This paper is an overview of results on the subword complexity of various classes of languages obtained by iterating a mapping, and more precisely on the asymptotic behaviour of this complexity.

In the first section we consider iterated morphisms, that is D0L languages. In this area the subword complexity falls into one of five classes, and class membership can be determined. The second section is devoted to homomorphic images of D0L languages, that is HD0L languages. In this case there is an infinite hierarchy of complexity classes, and many questions remain unsolved. In section 3 we are concerned with the effect of various combinatorial restrictions (square-freeness, constant distribution) on the subword complexity. Finally in section 4 we give some results concerning iterated sequential mappings (d.g.s.m.).

Most proofs are omitted if they appear somewhere else, or only sketched. The reader is assumed to be familiar with basic D0L language theory (see [RS]).

1. Subword complexity and iterated morphisms

The subword complexity (or complexity for short) of a language \( L \) is the function \( f_L \) where \( f_L(n) \) is the number of distinct subwords of \( L \) with length \( n \). An infinite word \( S \) is a mapping from the non-negative integers to some finite alphabet. One can define the complexity of \( S \) by the complexity of the language of its prefixes.

A D0L system is a 3-tuple \( \langle X, g, \alpha \rangle \) where \( X \) is a finite alphabet, \( g : X^* \rightarrow X^* \) a morphism and \( \alpha \in X^+ \) the axiom. The D0L language generated by \( D \) is

\[
L = \{g^i(\alpha), \; i \geq 0\}.
\]

---

(1) J.-J. PANSIOT, Université Louis-Pasteur, Centre de Calcul de l’Esplanade, 7, rue René Descartes, 67084 Strasbourg Cedex, France
The complexity of DOL languages has been studied by Ehrenfeucht, Lee and Rozenberg [ELR]. The main result is that the complexity depends on the type of the morphism \( g \). A morphism \( g \) is \( \epsilon \)-free or propagating if \( g(x) \neq \epsilon \) for all \( x \). It is (everywhere) growing if \( |g(x)| > 1 \) for all \( x \). Finally it is uniform, of modulus \( m \), if \( |g(x)| = m \) for all \( x \). The corresponding classes of DOL languages are denoted respectively PDOL, GDOL, UDOL.

**Theorem 1.1 [ELR].**

a) A DOL language has complexity at most \( cn^2 \) for some \( c \), and there exist PDOL languages with complexity greater than \( dn^2 \) for some \( d > 0 \).

b) A GDOL language has complexity at most \( cn \log n \) for some \( c \), and there exist GDOL languages with complexity greater than \( dn \log n \) for some \( d > 0 \).

c) A UDOL language has complexity at most \( cn \) for some \( c \), and there exist UDOL languages with complexity greater than \( dn \) for some \( d > 0 \).

It is also known that there is no sublinear complexity besides constant complexity (periodic languages)[R]. From these results one can ask whether or not there are other complexity classes, and if it is possible to find lower bounds. These questions are easier to solve in the framework of infinite words generated by iterated morphisms. Let \( D = \langle X, g, x_0 \rangle \) be a DOL system such that \( g \) is prolongeable in \( x_0 \), that is \( g(x_0) = x_0 u \). We have

\[
g^i(x_0) = x_0 u g(u) \ldots g^{i-1}(u)
\]

So \( g^i(x_0) \) is a prefix of \( g^{i+1}(x_0) \), and \( D \) defines a unique word, in general infinite, denoted by

\[
g^\omega(x_0) = x_0 u g(u) \ldots g^i(u) \ldots
\]

In order to get more precise bounds one has to introduce new types of morphisms. The important criterion is the relative order of growth of letters, that is the function \( |g^n(x)| \), for \( n \geq 0 \). This function grows asymptotically as \( n^{a x b^n} \) for some integer \( a_x \) and number \( b_x \geq 1 \).

A morphism is quasi-uniform if all letters have the same order of growth, \( b^n \), for some \( b > 1 \). Note that a uniform morphism of modulus \( m \) is quasi-uniform, with \( m = b \). A morphism is polynomially diverging if there exists a number \( b > 1 \) such that the order of growth of any letter \( x \) is \( n^{a_x b^n} \) and \( a_x \geq 1 \) for some \( x \). A morphism is exponentially diverging if every letter \( x \) has order of growth \( n^{a_x b^n} \) with \( b_x > 1 \), and not all \( b_x \) are equal. Note that all these morphisms are growing. In [P84b] the next theorem is established.

