Institut Camille Jordan
UMR 5208 du CNRS

Local and Global Regularities in Infinite Words

Mickaël Postic
Thèse de doctorat

Université Claude Bernard Lyon 1

Université Claude Bernard Lyon 1
École doctorale InfoMaths (ED 512)
Spécialité: Mathématiques
n^{o}. d'ordre: 2020LYSE1248

Local and Global Regularities in Infinite Words

Thèse présentée en vue d'obtenir le diplôme de
Doctorat de l'Université de Lyon

soutenue publiquement le 17 décembre 2020 par
Mickaël POSTIC

devant le jury composé de :

Gabriele FICI	(Università di Palermo)	Directeur
Anna FRID	(Université Aix-Marseille)	Examinatrice
Edita PELANTOVÁ	(Czech Technical University)	Rapporteuse
Narad RAMPERSAD	(University of Winnipeg)	Examinateur
Frank WAGNER	(Université Lyon 1)	Examinatrice
Luca ZAMBONI	(Université Lyon 1)	Directeur

suite aux rapports de :

Edita PELANTOVÁ	(Czech Technical University)
Jeffrey SHALLIT	(University of Waterloo)

Remerciements

L'ensemble des interactions qui ont contribué à amener à l'écriture de cette thèse étant trop important pour rentrer dans les quelques lignes que comporteront ces remerciements, je serai obligé de m'en tenir aux quelques-unes les plus importantes ou les plus reliées à cette thèse proprement dite. Je n'oublie évidemment pas pour autant tous ceux et toutes celles qui n'y figurent pas !

La logique veut que je commence ces remerciements par mes directeurs de thèse, sans lesquels il est certains que, si thèse il y avait eu, elle serait bien différente du manuscrit actuel ! Luca tout d'abord, merci avant tout pour les mathématiques, pour tout ce que tu as pu m'apporter, tant ton aide sur l'ensemble des résultats de cette thèse que sur les sujets que tu m'as faits découvrir, discuter mathématiques avec toi fut toujours un plaisir et une source de motivation. Mais nos discussions ne se sont pas contentées de parler de mathématiques, merci de m'avoir apporté ton point de vue riche de la connaissance d'un autre système sur nombre de sujets. Merci aussi de m'avoir soutenu dans ma croisade ferroviaire contre l'avion lors de mes déplacements, en particulier celui à Pise, ou encore de ton aide précieuse lors des démarches administratives. J'espère que les occasions d'échanger seront encore nombreuses à l'avenir. Gabriele, merci pour ces deux séjours en Sicile, les conseils sur place, les parties d'échecs et bien sur les discussions mathématiques. J'espère bien revenir lorsque la situation se sera tassée !

L'autre mathématicien sans lequel cette thèse n'aurait certainement pas existé, c'est Jean-Paul Allouche, qui m'a introduit à la combinatoire des mots avec un tel brio qu'il a réussi à me motiver à nouveau aux mathématiques, et ce n'était pas gagné. Jean-Paul, merci en particulier de m'avoir encadré pendant mon stage de M2, et de ta gentillesse lors de chacune de nos discussions, ainsi que pour m'avoir recommandé auprès de Luca.

Relire une thèse en anglais écrite par un francophone, pas particulièrement doué pour les langues qui plus est, n'a probablement rien d'une sinécure ; pourtant, vous le fîtes avec rigueur et précision, à en juger par vos rapports. Un grand merci donc, Edità et Jeffrey. Mes remerciements aussi aux membres de mon jury, Edità encore, Narad, Ana et Frank. Frank, merci aussi d'avoir fait partie de mon CST, merci aussi à Alessandro d'avoir accepté cette charge.

Les conférences m'ont permis de rencontrer de nombreux jeunes ou moins jeunes chercheurs de mon domaine, sans la présence de qui ces conférences auraient été bien moins attrayantes, merci en particulier à Julien Cassaigne pour les discussions sur des sujets aussi divers que le vin et les voyages en transsibérien (l'amour du train est réellement le point fort de la communauté de combinatoire des mots !)

Finalement, d'un point de vue académique, j'aimerais remercier trois de mes enseignants, Messieurs Garcin, Cavaillès, aka JAC, et Cognet. Monsieur Garcin, merci de m'avoir donné goût à ce point aux mathématiques, quand j'enseigne, c'est vous mon exemple, et ma vocation d'enseignant, je vous la dois largement. Monsieur Cognet, merci de m'avoir montré qu'on peut introduire un grain de folie dans son enseignement, je n'y excelle pas autant que vous mais je m'y emploie !

