BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 17, Number 1, July 1987

Theory of codes, by Jean Berstel and Dominique Perrin, Pure and Applied Mathematics, Vol. 117, Academic Press, Inc., New York, 1985, xiv + 433 pp., $60.00. ISBN 0-12-093420-5

The research of Shannon on information devices in the late 1940s, and in particular his paper [**Sha**] in 1948, formed a basis of quite extensive studies. The research inspired by his work has spread into several directions which by now are quite independent although there are, or at least should be, some connections. The theory of entropy is one of the directions and forms nowadays a branch in probability theory. The theory of error-correcting (and detecting) codes is another direction, and the theory of variable-length codes, which is the topic of this book, forms the third direction.

It should be made clear that despite their common origin the theory of error-correcting codes and the theory of variable-length codes have very little in common. The former is a beautiful application of commutative algebras, in particular the theory of finite fields, while the latter is connected to noncommutative structures such as free semigroups.

A systematic study of variable-length codes, or briefly codes, was initiated by M. P. Schützenberger in the mid 1950s, cf. [**Sc 1**]. It is not too much to say that without Schützenberger's contributions the theory of codes would not exist in the extent we know it today. Not only many of the major results of the theory are due to him, as is seen from the bibliographical notes of this book, but he also showed the direction in which to continue, via his original approaches in solving problems and via his conjectures.

In the past 30 years the theory of codes has developed into an interesting branch of discrete mathematics which provides a number of nice and deep results, as well as challenging problems. Certainly the theory is connected to many other areas of mathematics. In a broad sense the whole theory can be considered as a part of theoretical computer science, and its connections to areas like combinatorics on words, automata theory, formal language theory and semigroup theory are close and manifold.

It is a surprise to notice that *Theory of codes* is the first book devoted to the field. Of course parts of the theory have been included in other books, but so far it has been typical (and unfortunate) of the whole field that results have been seattered in the literature, and even some very important ones have not been easily available. In addition, many results have earlier been published only in French. So there definitely exists a need for a book on codes, and it is, in my opinion, very important that this book has appeared in such a highly respected series of mathematical textbooks.

In order to discuss the book we recall that a *code* over a finite alphabet A is any subset X of the free monoid A^* generated by A which satisfies:

$$\text{If } x_1 \cdots x_n = y_1 \cdots y_m \quad \text{with } x_i, y_j \in X,$$
$$\text{then } n = m \text{ and } x_i = y_i \quad \text{for } i = 1, \ldots, n,$$

or equivalently, any subset X of A^* which is of the form $h(B)$, where h: $B^* \to A^*$ is an injective morphism (and B need not be finite). So the theory of codes can be considered as the study of injective morphisms of free semigroups. This simple observation does not diminish the attractiveness of the theory. In fact, I would say, its effect is the opposite: It makes many problems easy to state, but not to solve, and thus mathematically challenging.

The basic question to be asked is "When is a given subset X of A^* a code?" This was answered by Sardinas and Patterson, cf. **[SP]**, at a very early stage of the research: Define recursively subsets U_n of A^* as follows:

$$U_0 = X^{-1}X - \{1\},$$
$$U_{n+1} = U_n^{-1}X \cup X^{-1}U_n \quad \text{for } n \geqslant 0,$$

where 1 denotes the identity of A^* and e.g. $X^{-1}X = \{y \mid \exists z \in X\colon\ zy \in X\}$. Then X is a code if and only if 1 is not in U_n for any $n \geqslant 0$. Clearly, in the case of finite X the above criterion yields an algorithm, referred to as the *Sardinas-Patterson algorithm*, for testing whether a given set X is a code. Actually, an algorithm is also obtained in the case when X is *recognizable*, i.e., recognized by a finite automaton. The above Sardinas-Patterson algorithm is simple. However, and this is interesting, a detailed proof of its correctness is surprisingly involved (although elementary).

