Library and Information Science No. 9 1971

Identification of Languages with Short Sample Texts
—— A Linguometric Study ——

短い試料における言語の判定法
—— 計量的言語研究の一方法 ——

Yukio Nakamura
中　村　幸　雄

要　　旨

文字で書いた自然語試料の言語を，一定の論理式によって判定する方法を見出す目的で，文字および単語の出現について定量的な研究を行なった。第I部では，まず意味を持った自然語文章中における，文字，文字の組合わせ，単語（これらを特徴と総称する）のふるまいを定量的に論ずるに必要な，出現率，出現間隔等の量を定義し，実例につきそれらの性質を明らかにしたうえ，複数の特徴の組合わせについて出現率および出現間隔を求める方法を導いた。

第II部では，実際に言語を判定するため，対象をラテン文字を使う25言語（26表記法）に関して判定条件を求める方法と判定式を示した。この言語の中には同一語族に属し表記法も極めて近いために，判定が困難な場合の対策を含めてある。判定条件には肯定的条件と否定的条件とを併用することが必要であり，また合計261種類の特徴を利用した。判定は手操作の場合パターンシートによる一致法を使うのが便利であるが，パターンのもととなる論理式はプログラム化も容易に行なえる。

なお付録として確定的ではない判定要素を多数個使う統計的判別法の利用も可能であることを示した。

PART I

1.1 Introduction

In many cases of information processing of documents, the printed texts in a natural language is the object for processing. The author intends to treat the written text in natural language in statistical way, that is to say, to consider the behavior of the occurrence of a character or a word as a quantity governed by some statistical rule.

The text in consideration is one with some meaning or other, thus a string of characters or words arbitrarily chosen is not included. One may insist that he can make a sentence or phrase with a certain meaning with or without a specified character or word so that the behavior of the text cannot be regarded as of statistical nature. This may be true only if we restrict ourselves to some short sentences or phrases. A short sentence can be constructed, for instance, without using the character "e" in English. One can not, however, make a text of 1,000 words without using the character "e."

中村幸雄：技術士．日本通信協力株式会社常務取締役．
Yukio Nakamura, Authorized Consulting Engineer (Information), General Manager, Nippon Telecommunications Consulting Co., Ltd., Tokyo, Japan.

The word "natural language" is contrasted to "machine language" or "meta language," and such language as Esperanto which is certainly created by a man, and some national languages which were created by a group of men to unify local languages or to adapt an existing languge to some special situation, are included in the category of natural language, as far as they are spoken by a good number of people.

In the following, a character or any combination of consecutive characters are expressed, where it is necessary, by parentheses, (t), or (th), and a word by brackets, as [and]. For instance, [i] is an independent word (for instance, in Swedish or in Slavonic languages) whereas (i) stands for any character "i" in any word, for instance, those in [in], [si], or [sit]. Endings are sometimes shown as (-ed), (-ing) and (-ty).

Definition of a word is different according to languages concerned, e.g., the word [l'express] is considered as two words in French but the word [air-to-ground] in English, as a single word. For the sake of simplicity, the author defines a word as the "string of characters put between two consecutive blank space." Thus l'express, rendez-vous, air-to-ground are all single words.

1.2 Character Interval

In a written text, two identical characters X's will appear interposing a certain number of words. This number n is interpreted as the "interval" of two X characters in a string of words. The interval can also be expressed by the number of interposing characters. To be clear about the distinction between two kinds of expression in interval, the former is designated as "character interval in words (CIW)," whereas the latter as "character interval in characters (CIC)."

Notation M_gX is used for the CIW of the character X and M_zX is for the CIC of the character X. In some cases, the same character appears twice or more in a word (e.g., the character (a) and (r) in the word [character]), the interval is expressed as $M_g=0^*$ (Fig. 1).

We will examine, at first, the case of low appearance rate of characters. Fig. 2 is the result of measurement for the character (å) in Swedish and for (ß) in German. The measurement is made for 200 interval cases of both characters, each taken from the texts of different nature (newspaper, literature, technical journal, etc.).

Plotting the frequency of appearance on logarithmic scale against M_g on linear scale, we can recognize clearly an exponetial decrease except for the area where the number of sample is not sufficiently large, for (å), the area beynd $M_g=40$, and for (ß), the area beyond $M_g=120$.

We can estimate, from this distribution, the ultimate distribution curve $F(M_g)$ for a very large number of sample, and then by choosing a constant a in such way as to satisfy

$$\int_g^\infty aF(M_g)dM_g=1, \tag{1}$$

we get the normalized distribution

$$f(M_g)=a\,F(M_g). \tag{2}$$

In the following, the word distribution is used for the normalized distribution unless otherwise stated.

The probability that a specific character X will appear at least once in a text of consecutive N words is the sum of the probabilities that the character interval in words (CIW), M_g, of the character X is equal to 0, 1, 2,... N-1. Thus the probability is equal to

$$\int_0^{N-1} f(M_g)dM_g \tag{3}$$

The probability that the character X does

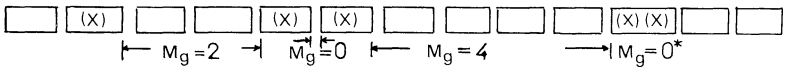

Fig. 1. Example M_g.

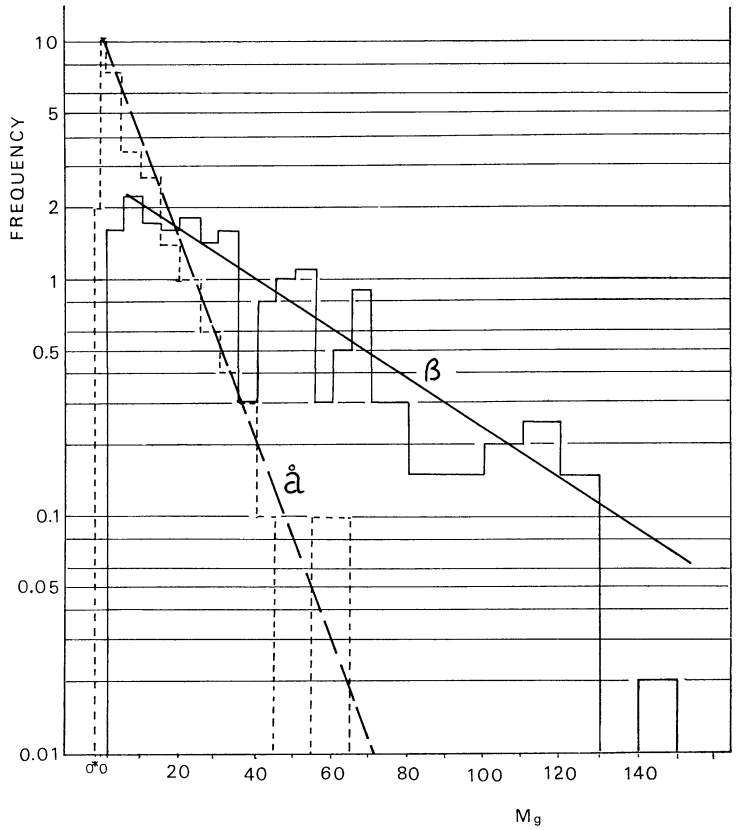

Fig. 2. Distribution of M_g for å (Swedish) and ß (German)

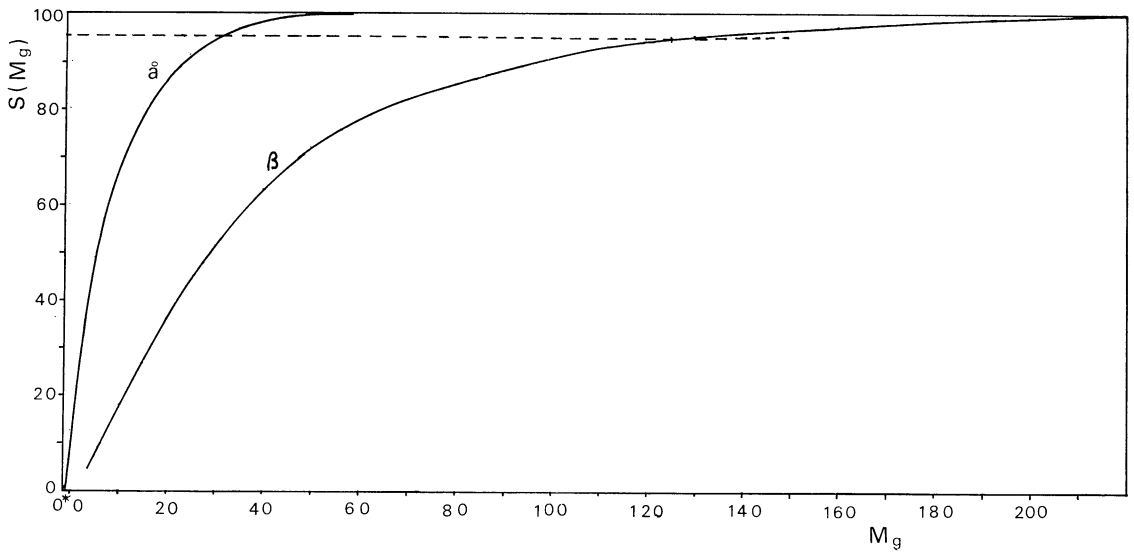

Fig. 3. Accumulated Distribution of M_g.

never appear in the same length of word string, is equal to

$$1-\int_0^{N-1} f(M_g)dM_g=\int_0^{\infty} f(M_g)dM_g-\int_0^{N-1} f(M_g)dM_g$$

$$=1-\int_0^{N-1} f(M_g)dM_g \qquad (4)$$

To obtain a minimum length of word string where at least one character X is included, say, at a significance level of 5 percent, the value $N-1$ which satisfy the relation

$$\int_{N-1}^{\infty} f(M_g)dM_g=0.05 \qquad (5)$$

or

$$\int^{N-1} f(M_g)dM_g=0.95 \qquad (5)$$

must be found.

