THE SYNTACTIC INFERENCE PROBLEM FOR DOL-SEQUENCES

P.G. DOUCET

Afdeling Theoretische Biologie, Rijksuniversiteit Utrecht.

SUMMARY

The syntactic inference problem consists of deciding, for a given set of words, whether there exists a grammar such that its language includes these given words; and also of actually finding any such grammars. In this paper, the problem is considered for DOL-systems. The stress is on the second, constructive, part of the problem.
The initial information may have various forms. Most of the results deal with cases in which

- the words are given as a sequence (i.e., with their rank order numbers), which may be either consecutive or scattered.
- the size of the alphabet is given.

From the decidability point of view most of the results are not new. The proposed decision method, however, represents a considerable speed-up by passing the initial data through a number of algebraic "sieves" which turn out to be quite dense.
The method depends on there being enough information to establish a linear dependence relation between the Parikh-vectors of the given words.
Several variants of the problem are discussed. One subcase of a hitherto open problem is solved; other problems remain open.

0. INTRODUCTION

Suppose a not-too-large set of words, say, $S = \{ab, aabc, a^6bbc\}$ is given. One can ask whether there exists a DOL-system G such that $S \subseteq L(G)$. This is perhaps the simplest form of the syntactic inference problem. It is one of the 36 such problems for L-systems posed by Feliciangeli and Herman [28] , and one of the 6 which are still open. There may, however, be some additional information. The alphabet may be given, and there may be some information on the order of appearance of the words; either general (only the order) or specific (the precise rank order numbers). Feliciangeli and Herman make a different distinction. They only consider ordered, but not precisely numbered, sets; within this domain they distinguish sets of consecutive words, sets of scattered, but equally spaced, words and unspecified ordered sets of words.

I shall mainly discuss those cases where the aphabet is given as well as the rank order numbers. Actually, the inference problem is known to be decidable as soon as the alphabet is given: it is not difficult to see that in that case the number of (reduced) DOL-systems is finite, and one can simply try them all out. I intend to present an algorithm which is able to discard the vast majority of combinations at an early stage. It can be roughly described by the following sequence of steps:

1 take the Parikh-images of the given words
2 find a linear dependence relation and its associated polynomial $\psi(x)$
3 find a divisor of $\psi(x)$
4 find a growth matrix
5 find a set of production rules.

Essentially, the method makes use of the number of letters present in each given word (as opposed to their order) for as long as possible. This allows one to apply algebraic methods to the resulting vectors, these being generally more powerful than the combinatorial approach. The method works only if sufficient words are given to establish a linear dependence relation between their Parikh-vectors. Even then, a certain amount of trial-and-error work is necessary.

1. PRELIMINARIES

I shall assume the reader to be acquainted with the notion of a DOL - system such as outlined by Rozenberg and Doucet [91] or Salomaa [102]. I shall denote a DOL-system G by the triple $< \Sigma,P,w_0 >$, where Σ is

the alphabet, P the set of production rules, and w_0 the axiom. $\mathcal{E}(G)$ denotes the infinite sequence of words generated by G, in order of appearance. If a sequence of words is equal to $\mathcal{E}(G)$ for some DOL-system G, it is called a DOL-sequence. Any subsequence of a DOL-sequence is called a DOL-subsequence. Most DOL-subsequences occurring in the sequel will be finite. They may either be consecutive (such as w_4,w_5,w_6,w_7)or scattered (such as w_0,w_3,w_4,w_{20}).
Unless specified otherwise, any sequence will be a sequence of numbered words, i.e. a subset of $\mathbb{N}\times\Sigma^*$; it may be finite or infinite. Sequences will be denoted by script letters.
$\#$ S denotes the number of elements of a set S.
$|w|$ denotes the length (= number of letters) of a word w.
If $\Sigma = \{\sigma_1,\ldots,\sigma_k\}$, then the Parikh-vector $\bar{w}$ assigned to a word w is defined as a vector in $\mathbb{N}^k$ with its i^{th} coordinate equal to the number of occurrences of σ_i in w. Example: if $\Sigma = \{a,b,c\}$ and w = cacaa, then $\bar{w} = (3,0,2)^T$. The superscript denotes the transposition operator, since vectors will be written as column vectors. All vectors will be distinguishable by a bar.
The length of a vector $\bar{a} = (\alpha_1,\ldots,\alpha_k)^T$ is defined as $|\bar{a}| = \Sigma\alpha_i$. This definition is compatible with the earlier definition of word length: $|w| = |\bar{w}|$.
Without the details of a formal definition it will be clear that in a similar way the set P of production rules can be mapped into a k×k matrix $A_P = ((c_{ij}))$, where c_{ij} gives the number of occurrences of σ_i in $P(\sigma_{ij})$; in other words, the j-th column of A_P equals the Parikh-vector of $P(\sigma_j)$. A_P is called the growth matrix of P or G; it is also called the production matrix. If no confusion is likely, A_P is also written A.
If $\mathcal{S}$ is any sequence of words $w_{i_1},w_{i_2},\ldots$, then its Parikh-image $\bar{\mathcal{S}}$ denotes the sequence of vectors $\bar{w}_{i_1},\bar{w}_{i_2},\ldots$. Similarly, $\bar{\mathcal{E}}(G)$, the Parikh-sequence of G, is defined as $\bar{w}_0,\bar{w}_1,\bar{w}_2,\ldots$. $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ denote the sets of natural, integer, rational, real and complex numbers, respectively. R[x] denotes the set of all polynomials in x with coefficients in the set R.