Merci aussi à l'ensemble des personnels de l'université que j'ai pu croiser pendant ces trois $+ \varepsilon$ années, à travers les enseignements ou l'administratif, qui ont été d'une efficacité et d'une gentillesse constante. Je remercie en particulier Fabienne Macro et Laurent Azema, qui m'ont aidé à arriver d'u point de vue technique et administratif à cette soutenance, malgré les obstacles parfois nus sur la route.

Ce qui a fait l'attrait de la thèse, à mes yeux, tient pour bonne part aux doctorant.e.s que j'ai pu croiser. Tout d'abord, celles et ceux que j'ai eu la chance d'avoir dans mon bureau, et en particulier merci Marion pour les repas de bureau quand je suis arrivé, ça aide pour l'intégration ! Merci Kiki pour ta bonne humeur et tes sourires permanents, merci David pour ton rire si prompt et le vin, merci Théo pour toutes les discussions et les casse-têtes, ton amour communicatif des maths, merci Tingxian et Jiao pour les découvertes culinaires. Gwladys, malgré nos divergences de points de vue, ta bonne humeur était un rayon de lumière dans mes journées. Olga, after you, the whisky will never taste the same anymore.

Quelque part, au moins dans mon coeur, le bureau 110 fut un peu le mien aussi, merci d'y avoir toléré mes incursions si fréquentes. Merci Sam, pour l'Ardèche, et pour être qui tu es (bon, sauf au volant). Ta façon d'être m'a marqué pour un moment. J'espère avoir l'occasion de retourner chez ce vieux Georges ! Mélanie, merci pour les jeux (que sera ma pause midi sans Schotten Totten), ta personnalité toujours à l'écoute, ta présence dès qu'on a besoin d'un service. Merci de m'entretenir des aventures de Patrick ! Rémi, merci d'avoir ajouté une bonne dose d'aléatoire dans nos parties de Hanabi, c'eût été trop facile sans. Lola, merci pour tes sourires si lumineux à chaque fois que je venais vous déranger. Vincent, sans toi, ma vie à Lyon, ça n'aurait clairement pas été aussi sympa. Alors merci. Pour le jonglage, pour l'ultimate, pour les bouffes ensemble, pour les randos, les discussions, pour m'avoir présenté Loreena, Théo, Nick et Capu, et j'en passe ! C'est dur de résumer en deux mots tout ce que je te dois depuis que je suis à Lyon, mais merci pour tout ça (enfin, sauf *Old Boy*). Enfin, Marina, je ne sais même pas par où commencer... Et pourtant, c'est toi qui le dis, j'ai habituellement bonne blague ! Merci de m'assurer, déjà, grâce à toi même quand la vie ne tient qu'à un fil, j'ai confiance ! Merci aussi pour les discussions, les randonnées, l'enrichissement du vocabulaire, et le palais idéal du facteur cheval !

La transition est évidente : merci Simon A. pour le Beaufort, la montagne, l'impression d'avoir un pied-à-terre à Grenoble, les soirées avec ta mère, les chatons, et puis ta gentillesse et ta douceur qui me font tant de bien. Puisqu'à l'ICJ, Simon est décidément un de mes prénoms favoris, merci Simon Z. pour les vins d'Alsace, l'escalade, l'aide à la cueillette, ton approche résolument optimiste qui me change de mon propre pessimisme. Comment ne pas mentionner aussi ta cuisine ! Cat, merci pour le jardinage. Hugo, merci pour ta mesure dans chaque discussion.

Merci enfin à tous les autres, Octave pour toujours me surveiller, Simon (encore un) chapeau d'être si vert, JC, Thomas G, merci pour toutes les discussions liées à l'agreg et ton humour si abrasif, Félix, Christian, ...

Tant que je suis dans les matheux, j'ai aussi une pensée pour ceux de l'ENS, qui m'accompagnent de plus ou moins loin depuis un bon moment maintenant ! Batman, Guillaume, Xavier, Abel, Julie, je suis ravi d'avoir partagé votre existence, que ce soit pour des jeux de société, des visites guidées de Marseille ou Montpellier, ou des randos pleines de particules intriquées et de containers de nuggets. Merci particulièrement à Maxence de m'avoir accompagné das la gestion du BDS il y a de ça quelques années, et à Benoît de m'avoir appris où se situait le sud-est.

Ma vie en dehors des maths est animée par un autre thème majeur : le badminton. Au-dela du plaisir que j'éprouve à jouer, c'est aussi mon lien social avec le monde non mathématique depuis maintenant plus de 10 ans. Les gens que j'ai pu y rencontrer furent le vrai sel de ce sport : que ce soit à Carrières, Montrouge, Oullins, Lyon ou Villeurbanne, que des belles rencontres. Et bien évidemment, je n'oublie pas Alexia, Fabien, Popo, Ludo et tous les autres joyeux lurons qui me donnent envie de retourner à Chauffailes tous les ans malgré les pyramides par 35° !

