Rajmund Ohly

SELECTED SEMIOTIC ASPECTS OF TEXTS (ON THE EXAMPLE OF SWAHILI)

Originally published as "Niektóre aspekty semiotyczne tekstu (na przykładzie języka suahili)," Studia Semiotyczne 2 (1971), 205–228. Translated by Maja Wolsan

We assume that a text is a certain sequence of signs selected from a predetermined repertoire. This sequence of signs, being a linear combination of the selected elementary signs of a finished set of signs (alphabet), forms a semiotic system. A finished set of those signs is then divided into sequences of word-shaped signs in accordance with some established rules (word formation), which are then combined into sentence-shaped sequences in line with the existing system of rules (syntax). If we attribute meaning to those sequences, which takes place when we interpret them, the sequences of signs form words and sentences of a given language (Klaus 1967: 645).

This article will try to answer two main questions: (a) To what extent does the divergence between the contents of two texts influence the proportionality of signs (sequences) used in these texts?; (b) To what extent does the distribution of signs (sequences) reflect the content of the texts?

As a preliminary study, we have confronted two texts: a philosophical parable by Shaaban Robert entitled *Uzuri* (*Beauty*) from the collection *Kielezo Cha Insha* (*Example of a Sketch*) (Robert 1954) was compared with a detective story by Mohd I Faiz, entitled *Uwanja wa Mwizi Umekwisha* (*The End of the Criminal Game*), published in No. 658 of the periodical "Mambo Leo" (Faiz 1962).

Shabaan Robert is considered the most eminent modern writer of the Swahili language area; his works can serve as a model for the creative use of modern Swahili. The analysed work belongs to the genre of moral-philosophical parable stemming directly from classical Swahili moralising

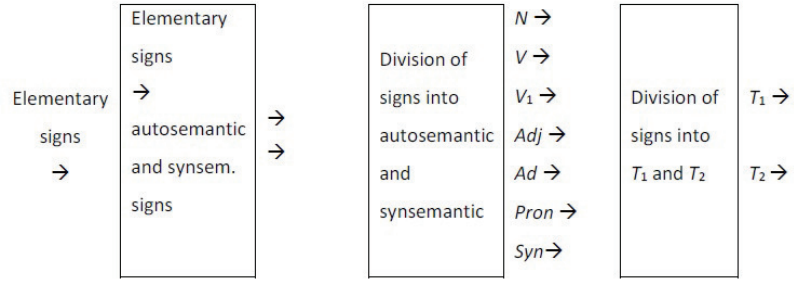
poetry. This genre, while typical of the collection, is quite new in prose as such. The story by Mohd I Faiz, in turn, is one of numerous works inspired directly by the genre of crime-story, massively popular and widespread in Europe. The author's name is not, as it seems, widely known. It should be stressed, however, that "Mambo Leo," which published the story, is one of those periodicals that pay much attention to linguistic correctness.

A. DISTRIBUTION OF SIGNS IN SETS

1. Let us take two texts with an equal number of sentences (32). T_1 (= Shaaban Robert's text) is composed of 1510 elementary signs, T_2 (= Mohd I Faiz's text) of 2102 signs. Compared with the number of sentences in these texts, this gives us on average 47 elementary signs per sentence in T_1 and 65 signs per sentence in T_2. Thus the first difference between the sets stems from the unequal use of elementary signs (T_1 — 100% : T_2 — 129%).[1]

2. Elementary signs were used to create word-shaped sequences, semantic signs, which made up the following subsets: $(N + V + V_1 + Adj + Ad + Pron + Syn^2) \in T_1$ or T_2.

The process can be illustrated as follows:



3. Deviations from the quantitative average of the elementary composition of semantic and synsemantic signs do not exceed 1 elementary sign in the texts. Only the composition of V_1 in T_2 is larger than that in T_1 by 1

[1]The reason for this is that the constitutive sentences in T_1 are simple sentences, while in T_2 they are compound or complex sentences.

[2]These sets can be presented as follows:

$T_1 =$ $N_1, N_2 \ldots N_{118}$; $V_1, V_2 \ldots V_{92}$; $V_{(1)1}, V_{(1)2} \ldots V_{(1)25}$; $Adj_1, Adj_2 \ldots Adj_{25}$; $Ad_0 \ldots Ad_0$; $Pron_1, Pron_2 \ldots Pron_{17}$; $Syn_1, Syn_2 \ldots Syn_{77}$

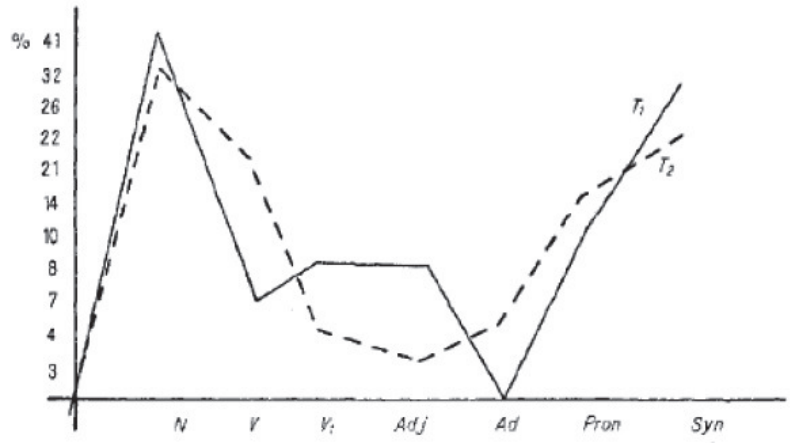
$T_2 =$ $N_1, N_2 \ldots N_{122}$; $V_1, V_2 \ldots V_{83}$; $V_{(1)1}, V_{(1)2} \ldots V_{(1)18}$; $Adj_1, Adj_2 \ldots Adj_{13}$; $Ad_1, Ad_2 \ldots Ad_{18}$; $Pron_1, Pron_2 \ldots Pron_{55}$; $Syn_1, Syn_2 \ldots Syn_{83}$

List of abbreviations: symbolising sings — N = noun, V = verb, V_1 = auxiliary verb, Adj = adjective, Ad = adverb; indicative sings — $Pron$ = pronoun; Syn = synsemantic signs.

