

Origins of combinatorics on words Lyon 2008

Dominique Perrin

November 17, 2008

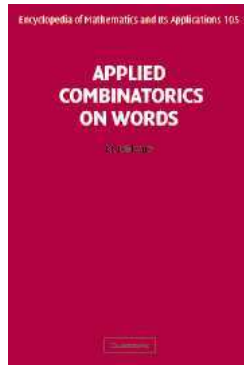
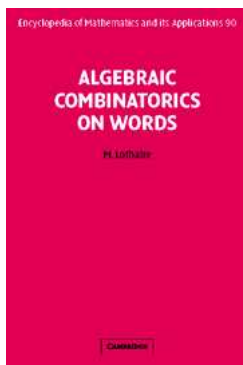
Outline

A gallery of portraits with variations on

- Unavoidable regularities
- Symbolic dynamics
- Necklaces
- Path encodings

Source: Jean Berstel, D.P., The origins of combinatorics on words, *European Journal of Combinatorics*, **28**, 2007, 996–1022.

The Lothaire series





The **Thue-Morse word** $t = abbabaab \cdots$ is obtained by iterating the substitution $a \mapsto ab$, $b \mapsto ba$ starting with the letter a . One obtains

abba baab baab abba baab abba abba baab \cdots

Thue proved that t is **overlap-free**, in the sense that it has no factor of the form $uvuvu$ for some words u, v , with u non empty.

Thue also proved a nice relationship between **square-free** words on three letters and overlap-free words. Indeed, for any infinite overlap-free word x over two letters a, b , the inverse image of x under the substitution

$$a \mapsto abb, \quad b \mapsto ab, \quad c \mapsto a$$

is a square-free word on three letters a, b, c without the factors aba or cbc , and conversely.

Starting from the Thue-Morse word t , one obtains the square-free word on three symbols

$$m = abcacbabcbac \cdots .$$

Prouhet

Indices of a less than 8 in $abbabaab$: 0,3,5,6.

$$\begin{aligned}0 + 3 + 5 + 6 &= 1 + 2 + 4 + 7 \\0^2 + 3^2 + 5^2 + 6^2 &= 1^2 + 2^2 + 4^2 + 7^2\end{aligned}$$

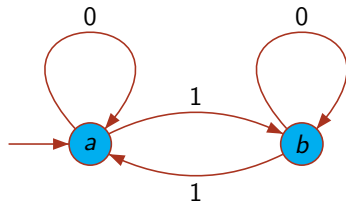
The Thue Morse word is a solution of the problem raised by Prouhet in 1851 (the Tarry Escot problem): find the nontrivial solutions of

$$a_1^k + a_2^k + \dots + a_r^k = b_1^k + b_2^k + \dots + b_r^k, \quad (1 \leq k \leq m)$$

with m as large as possible.

Automatic words

A word is called **automatic** if its n -th letter is the state reached by a finite automaton reading the representation of n in base k . The n -th letter of the Thue-Morse word is the parity of the number of 1 in the binary expansion of n .



The idea of automatic words appears in the work of **Cobham** in 1972, who proved that an infinite word is k -automatic iff it is the image, under a length-preserving morphism, of a fixed point of a substitution of constant length k .

Unavoidable regularities



Finding infinite words that avoid repetitions has its roots in the work of Thue and has been pursued, in particular in connection with problems of algebra.

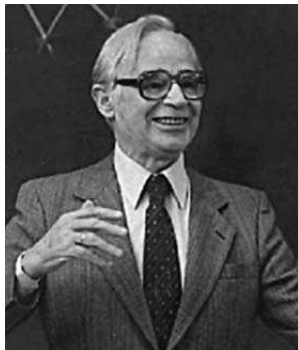
The general idea of unavoidable regularities is actually much wider and does not have a precise definition. One direction is that initiated by Ramsey in 1930.

Applied to infinite words, his famous theorem implies that if we color the factors of an infinite word x in r colors, there is a factorization

$$x = v u_0 u_1 u_2 \cdots$$

such that all factors except the first have the same color.

Van der Waerden



A related unavoidable regularity is given by **van der Waerden's** theorem: if the positive integers are partitioned in r classes, then at least one of the classes must contain arbitrary long arithmetic progressions. This theorem had been conjectured by I. Schur.

