front matter at the end
please see bookmarks

GIORGIO BUCCELLATI

THE OLD BABYLONIAN LINGUISTIC ANALYSIS PROJECT: GOALS, PROCEDURES AND FIRST RESULTS

PREFACE

The Old Babylonian Linguistic Analysis Project (hereafter OBLAP) has been supported since 1972 by a grant from the National Endowment for the Humanities, Washington, D.C. Before then it had grown slowly over a period of three and a half years in the form of a pilot project sponsored by the Academic Senate of the University of California, Los Angeles, through its Committee on Research.

It consists of a data bank of Akkadian cuneiform texts with complete grammatical analysis, a system of programs for automatic grammatical encoding and for various types of analysis, and various series of publications embodying the results of such analysis.

The present paper is the first published report about the Project (though oral presentations were made at the 17th Rencontre Assyriologique Internationale in Bruxelles on July 4th 1969 and at the 183rd Meeting of the American Oriental Society in Washington, D.C., on March 22nd 1973). It aims at describing the major phases of operation, the types of procedure chosen, the goals as presently implemented.

Primary goal of the Project has remained, throughout, the analysis of the textual material; the scope of programming was intentionally subordinated to the goal of obtaining concrete results in an efficient and economic way. As a result, our systems are wholly oriented toward practical applications. In line with this concern, the present description of the Project will appear at about the same time as the first four volumes of our series go to press (MLA I, GAC I, see below, section 8).

1. DATA BASE

Some 70 years ago so eminent a Semitist as THEODOR NÖLDEKE prefaced his first volume of studies in comparative Semitics by admitting his own inability to incorporate Akkadian within the remainder of the Semitic languages studied by him. Concern for method and perhaps a slight touch of irony are equally present in his remarks: on the one hand he says that he declines to deal with a language for which he could only refer to grammars and dictionaries (i.e. without mastery of the texts themselves); on the other, he expresses his suspicion that even some of the basic points of the language still defied true understanding. (It seems worthwhile to quote his words in full:

> Als grossem Mangel empfinde ich es, dass ich vom Babylonisch-Assyrischen fast ganz absehen musste. Und ich hätte ja mir so leicht durch Ausbeutung der Grammatik und des Wörterbuches von Delitzsch sowie einige anderer bequemer Hilfsmittel das Ansehen geben können, als wenn ich von der Sprache etwas verstünde! ... Ich kann übrigens noch immer den Verdacht nicht los werden, dass die wirkliche Lautgestalt der babylonisch-assyrischen Inschriften im einzelnen nicht so genau ermittelt ist wie ihr Sinn.
>
> *Beiträge zur Semitischen Sprachwissenschaft*, Strasbourg, 1904, p. V)

Fortunately changes have intervened in the meantime, and no Semitist today would allow himself to be as candid as Nöldeke was at the beginning of the century. True, the change may not be a complete reversal of the situation: as late as 1961 W. VON SODEN could write that some Semitists and even Assyriologists still seem to think of Akkadian as a language without grammar (*Akkadisch*, in G. LEVI DELLA VIDA (ed.), *Linguistica Semitica: Presente e Futuro*, Rome, 1961, p. 34). Yet it goes without saying that our grammatical understanding of the language is by now firmly rooted, and that the time has come for specialized and sophisticated research, especially in the area of dialects.

The fundamental importance of Akkadian for linguistic research needs hardly to be emphasized. In the first place, there is a large amount of textual material which awaits proper and full analysis. While no statistical data are available, a rough estimate may be based on a partial word count, by considering that the Chicago Assyrian Dictionary, which includes substantially all the relevant and distinctive passages in which a word occurs, has now reached a total of 3,731 quarto pages for

a total of 6,572 words corresponding to somewhat more than 2/5 the total Akkadian lexicon. This allows a projection of about 17,000 words for the total Akkadian lexicon, as compared, for instance, to about 7,000 for the lexicon of Biblical Hebrew. This material, both lexical and textual, is steadily increasing due to the work on unpublished texts stored in museums and the discovery of new texts in archaeological excavations.

The quantity of the material is only paralleled by its uniqueness in terms of cultural and linguistic history. The texts cover all aspects of life, and in the most minute details, for a period of well over two millennia. Linguistically, Akkadian is by far the oldest attested Semitic language, its earliest texts dating to around 2350 B.C., which makes them about 1,000 years older than the texts of any other Semitic language. (Linguistic evidence for Amorite, the second oldest Semitic language, is only a few centuries later than the earliest evidence for Akkadian, but is limited to personal names, without any texts.) More important, perhaps, is the fact that Akkadian represents a linguistic sub-family by itself, known as East Semitic, and is therefore of decisive importance in reconstructing the linguistic history of the Semitic family, or even the Afro-Asiatic group as a whole. If the effect of the study of Akkadian has not yet been fully felt in the field of Semitics, it is in part because much of the textual material is not readily available to non-Assyriologists, and in part because serious and detailed linguistic analysis has been lagging behind.

Purpose of OBLAP is to answer these needs by preparing a thorough linguistic analysis of Old Babylonian letters. The choice of Old Babylonian was suggested by the fact that this is traditionally regarded as the classical dialect of Akkadian, and that we have a considerable and representative amount of material. The further restriction to epistolary material is due to the fact that this genre comes the closest to spoken language, and thus is the richest in internal dialectal differentiation. It is paradoxical that, even though all general grammars of Akkadian are primarily based on Old Babylonian, no exhaustive grammar of Old Babylonian as such exists, exception made for peripheral Old Babylonian dialects. OBLAP is aimed at providing the groundwork for such an Old Babylonian grammar, and will provide an almost inexhaustible tool for sophisticated research on specialized topics of historical and linguistic analysis.