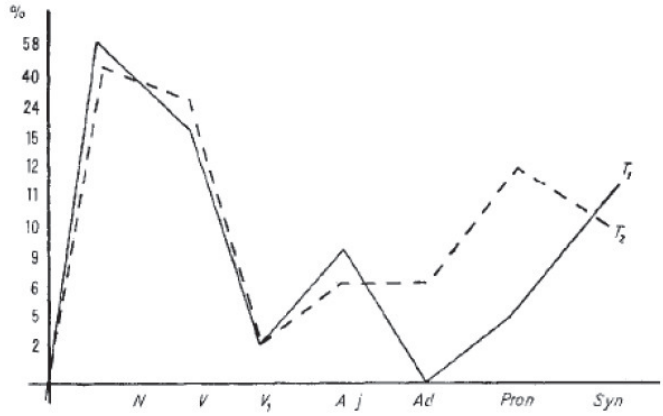
sign, therefore from now on we will leave the issue of elementary signs on the side.

4. In total, T_2 contains 97 semantic and synsemantic signs more than T_1. This quantitative difference is particularly large for verbs (60 units more), pronouns (28 units more) and adverbs (18 units more), but T_1 contains more adjectives than T_2 (12 units more).

The quantitative distribution of the parts of speech (signs) in both texts suggests, however, a tendency to even out the proportions. In fact, the percentage deviation in the distribution of the two texts only once exceeds 10% (T_2 has 14% more V) and once is close to 10% (T_1 has 9% more N); the other deviations do not exceed 5%, as shown on the graph below:



5. The vocabulary of T_2 is richer than that of T_1 by 71 units, which is 35%. The quantitative advantage of T_2 is particularly large in verbs (31 units), pronouns (18 units) and adverbs (12 units). T_1 has only a few more auxiliary verbs and adjectives. The percentage share of signs in both vocabularies shows that the share of N in the vocabulary of T_1accounts for 58% and that of T_2 for 40% of the total number of signs, while for V it is 15% and 24% respectively. Thus, the essential difference between the distributions of signs comes down to an 18% advantage of T_1 over T_2 in terms of the number of N and a 9% advantage of T_2 over T_1 in terms of the number of V. The other deviations do not exceed 7%, as illustrated by the following graph:



6. From the perspective of the frequency of use of each sign from the vocabulary, signs used only once constitute in total 65% of the T_2 vocabulary and 67% of the T_1 vocabulary. As for individual signs, the distribution is as follows:

As regards N: in T_1 75% and in T_2 77% were used once; in T_1 18% and in T_2 13% were used twice; in T_1 3% and in T_2 6% were used three times; and in both texts 4% of the total number of N were used more than three times.

As regards V: in T_1 90% and in T_2 — 59% were used once; in T_1 10% and in T_2 26% were used twice; while 15% were used three times and more, exclusively in T_2.

The frequency in use of V_1 is regular in both texts: 25% of V_1 were used in T_1 once, five, seven and twelve times, while in T_2 33.3% of V_1 were used once, five and twelve times.

As regards *Adj*: T_1 uses 54% and T_2 uses 92% of the vocabulary once,; twice: T_1 30%, T_2 8%; three times and more: 16%, exclusively in T_1.

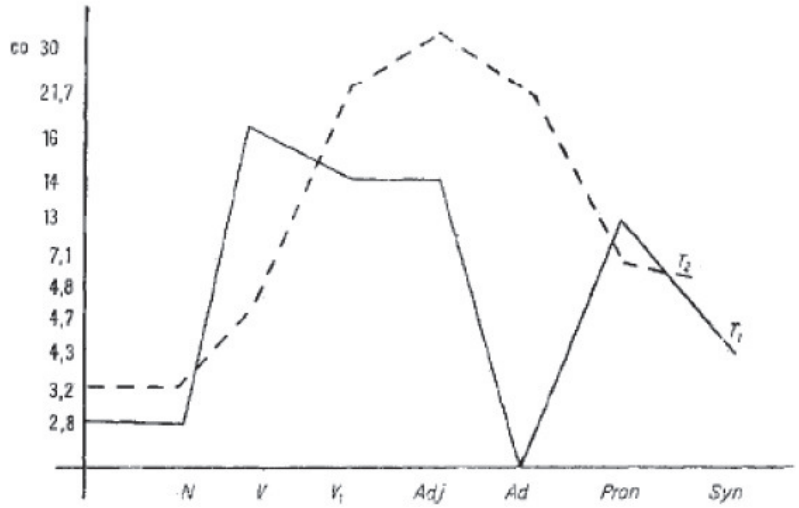
Proper *Ad* practically do not occur in T_1. The frequency of use of *Ad* in T_2 is as follows: once — 67%, twice — 25%, four times — 8%.

As regards *Pron*: 29% in T_1 and 45% in T_2 were used once; 32% were used twice in T_2 (exclusively); 29% in T_1 and 12% in T_2 were used three times; 13% were used fourtimes in T_1 (exclusively); 29% in T_1 and 4% in T_2 were used five times; and 4% were used fifteen times in T_2 (exclusively).

The frequency of *Syn* exhibits the largest distribution in both texts: 38% in T_1 and 35% in T_2 were used once; 23% in T_1 and 30% in T_2 were used twice; 15% were used three times in both T_1 and T_2; 5% were used four times in T_2 (exclusively); 5% were used nine times in T_2 (exclusively); 24% in T_1 and 10% in T_2 were used thirteen times and more.

7. The average frequency of occurrence can be illustrated by the following

amplitude:



B. REGULARITY OF DISTRIBUTION OF SIGNS IN SETS

If we assume that all parts of speech (*N* , *V*, V_1, *Adj*, *Ad, Pron*, and *Syn*) occur regularly in each sentence, then the degree of indeterminateness of their occurrence in each sentence would equal log 32 = 1.5051, and the degree of probability of occurrence of each of them would be $1/32$[3]; the indeterminateness of each result would be log 32/32 = 0.0457. However, the actual indicators of occurrence of the parts of speech show deviations from both the conventional average indeterminateness and between the two compared texts, e.g. in T_1 $N = 2.0719$ and in T_2 $N = 2.0864$, in T_1 $V = 1.3424$ and in T_2 $V = 1.9138$, etc.