**Theorem 1.2. —** Let \( S = g^\omega(x_0) \) be an infinite word, not ultimately periodic, with complexity \( f_S \), then

a) If \( g \) is growing \( f_S(n) \) grows as \( n \) (resp. \( n \log \log n \), \( n \log n \)) if and only if \( g \) is quasi-uniform (resp. polynomially, exponentially diverging).

b) If \( g \) is not growing then either \( f_S(n) \) grows as \( n^2 \), or \( S = h(g'(\omega(x_0))) \) where \( h \) is \( \epsilon \)-free and \( g' \) growing. In this case the complexity of \( S \) and \( S' \) grows as \( n, n\log\log n \) or \( n\log n \), depending on the type of \( g' \).

From the proof of this theorem one can decide in which complexity class falls \( S \). The next step is to generalize this result to arbitrary D0L languages. However, the situation is a little more complicated. In an infinite word, two letters occurring infinitely often are infinitely mixed. This is not the case for D0L languages such as

\[
L = \{a^{2^n}b^{3^n}, \ n \geq 1\}.
\]

Note that the morphism \( g \) generating \( L \), \( g : a \rightarrow a^2, \ b \rightarrow b^3 \) is exponentially diverging but \( L \) has linear complexity.

In the following we assume that \( L \) is an infinite D0L language generated by \( D = < X, g, \alpha > \), and \( L \) is not ultimately periodic.

**REMARK 1.** — Let \( x_0 \) be a new letter, \( X' = X \cup \{x_0\} \) and define

\[
g' : x_0 \rightarrow x_0\alpha, \ z \rightarrow g(z) \text{ for } z \in X.
\]

We have

\[
S = g'(\omega(x_0)) = x_0\alpha g(\alpha) \ldots g'(\alpha) \ldots
\]

Therefore all subwords of \( L \) are also subwords of \( S \) and morphisms \( g \) and \( g' \) are of the same type.

**LEMMA 1.3.** — If \( g \) is quasi-uniform then \( L \) has linear complexity.

**Proof.** — (Sketch). The upper bound is obtained by Theorem 1.2 and Remark 1, the lower bound by non periodicity.

**LEMMA 1.4.** — If \( g \) is polynomially diverging then the complexity of \( L \) grows as \( n\log\log n \).

**Proof.** — (Sketch) The upper bound is given by Theorem 1.2 and Remark 1. The lower bound comes from the fact that if a letter has order of growth \( n^{a_x}b^n \) with \( a_x \geq 1 \), then images of \( x \) contain letters with order of growth \( b^n \). Therefore iterating \( g \) on \( x \) generates an infinite word (possibly bi-infinite) with complexity \( n\log\log n \).

A word \( xux' \) is mixing if and only if the order of growth of letters \( x \) and \( x' \) is at least \( a^n \); and each letter of the non empty word \( u \) has order of growth at most \( b^n \) with \( b < a \).

**LEMMA 1.5.** — If \( g \) is exponentially diverging and \( L \) contains a mixing subword, then its complexity grows as \( n\log n \). If it contains no mixing subword, then its complexity grows as \( n\log\log n \) (resp. \( n \)) if there are (resp. there are no) letters with order of growth \( n^ab^n \) and \( a \geq 1 \).
Proof. — (Sketch) The first part of the lemma is similar to the case of infinite words. These always contain mixing subwords. If there is no mixing subword, then \( L \) is essentially of the form
\[
\{ g^i(\alpha_1)g^i(\alpha_2)\ldots g^i(\alpha_n), \ i \geq 0 \}
\]
where each \( L_j = \{ g^i(\alpha_j), \ i \geq 0 \} \) is polynomially diverging or quasi-uniform. Moreover the order of growth of letters from \( L_j \) is of the form \( n^{a_j}b_{j_0} \) with
\[
b_1 < b_2 < \ldots < b_{j_0} > \ldots > b_{n-1} > b_n.
\]
In this case the complexity of \( L \) grows as the maximum of the complexities of the \( L_j \)'s.

A DOL language \( L \) contains a linear sequence if there exists a sequence of subwords of the form \( \alpha_n\beta^n+d\gamma_n \). Where \( \beta \) is not empty, the lengths of \( \alpha_n \) and \( \gamma_n \) grow at least linearly in \( n \), and \( \beta \) is not a suffix of \( \alpha_n \) nor a prefix of \( \gamma_n \). Note that a DOL language with a linear sequence cannot be growing.

**Lemma 1.6.** — A DOL language \( L \) has quadratic complexity if and only if it contains a linear sequence. If it contains no linear sequence then its complexity grows as the complexity of a growing DOL language.