As with all dead languages for which electronic data retrieval has been made possible, there is the special added advantage of providing

the best possible substitute for the linguist's ideal (i. e. living!) informant. Naturally the validity of this assumption is directly proportional to the quality of both encoding and programming. The more differentiated and flexible these are, the more information one can derive with regard to specific questions on the part of the researcher. Even though simple lack of attestation would not preclude the possibility of the existence of a supposed form (a fact which, precisely, only a living informant could verify), yet one will be able at least to draw some conclusions within the realm of probability depending on the nature and size of the sample analyzed.

Another one of the linguist's concerns vis-à-vis dead languages is the role of statistics – an approach which is of course the most natural outgrowth of a computerized data base. Besides serving as a guide for such phenomena as the frequency (and hence productivity) of given features, statistics is often the only guide for deciding even basic questions of linguistic analysis, particularly as they relate to the writing system which, in case of Akkadian, is often elusive and still imperfectly understood. Considering that, with dead languages, writing is the only link possible between a linguist and the data, the most thorough investigation of the graphemic system is an indispensable precondition for linguistic understanding – particularly if the individual graphemes, as in the case of Akkadian, are highly polyvalent in both a phonological and logographic sense. Such investigation is precisely one of OBLAP's goals, on a scale never attempted before and, in fact, unimaginable without the aid of the computer.

The philological difficulties, deriving only in part from the writing system, make it difficult to prepare an adequate input for OBLAP when texts are published in cuneiform only or in obsolete editions; even the good editions (which may be estimated as covering less than half the total data base) present considerable problems of standardization and normalization, from phonological to lexical features. In addition, there is a considerable number of known unpublished Old Babylonian letters, some of which are being published for the first time within the framework of OBLAP (possibly around 10 % of the entire data base). As a result, the task of editing the texts for keypunching is more complex than it might seem at first. Naturally this means that the input is relatively limited from a quantitative point of view, when compared with data banks of other languages. But the fact that we strive for qualitative uniformity for the entire corpus on a par with the latest and best editions compensates for the lesser size of our corpus.

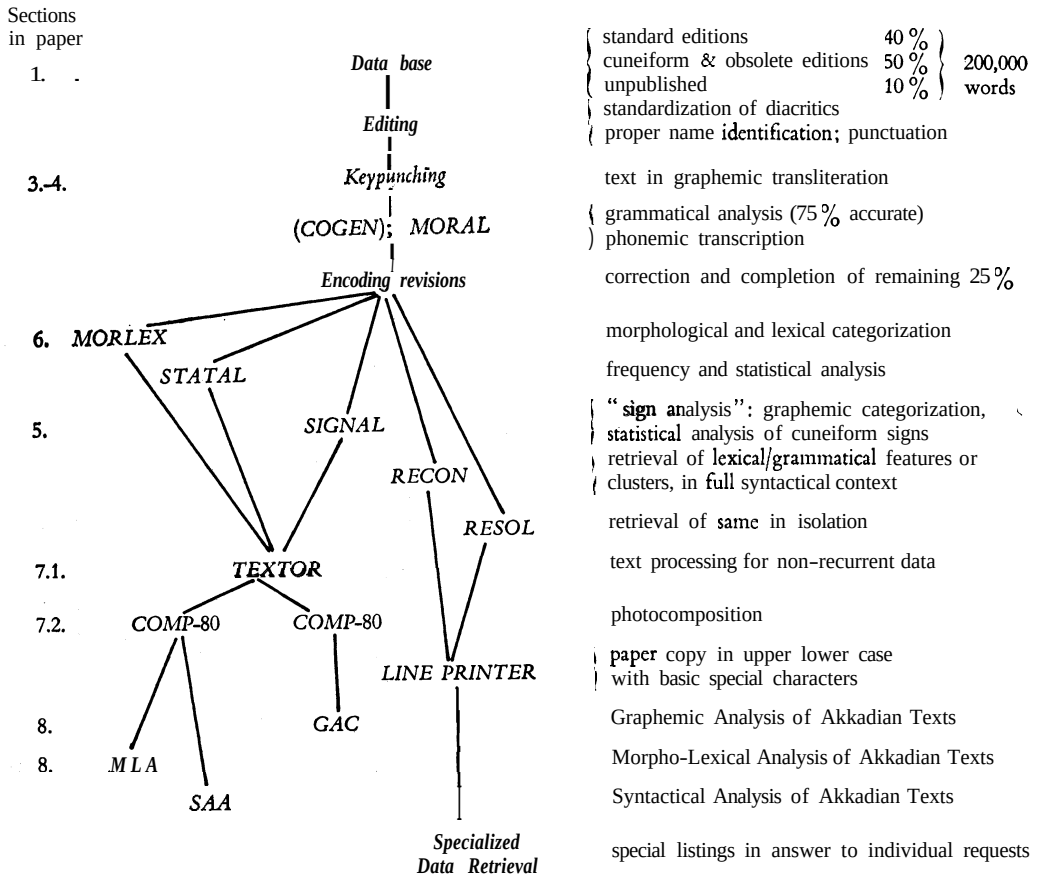


Fig. 1 OBLAP *Flow-Chart*

Up to the present, approximately 100,000 words have been keypunched, in various stages of encoding. This includes five volumes published in the series *Altbabylonische Briefe*, Leiden, 1964, and eight volumes in the series *Archives Royales de Mari*, Paris, 1950, plus various other scattered publications. The corpus presently available is thus about half the size of the total data base, which is estimated at about 200,000 words.

Naturally the system can be extended beyond the confines of Old Babylonian letters to any other genre or dialect of Akkadian. A recent agreement with the Istituto di Studi Micenei ed Egeo-Anatolici of Rome, Italy, and the Sezione Linguistica of CNUCE, Pisa, Italy, has paved the way for precisely such an expansion: as indicated more in detail in a separate report by Dr. Claudio Saporetti, encoding and programming systems elaborated by OBLAP will be made available for the analysis of Assyrian texts, as a joint project between OBLAP and the Italian Institutes. Implementation of the agreement is underway and the first results will be available to scholars within the near future.

