(1987/2017). In A. Bunger (Ed.)., *Special volume: Reissue of Innovations in Linguistic Education Volume 5. IULC Working Papers in Linguistics*, Vol. 17, No. 6.

LEXISTAT:
A PASCAL PROGRAM FOR
CREATING LEXICOSTATISTICAL EXERCISES

J. Jerome Smith

Department of Anthropology
University of South Florida
Tampa, Florida 33620
(813) 974-2150

Introduction

Glottochronology and the probabilistic model of lexicostatistics have been praised and damned over the years since their formulation and development by Morris Swadesh in the 1950s (e.g., 1950, 1951, 1952, 1955). A host of summaries, critiques, refinements and applications throughout the 1950s, '60s, and '70s testify to the interest and skepticism aroused by this statistical approach to the study of language history through the calibration of lexical loss (e.g., Lees 1953, Hoijer 1956, Olmstead 1957, Hymes 1960, Bergland and Vogt 1962, Chretien 1962, Dyen 1962, Diebold 1964, Dyen 1975, Georges 1975, Munro 1978).

Research relying solely or largely on lexicostatistics has tapered off considerably in the past ten years--for a variety of reasons related to continuing criticisms of the validity of its assumptions and the reliability of its empirical base. However, the purpose of this paper is not to raise and review these issues but to offer my colleagues in linguistic instruction an additional resource for dealing with them in the classroom.

What I want to do in this paper, then, is to describe a program I have written in Pascal for generating exercises in lexicostatistics. In a general course I teach in anthropological linguistics, the program has provided an interesting point of departure for discussing a number of issues in historical linguistics and in the interconnections of linguistics, cultural anthropology and archaeology.

What follows should be considered a user's manual for LEXISTAT--a mircrocomputer program that simulates a model of lexical change and lexicostatistical analysis presented by Sarah Gudschinsky 1956 in her classic paper, 'The ABCs of lexicostatistics (glottochronology)'. LEXISTAT is available from me on request in the hope that it will be used and improved upon by others who share my interest in the instructional uses of computers.

What LEXISTAT can do for you

LEXISTAT creates its own 'protolanguage' and three daughter 'languages' (PAM, SUE and LIZ) for lexicostatistical analysis. You may request new data or call up existing data for repeated use. You may save data for future use. The data are presented on three screens--one for each pair of languages. You may request printouts of data for use as worksheets and/or answer sheets.

LEXISTAT operates both interactively and automatically. At each step of analysis--from establishing cognate pairs to analyzing for depths of divergence--you are free to enter decisions and information from the keyboard or to ask LEXISTAT to do its own work for your review. Several menus give you easy access to ANY part of the exercise at any time--generally without altering work going on in other parts of the exercise. Thus, LEXISTAT encourages you and your students to work independently with the word lists while it stands ready at any moment to provide ITS assessment of the data for your review.

Using LEXISTAT's ability to save and recall exercises and to produce hardcopy worksheets and answer sheets, the program need not be run in the classroom at all. However, the quick reinforcement it provides for student efforts recommends it for in-class use.

One classroom approach is to divide students into three groups and assign a pair of languages to each, using hardcopy worksheets. LEXISTAT, working with the same data, can be queried for assistance and reinforcement as the students progress through the exercise. When the groups have completed their work, their data can be entered directly into LEXISTAT (as described below) to see how well each group has performed.

The data created by LEXISTAT

The data are three 100-word lists of core vocabulary--matched for semantic correspondence and representing the contemporaneous languages PAM, SUE and LIZ. Words are in a constant form of CVCVC. The full consonant set consists of nine cycling consonants:

/b ==> p ==> f/ (labials)
/d ==> t ==> s/ (alveolars)
/g ==> k ==> x/ (velars)

and for non-cycling consonants:

/m n l r/ (resonants)

There are five non-cycling vowels:

/a e i o u/

The protolanguage is created using a reduced 'upper case' set of twelve consonant and vowel phonemes;

/B D G M N L R A E I O U/

The general historical pattern of divergence for the three languages is shown in Figure 1. The maximum depth of the primary divergence of PAM:SUE/LIZ from PROTOLANGUAGE is 5000 years. The minimum depth of the intermediate divergence of SUE/LIZ from PROTOLANGUAGE is 2000 years.

