ANDRÉE TRETIAKOFF

# RESULTS OBTAINED WITH A NEW METHOD FOR THE AUTOMATIC ANALYSIS OF SENTENCE STRUCTURES

We present in this paper a method for the automatic analysis of sentence structures.

Our purpose is to constitute a frequency dictionary of the different structures used in the language. This dictionary will enable us to select the most useful sentence structures in order to recommend their exclusive use for the writing of texts intended for automatic translation.

We think that the automatic translation will be possible only if the texts are submitted to rules which limit the complexity of their syntax. These limitations will be the less noticed by an author as only the most unusual structures would have been left out. Of course the number of permitted structures will increase as the automatic translation codes are improved.

The sentence structures are obtained by a statistical analysis of the word strings according to procedures developed in the information theory.

In the present paper we have analysed only groups of two consecutive words as an example of our method.

The same type of analysis can be generalized by considering non-consecutive words and groups of more than two words.

## 1. GROUPS

The first step of the analysis is to put the words into groups according to their grammatical properties, for example: noun, adjective, article and so on. The number of groups has been limited to keep significative frequencies with respect to the length of the corpus (3500 words). In the text under study, we have used 67 groups. A list of these groups is given in Table 3.

Of course, our classification is somewhat arbitrary as it is based on a preliminary knowledge of the language. We will show later how the results of the analysis can help us to detect inadequate classifications.

Each word of the corpus has been replaced by a symbol (two figures integer) representing its grammatical group. We consider the words inside the sentence, that is to say between two strong punctuation signs (. ; ! ?). Inside the sentence all punctuation signs are suppressed.

We will call now " words " these symbols.

## 2. DICTIONARY OF STRINGS

The second step is the constitution of a string dictionary.

A sentence containing $N$ words produces $(N-1)$ strings. For instance, the sentence *Her daughter gave me an Italian lesson every day* represented by the string " 55 04 01 44 45 05 04 85 04 ", produces the following strings:

| | | | | |
|---|---|---|---|---|
| 156 | | 55 | HER | ** |
| | 2.173 | 1 | | *** |
| 157 | | 4 | DAUGHTER | ** * |
| | 0.024 | 2 | | ***** |
| 158 | | 1 | GAVE | ** * * |
| | 1.267 | 1 | | *** * |
| 159 | | 44 | ME | ** * |
| | -0.702 | 4 | | ** |
| 160 | | 45 | AN | ** * |
| | 2.379 | 1 | | *** * |
| 161 | | 5 | ITALIAN | ** * * |
| | 1.860 | 2 | | *** * |
| 162 | | 4 | LESSON | **** * * |
| | -0.421 | 3 | | *** |
| 163 | | 85 | EVERY | ** * |
| | 2.194 | 1 | | ***** |
| 164 | | 4 | DAY | ** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| First (complete) string | 55 | 04 | 01 | 44 | 45 | 05 | 04 | 85 | 04 |
| Second string | | 04 | 01 | 44 | 45 | 05 | 04 | 85 | 04 |
| etc... | | | 01 | 44 | 45 | 05 | 04 | 85 | 04 |
| | | | | 44 | 45 | 05 | 04 | 85 | 04 |
| | | | | | 45 | 05 | 04 | 85 | 04 |
| | | | | | | 05 | 04 | 85 | 04 |
| | | | | | | | 04 | 85 | 04 |
| Last string . . . . . | | | | | | | | 85 | 04 |

Each string is obtained by suppressing the first word of the preceding string.

The dictionary emphasizes the identical strings *whatever their position* in the sentence might be. A sample of the dictionary is given in Table 1.

For example, the string 05 04 which means an adjective followed by a common noun at the end of a sentence has the rank number 244, occurs 9 times in the sentences number 9, 10, 35 and so on.

All the strings beginning by the groups 05, 04 are also listed.

## 3. SENTENCE STRUCTURE

The last step of the analysis is the production of sentence structures, using the correlations between two consecutive words.