For this purpose one can use profitably the Fig. 3 which shows accumulated distribution $S(M_g)$ on logarithmic scale rather than to use distribution curve shown in Fig. 2.

Looking at Fig. 3, we obtain the value $N-1=31$ or $N=32$ words for (å) and $N=125$ for (ß). Mean value of Mg for (å) and (ß) are 10.2 and 48, respectively.

It is easy to see that the same principle can be applied to characters instead of words. Thus we can define the "character interval in characters (CIC)," M_z, in the same way.

1.3 The Appearance Rate of Characters and Words

1.3.1 *The Appearance Rate and CIW*

In a string of consecutive n words, if a specific character X appears p times, the ratio p/n can be defined as the "character appearance rate in words (CARW)," r_g, and the value is shown in percent.

Similarly, in a string of consecutive m characters, if a specified character X appears p times, the ratio p/m is defined as the "character appearance rate in chracters (CARC)," r_z, and the value is also expressed in percent.

We can see the relation between these two appearance rates, r_g and r_z, as

$$r_g=p/n=\bar{w}\cdot p/\bar{w}\cdot n=\bar{w}\cdot p/m=\bar{w}\cdot r_z$$

where $\bar{w}$ is the mean length of word measured in the number of characters. This gives a very simple relation

$$r_g=\bar{w}\cdot r_z\,. \qquad (6)$$

The appearance rates r_g and r_z are also statistical quantities. They may deviate from one

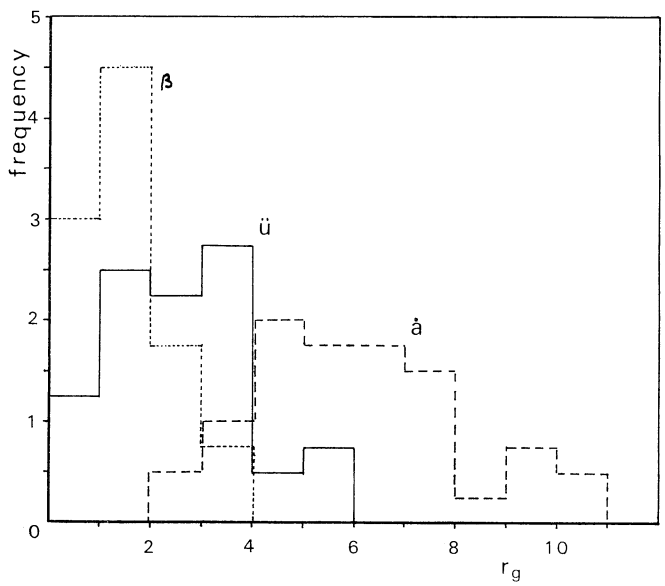

Fig. 4. Distribution of r_g

text to another. Only when the size of sample is large enough, the value of r_g and r_z is reduced to a definite value.

The curve in Fig. 4 shows the distribution of appearance rates. The curve å (Swedish), ü (German) and ß (German) are got from 40 texts in German and Swedish each containing consecutive 50 words.

The mean value of appearance rate does not change with the sample length n or m but the dispersion of appearance rate depends upon n or m. If we represent this dispersion by standard deviation σ, and if the distribution is normal, σ should be proportional to $n^{1/2}$ or $m^{1/2}$.

To ascertain this σ to n (or m) relation, samples of German and English sentences are chosen, at random, changing the value of n. Observations are made for a lot of ten samples for each of three different n values. The result (Fig. 5) shows, for every one of three characters, that

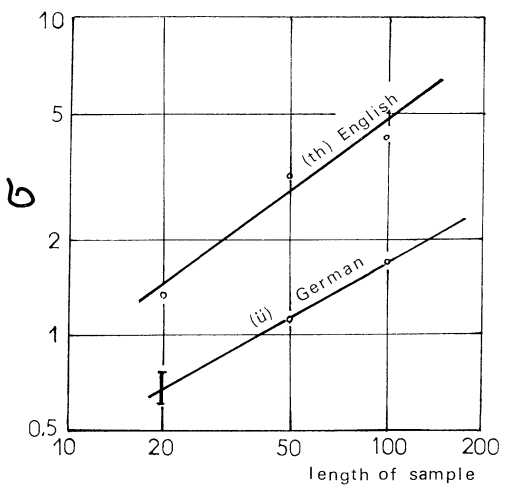

Fig. 5. Dispersion of r_g

$$\sigma \propto n^p \qquad (7)$$

and the value of p is nearly equal to 1/2. This proves that the distribution of dispersion is almost normal.

Next item of interest is the relationship between the appearance rate and the appearance interval of a specific character X. Assuming that the value of appearance rate is known, a method of calculating the distribution of CIW will be established.

A first-order model of the appearance of characters is now considered where the individual appearance of character X is independent (i.e., no correlation is supposed to exist) and the appearance is ruled by a certain probability. This means the formation of word from characters is stochastic if the sample text is long enough.

Under this assumption, following the appearance of character X in a word, whether the same character X appears in the next word is subject to the probability r_g. The probability that no character X appears in the first following word and X appears in the second following word is $(1-r_g)^2 r_g$. The probability that X does not appear in the first and second following words and appears only in the third following word is, similarly, $(1-r_g)^2 r_g$. If we represent the appearance of X in a word by $\langle X\rangle$, and non-appearanec by $\langle \bar{X}\rangle$, the probabilities for the occurence of the trains of $\langle \bar{X}\rangle$ and $\langle \bar{X}\rangle$ are given in Table 1 below.

Table 1.

train of words	probability
$\langle X\rangle\langle X\rangle$	$r=r_g$
$\langle X\rangle\langle \bar{X}\rangle\langle X\rangle$	$r=r_g(1-r_g)$
$\langle X\rangle\langle \bar{X}\rangle\langle \bar{X}\rangle\langle X\rangle$	$r=r_g(1-r_g)^2$
⋮	⋮
$\langle X\rangle\langle \overbrace{\bar{X}\rangle\langle \bar{X}\rangle\cdots\langle \bar{X}}^{n}\rangle\langle X\rangle$	$r=r_g(1-r_g)^n$

This table predicts that the probability r decrease exponentially with n, the number indicating CIW. The validity of this model can be checked if $\log r$ is linear against n, since

$$\log r=\log r_g+n\cdot\log(1-r_g) \qquad (9)$$

and, as $1-r_g<1$, $\log(1-r_g)$ is negative, and this relation gives a straight line decreasing with n.

In an observation for the character combination "ch" in German, a CARW of 10% is

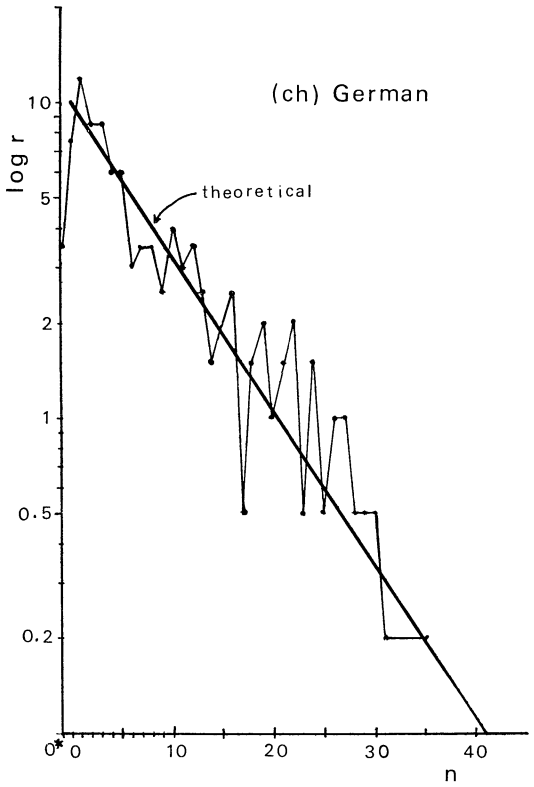

Fig. 6. Relation $\log r$ to n

obtained. The thick straight line in Fig. 6 is drawn for the relation (9) with $r_g=10\,\%$. The CIW for German (ch) is observed for 200 interval cases and the distribution of CIW is plotted against n. The observed distribution has slight deviation but the general trend of distribution agrees well with the theoretical straight line given by (9).