A particular property of this disproportion is the size of both texts: T_2 is much larger than T_1, it introduces 97 units more (while keeping the same length of text measured according to the number of sentences). Similar quantitative deviations are observable in individual subsets of signs.

However, while examining the internal volume proportions in the distribution of signs in T_1 and T_2, we can notice some regularities.

1. Among obvious regularities of the texts we should mention the elementary composition of semantic and synsemantic signs. Synsematic signs have the lowest average composition (3.45 phonemes), then in the rising order there are the compositions of pronouns (3.9), auxiliary verbs (4.3), adverbs (4.4), adjectives (5.2), nouns (5.5) and verbs (7.95). Only in one

[3]I.e. $\log_2 32 = 5$.

case do deviations reach 1 phoneme (advantage of T_2), which results from the proportion of the use of full and partial forms of auxiliary verbs.[4]

2. The next regularity is the predominating share of nouns in the distribution of words in the text, in T_1 — 41% and in T_2 — 32% of the text volume; synsemantic morphemes take second place (T_1 — 26%, T_2 — 22%). Consequently, nouns and synsemantic morphemes account for 67% of T_1 and 54% of T_2. Verbs, which are sentence-forming elements, take third place in T_2 (21%), and only last place in T_1 (7%). Even if we add nouns and auxiliary nouns together, their percentage share in T_1 will be 15% and in T_2 — 25%.

3. Despite the quantitative differences in the distributions, only in one case do they reach 14% (advantage of T_2 in terms of V), deviation index 3.742, and in one other case 9% (advantage of T_1 in terms of N), deviation index 0.9487; the others do not exceed 5% (index 0.7071). Hence we can speak of a further regularity: the share of individual parts of speech, regardless of the number of parts of speech used, is more or less constant in a text. Deviations reaching no more than 14% are observed only in nouns and verbs.

4. This regularity also concerns the vocabulary of the texts despite changes in quantitative distribution. Nouns have the largest quantitative share in the vocabularies of both texts (T_1 — 58%, T_2 — 40%), just as in the general volume distributions. Second place is taken by verbs (T_1 — 15%, T_2 — 24%); while synsematic morphemes are in third place in T_1 (11%) and fourth in T_2 (10%) —after pronouns (12%). Thus nouns and verbs constitute 73% of the vocabulary of T_1 and 64% of T_2.

This phenomenon is strictly related to the fourth regularity of the texts: the directions of use of the vocabulary are the same in both of them.

	%	N	V	V_1	Adj	Ad	Pron	Syn
Vocabulary	T_1	58	15	2	9	0	5	11
Volume	T_1	41	7	8	8	0	10	26
Vocabulary	T_2	40	24	2	6	6	12	10
Volume	T_2	32	21	4	3	4	14	22
Volume	T_1	–	–	+	–	0	+	+
Volume	T_2	–	–	+	–	–	+	+

As we can see, the use of a vocabulary for the composition of a text reduces the percentage share of nouns, verbs, adjectives and adverbs in the volume distribution of the text, while the percentage share of auxiliary verbs, pronouns and synsemantic morphemes increases.

[4] V_1 = *kawa* or *ni*.

5. The above rule is justified by the frequency of use of each word. Over 50% share of parts of speech whose vocabulary has been used once decreases in the overall volume of the text, while a lower than 50% share of parts of speech whose vocabulary has been used once increases in the overall volume of the text.

There is also a sixth regularity: the vocabulary of a text used only once in the text accounts on average for 66% of the vocabulary (in T_1 — 68%, in T_2 — 65%) and on average 32% of the overall volume of the text (T_1 — 34%, T_2 — 30%).

6. The next regularity is highlighted by the relation between the volume of the texts (= n) and the vocabulary of the texts (= v): the vocabulary accounts for 45% of the volume of T_1 and for 52% of T_2. Using the formula $v/\sqrt{n}$we receive the value of the index of quantitative diversity of the vocabulary[5] — 24.9 for T_1 and 29.4 for T_2. The deviation is relatively high — 4.5 units. But inclination, calculated according to the formula $a = \log n/\log v$ is 1.12 for T_1 and 1.09 for T_2. The deviation value is 0.03, therefore the values are quite similar. Similarly, the distribution index of the texts,[6] calculated according to the formula $v_1(f/v)$, shows substantial closeness of the values: T_1 — 1.47, T_2 — 1.52.

7. The degree of entropy,[7] according to the formula $H = \sum_{i=1}^{i=k} Pa_i log Pa_i$, is 0.1117 for T_1 and 0.0718 for T_2; thus the degree of indeterminateness in T_1 is larger than in T_2.

C. DISTRIBUTION OF SIGNS AND THE CONTENT OF SETS

The question that comes to mind is whether it is possible to determine, on the basis of data on the distribution of signs and the nature of the content, the semantic structure of the texts.

1. Based on the distribution of semantic and synsemantic signs, we can establish the speed with which information is conveyed and the average information content. The matter of speed will not be important further on in this article. Let us only note, as a curiosity, that if we assume the average composition of elementary signs for each semantic/synsemantic sign (M) in both texts of $5f/M$, i.e. 2.32 bit/M, and that the average time needed to read a text composed of 31 lines (with 65 signs per line) is 180 seconds,[8] then the average speed with which information is conveyed in both texts

[5]Formula according to P. Guiraud (1966: 96f).

[6]v_1 = a set of words used once.

[7]Formula according to Z. Rowieński, A. Ujemow and I. Ujemowa (1963: 74f).

[8]Second = *t*.

is 2.2 M/t, i.e. 1.1 bit/sec; given that the average number of words in T_1 is 302, which gives us 8.3 bits, and in T_2 it is 420, i.e. 8.7 bits, the speed with which information is conveyed is 18.3 bits/t for T_1 and 19.1 bits/t for T_2. These results are close to the standard obtained in other studies (Pierce 1967: 301f).[9] Deviations for both texts are minimal.