We can compare the probability $P_j$ of a word $j$ in the corpus and the conditional probability $P_j$ (if $i$) of the same word when the preceding word is given equal to $i$. We shall call in this paper "degree of correlation" the logarithm of the ratio of the conditional probability and the probability:

$$C_{ij} = Log_2 \, P_j \text{ (if } i)/P_j$$

The degree of correlation will be positive when the probability to get a word is increased by the knowledge of the preceding word, and negative when this probability is decreased. It is a measure of the "affinity" of two consecutive words.

This procedure can be generalized by considering groups of more than two words, not necessarily consecutive.

For each sentence of the corpus we can build a structure based on the correlation between two consecutive words in the following way. Inside the sentence, consecutive words are connected two by two in order of decreasing degree of correlation. For instance in the sentence:

*She loved a good laugh*

we have the following degrees of correlation:

*She loved* = 2.56
*loved a* = 1.23
*a good* = 2.38
*good laugh* = 1.86

Therefore the first words to be connected are *She* and *loved* then *a* and *good*. We will consider that their union is the first level. Then the word *laugh* will be connected to the group *a good*. This union will be a second level and finally the two halves of the sentence are connected and this union will be the third level.

This structure can be represented by the following graph, automatically produced by the computer, and by the string 1 3 1 2 obtained by writing the sequence of the successive levels.

SENTENCE NO 9

| | | | | | |
|---|---|---|---|---|---|
| 231 | | 44 | SHE | ** | |
| | 2.564 | 1 | | ***** | |
| 232 | | 1 | LOVED | ** | * |
| | 1.232 | 3 | | | ** |
| 233 | | 45 | A | ** | * |
| | 2.379 | 1 | | *** | * |
| 234 | | 5 | GOOD | ** | * * |
| | 1.860 | 2 | | | *** |
| 235 | | 4 | LAUGH | **** | |

Degrees of correlation:

| | |
|---|---|
| *She loved* | 2.56 |
| *loved a* | 1.23 |
| *a good* | 2.38 |
| *good laugh* | 1.86 |

String of groups: 44 01 45 05 04
String of levels: 1 3 1 2

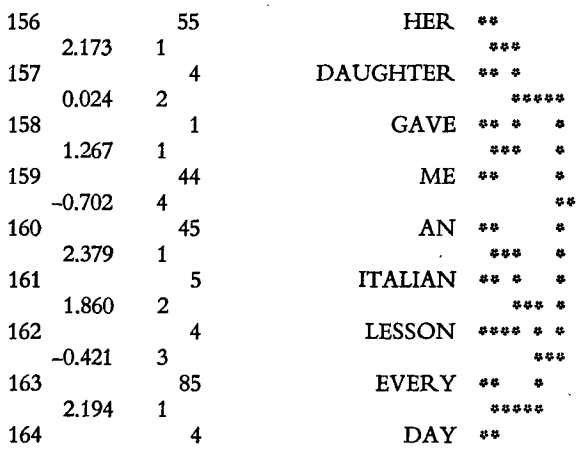
## 4. DICTIONARY OF STRUCTURES

This procedure has been applied for all the sentences of the text, producing strings of numbers which represent the structure of these sentences.

For each string of numbers, by suppressing the highest number we obtain 2 strings representing 2 substructures of this sentence. We carry on this procedure till the string has only 1 number, that is to say represents the structure of a group of 2 words.

For instance the structure of the sentence *Her daughter gave me an Italian lesson every day* is represented by the following string of numbers:

SENTENCE NO 5



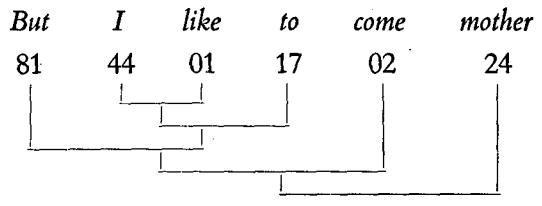

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Complete string: | 1 | 2 | 1 | 4 | 1 | 2 | 3 | 1 | (level 4) |
| 1 substring : | | | | | 1 | 2 | 3 | 1 | (level 3) |
| 2 substrings: | 1 | 2 | 1 | | 1 | 2 | | | (level 2) |
| 3 substrings: | 1 | | 1 | | 1 | | | 1 | (level 1) |

All the structures and substructures are classified in a dictionary, giving their frequencies and the positions of the sentences containing the corresponding word strings (Table 2).

