[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 7: THE DICTIONARY

AUTOMATIC AFFIX INTERPRETATION AND RELIABILITY OF THE HARVARD AUTOMATIC DICTIONARY[1]

Murray E. Sherry
Harvard University

1. Affix Interpretation

In the terminal phase of lookup in any stem dictionary, it is necessary to associate the stem and affix of each word, in order to determine the case, number, and gender of nouns and adjectives, and the person, number, gender, tense, mood, and voice of verbs. Figure 1 shows sample output for six morphological word types in the Harvard Automatic Dictionary: noun, adjective, verb, pronoun, numeral, and undeclinable.

The data in columns 1-5 and 8-9 consist of information stored in the dictionary entries. For each word, noun-, adjective-, and verb-analyzer programs associate the affix, class marker (column 2), and occasionally some of the grammatical symbols in columns 5 and 8, to determine the information that is then stored in columns 6 and 7.

The case and number of nouns and adjectives is entered into column 6 where a character position is reserved for each case and number combination (Table 1). The case coding is mnemonic (except "C" for dative) and the machine word is divided into two sections to express number, the first six characters representing the singular and the last six, the plural.

The gender is inserted into column 7 in the character positions corresponding to the related information on case and number (Table 2). The unused characters are filled with dashes for ease of reading. Potential multiple usage is indicated by the presence of more than one identifying character in columns 6 and 7.

The markings for verbs necessarily differ from those for nouns and adjectives (Table 3). The first six character positions are reserved for person and number which are indicated in the present and future tenses by an appropriate character in any one of the first

[1] This work has been supported in part by the National Science Foundation and the Rome Air Development Center, Air Research and Development Command.

	Columns								
Word Type	1	2	3	4	5	6	7	8	9
Noun	SENSI-TIVITY	N06.00	CHUVSTVI-TEL'NOST-I	00A-0039	NDIIFI00ΔΔΔΔ	-G-C-FN-A---	-F-F-FF-F---		213910000000
Adj.	CARRYING OUT	A04.00	VYPOLNJA-JUSHCH-EGO	00A-0733	AD0100ΔΔΔ4ΔΔ	-GA---------	-BM---------		035190000000
Verb	PRODUCE	V04.00	PROI%VOD-ITSJA	00A-0752	VNΔOP3000000	--T---BADRΔΔ		BIB4B5	162240000000
Pron.	WHICH	P01.00	CHT-O	00A-1015	FNCIΔSTRIΔOΔ	N-A---------	N-N---------		213847500000
Num.	ONE	D01.00	ODN-OJ	00A-1340	DBXRFSJPKΔΔΔ	-G-CIP------	-F-FFF------	-T-TTT------	124420555554
Unde.	WITH	I01.00	S-	00A-0804	RΔΔΔΔΔΔΔΔΔΔΔ	-GA-I--GA-I-		GIAROOBA1111	178910000000

Sample Dictionary Output for Six Morphological Word Types

Figure 1

Character	1:	N -	if nominative singular
Character	2:	G -	if genitive singular
Character	3:	A -	if accusative singular
Character	4:	C -	if dative singular
Character	5:	I -	if instrumental singular
Character	6:	P -	if prepositional singular
Character	7:	N -	if nominative plural
Character	8:	G -	if genitive plural
Character	9:	A -	if accusative plural
Character	10:	C -	if dative plural
Character	11:	I -	if instrumental plural
Character	12:	P -	if prepositional plural

Format of Column 6 of Dictionary Output with
Information on Case and Number for Noun
and Adjective Morphological Types
TABLE 1

M -	masculine
F -	feminine
N -	neuter
B -	masculine or neuter (adjective types only)
A -	masculine, feminine, or neuter

Allowable Characters in Column 7 of Dictionary
Output for Gender of Noun and Adjective
Morphological Types
TABLE 2

Session 7: THE DICTIONARY

Characters 1-6

Option A: (present and future tenses)

V in character position 1 = 1st person singular
Z in character position 2 = 2nd person singular
T in character position 3 = 3rd person singular
V in character position 4 = 1st person plural
Z in character position 5 = 2nd person plural
T in character position 6 = 3rd person plural

Option B: (past tense)

SSS in character position 1-3 = 1st, 2nd, or 3rd person singular
PPP in character position 4-6 = 1st, 2nd, or 3rd person plural

Characters 7-12

7: A = past (tense)
B = present
C = future
X = present or future

8: M = masculine (gender)
F = feminine
N = neuter
A = any

9: D = indicative (mood)
E = imperative
F = infinitive
G = gerund

10: R = reflexive (voice)
O = non-reflexive

11: X = Special situation among some verbs with affix ите which can be both 2nd person plural indicative and plural imperative.

12: not used

Format of Column 6 of Dictionary Output with Information on Person, Number, Tense, Gender, Mood, and Voice for Verb Morphological Types

TABLE 3

six character positions. Since the person cannot be determined from the morphological characteristics for verbs in the past tense, either all of the first three or all of the second three character positions are filled to designate number. For all verbs, the tense is given in character position 7, the gender in character position 8, the mood in character position 9, and the voice in character position 10.