2. When examining the average content of information in both texts solely on the basis of the distribution of parts of speech according to the formula[10]

$$(\frac{1}{N} \times log_2 N + \frac{1}{V} \times log_2 V \ldots \frac{1}{Syn} \times log_2 Syn)$$

we get for $T_1 = 0.88$ bit/M and for $T_2 = 1.06$ bit/M, which confirms the observations that the information content in T_2 is bigger than in T_1. Consequently, the distribution of parts of speech in T_2 is better because it conveys the content intended by the author more effectively.

3. In this case, our suppositions on the nature of the content of the two texts can be based on the distribution of sentence-forming elements. If we consider only symbolic names, it turns out that verbs account for 11% and auxiliary verbs for 13% of T_1, while in T_2 it would be 33% and 5% respectively. This shows the significant share of defining presuppositions in T_1, of expressions describing the fact of being something or having a certain quality, which are predominant over assertions of some activity or state of certain objects/subjects (Whiteley 1961: 148, 2n). In T_2, the share of defining presuppositions is negligible in the structure of the entire text and is clearly overshadowed by synthetic sentences. If we additionally take into account the factor of importance, it turns out that 90% of verb units in T_1 and 59% in T_2 occur once, while among adjectives 54% in T_1 and 92% in T_2 were used once. We can conclude that T_1 attaches the same weight to adjectives (high frequency of use) than T_2 does to verbs. We can further suppose that the content of T_1 is semantically static, which means that it has few words describing actions, many words for the being of objects, defining presuppositions, while the content of T_2 is dynamic, i.e. it is important that some things are happening in time and space.

4. An analysis of the lexical and grammatical forms (B)[11] occurring in both texts allows us to detect other significant regularities: first, deviations

[9] Our data is understated, which results from the conventional average (31 lines/180 sec.).

[10] Formula according G. Klaus (1967: 27).

[11] See Annex.

in simple, adjectival and verbal nouns and compounds belonging to the vocabulary do not exceed 13%, and the fact that there is a greater number of N_s in T_2 is countered by the fact that there is a greater number of N_{adj} in T_1 (Table 7); second, in both vocabularies abstracts account for 45% of N and the predominance of general names (Table 8) in T_1 is balanced by the predominance of individual names in T_2 (deviation up to 8%); third, deviations in the distribution of nouns according to noun classes reach 10% only once, and in the other cases fall below 6% (Table 10); fourth, the distribution of nouns, both derivative and simple, is almost identical in both texts (deviation of 2%) (Table 11).

The distinguishing elements cast more light on the specificity of the texts: in terms of nouns, Table 9 shows that the advantage of T_2 in simple nouns results first of all from the preference for individual names, while the advantage of T_1 in adjectival nouns is based mostly on adjectival abstracts. The sequence of nouns used in T_1, in line with the rule N_{abst} — N_{ind} — N_{gen} (13%) deviates from the sequence in T_2 N_{ind} — N_{abst} — N_{gen} (5%) in the lesser (in percentage) vocabulary of T_2 in general names. Hence abstracts and general names constitute 58% of T_1 and 50% of T_2, while individual names constitute 50% of T_2 and 42% of T_1. The assumption that T_2 is more concrete is confirmed by the fact that in T_2 *verba sentiendi et dicendi* and *verba affectus* constitute 27% of V, while in T_1 as much as 50% of V.

At the same time, we should notice borrowings, which come down solely to Arab words in T_1, while T_2 includes Arab, English, Portuguese, and Hindi words. Borrowings constitute 56% of nouns in T_2 (in T_1 — 33%); the degree of borrowings in the noun vocabulary of T_2 deviates far from the generally accepted norm (35%) but is balanced in the overall vocabulary (35.3%, Table 15). The presence of English borrowings in T_2 makes it possible to immediately establish the period of setting of T_2 as after 1905, which is not possible for T_1.

D. DISTRIBUTION AND THE SEMANTIC ASPECT OF SIGNS

According to the theory of information (Shannon), the concept of information is in fact a statistical concept and from the semantic perspective covers only the syntactic aspect of signs or sets of signs. However, sings enter into relations not only with other signs but also with their meanings (the semantic aspect) (Klaus 1967: 721f).

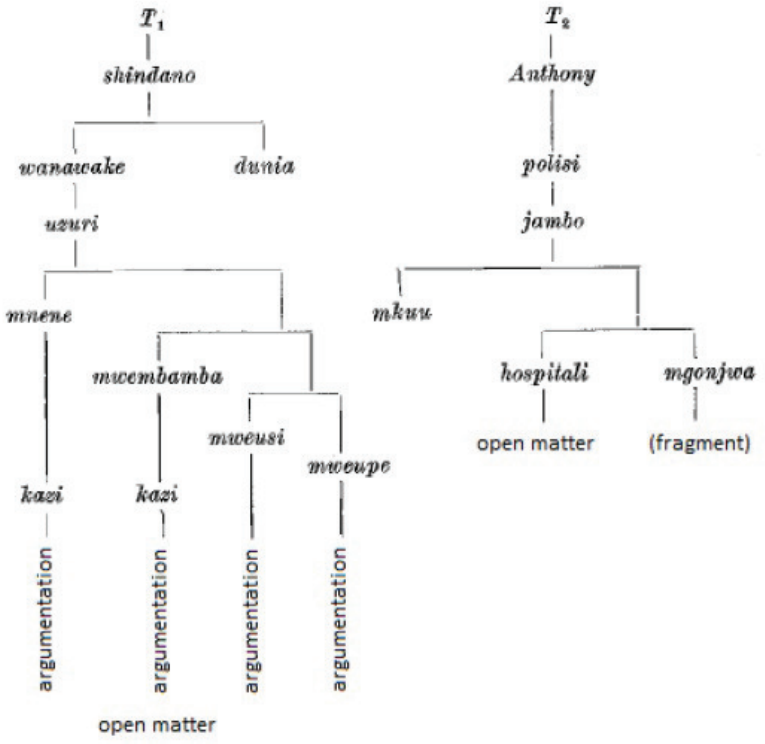
1. The analysis of the vocabularies (in the section A 5) shows a considerable percentage of common vocabulary as regards auxiliary verbs, synsemantic morphemes and pronouns, which is obvious, while for adjectives and

verbs it is 11—12% and for nouns only 4%. This might additionally point to the different contents of the texts. The common nouns include: *kitabu* 'book', *wakati* 'time', *mtu* 'man', *haraka* 'haste', *hewa* 'air', *mkono* 'hand' and *macho* 'eyes'; the verbs include: *toka* 'go out', *weza* 'can/be able to', *fanya* 'make/do', *taka* 'want', *sema* 'speak', *pa* 'give', *angalia* 'watch out', *patikana* 'receive'; adjectives: *gumu* 'heavy', *ema* 'good'; intensifiers: *sana* 'very'. Only some of these words, however, belong to the most frequently used words, therefore they do not determine the content of the texts. According to the rule that the most frequently used words are the most important in a given text, we have the following distribution of nouns and verbs:

T_1			T_2		
mwanamke (*wanawake*)	'woman/women'	17 x	*Anthony*	Anthony	11 x
			jambo	'case'	5 x
uzuri	'beauty'	7 x	*mkuu*	'superior'	5 x
dunia	'world'	4 x	*wakati*	'time'	3 x
shindano	'competition'	3 x	*polisi*	'police'	3 x
kazi	'work'	3 x	*sigara*	'cigarette'	3 x
weza	'be able to'	2 x	*hospitali*	'hospital'	3 x
taka	'want'	2 x	*mgonjwa*	'patient'	3 x
			sema	'speak'	7 x
			lala	'sleep'	3 x
			fika	'arrive'	3 x
			jibu	'ask'	3 x
			wasili	'arrive'	3 x
			ita	'summon'	3 x
			toka	'go out'	3 x
			ingia	'enter'	3 x

These words confirm the rule mentioned above as they are a clear reflection of the main motif of the content of each of the texts: in T_1 the main characters are women who believe that their beauty is the greatest in the world; therefore they compete to prove their superiority, for example, by proving that they can do any work better than others if they want to.

In T_2 the main character is Anthony, a police detective who is suddenly summoned by his superior in a criminal case, from where, after a short briefing, he goes to a hospital in order to question an injured patient. The cigarette is the usual attribute of hectic actions (the first part of the text). The following structures illustrate the composition of the motifs in the text:



In T_1, presenting the key motif requires the introduction of additional elements ('chubby', 'slim', 'black', 'white'), which are not among the most frequently used words, while in T_2 these words are sufficient to develop the motif. This is another proof of better distribution of T_2.

2. Syntactical data confirm the earlier observations. The structures of compound and complex sentences (Table 19) clearly show the dynamic of T_2: it uses twice as many syntactic forms introducing various logical connections and enriching the line of thought by providing the effects, causes, goal, relativity, contrast, and result of an action, as well as its duration; while in T_1 we can observe the unity of time and place, taking into account the cause, conditionality and relativity of action. T_2 has the greatest advantage in relative, purpose, contrasting and time clauses. The latter have additional support in structures with compound tenses, which do not occur in T_1.

Another regularity of both texts can be observed in the use of verb forms — a preference for the basic grammatical form of *cl.T.V* verbs, which is used in around 50% of the structures in the texts.

By analysing the syntax we can determine the syntactic index,[12] i.e. the nominal to adjectival attributes ratio. T_1 has 25 adjectival structures

[12]Formula according to Guiraud (1966: 96).

(adjectives in primary function), T_2 has 13, while nouns in the attributive function appear 28 times in T_1 and 20 times in T_2 (the frequency of use of a synsemantic morpheme — a). Therefore the syntactic index for T_1 is 0.89 and for T_2 — 0.65. These indexes are relatively high, which results from the limited basis of proper adjectives in Swahili and the resulting need to use nouns to express attributes. However, the fact that T_1 has a higher index despite the significant predominance in the use of adjectives additionally confirms the tendency of T_1 to emphasise the characteristics of the subjects rather than action.

The distribution of nouns according to their syntactic function shows another regularity of Swahili texts: nouns in their primary function account on average for 21% of text and in their secondary function on average for 64% (deviations around 2—3%). The further distribution of nouns according to the function of the base, the possessive attribute, the object attribute, and the prepositional phrase illustrates another regularity:

T	Base	Possessive	Object attribute and prepositional phrase
T_1	22	28	68
T_2	27	20	75

Ratio (%)

T_1	19	23	58
T_2	22	16	62

The deviations do not exceed 7%.

The deviations do not exceed 7%.

*

Statistics are critical for studies based on small samples. There is no doubt that this is indeed justified considering that the error rate decreases proportionally to the increase of the quantitative volume of the sample and that the increase not only makes the information more accurate but also enriches it. The error rate in the above analysis, namely deviations from the norm, is certainly considerable. The thing is that it is hard to verify it because there are no pre-existing statistical linguistic norms for Swahili. The only available data is the percentage of borrowings in the vocabulary of the language, which is 35%. The results of our analysis are consistent with this norm. Moreover, if we compare the data stemming from our analysis

with general linguistic data calculated by Yule (Guiraud 1966: 97), it turns out that there are actual similarities in proportions. It is assumed that the vocabulary of a given text is proportional to the square root of its length; the mean value is $v/\sqrt{n}$= 22. Yule gives the following example: n= 2000, v= 940, a = $\frac{logn}{logv}$ = 1.11, $v/\sqrt{n}$= 21. Thus $v$equals 47% of n; for T_1 this ratio is 45% and for T_2 — 52%; hence in our texts the quantitative richness index is 24.9 and 29.4 and a (inclination) = 1.12 and 1.09, respectively. In another work, Guiraud (1954: 37) presents the model that nouns from a 1000-unit vocabulary constitute 62% of the overall number of nouns of any text, while nouns used once account for 69% of the nouns in the vocabulary. For T_1 these indexes are 67% and 74% respectively. Examples of this kind are many. It should be concluded, as it seems, that despite the relatively small samples, the data obtained from our analyses will be close to the data obtained in the future for much longer texts. At the same time, the relatively minor deviations between the two texts and the observable regularities seem to exclude any randomness of the data. For now, the regularity indexes for Swahili texts proposed in our materials can serve as approximate reference data.