In this category without entirely belonging there, thank you Susie for all those boardgames nights (well, apart from those revolving around Dobble !)

C'est rare que je prenne le temps de les remercier, et pourtant je peux toujours compter sur eux, et ce depuis toujours : merci Papa et Maman de m'avoir soutenu jusqu'ici, et merci d'avoir essayé de répondre à mes nombreuses questions quand j'étais petit, ce n'est sûrement pas étranger à mon amour des sciences. Je sais qu'en cas de besoin, vous êtes toujours à mes côtés, et ce même si vous ne comprenez pas grand-chose à cette thèse ! Seb, merci d'avoir essuyé les pots cassés pour nous, et merci à toi et Fiona de me doter d'une nièce à chouchouter, Nico, Corentin, sans vous l'enfance aurait été moins riante, les repas de famille perdraient quelques décibels, et ça serait franchement triste. Bien qu'on n'ai pas de lien génétique, à mes yeux, vous êtes aussi ma famille : merci Ju et Dede d'être mes meilleurs amis depuis douze ans, et d'être là dans tout ce que j'aime, du badminton aux jeux de société en passant par les randos vélo et autre. Bien sûr, j'ai aussi une pensée pour mes grands-parents, qui réussissent l'exploit de nous nourrir et nous chérir tous malgré nos inévitables différences : un total de 34 petits enfants d'un côté, 9 de l'autre, ç'a dû être du boulot !

Le meilleur pour la fin : Célia, merci de me supporter, dans tous les sens du terme, depuis maintenant 8 ans. Merci de m'accepter comme je suis, et de partager tant de mes points de vue, de mes envies et de mes projets. Merci pour les bons moments, et les moins bons, merci de ne pas t'enfuir quand ça ne va pas !

Enfin, il reste une personne que je n'ai pas encore remerciée, et qui malheureusement n'aura jamais l'occasion de lire cette thèse : merci Papy de ton soutien indéfectible toutes ces années, et merci pour les arcs !

Introduction en Français

Cette introduction en français est destinée aux non-mathématiciens, et tente de rester accessible, mettant pour cela sous le tapis la rigueur et les définitions très précises. Le lecteur intéressé (et comprenant l'anglais) est renvoyé vers l'introduction en anglais, plus riche et plus précise.

La combinatoire des mots est une branche relativement récente des mathématiques, l'article fondateur du domaine étant généralement considéré comme étant celui d'Axel Thue en 1906. Cependant, malgré d'autres avancées telles que celles de Marston Morse sur les mots sturmiens, la combinatoire des mots n'a véritablement pris son essor que sous l'impulsion de Marcel-Paul Schützenberger et de ses coauteurs, entre autre aussi grâce à l'auteur fictif M. Lothaire, dans les années 1970 et 1980. Depuis cette époque, la combinatoire des mots est devenue un sujet de recherche très actif, notamment grâce aux liens très étroits du domaine avec l'informatique théorique : on peut par exemple citer les suites automatiques, grâce auxquelles j'ai découvert cette branche des mathématiques et sur lesquelles a porté mon stage de Master 2.

La combinatoire des mots est la branche des mathématiques traitant de l'étude des propriétés des mots finis ou infinis. Les propriétés étudiées sont variées, souvent liées à d'autres domaines des mathématiques comme la théorie des nombres ou la théorie des groupes. D'autres questions sont de nature plus algorithmique, puisque le domaine a, évidemment, un lien fort avec les questions informatiques d'étude de texte, comme par exemple la recherche de motifs.

Cette thèse regroupe des travaux ayant pour point commun des études de régularités globales et locales dans des mots infinis, bien que les sujets des différents chapitres soient globalement indépendants les uns des autres. Ce type de questions est fréquent en combinatoire des mots, l'article fondateur d'Axel Thue traitait justement d'un problème de régularité inévitable.

Le premier des trois chapitres présentés ici porte sur une généralisation des mots de Lyndon, le second sur l'étude de la complexité ouverte et fermée des mots infinis et le troisième sur les antipuissances dans les mots infinis. Le premier et le second chapitres sont peu ou prou des reproductions d'articles écrits au cours de la thèse, essentiellement modifiés au niveau de l'introduction. Le dernier chapitre contient un article publié ainsi qu'un résultat ayant donné lieu à un article sur arXiv, ainsi que plusieurs autres résultats encore trop épars pour une publication.