1.3.2 *Mean value of M_g and r_g*

In a text of consecutive D words, suppose that there are a words before the first word containing X and p words after the last word containing X (Fig. 7). Between the first and the last words containing X, suppose also that n words containing X are situated at equal

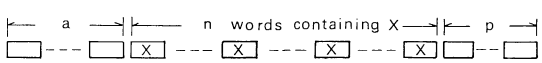

Fig. 7. Word String

interval $\bar{M}_g$. Then

$$(n-1)\bar{M}_g+n+a+p=D.$$

It is easy to see that

$$(n-1)\bar{M}_g+n<D,$$

then

$$\bar{M}_g<\frac{D-n}{n-1}. \tag{10}$$

It is understood that a and p are, in most cases, smaller than $\bar{M}_g$, thus

$$a+p\leqslant 2\bar{M}_g.$$

On the other hand

$$(n-1)\bar{M}_g+n+2\bar{M}_g>D,$$

then

$$M_g>\frac{D-n}{n+1}. \tag{11}$$

Combining (10) and (11)

$$\frac{D-n}{n-1}>\bar{M}_g>\frac{D-n}{n+1}. \tag{12}$$

Then, as an approximate value of $\bar{M}_g$, we get

$$\frac{D-n}{n}=\bar{M}_g \tag{13}$$

If $D\gg n$, this expression can be reduced to

$$\bar{M}_g=\frac{D}{n}.$$

This is an approximate value which is also an intuitive definition of $\bar{M}_g$.

Now, CARW r_g is, by definition,

$$r_g=\frac{n}{D}. \tag{14}$$

then, substituting (14) in (13), we get

$$\bar{M}_g=\frac{1}{r_g}-1. \tag{15}$$

Example: for "ch" in German, as $r_g=10\%$, $\bar{M}_g$ is calculated as 9.0 which agrees approximately with a measured value of 8.0.

1.4 Combinational Appearance of Different Characters

In the case of combinational appearance of two different characters X and Y, it is needed to consider two cases,

a) either X or Y will appear in a string of characters,

b) both X and Y will apear in a string of characters.

1.4.1 *Appearance of at least one of two characters*

Assuming the CIW's of characters X and Y as M_{gX} and M_{gY} and their mean values as $\bar{M}_{gX}$ and $\bar{M}_{gY}$, we get these mean values as

$$\bar{M}_{gX}=\frac{D-n_X}{n_X}, \text{ and } \bar{M}_{gY}=\frac{D-n_Y}{n_Y} \tag{16}$$

The mean CIW for the appearance of logical sum $X+Y$ can be considered as

$$\bar{M}_g(X+Y)=\frac{D-(n_X+n_Y)}{n_X+n_Y}. \tag{17}$$

Now (16) can be transformed into

$$n_X=\frac{D}{1+\bar{M}_{gX}} \qquad n_Y=\frac{D}{1+\bar{M}_{gY}} \tag{18}$$

and by substituting expression (18) in (17), we get

$$\bar{M}_g(X+Y)=\frac{(\bar{M}_{gY}+1)(\bar{M}_{gY}+1)}{\bar{M}_{gX}+\bar{M}_{gY}+2}. \tag{19}$$

In the case where M_{gX} and M_{gY} are much greater than unity, the expression above can be reduced to

$$\bar{M}_g(X+Y)\doteqdot\frac{\bar{M}_{gX}\cdot\bar{M}_{gY}}{\bar{M}_{gX}+\bar{M}_{gY}}. \tag{20}$$

Example: In German $\bar{M}_g$(ch)=7.9 words and $\bar{M}_g$(sch)=12.8 words. Using (19), $\bar{M}_g$(ch+sch) =4.9 words. The observed value is $\bar{M}_g$(ch +sch)=5.1 words, and the deviation is about 4 per cent.

In a text of D words, if characters X and Y appear n_X and n_Y times, respectively, the appearance rate of "either X or Y" is, by definition,

$$r_g(X+Y)=\frac{n_X+n_Y}{D}.$$

Then it is easy to see that

$$r_g(X+Y)=r_{gX}+r_{gY} \tag{21}$$

1.4.2 *Appearavce of both of two characters*

The problem of the simultaneous appearance of two kinds of characters in the same string of characters can be solved by using CIW's.

Now, CIW's for characters X and Y are designated as M_{gX} rnd M_{gY} respectively. Let us assume, for convenience, $M_{gX}>M_{gY}$. The appearance of X and Y are schematically represented in Fig. 8 which is divided into two cases.

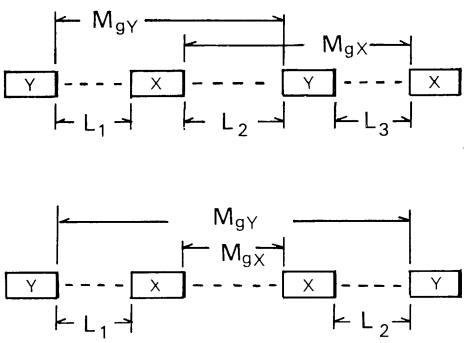

Fig. 8. Two cases of $M_{gx}-M_{gy}$ Relation

CASE A. This case represents one where the magnitude of M_{gX} and M_{gY} are of the same order. Concerning the length of word string which contains at least one X and one Y,

Lower limit value is found among

$$L_1+2,\ L_2+2,\ L_3+2,\ \ldots \tag{22}$$

where

$$L_1+L_2+1=M_{gY},\ L_2+L_3+1=M_{gX}.$$

(It is noted that these lower limit values can be reduced to zero.)

Upper limit value is

$$L_1+L_2+L_3+2=M_{gY}+L_3+1=M_{gX}+L_1 1 . \quad (23)$$

Among the upper limit values, the maximum value is found as

$$M_{gY}+M_{gX} \quad (24)$$

in the case where

$$L_2=0 .$$

CASE B. This is the case where M_{gY} is much greater than M_{gX}. In this circumstance, the length of word string in which at least one X and one Y is included is:

Lower limit values are found among

$$L_1+2,\ L_2+2,\ L_3+2,\ \ldots \quad (25)$$

where interval L_1, L_2, etc. can be zero.

Upper limit value is

$$M_{gY}+1\approx M_{gY} . \quad (26)$$

And the last approximate expression is valid when $M_{gY}\gg 1$.

In actual texts, both cases, A and B, may occur simultaneously, it is therefore, enough to conclude that the estimation of the upper limit of word string length which contains both X and Y, or the occurrence of the logical product of X and Y (represented as $X*Y$) is

$$\left.\begin{aligned} &M_{gX*Y}=M_{gX}+M_{gX} \quad \text{when } M_{gX}\approx M_{gY} \\ &\text{and} \\ &M_{gX*Y}=M_{gY}+1 \quad \text{when } M_{gX}\ll M_{gY} \end{aligned}\right\} \quad (27)$$

PART II

In some documentation and library works, one must handle many documents in many different languages. The library staff and clerks are not expected to be understand every language used in documents. However, it is very desirable that the personnel in documentation can identify the language he or she is confronting by some simple process even if he or she cannot read words nor understand the meaning of the text handled.

This Part II is devoted to find out a process of identifying the language in use of a document by only looking at the characters and simple words. Further it is hoped the process is applicable to computerization without difficulty. To this end, a very general description of Roman alphabets in different languages is given so as to be able to treat languages in a unified way.

2.1 Generalities on Roman Alphabet

2.1.1 *Alphabet*

Roman or Latin alphabet consists of basic 26 characters, additional special characters and characters with marks. Very faw languages use only basic 26 characters, for instance, English (except for loan words and less fashionable use of diacritics, such as 'coördinate' and Dutch.

Many languages use some additional characters and these can be classified into three categories:

(i) Independent additional characters···Examples are German ß, Icelandic þ and ð, Turkish ı (lower case), etc.,

(ii) Joint characters···Examples are œ in French and Latin, æ in Danish and Latin, ch in German, etc. Joint characters used by printers fi, ff, and ffi are not considered here.

(iii) Combined characters···Examples are 11 in Spanish, gy in Hungarian, etc. In these languages, these combined characters are regarded as independent alphabets and they have their proper position in alphabetical order.

However, we neglect this third category because they are not needed to distinguish from simple combination of basic characters.

2.1.2 *Diacritical signs*

These are added to basic characters and they can be divided into three clases, i.e., upper, middle and lower as shown in Table 2 below.

Table 2. Examples of Diacritical Signs

Marks	Example
Upper marks	é à ō a č ŝ ñ ż
Middle marks	ø đ ł
Lower marks	ç ș ą ę ạ ọ

2.2 Identification Techniques

2.2.1 *Use of single characteristic*

The writen text in a language, espcially in printed form, has many useful characteristics in indentifying that language from others. These are, in the order of easiness in finding out,

(a) independent additional characters and joint characters,
(b) characters with diacritical marks,
(c) hehavior of special character combination,
(d) simple short words,
(e) lack of specified characters,
(f) vowel-to-consonant character ratio at endings.

Among these, some characteristics are very easily found, but others not. For instance, the presence of characte (ñ) (category b) is very unique in identifying Spanish. However, its CIW is very large (M_0=360 words). It is, therefore, of little use in the identification of Spanish, unless we have a sample of 300 to 500 words in length.

The presence of character ß (category a) is decisive in identifying German. The mean value of CIW is, as is mentioned in 1.2, about 22. It is also shown that at least one character can be found in any sample of 125 words or more with a significance level of 5 percent (Fig. 3). However, this is not practical and we need to use other more frequent characteristics or to combine both in practical cases.

One might think that the appearance rate of some characters or words is useful for identification purpose. However, in practical cases, sample to be identified are hoped to be as short as possible. We have learnt in the Part I that the shorter the length of a sample is, the greater dispersin of appearance rate results.

Then, if a certain character appears in two or more languages and even if the appearance rates measured with a very large population is different for each language, the trial to inentify languages by the value of appearance rate observed with a short sample will fail in most cases.