The main aim of this work was to compare two texts with different content and to detect possible differences, which could be used to identify, based on distribution data, those differences in the distributions that result from the respective contents. The distribution data was illustrated by tables and analyses. We should only add, using the criteria of linguistic universalism great caution, that the distributions provided by Guiraud for French abstract prose (*prose abstraite*) and fiction/concrete prose (*prose concrète*) (Guiraud 1954: 39) allow us to notice similar general phenomena in both languages: both in French abstract prose of the 20^{th} century and in Shaaban Robert's philosophical parable the percentage share of nouns and adjectives is much larger and that of verbs and adverbs is much smaller than in French fiction and in the detective story by Mohd I Faiz. This seems to be the rule for the distribution of parts of speech in these literary genres.

ANNEX

A. DISTRIBUTION OF PARTS OF SPEECH

1. Shaaban Robert's work (= T_1) is composed of 32 sentences, constituting a closed whole. From the story by Mohd I Faiz (= T_2) we have selected 32 sentences accordingly, which constitute around 1/3 of the whole text, starting from the first sentence to the thirty-second.

2. Sentences were ordered according to the following rule: simple sentences: (a) short (Subject + Verb) (b) long (with additional other elements); non-simple sentences: (a) complex, (b) compound, (c) compound-complex.

Table 1

T	Simple sentence		Overall	Non-simple sentence			Overall	Total
	short	long		complex	compound	compound-complex		
T_1	6	14	20	6	5	1	12	32
T_2	2	8	10	5	2	15	22	32

Divergence I

T_1	4	6	10	1	3	0	0	0
T_2	0	0	0	0	0	14	10	0

Ratio (%)

T_1	19	43	62	19	16	3	38	100
T_2	6	25	21	16	6	47	69	100

Divergence II

T_1	13	18	31	3	10	0	0	0
T_2	0	0	0	0	0	44	31	0

3. T_1 is composed of 1510 phonemes, T_2 of 2012 phonemes. On average, 1 sentence in T_1 equals 47 phonemes, in T_2 — 65 phonemes.

Divergence III

T_1 — 100% of phonemes: T_2 — 129% of phonemes.

4. We have obtained the following distributions, taking the following as parts of speech: symbolic names: nouns (= N), verbs (=V), auxiliary verbs[13] (= V_1) adjectives (Adj), adverbs (= Ad); indicative names: pronouns (= $Pron$); and synsemantic morphemes (= Syn):

[13] V_1 are the forms of 'to be' — *kuwa*, *ni*, as well as *mo*, *na*.

Table 2

T	*N*	*V*	V_1	*Adj*	*Ad*	*Pron*	*Syn*	Total
T_1	118	22	25	25	–	27	77	294
T_2	122	82	18	13	18	55	83	391

Divergence IV

T_1	0	0	7	12	0	0	0	0
T_2	4	60	0	0	18	28	6	97

Ratio (%)

T_1	41	7	8	8	–	10	26	100
T_2	32	21	4	3	4	14	22	100

Divergence V

T_1	9	0	4	5	0	0	4	0
T_2	0	14	0	0	4	4	0	0

5. For the construction of the parts of speech we have used the following numbers of phonemes.

Table 3

T	*N*	*V*	V_1	*Adj*	*Ad*	*Pron*	*Syn*	Total without *Syn*	Total without *Syn* and *Pron*	Total
T_1	5.8	7.9	3.8	5.5	–	4.1	3.5	5.4	5.7	5.1
T_2	5.3	8.0	4.8	4.9	4.4	3.7	3.4	5.2	5.4	4.9
av.	5.5	7.9	4.3	5.2	4.4	3.9	3.45	5.3	5.55	5.0

Divergence VI

T_1	0.5	0	0	0.6	0	0.4	0.1	0.2	0.3	0.2
T_2	0	0.1	1.0	0	–	0	0	0	0	0

6. In T_1 there is a vocabulary of 122 words and 13 synsemantic morphemes, in T_2 there are 186 words and 20 synsemantic morphemes. The vocabulary has the following distribution:

Table 4

T	*N*	*V*	V_1	*Adj*	*Ad*	*Pron*	*Syn*	Total	Without *Syn*	Without *Syn* and *Pron*
T_1	78	20	4	13	–	7	13	135	122	115
T_2	83	51	3	12	12	25	20	206	186	161

Divergence VII

T_1	0	0	1	1	0	0	0	0	0	0
T_2	5	31	0	0	12	18	7	71	64	46

Ratio (%)

T_1	58	15	2	9	–	5	11	100
T_2	40	24	2	6	6	12	10	100

Divergence VIII

T_1	18	0	0	3	0	0	1	0
T_2	0	9	0	0	6	7	0	0

7. The frequency of use of the vocabulary is as follows:

Table 5

	N		*V*		V_1		*Adj*		*Ad*		*Pron*		*Syn*	
x	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
1	58	64	18	30	1	1	7	11	–	8	2	12	5	7
2	14	11	2	13	–	–	4	1	–	3	–	8	3	6
3	2	5	–	7	–	–	1	–	–	–	1	3	2	3
4	1	–	–	–	–	–	1	–	–	1	1	1	–	1
5	–	2	–	–	1	1					–	1	–	–
6	–	–	–	–	–	–					–	–	–	–
7	1	–	–	1	1	–					–	–	–	–
8	–	–			–	–					–	–	–	–
9	–	–			–	–					2	–	–	1
10	–	–			–	–					–	–	–	–
11	–	1			–	–					–	–	–	–
12	–	–			1	1					–	–	–	–
13	–	–									–	–	1	–
14	–	–									–	–	1	–
15	–	–									–	1	–	–
17	1	–											–	–
20													–	1
21													–	1
28													1	–