2. RECURRENCE RELATIONS SATISFIED BY PARIKH-SEQUENCES.

Let $G = < \Sigma,P,w_0 >$ be a DOL-system with $\#\Sigma = k$, and let A be G's growth matrix. The Parikh-mapping $\sigma_1 \mapsto (1,0,\ldots,0)^T,\ldots,\sigma_k \mapsto (0,\ldots,0,1)^T$ maps words over the alphabet Σ into the k-dimensional vector space R over the field $\mathbb{Q}^{*)}$. The growth matrix A is a linear mapping of R into itself.

In G's Parikh-sequence $\bar{\mathcal{E}} = \bar{w}_0, \bar{w}_1, \bar{w}_2, \ldots$ some vectors are linearly dependent. Let such a dependence be given by a recurrence relation like

$$\bar{w}_8 + \bar{w}_6 - 3\bar{w}_5 + 4\bar{w}_0 = \bar{0}. \qquad (1)$$

This can also be written as

$$(A^8 + A^6 - 3A^5 + 4I)\bar{w}_0 = \bar{0}$$

(I is the k×k identity matrix), or

$$\psi(A)\bar{w}_0 = \bar{0},$$

where $\psi(x) = x^8 + x^6 - 3x^5 + 4$.

I shall call $\psi(x)$ the associated polynomial of the recurrence relation (1), and vice versa. (The customary term "characteristic polynomial" may lead to confusion).

This section deals with the question: what recurrence relations obtain in $\bar{\mathcal{E}}(G)$? Most of the answers come by way of their associated polynomials.

Three polynomials connected with A or G are of special importance.

First, the characteristic polynomial of A, $\phi_A(x)$, defined by $\phi_A(x) = \det(xI - A)$.

Second, the minimal polynomial of A, $m_A(x)$, defined as the lowest-degree monic**) polynomial in $\mathbb{Q}[x]$ for which $m_A(A) = O$ (the null matrix).

Third, the minimal polynomial of G, $\mu_G(x)$, defined as the lowest-degree monic polynomial in $\mathbb{Q}[x]$ for which $\mu_G(A)\bar{w}_0 = \bar{0}$.

Observe that $\phi_A(x)$ and $m_A(x)$ depend on A only; $\mu_G(x)$ depends on both A and $\bar{w}_0$.

The following lemmas summarize a few standard facts from matrix theory.

Lemma 2.1 $\phi_A(x)$, $m_A(x)$ and $\mu_G(x)$ are unique.

Lemma 2.2 $m_A(x)$ contains the same linear factors $x-\lambda_i$ ($\lambda_i \in \mathbb{C}$) as $\phi_A(x)$; their multiplicaties may, however, be lower. If $\phi_A(x) = 0$ has no multiple roots in $\mathbb{C}$, then $m_A(x) = \phi_A(x)$.

*) It might be more elegant to construct the more restricted module R over the ring $\mathbb{Z}$, but the present approach will do.

**) i.e., with leading coefficient 1.

Lemma 2.3 $m_A(x)$ divides every polynomial $\psi(x)$ for which $\psi(A) = O$.
Lemma 2.4 $\mu_G(x)$ divides every polynomial $\psi(x)$ for which $\psi(A)\bar{w}_0 = \bar{0}$.
Lemma 2.5 There exist algorithms for finding $m_A(x)$ and $\mu_G(x)$.

For further information, I refer the reader to standard texts like Gantmacher, Chs. IV and VII*). Furthermore, I will need two theorems from algebra:

Lemma 2.6 If factorization in an integral domain R is unique, so is factorization in $R[x]$.

Lemma 2.7 A polynomial in $\mathbb{Z}[x]$ which can be factored in polynomials in $\mathbb{Q}[x]$ can already be factored in polynomials in $\mathbb{Z}[x]$.

Both lemmas can, e.g., be found in Birkhoff and Maclane, Ch. III**). Since $\phi_A(x) \in \mathbb{Z}[x]$, $\mu_G(x) \in \mathbb{Q}[x]$, $m_A(x) \in \mathbb{Q}[x]$, and $\mathbb{Z}$ has unique factorization, lemmas 2.6 and 2.7 can be applied, giving

Lemma 2.8 $m_A(x)$ and $\mu_G(x)$ have integer coefficients.

Theorem 2.9 If in a DOL-sequence $\mathcal{E}(G)$ some vectors from $\bar{\mathcal{E}}(G)$ satisfy a recurrence relation, then they also satisfy a monic recurrence relation with integer coeeficients.

Proof Suppose the recurrence relation $w_r + \alpha_{r-1}\bar{w}_{r-1}+\ldots.+\alpha_0\bar{w}_0 = \bar{0}$ is given, with all $\alpha_i \in \mathbb{Q}$.

If m is the degree of G's minimal polynomial $\mu_G(x)$, then $r < m$ is impossible; for $r = m$ the theorem is trivially true (by lemma 2.1), so $r > m$ remains to be examined.