For instance, (ä) appears in German and Swedish and CARW's for German is 4.5% and that for Swedish is 11.3 %. This appreciable difference will induce us to the use of CARW value in a sample for identification. In a sample of 50 words, if we find eight or more (ä)'s, we can judge safely that this text is in Swedish but if the number is less than six, we cannot judge whether the language is Swedish or German. Fig. 9 shows the dispersion of observed CARW values for both languages with 20 samples of 50 and 100 words

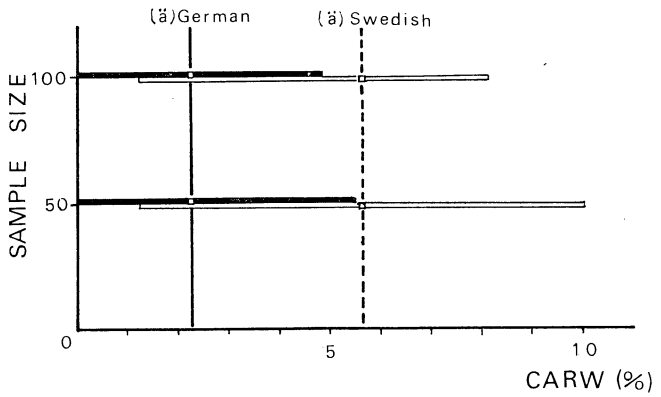

Fig. 9. Dispersion of CARW

Table 3. CARW of English Characters

Vowel characters			Consonant Characters								
	Total	Endings		Total	Endings		Total	Endings		Total	Endings
a	8.4	3.4	b	2.1	0	k	0.7	0.8	s	7.1	12.8
e	10.7	18.2	c	2.6	1.6	l	3.9	2.6	t	10.2	8.6
i	8.0	0	d	3.7	11.6	m	2.4	2.0	v	0.8	0
o	8.0	4.4	f	2.0	5.2	n	7.3	9.0	w	2.2	1.4
u	3.1	0.1	g	1.9	2.4	p	1.6	0.6	x	0.2	0
y	2.1	5.4	h	5.1	4.2	q	0.1	0	z	0.2	0
Total	40.3	31.6	j	0.1	0	r	5.5	5.6	Total	59.7	68.4

There still exist other useful characteristics; these are special combination of characters in the first and last syllable of words. For instance, the ending (-ing) is abundant in English and Swedish and (-ung) in German. Also (-um) and (-us) are characteristics effective for the identification of Latin and (-sjon) is peculiar to Norwegian.

In most languages the CRAW or CARC at the word ending is different from that in the words. Table 3 shows an example for English. This fact is based on several reasons:

(i) in many languages, plural from is expressed by adding a certain endding or endings, e.g., (-s) in English and French,

(ii) in Scandinavian languages definite article is expressed by endings (-en) or (-n) and independent definite artice does not exists,

(iii) in some languages, passive voice is much used, thus a particular ending for participle, e.g., (-ed) in English' appears very frequently,

(iv) in some languages majority of syllables are open so that only vowel characters appear in the endings, e.g., in Italian and transliterated Japanese, and the ratio vowel-to-consonant (V-C ratio) at the ending is much greater than other languages. (see Table 4).

Table 4. V/C for Typical languages

Language	Vowel-to-Consonant Ratio (V/C)		
	Mean value	Standard deviation	2σ range
German	0.30	0.15	0 - 0.60
Englsh	0.45	0.10	0.25- 0.65
Italian	7.2	2.7	1.8 -12.6
Japanese		2.8	2.5 - 8.1

2.2.2 *Prominent simple words*

One can say that all words in a language have some features. The adoption of words as characteristics of a language is, therefore, desirable but, on the contrary, it necessitates some knowledge in that language and this is not permissible for our case. Then we must restrict ourselves to the use of very short, frequent and easily recognizable words. Some examples are shown below:

(a) Words with one character

I (English), the appearance rate differs appreciably according to the style of texts. Also the word i exists in Italian, Scandinavian and Slavonic languages.

e (Italian), $r_g=2.0$ %

y (Spanish), $r_g=3.0$ %, this word exists also in French but its appearnce rate is as small as $r_g=0.1$ %.

v (Slavonic language),

etc.

(b) Words with two characters
of (English), $r_g=3.7\%$,
et (Latin), $r_g=3.4\%$; (French), $r_g=2.1\%$,
og (Danish and Norwegian), $r_g=3\%$,
etc.

(c) Words with three characters
and (English), $r_g=2.6\%$,
och (Swedish), $r_g=3.4\%$,
van (Dutch), $r_g=4.2\%$,
und (German), $r_g=8\%$,
the (English), $r_g=7.4\%$,
etc.

All these words are very probable candidate for the criteria of language identification.

2.2.3 *Combination of plural characteristics*

The important factor to evaluate a language identification scheme is precision and the applicability to small samples. For this last purpose the use of *logical sum* is very useful.

In 1.4.1 we have got formula (20) which gives mean CIW of characters X OR Y when CIW's for both characters are known. This formula can be extended to any number of characteristics to be combined by logical sum.

Example: Character (ą) and (ę) are characteristic feature of Polish. However the CARW's for both characters are relatively small; in fact, r_g(ą)=5% and r_g(ę)=8%. If we want to handle a sample of 10 to 20 words, it is necessary that CARW of a criterion for identification is greater than 10%. If we combine, therefore, (ą) and (ę) by an OR, the resultant r_g becomes 14%, thus the condition $r_g>10\%$ is satisfied.

The precision of identification is improved by the use of *logical product*. For instance, the word [et] appears both in French and in Latin and WARW's are large in both languages. In French r_g[et]=2.1% and $\bar{M}_g=47$ words and upper limit of M_g at five percent significance level, that is $\bar{M}_g+2\sigma$, is 110 words. In Latin, on the other hand, $r_g=2.6\%$, $\bar{M}_g=37.5$ words and upper limit is 116 words. If we find [et] in a text, we can say it is either in French or in Latin but, being r_g's are nearly equal, we cannot distinguish French from Latin by the value of r_g or M_g.

In French and in many other languages excluding Latin, character (é) exists and $r_g=14\%$, and $\bar{M}_g=6.1$ words in French. In Latin, on the other hand, the endings (-um) and (-us) appears frequently, as mentioned earlier, and $r_g\{$(-um)+(-us)$\}=7\%$. In French this two endings appear in some loan words from Latin but their number is quite small. Then we can propose tentative criteria for identifying French and Latin as:

for French: existence of word [et] and character (é), which can be expressed as the existence of [et]*(é),

for Latin: existence of word [et] and endings (-um) or (-us) which can be expressed as the existence of [et]$*\{$(-um)+(-us)$\}$.

By these criteria we can identify, without ambiguity, French and Latin texts. The number of words necessary for their application is calculated as follows:

for French: $\bar{M}_g$[et]=47, $\bar{M}_g$(é)=6, then by (27) in 1.4.2 and considering that $\bar{M}_g$[et] $\gg\bar{M}_g$(é), then $\bar{M}_g$[et]*(é)=48 words,

for Latin: $\bar{M}_g$[et]=28, $\bar{M}_g\{$(-um)+(-us)$\}=12$, using (27) in 1.4.2, we get $\bar{M}_g$[et]$*\{$(-um)+(-us)$\}=29$ words.

This result shows that the criteria is valid but they are not quite practical yet.

2.3 Establishment of Criteria for Identification

In this section, the way of estrblishing criteria by choosing adequate characteristics is given, based on some examples.

2.3.1 *Measurement of appearance rate*

Given a language, the first step is to choose characteristic characters and after measuring their CARW's we will place the characters in order of CARW value. Taking German as example, the characteristic characters are chosen as (ä), (ö), (ü) and (ß). The CARW's are shown in the Table 5.

Secondly we choose characteristic combina-

Table 5. CARW of Some Frequent Characters and Character Combinations and WARW of Frequent Words

character	CARW	chracter combination	CRAW	Word	WARW	Word	WARW
(ä)	3.2%	(ch)	11.5%	[die]	3.8%	[zu]	1.0%
(ö)	2.2	(sch)	6.3	[der]	3.5	[von]	1.1
(ü)	4.1	(ng)	4.6	[und]	3.4	[dem]	1.0
(ß)	2.6	(ung)	3.3	[in]	2.2	[auf]	0.9
				[den]	1.3	[es]	0.9
				[mit]	1.1	[das]	0.8
				[sie]	1.1	[des]	0.5

tion of characters. In German (ch), (sch), (ng) and (ung) are chosen. Their CARW's are also shown in Table 5.

Thirdly we choose some simple words which are likely to be characteristic. In German, for instance, we have [und], [der], [des], [dem], [den], [die], and [des].

2.3.2 *Use of logical sum and product*

By combining such characteristics as given above, we can establish criteria and check if these are applicable for a small sample, i.e., to check $\bar{M}_g$'s.

The trial in establishing criteria will proceed, in the case of German, like following:

a. (ß) This character is very unique and can be used as criterion. However, the CARW is low, r_g=2.2%.

b. (ä)*(ö)*(ü) The characters (ä), (ö) and (ü) appear in many languages. But simultaneous appearance of these three occurs only in German. Then the logical product of these three characters can also be a criterion, if we tolerate that its CARW is rather low.

c. [und]*[der] This is also a correct criterion for German. We regret that its WARW is low, r_g[und]*[der]=1.5%.

All these three criteria are valid, we only regret r_g's are low for practical use. Now we will try to increase r_g's by using logical sum technique. For instance, we can replace [der] by [der]+[des]+[dem]+[den]+[die]+[das], the CARW increases up to 10.9%. However, with the single criterion we cannot judge that the text is in German because the word [die] exists in English and Dutch and [des] in French.

