Project DOC: Its Methodological Basis[1]

William S-Y. Wang

I.

Dictionary on Computer, hereafter DOC, is part of an overall effort to harness an on-line computer for phonological research. For certain problems the linguist finds it necessary to organize large amounts of data, or to perform rather involved logical tasks -- such as checking out a body of rules with intricate ordering relations. In these situations a computer can be invaluable in that it forces the linguist to think through his problems with great precision and in that it can do certain jobs with a speed and accuracy not otherwise possible.

The overt aim of DOC is to reconstruct the phonological histories of the major Chinese dialects. At a deeper level our interest is to find out more about how phonological structures change in general and the relation between these changes and the synchronic systems they lead to. To achieve these objectives we must attempt to account for oceans of data (the regular and irregular developments of thousands of morphemes in dozens of dialects). The hypotheses we posit, i.e., the reconstructed forms and the associated rules, are likewise numerous and complex. The project is further complicated by the tens of thousands of logographs which must be contended with, especially when we involve the rime dictionaries and the rime tables. These considerations lead naturally to the use of a computer in our work.[2]

To construct significant hypotheses and to generalize them beyond the confines of our data pool, to grasp the theoretical import of each discovery, clearly these are creative tasks that cannot be mechanized. Nonetheless a well-hewn tool, such as we hope to develop DOC into, can contribute substantially to facilitate these creative tasks.

Of the many language families in the world, Chinese offers an ideal laboratory within which to study phonological change for many reasons. Chief among these are two: (1) its unrivaled wealth of materials, and (2) its distinctive phonology and orthography.

(1) The earliest extant materials date back to ca. 1500 B.C. in the form of oracle inscriptions. We have a virtual time depth of some three and one-half millenia of literature. This literature includes not only such works as rime tables and rime dictionaries, but also extensive contributions from a tradition of philological scholarship that arose in the early Song and reached considerable sophistication in the Qing period. Indeed the view has been expressed that in China the methods of scientific reasoning were primarily developed in the hands

of the Qing philologists (as opposed to Europe where they originate in the physical sciences). Few language groups compare with Chinese with respect to this immense treasure of literature to work with.

(2) By far the greatest bulk of the present knowledge of phonological change comes from investigations of Indo-European languages. It is not unlikely, then, that the present theories and methods are skewed in the direction of characteristics found in these languages. Studying a language family with a very different structure will help us balance this skewed perspective. Indeed Meillet must have had something like this in mind back in 1913 when, in discussing the comparative method, he wrote:

"Les rapprochements reçoivent des confirmations utiles quand on peut constater que des concordances grammaticales s'ajoutent à la concordance du son and du sens....Les langues qui, comme les langues indo-européenes...ont des particularités grammaticales attachées à certain mots se prêtent donc mieux à la démonstration de l'étymologie que les langue où tous les mots se conforment aux même règles grammaticales. La difficulté qu'on éprouve à poser la grammaire comparée de certaines langues, notamment en Extrême-Orient, vient en partie de là." (p. 32).

A language of the Chinese type is distinct in that (a) it has no inflectional paradigms and no morphophonemic alternations to speak of, (b) it has a very simple syllabic structure, (c) it has tones, and (d) its orthography is logographic. These characteristics all have implications for research on phonological change.

(a) Current views of diachronic phonology invariably emphasize the importance of paradigmatic analogy as one of the two major forces of phonological change (the other major force being phonetic). The formalization of analogy may be in terms of proportionality in a structuralist framework, or in terms of rule simplification within the context of generative phonology. It would be of considerable theoretical interest to examine these views with respect to Chinese, which has virtually no paradigms. In particular we would want to investigate what are the mechanisms whereby a change diffuses lexically[3] in Chinese, where word classes are not related by morphophonemic alternations. An understanding of these mechanisms is crucial toward answering the question of whether phonological change is or is not phonetically actuated.