It is worth noticing that already this basic question can be interpreted and solved in terms of automata theory, a connection which is clearly visible throughout the book. In these terms the problem is whether a given automaton is unambiguous, which is a very natural and well-known problem in automata theory. This approach yields an alternate solution to the Sardinas-Patterson algorithm: A finite set $X \subseteq A^*$ is a code if and only if

$$\bigcup_{\substack{x,y \in X\\ x \neq y}} (xX^* \cap yX^*) = \varnothing .$$

Hence, the problem is reduced to the emptiness problem of finite automata. Now the proof of the existence of an algorithm is easy, but the algorithm itself is extremely ineffective.

The main goal of the theory is to try to give structural characterizations for families of codes as well as to provide effective methods for constructing all codes of certain types. In general, these problems are still very much open. In the case of two-element codes such a characterization is well known and easy to obtain: The set $X = \{x, y\} \subseteq A^*$ is a code if and only if x and y are not powers of a word which, in turn, holds if and only if $xy \neq yx$. This is a special case of a more general result known as the Defect Theorem. On the other hand, already in the case of three-element codes no such characterization is known!

It follows that it is reasonable to study certain special classes of codes. For *prefix codes*, i.e., for codes $X \subseteq A^*$ satisfying $XA^+ \cap X = \varnothing$, where $A^+ = A^* - \{1\}$, a structural characterization is obvious since such codes can be presented as trees. For finite *biprefix codes* (which are "prefix codes both from left to right and from right to left") a structural characterization can be achieved as well; however, it is much more complicated and, in fact, forms one of the highlights of the whole theory. This will be discussed more in a moment.

A well-motived family of codes is that of maximal codes: A code $X \subseteq A^*$ is *maximal* if it is not properly included in any code over A. The study of maximal codes (of a certain type) is one of the main trends in this book. There exist at least two reasons to study maximal codes. Firstly, since each code is included in a maximal one (cf. below), if something is known about the structure of maximal codes of a certain type, this also tells something about all codes of the same type. Secondly, from the point of view of applications the maximality is an important property: It tells that the full capacity of a code is utilized.

There is a surprisingly simple criterion, due to Schützenberger, cf. [**Sc 1**], to test whether a given finite code $X \subseteq A^*$ (but not a set) is a maximal code. Let $\pi : A^* \to (\mathbf{R}_+, \cdot)$, where $(\mathbf{R}_+, \cdot)$ denotes the multiplicative monoid of nonnegative real numbers, be a morphism satisfying $\pi(A)\ (= \Sigma_{a \in A} \pi(a)) = 1$. Then X is a maximal code if and only if

$$\pi(X)\ (= \Sigma_{x \in X} \pi(x)) = 1.$$

In one direction the proof is very easy, while in the other direction it is based on another characterization result of finite maximal codes which is as follows: A finite code $X \subseteq A^*$ is maximal if and only if it is *complete*, which by definition means that, for each word x in A^*, $A^*xA^* \cap X^* \neq \varnothing$, or equivalently, X^* meets every two-sided ideal of A^*.

The above criterion does not hold for all codes, a counterexample being the prefix code $X = \{u \in \{a, b\}^+ \mid |u|_a = |u|_b$ and for each proper prefix v of u $|v|_a \neq |v|_b\}$, which is maximal (and complete) but satisfies the condition $\pi(X) = 1$ only for *one* morphism π of the above form, namely for the morphism π such that $\pi(a) = \pi(b) = \frac{1}{2}$. (Here $|u|_a$, for example, denotes the number of occurrences of a letter a in u.) On the other hand, the assumption of X being finite can be weakened quite a lot. Indeed, the criterion holds true for all recognizable and also for all thin codes, which are defined as follows: A code $X \subseteq A^*$ is *thin* if there exists a word $x \in A^*$ such that $A^*xA^* \cap X = \varnothing$, or equivalently, $A^* - (A^*)^{-1}X(A^*)^{-1} \neq \varnothing$. Of course, each finite code is thin, and it is not too difficult to see that also each recognizable code is thin.

This poses a question on the level of generality at which results should be presented in a book. The above is not the only example of a result which has been first discovered for finite codes, and later observed not only to be true but also provable with almost no extra effort for recognizable codes as well. This suggests that at least recognizable codes should be considered in the general presentation of the theory, and this point is strongly adopted in the book. On the other hand, moving from recognizable codes to more general codes often leads to complicated considerations and difficulties. However, thin codes are very suitable in the algebraic treatment of the theory. With this in mind the level of thin codes is the level of generality most often chosen in this book.

