Eodermdromes:
A Graph-Theoretical Tool for Linguistics

G. Bloom, J. Kennedy, A. Gewirtz, P. Wexler

Abstract. We apply graph theoretical techniques to a representation of linguistic structures called *spelling nets*. Studying *eodermdromes* is an initial step in determining how to extract from spelling nets, some quantifiable characteristics of linguistic structure. Generating *minimal eodermdromes* is seen to be an intellectually challenging new 'word game' which is now being attacked by man and computer. We also note in this paper, how the underlying practice problem gives impetus to studying open problems in graph theory.

Introduction

Some relationships which are central to linguistics have traditionally lent themselves to graphical representations of an elementary kind [1]. We are presently in the process of applying graph theoretical ideas to one of these representations, namely *spelling nets*, in order to suggest ways of *quantifying* characteristics of complex linguistic structures [2] used to compare languages and to trace language development. In this paper we will present these ideas in three steps. Initially, we will introduce a special class of spelling nets that have been dubbed *eodermdromes* and illustrate their properties in the context of an appealing new class of word games [3]. Secondly, we link the properties of eodermdromes to 'serious' problems of linguistics. Finally, we show how these applications motivate research into areas of graph theory about which relatively little is currently known.

1 Spelling Nets, Eodermdromes and Word Games

A *spelling net* is a labelled graph whose points represent *characters* (i.e. letters, words, etc.) and whose lines represent character adjacencies. To construct a spelling net, using letters as characters,

1. write down the distinct letters of a word or phrase,
2. connect pairs of letters using the sequence of lines comprising the Eulerian path spelling out that word or phrase.

Example: The spelling net for THIS IS A SPELLING NET is shown in Figure 1, but it is not an eodermdrome.

An *eodermdrome* is defined to be a non-planar spelling net. We show the spelling net for EODERMDROME in Figure 2.

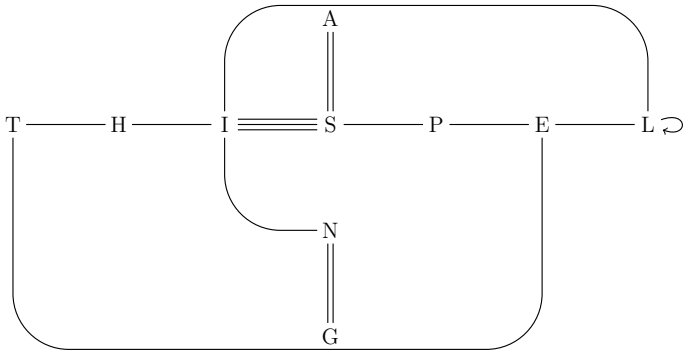

Fig. 1. The spelling net for THIS IS A SPELLING NET

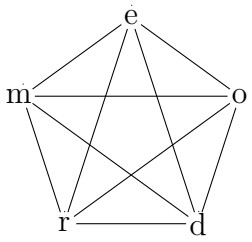

Fig. 2. The spelling net for EODERMDROME

The *size* of a spelling net is the number of lines in it. Although $K_{3,3}$ with its nine lines is the smallest non-planar graph, it cannot represent a spelling net because it has no Eulerian path. At least two lines need to be added to $K_{3,3}$ to generate a graph containing an Eulerian path, consequently a minimal eodermdrome is represented by the 10-line, Eulerian, non-planar graph K_5.

We have amused ourselves by producing minimal eodermdromes. For example, each line of the following 'poem' is a minimal eodermdrome:

TEARS AT REST

Stray satyrs,
Dense and sad,
Tip tan paint.

Teaser's tart
Pursue prep, ...
Yearly relay.

Sweat wastes ...
Science sins ...
Ah, ... rather tea.

