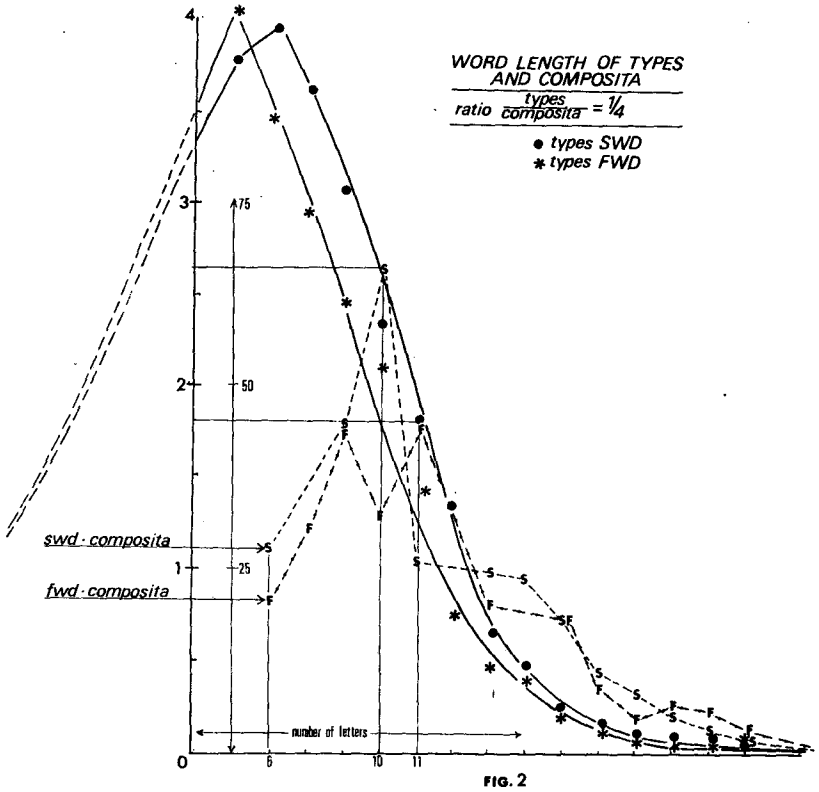

FIG. 2

4. An interference model

The interference model presented here consists of two parts : the syntactic one, containing also the word-formational devices, which may be thought of as a generative device of the kind described by N.CHOMSKY and other generativists; the second part, called *the lexical morpheme store*, is thought of as consisting of entries "written down" in terms of conceptual symbols, provided with actual linguistic interpretations. These "interpretations", which in a very simplified manner may be identified with words *tout court*, are picked out of the store and "fitted" into previously constructed sentence forms. In other words, we assume that the sentences are formed according to semantic requirements before the actual words have been chosen. This last routine goes on in a semi-automatic way, which may be visualized as picking the required lexemes - according to the entries in terms of conceptual symbols - out of a magnetic tape gliding under a reading device of some sort.

For the case of a bilingual speaker, we can imagine the procedure as a tape with three different tracks, the middle one containing the "entries" , the other two the respective actual morphemes, in casu Dutch and French (D and F in *fig.3*). Speaking in one of the two languages demands a switch-over to one of the external tracks. It may be assumed that, in the case of a monolingual Dutch speaker, the cells contain the parallel French and Dutch words in an unordered manner, whereas with a francophone a bias exists towards the French loan-word (e.g. column 1 on fig.3 : *phénomène > fenomeen (verschijnsel)*). This explains the predilection for loan-words even within the limits of the "basic vocabulary" and the more so with words of low frequency. Other variants of speech production behavior are possible; for instance the hypercorrect option 1 → *1* in column 2, where the speaker consciously reaches for the more distant lexeme, and the case of pure borrow-

ing, which may be conceived of as an automatic switch-over to the French side, wherever the Dutch track is blank or whenever the bilingual's competence fails to furnish a good Dutch word or synonym. In this process the French lexeme is placed in the cell on the Dutch side (cf.column 3 where ∅ is the lacking word).