2. OUTLINE OF PROCEDURES

OBLAP is a complex system of procedures ranging from editing and encoding to data retrieval, analysis and publication. The enclosed flow chart outlines the various stages of the system. Most of these will be discussed individually in the pertinent sections of the paper as indicated on the left margin of the flow chart (see fig. 1).

3. ENCODING: INTERNAL SYNTAX

I will not go here into detail in describing the encoding system or the linguistic analysis of Akkadian on which it rests; the encoding manual contains 28 pages of codes which are obviously not within the purview of this Conference. A simple description of the basic outline will be sufficient to provide an insight into the internal syntax on which the programs operate.

In the initial keypunching words are entered in continuous succession in a graphemic transliteration which renders very closely the cuneiform text, e.g.:

(1)

```
                                                                    73-80
←————————————————————————— 1-72 ———————————————————————————→ ←——→
#1 A-NA BE-LIZIA # QI2-BI2-MA # UM-MA   1_D-*UTU+HA-ZI-IR-MA     00001
```

Spaces and hyphens represent boundaries between cuneiform signs. Symbols added to the transliteration found in good text editions are minimal; in fact, apart from the standardization of diacritics and a plus sign to divide elements of personal names, the only true addition consists of the digit 1 with a break character before personal names. (The number sign stands for line division, and the asterisk for capital letters, both of which features are indicated in a standard edition.)

When the encoding program is applied (MORAL, see below, section 4), the text is divided so that only one word appears on each card; in compound words of various type (especially proper names, suffixes and enclitics) the individual elements are isolated on separate cards.

The textual part appears on the right half of the card (cols. 41–73), arranged in free format. The graphemic transliteration given in the input is now provided automatically with a phonemic transcription – which is not an easy task because the relationship between the two is not isomorphic; for instance vocalic length is normally not expressed in the writing (e.g. the writing BE-LI2-IA represents phonemic/be:li:ya/), and logograms alternate with phonograms (e.g. *DINGIR-*UTU stands for /Šamaš/).

Columns 1–39 are reserved for grammatical encoding, in fixed column format. The encoding system is exhaustive for morphology (cols. 10–35), and nearly exhaustive for syntax (cols. 36–39, plus occasional columns in the morphology section).

Columns 1–4 are reserved for footnotes, and 5–9 for line and word order sequence.

Columns 40 is reserved for multiple interpretations or multiple readings; e.g. the word +HA-ZI-IR in our sample text is susceptible of two interpretations, hence it is repeated a second time, with a dash in column 40; this can be repeated as often as desirable, thus allowing for an unlimited number of multiple interpretations. Naturally in a connected text the word only appears once.

Columns 74–80 are reserved for identification purposes, with a different number for each card or record within a given corpus.

A completely encoded sample is given below, using the same text given in (1) as sample for the initial input. About 75 % of the additional information in (2) with respect to (1) is provided by a program (MORAL,

see below), while the remaining 25 %, especially the syntactical codes, are entered by encoders:

footnotes	*Reference*	*Morphology*	*Syntax*	*Text*	*Sequence*
	1– 1	X DT ANA		A-NA/ANA	00001
*	1– 2	NCGSM *B:L EWEPE:S-	HS	BE-LI2/BE:LI:/+	00002
	1– 2	PSG1SM	GOJ	+IA/YA/	00003
	2– 3	VB-L-S2M*QB:IW-		QI2-BI2/QIBI:/+	00004
	2– 3	E" MA		+MA/MA/	00005
	3– 4	A "- UMMA		UM-MA/UMMA/	00006
	3– 5PDD		S	*DINGIR+	00007
	5– 5PDNA SM	-&AM&- PARS-	S	/+*UTU/&AMA&/+	00008
	3– 5P NP	>HA:ZIR- A		+HA-ZI-IR/HA:ZIR/+	00009
	3– 5P PSN3SM			-+HA-ZI-IR/HA:ZIR/+	00010
	3– 5 E"	MA		+MA/MA/:	00011

4. ENCODING: PROGRAMMING (BY JOHN L. SETTLES)

MORAL (MORpho-lexical AnaLysis) processes transliterated texts punched in a continuous stream on cards. The only diacritics required serve to identify proper names, word boundaries within proper names and line changes in the original text. In the input a hyphen is used to separate signs within a word, while a space separates words.

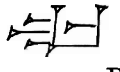
MORAL scans the input text and isolates the transliteration representing one-word. The transliteration is reduced to an " internal " transcription– internal because no attempt is made to resolve long vowels or consonants where they are not evident in the transliteration. For example, LI-IZ-ZI-ZU is reduced to *lizzizu*; the final *u* may be long and the final *z* may be ~~short~~, but for the purposes of internal transcription they are short. Internal transcription is obtained by removing all hyphens, break indications, and other punctuation from the transliteration. The resulting string is then scanned and identical vowels in juxtaposition are assimilated. Such predictable anomalies of the writing system as I-DIN-NAM for *iddinam* are handled as special cases.

long

The internal transcription is first used as a search argument against a table of invariable forms, which contains prepositions, adverbs, coordinating and subordinating particles, pronouns. If a match is ~~not~~ found, the analysis is retrieved from the table. If a match is not found, the internal

transcription is analyzed to determine if it is a noun or a verb. This is done by reference to two additional tables which contain the verbal and nominal prefixes and suffixes. By use of these tables, MORAL separates the strings into all possible combinations of prefix-base-suffix. The isolation of the prefix and suffix furnishes some of the morphological analysis – in the case of verbs, the indications of person, number and mood. In cases where more than one set of compatible prefix-suffix combinations are isolated, multiple analyses are produced (which are then resolved manually on the basis of syntactical considerations). The isolated base is converted to a string representing its consonantal and vocalic form, and is searched in still another table which identifies the internal inflection. The radicals are then isolated from the base and are used to obtain lexically determined data (root vowel, pattern vocalism, etc.). All of the tables contain indications as to compatible and incompatible entries in other tables; for example, each prefix in the prefix table has an indication as to the suffixes which can occur with it and those which may not. The lexical table contains data as to the attested forms for each root or base. It is updated by a separate program which reads the analyzed texts after they have been reviewed and corrected manually.