Although each of these lines has a K_5 spelling net, there are structural differences between them. In Figure 3A we show for lines 2, 4 and 7 of the poem the first polygon generated in tracing out each of their spelling nets. In Figure 3B the letters of these spelling nets have been replaced by integers representing the order of occurrence of those letters. It is apparent that line 4 of the poem is structurally different from lines 2 and 7. The sequences of Figure 3 are completed in an obvious way to (1 2 3 4 5 1 4 2 5 3 1) and (1 2 3 1 4 5 3 4 2 5 1), corresponding to lines 2 (or 7) and 4 respectively of the poem. Such sequences illustrate a canonical representation for the underlying structural patterns in minimal eodermdromes. For K_5 eodermdromes there are 22 sequences each corresponding to one of the distinct Eulerian circuits. A complete listing of these sequences is given in Table 1 along with a representative eodermdrome corresponding to each sequence.

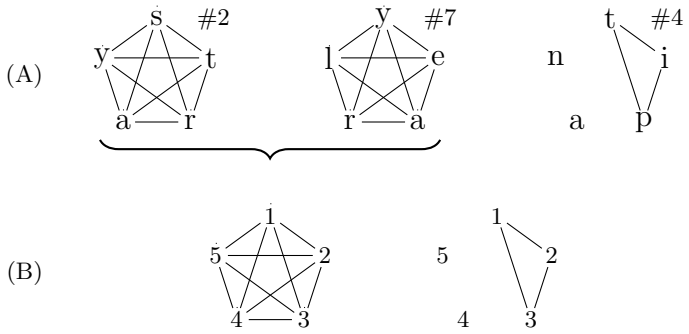

Fig. 3. (A) Starting polygonal sequences of the spelling nets of lines #2, #4, and #7 of the poem. (B) Ordinal representation of those starting sequences: 1 2 3 4 5 1 & 1 2 3 1.

The examples in Table 1 and in the poem were contributed by eodermdromophiles acknowledged in reference [3] where a more exhaustive list of minimal eodermdromes is given.

The number of eodermdromic sequences which can be formed from the 26 letters of the Roman alphabet is large, viz. $\binom{26}{5} \times 5! \times 11 = 86,829,600$. All of these can be computer generated, but very few form meaningful English phrases. The task of extracting the meaningful phrases from this list is formidable. By incorporating sufficiently many orthographic and syntactic constraints one can hope to reduce the task of manually searching for meaningful phrases in the computer-generated list to a manageable level. For example,

1. Q cannot appear, since, in its common English usage, it must always be followed by a U. However, every letter must occur at least twice in a minimal eodermdrome and no letter pair may be repeated.
2. In English, certain letter sequences do not occur [4], thus, for example, there is no word starting with B whose second letter is C, F, G, I, J, K, M, N, P, Q, S, T, V, X, or Z. Similarly, every word starting with IB must have as its next letter A, E, L, O, R, U, or Y.

1.	1 2 3 1 4 2 5 3 4 5 1	SIN SCIENCES
2.	1 2 3 1 4 2 5 4 3 5 1	AH, TAR HER TEA
3.	1 2 3 1 4 3 5 2 4 5 1	HOT HAT COACH
4.	1 2 3 1 4 3 5 4 2 5 1	TARTERS EAST
5.	1 2 3 1 4 5 2 4 3 5 1	REAR PIE PAIR
6.	1 2 3 1 4 5 3 4 2 5 1	RED RUM DUE, MR.
7.	1 2 3 4 1 3 5 2 4 5 1	SHOES ON HENS
8.	1 2 3 4 1 3 5 4 2 5 1	KIOSK ON SINK
9.	1 2 3 4 1 5 2 4 5 3 1	TRACTOR COAT
10.	1 2 3 4 1 5 4 2 5 3 1	TREAT: BAR BET
11.	1 2 3 4 2 5 1 3 5 4 1	TORN OUT RUNT
12.	1 2 3 4 2 5 1 4 5 3 1	DREARY DAY, ED
13.	1 2 3 4 2 5 3 1 4 5 1	DENSE AND SAD
14.	1 2 3 4 2 5 3 1 5 4 1	SPIN POISONS
15.	1 2 3 4 2 5 4 1 3 5 1	PURSUES PREP
16.	1 2 3 4 2 5 4 1 5 3 1	ORDER NEON, DO
17.	1 2 3 4 5 1 3 5 2 4 1	GIANT GATING
18.	1 2 3 4 5 1 4 2 5 3 1	YEARLY RELAY
19.	1 2 3 4 5 2 4 1 3 5 1	SCIENCE SINS
20.	1 2 3 4 5 2 4 1 5 3 1	SWEAT WASTES
21.	1 2 3 4 5 3 1 4 2 5 1	TEARS AT REST
22.	1 2 3 4 5 3 1 5 2 4 1	EARLY, RE: YALE