All α_i are rational, so one can find a number M such that

$$M\bar{w}_r + M\alpha_{r-1}\bar{w}_{r-1}+\ldots+M\alpha_0\bar{w}_0 = \bar{0} \qquad (2)$$

has only integer coefficients. $\bar{w}_r$ can be expressed in $\bar{w}_{r-1},\ldots,\bar{w}_{r-m}$ by means of the recurrence relation associated with $\mu_G(x)$, which has integer coefficients. By subtracting this relation M-1 times from (2), one obtains a relation of the required form. $\boxtimes$

The subspace of Q^k spanned by the vectors $\bar{w}_0,\bar{w}_1,\bar{w}_2,\ldots$ is of dimension m, since, from rank order number m on, each vector linearly depends on the previous vectors. Hence

Theorem 2.10 If $\mu_G(x)$ has degree m, then any set of m vectors from $\bar{\mathcal{E}}(G)$ is linearly dependent.

*) F.R. Gantmacher, The theory of matrices. New York: Chelsea, 1959 (Translated from the Russian).

**) G. Birkhoff and S. Maclane. A survey of modern algebra. New York: Mac Millan, 1953.

Let G be a DOL-system with minimal polynomial $\mu_G(x)$, of degree m. Then $\mathcal{E}(G) = w_0, w_1, w_2, \ldots$. All Parikh-vectors $\bar{w}_r$ from $\bar{\mathcal{E}}(G)$ can be expressed in terms of the first m Parikh-vectors $\bar{w}_0, \ldots, \bar{w}_{m-1}$. In the sequel, this initial subsequence will be denoted by $\bar{\mathcal{E}}_0$. These m vectors can also be collected in the matrix E_0, with k rows and m columns. Thus each vector $\bar{w}_r \in \bar{\mathcal{E}}(G)$ can be written as

$$\bar{w}_r = E_0 \bar{c}_r,$$

where $\bar{c}_r$, the <u>coefficient vector</u> of the word w_r (or the vector $\bar{w}_r$) is an m-vector. So $\bar{c}_r$ is just another way of representing the recurrence relation by which $\bar{w}_r$ can be expressed in the first m Parikh-vectors. Observe that

(i) For each r, $\bar{c}_r$ is unique and has integer coefficients.

(ii) The associated polynomial of $\bar{c}_r$ (with obvious definition) is equal to $x^r (\mathrm{mod}\ \mu_G(x))$.

(iii) If some set of Parikh-vectors satisfies a recurrence relation

$$\bar{w}_r + \alpha_{r-1}\bar{w}_{r-1} + \ldots + \alpha_0 \bar{w}_0 = \bar{0},$$

then, since E_0 is non-singular, their coefficient vectors satisfy the same relation:

$$\bar{c}_r + \alpha_{r-1}\bar{c}_{r-1} + \ldots + \alpha_0 \bar{c}_0 = \bar{0}.$$

3. THE BASIC INFERENCE PROBLEM

The simplest problem is the following.

Given an alphabet of size k and a sequence $\mathcal{S} = w_0, \ldots, w_k$, find all DOL-systems G for which $\mathcal{S}$ is the initial subsequence of $\mathcal{E}(G)$.

To solve the problem, first form the Parikh-images of the words: $\bar{\mathcal{S}} = \bar{w}_0, \ldots, \bar{w}_k$. If $\mathcal{S}$ is to be the initial subsequence of some $\mathcal{E}(G)$, then the following relation must hold:

$$A \begin{pmatrix} \bar{w}_0 & \cdots\cdots & \bar{w}_{k-1} \\ | & & | \\ | & & | \end{pmatrix} = \begin{pmatrix} \bar{w}_0 & \cdots\cdots & \bar{w}_k \\ | & & | \\ | & & | \end{pmatrix} \qquad (3)$$

or

$$AS = T \qquad (4)$$

(A is the growth matrix of G).

One may now distinguish two cases, depending on S.

First, if S is non-singular (which, incidentally, means that $\phi_A(x) = m_A(x) = \mu_G(x)$), then A is uniquely determined by

$$A = TS^{-1}$$

Second, if rank (S) = k-r with $r > 0$, then A can be written

$$A = A_0 + \lambda_1 A_1 + \ldots . + \lambda_q A_q,$$

where - A_0 is some solution of AS = T

- $A_1, \ldots, A_q$ are mutually independent solutions of AS = O; $q \leqslant k.r.$
- $\lambda_1, \ldots, \lambda_q$ are otherwise arbitrary numbers such that A has only positive elements.

The last condition leads to q linear unequalities in $\lambda_1, \ldots, \lambda_q$, with finitely many solutions (perhaps none). Properly speaking, the number of solutions can indeed be infinite, but only if (and as far as) letters not occurring in $\mathcal{S}$ are concerned. The complication is completely formal, since such a letter will never appear at all if it did not appear in $\mathcal{S}$; it may, if desired, be formally eliminated by only admitting reduced G's.

Once an admissible growth matrix A is found, one returns from Parikh-vectors to words.

Lemma 3.1 The growth matrix and the first $\# \Sigma+1$ words of a DOL-sequence uniquely determine the useful production rules.

Proof The restriction to useful production rules (i.e. rules which are at all applied in &) should be obvious.

Let $w_0 = \sigma_{01} \ldots . \sigma_{0p}$.

One can then parse w_1 in p subwords, starting from the left-hand side:

$$w_1 = P(\sigma_{01}) * \ldots . * P(\sigma_{0p}),$$

where the subword lengths $|P(\sigma_{0i})|$ can be looked up in A, being equal to the lengths of the column vectors $\overline{P(\sigma_{0i})}$ of A.