(b) The simple syllabic structure of the morphemes and the even accentual structure of the sentences are also of special interest. Whereas many recurrent types of change outside of Chinese involve the reduction of consonant clusters into geminates, or the breaking up of clusters by vowel epenthesis, or the reduction of syllabic elements due to stress shifts, these changes hardly occur at all in Chinese. The pervasive themes found in phonological structures like Chinese are palatalization, dento-labialization, reduction of post-vocalic obstruents, and various complex interplays between the segmental syllable and tones.

Research on phonological change now suffers from a severe lack of a systematic catalog of carefully documented changes. Given that X has the reflex Y, we need to know if this change was induced within the system or if it was actuated by another linguistic system; did X go directly into Y or were there intermediate phonemic stages; if each direct change was abrupt or gradual, phonetically and lexically. Only when a sufficient fund of such information is available can one successfully meet the challenge of phonetic and other types of explanations, and only then can phonology make the exciting transition from a descriptive effort into an explanatory science. DOC is designed to facilitate the gathering of this fund of information.

The tones of Chinese have intrigued students of language for many years. They are of interest to phonological theory because they form a relatively self-contained subsystem in the sound structure that can serve as a relatively independent testing ground for the theory. The Chinese have had a categorical (though not physical) understanding of the tones of their language for well over 1500 years. During this period although the morpheme membership of the tonal categories has been relatively stable, the physical manifestations of the tones have undergone considerable changes. Some of these changes, it appears, are intricately connected to segmental features. The investigation of these changes can contribute much toward our understanding of the inter-relationships between phonation and articulation.

Lastly, the logographic system of writing has certain unique implications. Since the logographs are much more distantly related to the sounds of the language than are the alphabets of the European languages, one can assume that they have exerted very little influence on the developments of the various sound systems. In other words, we have fewer cases of historical confusion due to spelling pronunciation to contend with. By the same token the logographs themselves have an amazing longevity, so that we can make many inferences about their phonetics for as far back as three thousand years ago.

In sum, then, DOC is being developed as a powerful too l that will give phonological research a speed and precision not otherwise attainable. Its creative use can lead us to a deeper understanding of phonological structure and change on a quantitative basis. At present this tool is being developed within the context of Chinese, for the reasons outlined in the foregoing paragraphs. We expect that the methods we will have worked out will be largely applicable to the study of the phonology of any language group. Indeed, it is to be hoped that a field like

Indo-European may one day be subjected to the rigors of this tool, and its results validated on a quantitative and objective basis.

II.

At present the primary source of data is the Hanyu Fangyin Zihui.[4] The 17 dialects reported in the Zihui are now available on Linc tape. Outside of the Zihui, we have the complete Kan-on, Go-on, Sino-Korean and portions of the Zhongyuan Yinyun.

For a variety of reasons, such as ease of tape-punching, ease of proof-reading and error-correction, and ease of writing of utility programs, the data are stored in several formats. These formats are related to each other by a set of supporting programs, as shown in Figure 1. The rectangles indicate data formats and the circles indicate supporting programs.

The first stage in the data collection is the punching of paper tape on the teletype. A standard entry requires 24 punches:

1.	space
2-5.	telegraphic code (G)
7.	dialect identification (D)
8-9.	tone (T)
10-13.	initial (I)
14-15.	medial (M)
16-20.	nucleus (N)
21.	ending (E)
22.	literary (L)
23.	carriage return (R̸)
24.	line feed

Punches 23 and 24 are discarded by the supporting program RDPT (Read Paper Tape). After using RDPT the resulting Dialect tape should have the structure illustrated in Figure 2.

For proof-reading and error correcting, the Dialect tapes may be converted into File tapes in the format of LAP 6 D, as shown in Figure 3. In this format each entry has 20 characters (or half-words) which is the maximum number that can be displayed per line on the scope; each entry is further followed by a CASE (Linc code 23) that is disregarded by FILEDOC and a R̸ (Linc code 12) that shifts the display to the next line. So each entry on the Linc tape is still 22 characters or 11 words long, even though only 20 characters are displayed. The spaces (Linc code 14) are converted into periods (Linc code 20) for ease of reading.