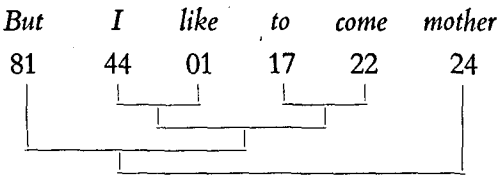
For example, the structure 1 4 2 1 3 has the rank number 65 and is found 5 times in the sentences number 12 16 21 24 41.

## 5. CLASSIFICATION ERRORS

If the structure of a sentence is unsatisfactory, this can be due to an error in the classification of a word of this sentence. This observation is used to detect and correct classification errors. For example in the sentence:



the word *come* had been classified in a wrong group 02 (indicative of intransitive verbs). When corrected (22 = infinitive of intr. verbs) we obtain the following structure:




Another way to check the classification of words into groups is to use the quantity of information associated to the law of succession of two consecutive words. It is known from communication theory that the average amount of information by word is reduced when we know the law of succession of two consecutive words. This reduction is precisely equal to the average degree of correlation of all the groups:

$$C = \sum_{ij} P_{ij} C_{ij}$$

We shall call it quantity of information associated to the law of succession of two consecutive words.

In order to check the validity of the choice of the grammatical group for a word, the quantity of information associated to the law of succession of the groups is measured. Then, changing the choice of the group, the quantity of information is measured again for this new classification. The greater the quantity of information associated to a law of succession of the groups, the better the distribution of these words into these groups.

## 6. CONCLUSION

The sample chosen here (a novel by S. Maugham of 3500 words) is too short to obtain significant frequencies for the different structures.

This sample contains 200 sentences of an average length of 17 words.

In spite of the simplicity of the method of analysis employed, 72 sentences of an average length of 10 words have been correctly analysed.

This shows that the correlation of 2 consecutive words, although insufficient, will play an important part in the more elaborated methods of analysis that we are now developing.

TABLE 1.

```
244   9   5  4
              9  10  35  37  41  49  54  66  72
245   1   5  4  1  17  22  97  5  4
              35
246   1   5  4  1  35  4
              30
247   1   5  4  1  35  4  7  55  4  85  4
              40
248   1   5  4  1  44  95  4  7  74
               1
249   1   5  4  3  17  23  35  4  97  45  5  4  97  35  4
              16
250   1   5  4  9  44  3  41
              68
251   1   5  4  16  24  9  44  1  54
              14
252   1   5  4  80  95  4  16  55  4  44  2  26
               2
253   1   5  4  85  4
               5
254   2   5  4  97  24
              13     42
255   1   5  4  97  35  4
              16
256   1   5  4  97  65  5  4  1  35  4  7  55  4  85  4
              40
257   1   5  5  4
              41
258   1   5  5  4  1  35  4
              30
259   1   5  17  21  45  4
               17
260   1   5  17  21  66  85  4
               50
261   1   5  17  22  16  35  4  7  45  5  4
               72
262   1   5  17  22  26
               44
263   1   5  27  44
               63
264   1   5  27  55  4
```

TABLE 2.