The combination by logical sum

$$(ä)+(ö)+(ü)+(ß)$$

gives an r_g of 12.1%. However, this single criterion is not valid because (ä), (ö) and (ü) exist separately in many languages.

Now, taking the logical product of two criteria, i.e.,

$$\{[der]+[des]+[dem]+[den]+[die]+[das]\} *\{(ä)+(ö)+(ü)+(ß)\} \quad (D4)$$

is very promising criterion for German. The r_g for this combination becomes 6.1% (see 1.4 for the method of calculation).

We can construct similar combination of logical product of which r_g exceeds 10%. By using such expressions as

α=[der]+[des]+[dem]+[den]+[die]+[das]
β=(ä)+(ö)+(ü)+(ß)
γ=[und]+[von]+[vom]+[vor]+[ist]+[hat]
δ=(ch)+(sch)

and taking product of α, β, γ and δ, we get criteria

$$\alpha * \beta \quad (D4)$$
$$\alpha * \gamma \quad (D5)$$
$$\alpha * \delta \quad (D6)$$
$$\beta * \gamma \quad (D7)$$

Table 6. Identification by the Criteria for German

Test Sample Number	α	β	γ	δ	D_4	D_5	D_6	D_7	D_8	D_9	Minimum among D's
1	10	6	1	5	10	6	6	10	10	**5**	5
2	X	X	2	3	X	X	X	X	X	**3**	3
3	15	6	7	4	15	7	**6**	15	15	7	6
4	4	20	X	14	20	X	20	**14**	X	X	14
5	13	15	X	8	15	X	15	**13**	X	X	13
6	6	9	X	11	**9**	X	11	11	X	X	9
7	5	7	8	13	**7**	8	13	13	8	13	7
8	1	X	3	5	X	X	X	5	**3**	5	3
9	4	11	18	7	**11**	18	**11**	7	18	18	11
10	4	19	X	6	19	X	19	**6**	X	X	6
11	1	X	7	2	X	X	X	**2**	7	7	2
12	1	6	7	**6**	**6**	7	**6**	**6**	7	7	6
13	7	2	4	2	7	4	**2**	7	7	4	2
14	10	6	2	5	10	6	6	10	10	**5**	5
15	9	X	8	5	X	X	X	9	9	**8**	8
16	1	13	3	15	13	13	15	15	**3**	15	3
17	1	12	6	5	12	12	12	**5**	6	6	5
18	1	3	6	9	**3**	6	9	9	6	9	3
19	X	1	6	9	X	**6**	9	X	X	9	6
20	X	12	2	3	X	12	12	X	X	**3**	3
21	1	2	X	X	**2**	X	X	X	X	X	2
22	1	8	3	17	8	8	17	17	**3**	17	3
23	1	X	14	2	X	14	X	**2**	14	14	2
24	1	10	8	6	10	10	10	**6**	8	8	6
25	17	9	X	7	17	X	**9**	17	X	X	9
26	2	11	14	9	11	14	11	**9**	14	14	9
27	4	7	12	8	**7**	12	8	8	12	12	7
28	3	12	6	8	12	12	12	8	**6**	8	6
29	3	10	X	5	10	X	10	**5**	X	X	5
30	13	5	10	8	13	10	**8**	13	13	10	8
31	2	1	X	5	**2**	X	5	5	X	X	2
32	4	X	2	2	X	X	X	4	4	**2**	2
33	X	7	X	2	**7**	X	**7**	X	X	X	7
34	1	3	X	4	**3**	X	4	4	X	X	3
35	7	19	3	1	19	19	19	7	7	**3**	3
36	1	X	5	3	X	X	X	**3**	5	5	3
37	17	1	X	1	17	X	**1**	17	17	X	1
38	4	5	6	5	**5**	6	**5**	**5**	6	6	5
39	2	16	1	8	16	16	16	8	**2**	8	2
40	3	2	X	4	**3**	X	4	4	X	X	3

$$\beta * \delta \qquad \text{(D8)}$$

$$\gamma * \delta \qquad \text{(6D)}$$

We will check how these criteria are applicable to actual samples. A group of 40 samples of 20 words each have been chosen and criteria (D4) to (D9) are applied to this group of 40. Results are shown in Table 6. The effectiveness of criteria (D4) to (D9) is clearly shown in the Table 6. Among 40 samples, 32 or 80 percent, are identified by (D6) and the average number of words needed to identify is 9.9. Those samples marked with X in the Table 6 are too short for identification by a single criterion. If we use, in parallel, all of the criteria (D4) to (D9), i.e., to combine these six criteria with OR's, all 40 samples are identified and the average number of words needed to identify is reduced to 5.4; this shows the effectiveness of logical sum technique. The last column in Table 6 shows the minimum length for each sample to be identified successfully.

2.4 Error in Identification

In the preceding section, we have established criteria of identification for German, there still remain problems about the misidentification made for samples in non-German languages. To ascertain this, ten samples of 20 words were prepared for each of 25 languages and test was made whether any among these samples is misidentified as German.

The result shows that misidentification took place for two cases, i.e.. a Dutch sample is misidentified as German by (D6) and a Swedish sample by (D8). The reason for these misidentification is discussed below.

2.4.1 *Discussion on the first error*

The text which caused the first error is shown below:

> We hebben in onze gesprekken nogal wat klachten gehoord over de druk, soms dwang, die vanwege het departement···

It is easy to find out the reason: Dutch text contains (ch) and [die]. Character combinations (ch) and (sch) are very common in both German and Dutch; these characteristics are, therefore, misleading. However these two appear very frequently and, therefore, they are useful for identification use, if we combine these with other elements to prevent misidentification.

The prevention of this error is simple if we add negative condition to exclude Dutch texts, i.e., to combine the considered criterion (D6) with the term

$$\overline{[\text{de}]+[\text{een}]+[\text{het}]+[\text{van}]}$$

where the notation " — " stands for negation.

We can expect to have similar error, though this was not the case with our test samples, with English text which contains the English word [die] and character combinations (ch) and (sch). To prevent this possible error, we can extend the the negative condition shown above to exclude English too by giving the form

$$\overline{[\text{de}]+[\text{een}]+[\text{het}]+[\text{van}]+[\text{and}]+[\text{of}]+[\text{the}]}.$$

The criterion now becomes

$$\alpha * \delta * \{\overline{[\text{de}]+[\text{een}]+[\text{het}]+[\text{van}]+[\text{and}]+[\text{of}]+[\text{the}]}\} \qquad \text{(D6')}$$

2.4.2 *Discussion on the second error*

The second error occured on a Swedish sample which contains (ch) and (ö). The combination (ch) occurs very frequently in Swedish (r_g(ch)=10%.) but this (ch) appears only in the word [och], meaning "and." Because we have no German word [och], we can avoid the difficulty by adding the negative condition that "no [och] appears" to the criterion for German. By using the same notation as is shown in 2.4.1, the new criterion for German is now

$$\beta * \delta * \overline{[\text{och}]} \qquad \text{(D8')}$$

2.5 Cases of Hardly Identifiable Languages

In establishing criteria for each language in the way shown in 2.4, we find some difficulties in some languages which have strong similari-

ty. The cases are relations Danish-Norwegian and Serbo-Croat-Slovene-ISO-transliterated Russian. In these languages, the difference in characters are very few and frequent words are almost identical since these languages belong to the same family.

2.5.1 *Case of Danish-Norwegian*

Especially the case of Norwegian is difficult. Written Norwegian was almost common with Danish in the 18th century. The difference became significant after the change of political situation in 1814 and successive reform in orthography in this century. Moreover, even now, Norwegian has two forms of language Riksmål and Landsmål as national language.

Table 7. Criteria for Danish and Norwegian

Criteria	
Dan 1	{(å)+(aa)+(ø)+(æ)} * **D** * $\bar{\mathbf{I}}$*{$\overline{(ä)+(ö)}$}
Dan 2	{[og]+[at]+[i]+[det]}*$\mathbf{D}$*$\bar{\mathbf{I}}$*{$\overline{(ä)+(ö)}$}
Nor 1	{(å)+(aa)+(ø)+(æ)} * **N** * $\bar{\mathbf{I}}$*{$\overline{(ä)+(ö)}$}
Nor 2	{[og]+[at]+[i]+[det]}*$\mathbf{N}$*$\bar{\mathbf{I}}$*{(ä)+(ö)}

where

D = {[ad]+[af]+[blev]+[efter]+[ind]+[mig]+[mellem]+[nu]+[op]+[sig]+(øj)+r_{θ}(æ)>6%+(tion)+(-hed)}
I = (þ)+(ð)+(v́)−(é)
N = [av]+[a]+[ble]+[enu]+[ett]+[etter]+[inn]+[meg]+[mellom]+[nå]+[opp]+[seg]+(øy)+(sjø)+(-**cc**)+(sjon)+(-het)

The criteria chosen are shown in Table 7. Here, the identification process is divided into four steps. The first terms

$$\{(å)+(aa)+(ø)+(æ)\}*\overline{\{(ä)+(ö)\}}$$

in Dan-1 and Nor-1 and the first terms

$$\{[og]+[at]+[i]+[det]\}*\overline{\{(ä)+(ö)\}}$$

in Dan-2 and Nor-2 are common, respectively, that means we first identify the sample under test by checking if this is in (Danish or Norwegian) or non-(Danish or Norwegian).