5. GRAPHEMIC CATEGORIZATION (JOHN L. SETTLES)

The polyvalence of signs in a cuneiform system makes a concordance of the signs by shape extremely useful. For example the sign  is read KA, ZU_2, DUG_4, INIM, etc., depending on the context. By grouping all occurrences of the sign, without regard to reading, it is possible to compare and contrast the various environments in which the sign is attested.

Such a concordance is the primary output of SIGNAL, the SIGN Analysis function of OBLAP. In addition frequency distributions are produced by sign, reading and homophonous value.

SIGNAL consist of two programs (written in ASSEMBLER language for the IBM System/370) and an intervening sort. The initial program constructs a continuous stream of transliterated text from the input corpus, adding text on the right while shifting and truncating on the left. This technique of retaining only a limited segment of text in memory at any given instance allows SIGNAL to process large corpora in a minimum of main storage. As each sign value reaches the center of the

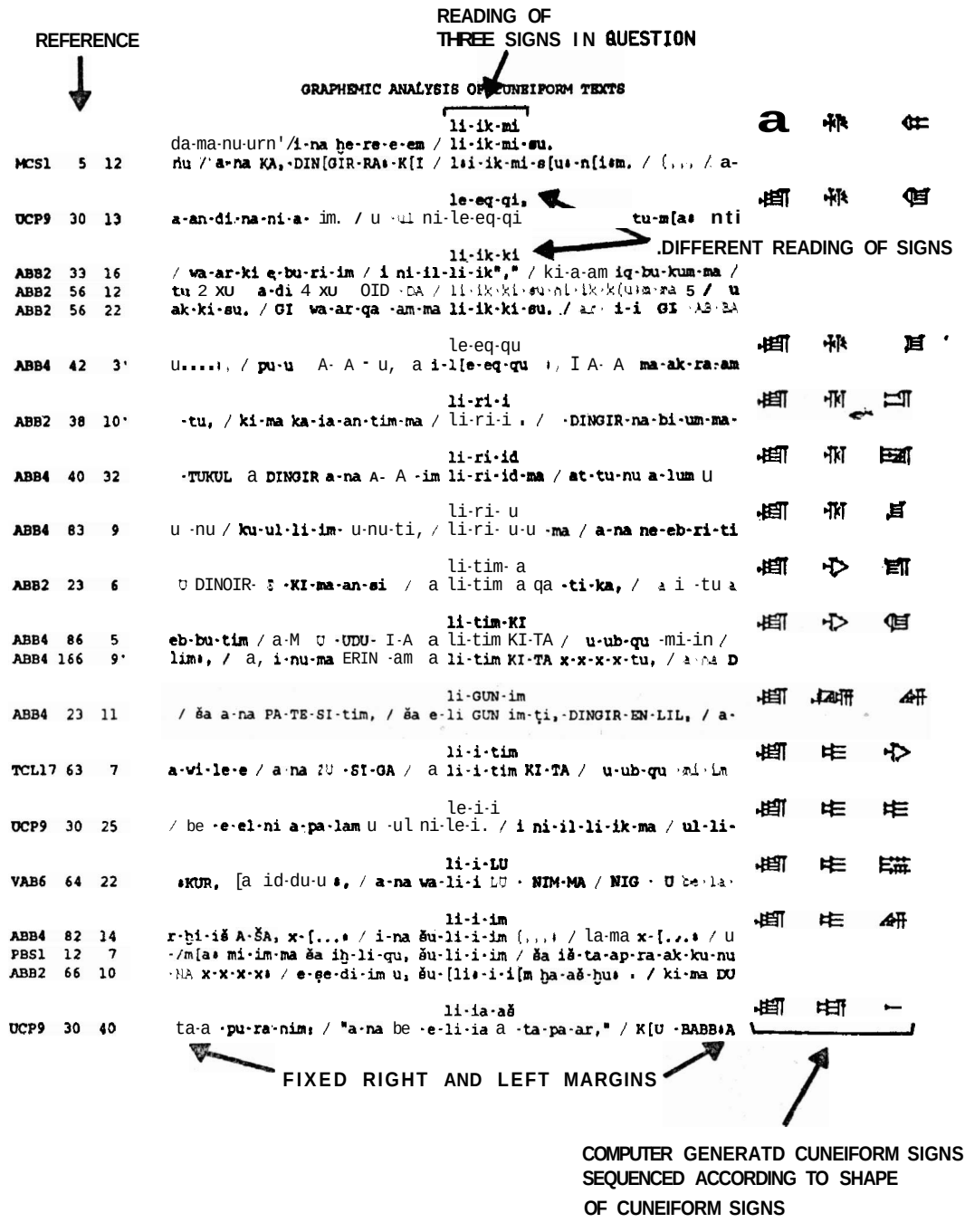

READING OF
THREE SIGNS IN QUESTION
REFERENCE
GRAPHEMIC ANALYSIS OF CUNEIFORM TEXTS
li-ik-mi
da-ma-nu-urn'/i-na ḫe-re-e-em / li-ik-mi-šu.
MCS1 5 12
li-eq-qi,
UCP9 30 13
a-an-di-na-ni-a- im. / u -ul ni-le-eq-qi tu-m[a nti
li-ik-ki
.DIFFERENT READING OF SIGNS
ABB2 33 16
/ wa-ar-ki ṣ-bu-ri-im / i ni-il-li-ik"," / ki-a-am iq-bu-kum-ma /
ABB2 56 12
tu 2 XU a-di 4 XU OID -DA / li-ik-ki-su-nu-ik-k(u)m-ma 5 / u
ABB2 56 22
ak-ki-su. / GI wa-ar-qa -am-ma li-ik-ki-su. / ar- i-i GI -AŠ-BA
le-eq-qu
ABB4 42 3'
u....., / pu-u A- A - u, a i-l[e-eq-qu , I A- A ma-ak-ra-am
li-ri-i
ABB2 38 10'
-tu, / ki-ma ka-ia-an-tim-ma / li-ri-i . / -DINGIR-na-bi-um-ma-
li-ri-id
ABB4 40 32
-TUKUL a DINGIR a-na A- A -im li-ri-id-ma / at-tu-nu a-lum u
li-ri- u
ABB4 83 9
u -nu / ku-ul-li-im- u-nu-ti, / li-ri- u-u -ma / a-na ne-eb-ri-ti
li-tim- a
ABB2 23 6
U DINOIR- I -KI-ma-an-si / a li-tim a qa -ti-ka, / a i -tu a
li-tim-KI
ABB4 86 5
eb-bu-tim / a-M U -UDU- I-A a li-tim KI-TA / u-ub-qu -mi-in /
ABB4 166 9'
lim , / a, i-nu-ma ERIN -am a li-tim KI-TA x-x-x-x-tu, / a-na D
li-GUN-im
ABB4 23 11
/ ša a-na PA-TE-SI-tim, / ša e-li GUN im-ṭi,-DINGIR-EN-LIL, / a-
li-i-tim
TCL17 63 7
a-wi-le-e / a-na ?U -SI-GA / a li-i-tim KI-TA / u-ub-qu -mi-in
le-i-i
UCP9 30 25
/ be -e-el-ni a-pa-lam u -ul ni-le-i. / i ni-il-li-ik-ma / ul-li-
li-i-LU
VAB6 64 22
KUR, [a id-du-u , / a-na wa-li-i LU - NIM-MA / NIG - U be-la-
li-i-im
ABB4 82 14
r-ḫi-iš A-ŠA, x-[...] / i-na šu-li-i-im (...) / la-ma x-[...] / u
PBS1 12 7
-/m[a mi-im-ma ša iḫ-li-qu, šu-li-i-im / ša iš-ta-ap-ra-ak-ku-nu
ABB2 66 10
-NA x-x-x-x / e-ṣe-di-im u, šu-[li-i-i[m ḫa-aš-ḫu . / ki-ma DU
li-ia-aš
UCP9 30 40
ta-a -pu-ra-nim / "a-na be -e-li-ia a -ta-pa-ar," / K[U -BABB A
FIXED RIGHT AND LEFT MARGINS
COMPUTER GENERATD CUNEIFORM SIGNS
SEQUENCED ACCORDING TO SHAPE
OF CUNEIFORM SIGNS

Fig. 2.

constructed text stream that sign becomes the sign in question. The value is extracted from the embodying punctuation and textual diacritics and is used to access a table containing a correlation of sign values to a set of standard Assyriological numbers identifying signs and their cuneiform shapes. In addition to the correlation of the transliterated values to the cuneiform signs the sign in question is categorized in relation to the two adjacent signs to the right. A record containing the numerical equivalent of the three signs and their context is written to an intermediate file, which after being sorted, is formatted by the final program to produce the concordance.