LAP 6 D is modified from LAP 6 in two ways. The left margin is moved 4 positions to the left so that each entry will exactly fit one line. More space is allotted for files on both the systems tape and the file tape. According to my present understanding, each dialect tape is just about the size that a single file can accommodate.

The uses of LAP 6 D files are obvious. We can use the full set of meta commands for such files as well as the editing conveniences.

The AC tape contains the Qiè-Yùn information for the logographs as these are recorded in the Zihui, as shown in Figure 4. The use of this tape makes it possible to add this information to any dialect tape by matching the telegraphic codes via the ACCODE program. The resultant AC-Dialect tape has the structure also shown in Figure 4. Notice that positions 17 through 32 correspond to 7 through 22 in entry structure of Dialect tape illustrated in Figure 2. As shown in (D) in Figure 4, the number of dialect forms for each entry can be easily increased.

Finally, it will be useful for certain problems to have the result in the form of a set of logographs. At present our computer can only give us the telegraphic code of the characters. With LOGOTAB we hope to be able to display the logographs on the scope and/or print them out by means of a special purpose computer. The 16 x 16 matrix representations of several thousand logographs have already been designed by Susumu Kuno's group at the Harvard Computation Laboratory, cf. Hayashi, et al., 1968. LOGOTAB will be essentially a table look-up program that will translate telegraphic codes into those matrix representations. We are also giving thought to a similar logograph input device as that used by the Harvard group.

III.

Although the ideas for DOC were first conceived in 1966, it is only in late spring 1969 that the project began to be operational. Several linguistic programs have been written for it, especially with respect to the seven Mandarin dialects. In Figure 5 we see a correlation program that quantifies the development of the Ancient Chinese tones into each of the Mandarin dialects. The points of greatest interest are of course with the cells which show the small number of exceptional developments. Are they due to borrowing from other dialects, residue from changes that have not yet completed their course, or are they due to the inception of new changes yet to be systematized?

As the data pool becomes richer and richer with the addition of each new dialect or rime dictionary, it became increasingly obvious that our little laboratory Linc would not be able to cope with all the problems efficiently. Since the beginning of the summer, Tom McGuire of the Phonology Laboratory has helped us establish a remote terminal that connects to the CDC 6400 in the University Computer Center. Some of our materials have already been converted into magnetic tape that is compatible with that computer. An example of the new format of DOC is shown in Figure 6.