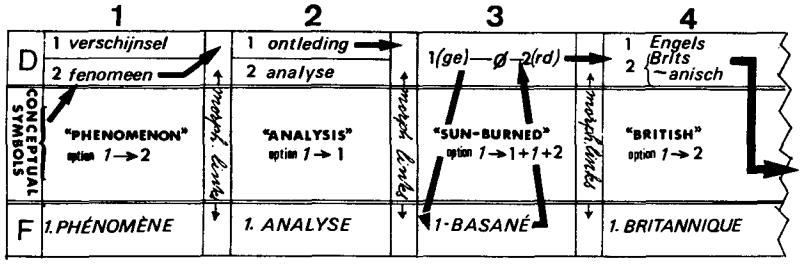

FIG. 3

We assume that the word-formational rules belong to the syntactical part. Thus the reshaping of new French borrowings (cf. the loan-adjective *gebasaneerd* composed of the French *basané* , whose counterpart is lacking in the Dutch track, and of two Dutch affixes *ge-* and *-d*) is done in the grammatical part of our model. As a matter of fact, this assumption is a heuristic over-simplification, because certain grammatical morphemes are in fact borrowed, cf. the endings *-eren, -atie, -age* etc. In order to explain this phenomenon, one could argue on the fact that in many cases whole word-items are introduced to the lexical store and activate the analogy mechanism, - but this problem would lead us beyond the scope of the present investigation.

5. *A code-switching theory*

There has been much speculation about the possible principle of lexeme order in the store, some ordering being a necessary condition of efficient re-coding. Much discussion, too, has

been devoted to the so-called ZIPF-law [3]. The most convincing explanation was that suggested by HERDAN [4], namely that an ordering of items by decreasing frequency would diminish the number of operations necessary to identifie a given item. *"Let us ... assume that the arrangement of the entries is systematic according to frequency of occurence in descending order of frequency, so that the most frequent word has rank 1, the second most frequent word rank 2, and so on. If in such a dictionary, that is one in which words are arranged in order of decreasing frequency and increasing order of rank, the look-up procedure is one of successive comparison, the word of rank r will require r look-up operations, and since this word occurs - the Zipf-law assumed - C/r times, the total number of look-up operations required to locate a word is C (the constant in the Zipf-law, formulated as* $r.f_r = C$ *). Thus for n words contained in the dictionary, nC look-up operations will be required. On the other hand, we know that for the Zipf-law the total number of occurences (the text length in terms of word number) and thus the total number of words to be searched, is given by*

$$N = \int \frac{C\ dr}{r} = C\ \log_e n$$

It follows that the average number of look-up operations per word is

$$A_n = nC/C\ \log_e n = n/\log_e n$$

(...) This compares favourably with the n/2 look-up operations which would be needed under the scheme described above, which makes no use of the frequency element.")

Within the framework of our model it would mean that the winding and unwinding of the tape takes considerably less time than in the case of wholly random distribution. The question remains of what principle underlies the differentiation of item possibility. Here too, the concept of "pigeon-holing" of semantic information proposed by HERDAN [4], seems to be the most plausible.

In other words, the "conceptual symbols" do not represent separate pieces of the *univers de discours* taken at random, but are probably ordered by some classificational system, resembling the biological classification.

6. *Word content and entropy*

To test this hypothesis we divided the FWD material into three frequency-classes (group I: absolute frequency 1, group II: frequency 2 and 3, group III : frequency above 3) and examined the samples of these groups according to their distribution within the classificational system applied by L.BROUWERS in his Dutch thesaurus *HET JUISTE WOORD* [5]. The supposition was that in the event of ordering of some kind, the distribution of items among the "content classes" in the thesaurus (expressed as entropy and redundancy) would be different for various frequency groups, and further, that in the event of the "pigeon-holing" suggested by HERDAN, the redundancy should increase for groups of items with higher frequencies. Such an increase was in fact observed, as the reader can conclude from the following table:

	FREQUENCY 1	*FREQUENCY 2-3*	*FREQUENCY > 3*
H	*5.099*	*4.892*	*4.854*
R	*0.15*	*0.18*	*0.19*

Thus it seems that some "natural order", reflecting a classification of concepts according to their content, is at least one of the causes differentiating the relative frequencies of words. This result is compatible with the fact that the diversity of subject-matter (cf.1) does not considerably alter the growth rate of vocabulary. This statement need not