At the time of formatting a determination is made as to whether the output is to be directed to an online printer or is to be photo-composed by the Information International COMP-80. In the latter case references are converted to font and graphic codes: these refer to a previously defined set of instructions stored in the COMP-80 which enable it to produce on film the actual cuneiform shapes. (For an illustration of the output see fig. 2).

6. MORPHOLOGICAL AND LEXICAL CATEGORIZATION (ARTHUR R. SORKIN)

As described above, the encoded data cards contain the grammatical analysis, transliteration, and transcription. The first phase of the program reads the data cards into main memory. As they are read, the analysis, transliteration and transcription for each card are separated. The transliteration and transcription are converted into special data codes (as required for special Akkadian characters). Then they are concatenated onto the ends of the strings of text consisting of respectively all of the previously read transliteration and transcription. When the entire corpus has been read, there are in storage two complete texts, one in transliteration and one is transcription.

At the same time, the grammatical analysis is examined as each card is read to determine if the word falls into one of the desired grammatical categories. If it does, then the analysis is stored, along with reference data and pointers to the locations of the word in the copies of the transliteration and transcription that are being constructed.

As an illustration, assume that the entire corpus consisted of the cards in (2) (above, section 3). Further assume that only verbal roots were of interest. Then there would be three different areas in memory which would contain the following:

(transcription):

ana belīya qibī-ma umma šamaš-hāzir-ma:

(transliteration):

a-na be-li2-ia qi2-bi2-ma um-ma DINGIR-UTU-*ha-zi-ir-ma:*

(analysis):

1-2 NCGSM *B:L EWEPE:S- HS location of *bēlīya, be-li2*
2-3 VB-1-2SM*QB: IW- location of *qibī-ma, qi2-bi2*

Also read and stored at this time are the comment type footnotes, which are retrieved during the output phase.

When the entire corpus has been read, and the desired grammatical analyses have been stored, the sorting phase of the program starts. Because of the speed characteristics of the computer being used, an IBM System 360 model 91 with 4,000,000 bytes of main memory, it was decided to use the internal sorting algorithm of Shell (WILLIAM A. MARTIN, *Sorting*, in « Computing Surveys », (Dec. 1971) 3-4, pp. 147-174). It was chosen as being reasonably fast and simple to program. In the sorting phase only the stored analyses are sorted. The copies of the text are not moved.

First, the analyses are sorted by major category: verbal roots, non-verbal nouns, foreign words, pronouns, invariables, etc. Then each of the major categories is sorted according to the type of analysis that it has. For example, in the verbal root section the analyses are sorted alphabetically by root. Then items with the same root are sorted as to whether they have a stem or a pattern. Then the various stems are sorted into the desired order. Likewise, the patterns are also sorted into order. Items with the same stem (or same pattern) are further sorted according to the rest of the grammatical analysis, e.g. case, state, mood, person, number, gender, as appropriate. Finally, items with identical analysis are alphabetized by the word itself. Items with identical analysis and word are finally sorted by chronological data and reference.

The processing of the other major categories proceeds similarly, using the appropriate grammatical parts. For example, the non-verbal nouns with full inflection are first sorted alphabetically by base.