Van der Waerden heard of the conjecture through Baudot, a student at Göttingen at the time, and referred to his result as Baudot's conjecture. This applies directly to words, with the equivalent formulation that for any infinite word x there are arbitrary large integers k such that for some n, m , we have $x_n = x_{n+m} = \dots = x_{n+km}$.

Symbolic dynamics

The study of dynamical systems is derived from the work of Newton on the laws of motion applied in particular to the planetary system.



Words come in with what was called by **Morse** (1892-1977) and **Hedlund** a *symbolic flow*. The elements, called *symbolic trajectories* are just infinite words. The idea goes back to Hadamard, who first used sequences of symbols to describe qualitatively the infinite geodesic curves on a surface. Morse proved the existence of a non-periodic uniformly recurrent point by considering the Thue-Morse word.

Sturmian words

An infinite word over a binary alphabet is called **Sturmian** if for all $n \geq 0$, the number of its factors of length n is $n + 1$.



As an example, the **Fibonacci word** $f = abaababaabaab \cdots$ which is defined as the unique fixed point of the substitution $(a \mapsto ab, b \mapsto a)$ is Sturmian.

525-2



COURS D'ANALYSE

DE

L'ÉCOLE POLYTECHNIQUE,

PAR M. ^{Charles} STURM,
Membre de l'Institut,



PUBLIÉ D'APRÈS LE VOEU DE L'AUTEUR

PAR M. E. PROUHET,
Professeur de Mathématiques.

TOME SECOND.

PARIS,

MALLET-BACHELIER, IMPRIMEUR-LIBRAIRE

DU BUREAU DES LONGITUDES, DE L'ÉCOLE IMPÉRIALE POLYTECHNIQUE,
QUAI DES AUGUSTINS, 55.

1859

(Mademoiselle Anna Sturm, propriétaire des Œuvres posthumes de son frère, et M. Mallet Bachelier, éditeur, se réservent le droit de traduction.)

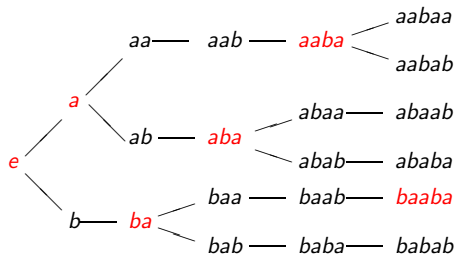


Figure: The factors of the Fibonacci word.

Mechanical words

There is also an alternative definition of the Fibonacci word using approximations of irrationals by rationals. Let α be some irrational with $0 < \alpha < 1$, and let $s(\alpha) = (s_n)$ be the sequence

$$s_n = \begin{cases} a & \text{if } \lfloor (n+1)\alpha \rfloor = \lfloor n\alpha \rfloor, \\ b & \text{otherwise} \end{cases}$$

For $\alpha = 2/(3 + \sqrt{5})$, one has $s(\alpha) = af$. This formula shows that the symbols s_n can be interpreted as the approximation of a line of slope α by points with integer coordinates. It is a theorem due to Morse and Hedlund that Sturmian words can be defined equivalently by a formula as above with α an irrational number.

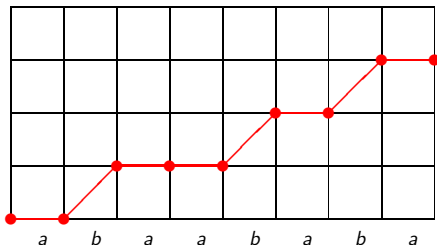


Figure: The graphical representation of the Fibonacci word

Mechanical words (also called Christoffel words) have historical roots in the work of the astronomer Jean **Bernoulli** III who studied these words in connection with continued fractions.

Continued fractions

The connection with continued fractions can be summarized as follows. Let us denote by $[n_0, n_1, n_2, \dots]$ the **continued fraction**

$$n_0 + \frac{1}{n_1 + \frac{1}{n_2 + \dots}}$$

Let $[0, 1 + d_1, d_2, \dots]$ be the continued fraction expansion of some irrational α with $0 < \alpha < 1$. Let w_n be the sequence of words defined by

$$w_{-1} = b, \quad w_0 = a, \quad w_n = w_{n-1}^{d_n} w_{n-2} \quad (n \geq 1).$$

Then $s(\alpha) = \lim w_n$. In particular, let us consider the case of the Fibonacci word. We have $s(\alpha) = af$ with $\alpha = 2/(3 + \sqrt{5})$. Actually, $2/(3 + \sqrt{5}) = [0, 2, 1, 1, \dots]$ in accordance with the fact that the sequence w_n is the sequence of Fibonacci words.