Proof. — (Sketch) Again the proof of the first part is similar to the case of infinite words. If there exist no linear sequence, then either \( L = h(L') \) where \( h \) is \( \epsilon \)-free and \( L' \) a growing DOL-language, or it is a concatenation of simpler DOL languages, as in the proof of Lemma 1.5.

**Remark 2.** — For a DOL language one can decide the existence of mixing factor and of linear sequence.

**Theorem 1.7.** — The complexity of a DOL language \( L \) grows as \( 1, n, n\log\log n, n\log n \) or \( n^2 \). Moreover one can decide which order of growth apply to the complexity of \( L \).

The complexity of DOL-languages is rather well understood. We will see in the next section that this is less true for HDOL languages.

**2. Homomorphic images**

An HDOL system is a 5-tuple \( D = < X, Y, g, h, \alpha > \) where \( < X, g, \alpha > \) is a DOL system, and \( h : X^* \to Y^* \) a morphism. The HDOL language generated by \( D \) is
\[
\{ h(g^i(\alpha)), i \geq 0 \}.
\]
A tag system (see [C]) is an HDOL system where \( \alpha \) is reduced to a letter \( x_0 \), \( g \) is prolongable in \( x_0 \), and \( h \) is literal (that is \( h(x) \in Y \) for all \( x \in X \)). A tag system generates an infinite word \( h(g^\omega(x_0)) \).
From [P83] one can observe that the classes of quasi-uniform, polynomially and exponentially diverging tag systems are closed by \( \varepsilon \)-free morphisms. The class of arbitrary tag systems is closed by arbitrary morphisms. Since a literal morphism cannot increase the complexity, and since Remark 1 can be extended to HD0L languages, we have the following theorem (parts of it were proved directly for HD0L languages in [ER82]).

**Theorem 2.1.** — Let \( L \) be the language generated by the HD0L system \( < X, Y, g, h, \alpha > \). Then

a) If \( h \) is \( \varepsilon \)-free and \( g \) quasi-uniform then the complexity of \( L \) is bounded by \( cn \) for some \( c \).

b) If \( h \) is \( \varepsilon \)-free and \( g \) polynomially diverging then the complexity of \( L \) is bounded by \( cn \log \log n \) for some \( c \).

c) If \( h \) is \( \varepsilon \)-free and \( g \) exponentially diverging then the complexity of \( L \) is bounded by \( cn \log n \) for some \( c \).

d) If \( h \) and \( g \) are arbitrary then the complexity of \( L \) is bounded by \( cn^2 \) for some \( c \).

Since these upper bounds are identical to upper bounds for D0L languages, it is tempting to conjecture that HD0L languages fall into the same five complexity classes. However this is not true, as can be seen with the following HD0L system:

\[
\begin{align*}
  x_0 &\to x_0az \\
  z &\to xz \\
  a &\to a \\
  g : y &\to yz \\
  h : y &\to 0
\end{align*}
\]

We have

\[
g^n(x_0) = x_0azg(a)g(x)\ldots g^{n-1}(a)g^{n-1}(x).
\]

Since

\[
|g^n(x)| = (n + 1)(n + 2)/2, \quad h(g^n(x)) = 0^{(n+1)(n+2)/2}.
\]

Similarly \( g^n(a) = a \) and \( h(g^n(a)) = 1 \). Therefore

\[
h(g^n(x_0)) = 11010^310^610^10\ldots10^{(n+1)/2}1.
\]

The complexity \( f_L(n) \) of this language verifies

\[
f_L(n + 1) = f_L(n) + b(n)
\]

where \( b(n) \) is the number of words \( u \) of length \( n \) such that both \( u0 \) and \( u1 \) are subwords of \( L \). These subwords are of the form \( 0^n \) or \( 0^i10^j \), \( i + j + 1 = n \), \( j = (t + 1)(t + 2)/2, i \leq t(t + 1)/2 \). Hence the relation

\[
(\sqrt{2} - 1)\sqrt{n} \leq b(n) \leq \sqrt{2}\sqrt{n} + 2.
\]
From this it follows that there exist constants $c_1$ and $c_2$, $0 < c_1 < c_2$ with

$$c_1 n^{\sqrt{n}} \leq f_L(n) \leq c_2 n^{\sqrt{n}}.$$  

This result can be generalized in the following lemma.

**Lemma 2.2.** There exists an HD0L language $L_k = \{h(g^i(x_0)), i \geq 0\}$ with complexity growing as $n^{\sqrt{n}}$ for all $k \geq 0$. Moreover one can choose $g$ $\epsilon$-free and $h$ literal.