Once the sort phase is done, then the output phase is started. The sorted analyses are examined, and the type of the item is determined. The analysis is examined to see if there is a comment type footnote for

the item. Using the pointers stored with the analysis, a line of context is constructed, with the word in question in transliteration, and the rest of the context in trascription. For items having lexical data, e.g., roots, stems, patterns, bases, that information is recovered at this time. Finally, the output page is constructed with the appropriate spacing, headings, etc., so that each item is printed along with any coded and comment type footnotes and any lexical information. The output is then written on magnetic tape. When all of the analyses have been processed, the finished tape is ready to be processed by the COMP-80 to produce the finished microfilm copy (see below, section 7). For an illustration of the printed output see figure 3.

7. TEXT PROCESSING AND PHOTOCOMPOSITION

7.1. Non-recurrent Text *Processing* (*John* L. Settles).

In addition to the information extracted from the data bank and formated for publication by the various programs of OBLAP, a certain amount of non-recurrent material is prepared for each publication. Of necessity, this material contains any or all of the special symbols occurring elsewhere in the volume. As a result, a hybrid system was developed, which uses the Format Management System (FMS) available through the Campus Computing Newtork at the University of California, Los Angeles.

FMS is a programming system designed to arrange free format text into a structured copy. A continuous stream of input text, consisting of text interspersed with format control commands, is punched into eighty column cards. The normal output medium of FMS is the printed page; however OBLAP has taken advantage of FMS's capability of producing a magnetic tape suitable for further processing.

The OBLAP "TEXTOR" system consists of a preprocessor and postprocessor to the FMS system. Special format control commands, undefined in FMS, were developed which allow the inclusion of such symbols as Ḫ (het), Š (shin), etc., into the text stream. The OBLAP preprocessor to FMS converts these special control commands into character configurations, which, although acceptable to FMS as text, normally result in a blank space. The combination of FMS and OBLAP commands is provided for, so that the submission of lower case as well as upper case is permissible.

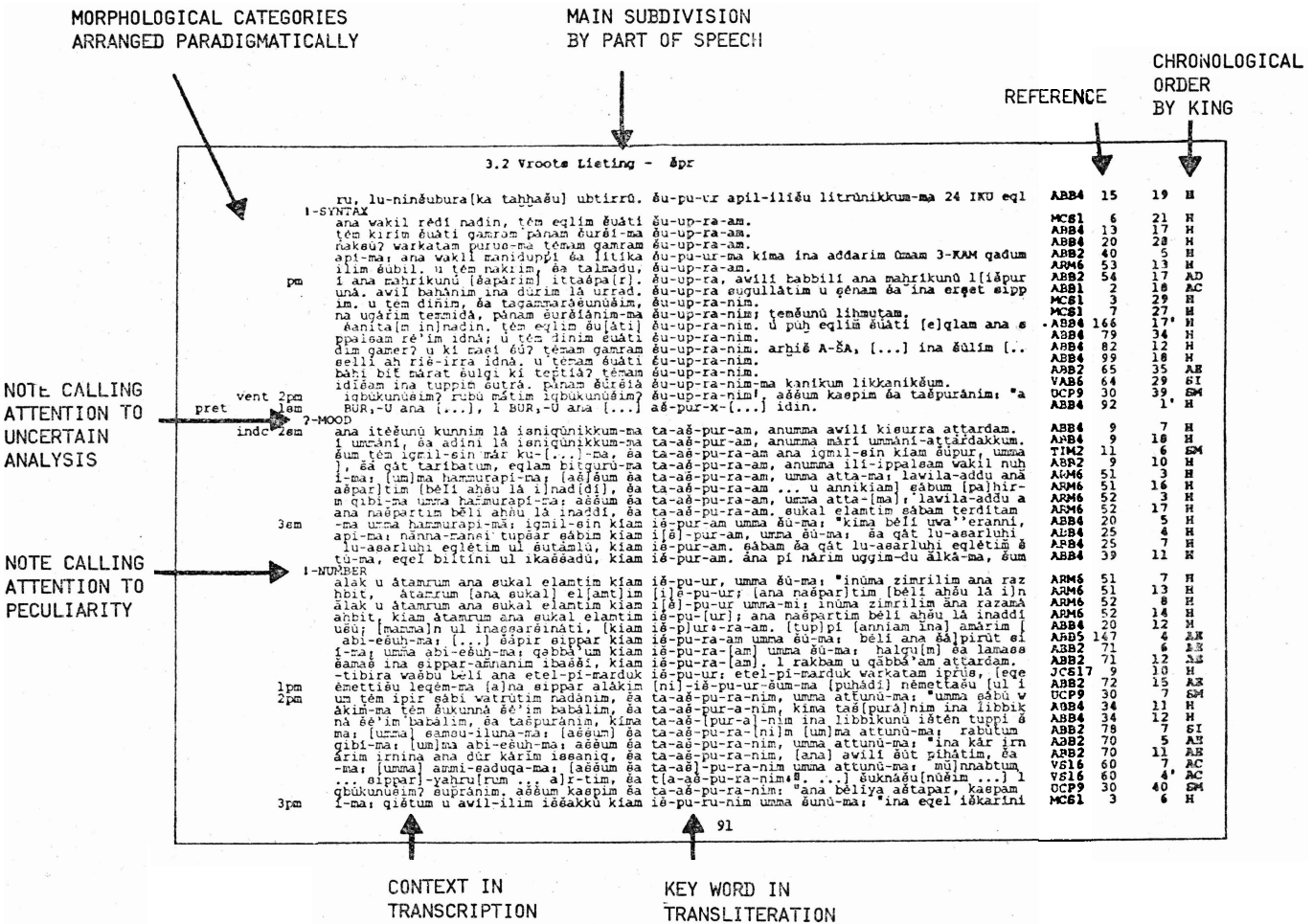

MORPHOLOGICAL CATEGORIES
ARRANGED PARADIGMATICALLY
MAIN SUBDIVISION
BY PART OF SPEECH
CHRONOLOGICAL
ORDER
BY KING
REFERENCE
3.2 Vroots Listing - šprNOTE CALLING
ATTENTION TO
UNCERTAIN
ANALYSIS
NOTE CALLING
ATTENTION TO
PECULIARITY
CONTEXT IN
TRANSCRIPTION
KEY WORD IN
TRANSLITERATION
91

Fig. 3.