Necklaces

A circular word, or **necklace**, is the equivalence class of a word under circular shift. A necklace of length n is primitive if its period is not a proper divisor of n .



Figure: A necklace with period 3, and a primitive necklace.

Enumeration of necklaces



The enumeration of necklaces of length n on k symbols has been known for a long time. It appears explicitly in a paper of **MacMahon** of 1892. The formula

$$M(n, k) = \frac{1}{n} \sum_{d|n} k^d \mu(n/d)$$

for the number $M(n, k)$ of primitive necklaces where μ is the Möbius function, is often called **Witt's formula**.

The formula for the total number of necklaces of length n on k symbols

$$N(n, k) = \frac{1}{n} \sum_{d|n} k^d \varphi(n/d)$$

where φ is Euler's function, is called **MacMahon's formula**.

de Bruijn cycles

A famous result is the existence, for each $n, k \geq 1$, of a de **Bruijn cycle** of length k^n on k letters. This is a necklace such that each word of length n over k letters appears exactly once as a factor. For example,

aaaaabaaabbaababababbbababbabbbb

is a de Bruijn cycle of length 32 for $k = 2, n = 5$. This result has a curious and interesting history. First, it was actually explicitly obtained many years before by **Flye Sainte-Marie** in 1894. He proved that the number of such necklaces of length 2^n on a binary alphabet is $2^{2^{n-1}-n}$.

The result had been conjectured by Rivière the same year as question 48:

Si l'on considère tous les arrangements n à n qu'on peut former avec deux objets, il est toujours possible de trouver un arrangement de 2^n termes (formé avec les mêmes deux objets) a_1, a_2, \dots, a_{2^n} , tel que les groupes

$$a_1, a_2, \dots, a_n; \quad a_2, a_3, \dots, a_{n+1}; \quad \dots$$

$$a_{2^{n-1}}, a_{2^n}, a_1, \dots, a_{n-2}; \quad a_{2^n}, a_1, a_2, \dots, a_{n-1};$$

représentent tous les arrangements n à n dont le nombre est évidemment 2^n . Cette proposition est vérifiée expérimentalement jusqu'à des limites suffisantes pour en présager l'exactitude. Est-elle déjà connue? Pourrait-on en donner une démonstration? Y a-t-il en général plus d'une espèce de solutions et dans ce cas combien?

Posthumus

The same problem was reintroduced by Martin in 1934, who first had the idea of producing the least possible such word by a greedy algorithm. Independently, the problem was raised again in 1943 by Posthumus (1902–1990), a radio engineer working at the Philips Research Laboratories. He found that the number of de Bruijn cycles for $n = 1, 2, 3, 4, 5$ was 1, 1, 2, 16, 2048. The cycles with $n = 5$ were of technical interest. Indeed, the **Baudot code**, patented in 1874 by Emile Baudot (1845–1903) use 5-bit words to encode 32 characters. The number of cycles was, however, not of any practical significance, and the interest in proving the formula seems to have been purely *pour le sport*. De Bruijn worked at this problem when he began to work himself at the Philips laboratories in 1944. He worked out by hand the case $n = 6$, finding 67108864 cycles and used the techniques he had developed to achieve the computational prowess to prove a formula answering the question of Posthumus.

de Bruijn graph

The corresponding problem for an alphabet with k letters seems to have been first raised and solved by Tanja van Aarden Ehrenfest and de Bruijn. The number is $k^{-n}(k!)^{n^{k-1}}$. The proof, as the original one by Flye Sainte-Marie and de Bruijn uses what is now called the de Bruijn graph of order n .