By this procedure, $P(\sigma_{0i})$ is found for all letters in w_0. The procedure is then continued for the one-step derivations $w_1 \Rightarrow w_2, \ldots, w_{k-1} \Rightarrow w_k$. Any letter which has not appeared by then will not appear at all. ⊠

Theorem 3.2 For a given k×k matrix A and a word sequence $\mathcal{S} = w_0, \ldots, w_k$ over a k-letter alphabet there is at most one reduced DOL-system G with the properties

(i) A is the growth matrix of G.

(ii) $\mathcal{S}$ is the initial subsequence of &(G).

G can be effectively constructed.

Proof The theorem follows from lemma 3.1 by observing that the construction of the set of production rules P from A and $\mathcal{S}$ does not depend on the fact that A is a growth matrix or $\mathcal{S}$ is a DOL-subsequence. $\boxtimes$

During the construction of P several things may happen, in this order:

1 The total length of $P(w_i)$ as found from A is not equal to the length of the given word w_{i+1}.

2 After parsing w_{i+1} in $|w_i|$ subwords the Parikh-vectors of the individual subwords are not equal to the appropriate columns of A, even though the lengths may match.

3 For some letter σ, $P(\sigma)$ as found from one instance of σ somewhere in the derivation may differ from $P(\sigma)$ as found from some other instance of σ, even though both Parikh-vectors are equal (and consistent with A).

In each case, A is rejected as a growth matrix for $w_0,\ldots,w_k$. Whether in case 3 $w_0,\ldots,w_k$ must also be rejected as a DOL-subsequence is not yet quite clear to me.

If the matrix A happens to be non-singular (which is the rule rather than the exception), theorem 3.2 has interesting consequences, which can be formulated in various ways. Let the *order* of a recurrence relation $\psi(A)\bar{w}_0 = \bar{0}$ be defined as the degree of the polynomial $\psi(x)$.

Corollary 3.3 If a DOL-subsequence $\mathcal{S}$ does not satisfy any recurrence relation of order lower than $\#\Sigma$, then there is only one G with $\mathcal{S} = \mathcal{E}(G)$.

Corollary 3.4 If a sequence $\mathcal{S}$ of k+1 words over a k-letter alphabet does not satisfy any recurrence relation of order lower than k, then there exists at most one DOL-system G such that $\mathcal{S}$ is the initial subsequence of $\mathcal{E}(G)$.

Corollary 3.5 Two different (and reduced) DOL-systems G and H can only produce the same sequence if

$$\mu_G(x) = \mu_H(x) \neq \phi_G(x) = \phi_H(x).$$

4. Inference from a scattered sequence.

Like in the previous section, the alphabet Σ is regarded as given; $\#\Sigma = k$. The given sequence of words, however, has the form $\mathcal{S} = w_{i_1}, w_{i_2}, \ldots, w_{i_p}$, with $i_0 < i_1 < \ldots < i_p$ and p arbitrary, instead of $\mathcal{S} = w_0, w_1, \ldots, w_k$. After a few remarks on notation I shall first describe the algorithm which produces all possible G's from $\mathcal{S}$, then go into its justification, and next give two examples. The section is concluded by a flow diagram indicating the acceptance/rejection structure of the algorithm.

Three different sequences will appear in the sequel:

- the given sequence $\mathcal{S} = w_{i_1}, w_{i_2}, \ldots, w_{i_p}$.
- the initial D0L-subsequence $\mathcal{E}_0 = w_0, w_1, \ldots, w_{m-1}$, where m is the degree of G's minimal polynomial.
- the next-to-initial D0L-subsequence $\mathcal{E}_1 = w_1, w_2, \ldots, w_m$.

Each of these sequences can be collected in a matrix. They will be denoted by S, E_0 and E_1, respectively; they are elements of $\mathbb{N}^{k\times p}$ and $\mathbb{N}^{k\times m}$ (twice).

As described at the end of section 2, the elements of $\overline{\mathcal{S}}$ can all be expressed in the elements of $\overline{\mathcal{E}_0}$, each $\overline{w}_j$ by means of its coefficient vector $\overline{c}_j$. The coefficient vectors of $\mathcal{S}$ can again be collected in a matrix, C, which is an element of $\mathbb{Z}^{m\times p}$.

The construction procedure now runs as follows:

1. Find some recurrence relations within $\overline{\mathcal{S}}$. Determine the greatest common divisor of their associated polynomials, say, $\psi(x)$.
2. Find a monic divisor of $\psi(x)$, with integer coefficients and degree $\leqslant k$. Let $\chi(x)$, with degree m, be such a divisor. The next steps will investigate whether $\chi(x) = \mu_G(x)$ for some G such that $\mathcal{S} \subseteq \mathcal{E}(G)$.
3. Compute the coefficient matrix C from $\chi(x)$ and the index set of $\mathcal{S}$.
4. Find $E_0 \in \mathbb{N}^{k\times m}$ satisfying the matrix equation $S = E_0C$.
5. Determine $\overline{w}_m$ from $\overline{w}_0, \ldots, \overline{w}_{m-1}$ as found in E_0 and from $\chi(x)$'s associated recurrence relation. Now one can compose E_1 from E_0 and $\overline{w}_m$. Next find $A \in \mathbb{N}^{k\times k}$ satisfying the matrix equation $E_1 = A\ E_0$.
6. Determine all powers of the production rules P needed to produce the words of $\mathcal{S}$ from one another by first computing the appropriate powers of A and then using these to parse the words of $\mathcal{S}$ (as in the basic inference problem from section 3).
7. By combinatorial means, find the set of production rules P from the growth matrix and the various powers of P found in step 6.