The postprocessor to FMS converts the magnetic tape file containing the FMS structured text into a format acceptable to the Information International COMP-80 video microfilm recorder. Each occurrence of a character configuration defined as a special graphic, is converted to the COMP-80 codes necessary to "draw" the appropriate symbol.

7.2. Photocomposition (Sal *J. Fallone*).

An association between technological methods of space exploration and the production of cuneiform text by computer may not be readily apparent. Lunar exploration requirements to produce highly detailed video images fostered development of video systems having resolutions of 16,384 lines as compared to the more familiar 1,200 line resolution of ordinary television. The addition of small computers to these video systems enabled development of a device which is capable of illuminating any point in a square matrix of 16,384 by 16,384 points and allows the video beam to draw lines between any points in the matrix. With provision for mounting suitable cameras at the face of the video tube, the high resolution microfilm recorder came into existence as a modern photocomposition device.

Concurrent with this development, the printing industry made substantial improvements in phototypesetting. Offset platemakers using 35 mm. microfilm, rather than typeset, as an original image were perfected. The microfilm images are directly expanded onto either photosensitive paper or zinc offset press plates by means of a suitable lens system.

The Information International COMP-80 is a programmable video microfilm recorder used for production of OBLAP texts. The COMP-80 characters are constructed by invocation of a prepared set of instructions which drive the video beam in a predetermined pattern. Under the auspices of OBLAP, a complete set of instructions for each cuneiform sign occurring in the Old Babylonian corpus was created. The instructions for drawing the signs are for relative motions of the video beam only and do not specify the exact size or the location of the sign on the printed page. Figure 4 exemplifies the OBLAP method for creation of cuneiform character. The expanded sign is shown on a grid-like background to assist the encoder in the determination of errors in the code. A reduced image is displayed in the upper right hand corner in the size chosen for the final printed size. Below the reduced image, the identification number and the instruction set appear. The sets of instructions for all

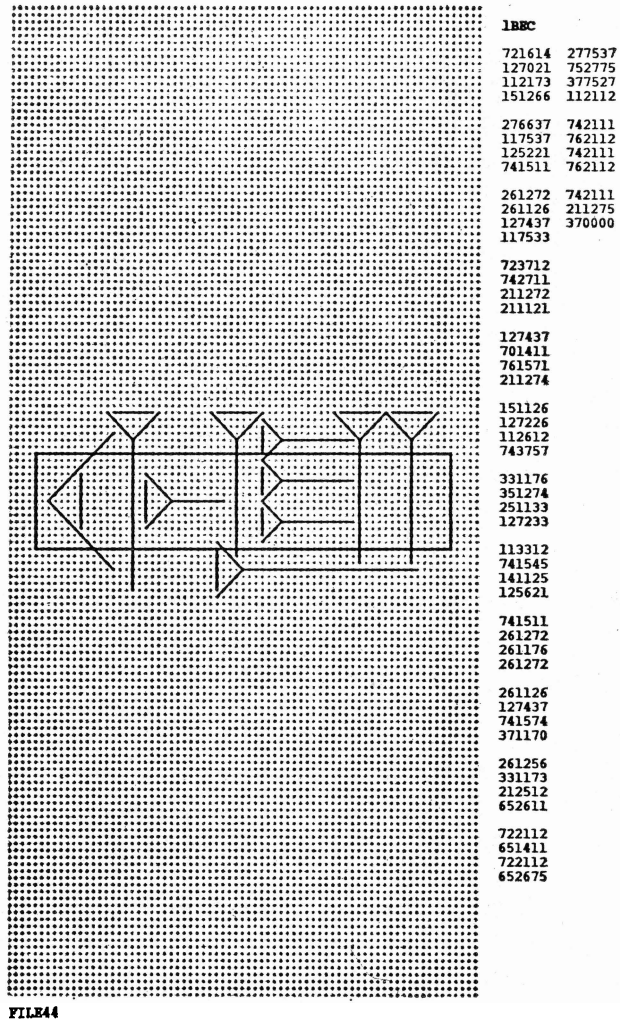


Fig. **4.**

signs, together with identification numbers, are assembled in a table which is in essence a conventional type font.

The typesetting of the OBLAP text is performed from a magnetic tape formatted by programs described above in section 7.1. The formatting consists of page change, line space, and character size instructions for each sign in the proper order, applies a size multiplication factor to the instructions, positions the starting point of the video scan,

turns on the beam, and " writes " the line of text. A page consisting of 64 such lines is set in less than 5 seconds. To produce the microfilm master, the tape is used with a 35 mm. camera mounted. A subsequent interpretation of the magnetic tape with a microfiche camera mounted creates the master microfiche.

The phototypesetting methods and techniques developed for OBLAP are directly applicable to any typeface and will be applied in the near future to other Ancient Near Eastern and Semitic languages.

8. PUBLICATIONS AND DATA BANK

The Data Bank is accessible to all interested scholars in the form of specialized outputs. Besides standard sort programs of single features or clusters of features (RESOL, not illustrated here), a more useful program, particularly for syntactical research, provides listings of features or clusters of features in full context, i.e. within sentence boundaries, with the relevant item underlined (RECON, see figure 5).