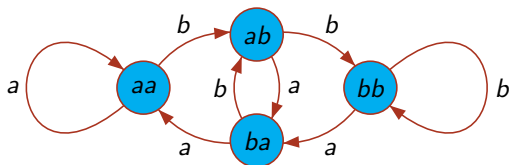
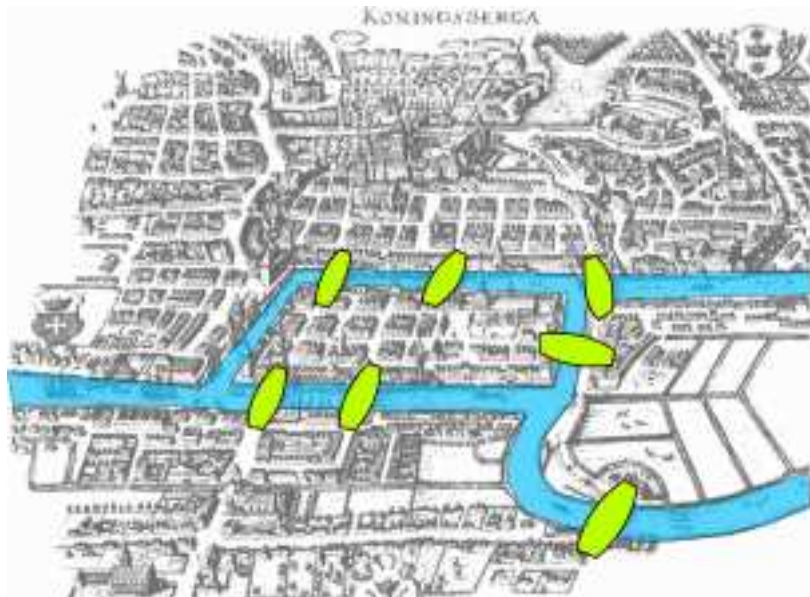


Figure: The de Bruijn graph of order $n = 3$.

Eulerian graphs

The de Bruijn graph is Eulerian since each vertex has indegree k and outdegree k . By Euler's theorem it has an **Eulerian cycle**. A graph is *Eulerian* if there is a cycle going through all edges exactly once. Euler's theorem states that a graph is Eulerian if and only if it is connected and every vertex has same indegree and outdegree. This theorem, published in 1736, can be considered the first theorem of graph theory. It is supposed to have been motivated by the puzzle of the seven bridges on the Pregel, in the city of Königsberg, now Kaliningrad. The label of an Eulerian cycle is a de Bruijn word.

The seven bridges of Königsberg



The BEST theorem

The formula has been generalized by a theorem sometimes called the BEST theorem proved independently by de Bruijn, van Aarden-Ehrenfest and by Smith, Tutte. It enumerates all Eulerian cycles in an Eulerian graph G with n vertices v_1, v_2, \dots, v_n by the formula

$$t_i(G) \prod_{j=1}^n (d(v_j) - 1)! \quad (1)$$

where $t_i(G)$ is the number of spanning trees rooted at v_i and with edges oriented towards v_i and $d(v)$ is the outdegree of the vertex v . The role played by spanning trees in this formula comes from the nice combinatorial property associating with each Eulerian cycle the tree of edges used to leave of a vertex for the last time.

As an example, Figure 5 represents the two possible spanning trees rooted in bb in the de Bruijn graph of order 3. Following the Eulerian path starting and ending at the root, we obtain the two possible de Bruijn words

$aaababbb$ and $abaaabbb$.



Figure: The two spanning trees of de Bruijn graph of order $n = 3$ with root bb .

Kirchoff laws

The introduction of the number of spanning trees of a graph and its characterization in terms of the adjacency matrix of the graph is itself much older, since it appears with the work of **Kirchhoff** on electrical networks.



The number of spanning trees with a given root in a graph is computed by a determinant related to the adjacency graph of the matrix.

The Matrix-Tree Theorem

For example, let us consider the de Bruijn graph of order 3. Let us consider the matrix $M = D - A$ where D is the diagonal matrix of degrees and A is the adjacency matrix (we do not take loops into account). We have

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

Then each principal minor of order 3 of this matrix has a determinant equal to 2, which is precisely the number of spanning trees with a given root. This theorem appears first in **Borchardt** in 1860.

The existence of a de Bruijn cycle (as we continue to call them) was actually rediscovered several times. In particular, **Good** used the Eulerian graph method to prove the existence of a de Bruijn cycle. Earlier, **Mantel** seems to have been the first one to use, for a prime value of k a linear recurrence relation corresponding to a primitive polynomial of degree n over the field with k elements to define a de Bruijn cycle of length k^n . For example, for $k = 2$ and $n = 3$, using the relation $u_n = u_{n-2} + u_{n-3}$ corresponding to the primitive polynomial $1 + x + x^3$, we obtain from the initial value $u_0 = 0, u_1 = 0, u_2 = 1$ the sequence of length 7 $(0, 0, 1, 0, 1, 1, 1)$. Adding a final 0 gives a de Bruijn cycle. The computation of a de Bruijn cycle in this form is often called a **shift-register sequence**, as a reference to the implementation of the recurrence relation in hardware using a register.