| *Rank* | *Frequency* | |
|---|---|---|
| 56 | 1 | 0 1 3 2 1 4 5 1 2 1 6 1 2 3 1 |
| | | 22 |
| 57 | 1 | 0 1 3 2 1 6 4 3 1 2 5 1 3 2 1 |
| | | 61 |
| 58 | 1 | 0 1 4 1 2 1 3 |
| | | 33 ..... sentence number |
| 59 | 1 | 0 1 4 1 2 1 3 1 2 |
| | | 48 |
| 60 | 1 | 0 1 4 1 2 1 3 5 1 |
| | | 33 |
| 61 | 1 | 0 1 4 1 2 1 3 5 1 6 3 1 2 4 |
| | | 33 |
| 62 | 1 | 0 1 4 1 2 3 |
| | | 15 |
| 63 | 1 | 0 1 4 1 2 3 1 |
| | | 63 |
| 64 | 1 | 0 1 4 1 2 3 5 1 2 3 4 2 1 3 1 |
| | | 15 |
| 65 | 5 | 0 1 4 2 1 3 |
| | | 12 16 21 24 41 |
| 66 | 2 | 0 1 4 3 1 2 |
| | | 20 42 |
| 67 | 1 | 0 1 4 3 1 2 5 1 2 |
| | | 20 |
| 68 | 1 | 0 1 5 1 2 1 3 2 1 4 |
| | | 18 |
| 69 | 1 | 0 1 5 1 2 1 3 2 1 4 6 |
| | | 18 |
| 70 | 1 | 0 1 5 1 2 1 3 2 1 4 6 7 1 |
| | | 18 |
| 71 | 1 | 0 1 5 1 4 2 1 3 |
| | | 24 |
| 72 | 1 | 0 1 5 1 4 2 1 3 6 1 2 3 4 |
| | | 24 |
| 73 | 1 | 0 1 5 2 1 3 4 |
| | | 28 |
| 74 | 1 | 0 1 5 2 1 3 4 6 1 |
| | | 28 |
| 75 | 1 | 0 1 6 1 2 5 4 3 1 2 |
| | | 52 |
| 76 | 1 | 0 1 6 1 2 5 4 3 1 2 7 1 |

SENTENCE NO 60

| | | | | |
|---|---|---|---|---|
| 3348 | | 65 | THOSE | ** |
| | 1.421 | 1 | | ***** |
| 3349 | | 4 | THINGS | ** * |
| | 0.268 | 3 | | ** |
| 3350 | | 3 | ARE | ** * |
| | 1.873 | 1 | | *** * |
| 3351 | | 36 | NOT | ** * * |
| | 1.804 | 2 | | *** |
| 3352 | | 31 | DONE | **** |

SENTENCE NO 56

| | | | | |
|---|---|---|---|---|
| 3255 | | 29 | WHAT | ** |
| | 4.345 | 1 | | *** |
| 3256 | | 3 | IS | ** * |
| | 1.121 | 2 | | *** |
| 3257 | | 5 | WRONG | **** * |
| | -0.095 | 3 | | ** |
| 3258 | | 27 | WITH | **** * |
| | 2.026 | 2 | | *** |
| 3259 | | 55 | HIS | ** * |
| | 2.173 | 1 | | *** |
| 3260 | | 4 | MORALS | ** |

SENTENCE NO 58

| | | | | |
|---|---|---|---|---|
| 3323 | | 54 | THAT | ** |
| | 2.306 | 1 | | *** |
| 3324 | | 2 | SOUNDS | ** * |
| | 1.570 | 2 | | *** |
| 3325 | | 66 | QUITE | ** * * |
| | 2.689 | 1 | | *** * |
| 3326 | | 5 | IMPOSSIBLE | ** * |
| | -2.392 | 3 | | ** |
| 3327 | | 1 | CRIED | ** * |
| | 1.663 | 1 | | ***** |
| 3328 | | 24 | KITTY | ** |

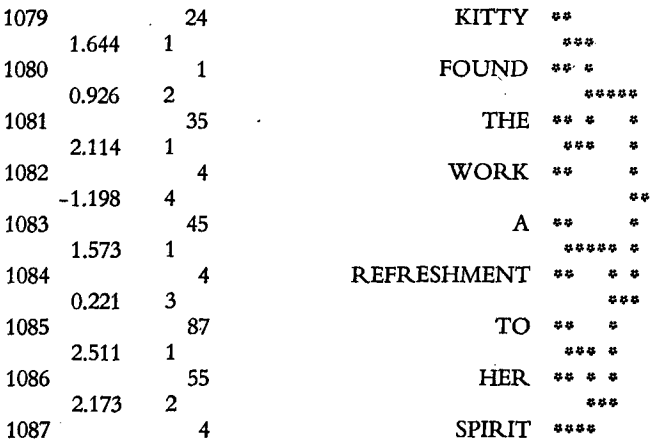
SENTENCE NO 36

| | | | | |
|---|---|---|---|---|
| 1839 | | 44 | SHE | ** |
| | 2.564 | 1 | | *** |
| 1840 | | 1 | FORCED | ** * |
| | 1.859 | 2 | | *** |
| 1841 | | 55 | HER | ** * * |
| | 2.173 | 1 | | *** * |
| 1842 | | 4 | LIPS | ** * |
| | 1.012 | 3 | | ** |
| 1843 | | 7 | INTO | **** * |
| | 1.180 | 2 | | *** |
| 1844 | | 45 | A | ** * |
| | 1.573 | 1 | | *** |
| 1845 | | 4 | SMILE | ** |