Définissons quelques notions incontournables à la compréhension du lecteur : un $mot\ fini$ est, comme dans le vocabulaire courant, une suite finie de $lettres$, c'est-à-dire d'éléments d'un $alphabet$ $\mathbb{A}$. Cependant, rien n'oblige l'alphabet à être l'alphabet romain, grec, japonais ou que sais-je, auxquels nous sommes habitués : l'alphabet est en fait un ensemble de symboles, souvent, mais pas toujours, fini. Ainsi, les alphabets considérés peuvent être par exemple $\mathbb{A} = \{0, 1\}$ dans le cas où l'on voudrait étudier du code informatique, ou bien $\mathbb{A} = \{A, C, T, G\}$ si l'on s'intéresse à l'étude de l'ADN. Dans certains cas, l'alphabet peut être infini, comme nous le verrons lorsque nous parlerons du mot de Zimin, défini sur l'alphabet $\mathbb{A} = \mathbb{N}$. Pour un mot fini u, on peut définir sa longueur $|u|$ comme le nombre de lettres qui le composent. L'ensemble des mots de longueur n est alors $\mathbb{A}^n = \{a_1 \cdots a_n | \forall i \in [[1, n]], a_i \in \mathbb{A}\}$ et l'ensemble des mots fini est $\mathbb{A}^+ = \bigcup\limits_{n=1}^{+\infty} \mathbb{A}^n$. D'autre part, il est aussi possible de définir l'ensemble des mots infinis sur $\mathbb{A}$, c'est-à-dire des suites à valeurs dans $\mathbb{A}$: on le dénote $\mathbb{A}^{\mathbb{N}}$ et on a alors $\mathbb{A}^{\mathbb{N}} = \{a_1 a_2 \cdots | \forall i \in \mathbb{N}^*, a_i \in \mathbb{A}\}$. À cet ensemble de mots, on rajoute généralement le mot vide ϵ, qui est le mot de longueur 0 (ne contenant aucune lettre) afin d'avoir un monoïde $\mathbb{A}^* = \mathbb{A}^+ \cup \{\epsilon\}$. Enfin, pour un mot fini ou infini w, on dit que u est un *facteur* de w si $w = pus$ pour p et s des mots, possiblement vides. Si $p = \epsilon$ on dit que u est un *préfixe* de w, et si $s = \epsilon$ on dit que u est un *suffixe* de w.

Le premier chapitre de cette thèse s'intéresse à une généralisation des mots de Lyndon et énonce le résultat suivant :

Théorème 1. *Pour tout alphabet fini $\mathbb{A}$, tout mot infini $x \in \mathbb{A}^{\mathbb{N}}$ admet une factorisation unique en produit décroissant de mots de Lyndon généralisés.*

Un mot de Lyndon fini est un mot qui est plus petit, au sens de l'ordre du dictionnaire (l'*ordre lexicographique*), que toutes les rotations de ce mot. Par exemple, $mot < otm < tmo$ donc mot est de Lyndon pour l'ordre du dictionnaire. On constate ici que $mot < ot$ et $mot < t$, donc mot est plus petit que ses suffixes pour l'ordre lexicographique. En fait, ceci est une propriété donnant une définition alternative des mots de Lyndon :

Définition 1. Un mot $u \in \mathbb{A}^+$ est de Lyndon s'il est strictement plus petit que tous ses suffixes propres.

On constate alors que cette définition peut s'étendre aux mots infinis. De la même façon, on définit donc les *mots de Lyndon infinis* :

Définition 2. Un mot $u \in \mathbb{A}^{\mathbb{N}}$ est de Lyndon s'il est strictement plus petit que tous ses suffixes propres.

Les mots de Lyndon ont de nombreuses propriétés et applications, l'une des plus importantes est la suivante : tout mot fini ou infini peut s'écrire de façon unique comme un produit décroissant de mots de Lyndon.

Reutenauer, d'abord seul puis avec Dolce et Restivo, a proposé de généraliser la définition des mots de Lyndon à un ensemble plus large d'ordres sur les mots : plutôt que de prendre l'ordre lexicographique du dictionnaire, ils ont étudié des ordres où la comparaison entre deux lettres dépend de la position considérée dans le mot. Par exemple, on pourrait considérer l'ordre entre deux lettres comme l'ordre de l'alphabet habituel si ces lettres sont en position impaires, et l'ordre inverse si elles sont en position paires. Ainsi, il serait possible d'avoir $ab < aa < ba$ dans notre nouveau dictionnaire. Un tel dictionnaire serait par ailleurs toujours fonctionnel puisque pour trouver un mot, il suffirait, comme dans notre dictionnaire actuel, de chercher lettre à lettre. Par exemple, pour mot, se rendre dans la section des m puis chercher la sous section des mo et ainsi de suite.