Figure 1: LEXISTAT's general pattern of language divergence

LEXISTAT creates its word lists by invoking 1) the assumptions of lexicostatistics (especially the principle of a constant rate of retention per millenium across languages), 2) a rate of phonemic shift for the labial consonants (voiced ==> voiceless ==> fricative), and 3) a rate of random change for vowels and resonants.

Once the primary and intermediate depths of divergence have been randomly established for the PROTOLANGUAGE and LIZ/SUE branchings, respectively, the proportion of retained words in a list is calculated by X^Y, where X is Gudschinsky's recommended constant rate of retention (.805) and y is the depth of divergence in millenia. For every word in a list, if a random number is less than X^Y, then the word is randomly changed. The phonemes of each retained or new word are then examined in a similar manner for the likelihood of further random or systematic shifts. In this fashion an attempt is made to roughly simulate just the language dynamics students are being asked to explore.

The protolanguage is created using a reduced 'upper-case' set of twelve consonant and vowel phonemes:

/B D G M N L R A E I O U/

The general historical pattern of divergence for the three languages is shown in Figure 1. The maximum depth of the primary divergence of PAM:SUE/LIZ from PROTOLANGUAGE is 5000 years. The minimum depth of the intermediate divergence of SUE/LIZ from PROTOLANGUAGE is 2000 years.

Figure 1: LEXISTAT's general pattern of language divergence

LEXISTAT creates its word lists by invoking 1) the assumptions of lexicostatistics (especially the principle of a constant rate of retention per millenium across languages), 2) a rate of phonemic shift for the labial consonants (voiced ==> voiceless ==> fricative), and 3) a rate of random change for vowels and resonants.

Once the primary and intermediate depths of divergence have been randomly established for the PROTOLANGUAGE and LIZ/SUE branchings, respectively, the proportion of retained words in a list is calculated by X^y, where X is Gudschinsky's recommended constant rate of rentention (.805) and y is the depth of divergence in millenia. For every word in a list, if a random n is less than X^y, then the word is randomly changed. The phonemes of each retained or new word are then examined in a similar manner for the likelihood of further random or systematic shifts. In this fashion an attempt is made to roughly simulate just the language dynamics students are being asked to explore.

The result is a set of data that is highly normalized and biased in the direction of a 'successful' solution--but complex enough to be interesting and instructive. Suggestions for other approaches to data generation are actively solicited.

How to use LEXISTAT

LEXISTAT is available from the author. It is written in Pascal for a Kaypro II (CP/M) using a Heath/Zenith H-19 terminal. Other systems will require some modification and recompiling. Since the source code (lexistat.pas) is provided with ample commentary (disk files and hardcopy), conversion to other systems with Pascal compilers is straightforward. Attention to screen control codes is all that should be required for system transportation. Include a blank 5 1/4" (single-sided double-density) disk with all requests, unless hardcopy listings alone are sufficient to your needs.

LEXISTAT consists of a main program (lexistat.com) and two included files (lexcount.inc and lexidata.inc). After modifying and compiling as necessary, run 'lexistat' from the operating system prompt. Except for supplying filenames in save and load routines, all requests for choices and other information respond to single keystrokes. <CR> is not necessary after input.

The title screen presents two options:

Use (1) existing data
    (2) new data

Do NOT request existing data unless a language data file has been created through previous use of LEXISTAT. ('lexidat1' is on disk for those who wish to use it.) No directory is currently available, so be sure to have appropriate filenames at your disposal while the program is running. New or existing data will be brought into use while the screen displays information on the assumptions and procedures to be used in lexicostatistical analysis. After the data have been loaded and you have discussed the information screen with students, <CR> ('return' or 'enter' keys) will take you to the options menu of LEXISTAT.