Lyndon words



A **Lyndon word** is a primitive word that is minimal in its conjugacy class.

For example, the list of Lyndon words of length 6 on the alphabet $\{a, b\}$ reads

aaaaab, aaaabb, aaabab, aaabbb, aababb, aabbab, aabbbb, ababbb, abbbbb.

The number of Lyndon words of length n on k symbols is, of course, $M(n, k)$.

Standard factorization

One of the basic properties of Lyndon words is that any Lyndon word x that is not a letter can be written $x = yz$ where y, z are Lyndon words with $y < z$. This factorization is not unique since, for example $(a)(abb) = (aab)(b)$ but there is a unique one, called **standard**, which corresponds to the choice of the longest possible second term.

Lyndon words give rise to commutators through a process of iterated dichotomy, using the notion of standard factorization. For example, the Lyndon word $aababb$ with standard factorization $(a)(ababb)$ gives rise to the commutator $[a, [[a, b], [[a, b], b]]]$. Lyndon words give rise to a basis of the free Lie algebra.

Back to de Bruijn cycles

There is a surprising connection with de Bruijn cycles, that was discovered in 1978 by **Fredericksen** and **Maiorana**. The concatenation in increasing order of Lyndon words of length dividing n is, for any $n \geq 1$, a de Bruijn word (which is actually the first one in lexicographic order and coincides with the word produced by Martin). For example, for $n = 4$, we obtain

a aab aabb ab abbbb

Since Lyndon words can be generated efficiently, this characterization also provides a linear-time algorithm that requires only logarithmic additional space for computing one de Bruijn word.

Path encodings

The enumeration of paths in the plane was studied early in connection with probability theory. The first significant result, a solution to the *ballot problem*, is due to **Bertrand** in 1887 . It says

Suppose that, in a ballot, candidate P scores p votes and candidate Q scores q votes, where $p > q$. The probability that throughout the counting there are always more votes for P than for Q equals $(p - q)/(p + q)$.

The proof of this result consists in enumerating the paths going from the origin to the point (x, y) using up and down diagonal moves and staying in the first quadrant as on Figure 6, with $x = p + q, y = p - q$.

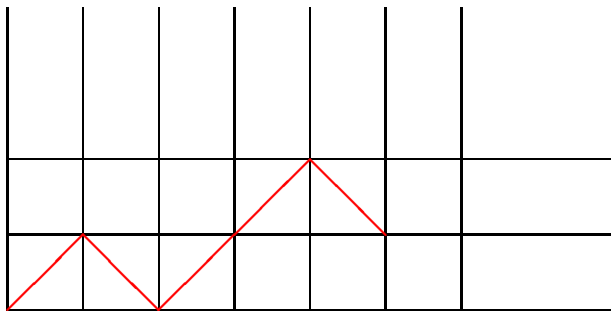


Figure: A path labeled *abaab*

The reflexion principle

The proof uses the *reflexion principle* credited to Désiré André in 1887. This principle says that, given two points A, B in the first quadrant, the number of paths going from A to B which touch or cross the x -axis is equal to the total number of paths from A' to B where A' is symmetrical to A with respect to the x -axis .

Using this lemma, it is easy to prove the ballot theorem. Indeed, let $N_{x,y} = \binom{x}{p}$. The number of paths from the origin to (x, y) that never touch the x -axis (except for the origin) is the same as the number of paths going from the point $(1, 1)$ to the point (x, y) which never touch the x -axis. By the reflexion lemma, this number is

$N_{x-1,y-1} - N_{x-1,y+1} = y/x N_{x,y}$, and thus the probability of the paths staying above the x -axis is y/x .