Munis d'un nouvel ordre basé sur cette idée (avec de petites modifications pour arriver à leurs fins), Reteunauer et ses coauteurs définissent les mots de Lyndons généralisés comme ceux étant plus petits que leur suffixes propres. Ils parviennent alors à démontrer que tout mot fini admet, là encore, une factorisation décroissante en produit de mots de Lyndon généralisés. Ils laissent ouverte la question de l'existence et l'unicité d'une telle factorisation dans le cadre des mots infinis. Avec Luca Zamboni, après avoir encore un peu plus généralisé la classe des ordres sur lesquels nous définissons les mots de Lyndon, pour ne garder que la propriété stipulant que pour chercher un mot, il suffit de chercher ses préfixes consécutifs, nous avons réussi à répondre à leur question ouverte, avec le Théorème 1.

Le second chapitre de cette thèse propose un analogue d'un théorème majeur de la combinatoire des mots, le théorème de Morse-Hedlund. Ce théorème s'intéresse à la *fonction de complexité* des mots infinis, c'est-à-dire la fonction comptant le nombre de facteurs de longueur n d'un mot donné. Précisément, il stipule qu'un mot w est *apériodique*, c'est-à-dire qu'aucun de ses suffixes n'est *périodique*, donc de la forme $(a_1 \cdots a_n)^\omega = a_1 \cdots a_n a_1 \cdots a_n a_1 \cdots$ si, et seulement si, sa fonction de complexité est non bornée.

L'objectif de ce chapitre est d'étendre ce résultat à deux nouvelles fonctions de complexité, les complexités ouvertes et fermées. Le résultat auquel ce chapitre aboutit est le suivant :

Théorème 2. *Soit $x \in \mathbb{A}^{\mathbb{N}}$ un mot infini sur un alphabet fini $\mathbb{A}$. Les propositions suivantes sont équivalentes :*

1. *x est apériodique ;*
2. $\limsup\limits_{n\to+\infty} \mathrm{Cl}_x(n) = +\infty$ *;*
3. $\liminf\limits_{n\to+\infty} \mathrm{Op}_x(n) = +\infty$.

La fonction Cl_x (respectivement Op_x) est la *fonction de complexité fermée* (resp. *ouverte*) qui compte le nombre de *facteurs fermés* (resp. *ouverts*) de longueur n de x.

Définition 3. Un facteur u est *fermé* s'il a un préfixe non vide qui est aussi un suffixe et n'apparaît que deux fois comme facteur de u, comme préfixe et comme suffixe. Cette notion est étudiée depuis longtemps, particulièrement en lien avec les système dynamiques, et connue sous le nom de *mot de premier retour*. Un facteur est $ouvert$ s'il n'est pas fermé.

L'article sur lequel se base ce chapitre est un travail joint avec Olga Parshina.

Le troisième chapitre traite de la notion qui m'a occupé pendant l'essentiel de ma thèse, les *antipuissances*. Les premières questions en combinatoires des mots, traitées par Axel Thue, sont des questions de *puissances* évitables ou inévitables : un mot binaire de longueur supérieure à trois ne peut éviter de contenir un carré, c'est-à-dire un facteur se répétant deux fois consécutivement. Par exemple, $abba$ contient le carré *bb*. Un carré est une puissance d'$ordre$ 2. Axel Thue a construit un mot infini sur un alphabet de trois lettres ne comptant aucun carré, donc les carrés sont évitables sur un alphabet trois lettres. Cette notion de puissance a rapidement été étendue au cas abélien, c'est-à-dire dans lequel deux mots sont équivalents s'ils ont le même nombre d'occurrences de chaque lettre : ab est donc équivalent à *ba*. Dans ce contexte, $abba$ devient alors un carré abélien. Savoir si les puissances abéliennes sont évitables dans un mot infini est une question qui a généré beaucoup de travaux, mais s'est finalement conclue par l'affirmative : certains mots infinis sur un alphabet de quatre lettres évitent les carrés abéliens.

Dans un article récent, Fici, Restivo, Silva et Zamboni ont mis en évidence des motifs apparaissant inévitablement dans tous les mots ne comptenant pas des puissances de tout ordre : les *antipuissances*.

Définition 4. Une antipuissance d'ordre k est, en quelque sorte, l'opposé d'une puissance d'ordre k : c'est une succession de k blocs de même longueur deux-à-deux distincts.

Par exemple, $abba$ est une 2-antipuissance, tandis que $abbaab$ n'est pas une 3-antipuissance. Dans leur article introduisant cette définition, Fici et al montrent que tout mot infini contient soit des antipuissances de tout ordre, soit des puissances de tout ordre. Plus précisément, pour tout k, ils exhibent une borne sur la longueur d'un mot fini ne contenant ni antipuissance d'ordre k, ni puissance d'ordre k.

