Theoretical Computer Science 38 (1985) 137-142
North-Holland

NOTE

UNE GÉNÉRALISATION DES THÉORÈMES DE HIGMAN ET DE SIMON AUX MOTS INFINIS

A. FINKEL

Université de Paris-Sud, L.R.I., Bâtiment 490, 91405 Orsay Cedex, France

Communicated by D. Perrin
Received February 1984
Revised October 1984

Résumé. Nous montrons que d'une suite infinie de mots finis ou infinis on peut extraire une sous-suite croissante pour le pré-ordre 'être un sous-mot de'; ceci constitue le théorème de Higman étendu aux mots infinis. Puis nous généralisons le théorème de Simon aux mots infinis.

Abstract. We prove that from an infinite sequence of finite or infinite words on a finite alphabet one can extract an increasing subsequence for the quasi-order 'be a subword of'; that is, the extension of Higman's theorem to infinite words. Then we generalize Simon's theorem for infinite words.

1. Introduction

Le but de cette note est d'étendre le théorème de Higman [1, 2, 3] et le théorème de Simon [3, 6] aux mots infinis. Autrement dit nous montrons que de toute suite infinie de mots finis ou infinis on peut extraire une sous-suite croissante pour le pré-ordre 'être un sous mot de', noté / . De plus nous montrons qu'étant donné deux mots u et v ayant le même ensemble de sous-mots de longueur n il existe un mot w tel que u/w, v/w et tel que w ait les mêmes sous-mots de longueur n que u et v.

La propriété, pour un ensemble infini muni d'un pré-ordre, de contenir deux éléments comparables a été étudiée par beaucoup d'auteurs et a fait l'objet d'un historique [2]. Citons, pour exemple, deux ensembles pré-ordonnés sur lesquels la propriété précédente est vérifiée: sur l'ensemble des vecteurs d'entiers positifs muni de la relation d'ordre habituelle cette propriété est connue comme le lemme de Dickson; sur l'ensemble des mots finis (sur un alphabet fini) muni de la relation de pré-ordre 'être un sous mot de', il s'agit d'une application du théorème de Higman aux mots finis.

La preuve du théorème de Higman, donnée, par exemple, dans [3], ne se généralise pas de facon immédiate aux mots infinis: le problème est qu'on ne peut pas parler

d'un mot infini u plus court qu'un mot infini v. L'extension de la preuve, aux mots infinis, se fait en remarquant que tout mot infini u peut s'écrire sous la forme $u = u'u''$ où u' est un mot fini (éventuellement vide) et u'' un mot infini dit 'équitable' tel que toute lettre qui apparaît dans u'' y apparaît une infinité de fois.

Nous avons appris, après avoir écrit nos preuves, que Erdös, Higman, Milner, Nash-Williams, et Rado avaient démontré [4, 5] de facon très générale le théorème de Higman étendu aux mots infinis à supports ordinaux. Notre démonstration reste intéressante pour deux raisons:

- elle est plus simple que celle donnée en [4],
- elle fait apparaître la décomposition d'un mot infini en un mot fini et en un mot infini équitable.

Cette décomposition, canonique, permet la démonstration du théorème de Simon [6] qui, lui, n'a jamais été généralisé aux mots infinis.

Le théorème de Simon énonce que pour tous mots finis u et v, pour tout entier $n > 0$, si u et v ont les mêmes sous-mots de longueur n il existe un troisième mot w tel que u et v soient sous-mots de w et tel que w ait les mêmes sous-mots de longueur n que u et v.

L'extension du théorème de Simon repose sur deux faits: premièrement l'ensemble des sous-mots de longueur n d'un mot infini u peut être obtenu comme l'ensemble des sous mots de longueur n d'un facteur gauche u' de u; deuxièmement si deux mots infinis u et v possèdent le même ensemble $S(n, u) = S(n, v)$ de sous-mots de longueur n, alors $S(n, u)$ contient l'ensemble des mots de longueur n construits sur l'ensemble $\mathrm{equ}(u, v)$ des lettres qui apparaîssent une infinité des fois dans u ou dans v.

Un mot infini w qui a comme diviseurs deux mots infinis u et v, possédant le même sous-ensemble de mots longueur n, s'obtient alors en appliquant le théorème de Simon (pour les mots finis) à deux facteurs gauche u' et v' de u et v correctement choisis. Soit w' le mot fini ainsi obtenu qu'on complète alors par un mot infini équitable w'' quelconque sur l'alphabet $\mathrm{equ}(u, v)$ défini ci-dessus. On a alors $w = w'w''$.