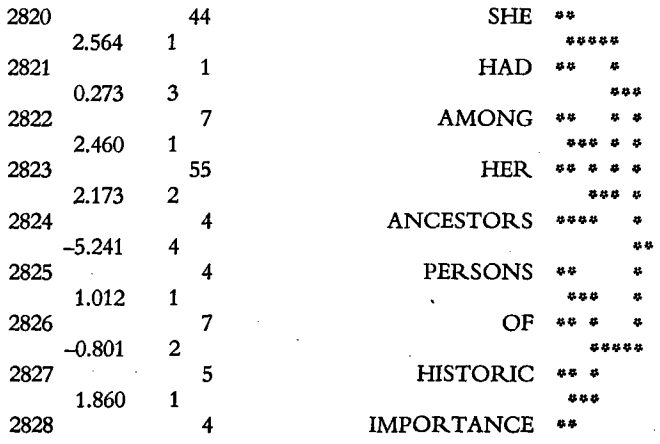
SENTENCE NO 20

| | | | | |
|---|---|---|---|---|
| 1121 | | 35 | THE | ** |
| | 2.114 | 1 | | ******* |
| 1122 | | 4 | MOTHER | ** * |
| | 0.024 | 4 | | *** |
| 1123 | | 1 | GAVE | ****** * * |
| | 0.273 | 3 | | *** * |
| 1124 | | 7 | INTO | ** * * |
| | 2.460 | 1 | | *** * * |
| 1125 | | 55 | HER | ** * * * |
| | 2.173 | 2 | | *** * |
| 1126 | | 4 | CARE | **** * |
| | -1.604 | 5 | | ** |
| 1127 | | 35 | THE | ** * |
| | 3.863 | 1 | | *** * |
| 1128 | | 25 | SMALLER | ** * * |
| | 2.087 | 2 | | ******* |
| 1129 | | 4 | CHILDREN | **** |