Mes résultats liés à cette notion d'antipuissance sont de deux types : l'un d'entre eux améliore la borne sur la longueur assurant la présence d'antipuissances d'ordre k dans le cadre de mots points fixes de morphismes $reconnaissables$ tandis que les autres résultats s'attachent à des généralisations de la notion d'antipuissance au cas abélien.

La notion de reconnaissabilité est essentielle en théorie des substitutions, l'idée étant, pour un facteur quelconque d'un mot w point fixe d'un morphisme, de "désubstituer" ce facteur, c'est-à-dire de l'exprimer comme facteur de l'image par le morphisme d'un autre facteur le plus court possible de w. Je montre que dans le cadre des points fixes apériodiques de morphismes uniformes, on peut en toute position trouver des k-antipuissances de longueur proportionnelle à k :

Théorème 3. *Si σ est un morphisme primitif et m-uniforme, avec un point fixe apériodique x, il existe une constante $C = C(\sigma)$ telle que : $\forall y \in X(\sigma)$, $\forall n, k \in \mathbb{N}$, y contient une k-antipuissance dont la longueur des blocs est au plus Ck commençant en position n.*

Enfin, après avoir défini les antipuissances abéliennes comme des antipuissances pour la relation d'équivalence abélienne, nous avons, avec Gabriele Fici et Manuel Silva, montré les deux résultats suivants, qui fournissent des exemples de mots connus contenant ou non des antipuissances abéliennes :

Théorème 4. *Le mot de Sierpiǹski ne contient aucune 11–antipuissance, et en particulier il ne contient aucune 11–antipuissance abélienne.*

Théorème 5. *Tout mot de pliage contient des antipuissances abéliennes de tout ordre.*

Contents

Chapter 1

Introduction

1.1 Preamble and outline

Combinatorics on words is a relatively new branch of mathematics, which dates back to Axel Thue's article in 1906 on the avoidability of some patterns in infinite words. Alas, Thue's results were more or less forgotten for more than half a century, and apart from some results, especially those of Marston Morse and Gustave Hedlund, the field was not particularly active. It is mainly thanks to the work of Marcel-Paul Schützenberger and his colleagues that the topic started to be active again, partly thanks to their collective book under the fictive name M. Lothaire, *Combinatorics on Words*, published in 1983.

Combinatorics on words is the area of mathematics that focuses on the study of sequences of symbols. It has links with many other branches of mathematics, e.g. algebra, number theory, probability, symbolic dynamics and, of course, combinatorics. It is also linked to computer science, and is sometimes even classified as theoretical computer science rather than mathematics. One example of the links between computer science and combinatorics on words is automatic sequences, whose study supervised by Jean-Paul Allouche was my entry point in combinatorics on words.

This thesis focuses on local and global regularities arising in infinite words. Those questions are at the core of combinatorics on words; in fact, Axel Thue's results were studies of regularities arising, or not, in an infinite word. This manuscript is divided in three mutually independent chapters. The first chapter deals with a generalization of Lyndon words; it is based on an article co-written with Luca Zamboni and published in *Theoretical Computer Science.* The second chapter considers an analogous of the celebrated Morse-Hedlund theorem in the case of open and closed complexity functions; it is based on work done jointly with Olga Parshina and submitted but yet to be published. Finally, the third chapter is a study of the new notion of antipower, improving some existing bounds for certain class of words and introducing a possible generalization of the notion. The main results of this chapter are based on an article written jointly with Gabriele Fici and Manuel Silva and published in *Advances in Applied Mathematics*.

In the rest of this introduction, I will start by giving some usual definitions, and then I will discuss the topic of each chapter in more details.

1.2 Definitions and notation

The fundamental definition in this area of mathematics is that of $words$. Words are strings of $letters$ taken from an $alphabet$, a finite or infinite set that will always be denoted $\mathbb{A}$ in this thesis. Examples of alphabets are many; here are some examples of alphabet with some of their applications to fields outside combinatorics on words:

$$\mathbb{A} = \{0, 1\} \text{ is the alphabet used to study most of computer code,}$$

$$\mathbb{A} = \{A, C, T, G\} \text{ is the alphabet used to study DNA,}$$

$$\mathbb{A} = \{a, b, c, \dots, z\} \text{ is the classical alphabet and can be used to study words,}$$

$$\mathbb{A} = \mathbb{N} \text{ is an example of infinite alphabet we are using at some point of this thesis.}$$

Amongst other alphabets, we could also consider the set of UTF-8 characters that allows us to study text.