2. Définitions et notations

On note A^* le monoide libre engendré par l'alphabet A, et A^ω l'ensemble des mots infinis sur A; on pose $A^\infty = A^* \cup A^\omega$. On désigne par $|A|$ le cardinal de A et par $|u|$ la longueur du mot u. Le mot vide a une longueur nulle et est noté λ. A^n désigne l'ensemble des mots de longueur n sur l'alphabet A et $A.B$ est l'ensemble défini de la manière suivante $A.B = \{a.b \mid a \in A \text{ et } b \in B\}$. Si u est un mot et p un entier positif, on notera $u[p]$ le facteur gauche de u de longueur égale à p.

Définition. Soient u et v deux mots de A^*. u *divise* v (on note u/v) s'il existe un mot w de A^* tel que $v = w_1u_1w_2u_2 \ldots w_nu_nw_{n+1}$ avec $u = u_1 \ldots u_n$ et $w = w_1 \ldots w_{n+1}$,

les u_i pour $i = 1, \ldots, n$ étant des lettres de A, les w_j pour $j = 1, \ldots, n+1$ étant des mots de A^*.

On étend de facon naturelle cette relation sur A^∞. Remarquons que / est une relation d'ordre sur A^* mais seulement un pré-ordre sur A^ω: en effet, cette relation n'est plus antisymétrique sur A^ω comme le montre l'exemple suivant: $(ab)^\omega/(aab)^\omega$ et $(aab)^\omega/(ab)^\omega$ mais $(ab)^\omega$ n'est pas égal à $(aab)^\omega$.

Lemme 2.1. *Soient f et g deux mots finis et u et v deux mots de A^∞. Si f/g et u/v, alors fu/gv.*

Notations. Si $u \in A^\infty$, on note $\mathrm{equ}(u)$ le sous-ensemble de A définit par $\mathrm{equ}(u) = \{a \in A \mid u_a = \infty\}$ et $\mathrm{alph}(u)$ le sous-ensemble de A définit par $\mathrm{alph}(u) = \{a \in A \mid u_a \neq 0\}$ où u_a désigne le nombre d'occurrences de la lettre a dans le mot u.

Définition. Un mot u de A^∞ est équitable si $\mathrm{equ}(u) = \mathrm{alph}(u)$.

Lemme 2.2. *Tout mot, fini ou infini, u de A^∞ se factorise de façon canonique en un mot fini u' de longueur minimale et un mot infini u'' équitable (éventuellement vide).*

Exemple. Le mot $u = abac$ admet une factorisation $u = u'u''$ avec $u'' = \lambda$. Le mot $u = ab^5a(ca)^\omega$ s' écrit $u = u'u''$ avec $u' = ab^5$ et $u'' = (ac)^\omega$; en effet, $ab^5a(ca)^\omega$ n'est pas la factorisation cherchée car la longueur de ab^5a est strictement plus grande que celle de ab^5.

Donnons une caractérisation complète des mots qui se divisent l'un l'autre.

Lemme 2.3. *Soient $u, v \in A^\infty$. u/v et v/u si et seulement si il existe un mot fini u' et un sous-alphabet $B \subseteq A$ tels que $u = u'u''$, $v = v'v''$ avec u'' et v'' deux mots équitables sur B.*

Preuve. L'implication de droite à gauche étant évidente, montrons la réciproque. Par hypothèse, u/v et v/u donc u et v ont la même longueur. Si la longueur de u est finie, alors $u = v$ et le lemme est trivialement vérifié. Supposons donc que $|u| = \infty$; en utilisant le Lemme 2.2, on sait qu'il existe deux sous-alphabets B_1 et B_2 de A tels que $u = u'u''$, $v = v'v''$ avec u', $v' \in A^*$ et u'', v'' équitables sur B_1, B_2.

Montrons que B_1 est inclus dans B_2.

$a \in B_1 \Leftrightarrow u = u'au_1au_2a \ldots au_na \ldots$ avec u', $u_i (i > 0)$ des mots de A^*. Comme u/v, $a \in B_2$. D'où par symétrie on obtient $B_1 = B_2 = B$.