Table 1. The 22 canonical minimal eodermdrome sequences with examples

To explore this approach, we are currently developing a computer program which will generate only the 'pentagonally starting' eodermdromic sequences #17 and #18, in which each of the five distinct letters appears exactly once in the first five positions followed by the re-occurrence of the first letter. Generating only this restricted set allows one to impose in the program a variety of additional linguistically-based constraints. These constraints significantly reduce the size of the computer-generated list which must be manually scanned. At the same time virtually no 'meaningful' eodermdromes are eliminated, so that their presence will be easier to detect in the reduced sample. An example of the type of linguistically-based rule that can be utilized here is that sequences #17 and #18 cannot represent meaningful English eodermdromes if they start with X or J. Moreover, we have decided to restrict 'acceptable' eodermdromes to those which do **not** contain either obsolete words or abbreviations. Thus, we do not consider as acceptable eodermdromes either SUTRAS TAURS ("a philosophy of bulls") or SULKY, SLY U.K.'S ("certain people from the United Kingdom"). We also intend to have the students at Daytop Miniversity of Brooklyn College examine the feasibility of extending the procedure to the rest of the list of minimal eodermdromic sequences.

A quite different, but very tractable computer approach to finding minimal eodermdromes is to give a program a set of vocabulary words from which it will generate eodermdromes. This vocabulary may be any choice of words all

of which are formed using a set of 5 distinct letters. In the currently running preliminary version of the program, it is necessary for the individual choosing one of the $\binom{26}{5}$ possible English letter sets also explicitly to choose the set of words incorporating those letters. Typically, this has meant that the vocabulary sizes for the letter sets that have been considered to date range from 30 to 90 words. In running the preliminary version of the program on a vocabulary of 47 words, the program produces 122 eodermdromic word sequences of which 5 were clearly 'meaningful' and another 9 made sense if one loosened the requirements for being meaningful to allow contractions and strange word orders. This program can become extremely useful if it is coupled to a computer-based dictionary, so that **all** legal vocabulary using a desired letter set can be included in its input.

The algorithm for producing minimal eodermdromes from given vocabulary is based on obtaining all one's in a 'sum' vector whose components are the number of times each possible letter adjacency pair occurs in a phrase. The procedure is easily illustrated by showing how the words PAINT, TAN, and TIP can be properly combined. When the words are entered in the program, they are encoded as integers so that, for example, 1 = A, 2 = I, 3 = N, 4 = P, 5 = T. Each word is stored as a vector whose first component is the first letter and whose second component is the last letter of the word. The remaining 10 components of the vector have 0 or 1 entries depending on whether the letter adjacency pairs that they represent are present in the word. The order of the components is lexico-graphical, that is, (1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5), (3,4), (3,5), (4,5). Thus,

$$\begin{aligned} \text{PAINT} &= (4,5,1,0,1,0,1,0,0,0,1,0) \\ \text{TAN} &= (5,3,0,1,0,1,0,0,0,0,0,0) \\ \text{TIP} &= (5,4,0,0,0,0,0,1,1,0,0,0) \end{aligned}$$

The sum of the final 10 components is $(1,1,1,1,1,1,1,0,1,0)$. When the words are placed in the order TIP TAN PAINT, the final letters of TIP and TAN combine with the initial letters of TAN and PAINT respectively. Consequently, ones are entered into the sum vector for components (4,5) and (3,4), giving an all-one's sum vector and an eodermdrome. The phrase TIP TAN PAINT is also eodermdromic, but none of the other four permutations of these words are eodermdromes.