Ad 1. $\overline{\mathcal{S}}$ may or may not satisfy a recurrence relation. Of course it always does if $\mathcal{S}$ contains k+1 or more words, but this is not a necessary condition. If $\bar{\mathcal{S}}$ does not satisfy a recurrence relation, the whole procedure simply doesn't work. If it does, there may be several, and it is helpful (though not necessary) to find them all. If $\mathcal{S}$ is part of a D0L-sequence $\mathcal{E}(G)$, then the associated polynomials of these recurrence relations are all multiples of $\mu_G(x)$; so is their greatest common divisor, $\psi(x)$. If the associated polynomial of any of the discovered recurrence relations does not (in its monic form) have integer coefficients, then $\mathcal{S}$ is no D0L-subsequence (by theorem 2.9).

Example: If $\mathcal{S}$ consists of w_2 = acd, w_3 = abba and w_5 = acbbdaa, then $2\bar{w}_2 + \bar{w}_3 - 2\bar{w}_5 = \bar{o}$; the associated polynomial (in monic form) is $x^5 - \frac{1}{2}x^3 - x^2$, which does not satisfy theorem 2.9. Hence $\mathcal{S}$ is not a D0L-subsequence.

<u>Ad 2</u>. $\mu_G(x)$ must have the following properties:

(i) it divides $\psi(x)$

(ii) it is monic and has integer coefficients (by lemma 1.8).

(iii) it has degree k or less.

Step 2 consists of finding all polynomials $\chi(x)$ with these properties, by trial and error. That this is a finite enterprise is ensured by

<u>Lemma 4.1</u>. For a given polynomial $\psi(x) \in \mathbb{Z}[x]$ bounds can be found for the coefficients of all $\psi(x)$'s divisors of given degree m.

<u>Proof</u>. By a well-known theorem from algebra, all complex roots y_j of a polynomial

$$\psi(x) = x^n + \alpha_{n-1}x^{n-1} + \ldots \alpha_1 x + \alpha_0$$

are either smaller than 1 or are bounded by

$$|y_j| \leqslant M = n \cdot \max_i \{\alpha_i\}.$$

Since in our case $|\alpha_i| \geqslant 1$ for all i, $|y_j| \leqslant M$ holds in all cases. Any divisor $\chi(x)$ of $\psi(x)$ of degree m can be written as

$$\begin{aligned}\chi(x) &= (x-y_1)\ldots(x-y_m)\\ &= x^m + \beta_{m-1}x^{m-1} + \ldots + \beta_0,\end{aligned}$$

where

$$|\beta_{m-1}| = |y_1+\ldots+y_m| \leqslant mM$$

$$|\beta_{m-2}| = |\sum_{i\neq j} y_i y_j| \leqslant m(m-1)M$$

etc.

Thus each β_i of $\chi(x)$ can be bounded in terms of m and M. (To be sure, the bounds so obtained are often not very practical, and may be considerably improved by using the fact that $\beta_i \in \mathbb{Z}$; for example, β_0 must, by lemma 2.7, be a factor of α_0). ⊠

Ad 3. The construction of the various $\bar{c}_i$ was described at the end of section 2.

Ad 4. The matrix C may be singular, so there may be several (though only finitely many) E_0 satisfying the equation $S = E_0C$.

Ad 5. Like in the basic inference problem of section 3, several (though only finitely many) growth matrices A may be found.

Ad 6. Knowing A, one can now parse w_{i_1} into $|w_{i_0}|$ subwords, each of length found from the appropriate column vector length of $A^{i_1-i_0}$, and thus infer some rules from $P^{i_1-i_0}$. In contrast with the basic problem, the absence of a letter in $\mathfrak{J}$ does not mean that it is never used at all. It may have been used in the words in between the given words, and it may be indispensable.

Ad 7. The information obtained from step 6 does not always uniquely determine P.

As an example, consider the following problem:

Find all D0L-systems over the alphabet {a,b,c,d} such that &(G) includes w_1 = d, w_3 = dac, w_5 = accbd, w_9 = cbddacdacaccbd. To solve the problem, follow the steps of the procedure:

1. The sequence $\mathfrak{J}$ consists of $\bar{w}_1 = (0,0,0,1)^T$, $\bar{w}_3 = (0,1,1,1)^T$, $\bar{w}_5 = (1,1,2,1)^T$, $\bar{w}_9 = (3,2,5,4)^T$.
The only recurrence relation obtaining in $\mathfrak{J}$ is $\bar{w}_9-3\bar{w}_5+\bar{w}_3-2\bar{w}_1 = \bar{o}$. Its associated polynomial is $\psi(x) = x^9-3x^5+x^3-2x$.

2. $\psi(x) = x(x^2+2)(x^3-x-1)^2$. Its set of divisors of degree $\leqslant 4$ (= Σ) exhausts the possibilities for G's minimal polynomial; it consists of x, x^2+2, $x(x^2+2)$, x^3-x-1 and $x(x^3-x-1)$.
Of these x can be immediately discarded. So can x^2+2 (which is associated with the impossible recurrence relation $\bar{w}_2 = -2\bar{w}_0$) and $x(x^2+2)$. Two polynomials remain, x^3-x-1 and x^4-x^2-x.