Let us go back to maximal codes for a moment. It is an obvious consequence of Zorn's Lemma that each code is included in a maximal one. However, this does not say anything about the problem of how to extend a given code to a maximal one. For example, can every finite code be extended to a finite maximal code? This question was answered negatively by Restivo [**Res**]; probably the simplest counterexample is the code $\{a^5, ab, b, ba^2\}$. Quite recently a *method of completing* a finite code $X \subseteq A^*$, that is, extending it to a complete code, was introduced by Ehrenfeucht and Rozenberg [**ER**]. This method goes as follows. Let $X \subseteq A^*$ be a noncomplete finite code. Then there exists, as is not difficult to see, a word $w \in A^*$ such that

(i) w is unbordered, i.e., $wA^+ \cap A^+w = w \cup wA^*w$, and

(ii) $A^*wA^* \cap X^* = \varnothing$.

Now, define

$$U = A^* - X^* - A^*wA^*$$

and set

$$Y = X \cup y(Uy)^*.$$

Then Y is recognizable, and it is not very difficult to conclude that Y is a complete code. Consequently, X can be effectively completed to a recognizable complete code, and hence also to a recognizable maximal code. Here again it is worth noticing that it does not make any difference whether X is finite or recognizable. In any case, Y is recognizable and maximal. In this particular problem the proof goes without any changes for thin codes as well—only the result is not guaranteed to be recognizable, of course.

As an evidence of the up-to-dateness of this book it can be mentioned that the above Ehrenfeucht-Rozenberg method appeared (in the generalized form described above) in this book earlier than in a scientific journal!

Although we have been brief in our presentation, it should be clear by now that the combinatorics of words constitutes an essential part of the theory of codes. Indeed, all the results discussed so far are proved by such methods. We mention two more such examples.

As a generalization of prefix codes we define codes having a finite deciphering delay $d \geqslant 0$. A set $X \subseteq A^*$ (not necessarily a code) is said to have a *finite deciphering delay d* if it satisfies:

$$\forall x,\ x' \in X, \forall y \in X^d, \forall u \in A^*: \quad xyu \in x'X^* \Rightarrow x = x'.$$

It follows immediately that X is a code. The importance of this notion comes from the fact that this property allows an easy decoding: If X is finite it is enough to have a finite look-ahead in order to decode (from left to right) a coded message. Now, an interesting result, again due to Schützenberger, says that any finite maximal code with a finite deciphering delay is a prefix code [**Sc 2**].

In the second example we return to biprefix codes. There exists a remarkable characterization of maximal finite biprefix codes due to studies initiated by Schützenberger and completed by Césari, cf. [**Sc 3**] and [**C**]. Let $X \subseteq A^*$ be a maximal finite biprefix code. A *parse* of a word $w \in A^*$ is a triple (v, x, u), where $w = vxu$ with $v \in A^* - A^*X$, $x \in X^*$, and $u \in A^* - XA^*$. Then it can be verified that the number of different parses a word w may have is bounded (independently of w), and the smallest upper bound is called the *degree* of X. Further the *kernel* of X is the set $X \cap (A^+)^{-1}X(A^+)^{-1}$, that is to say, the set of those words in X which occur also as internal factors in X. Thus we have associated each maximal finite biprefix code with two parameters, its degree and kernel. The surprising result is that these two parameters uniquely define X! Moreover, a transformation, called an *internal transformation*, can be defined on the family of finite maximal biprefix codes over A having a given degree n in such a way that all biprefix codes in this family can be obtained from the uniform code A^n via these transformations. In addition, these families are finite, although their cardinality grows very rapidly when d grows: In a binary A there exist 5056783 such codes having degree 5!

The above theory of finite biprefix codes is probably mathematically the most beautiful part within the theory of codes. In this book it is very clearly presented, and also partially reworked. Moreover, again the results are formulated (whenever possible) not only for finite codes but also for thin codes.