Even with our preliminary version of this program running, we have not as yet been able to answer many of the fascinating word-game questions for minimal eodermdromes that naturally come to mind. We list some of these questions here; others can be found in [2,3].

i) Anagrammatic sets of minimal eodermdromes with different letter sequences exist. For example, sequences #6, #14, and #16 are represented by the anagrams SEA'S TARTARS, STAR TEASERS, and STARTER SEAS respectively. Can one set of letters yield all 22 patterns with (reasonably) meaningful minimal eodermdromes?

ii) If complete anagrammatic sets of minimal eodermdromes cannot be found, then is it possible to fill in the missing patterns with eodermdromes constructed from the same set of 5 letters? For example, minimal eodermdromes that are *homolexical* with the anagrammatic ones cited above are TEASER'S TART and TEARS AT REST which exhibit patterns #15 and #21 respectively.

iii) We noted above that the K_5 structure and English orthography preclude finding a Q in a minimal eodermdrome. Clearly, no palindromic constructions can be used, since letter pairs may not recur in any order. Are any other language artifacts ruled out by the K_5 structure that we are considering?

iv) In the list of minimal eodermdromes in Table 1, only #5, REAR PIE PAIR, contains 3 vowels, the others all contain 2 vowels. We know of no examples of minimal eodermdromes containing 1 or 4 vowels. Do any exist?

v) What examples exist in other languages? In an old form of Polish WROG WARGA ÓW (which means "the enemy of Warga") exemplifies pattern #9.

vi) None of the words found in the 2nd and 3rd editions of the Webster's Unabridged Dictionaries is a minimal eodermdrome. This was one of the motivations for coining the word EODERMDROME which is a minimal eodermdrome. Are any words that are not in those dictionaries, such as scientific terms, minimal eodermdromes?

vii) Letters are not the only units that can be used in constructing minimal eodermdromes. The following example exhibits pattern #2 when the *words* in the paragraph are taken as the units.

MARCH IN STUDENTS. MARCH PAST IN MAY. PAST STUDENTS MAY MARCH.

The meaning of this paragraph is clear, if one interprets it as processional instructions being given to spring graduates and alumni.

2 Linguistic Considerations

Although our investigations are still in a preliminary stage, we suspect that eodermdromes may prove useful in several linguistic contexts. This is based on the obvious observation that some languages are far more 'rigid' than others. The rigidity can evince itself in various ways: (i) in orthographic rules governing letter order (such as the English Q being followed by U); (ii) in rules precluding or requiring various sound sequences; and (iii) in rules regulating the order of words in sentences. It is a well-known linguistic principle that languages which extensively modify words to show their 'parts of speech' (nouns, verbs, etc.) and to give other information (tense, gender, number, etc.) have little need for rigidity of word order. (Similarly, a language which uses word order to convey grammatical function has less need for, and tends to lose, the endings and internal modifications that tell the user of a word what its grammatical function is.)

As a consequence of the differences between allowable sequencing of units in different languages, we anticipate that comparable sets of spelling nets will

differ statistically from one another in various languages. We are currently at the beginning of a research program that will attempt to determine what features of spelling nets convey the greatest amount of linguistic information. We suspect that with the 'right' statistical measures and with sets of nets using the various possible units, that a quantitative system of describing languages can be developed. With such a scale one can investigate several exciting possibilities.

1. Do modern languages differ significantly from one another on such a structural scale? Are such differences significantly pronounced so that knowing the value of the structural variable is sufficient to identify the language (or at least the family of languages) whose value is given?
2. Can one meaningfully indicate in a numerical way how a modern language has evolved, by tracing the functional behavior of spelling net statistics?
3. Do the statistics for spelling nets of individual or group language usage differ sufficiently so that speakers or writers of one language can be distinguished from one another?