SENTENCE NO 19




SENTENCE NO 49



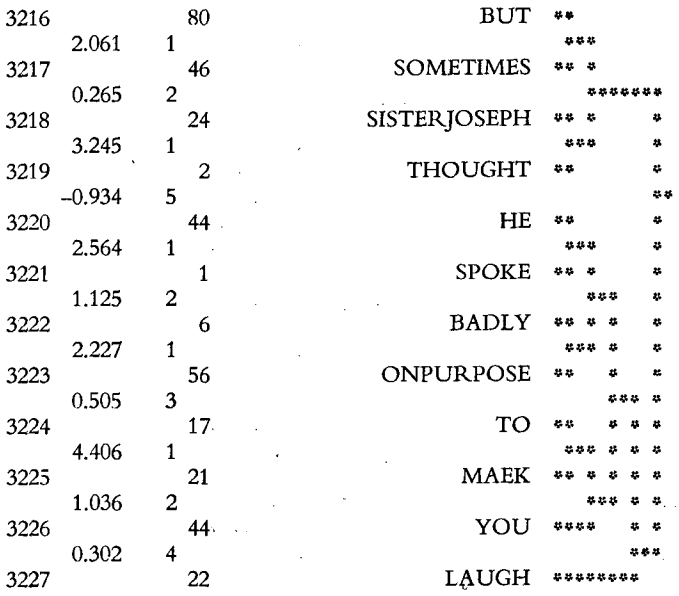
SENTENCE NO 50

| | | | | |
|---|---|---|---|---|
| 2830 | | 94 | IT | ******** |
| | 2.047 | 4 | | *** |
| 2831 | | 68 | MUST | **** * * |
| | 5.493 | 2 | | *** * * |
| 2832 | | 28 | HAVE | ** * * * * |
| | 7.456 | 1 | | *** * * * |
| 2833 | | 33 | BEEN | ** * * * |
| | 2.742 | 3 | | *** * |
| 2834 | | 5 | HARD | ****** * |
| | 0.490 | 5 | | ** |
| 2835 | | 17 | TO | ** * |
| | 4.406 | 1 | | ***** * |
| 2836 | | 21 | LEAVE | ** * * |
| | 0.540 | 3 | | ***** |
| 2837 | | 66 | SO | **** * |
| | 0.870 | 2 | | *** |
| 2838 | | 85 | MUCH | ** * |
| | 2.194 | 1 | | *** |
| 2839 | | 4 | GRANDEUR | ** |

SENTENCE NO 52

| | | | | |
|---|---|---|---|---|
| 2850 | | 44 | SHE | ** |
| | 2.564 | 1 | | ********** |
| 2851 | | 1 | HAS | ** * |
| | 0.926 | 6 | | *** |
| 2852 | | 35 | THE | ** * * |
| | 3.863 | 1 | | *** * * |
| 2853 | | 25 | MOSTBEAUTIFUL | ** * * * |
| | 2.087 | 2 | | ******* * * |
| 2854 | | 4 | HANDS | **** * * * |
| | 1.192 | 5 | | *** * |
| 2855 | | 9 | THAT | ******** * * |
| | 2.438 | 4 | | *** * |
| 2856 | | 44 | I | ****** * * |
| | 2.727 | 3 | | *** * |
| 2857 | | 8 | HAVE | ** * * |
| | 3.709 | 1 | | *** * * |
| 2858 | | 46 | EVER | ** * * * |
| | 2.804 | 2 | | *** * |
| 2859 | | 31 | SEEN | **** * |
| | -0.655 | 7 | | ** |
| 2860 | | 1 | SAID | ** * |
| | 1.663 | 1 | | ********** |
| 2861 | | 24 | KITTY | ** |