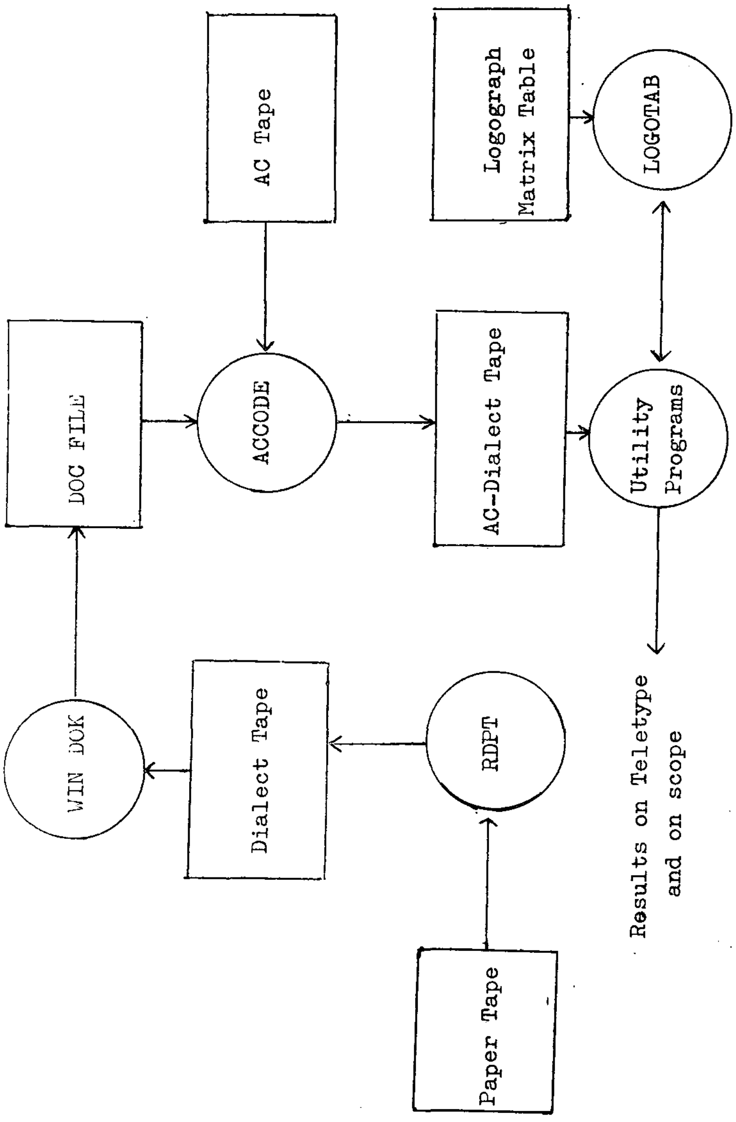

Figure 1: Organization of DOC (a first approximation).

(A) Entry Structure

Each entry has 22 half-words, as follows:

#	G	G	G	G	G	D	T	T	I	I	I	I	M	M	N	N	N	N	N	E	L
1	2 - 6					7	8	9	10-13				14	15	16-20					21	22

(B) Tape Structure

Each block on the Linc tape contains 23 entries, with the last three words filled by 5555.

BN	Address	Content
000	000-012	Entry 1
	013-025	Entry 2
	026-040	Entry 3
	.	.
	.	.
	.	.
	362-374	Entry 23
	375-377	5555

Figure 2: Dialect Tape.

(A) CANTON: structure of a single file (23 centuries per block)

BN	Address	Content
000	000	2065
	001	5712
	002-014	Entry 1
	015-027	Entry 2
	⋮	⋮
	364-376	Entry 23
	377	5555
001	⋮	
	000	5555
	001	5555
n	002-014	Entry 1
	015-027	Entry 2
	⋮	⋮
	364-376	last entry
	377	7777

(B) Entry Structure

G	G	G	G	G	T	T	I	I	I	I	M	M	N	N	N	N	N	E	L	Case	K̸
1 - 5					6	7	8-11				12	13	14-18					19	20	21	22

Figure 3: DOC File.

(A) AC Tape Entry Structure:

G	G	G	G	she	h/k	D	T	Rime	Initial
1 - 5				6-7	8	9	10	11-12	13-16

(B) AC Tape: each entry takes 20_8 words

000	000-017	Entry 1
	020-037	Entry 2
	•	•
	•	•
	•	•

(C) AC-Dialect Tape: entry structure

G	G	G	G	G	AC info	D	T	T	I	I	I	I	M	M	N	N	N	N	N	E	L
1 - 5					6-16	17	18	19	20-23				24	25	26-30					31	32

(D) AC-Dialect Tape:

000	000-007	Telecode and AC	Entry 1
	010-017	D-1	
	020-027	D-2	
	∘	∘	
	∘	∘	
	∘	∘	
		D-n	

Figure 4: AC Tape and AC-Dialect Tape

Distribution of AC tones - 1

	DF AC	1a	1b	2	3	4a	4b	total
Peking	I	491	454	9	3	0	0	957
	II	9	6	376	98	0	0	489
	III	7	4	10	600	0	0	621
	IV	91	81	28	142	0	0	342
	Total	598	545	423	843	0	0	2409
Jì-nán	I	485	437	10	15	0	0	947
	II	21	11	356	102	0	0	490
	III	14	4	10	592	0	0	620
	IV	175	86	19	94	0	0	374
	Total	695	538	395	803	0	0	2431
Xī-ān	I	479	447	10	10	0	0	946
	II	12	7	366	101	0	0	486
	III	14	6	22	574	0	0	618
	IV	253	87	11	19	0	0	370
	Total -	758	547	409	704	0	0	2420