The core of LEXISTAT from the user's point of view is the options menu screen (Table 1).

Table 1: LEXISTAT options menu

The data screens

Options 1-3 take you to one of three data screens for cognate counting. Each pair of lists in presented in a 20-row by 5-column format of semantically equivalent items, as partially shown in Table 2.

Table 2: Part of a LEXISTAT data screen before cognate counting

Instructions for cursor movement and cognate decisions are at the bottom of the screen. A running total of the cognate count is maintained at the lower right corner of the screen.

The cursor rests on cognate decision positions (initially all '0') and may be moved among them with vertical and horizontal wrap-around in two ways:

1) Arrow keys (or their equivalent) move the cursor without affecting cognate counts. Strict circular (either vertical or horizontal) wrap-around is maintained in this mode of cursor movement.

2) '0' or '1' may be entered to indicate that a pair is cognate ('1') or non-cognate ('0'). This input gives a line feed to the next row in the column. At the bottom of a column the feed is to the top of the column to the right. At the bottom of the rightmost column the feed is to the top of the leftmost column.

<CR> returns the user to the options menu without affecting cognate counts, which are shown to the right of options 1-3, respectively. The user is free to go back to any data screen for further work before analyses are performed.

Assumptions and procedures

Options 4 (Table 3) and 5 (Table 4) bring up assumption and procedure screens, respectively, as seen at the outset of the program. These follow Gudschinsky 1956 closely.

Table 3: LEXISTAT assumptions screen

Table 4: LEXISTAT procedures screen

After reading Gudschinsky and discussing the assumptions and procedures for doing lexicostatistical analysis, the student should be able to determine the cognate status of semantically matched lexical pairs, so that the data in Table 2 now appear as shown in Table 5.

Table 5: Part of a LEXISTAT data screen after cognate counting

Analyses

Options 6-8 allow for various analyses of the data. Option 6, for instance, permits the introduction of cognate counts independently of those accumulated through work with the data screens. This is especially

useful if students are working with worksheets and supply suggested counts for testing. The option leads to an analysis of divergence depths, so be sure each count is between 1 and 99. Zero or empty counts will lead to 'division by zero' errors. The program will 'crash' and return to the operating system.

Option 7 provides a program-generated cognate count, replacing any work you may have done on the data screens with LEXISTAT's own assessment. After noting identical items, LEXISTAT actually examines the five corresponding phonemes of each lexical pair in each pair of lists, as if it were a student (nearly 1500 comparisons in less than 10 seconds). This option is most useful for checking student work and producing 'answer' sheets.

Option 8 performs analyses of the depths of divergence of the three languages. The formulas used are presented on the screen before LEXISTAT's results are displayed as shown in Table 6. A typical analysis is shown in Table 7.

Table 6: LEXISTAT analysis formulas

Table 7: A typical LEXISTAT analysis

Saving data and quitting LEXISTAT

Option 9 allows the instructor to save existing data for future use. After the data have been saved an option to print hardcopy is offered. (Be sure the printer is online with paper set at the top of a page.) Since the data saved include the current cognate status of each lexical pair ('0' or '1'), saving and printing new data (where cognate status is initialized at '0' for all pairs) will create a worksheet that can be reproduced for student use. Saving and printing after cognate pairs have been determined (manually or through option 7) will provide an answer sheet for grading and other purposes.

The user quits the program using control-C. Exit is to the operating system with no data retention. The command may vary from system to system and should be modified as appropriate.

Additional comments

I am myself skeptical of the reliability and validity of lexicostatistics, especially when it is viewed as a mechanical procedure for the determination of absolute chronology. However, it represents a major effort to enhance the study of language dynamics with statistical models and students should be aware of it.