3. First try $\chi(x) = x^3-x-1$. $\chi(x)$ is associated with $\bar{w}_3 = \bar{w}_1 + \bar{w}_0$, and iteration produces

$$\bar{w}_4 = \bar{w}_2 + \bar{w}_1$$
$$\bar{w}_5 = \bar{w}_3 + \bar{w}_2 = \bar{w}_2 + \bar{w}_1 + \bar{w}_0$$
$$\vdots$$
$$\bar{w}_9 = 3\bar{w}_2 + 4\bar{w}_1 + 2\bar{w}_0$$

So $\bar{c}_1 = (0,1,0)^T$, $\bar{c}_3 = (1,1,0)^T$, $\bar{c}_5 = (1,1,1)^T$ and $c_9 = (2,4,3)^T$;

$$c = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 1 & 1 & 4 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

4. E_0 must now be solved from $S = E_0C$, or $\begin{pmatrix} 0 & 0 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 1 & 2 & 5 \\ 1 & 1 & 1 & 4 \end{pmatrix} = E_0 \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 1 & 1 & 4 \\ 0 & 0 & 1 & 3 \end{pmatrix}$.

E_0 must also be in $\mathbb{N}^{4\times 3}$.

There happens to be precisely one solution : $E_0 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

This supplies the three initial vectors of $\bar{\mathcal{E}}(G)$.

5. The fourth vector of $\bar{\mathcal{E}}(G)$ is found from the first three and from G's minimal recurrence relation : $\bar{w}_3 = \bar{w}_1 + \bar{w}_0 = (0,1,1,1)^T$. Now E_1 is also known, and the growth matrix $A(\in \mathbb{N}^{4\times 4})$ can be solved from

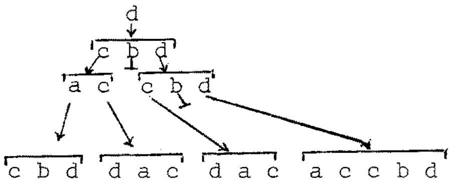
$$AE_0 = A_1, \text{ or } A \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

There are solutions, $A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ and $A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$

6. First try the first A. The powers of A relevant for are A^2(for $\bar{w}_1 \overset{2}{\Rightarrow} \bar{w}_3$ and $\bar{w}_3 \overset{2}{\Rightarrow} \bar{w}_5$) and A^4(for $\bar{w}_5 \overset{4}{\Rightarrow} \bar{w}_9$).

$$A^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ and } A^4 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 2 \\ 1 & 0 & 1 & 1 \end{pmatrix} .$$

These induce the following parsing in $\mathcal{S}$:



The parsing is consistent, both internally and with A, and produces the following information :

$$P^2(d) = cbd \; ; \; P^4(a) = cbd, \; P^4(c) = dac, \; P^4(d) = accbd.$$

7. Combining the data from step 6 with the growth matrix A, one obtains one set of production rules :
$\{a \to cb,\ b \to \lambda,\ c \to d,\ d \to ac\}$ and two possible axioms: $w_0 = b_c$ or $w_0 = cb$.

6,7 Now try the other A left over from step 5. In the same fashion, one set of production rules is produced, again with two possible axioms :

$P = \{a \to cbd,\ b \to d,\ c \to \lambda,\ d \to ac\}$; $w_0 = bc$ or $w_0 = cb$.

3,4,5,6,7 One more $\chi(x)$ was left over from step 3 : $\chi(x) = x^4 - x^2 - x$.

It first produces $C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & 2 \end{pmatrix}$, then E_0 $\begin{pmatrix} \alpha_0 & 0 & 1 & 0 \\ \alpha_1 & 0 & 0 & 1 \\ \alpha_2 & 0 & 1 & 1 \\ \alpha_3 & 1 & 0 & 1 \end{pmatrix}$ with the α_i (the axiom's coordinates) arbitrary.

In step 5, two solutions for A appear, together with further restrictions on the axiom :

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \text{ with } \overline{w}_0 = (1,\lambda,0,0)^T \ (\lambda \text{ arbitrary})$$

or

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \text{ with } \overline{w}_0 = (1,0,\lambda,0)^T \ (\lambda \text{ arbitrary}).$$

In steps 6 and 7 the same two P's as before are produced ; the axioms are, however, sligthly different :

$P = \{a \to cb,\ b \to \lambda,\ c \to d,\ d \to ac\}$ with $w_0 = b^m cb^n$ (m,n arbitrary),

$P = \{a \to cbd,\ b \to d,\ c \to \lambda,\ d \to ac\}$ with $w_0 = c^m bc^n$.

This concludes the example.

Another example may show the procedure's speed to advantage.