Next step is to distinguish from Norwegian in a mixture of Danish and Norwegian. The third terms **D** and **N** contain character combinations and words which are similar but not identical in the two languages. The third term $\bar{\mathbf{I}}$ is to prevent the misidentification of Icelandic to Danish or Norwegian, and the fourth term is to exclude Swedish.

2.5.2 *Case of three slavic languages*

Three languages among Slavic family, Serbocroat (especailly Croat which employs Roman alphabet), Slovene and ISO-transliterated Russian use the same alphabet which contain č, š, ž, etc. Croat uses other characters such as ć and đ but CARW's for these are low. Other member of Slavic languages with Roman alphabet, e.g., Polish and Czech use quite different orthography and they are easily recognizable.

The trial was made for these three languages to establish criteria using words. The abbreviations Hrs, R-i and Slv are used to represent Croat (Hrvatska), Slovene, ISO-transliterated Russian, respectively. Another way of Russian transliteration standardized by the British Standards and the American Standards is represeted as R-b. The expressions [Hrs], [Slv], [R-i] represent any word in these three languages, respectively, and, if needed, they are accompanied by a suffix. The words common in two languages are represented, for example, by $[\text{Hrs-Slv}]_i$, $[\text{Slv-Ri}]_j$, etc.

Now, if we have a sample satisfying the relation:

$$[\text{R-i-Hrs}]_i*[\text{Hrs-Slv}]_j\,, \qquad \text{(S1)}$$

it is Croat. If we employ only this relation (S1), the applicability is very low. So we need to employ the OR combination of such relations

$$\sum_{i,j}[\text{R-i - Hrs}]_i*[\text{Hrs - Slv}]_j\,.$$

Next, we define three expressions as below

$$[\mathbf{R\text{-}i - Hrs}] = \sum_i [\text{R-i - Hrs}]_i \qquad \text{(S2)}$$

etc.

the criteria for each of three languages become

for Russian(ISO) **S*[R-i-Hrs]*[R-i-Slv]** (S3)

for Croat **S*[R-i-Hrs]*[Hrs-Slv]** (S4)

for Slovene **S*[R-i-Slv]*[Hrs-Slv]** (S5)

where **S** is the term to represent criterion common to these three languages.

It is, of course, possible to find out words which appear only in a language. If we represent these words by **[Ri]**, **[Hrs]**, and **[Slv]**, we can establish criteria for identification as

for Russian(ISO) **S*[R-i]** (S6)

for Croatian **S*[Hrs]** (S7)

for Slovenian **S*[Slv]** (S8)

After these considerations, criteria have been established as shown in the Table 8. Check was made for 15 samples of 30 words in each of three languages. The result is shown in Table 9.

Table 9 shows that these criteria give fairly good results. In general the criteria (S3) to (S5) require more number of words than the criteria (S6) to(S8). The samples marked with *X* show that the identification was impossible with 30 words. If we use both series of criteria, the identification was always possible with 30 words.

Table 8. Criteria for the Three Slavic Languages

(**S**) = {(č)+(š)+(ž)}*{(v̌)+(v̆)+(d̆)+(ň)+(ř)+(ť)
(**R-i**) = [ėti]+[ėto]+[ėtu]+[gde]+[kto]+[vy]
(**Hrs**) = [koji]+[sa]+[su]+[što]
(**Slv**) = [in]+[bo]+[ki]
(**R-i-Hrs**)=[i]+[ili]+[kod]+[u]
(**Hrs-Slv**)=[bi]+[je]+[pa]+[od]+[pri]+[se]+[sve]
(**R-i-Slv**)=[ko]+[so]+[v]+[vse]+[z]+[že]

2.6 Identification of Languages

2.6.1 *Target languages and criteria for identification*

The number of languages under consideration is limited to 22 national languages and an international language both using Roman alphabet and being in use to convey scientific and techni-

Table 9. Test Results for Three Slavic Languages

Criteria	Sample Number 1	2	3	4	5	6	7	8	9	10	Language of Sample
[S]*[R-i - Hrs]*[R-i - Slv]	17	X	20	X	11	X	7	22	18	10	R - i
[S]*[R-i - Hrs]*[Hrs- Slv]	19	23	11	23	2	19	20	18	X	X	Hrs
[S]*[R-i - Slv]*[Hrs - Slv]	15	X	6	X	13	29	—	—	—	—	Slv
[S]*[R-i]	5	5	4	12	6	5	6	2	2	7	R-i
[S]*[Hrs]	7	14	11	23	20	4	X	8	8	50	Hrs
[S]*[Slv]	39	14	17	17	22	38	—	—	—	—	Slv

Table 10. List of Considered Languages

Croat	French	Latin	Slovene
Czech	German	Norwegian	Spanish
Danish	Hungarian	Polish	Swedish
Dutch	Icelandic	Portuguese	Turkish
English	Indonesian	Rumanian	Vietnamese
Esperanto	Italian	Russian (ISO and BS transliterated)	
Finnish	Japanese (transliterated)		

cal knowledge and information as well as two other languages having transliteration rules into Roman alphabet and are appearing frequently in actual library catalogs and index journals. Their names are listed in Table 10.

Local languages such as Scottish, Welsh, Basque and Yiddish are not considered since we can hardly imagine that scientific or technical documents appear in these languages. On the contrary, two or more authorized languages used in a federal nation are considered. Lithuanian, Latvian and Estonian should be included but these are excluded because of the difficulty in the present author's situation.

It should be pointed out that the principle involved in this paper can be applied to other languages employing non-Roman alphabet, for instance, to the case of Cyrillic alphabet.

The criteria for identification are, as it is seen in Table 11, sometimes very simple (e.g., the case of Icelandic and Hungarian) but in most cases two or three alternative criteria are combined by OR relation to increase the applicability to short samples. The most complicated case is for Danish and Norwegian. For the three Slavic languages mentioned above (2. 5. 2), the criteria finally chosen are simplified than that discussed earlier.

The relation among Roman languages—Spanish, Portuguese, and Italian—is somewhat complicated because of added negative conditions to prevent misidentification among them and Esperanto which is similar in some respect.

2. 6. 2 *The identification process*

Though the identification process seems to be complicated, it is, however, rather simple if we use pattern sheet shown in the Fig. 10. The "Identification or ID sheet" shown in Fig. 10 contains all of 261 characteristics employed, including some characteristics which are not use in criteria adopted in Table 11, divided into categories of vowel characters, consonant characters, special additional characters, combinations of characters, endings, and words. Manual identification process with this sheets is explained below.

Text to be identified is examined at sight and if the text contains specified characteristics, the positions corresponding to the characteristics are marked black with a thick pencil or felt pens, as is done in any mark sheet or mark sense card. To examine all characters and words for a text of 10 to 20 words, it takes about 3 minutes.

On the other hand, "pattern sheets" are provided for each language. Each position of characteristic is punched in round hole if this characteristic is used in the criteria for identification. The logical relation is shown on the pattern sheets by colored lines (in the Fig. 11 colored lines are replaced by different kind of black lines). If holes are connected by the line of a color, they are linked with OR relation. The logical product is shown on the pattern sheets by an asterisk placed between lines of different colors. For negative characteristics, holes are in rectangular form so as to be able to recognize them easily. The logical formula are also printed in lower part of the sheet to help users.

If a text is given, characteristics given in the ID sheet is examined and relevant positions on the ID sheet are marked, then pattern sheets are placed on the ID sheet one after the other and if black positions appear through holes, we can judge these black marks conform to the logical equation of identification.

As an example, an unknown text shown below is presented.

> Besonders gefährlich und deshalb unbedingt zu vermeiden sind irgendwelche starren···
>
> (10 words)

We examine the text and relevant positions are marked (see Fig. 10). When pattern sheet for German is placed over the ID sheet (see Fig. 11), we see that in the chain 1 we have one black mark in a hole, that means criterion 1 is satisfied (in fact, (ä) exists). In the chain 2 we have one black mark (in fact, (ch) exists) and negative condition are also satisfied as we see no mark in rectangular hole). We can

I D S H E E T

No. 239

A Characters

Vowel characters

●	â	á	à	å	ã	ă			ą				
ë	ê	é	è		ě	ė			ę				
ï	î	í	ì			ǐ	İ						
ö	ô	ó	ò		õ			ő	ø				
ü	û	ú	ù	ů	ũ	ŭ		ű					

Consonant characters

č	ĉ	ć			ç				ķ	š	ŝ	ś		ș
ď			đ	ď	ḍ		l'	ł	ļ	ť				ț
ğ	ĝ					ň	ń	ñ	ņ	ž		ź	ż	
ĥ	ĵ													
	y	ý	ȳ		●									

B Additional and Joint characters: æ œ ij ł ð þ ß

C Combined characters

aa	ää	ão		ee	ée	eu		ii	ij		
õe	oo	öö	øj	øy		uu	uw	ya	yo	yu	yy
●	cs	cz		dj	dz	dzs		gn	gy		kh
lj		ngh	nj	ny		qu	rz	sch	●	sj	sjø sz
tch	th	tion		tj		zh					

D Endings

-d	-ed	-g	-gg	-hed	-het	-ii	-ing	-iya	-jn	-m	-nn	-no
-oi	-ogo	-oi	-tt	-um	-ung	-us	-y	-z				

E Words

a	å	ad	af	an	and	at	att	av	az	bir	ble	blev
che	com	da	dan	das	de	del	dem	der	des	det	di	die
do	du	e	è	een	efter		el	em	en	enn	és	est
et	ett	etter		för	ga	ha	hat	het	i	I	în	in
ind	inn	ist	itu	jang	kaj	ke	ki	la	meg	mellem		mellom
mig	na	ni	no	nu	o	och	of	og	oléh	op	opp	
på	que	qui	s	sa	seg	şi	sig	the	to	u	U	um
uma	un	una	und	upp	v	van	ve	vom	von	vor	wa	y
z	za	že	●									

F V/C <0.7 0.7-1.3 >1.3

G Identification:

Fig. 10. ID Sheet

judge that criteria 3 (i.e., ①＊②＊N_1＊N_2) is satisfied, then the text is in German.