I find the tightly argued assumptions of lexicostatistics to be especially helpful in discussing linguistic relativity and lexical semantics. Anthropology students are particularly intrigued by the linkages of linguistic and archaeological research suggested by the lexicostatistical approach.

In addition to its pedagogical interest, I have wanted to share LEXISTAT as a demonstration of the value of writing such programs in Pascal. Most of us developing simulations and exercise generators for microcomputers probably work in BASIC, because it is supplied with our machines. This is unfortunate because it is so difficult to transport our programs to other systems, as anyone who has tried it will testify.

The advantage of a stable compiled language like Pascal is that the source code is readily transferred from one system to another, where the particular Pascal compiler then performs the conversion directly. Typically, only cursor and screen control codes need be manually modified. My hope is that others will be encouraged to follow the example set here--to do instructional programming with an eye toward sharing the results widely with colleagues.

Be that as it may, my students in linguistics and anthropology courses thoroughly enjoy getting their hands on language data--however ersatz--and discovering that the abstract models so typical of linguistic theory can be put to work in concrete instances. If LEXISTAT and other exercise generators encourage this sense of discovery in students they have much to recommend them.

FIGURE 1: LEXISTAT's general pattern of language divergence

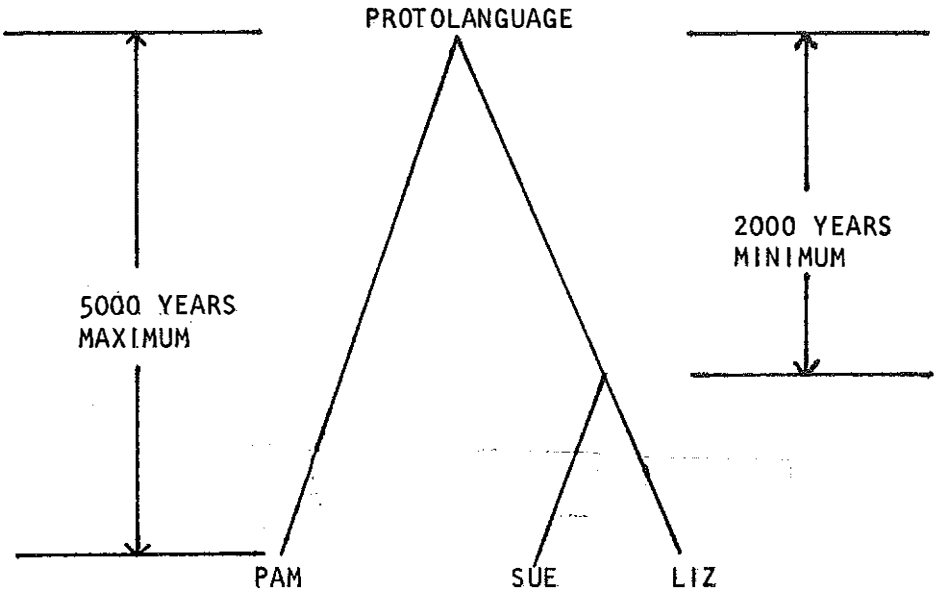

TABLE 1: LEXISTAT options menu

1. Examine PAM and SUE.
2. Examine PAM and LIZ.
3. Examine SUE and LIZ.
4. Review Assumptions.
5. Review Procedures.
6. Analyze other cognate counts.
7. Request computer determination of cognate counts.
8. Cognate counts complete. Begin analysis of relations.
9. Save names, dates, and lists to disk.

·C (control-C) Quit LEXISTAT (no data retained).

TABLE 2: Part of a LEXISTAT data screen before cognate counting

	PAM		SUE	PAM		SUE
		A			B	
1	mabal	0	tinak	gorog	0	rimum
2	radur	0	ratur	nodem	0	rumuk
3	momir	0	momir	gamug	0	kamuk
4	derim	0	terer	nugar	0	kapar

TABLE 3: LEXISTAT assumptions screen

ASSUMPTIONS

1. Core vocabulary is least subject to change.
2. The rate of retention (r) in core vocabulary is stable.
3. The rate of retention (r) is constant across languages.
4. Divergence is calculated by relating (r) to cognate counts.