Remark 1. For $\mathbb{A}$ and $\mathbb{B}$ two finite alphabets with same cardinality, there are bijections from one to the other. For this reason, from a combinatorial point of view, $\mathbb{A} = \{0, 1\}$ and $\mathbb{A} = \{a, b\}$ are the same, and instead of specifying the alphabet, it is enough to specify its cardinality. In the previous case, we just refer to both as the $binary$ alphabet.

Now let us turn to $words$: words are finite or infinite strings of symbols taken from $\mathbb{A}$. For a finite word u, we call the $length$ of u and denote $|u|$ the number of symbols that constitute u.

Example 1.2.1. On the classical roman alphabet, the word $four$ is of length $|four| = 4$, as is the word $five$, while $|six| = 3$.

Hence, labelling $\mathbb{A}^n$ the set of words of length n over $\mathbb{A}$, we get:

$$\mathbb{A}^n = \{a_1 \cdots a_n | \forall i \in [[1, n]], a_i \in \mathbb{A}\}.$$

We also define the empty word as a word of length 0, denoted ϵ.

The set of nonempty finite words over $\mathbb{A}$ is denoted by $\mathbb{A}^+$, and the set of finite words is $\mathbb{A}^*$, we then have:

$$\mathbb{A}^+ = \bigcup_{n=1}^{+\infty} \mathbb{A}^n \text{ and } \mathbb{A}^* = \mathbb{A}^+ \cup \{\epsilon\}.$$

We will also consider infinite words in this thesis. Infinite words can be seen as sequences of elements of $\mathbb{A}$. Their set is denoted $\mathbb{A}^{\mathbb{N}}$:

$$\mathbb{A}^{\mathbb{N}} = \{a_1 a_2 \cdots | \forall i \in \mathbb{N}^*, a_i \in \mathbb{A}\}.$$

In fact, those words are sometimes called *right-infinite words* since it is also possible to define *bi-infinite words*:

$$\mathbb{A}^{\mathbb{Z}} = \{\cdots a_{-1}a_0a_1\cdots | \forall i \in \mathbb{Z}, a_i \in \mathbb{A}\}.$$

In this thesis, the infinite words we are considering will always be right-infinite words, I will use either of the denomination without distinction.

For the rest of this thesis, I will try to use a and b for letters, while u and v will refer to finite words and x and y will be used for infinite words. I will use w and z for words that are either finite or infinite. As is usually the case, n and m will mostly stand for integers.

One fundamental operation that can be applied to elements of $\mathbb{A}^*$ is the $concatenation$. For $u = u_1 \cdots u_n$ and $v = v_1 \cdots v_m$ two elements of $\mathbb{A}^*$, we write uv the concatenation of those two elements, that is, the word $u_1 \cdots u_n v_1 \cdots v_m$. This operation can be extended to $\mathbb{A}^* \times \mathbb{A}^{\mathbb{N}}$ in the same way: for $u = u_1 \cdots u_n$ and $w = w_1 \cdots$ one has $uw = u_1 \cdots u_n w_1 \cdots$.

Example 1.2.2. For $u =$ lock and $v =$ down, we get $uv =$ lockdown. Also, we can see that the concatenation is not commutative, since $vu =$ downlock $\neq uv$.

A *factor* $f \in \mathbb{A}^*$ of a finite word u is a word such that there exist $p, s \in \mathbb{A}^*$ with $u = pfs$. If $p = \epsilon$ we say that f is a *prefix* of u, and if $s = \epsilon$ we say that f is a *suffix* of u. Those notions can be extended to infinite words: for $w \in \mathbb{A}^{\mathbb{N}}$ we say s is a *suffix* of w if there exists $p \in \mathbb{A}^*$ with $w = ps$; we say that $p \in \mathbb{A}^*$ is a *prefix* of w if for some suffix s of w one has $w = ps$. Then, a $factor$ of w is a factor of one of its prefixes. For any given word w, finite or infinite, we write $\mathrm{Fact}(w)$ the set of its factors, and for any integer n, we write $\mathrm{Fact}_n(w)$ the set of its factor of length n, so

$$\mathrm{Fact}_n(w) = \mathrm{Fact}(w) \cap \mathbb{A}^n.$$

If u is a finite word, we say that u is a *pure power* or simply a *power* if there exist a non-empty word v and an integer n with $n > 1$ such that $u = v^n$. Otherwise, we say that u is *primitive*. Two words u and v are said to be conjugates if there exist u_1 and v_1 such that $u = u_1v_1$ and $v = v_1u_1$.

Example 1.2.3. The word $papa$ is a pure power, while dad is a primitive word conjugated with add.