**Proof.** One can construct an HD0L language of the form

$$L_k = \{10^{p(1)}10^{p(2)}\ldots10^{p(n)}1, n \geq 0\}$$

where $p$ is a positive polynomial of degree $k$. The complexity of $L_k$ grows as $n^{\sqrt{n}}$.

So contrary to the case of DOL languages, there is an infinite hierarchy of complexity classes for HD0L languages. This raises the (open) questions of finding all complexity classes, and characterizing members of each class.

### 3. Subword complexity and combinatorial restrictions

A combinatorial restriction on words is a restriction on the sequences of letters or subwords that may appear in a word. Among these restrictions are square-freeness and constant distribution. A language is square-free if it contains no subword of the form $uu$ for some non empty word $u$. Similarly one can define $m^{th}$-power-free languages for $m > 1$. This can be extended to fractional powers: $u^t$ for $t$ rational is the prefix of $u^+$ verifying $|u^t| = t|u|$.

In [ER81a] it was shown that square-free DOL languages have at most complexity $cn \log n$ and at least $dn$. This result was strengthened in [ER83b], [ER83c] as follows:

**Theorem 3.1.** An $m^{th}$-power-free DOL language has complexity growing at least as $cn$ and at most as $dn \log n$. Moreover $m^{th}$-power-free languages ($m \geq 3$) on 2 letter alphabets and square-free language on 3 letter alphabets have linear complexity. Finally there exist third-power-free DOL languages on 3 letter alphabets and square-free languages on 4 letter alphabets with complexity growing as $n \log n$.

A language $L$ has a constant distribution if all sufficiently long subwords contain the same letters. The effect of constant distribution on the complexity is strong as shown in the following theorem [ER81b].

**Theorem 3.2.** A DOL language with constant distribution has complexity bounded by $cn$ for some $c$.

This property is not true anymore for HD0L languages.
Lemma 3.3. — There exist HD0L languages with constant distribution and quadratic complexity.

Proof. — The HD0L-system $< X, Y, g, h, x_0 >$

$\begin{align*}
x_0 &\rightarrow x_01 & x_0 &\rightarrow 000 \\
g : 1 &\rightarrow 01 & h : 1 &\rightarrow 010 \\
0 &\rightarrow 0 & 0 &\rightarrow 011
\end{align*}$

generates the language $h(L)$ where

$L = \{g^i(x_0), i \geq 0\} = \{x_01010^21\ldots10^i1, i \geq 0\}$

The language $L$ has quadratic complexity (cf Lemma 1.5) moreover $h$ is uniform and injective, hence the complexity of $h(L)$ is the same as the complexity of $L$ up to a constant factor.

In contrast, properties on the complexity of $m$th-power-free D0L languages can be partially extended to HD0L languages.

Lemma 3.4. — If $L = \{h(g^i(\alpha)), i \geq 0\}$ is an $m$th-power-free HD0L language, then its complexity grows at most as $cn \log n$ for some $c$.

Proof. — Assume $L$ has a complexity growing faster than $cn \log n$ for any $c$. Consider the infinite word

$S = \#h(\alpha)\#h(g(\alpha))\#\cdots\#h(g^i(\alpha))\#$

This word can be constructed using a tag system, that is $S = h'(g(x_0))$ where $h'$ is a coding. Therefore the complexity of $S' = g'(x_0)$ is at least the complexity of $S$ which is greater than the complexity of $L$. By Lemma 1.6, $S'$ must contain a linear sequence, that is arbitrarily large powers of some non empty word $u$. Obviously $S$ contains arbitrarily large powers of some non empty word $v = h'(u)$. Since $L$ is $m$th-power-free, $h(g^i(\alpha))$ does not contain $v^n$ for $n \geq m$. Hence $v^n$ must contain at least $n$ symbols #. From this it follows that $n$ values of $h(g^i(\alpha))$ for $i \geq 0$ must be equal, for arbitrarily large $n$. It can be shown that this implies the ultimate periodicity of the sequence, hence $L$ is finite, a contradiction.

The fact that D0L languages with very small repetitions have as few distinct subwords as possible is surprising at first. However this can be explained.

Lemma 3.5. — Let $L$ be an infinite DOL language on $k$ letters containing at most $t$th powers ($t > 1$), such that no infinite DOL language on $k-1$ letters has this property. Then $L$ has linear complexity.