In placing other ID sheets, e.g., that for English, we see the chain 1 of criterion 2 is satisfied but no other chain is. We conclude that the sample is not in English. Repeating this process for each language, we conclude that the text is in German and not in any other 24 languages.

The whole process takes no more than 3 minutes.

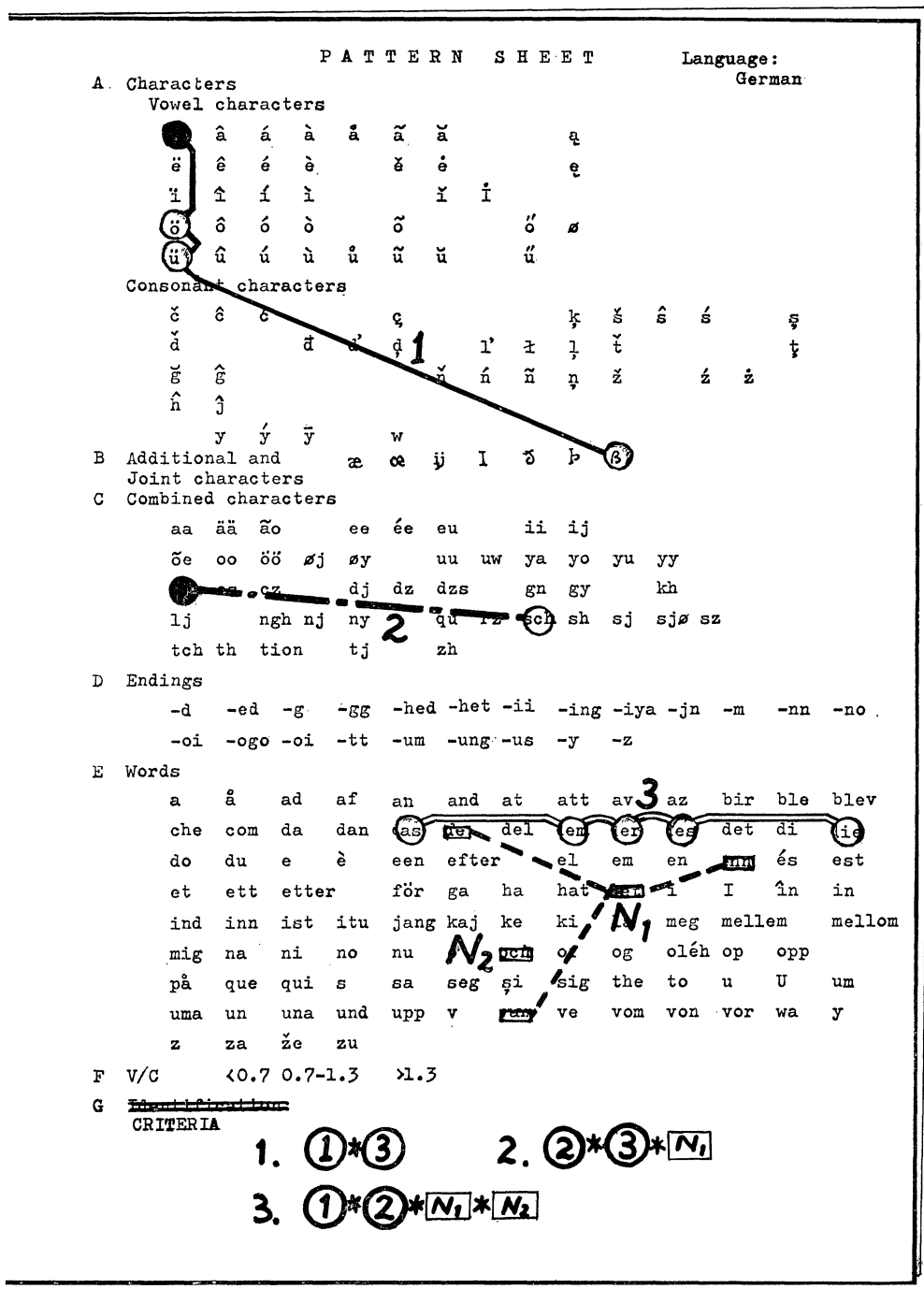

Fig. 11. Pattern Sheet for German Superposed on an ID Sheet

Table 11. Complete Criteria for 25 Languages Considered

( )	characters or combination of characters, [ ] words,
c	any consonant characters, **v** any vowel characters
$\mathbf{c}_d$	any consonant characters with diacritical signs,
$\mathbf{v}_d$	any vowel characters with diacritical signs,
(–)	endings, —— negation,
–	exclude, e.g., $\mathbf{v}_d$ – (é) any vowel characters with diacritical signs excluding character é.

Croat		{(č)+(š)+(ž)}*{(ć)+(đ)+[u]}*{**c'**)+(y)}
Czech	1	{(č)+(š)+(ž)}*{(á)+(ě)+(í)+(ó)+(ů)+(ý)+(ch)}
„	2	(ď)+(ň)+(ř)+(ť)
Danish	1	{(å)+(ø)+(æ)+(aa)}*{(øj)+(tion)+(-hed)+[ad]+[af]+[blev]+[efter]+[ind] +[mellem]+[mig]+[nu]+[op]+[sig]}*{(ä)+(ö)}*{(á)+(í)+(ó)+(ú)+(ð)+(þ)}
„	2	{[at]+[det]+[i]+[og]}*{(øj)+(tion)+(-hed)+[ad]+[af]+[blev]+[efter]+[ind] +[mellem]+[mig]+[nu]+[op]+[sig]}*{(ä)+(ö)}*{(á)+(í)+(ó)+(ú)+(ð)+(þ)}
Dutch	1	{(aa)+(ij)+(sch)}*{(ee)+(oo)+(-ing}*{$\mathbf{v}_d$- (é) - (ë)}
„	2	{[een]+[het]+[van]}*{$\mathbf{v}_d$- (é) - (ë)}
English	1	[and]+[of]+[the]
„	2	{(sch)+(ch)+(-ing)}*{[a]+[an]+[in]+[to]} *{(ä)+(ö)+(ü)+(ß)+[dem]+[der]+[das]+[een]+[het]+[van]+[de]}
Esperanto		{(ŭ)+(ĉ)+(ĝ)+(ĥ)+(ĵ)+(ŝ)}*{[de]+[kaj]+[la]}
Finnish		{(ä)+(ö)}*{(aa)+(ii)+(ee)+(uu)}+{(ää)+(öö)+(yy)}
French		{[de]+[des]+[du]+[et]+(qu)}*{(á)+(í)+(ó)+(ú)+(ã)+(õ)+(-ed) +(-y)+[di]+[do]+[e]+[è]+[in]+[na]+[no]+[o]+[una]}
German	1	{(ä)+(ö)+(ü)+(ß)}*{[das]+[dem]+[der]+[des]+[die]}
„	2	{(ch)+(sch)}*{[das]+[dem]+[der]+[des]+[die]}*{[de]+[een]+[het]+[van]}
„	3	{(ä)+(ö)+(ü)+(ß)}*[och]*{[de]+[een]+[het]+[van]}
Icelandic		{(á)+(é)+(í)+(ó)+(ú)}*{(ö)+(ü)+(ő)+(ű)}
Icelandic		{(ð)+(þ)}+{(á)+(é)+(í)+(ó)+(ú)+(ý)}*{(å)+(ö)+(æ)}
Indonesian	1	{(dj)+(nj)+(sj)+(tj)}*{$\mathbf{v}_d$–(é)}*{(-d)+(-g)+(-z)}
„	2	{[dan]+[itu]+[i]+[ke]+[oléh]+[jang]}*{$\mathbf{v}_d$- (é)}
Italian		{[che]+[da]+[del]+[di]+[è]+[in]+[ha]}*{(â)+(ê)+(î)+(ô)+(û)+(ã)+(õ)+(č) +(ć)+(ç)+(ǩ)+(ž)+(y)+(th)+(-ed)+[de]+[des]+[el]+[que]+[y]+[qui]}
Japanese		{(-ii)+(-no)+(w)+(y)}*{[ga]+[ni]+[no]+[o]+[to]+[wa]} *{(é)+($\mathbf{c}_d$)+(dj)+(nj)+(sj)+(th)+(tj)+(-y)}
Latin		{(-um)+(-us)+(-m)}*{[de]+[est]+[et]}*{$\mathbf{v}_d$+$\mathbf{c}_d$}
Norwegian	1	{(å)+(ø)+(æ)+(aa)}*{(øy)+(sjø)+(sjon)+(-het)+[å]+[av]+[ble]+[enn]+[ett] +[etter]+[meg]+[mellom]+[på]+[opp]+[seg]} *{(ä)+(ö)}*{(á)+(í)+(ó)+(ú)+(ð)+(þ)}
„	2	{[at]+[det]+[i]+[og]}*{(øy)+(sjø)+(sjon)+(-het)+[a]+[av]+[ble]+[enn] +[ett]+[etter]+[meg]+[mellom]+[på]+[opp]+[seg]} *{(ä)+(ö)}*{(á)+(í)+(ó)+(ú)+(ð)+(þ)}
Polish	1	(ą)+(ę)+(ł)+(ń)+(ś)+(ź)+(ż)
„	2	{(ć)+(cz)+(rz)+(sz)}*(w)*(y)