Figure 5A

Distribution of AC tones - 2

	DF AC	1a	1b	2	3	4a	4b	total
Taì-yúan	I	921	10	14	9	0	0	954
	II	23	0	368	100	0	0	491
	III	13	0	11	595	2	2	623
	IV	6	3	3	11	344	71	438
	Total	963	13	396	715	346	73	2506
Hàn-Kǒu	I	489	437	12	11	0	0	949
	II	11	8	372	102	0	0	493
	III	14	10	12	587	0	0	623
	IV	11	349	2	12	0	0	374
	Total	525	804	398	712	0	0	2439
Chéng-dū	I	485	441	9	11	0	0	946
	II	10	8	370	99	0	0	487
	III	18	7	15	582	0	0	622
	IV	7	362	1	12	0	0	382
	Total	520	818	395	704	0	0	2437
Yáng-shōu	I	484	449	12	9	0	0	954
	II	10	6	373	103	0	0	492
	III	11	3	14	595	2	0	625
	IV	5	1	1	8	361	0	376
	Total	520	459	400	715	363	0	2447

Figure 5B

```
  TAI-YUAN        1  K     U  AI
  HAN-KOU         1  K     U  AI
  CHENG-DU        1  K     U  AI
  YANG-ZHOU       1  K     U  E2

1042A        -    4  V     F3 WM 28
  PEKING          1B F        A
  JI-NAN          1B F        A
  XI-AN           1B F        A
  TAI-YUAN        4B F        A      Q
  HAN-KOU         1B F        A
  CHENG-DU        1B F        A
  YANG-ZHOU       4  F        A      Q

1042B        -    1  UZP   K3 CG 5
  PEKING          1B TSRH     E3     V
  JI-NAN          1B TSRH     E3     V
  XI-AN           1B TSRH     E3     V
  TAI-YUAN        1  TS H     F3     V
  HAN-KOU         1B TS H     E3     N
  CHENG-DU        1B S        F3     N
  YANG-ZHOU       1B TS H     F3     N

1044         -    4      H K3 NA 19
  PEKING          2           I
  JI-NAN          1           I
  XI-AN           1           I
  TAI-YUAN        4        I  E3
  HAN-KOU         1B          I
  CHENG-DU        1B          I
  YANG-ZHOU       4           I

1046         -    2  K     K3 NU 5
  PEKING          2  TCP   I  CU
  JI-NAN          2  TCP   I  OU
  XI-AN           2  TCP   I  CU
  TAI-YUAN        2  TCP   I  CU
  HAN-KOU         2  TCP   I  CU
  CHENG-DU        2  TCP   I  E3U
  YANG-ZHOU       2  TCP   I  C1U3

1048         -    2        K3 WC 2
  PEKING          2        I  E
  JI-NAN          2        I  E
  XI-AN           2        I  E
  TAI-YUAN        2        I  F2
  HAN-KOU         2        I  E
  CHENG-DU        2        I  F
  YANG-ZHOU       2           I1

1050         -    2  NJ    F3 NC 8
  PEKING          2  ZR       U
  JI-NAN          2  L        U
  XI-AN           2  V        U
  TAI-YUAN        2  Z        U
  HAN-KOU         2           Y
```