All the results discussed above have been established by using, in one way or another, combinatorial methods and combinatorial properties of words. There exists, however, another approach which leads to important results of codes. This method is more algebraic, and its starting point is the notion of an unambiguous automaton. Clearly, if $X \subseteq A^*$ is a code, then X^* is unambiguous. Moreover, it can be recognized by a (not necessarily finite) automaton, most naturally using a so-called *flower automaton* $\mathscr{A}_D^*(X)$ which, by definition, is as follows: For each word x in X, if $x = a_1 \cdots a_n$, with $a_i \in A$, then there exists in $\mathscr{A}_D^*(X)$ a distinct path of the form $q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} \cdots q_{n-1} \xrightarrow{a_n} q_0$, where q_0 is the only initial and final state. Of course, this automaton is unambiguous.

Associating a code $X \subseteq A^*$ with a monoid, namely with the monoid of state transformations of $\mathscr{A}_D^*(X)$, algebraic methods can be employed in studying codes. In order to be a bit more concrete let us consider the case of *very thin codes* $X \subseteq A^*$ which satisfy, by definition, the condition that there exists a word $x \in X^*$ such that it is not a factor of X, or equivalently, $X^* \cap (A^* - (A^*)^{-1}X(A^*)^{-1}) \neq \varnothing$. It is an immediate consequence of the definitions that every very thin code is thin, while the reverse is not true in general. However, it can be shown that for complete codes these two notions coincide. Now, it turns out that very thin codes satisfy a finiteness condition which allows us to associate such a code with a group, namely with the Suschkewitch group of the monoid described above. These lines provide an explanation why thin codes are so central in this book.

After finding the above connection between the theory of codes and classical algebraic theories, one is not surprised that the new theory can benefit from the older ones. This book contains many such examples. However, we state here only two results which are proved by using this approach. As we hinted earlier the implication "each thin complete code is maximal" can be proved by using probabilistic measures. It can also be proved, in a completely different way, using the algebraic approach. The second result we want to mention here yields a presentation for semaphore codes, that is, for codes $X \subseteq A^*$ which are of the form $X = A^*S - A^*SA^+$ with $S \subseteq A^*$. It is not difficult to conclude that these codes are maximal prefix codes, and also thin. Further let us call a maximal prefix code $X \subseteq A^*$ *synchronous* if there exists a word $x \in A^+$ such that $A^*x \subseteq X^*$. Now, a deep result of Schützenberger states that each semaphore code $X \subseteq A^*$ can be expressed in the form $X = Z^d$, where Z is a synchronous semaphore code and $d \geqslant 1$. The algebraic approach yields an explanation of d: It equals the degree of the group of X.

As a final matter we return to the question of characterizing all finite codes. We have already emphasized that this question is still very much open. As a step in this direction a notion of a factorizing code has been introduced. A code $X \subseteq A^*$ is said to be *factorizing* if there exist subsets P and Q of A^* such that each word $w \in A^*$ can be expressed uniquely in the form

$$w = qxp \quad \text{with } q \in Q,\ p \in P, \text{ and } x \in X^*.$$

This definition directly leads to formal power series over noncommuting variables. Indeed, associating each unambiguous set $Y \subseteq A^*$ with the power series $\underset{\cdot}{Y} = \sum_{y \in Y} y$, we conclude that $X \subseteq A^*$ is factorizing if and only if there exist subsets P and Q of A^* such that

$$(1) \qquad \underset{\cdot}{A}^* = \underset{\cdot}{Q}\underset{\cdot}{X}^*\underset{\cdot}{P}.$$

Taking inverses (remember that $\underset{\cdot}{Y}^* = 1/1 - \underset{\cdot}{Y}$), we can rewrite (1) as

$$1 - \underset{\cdot}{X} = \underset{\cdot}{P}(1 - \underset{\cdot}{A})\underset{\cdot}{Q}.$$

It follows directly from (1) that if a finite code $X \subseteq A^*$ is factorizing with P and Q finite, then X is complete and hence maximal. Also, as a consequence of a theorem of Schützenberger mentioned below, the converse holds in the following sense: If a maximal finite code is factorizing, then P and Q are finite. However, it is an open question—and this is probably the most important open question within the whole theory of codes—whether each maximal finite code is factorizing. An example of Shor [**Sho**] shows that not all finite codes are factorizing.