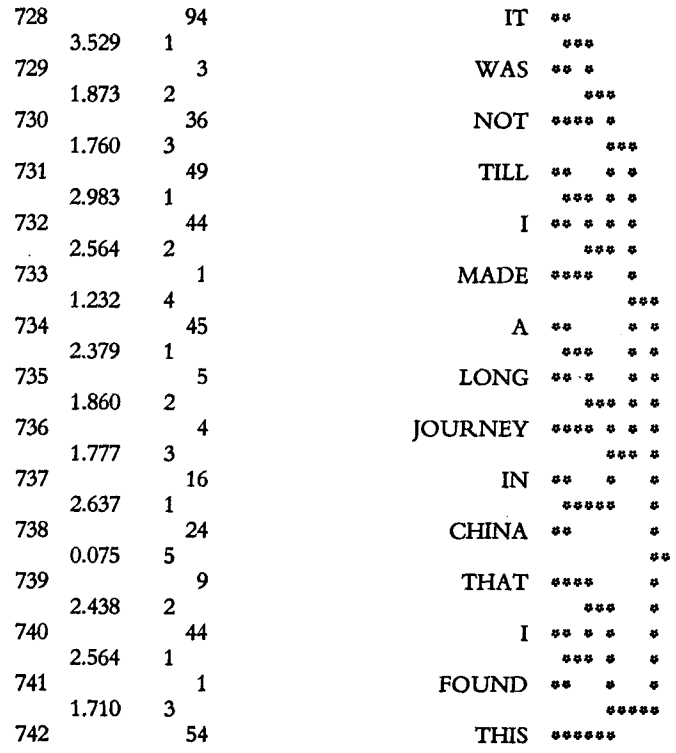
SENTENCE NO 55



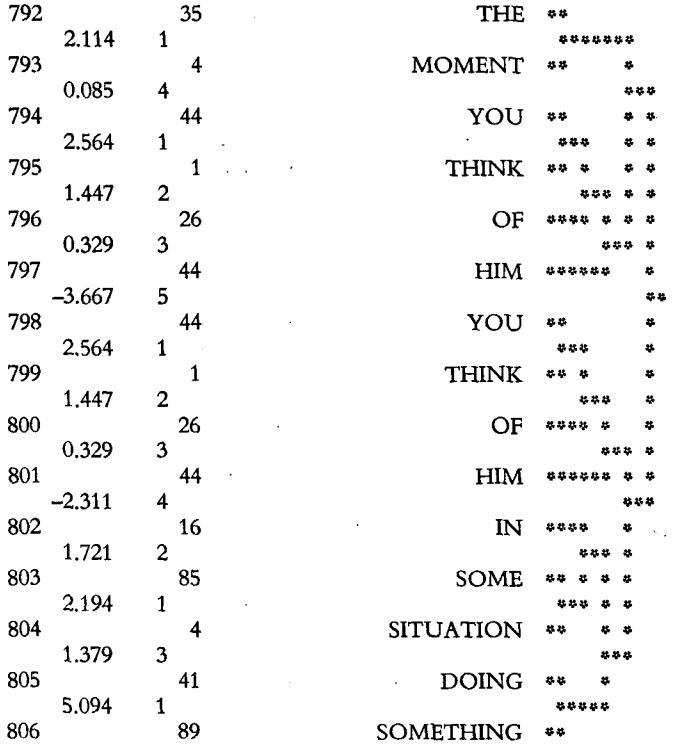
SENTENCE NO 32

| | | | | |
|---|---|---|---|---|
| 1734 | | 7 | BY | ** |
| | 2.122 | 1 | | *** |
| 1735 | | 35 | THE | ** * |
| | 2.114 | 2 | | *** |
| 1736 | | 4 | MERCY | ** * * |
| | 2.288 | 1 | | *** * |
| 1737 | | 97 | OF | ** * |
| | 0.113 | 3 | | *** |
| 1738 | | 4 | PROVIDENCE | ****** * |
| | 0.085 | 4 | | *** |
| 1739 | | 44 | I | ** * * |
| | 2.115 | 1 | | ***** * * |
| 1740 | | 3 | WAS | ** * * * |
| | 1.047 | 3 | | *** * |
| 1741 | | 7 | AT | ** * * |
| | 2.122 | 1 | | *** * * |
| 1742 | | 35 | THE | ** * * * |
| | 2.114 | 2 | | *** * |
| 1743 | | 4 | DOOR | **** * |
| | -1.074 | 5 | | ** |
| 1744 | | 66 | JUST | ****** * |
| | 1.696 | 3 | | ***** |
| 1745 | | 49 | AS | ** * |
| | 2.983 | 1 | | *** * |
| 1746 | | 44 | SHE | ** * * |
| | 2.212 | 2 | | *** |
| 1747 | | 2 | CAME | **** |

SENTENCE NO 14



SENTENCE NO 15



SENTENCE NO 22

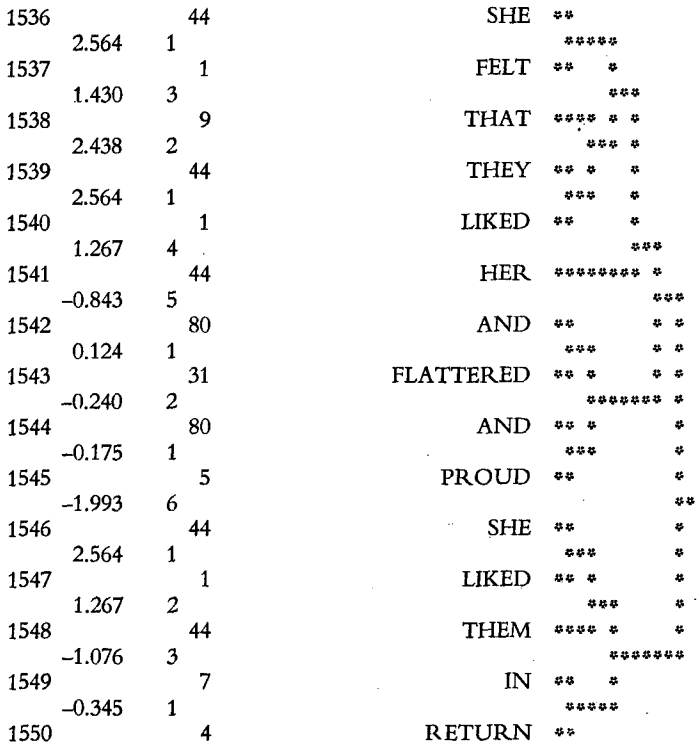


TABLE 3.