Figure 6A

```
  JI-NAN              18           L   A       V
  XI-AN               18  V            A       V
  TAI-YUAN            1   V            A2    Z
  HAN-KOU             18           L   A       V
  CHENG-DU            18           U   A       V L
  YANG-ZHOU           18           L   A1      V

0074            -     1   K        K2  WU  4
  PEKING              1   TCP      I   AU
  JI-NAN              1   TCP      I   C2
  XI-AN               1   TCP      I   AU
  TAI-YUAN            1   TCP      I   AU
  HAN-KOU       1     1   TCP      I   AU          L
  HAN-KOU       2     1   K            AU
  CHENG-DU            1   TCP      I   AU
  YANG-ZHOU           1   TCP      I   C2

0076            -     4            K3  XG  12
  PEKING              3                I
  JI-NAN              3                I
  XI-AN               1                I
  TAI-YUAN            4            I   E3      Q
  HAN-KOU             18               I
  CHENG-DU            18               I
  YANG-ZHOU           4            I   E3      Q

0079            -     1   K        K3  XG  1
  PEKING              1   TCP          I       V
  JI-NAN              1   TCP          I       V
  XI-AN               1   TCP          I       V
  TAI-YUAN            1   TCP          I       V
  HAN-KOU             1   TCP          I       N
  CHENG-DU            1   TCP          I       N
  YANG-ZHOU           1   TCP          I     Z

0080            -     1   C        K4  XG  13
  PEKING              18  T    H       I       V
  JI-NAN              18  T    H       I       V
  XI-AN               18  T    H       I       V
  TAI-YUAN            1   T    H       I       V
  HAN-KOU             18  T    H       I       N
  CHENG-DU            18  T    H       I       N
  YANG-ZHOU           18  T    H       I     Z

0081            -     3   L        K3  YG  7
  PEKING              3   L        I   A       V
  JI-NAN              3   L        I   A       V
  XI-AN               3   L        I   A       V
  TAI-YUAN            3   L        I   A2    Z
  HAN-KOU             3   N        I   A       V
  CHENG-DU            3   N        I   A       V
  YANG-ZHOU           3   L        I   A1      V

0086            -     1   NJ       K3  NN  11
  PEKING              18  ZR           E3      N
  JI-NAN              18  ZR           E     Z
```

Figure 6B

[1] The work reported here is supported in part by grants from the National Science Foundation and the American Council of Learned Societies.

[2] See Lyovin (1968) for a more detailed description of the beginnings of Project DOC.

[3] The hypothesis of lexical diffusion, i.e., phonological change operates gradually across the lexicon is admittedly controversial. The hypothesis would not be acceptable to theorists in the Neogrammarian tradition, e.g., L. Bloomfield. However, as I argue in Wang (1969), there are good reasons for thinking that this is indeed how changes are implemented within narrow time spans, i.e., morpheme by morpheme rather than phoneme by phoneme. The proof of the hypothesis requires large scale studies of the sort exemplified by DOC.

[4] The Zihui has many draw-backs, as pointed out in Lyovin's review (1969). However, it is obviously the best set of core materials to start the project on.

References

Dong, Tong-he. 1953. Zhongguo Yuyinshi. (History of Chinese Phonetics). Taiwan.

Dougherty, Ching-yi, Sidney Lamb and Samuel Martin. 1963. Chinese character indexes. 5 vols. Berkeley: University of California Press.

Hayashi, Hideyuki, Sheila Duncan and Susumu Kuno. 1968. Graphical input/output of nonstandard characters. Communications of the Association for Computing Machinery. 11.9.613-8.

Lyovin, Anatole. 1968. A Chinese dialect dictionary on computer: progress report. POLA 7. Berkeley.

______________. 1969. Review of Hanyu Fangyin Zihui. Language 45.3.

Meillet, Antoine. 1913. Sur la méthode de la grammaire comparée. Reprinted in his Linguistique Historique et Linguistique Générale. (Paris, 1965).

Peking University. 1962. Hanyu Fangyin Zihui. (Phonetic Dictionary of Chinese Dialects). Peking.

Wang, William S-Y. 1967. Phonological features of tone. International Journal of American Linguistics. 33.93-105.

________________. 1968. The many uses of F_o. POLA 8. Berkeley.

________________. 1969. Competing changes as a cause of residue. Language 45:1.9-25.

Wang, William S-Y. and Anatole Lyovin. 1969. Chinese Linguistics Bibliography on Computer. In press with Cambridge University Press.