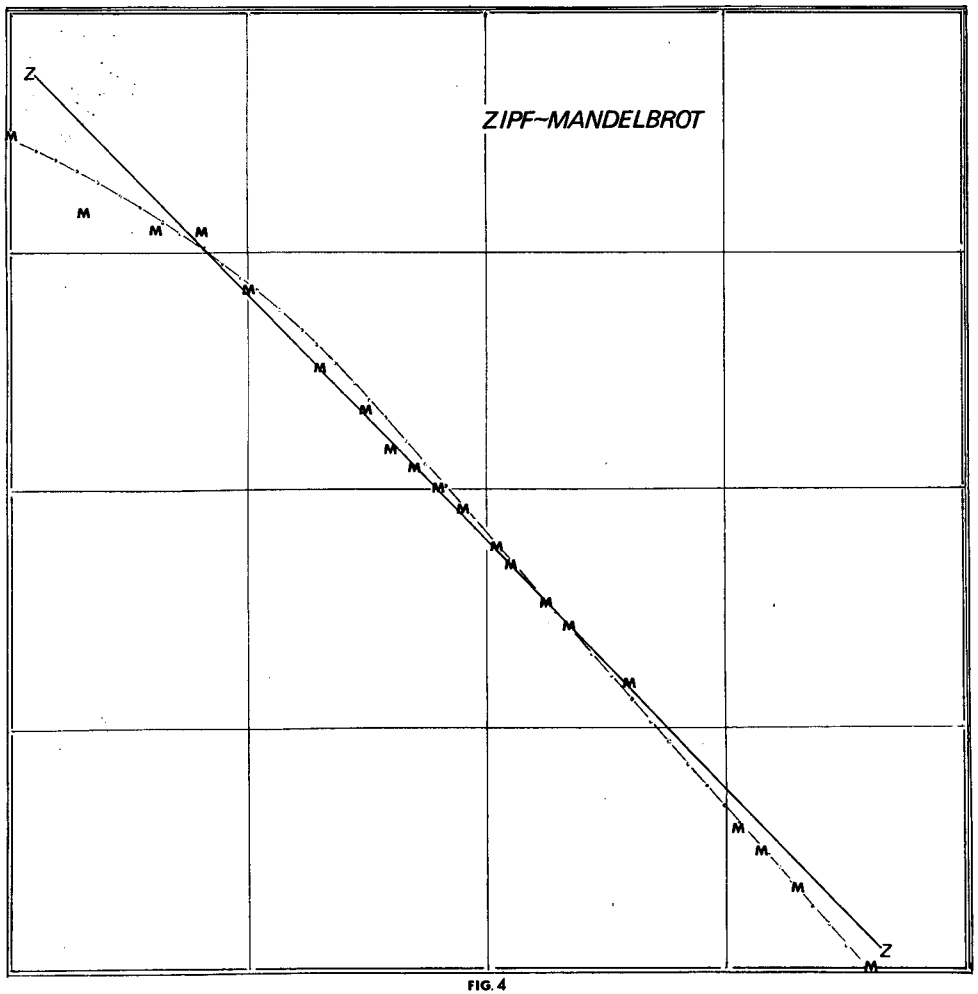

FIG. 4

rule out other devices allowing quick interconnections between words belonging to the same content-group but differing in frequency; (cf. the so-called *association of related concepts* suggested by P.A.KOLERS [6]). However, the basic principle of order seems to be of a statistical kind, as is proved by the perfect fit of the rank-frequency distribution with the theoretical distribution according to the ZIPF-MANDELBROT formulation (cf.fig.4). The correlation coefficient between the observed and the theoretical distribution is 0.993!

7. Consequences

The assumed model has consequences, which have been empirically tested:

1. The assumed model, and especially the process of *blank-filling* of the Dutch track with French morphemes, presupposes that in general the FWD-writers will use a greater number of foreign words than the SWD allows. This fact is already apparent from the overall percentage of foreign elements in FWD (cf. fig.5) In particular the foreign words should appear more frequently in proportion to the increase of text-length). The investigation of vocabulary growth rate has in fact shown that this is the case : the ratio of new foreign words to the total vocabulary remains stable (ca. 1/10) until a vocabulary of 3,000 items is reached. Thereafter it increases considerably. The sample described as Part 2 (fig.5) containing ca.50,000 words, has not been pre-edited; i.e. no orthographic mistakes or ommissions have been eliminated, as it was done manually in Part 1. Thus all orthographic idiosyncrasies have been counted as new types by the computer. We assume that the difference in the size of the so-called basic vocabulary (3,000 – 3,500) is mainly due to this fact.