If the affix cannot be associated with the stem, the corresponding dictionary entry is rejected as incompatible. If every dictionary entry with the same stem as a given word is incompatible, the text word is labeled with "INCOMPAT X" in column 6. In a similar manner, "INCOMPAT Z" is marked in column 6 whenever a word is found that has not been classified into one of the normal classes.

The case, number, and gender of pronouns and numerals are coded into the dictionary entries, once and for all, during dictionary compilation since pronouns and numerals are non-productive classes. The details of the coding of all the non-productive classes have been reported previously [1 , 2, 3] .

Since our work is essentially experimental, the layout of the dictionary output has been designed for flexibility and ease of reading. The maximization of operating speed and efficiency has been deliberately deferred until the system is ready for production operation.

A complete description of affix interpretation and of the method of operation of the analyzer programs is in preparation.

2. Reliability and Accuracy of the Harvard Automatic Dictionary

The output of "Frequency Runs" [4] is used as test material to determine the reliability and accuracy of the Harvard Automatic Dictionary. A list is kept containing every distinct inflected form in every text in our tape library together with its frequency of occurrence. The latest test, Frequency Run V, processed in January 1960, was based on 104, 097 words of text consisting of 14, 698 distinct inflected forms.

In addition to the main list of analyzed dictionary entries referred to by the distinct inflected forms (including all homographs), three supplementary lists were produced: a list of all the incompatible entries; a list of all homograph sets remaining after the incompatible entries have been removed, each homograph set consisting of two or more compatible dictionary entries; and a list of problem sets, each

consisting of one or more incompatible dictionary entries with no compatible entries. Every homograph set and every problem set consists of entries referred to by a single distinct inflected form.

All three supplementary lists were studied carefully to detect errors in the system. The list of incompatible entries indicated only one error in the analyzer programs which was also detected by the problem set listing. The information gleaned from the other two lists is summarized in Tables 4 and 5.

Table 4 indicates that only 2.4% of the distinct inflected forms refer to more than one entry in the dictionary, i. e. , to a homograph set. Of these homograph sets, almost half are essential [5] . Another 23% are due to short form adjectives whose existence is questionable. Since there is at present no reliable source of information on this subject, these short forms have been left in the dictionary. The remaining 30% are due to various types of errors in the dictionary and in the analyzer programs that can be easily corrected. This table also shows the same data as referred to actual text occurrences. Only 9% of the homograph sets, referred to by only 0.4% of all text occurrences, are obtained because of errors, while the overwhelming majority (84%) of homographs that occur are essential. An interesting and noteworthy property of Russian is that almost 5% of the text occurrences refer to homograph sets.

Table 5 displays the small number (1.0%) of distinct inflected forms that result in problem sets after dictionary lookup. The occurrence of the problem sets among words in texts is even rarer (0.3%). Less than 0.1% of the words in texts refer to problem sets due to errors in the dictionary. Now that these errors have been detected, it is a simple matter to institute the corrections.

The results of Frequency Run V indicate that, although scattered errors remain in the Harvard Automatic Dictionary and the Continuous Dictionary Run, the system is more accurate and reliable than any other known lookup scheme of comparable size.

	Distinct Inflected Forms		Text Occurrences	
Essential homographs	165	46%	4214	84%
Short form adjectives	83	23	360	7
Duplicates in dictionary	61	17	254	5
Reflexive coding errors	37	10	144	3
Other dictionary errors	7	2	12	
Words not in classes	4	1	6	1
Analyzer errors	1	-	15	
	358	99%	5005	100%
358 out of 14, 698 distinct inflected forms (2. 4%) 5,005 out of 104,097 words of text (4. 8%)				

Summary of Homograph List, Frequency List V,
January 1960
TABLE 4

	Distinct Inflected Forms		Text Occurrences	
Words missing from dictionary.	62	41%	203	56%
Typographical errors	56	37	61	17
Dictionary errors	23	15	80	22
Hyphenated words	7	5	11	3
Analyzer errors	2	1	9	2
	150	99%	364	100%
150 out of 14, 698 distinct inflected forms (1. 0%) 364 out of 104, 097 words of text (0. 3%)				

Summary of Problem Sets, Frequency Run V,
January 1960
TABLE 5

REFERENCES

[1] Matejka, L., "Grammatical Coding for Pronouns", Mathematical Linguistics and Automatic Translation, Report No. NSF-3, Section XVIII, Harvard Computation Laboratory, (August 1959).

[2] Matejka, L., "Grammatical Coding for Prepositions", Mathematical Linguistics and Automatic Translation, Report No. NSF-3, Section VI, Harvard Computation Laboratory (August 1959).

[3] Magassy, K., "Grammatical Coding for Numerals", (to be published).

[4] von Susich, S., "Frequency Runs - A System for Lexical Quality Control and Statistical Analysis", Mathematical Linguistics and Automatic Translation, Report No. NSF-3, Section VIII, Harvard Computation Laboratory (1959).

[5] Oettinger, A. G., "A Study for the Design of an Automatic Dictionary", Doctoral Thesis, Harvard University (1954).