Posons donc $u = u'u''$, $v = v'v''$ avec u', $v' \in A^*$ et u'', v'' équitables sur B. Montrons maintenant que $u' = v'$. Supposons les différents, donc, par exemple, u' différent du

mot vide. u' peut donc s'écrire $u' = u'_1 c$ avec $u'_1 \in A^*$ et $c \notin B$. Il est facile de voir que $v' \neq \lambda$. Posons donc $v' = v'_1 d$ avec $v'_1 \in A^*$ et $d \notin B$, Comme u/v on a $u'_1 c / v'_1 d$ donc aussi $v'_1 d / u'_1 c$ puisque v/u. D'où $u' = v'$ ce qui est en contradiction avec l'hypothèse choisie d'où $u' = v'$. □

Corollaire 2.4. *Pour tout $u \in A^\infty$ il existe deux mots finis f et g tels que $fg^\omega / u / fg^\omega$.*

Définition. Soit $n > 0$, on dit que deux mots u et v de A sont n-équivalents s'ils ont les mêmes sous-mots de longueur n. On notera $u \equiv_n v$ si pour tout $x \in A^\infty$ $(|x| = n \Rightarrow (x/u \Leftrightarrow x/v))$.

Lemme 2.5. *Soient u, $v \in A^\infty$ et $n > 0$. Soient $u = u'u''$ et $v = v'v''$ les factorisations selon le Lemme 2.2 et $n > \max\{|u'|, |v'|\}$. Alors $u \equiv_n v \Rightarrow \text{alph}(u'') = \text{alph}(v'')$.*

3. Théorèmes de Higman et de Simon appliqués aux mots infinis

Théorème 3.1. *Soit $\{u_n\}$ une suite à éléments dans A^∞. Alors il existe une sous-suite $\{u_{i_n}\}$ croissante pour le pré-ordre /.*

Preuve. Avec le Lemme 2.2 chaque u_n se décompose de facon canonique en un mot fini u'_n et un mot équitable infini u''_n. Comme A est fini, il ne contient qu'un nombre fini de parties; on peut donc extraire de la suite $\{u_n\}$ une sous-suite $\{u_{r_n}\}$ telle que tous les u''_{r_n} sont écrits sur le même sous-alphabet de A. On pose, pour tout n, $v'_n = u'_{r_n}$ et $v''_n = u''_{r_n}$.

On applique le théorème de Higman à la suite de mots finis v'_n donc il existe une suite $\{q_n\}$ d'entiers telle que, pour tout n, v'_{q_n} divise $v'_{q_{n+1}}$ (1). Comme d'autre part tous les v''_n sont équitables sur un même alphabet on utilise le Lemme 2.3 pour dire que: pour tout n, v''_{q_n} divise $v''_{q_{n+1}}$ (2). En réunissant (1) et (2) on obtient avec le Lemme 2.1.

Pour tout n, $v'_{q_n} v''_{q_n}$ divise $v'_{q_{n+1}} v''_{q_{n+1}}$; c'est-à-dire, pour tout n, $u_{n_{q_n}}$ divise $u_{n_{q_{n+1}}}$. La suite cherchée est donc $\{u_{n_{q_n}}\}$. □

Introduisons les notations suivantes.

Notations. $S(n, u)$ est l'ensemble des sous-mots de u de longueur n, $\text{equ}(u, v)$ désignera l'ensemble $\text{equ}(u) \cup \text{equ}(v)$. Rappelons que $u = u'u''$ est la décomposition canonique (selon le Lemme 2.2) d'un mot u en un mot fini u' de longueur minimale et un mot u'' équitable.

Théorème 3.2. *Soient u et v deux mots de A^∞ n-équivalents. Alors il existe au moins un mot w de A^∞ tel que u/w, v/w et tel que $u \equiv_n w \equiv_n v$.*

La preuve de ce théorème repose sur les deux lemmes suivants.

Lemme 3.3. *Pour tout mot u de A^∞ et pour tout entier n il existe un mot fini x, facteur gauche de u, tel que $u \equiv_n x$.*

Lemme 3.4. *En utilisant les notations introduites ci-dessus on a: $u \equiv_n v$ implique que l'ensemble des mots de longueur n sur l'alphabet* $\mathrm{equ}(u'', v'')$ *est inclus dans $S(n, u) = S(n, v)$.*

Preuve du Lemme 3.3. On a $u = u'u''$ avec $\mathrm{equ}(u'') = \mathrm{alph}(u'')$. Notons $\mathrm{equ}(u'') = \{a_1, \ldots, a_k\}$. Soit p le plus petit entier tel que $u''[p] = u_1^1 a_1 u_2^1 a_2 u_3^1 a_3 \ldots u_k^1 a_k u_1^2 a_1 u_2^2 \ldots u_k^2 a_k \ldots u_k^n a_k$ avec tous les $u_i^j \in A^*$ pour $i = 1, \ldots, k$ et $j = 1, \ldots, n$. Le nombre p existe ($p = 0$ si $u \in A^*$) car u'' est équitable, infini ou égal au mot vide, sur $\{a_1, \ldots, a_k\}$. Comme $S(n, u) = S(n, u'u'') = \bigcup_{q=0}^{m} S(q, u') . S(n-q, u'')$, avec $m = \min(|u'|, n)$, et que $S(n-q, u'') = (\mathrm{alph}(u''))^{n-q} = S(n-q, u''[p])$ pour tout $q \leqslant n$, on a alors $S(n, u) = S(n, x)$ avec $x = u'u''[p]$. □