| Group | |
|---|---|
| 01 | INDICATIVE (TRANSITIVE VERBS) |
| 21 | INFINITIVE (TRANSITIVE VERBS) |
| 31 | PAST PARTICIPLE (TRANSITIVE VERBS) |
| 41 | PRESENT PARTICIPLE (TRANSITIVE VERBS) |
| 51 | GERUND (TRANSITIVE VERBS) |
| 02 | INDICATIVE (INTRANSITIVE VERBS) |
| 22 | INFINITIVE (INTRANSITIVE VERBS) |
| 32 | PAST PARTICIPLE (INTRANSITIVE VERBS) |
| 42 | PRESENT PARTICIPLE (INTRANSITIVE VERBS) |
| 52 | GERUND (INTRANSITIVE VERBS) |
| 03 | INDICATIVE (STATE VERBS) |
| 23 | INFINITIVE (STATE VERBS) |
| 33 | PAST PARTICIPLE (STATE VERBS) |
| 43 | PRESENT PARTICIPLE (STATE VERBS) |
| 53 | GERUND (STATE VERBS) |
| 08 | INDICATIVE (AUXILIARY VERBS) |
| 28 | INFINITIVE (AUXILIARY VERB) |
| 68 | WOULD, SHOULD, WILL, CAN, MAY, HAVE TO |
| 78 | INDICATIVE (TO DO, AUXILIARY VERB) |
| 88 | INFINITIVE (TO DO, AUXILIARY VERB) |
| 04 | COMMON NOUN |
| 14 | COMMON NOUN (POSSESSIVE CASE) |
| 24 | PROPER NOUN |
| 34 | PROPER NOUN (POSSESSIVE CASE) |
| 44 | PRONOUN (PERSONAL) |
| 54 | PRONOUN (DEMONSTRATIVE) |
| 64 | PRONOUN (INDEFINITE) |
| 74 | PRONOUN (PERSONAL REFLEXIVE) |
| 94 | PRONOUN (IMPERSONAL) |
| 05 | ADJECTIVE (QUALIFICATIVE) |
| 15 | ADJECTIVE (COMPARATIVE) |
| 25 | ADJECTIVE (SUPERLATIVE) |
| 35 | ARTICLE (DEFINITE) |
| 45 | ARTICLE (INDEFINITE) |
| 55 | ADJECTIVE (POSSESSIVE) |
| 65 | ADJECTIVE (DEMONSTRATIVE) |
| 85 | ADJECTIVE (INDEFINITE) |
| 95 | ADJECTIVE (CARDINAL) |
| 72 | ADJECTIVE (PRESENT PARTICIPLE) |
| 73 | ADJECTIVE (PAST PARTICIPLE) |
| 93 | ADJECTIVE (ORDINAL) |
| 39 | ADJECTIVE (INTERROGATIVE) |

```
06   ADVERBS MADE FROM ADJECTIVES
16   ADVERB (PLACE)
26   POSTPOSITION
36   NOT
46   ADVERB (TIME)
56   ADVERB (MANNER)
66   ADVERB (QUANTITY)
76   AS, LIKE
86   ADVERB (REPETITION)
96   ADVERB (EXCLAMATIVE)

07   PREPOSITION
17   TO (INFINITIVE)
27   WITH, WITHOUT
87   TO
97   OF
89   NOTHING, SOMETHING
09   THAT
19   PRONOUN (RELATIVE)
29   PRONOUN (INTERROGATIVE)
49   CONJUNCTION (TIME)
59   CONJUNCTION (CAUSE)
69   CONJUNCTION (SUPPOSITION)
79   CONJUNCTION (COMPARISON)

80   AND
81   BUT, OR
/*
```