These are large and difficult questions for which the answers will not be immediate. However, even some easier questions can be investigated that will start us moving toward the larger questions. For example, in the English dictionary (choose one such), characterize the collection of orthographic spelling nets.

3 Applicable Graph Theory

We were able to answer the first graph theoretical question that we asked for this linguistic project. Indeed, it is easy to recognize that spelling nets have to be graphs with Eulerian paths. It is also quite direct to determine that K_5 is the smallest eodermdrome, and that there are 22 distinct Eulerian paths that can be traced in that graph. After those few insights, the questions become considerably more difficult.

Question I: How many Eulerian paths are there on an arbitrary spelling net? Or, more restricted, for odd m, how many Eulerian paths are there on K_m? There is 1 on K_3; there are 22 on K_5; therefore the question is open.

If one wants to use statistics of the spelling nets to characterize languages, it is necessary to find an appropriate function to partition the spelling nets into classes characterized by a numerical value of the function. What would prove useful? One possibility is to characterize the spelling nets by their genus. Unfortunately, this seems too crude, since there are not likely to be spelling nets of interest with genus greater than one. Consequently virtually all of the spelling nets would be in one of only two classes.

A better measure of the complexity of spelling nets would need to be more refined in the sense that it partitions the nets into more classes. Yet it should give some global measure of the underlying graph. Connectivity might be suitable, but on the other hand, it may be too sensitive and react too strongly to local subgraphs. For example, all of the complexity of a structure would be ignored if the spelling net happened to have a cutpoint in it. Thus, connectivity

measurements quantify what is least complex in a graph. A graph parameter that does the opposite is the 'crossing number' of a graph. The *crossing number* is simply the minimum number of edge crossings that occur in a graph when it is drawn as well as possible. When the crossing number is zero, we have the class of planar graphs. However, when the crossing number is one, only a subset of the genus one graphs are included. Graphs with crossing number two are members of both the genus one and genus two classes. And so on. Unfortunately, although it seems that the crossing number of a spelling net may be the most useful graph parameter to measure what is 'complex' about the net, little is known about the crossing numbers. Recently, White and Beineke [5] summarized what is known, mostly in the form of bounds, and cited a few references in the area.

Question II: How does one determine the crossing numbers for graphs of relatively low order? Does knowing that the graphs are spelling nets with no more than two points of odd degree simplify the calculation in any way?

Until more is known about crossing numbers it will be difficult to give general answers to our questions about languages. We expect that useful statistical measures of the complexity of the structure of a language (or some subset of a language) are its Mean Crossing Numbers. These are the mean crossing numbers for the language's orthographic, phonic, and lexiconographic spelling nets.

Finally, we wonder how well one can measure the complexity of language usage by determining how 'closely' constructions that are constrained to have low crossing number approximate 'normal' unconstrained speech and writing. We can consider the language constrained to have constructions with crossing number not greater than k to be the kth-order approximation to the unconstrained language. We suspect that for some very low value of k, in most cases, the approximation will converge to normal usage.

References

1. Stewart, A. (1979). *Graphical Representation of Models in Linguistic Theory.* Bloomington, IN: Indiana University Press.
2. Bloom, G., Kennedy, J. & Wexler, P. (1980). "Linguistic Complexity and Minimal Eodermdromes." *Linguistics* 18(1-2),pp. 3-16
3. Bloom, G., Kennedy, J., & Wexler, P. (1980). "Ensnaring the Elusive Eodermdrome." *Word Ways* 13(3), pp. 131-40
4. Book Club Associates. (1979). *The Compact Edition of the Oxford English Dictionary.* Oxford: Oxford University Press.
5. White, A. & Wilson, R. *Selected Topics in Graph Theory* (Eds. L.W. Beineke & R.J. Wilson), Academic Press, New York, 1979.