Preuve du Lemme 3.4. On procède par recurrence sur n.

$$n = 1 \Rightarrow u'u'' \equiv_1 v'v'' \Leftrightarrow \mathrm{alph}(u'u'') = \mathrm{alph}(v'v'') \Rightarrow \mathrm{equ}(u'', v'') \subseteq \mathrm{alph}(u'u'');$$

donc $\mathrm{equ}(u'', v'') \subseteq S(1, u) = S(1, v)$. Supposons maintenant que $(\mathrm{equ}(u'', v''))^n$ est inclus dans $S(n, u)$.

On a $u'u'' \equiv_{n+1} v'v'' \Rightarrow u'u'' \equiv_n v'v''$ et aussi $S(n+1, u'u'')$ contient l'ensemble $S(n, u'u'') . S(1, u'') \cup S(n, v'v'') . S(1, v'')$.

Avec l'hypothèse de recurrence on a: $(\mathrm{equ}(u'', v''))^n . (S(1, u'') \cup S(1, v'')) \subseteq S(n+1, u'u'')$, c'est-à-dire, $(\mathrm{equ}(u'', v''))^{n+1} \subseteq S(n+1, u'u'')$. D'où le résultat cherché. □

Preuve du Théorème 3.2. D'après le Lemme 3.3 il existe deux entiers p_1 et p_2 tels que $u'u''[p_1] \equiv_n v'v''[p_2]$. D'autre part en utilisant le Lemme 3.4 on obtient que $(\mathrm{equ}(u'', v''))^n \subseteq S(n, u)$. On applique alors le théorème de Simon aux deux mots finis suivants $u'u''[p_1]$ et $v'v''[p_2]$. Il existe donc un mot fini w' tel que $u'u''[p_1]$ et $v'v''[p_2]$ divisent w' et tel que $u'u''[p_1] \equiv_n w' \equiv_n v'v''[p_2]$. Pour tout mot infini w'' équitable sur $\mathrm{equ}(u'', v'')$ on a $S(n, w') = S(n, w'w'')$.

En effet, $S(n, w'w'') = \bigcup_{q=0}^{m} S(q, w') \cdot S(n-q, w'')$ avec $m = \min\{|w'|, n\}$. Or $S(n-q, w'') = (\mathrm{equ}(u'', v''))^{n-q}$ et on a vu (Lemme 2.2) que $(\mathrm{equ}(u'', v''))^n$ est inclus dans $S(n, w')$ donc $S(n, w'w'') = S(n, w')$.

De plus, $u/w'w''$, $v/w'w''$ donc $w = w'w''$ est le mot cherché. Remarquons que si le cardinal de $\mathrm{equ}(u'', v'')$ est plus petit ou égal à 1 alors il y a au moins un mot w; sinon il y a une infinité non dénombrable de mots infinis w. □

Remerciements

Les discussions que j'ai eu avec Jean Eric Pin et Dominique Perrin furent très fructueuses. Maurice Pouzet m'envoya une bibliographie sur le théorème de Higman et ses développements. Je remercie, enfin, les référees anonymes qui ont lu de près les premières versions de ce texte et m'ont permis par leur nombreuses remarques de l'améliorer grandement.

Bibliographie

[1] G. Higman, Ordering by divisibility in abstract algebras, *Proc. London Math. Soc.* **2** (1952) 326-336.
[2] J.B. Kruskal, The theory of well quasi-ordering: A frequently discovered concept, *J. Comput. Theory, Ser. A.*, **13** (1972) 297-305.
[3] M. Lothaire, Combinatorics on words, in: G.-C. Rota, ed., *Encyclopedia of Mathematics and its Applications Vol. 17* (Addison-Wesley, Reading, MA, 1983).
[4] C. St. J. A. Nash-Williams, On well quasi-ordering transfinite sequences, *Proc. Cambridge Phil. Soc.* **61** (1965) 33-39.
[5] R. Rado, Partial well-ordering of sets of vectors, *Mathematika* **1** (1954) 89-95.
[6] I. Simon, Piecewise testable events, in: H. Brakhage, ed., *Automata Theory and Formal Languages*, Lecture Notes in Computer Science **33** (Springer, Berlin, 1975) 214-222.