This result can be used to count the number f_{2n} of paths from the origin to the point $(2n, 0)$, or equivalently the words of length $2n$ of the **Dyck language** D , which is the set of words over $\{a, b\}$ such that both a and b occur n times and all the prefixes have more a 's than b 's. Indeed, such a word ends with a b and their number is also the number of paths from the origin to the point $(2n - 1, 1)$ which stay above the x -axis which, by the ballot theorem, is equal to

$$f_{2n} = \frac{1}{2n - 1} \binom{2n - 1}{n - 1} = \frac{1}{n} \binom{2n - 2}{n - 1}. \quad (2)$$

Let us remark that since $D = aD^*b$, the number of words of D^* of length $2n$ is

$$u_{2n} = \frac{1}{n + 1} \binom{2n}{n}. \quad (3)$$

Catalan numbers

We shall soon meet these numbers, known as **Catalan numbers**, once more, with a completely different proof of this formula using power series. It is of interest to remark that another combinatorial proof of Formula 3 is possible. It goes along the following lines. For a word w on $\{a, b\}$, let $\varphi(w)$ be the difference between the number of occurrences of a and of b . Let L be the set of words w such that $\varphi(w) = -1$ and $\varphi(u) \geq 0$ for any proper prefix u of w . It is easy to see that $L = D^*b$, or in other terms that the words of L are those of D^* with an additional b at the end. The set L is actually the **Lukasiewicz** language used to denote expressions in polish notation with a as a symbol of operand and b as a symbol of operator. Then, as first observed explicitly by Raney, any word w such that $\varphi(w) = -1$ has exactly one conjugate in L , whence Formula 3.

Words and trees



The coding of trees by words has its roots in the need for writing an algebraic term, which is actually a

tree, as a word. This appeared as a necessity at the beginning of formal logic. It was also a concern for formulas of chemistry. It was in this context that the **polish notation**, credited to Lukasiewicz, was introduced. It allows to encode a binary tree without using parenthesis. This is a small combinatorial miracle not always given its real value.

Catalan numbers

The enumeration of trees is a subject with an interesting history which gave rise to a flood of papers in the *Journal de Mathématiques Pures et Appliquées* during the years 1838–39. Catalan is credited for the formula which gives the number of binary trees with n internal vertices as the *Catalan number*

$$T_n = \frac{1}{n+1} \binom{2n}{n}.$$

This is also the number of planar trees with $n + 1$ vertices and the number of (one-sided) Dyck words with $2n$ parentheses. Actually, the formula goes back to Euler.

Binet indicated that Euler had communicated the formula to Segner, although he did not publish a proof. The object of the enumeration at that time was not binary trees, but the equivalent notion of a triangulation of a polygon.

The problem was considered by a number of analysts including Rodrigues, Lamé and Catalan who used all their skills to transform the nonlinear recurrence

$$T_{n+1} = \sum_{i=0}^n T_i T_{n-i} \quad (4)$$

into the simple

$$T_{n+1} = \frac{2n+2}{n+2} T_n.$$

The paper of Binet uses the generating series $T(z) = \sum_{n \geq 0} T_n z^n$ to derive the solution via the formula

$$T(z) = \frac{1}{2}(1 - \sqrt{1 - 4z}).$$

Prüfer code

Another classical result obtained early on the enumeration of trees, is the formula giving the number of labeled trees with n vertices as n^{n-1} . The formula itself was discovered by Borchardt and was known to Cayley.

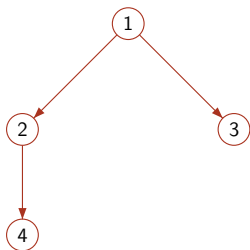


Figure: A tree with Prüfer code 1, 2, 1.

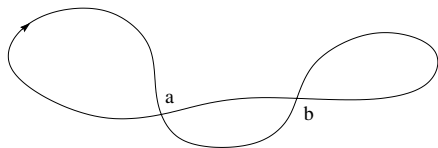
The derivation of this formula through the bijection now known as the **Prüfer code** was obtained independently much later.

Gauss codes



A closed curve in the plane is *normal* if it has only finitely many self-intersections and

these are transverse double points. Label the intersections of such a curve with distinct symbols from an alphabet A . The **Gauss code** of the curve is the word obtained by proceeding along the curve and noting each crossing point label as it is traversed. The word obtained is really a conjugacy class.



For example, the Gauss code of the curve above is $abba$. Gauss codes were introduced by Carl Friedrich Gauss in 1838. He observed that the distance between the two occurrences of each symbol in such a code is always an even integer. This property, however, is not characteristic.

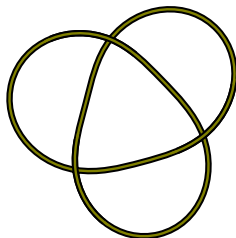


Figure: The trefoil knot