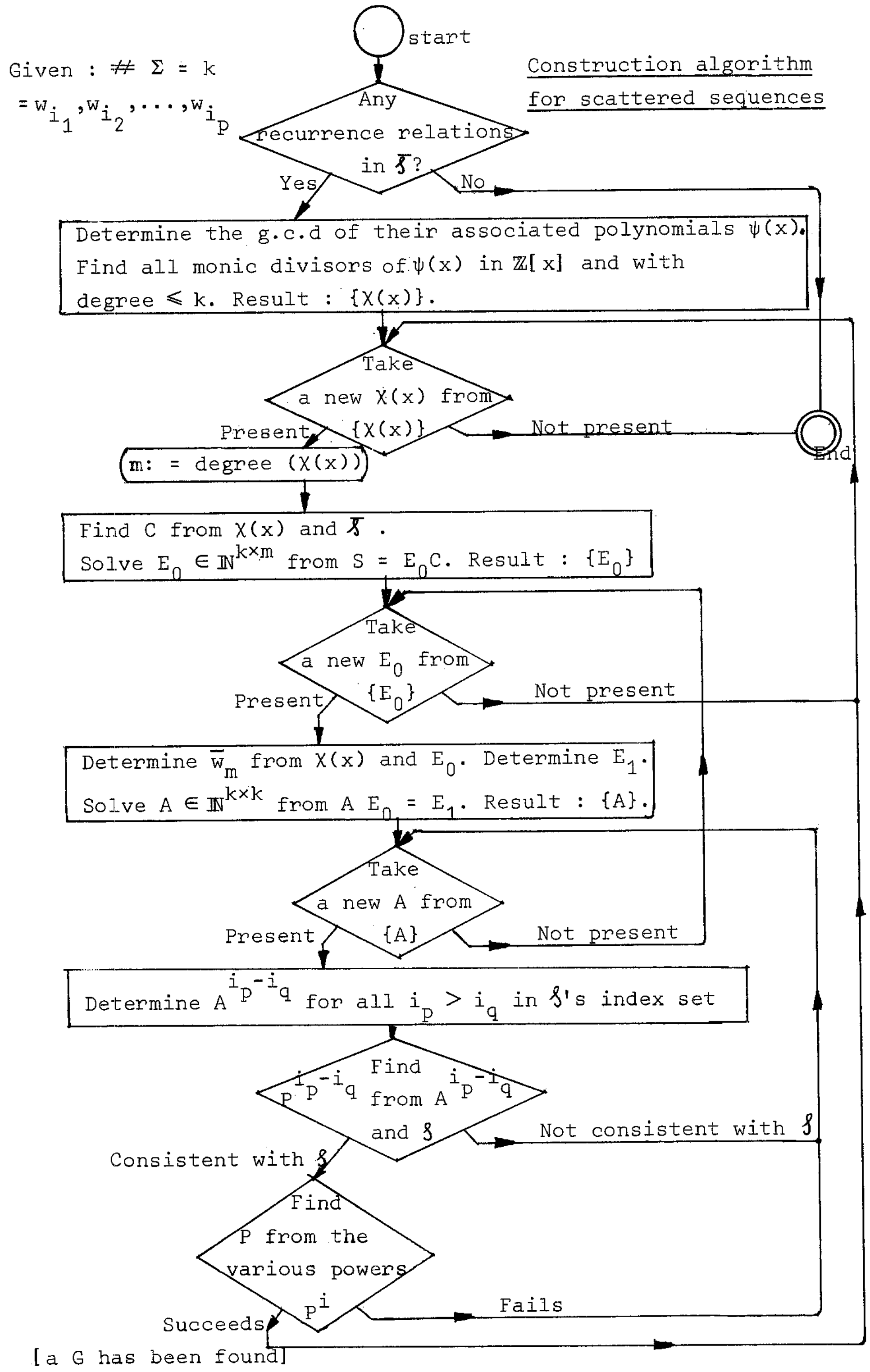
Let $\mathcal{S}$ be given, consisting of w_1 = b, w_4 = acb, w_6 = ddab, w_9 = abcaddab, and w_{11} = aabbccddabcd. $\Sigma = \{a,b,c,d\}$. Is $\mathcal{S}$ part of a D0L-sequence ?

Step 1 : $w_1 = (0,1,0,0)^T$, $w_4 = (1,1,1,0)^T$, $w_6 = (1,1,0,2)^T$, $w_9 = (3,2,1,2)^T$, $w_{11} = (3,3,3,3)^T$.

Now $\overline{w}_{11}$ is of course dependent on the other vectors, but not by a monic relation with integer coefficients.

Hence $\mathcal{S}$ is (by theorem 2.9) not a part of any D0L-sequence.

Observe that this remains true if the alphabet is not given.



Construction algorithm for scattered sequences

start
Given : $\# \Sigma = k$
$\mathfrak{S} = w_{i_1}, w_{i_2}, \ldots, w_{i_p}$
Any recurrence relations in $\mathfrak{S}$?
Yes
No
Determine the g.c.d of their associated polynomials $\psi(x)$. Find all monic divisors of $\psi(x)$ in $\mathbb{Z}[x]$ and with degree $\leqslant k$. Result : $\{\chi(x)\}$.
Take a new $\chi(x)$ from $\{\chi(x)\}$
Present
Not present
End
m: = degree $(\chi(x))$
Find C from $\chi(x)$ and $\mathfrak{S}$. Solve $E_0 \in \mathbb{N}^{k \times m}$ from $S = E_0 C$. Result : $\{E_0\}$
Take a new E_0 from $\{E_0\}$
Present
Not present
Determine $\bar{w}_m$ from $\chi(x)$ and E_0. Determine E_1. Solve $A \in \mathbb{N}^{k \times k}$ from $A E_0 = E_1$. Result : $\{A\}$.
Take a new A from $\{A\}$
Present
Not present
Determine $A^{i_p - i_q}$ for all $i_p > i_q$ in $\mathfrak{S}$'s index set
Find $P^{i_p - i_q}$ from $A^{i_p - i_q}$ and $\mathfrak{S}$
Consistent with $\mathfrak{S}$
Not consistent with $\mathfrak{S}$
Find P from the various powers P^i
Succeeds
Fails
[a G has been found]

5. Further extensions of the inference problem.

The problems discussed in the previous sections had the following properties in common :

(i) the alphabet Σ is given.