Proof. — (Sketch) Every growing letter must produce all other letters (otherwise one can construct an infinite language with less letters). It follows that $L$ has a constant distribution.
From this result one get that \( m^{th} \)-power-free languages on 2 letters, \( 7/4^{th} \)-power-free or square-free languages on 3 letters, or \( 7/5^{th} \)-power-free languages on 4 letters (see [P84a]) have linear complexity.

Another restriction has been studied by Ehrenfeucht and Rozenberg, that is locally catenative D0L languages [ER 83a]. This class has also linear complexity.

4. Iterated d.g.s.m.

Since D0L and HD0L languages have at most quadratic complexities, but arbitrary languages may have exponential complexities, one might ask for a generating device more powerful than iterated morphisms. This is the object of this last section.

An iterated deterministic generalized sequential mapping (i.d.g.s.m.) is a 6-tuple \( D =< X, S, \delta, \sigma, s_0, u_0 > \) where \( X \) is a finite alphabet, \( S \) a finite set of states, \( \delta : S \times X \rightarrow S \) the next state function, \( \sigma : S \times X \rightarrow X^* \) the output function, \( s_0 \) the initial state and \( u_0 \) the axiom. The language generated by \( D \) is

\[
\{ u_i = \sigma(s_0, u_{i-1}), \ i \geq 0 \} \cup \{ u_0 \}.
\]

One can define the type of an i.d.g.s.m. (uniform, growing, \( \epsilon \)-free) using the length of \( \sigma(s, x) \), in the same way as for morphisms. An i.d.g.s.m. is prolongeable in \( x_0 \) if \( \sigma(s_0, x_0) \) starts with \( x_0 \). In this case it defines a unique word, in general infinite. In [B] the following theorem is proved.

**Theorem 4.1.** — Let \( S \) be an infinite word generated by iterating a uniform d.g.s.m. of modulus \( m \), on \( k \) letters, with \( s \) states. Then the complexity \( f_S \) of \( S \) verifies

\[
f_S(n) \leq k^2 s m n^{1 + (\log s / \log m)}
\]

Moreover one can construct a uniform d.g.s.m. such that

\[
f_S(n) \geq \frac{1}{m} n^{1 + (\log s / \log m)}.
\]

This theorem can be extended to growing d.g.s.m. as follows [P84c].

**Theorem 4.2.** — Let \( S \) be an infinite word generated by the growing i.d.g.s.m.

\[
D =< X, S, \delta, \sigma, s_0, x_0 >.
\]

The subword complexity of \( S \) verifies

\[
f_S(n) \leq 2k^2 s m_2 n^{\log s m_2 / \log m_1}
\]

where \( m_1 \) and \( m_2 \) are lower and upper bounds for the length of \( \sigma(s, x) \) for \( s \in S \) and \( x \in X \).
Therefore a growing d.g.s.m. has a polynomial complexity, but of arbitrary large
degree according to Theorem 4.1. It is easy to construct an ε-free i.d.g.s.m. gen-
erating exponential complexity (see [P84c]). So there is a big jump in complexity
between growing and non-growing i.d.g.s.m.

REFERENCES

[B] BLEUZEN (N.). — Suites points fixes de transductions uniformes, Dépar-
tement de Mathématiques-Informatique, Université de Marseille, 1984, to
appear.

p. 64–192.

[ELR] EHRENFEUCHT (A.), LEE (K.) and ROZENBERG (G.). — Subword com-
plexities of various classes of deterministic developmental languages without
interactions, Theoretical Computer Science, t. 1, 1975, p. 59–75.

[ER81a] EHRENFEUCHT (A.) and ROZENBERG (G.). — On the subword com-
plexity of square-free DOL-languages, Theoretical Computer Science, t. 16,

[ER81b] EHRENFEUCHT (A.) and ROZENBERG (G.). — On the subword com-
plexity of DOL languages with a constant distribution, Information Process-

[ER82] EHRENFEUCHT (A.) and ROZENBERG (G.). — On subword complexities
of homomorphic images of languages, R.A.I.R.O. Informatique Théorique,

[ER83a] EHRENFEUCHT (A.) and ROZENBERG (G.). — On the subword com-
plexity of locally catenative DOL-languages, Information Processing Letters,

[ER83b] EHRENFEUCHT (A.) and ROZENBERG (G.). — On the subword com-
plexity of m-free languages, Information Processing Letters, t. 17, 1983,
p. 121–124.

[ER83c] EHRENFEUCHT (A.) and ROZENBERG (G.). — On the size of the alpha-
bet and the subword complexity of square-free DOL languages, Semigroup

[P83] PAN SIOT (J.-J.). — Hiérarchie et fermeture de certaines classes de tag-