There exist two important results in the direction of an affirmative answer of the above question. The first is again due to Schützenberger [**Sc 4**], and is as follows. For a formal power series $\underset{\cdot}{Y}$ over noncommuting variables let us denote by $\underset{\cdot\cdot}{Y}$ the corresponding power series over commuting variables. The theorem says that if $X \subseteq A^*$ is a finite maximal code, then $1 - \underset{\cdot\cdot}{A}$ divides $1 - \underset{\cdot\cdot}{X}$. The proof is long, but can be found in the book. The second result, due to Reutenauer, cf. [**Reu**] or [**BR**], is even more closely related to the above factorizing question, but is unfortunately not included in the book. The result says that for each maximal finite code $X \subseteq A^*$ there exist polynomials $\underset{\cdot}{P}$ and $\underset{\cdot}{Q}$ with *integer* coefficients such that $1 - \underset{\cdot}{X} = \underset{\cdot}{P}(1 - \underset{\cdot}{A})\underset{\cdot}{Q}$.

Besides being interesting in its own right the above question has important connections. Indeed, as is relatively easy to see, each factorizing code is *commutatively equivalent to a prefix code*—a property which is certainly interesting from the point of view of information theory. Hence, if every finite maximal code were factorizing it would also be commutatively equivalent to a prefix code. Again the example of Shor shows that a finite code need not be commutatively equivalent to a prefix.

So far we have talked about several problems which, according to our estimate, are among the most interesting in the theory of codes, and which are broadly discussed in this book. By no means have we tried to be exhaustive, nor have we gone into proofs of theorems. On the other hand, we have tried to give a flavor of the ideas and techniques used in the book. In particular, it should be noted that the theory has connections to many different parts of mathematics.

The amount of material covered by this book is very large. There is no way to include all of it in a one-year university course. The book is unique in the sense that a large portion of its contents cannot be found in any other existing book. However, it is not a collection of distinct important results: the authors have really reworked many theorems in order to create a uniform theory. The presentation is self-contained, although basic knowledge on automata theory makes the reading easier. The book contains detailed bibliographical notes as well as many exercises on different levels.

In my opinion this book is an excellent example of the phenomenon that a simple mathematical notion—an injective morphism of free semigroups—may lead to an interesting theory with challenging problems. Anybody interested in such things should be interested in *Theory of codes*.

References

[**BR**] J. Berstel and C. Reutenauer, *Rational series and their languages*, EATCS Monographs on Theoretical Computer Science, Springer-Verlag, Berlin, to appear.

[**C**] Y. Césari, *Propriétés combinatoires des codes biprefix*, in: D. Perrin (ed.), *Théorie des Codes*, LITP, Paris, 1979, pp. 20–46.

[**ER**] A. Ehrenfeucht and G. Rozenberg, *Each regular code is included in a regular maximal code*, RAIRO Theor. Informatics and Appl. **20** (1986), 89–96.

[**Res**] A. Restivo, *On codes having no finite completions*, Discret. Math. **17** (1977), 309–316.

[**Reu**] C. Reutenauer, *Noncommutative factorizations of variable length codes*, J. Pure Appl. Alg. **36** (1985), 167–186.

[**SP**] A. A. Sardinas and C. W. Patterson, *A necessary and sufficient condition for the unique decomposition of coded messages*, IRE Intern. Conv. Rec. **8** (1953), 104–108.

[**Sha**] C. E. Shannon, *The mathematical theory of communications*, Bell. Syst. Techn. J. **27** (1948), 379–423, 623–656.

[**Sho**] P. W. Shor, *A counterexample to triangle conjecture*, J. Combin. Theory Ser. A **38** (1985), 110–112.

[**Sc 1**] M. P. Schützenberger, *Une théorie algébrique du codage*, Séminaire Dubreil-Pisot, 1955–1956, Exposé No. 15.

[**Sc 2**] ______, *On a question concerning certain free monoids*, J. Combin. Theory **1** (1966), 417–422.

[**Sc 3**] ______, *On a special case of recurrent events*, Ann. Math. Statist. **32** (1961), 1201–1213.

[**Sc 4**] ______, *Sur certains sous-monoïdes libres*, Bull. Soc. Math. France **93** (1965), 209–223.

Juhani Karhumäki