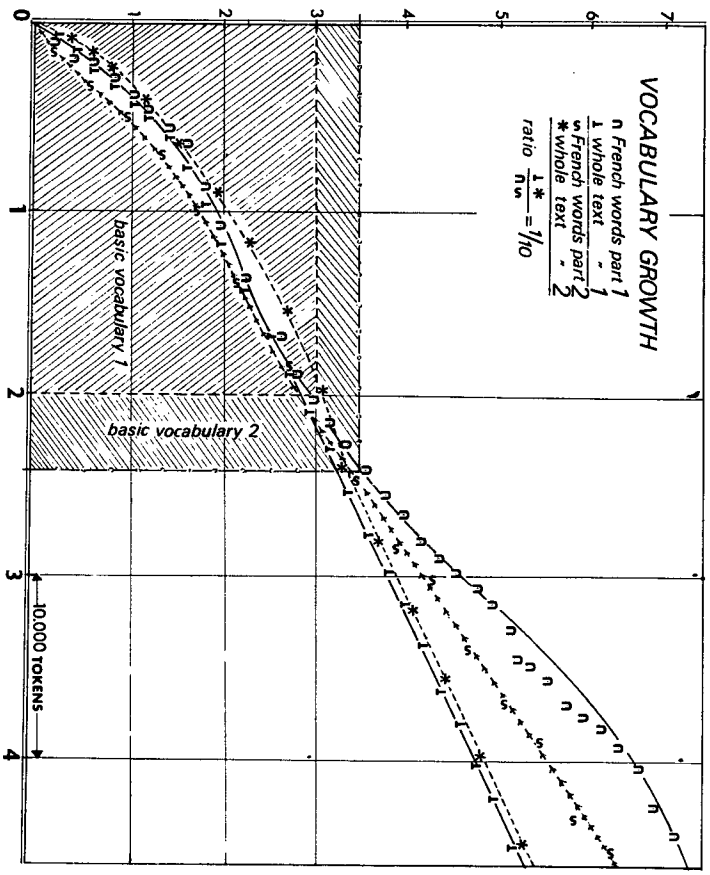
VOCABULARY GROWTH
n French words part 1
1 whole text - 1
s French words part 2
* whole text - 2
ratio 1/n = 1/10
basic vocabulary 1
basic vocabulary 2
10.000 TOKENS

FIG.5

2. As the choice of lexemes from the store takes place in terms of "conceptual symbols", the lexical diversity should not be substantially diminished on account of the limited vocabulary. The blank-fillings with French lexemes should allow the francophones to keep the overall diversity on a normal level, i.e. on that of the SWD-writers. In other words, we suppose that the greater number of foreign elements in FWD-texts is the consequence of the endeavor to "keep in pace" with the normal rate of language diversity.

8. CONCLUSIONS

a) The francophone bilinguals use more than twice as much words as the monolingual native speakers of Dutch.
b) This fact is connected with the tendency to keep the overall variety of vocabulary at a certain "normal" level of speech production. This variety is a bit smaller than in the case of native speakers (cf. the ratio $r=\frac{log\ V}{log\ N}$; for FWD 0.837, for SWD 0.880).
c) It can nevertheless be described as "normal" since the value of parameter B in MANDELBROT's formulation of the ZIPF-law is 1.03347.
d) The foreign lexemes are not equidistributed in the assumed word store; their number increases with the growing text length and this increase is quite evident above the first 3,000 words. This fact allows one to think of them as a "basic vocabulary", covering various subjects (two different multi-subject samples gave nearly identical values of the basic vocabulary).
e) The existence of the basic vocabulary and the good fit of empirical data with the theoretical distribution known as ZIPF-law, strengthens the assumption that the word-units in the store are ordered.
f) One of the ordering principles is the pigeon-holing of information according to some classificational system which takes into account the informational content of words.

REFERENCES

1. The terms "bilingual" and "bilingualism" are understood here in the meaning used by E.HAUGEN, Bilingualism in the Americas, Alabama 1956, p.9 : "Bilinguals (...) is a cover term for people with a number of different language skills, having in common only that they are not monolinguals". Cf. also the same author, The Norwegian language in America, Philadelphia 1953, p.7 :"Bilingualism is understood here to begin at the point where the speaker of one language can produce complete meaningful utterances in the other language".
2. VAN DALE, Groot Woordenboek der Nederlandse Taal, door Dr.C. Kruyskamp, M.Nijhoff, Den Haag 1961-8.
3. Cf.Mandelbrot, Structure formelle des textes et communication, Word, 10 (1954) pp.1-42 and G.Herdan, The Calculus of Linguistic Observation, Mouton & Co, The Hague 1962, pp.59-64.
4. G.Herdan, Type-Token Mathematics, Mouton, The Hague 1960,p.205.
5. L.Brouwers s.j., Het juiste woord. Betekeniswoordenboek der Nederlandse taal, Brepols, Brussel-Turnhout, 1965.
6. P.A.Kolers, Bilingualism and Information Processing, The Scientific American, vol.218, 3, 1968.

Institute of Applied Linguistics
University of Louvain
Vesaliusstraat 2, Louvain(Belgium)