Divergence IX

Divergence IX

	N		*V*		V_1		*Adj*		*Ad*		*Pron*		*Syn*	
x	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
1	0	6	0	12	0	0	0	4	0	8	0	10	0	2
2	3	0	0	11	–	–	3	0	0	3	0	8	0	3
3	0	3	0	7	–	–	1	0	–	–	0	2	0	2
4	1	0	–	–	–	–	1	0	0	1	0	0	0	2
5	0	2	–	–	0	0					0	1	–	–
6	–	–	–	–	–	–					–	–	–	–
7	1	0	0	1	1	0					–	–	–	–
8	–	–			–	–					–	–	–	–
9	–	–			–	–					2	0	0	1
10	–	–			–	–					–	–	–	–
11	0	1			–	–					–	–	–	–
12	–	–			0	0					–	–	–	–
13	–	–									–	–	1	0
14	–	–									–	–	1	0
15	–	–									0	1	–	–
17	1	0											–	–
20													0	1
21													0	1
28													1	0

8. The average frequency of occurrence[14] of parts of speech:

Table 6

T	*N*	*V*	V_1	*Adj*	*Ad*	*Pron*	*Syn*
T_1	2.8	16	14	14	0	13	4.3
T_2	3.2	4.7	21.7	30	21.7	7.1	4.8

Divergence X

T_1	0	11.3	0	0	0	5.2	0
T_2	0.4	0	7.7	16	21.7	0	0.5

B. DISTRIBUTION OF LEXICAL AND GRAMMATICAL FORMS

1. In T_1 there are 49 simple nouns, including 24 Arab borrowings, and 28 derivative nouns, including 12 adjectival nouns (= N_{adj}) and 16 verbal

[14] I.e. average spaces between the occurrences of the parts of speech in the text.

nouns (= N_v); apart from that, there is 1 compound (= N_c). There is 1 Arab borrowing among the N_{adj} and 1 among N_v.

In T_2 there are 63 simple nouns, including 44 borrowings: 31 from Arab, 9 from English, 2 from Portuguese, and 2 from Hindi, as well as 18 derivative nouns: 4 adjectival and 14 verbal. There is also 1 compound. Among N_v there are 2 Arab borrowings; the compound is composed of two borrowings (English and Arab).

Table 7

T	N_s	N_{adj}	N_v	N_c	Total borrowings
T_1	49	12	16	1	26
T_2	63	4	15	1	47

Divergence XI

T_1	0	8	1	0	0
T_2	14	0	0	0	21

Ratio (%)

T_1	63	16	20	1	33
T_2	76	5	18	1	56

Divergence XII

T_1	0	11	2	0	0
T_2	13	0	0	0	23

2. In T_1, N are composed of 18 abstract names (including 11 borrowings), 10 general names and 21 individual names; N_{adj} of 8 abstracts and 4 systemic names; N_v of 13 abstracts, including 3 *nomina actionis*, 1 *nomen patientis* and 3 *nomina agentis*.

In T_2, among N there are 25 abstract names (including 22 borrowings), 4 general names and 34 individual names; N_{adj} include 1 abstract and 3 systemic names; N_v are composed of 12 abstracts, of which 6 are *nomina actionis* and 2 are *nomina agentis*.

Table 8

T	abst. names	gen. names	ind. names
T_1	36	10	33
T_2	38	4	41

Divergence XIII

T_1	0	6	0
T_2	2	0	8

Ratio (%)

T_1	45	13	42
T_2	45	5	50

Divergence XIV

T_1	0	8	0
T_2	0	0	8

Table 9

	N_s				N_{adj}		N_v			
T	abst.	gen.	ind.	borr.	abst.	sys.	abst.	*act.*	*pat.*	*ag.*
T_1	18	10	21	11	8	4	7	13	1	3
T_2	25	4	34	22	1	3	6	12	–	2

Divergence XV

T_1	0	6	0	0	7	1	1	1	1	1
T_2	7	0	13	11	0	0	0	0	0	0

3. Distribution of N based on noun classes:

Table 10

	classes							
T	1	2	3	4	5	6	7	8
T_1	10	4	28	3	14	13	–	6
T_2	7	9	33	6	16	6	–	6

Divergence XVI

T_1	8	0	0	0	0	7	0	0
T_2	0	5	5	3	2	0	0	0

Ratio (%)

T_1	12	5	36	4	18	17	–	8
T_2	8	11	40	7.3	19	7.3	–	7.3

Divergence XVI

T_1	4	0	0	0	0	9.7	–	0.7
T_2	0	6	4	3.3	1	0	–	0

4. Among the 20 V in T_1,there are 7 derivative and 13 simple verbs, while among the 51 V in T_2,there are 19 derivative and 32 simple verbs.

Table 11

T	V_s	V_{der}	% of V_{der}
T_1	13	7	35
T_2	32	19	37

Divergence XVIII

T_1	0	0	0
T_2	19	12	2

Among the 20 V in T_1, there is 1 borrowing (Arab), while among the 51 V in T_2, there are 9 borrowings, which gives us the relation of 5% : 17%.

The distinguishing feature of V in T_1 are *verba sentiendi et dicendi* and *verba affectus*, i.e. 10 in 20 V; in T_2 there are 14 verbs of the same type in 51 V, which gives us a relation of 50% : 27%.

5. Among the 13 adjectives in T_1, there are 3 borrowings, i.e. 23%, while in T_2 among the 12 adjectives, there are 4 borrowings, that is 33%.

Six among the 13 *Adj* in T_1 express mental properties, 5 express physical properties (including 4 for volume or size, 2 for colours), 1 expresses quantity and 1 — intensification. Seven among the 12 *Adj* in T_2 express mental properties, 2 — physical properties (volume or size), 2 — quantity and 1 — intensification of a property.

Table 12

T	mental	physical	quant.	int.	borrowings	%
T_1	6	5	1	1	3	23
T_2	7	2	2	1	4	33

Divergence XIX

T_1	0	3	0	0	0	0
T_2	1	0	1	0	1	10

Percentage rate

T_1	46	38	8	8	23
T_2	58	17	17	8	33

Divergence XX

T_1	0	21	0	0	0
T_2	12	0	9	0	10

6. Among the 12 *Ad* in T_2, there are 5 adverbs of manner, 3 of place and 4 of time. In T_1 there are no adverbs.

Table 13

T	manner	place	time
T_1	–	–	–
T_2	5	3	4

7. Among the 6 *Pron* in T_1, there are 3 possessive, 1 indicative, 1 interrogative and 1 independent. Among the 25 *Pron* in T_2, there are 5 possessive, 12 indicative, 1 interrogative and 2 independent, as well as 5 dependent pronouns separate from a verb complex.