ša *ta-aš-pur-am.*

ABB2 1- 1 4 2

aš-šum ša *t[a-aš]-pur-am* um-ma at-ta-ma:

ABB2 4 3- 3

[a-d]i [t]e⁴-em-ka la *aš-pur-am*-[ma] Si-pi²-ir ID²/na:rim/-im Sa ih-he-ru-[u²] la i-mu-ru-nim mu-u² a-na Hi-ip-ri-im ga-am-ri-im la uš-ta-ar-du-u²

ABB2 4 1'- 4

iš-tu ID²/na:ram/ šu-a-ti te-eh-te-ru-u² ši-[i]p-ra-am ša *aš-pu-ra*-kum [e]-pu-uš

ABB2 5- 18- 7

aš-šum NAGAR-MEŠ/nagga:ri/ [ša]-k[a-nim] MA²-NI-DUB/maniduppi:m/ [e-pe²-Si-im] Sa *aš-pur-ak*-kum-[ma] um-ma at-ta-ma:

ABB2 8- 6- 7

. . . . . .

Cards Read: 14.932
Examples Printed: 86
Bibliography: ABB II: COMPLETE

Fig. 5.

Next to the Data Bank, specific publication series are also planned as part of OBLAP. They are based on selected corpora extracted from the main data base because of internal unifying criteria: each corpus consists of texts which are homogeneous in terms of geographical or social provenience, or of chronology, or the like – the purpose being to favor linguistic analysis of a coherent body of evidence in each case. It must be noted that such criterion is not customary in the field of Assyriology, where letters are normally gathered and published at best as archives, that is in virtue of their having been found together in the same archaeological locus (normally the addressee's tablet repository, a point of convergence which often does not correspond to linguistic homogeneity on the part of the senders). As a result, the establishment of the corpus requires in itself a good deal of bibliographical research and the determination of clear criteria for defining the nature and extent of homogeneity. The first corpus prepared includes all royal letters from Babylon, i.e. letters bearing the name of a king of Babylon as the sender: there is a total of 209 letters (approximately 12,000 words), culled from 17 publications, plus one unpublished text. The second corpus, about half the size of the royal letters, includes all local letters from Harmal, a provincial town to the northeast of Babylon; these texts, half of which are unpublished, are being edited for OBLAP by RIA DE J. ELLIS of the University Museum in Philadelphia. Other corpora on which work is currently under way are the local letters from Mari and the letters of royal officials from southern Babylonia.

Each one of these corpora will be processed by means of three basic program systems, yielding three different series: Graphemic Analysis of Cuneiform Texts (GAC), Morpho-lexical Analysis of Akkadian Texts (MLA), and Syntactical Analysis of Akkadian Texts (SAA). Work is at present being completed on the first two systems only, while the third system has yet to be developed, although the pertinent analysis is already encoded in the input; the publication will be similar in format to MLA, except that categorization of the data will be according to basic syntactical features (types of sentences and clauses; word order; noun phrase structure, e.g. nominalization; transformation, e.g. deletion).

Common to all series is the practical feature that they will be published both in microfiche and in hard copy. The much cheaper microfiche edition will bring our results within reach of scholars who intend to use the publication as an occasional reference tool, while the hard copy edition may prove to be more useful for specialized libraries and scholars engaged in closely related areas of research.

Since the first volumes of both GAC and MLA will appear in print at about the same time as the Proceedings of this Conference, and since the programming aspects of the underlying computational systems have already been discussed in the preceding sections, I will not give here a description of the format of the volumes except for calling attention to some of the more salient features.

GAC's most striking peculiarity is the fact that the cuneiform signs are reproduced in a standardized cuneiform shape (Figure 2) produced in the manner described in the preceding section. Reproduction of the original script is not a luxury, but corresponds to a precise need to which this series addresses itself. The high degree of polyvalence of the cuneiform signs can only be resolved with the help of the context. If the context is obscure or broken (as often happens with cuneiform tablets) only comparison with identical sequences of signs in known contexts can help. GAC provides the best opportunity for such comparison: by giving the signs in the cuneiform script, and by arranging them according to a sequence based on the shape of the cuneiform combinations rather than according to the alphabetic sequence of transliterated values, the signs can easily be retrieved even when their reading is unknown or uncertain. A list of signs based on the transliteration would prejudge the reading, and would make it more difficult to identify the desired sequences.

GAC will be of invaluable help when applied to literary texts: since many duplicate ancient copies exist of such texts, and since they are often reduced to small fragments, identification of fragments will be greatly facilitated once the existing texts are processed through the computer. It must also be noted that this system can be utilized for all syllabic cuneiform scripts, and not only for Akkadian; applications to Sumerian and Hittite are planned for the near future.

MLA provides an analysis of all words in the corpus arranged according to a complex hierarchy of lexical, morphological and chronological categories (figure 3). An important feature is that the listing of the text switches from graphemic transliteration to phonemic transcription depending on whether an item is the word in question or simply part of the context. This enables the reader to follow the context easily, while giving at the same time, for the word in question, the exact form as found in the text, with only a minimum of linguistic normalization. In this manner an adequate philological basis is preserved to enable the scolar to verify the validity of any given linguistic interpretation. Another useful feature is represented by the footnotes, both in code and in

prose form, which allow greater flexibility in indicating all that is necessary for an exhaustive linguistic analysis.

Various computations of frequency sequences and statistical correlations complete each volume in both series.

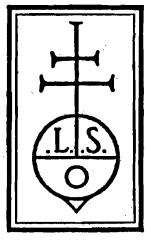
9. TECHNICAL SPECIFICATIONS

IBM System/360 Model 91
4 million bytes main memory
2314 direct access storage.

IBM System/370 Model 135
192K memory
3330 direct access storage.

GIORGIO BUCCELLATI

THE OLD BABYLONIAN LINGUISTIC ANALYSIS PROJECT: GOALS, PROCEDURES AND FIRST RESULTS



FIRENZE
LEO S. OLSCHKI EDITORE
MCMLXXVII

Estratto da:

COMPUTATIONAL
AND MATHEMATICAL LINGUISTICS

PROCEEDINGS OF THE INTERNATIONAL
CONFERENCE ON COMPUTATIONAL LINGUISTICS

Pisa, 27/VIII-1/IX 1973

* *

LEO S, OLSCHKI EDITORE - FIRENZE - MCMLXXVII