(ii) enough words are given to establish a recurrence relation within the given sequence.

(iii) not only a number of words are given, but their rank order numbers as well.

One can examine to what extent the method remains valid for problems not possessing these properties. In this section I shall discuss some of the seven remaining cases.

In general, one can say that the method hinges on determining $\mu_G(x)$ from a recurrence relation within $\mathfrak{S}$.

In the absence of such a relation (the cases (000),(001),(100),(101); in binary code , referring to the three properties) the method simply does not work; If $\# \Sigma$ is given ((100) and (101)) the problem can be solved by a laborious but finite exhaustive search. If $\# \Sigma$ is not given (subcases (000) and (001)) the problem is not so simple. In fact, I do not know whether the decision problem is at all solvable for these cases. The same goes for case (010).

In case (101) ($\# \Sigma = k$ and a numbered sequence $\mathfrak{S}$ are given ; but $\overline{\mathfrak{S}}$ satisfies no recurrence relation), two subcases may be distinguished: $\mathfrak{S}$ either does or does not certain a word with rank order number larger than k.

If $\mathfrak{S}$ contains at least one word w_p with $p \geqslant k$, then it is not difficult to see that any possible growth matrix $A = ((a_{ij}))$ is bounded by $a_{ij} \leqslant M$ (where M is the maximum number occurring in the vectors of $\overline{\mathfrak{S}}$), except for those numbers referring to mortal letters. As a result, the problem is bounded for all vital letters and not bounded (in a rather unimportant way) for mortal ones.

If no word w_p with $p \geqslant k$ is given, no such upper bound for the elements of A can be given, and the problem often has infinitely many solutions.

Case (100) would reduce to a finite number of the previous cases (101) if from $\# \Sigma$ and $\mathfrak{S}$ an upper bound for the rank order numbers could be deduced. This bound can indeed be found ; by a size argument if L(G) is finite, by a growth argument if L(G) is infinite.

If L(G) is finite, then a result by P. Vitányi [114] states that L(G) contains at most $k(1 + k^{n-1})$ words, where n is the number of different monorecursive letters in a certain, specified, word. Since obviously $n \leqslant k$, $\# (L(G)) \leqslant k(1 + k^{k-1})$. This number gives the required upper bound, since any higher rank order number refers to a duplicate of an earlier word.

If, on the other hand, L(G) is infinite, then $|w_{n+k}| \geqslant |w_n| + 1$ for every n. Consequently, an upper bound for the rank order numbers $\{i\}$ in the given set of words S is given by

$$i \leqslant k \cdot \max \{|w| : w \in S\}$$

So case (100) can be reduced to case (101) ; hence the case is solvable. The indicated procedure is of course not nearly a practical method.

In case (011) a recurrence relation can be found in ξ , but Σ may be larger than the "observed" Σ_{obs}.

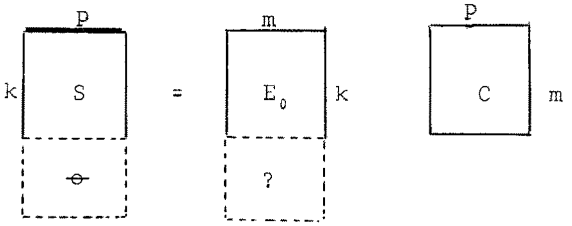
The inference algorithm of section 4 applies during the first steps, where no knowledge of Σ is necessary.

In step 4 problems may arise. First, the degree of $\mu_G(x)$ can never be larger then $\# \Sigma$. So, if it turns out that $\deg(\mu_G(x)) = m > \# \Sigma_{obs}$, Σ must be larger than Σ_{obs}. Now simply extend Σ_{obs} to Σ in the minimal way : namely, such that

$$\# \Sigma = k = \max (\# \Sigma_{obs}, m),$$ and then apply

step 4 : find a matrix $E_0 \in \mathbb{N}^{k \times m}$ satisfying $S = E_0 C$.

From the pictorial representation



it is not difficult to see that increasing $\# \Sigma$ beyond k cannot have any other effect than adding mortal letters to solutions already obtained with Σ. In other words : if there are no solutions for E_0 with this minimal Σ, then there are none.

That Σ can indeed be larger than Σ_{obs} can be seen from this very simple example : Find a DOL-system such that $w_0 = a$, $w_4 = aaa$. $\psi(x) = x^4 - 3$, with no other divisors. If $\psi(x)$ is to be G's minimal polynomial, then $\# \Sigma \geqslant 4$. In fact, $p = \{a \to b, b \to c, c \to d, d \to aaa\}$ provides a solution.

Case (110) can be regarded as a more favorable subcase of (100), involving considerably less guesswork.

Of the variants discussed, (011) may be the most interesting one, since it solves a subcase of the hitherto open problem (Feliciangeli and Herman [28]) of finding G from $\mathcal{S}$ if Σ is not given.