Table 14

T	poss.	ind.	interr.	indepe.	dep.
T_1	3	1	1	1	–
T_2	5	12	1	2	5

Divergence XXI

T_1	0	0	0	0	0
T_2	2	11	0	1	5

8. The proportions of synsemantic morphemes will be discussed in the section on syntax.

9. The common lexical composition of both texts is as follows:

Among the 161 *N* in both texts, the common vocabulary is 4%, among the 71 *V* — 11%, among the 7 V_1 — 85%, among the 25 *Adj* — 12%, among the 32 *Pron* — 15%, and among the 33 *Syn* — 21%. There are no common *Ad*.

The common elements constitute the following percentage of the two texts:

Table 15

%	*N*	*V*	V_1	*Adj*	*Ad*	*Pron*	*Syn*
T_1	8	40	75	23	–	70	53
T_2	8	15	100	25	–	20	35

Divergence XXII

T_1	0	25	0	0	0	50	18
T_2	0	0	25	2	0	0	0

10. Distribution of borrowings in the two texts according to parts of speech:

Table 16

%	*N*	*V*	*Adj*	Total
T_1	33	5	23	20.3
T_2	56	17	33	35.3

Divergence XXIII

T_1	0	0	0	0
T_2	23	12	10	15

C. DISTRIBUTION OF SYNTACTIC FORMS

1. The distribution of nouns used either in the form of a base, i.e. in the primary function (= 1) or as attributes, i.e. in secondary functions of the first (=2), third (=3) or fourth (=4) grade is as follows:

Table 17

N	1	2	3	4	Total
T_1	22	74	21	1	118
T_2	27	79	13	3	122

Divergence XXIV

T_1	0	0	8	0	0
T_2	5	5	0	2	4

Ratio (%)

N	1	2	3	4
T_1	19	63	17	1
T_2	22	65	11	1

Divergence XXV

T_1	0	0	6	0
T_2	3	2	0	0

2. The grammatical forms of verbs used in T_1 and T_2 are as follows: 0 = *V*, 1 = *cl. T. V*, 3 = *cl. T. rel. V*, 4 = *cl. T. ob. V*, 5 = *cl. T. rel. ob. V*,

6 = *cl. T. ref. V*, 7 = *neg. cl. T. V*, 8 = *cl. V R. pass. rel.*, 9 = *cl. V rel.*, 10 = *cl. T. V R. pass.*, 11 = *cl. T. rel. V R. pass*, 12 = *cl. T. ob. V R. pass.*[15]

The distribution of these structures in the texts is as follows:

Table 18

V	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
T_1	1	3	10	–	1	–	1	3	1	1	1	–	–	22
T_2	–	15	44	3	9	–		1	–	–	8	1	1	82

Divergence XXVI

T_1	1	0	0	0	0	0	1	2	1	1	0	0	0	0
T_2	0	12	34	3	8	0	0	0	0	0	7	1	1	60

Ratio (%)

T_1	4.5	14	46	–	4.5	–	4.5	14	4.5	4.5	4.5	–	–	
T_2	–	19	54	4	10	–	–	1.3	–	–	9	1.3	1.3	

Divergence XXVII

T_1	4.5	0	0	0	0	0	4.5	12.7	4.5	4.5	0	0	0	
T_2	0	5	8	4	5.5	0	0	0	0	0	4.5	1.3	1.3	

3. Structures with compound tenses (auxiliary verb + main verb) were used 5 times in T_2. These forms do not appear at all in T_1.

4. All adjectives in T_1 and T_2 are used in the primary function.

5. After reducing compound-complex sentences to compound and complex sentences we can identify 14 sentences of this kind in T_1 and 53 in T_2. T_1 uses 5 types of sentences, and T_2 uses 10.

[15] *V* = stem, *cl* = pre-verbal pronoun (verbal clitic), *T* = indicator of time and aspect, *rel.* = relative pronoun, *ob.* = object pronoun, *ref.* = reflexive marker, *neg* = negation marker, *VR* = root, pass. = passive marker.

Table 19

Complex and compound sentences

	compound			complex							
T	conjunctive	disjunctive	resultative	time	object	comparative	effect	causative	conditional	purpose	relative
T_1	6	–	–	–	–	1	–	3	2	–	2
T_2	22	4	1	6	3	1	2	2	–	6	6

Divergence XXVIII

T_1	0	0	0	0	0	0	0	1	2	0	0
T_2	16	4	1	6	3	0	2	0	0	6	4

Ratio (%)

T_1	43	–	–	–	–	7	–	22	14	–	14
T_2	42	7	2	11	6	2	4	4	–	11	11

Divergence XXIX

T_1	1	0	0	0	0	5	0	18	14	0	3
T_2	0	7	2	11	6	0	4	0	0	11	0

Bibliography

1. Robert, Shaaban (1954) ”Uzuri.” In *Kielezo Cha Insha*. Johannesburg: Witwatersrand U.P.

2. Faiz, Mohd I. (1962) ”Uwanja wa Mwizi Umekwisha.” *Mambo Leo* No. 658, Dar es Salaam.

3. Klaus, Georg (1967) *Wörterbuch der Kybernetik*. Berlin: Dietz.

4. Guiraud, Pierre (1954) *Les Caractères statistiques du vocabulaire*. Paris: PUF.

5. Guiraud, Pierre (1966) *Zagadnienia i metody statystyki językoznawczej*. Warszawa: PWN.

6. Rovenski, Zinovei Ilich, Uemov, Avenir Ivanovich and Ekaterina Andreevna Uemova (1963) *Filozoficzny zarys cybernetyki*. Warszawa: Książka i Wiedza.

7. Pierce, John R. (1967) *Symbole, sygnały i szumy*. Warszawa: PWN.

8. Whiteley, Wilfried H. (1961) "Further Problems in the Study of Swahili Sentences." *Lingua* 10[2]: 148—173.