Portuguese	{(ã)+(ê)+(í)+(ô)+(ó)+(õ)+(ç)}*{[com]+[em]+[na]+[um]+[uma]}
Rumanian	{(ă)+(î)+(ḍ)+(ş)+(ţ)}*{[a]+[e]+[de]+[la]+[în]+[să]+[şi]}
Russian(RS)	{(ch)+(sh)+(zh)+(kh)}*{(-ii)+(-iya)+(-ogo)+(-oi)+[i]+[na]+[s]+[v]+[za]}
Russian(ISO)	{(č)+(š)+(ž)}*{(v̆d)+(ď)+(ň)+(ř)+(ť)}{*{(y)+(c')}
Slovene	{(č)+(š)+(ž)}*{[ki]+[in]+[v]+[z]}*{(v̆d)+(ď)+(ň)+(ř)+(ť)+(đ)+(ć)+[u]+(y)+(c')}
Spanish	{[a]+[de]+[en]+[la]+[que]}*{(á)+(í)+(ó)+(ú)+[el]+[y]}*{(ĕ)+(ö)+(ő)+(ü)+(č)+(ď)+(š)+(ž)+[es]+[est]+[et]}
Swedish 1	{(ä)+(ö)}*(å)
„ 2	{[för]+[inn]+[och]+[upp]}
„ 3	{(ä)+(ö)}*{[att]+[av]+[det]+[en]+[i]}
Turkish 1	{(ö)+(ü)+(ı)+(İ)}*{(ğ)+(ç)+(ş)}
„ 2	{(ı)+(Î)}*{[bir]+[ve]}
Vietnamese	(v̊)+(v̆)+(v̛)+(v̂)+(v̌)+(v̊)+(ṽ)

2.6.3 *Some results*

Tests have been made for many samples and it is revealed that:

1) Texts with 10 or more words are identified correctly at a significance level of five percent.
2) If the number of word is less than 10, identification is sometimes impossible, that is the result of identification remains indeterminate.
3) If texts are not in 25 languages in consideration, misidentification takes place naturally. For instance a Hawaiian text is misidentified as Japanese, since one of the chosen text

   "O ke a*no o* na kanaka Hawa*ii* i ka *wa* i hiki mua mai ai o..."

   has sufficient characteristics (characters and words in italic) to be judged as Japanese (see also criteria in Table 11).

2.6.4 *Computerization*

The process shown above can be performed by a computer without difficulty. The author has not yet completed the programming of the identification process but he hopes to do so in the near future.

2.6.5 *Remarks*

The present paper is intended to show an application of the statistical treatment of language elements (characters, combination of characters, and a few other characteristics). For this purpose some basic treatments of characteristics are presented with simple mathematical techniques in the Part I. The present author thinks that this kind of operations can be termed as "linguometrics," similar to such terms as biometrics and econometrics. The application of linguometrics is not limited to the identification of languages but rather the linguometrical method has a wider range of applications.

APPENDIX

Criterion with non-deterministic characteristics

As it is shown in 2.5 and 2.6 an identification criterion takes, in general, the form of

$$(C_{11}+C_{12}+C_{13}+\cdots)*(C_{21}+C_{22}+\cdots)*\overline{(C_{m1}+C_{m2}+\cdots)}*\overline{(C_{n1}+C_{n2}+\cdots)}$$

In preceding discussions, any term of the form $(C_{i1}+C_{i2}+\cdots)$ should, it is assumed implicitly, be as short as possible. If we handle this identification process by computers, the number of characteristics is not necessarily limited to a small number. In this circumstance, the establishment of criteria can be made from a different point of view. The author gives an alternative method which is useful for pattern recognition in general.

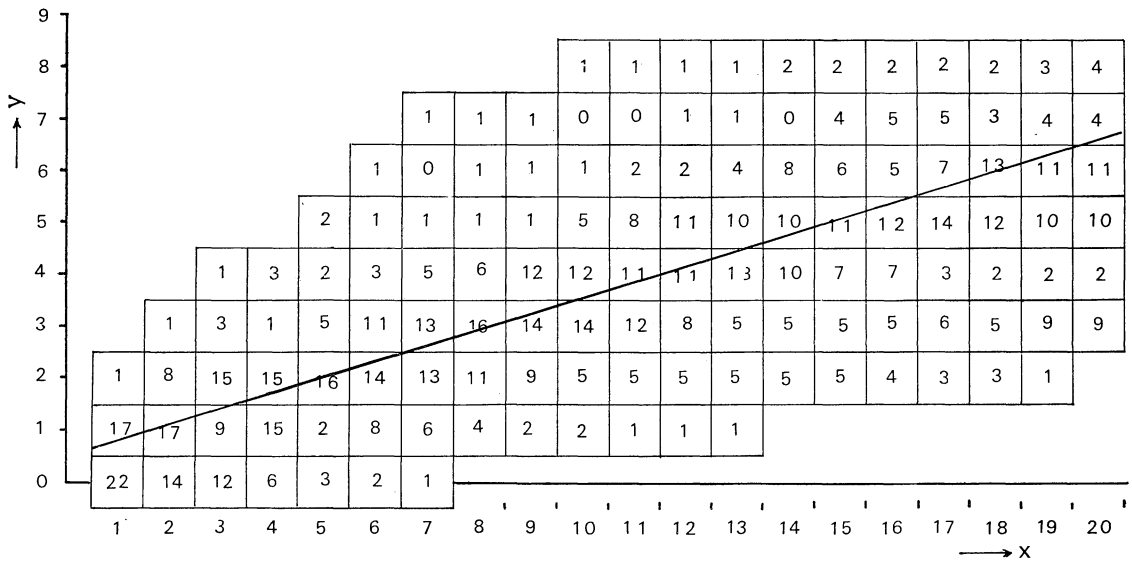

Fig. 12. Number of Characteristics *vs* Number of Words

The method proposed is described with an example for the identification of German.

In the criteria for German (D4 to D9), the simultaneous existence of two or more characteristics in any one of α, β, γ and δ is not effective for identification.

For instance, a single occurrence of [die] does not justify the identification as German nor the simultaneous existence of [die], [der] and [das]. But if we find these three words in the same sample, we can say that this sample is in German, i.e., the criterion is [die]*[der]*[das]. The existence of [die] in English, that of [des] in French, etc. make no interference. Also the criteria (ä)*(ö)*(ü), [der]*(ä)*(ö), etc. hold.

Then we can consider all the characteristics included in the α, β, γ, δ (that is, 16 characteristics in all), without group formation, and we will see whether any combination of three or more among 16 characteristics can be taken as criterion for identification. The trial was made for 40 samples of 20 words described in 2. 3. 2.

The result is shown in Fig. 12. Here the abscissa x is the number of words, the ordinate y is the number of characteristics appeared in the sample and the number in each square represents the number of samples which have y characteristics in x words. As an example, one can read in the column on the abscissa value of $x=10$, one to eight different characteristics have appeared in the beginning ten words of samples and the distribution is as given below:

Number of different characteristics (y)	0	1	2	3	4	5	6	7	8
Frequency of occurrence	0	2	5	14	12	5	1	0	1

Peaks of the distribution curve for each column are almost on a straight line $y=0.5+0.30x$. We can observe that to have three different characteristics without exception, the length of sample must be equal or greater than 20 words. The required sample length for identification by y different characteristics increases obviously with increasing number of characteristics.

If we assume that the simultaneous existence of m characteristics among 16 is the criterion for German, we have naturally some risk to misidentify English as German since four characteristics (ch), (sch), [die], [hat] can be found also in English though the CARW or

WARW of them are relatively small.

From the point of view of deterministic logic, this criterion is not permissible. However, it is permissible, if we take $m \geqslant 3$; the reason for this is given below. We assume, for sake of simplicity, that the CARW's of 16 characteristics are identical. Since we have four characteristics common to German and English, there are ${}_4C_3=4$ cases where these characteristics appear in three characteristics serving for identification. The total possible case amounts to ${}_{16}C_3=560$, then the degree of risk of misidentification is $4/560=0.71\%$ which is sufficiently low.

Since we take a sample of 10 words, we have the possibility of $33/40=82.5\%$ for having three or more characteristics according to the distribution shown above. But still have chances to have, in 10 word sample, only two or one characteristics for which the risk of misidentification is higher. The over-all risk can be calculated as

$$r=\sum_y p(y)q(y)/\sum_y p(y)$$

where $p(y)$ is the probability of having y characteristics in 10 words, and $q(y)$ is the probability of having risk of misidentification when y characteristics exist in 10 word sample.

Using the data given in the distribution show above, we get $r=2.1\%$ for $x=10$ which is tolerable considering our level of significance (5%). If we take $x=20$ and making similar calculation, we get $r=0.16\%$ which is very favorable for our new criterion.

We can, then, conclude that the simultaneous appearance of m characteristic among n characteristics chosen for a language, provided $n \gg 1$, can be taken as a statistically correct criterion even if there are, amon n characteristics, several of which are not unique in that language. It is certain that this way of statistical reasoning will facilitate the establishing of criteria for language identification.