ADDITIONAL ASSUMPTIONS FOR THIS SIMULATION

1. Once diverged, languages develop in TOTAL isolation.
2. Phonemic change is most likely to occur at a given point of articulation.
3. Vowels are least likely to experience systematic change.

TABLE 4: LEXISTAT procedures screen.

PROCEDURES

1. Eliminate borrowed words. (N/A for this simulation).
2. Establish semantic equivalence (given here).
3. Test for cognate status (3/5 agreement of phonemes).
   a. Phonemic identity.
      (/P/-word1 = /P/-word2)
   b. Phonetic similarity.
      (N/A for this simulation)
   c. Phonetic conditioning.
      (N/A for this simulation)
   d. Phonemic correspondance
      (/P/-word1 ^{2} /P/-word2)

TABLE 5: Part of a LEXISTAT data screen after cognate counting

	PAM		SUE		PAM		SUE
		A				B	
1	mabal	0	tinak		gorog	0	rimum
2	radur	1	ratur		nodem	0	rumuk
3	momir	1	momir		gamug	1	kamuk
4	derim	1	terer		nugar	0	kapar

cognate count = 4

TABLE 6: LEXISTAT analysis formulas

The analysis follows Gudschinky. Use the following formulas to complete your own calculations before proceeding:

T = time of divergence in millenia.
C = % cognates.
r = retention rate per millenium (.805).
n = list size (100).
S = standard error (@ 7/10 confidence level).
E = range of error.

T = log(C)/(2xlog(r))
S = sqroot ((Cx(1-C))/n)
E = T-(log(C+S)/(2xlog(r)))

TABLE 7: A typical LEXISTAT analysis

	Number of Cognates (Cx100)	Calculated Depth (Tx1000)	Actual Depth (years)	Range of Error (E)	Within Range (Y/N)
PAM/SUE	18	3953	4396	±446	Y
PAM/LIZ	17	4084	4396	±460	Y
SUE/LIZ	44	1892	2134	±246	Y

References

Bergland, Knut and Hans Vogt. 1962. On the validity of glottochronology. *Current Anthropology* 3.115-53.

Chretien, C. Douglas. 1962. The mathematical models of glottochronology. *Language* 38.11-37.

Diebold, A. Richard. 1964. A control case of glottochronology. *American Anthropologist* 66.987-1006.

Dyen, Isidore. 1962. The lexicostatistically determined relationship of a language group. *International Journal of American Linguistics* 28.153-61.

_____. 1975. *Linguistic Subgrouping and Lexicostatistics*. The Hague: Mouton.

Georges, Nia. 1975. Lexicostatistics of five Eastern Tukanoan languages. *Human Mosaic* 9.1-15.

Gudschinsky, Sarah C. 1956. The ABC's of lexicostatistics (glottochronology). *Word* 12.175-210.

Hoijer, Harry. 1956. Lexicostatistics: A critique. *Language* 32.49-60.

Hymes, Dell H. 1960. Lexicostatistics so far. *Current Anthropology* 1.3-34.

Lees, Robert B. 1953. The basis of glottochronology. *Language* 29.113-27.

Munro, Stanley R. 1978. Glottochronologic theory: Valid or not valid in Chinese languages? *The Canadian Journal of Linguistics* 23.55-65.

Olmstead, D. L. 1957. Three tests of glottochronological theory. *American Anthropologist* 59.839-42.

Swadesh, Morris L. 1950. Saliah internal relationships. *International Journal of American Linguistics* 16.157-67.

_____. 1951. Diffusional cumulation and archaic residue as historical explanations. *Southwestern Journal of Antrhopology* 7.1-21.

_____. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96.453-63.

_____. 1955. Towards greater accuracy in lexicostatistical dating. *International Journal of American Linguistics* 21.105-15.