It is possible to define a distance on the sets $\mathbb{A}^*$ and $\mathbb{A}^{\mathbb{N}}$. An usual way to do this is the following: for w and z words, we define the distance $d(w, z)$ by

$$d(w,z) = \begin{cases} 0 & \text{if } w = z \\ 2^{-|w|} & \text{if } w \text{ is a prefix of } z \\ 2^{-|z|} & \text{if } z \text{ is a prefix of } w \\ 2^{-j} & \text{where } j \text{ is the first position where } w_j \neq z_j \end{cases}.$$

This allows us to define a topology on the set of words, finite and infinite, and to say that a sequence of words $(u_n)_{n\in\mathbb{N}} \in (\mathbb{A}^+)^{\mathbb{N}}$ is such that $u_n \to x$ for $x \in \mathbb{A}^{\mathbb{N}}$ when for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ with $n \geq N \Rightarrow d(u_n, x) < \epsilon$. So we can define, for finite words $u \in \mathbb{A}^*$ and $v \in \mathbb{A}^+$ the word $uv^\omega = \lim\limits_{n\to\infty} uv^n$.

For w an infinite word, we say that w is *purely periodic* if there exists $u \in \mathbb{A}^+$ such that $w = u^\omega$. If there exists $u \in \mathbb{A}^+$ and $v \in \mathbb{A}^+$ such that $w = uv^\omega$ we say that w is *ultimately periodic*. If w is not ultimately periodic we say that w is *aperiodic*.

One way to construct infinite words is by using *morphisms*. For two alphabets $\mathbb{A}$ and $\mathbb{B}$, a morphism σ from $\mathbb{A}^*$ to $\mathbb{B}^*$ is an application such that $\sigma(uv) = \sigma(u)\sigma(v)$ for all $u, v \in \mathbb{A}^*$. If for every letter a in $\mathbb{A}$ we have $|\sigma(a)| > 1$ and if there exists a letter $a' \in \mathbb{A}$ such that a' is a prefix of $\sigma(a')$ it is easy to see that the sequence of words $(\sigma^n(a'))_{n\in\mathbb{N}}$ admits a limit in $\mathbb{A}^\mathbb{N}$ denoted by $\sigma^\omega(a')$. If there exists an integer m such that for every letter $a \in \mathbb{A}$ one has $|\sigma(a)| = m$ we say that σ is an m*-uniform morphism*.

Example 1.2.4. In his original article, Axel Thue considered the 2-uniform morphism

$$\mu : \begin{cases} 0 & \to 01 \\ 1 & \to 10 \end{cases}.$$

The word **t**$= \mu^\omega(0) = 01101001100101101001\cdots$ is called the *Prouhet-Thue-Morse* word (as it was used independently by those three mathematicians) and is widely used in combinatorics on words.

For an infinite word, a property that is interesting from many points of view is recurrence: a word $w \in \mathbb{A}^\mathbb{N}$ is said to be *recurrent* if each of its factors appears twice (which is the same as saying that every factor appears infinitely often, or saying that any suffix s of w verifies $\mathrm{Fact}(s) = \mathrm{Fact}(w)$).

For a recurrent or non recurrent infinite word w, it is always possible to define the set of its *recurrent factors* $\mathrm{RecFact}(w)$ as the set of factors that appear infinitely often. It is easy to see that on a finite alphabet this set is always nonempty. Using this set, it is easy to characterize recurrent words:

$$(w \text{ is a recurrent word}) \Leftrightarrow (\mathrm{Fact}(w) = \mathrm{RecFact}(w)).$$

A property stronger than recurrence is *uniform recurrence*: an infinite word w is said to be *uniformly recurrent* if for any $u \in \mathrm{Fact}(w)$, there exists $n \in \mathbb{N}$ such that for all $v \in \mathrm{Fact}_n(w)$ we have $u \in \mathrm{Fact}(v)$.

Example 1.2.5. Consider the word $1010010100001010010100000\cdots$ which is the limit of the sequence of words defined by $u_0 = 1$ and for all $n \in \mathbb{N}$, $u_{n+1} = u_n 0^{n+1} u_n$. It is not uniformly recurrent, but it is recurrent. The word $1000\cdots = 10^\omega$ is not recurrent (and hence, not uniformly recurrent).

Finally, a notion that I will use in this thesis is the notion of *abelian equivalence*. For any finite word u we can define, for every letter $a \in \mathbb{A}$, the number of occurrences of this letter in the word u. We will label $|u|_a$ this quantity. Then we say that two words u and v are *abelian equivalent* and we write $u \sim_{ab} v$ if $|u|_a = |v|_a$ for every letter a. Many of the definitions in combinatorics on words have a counterpart in the abelian setting; for example, an *abelian power* of order n is concatenation of n consecutive abelian equivalent words.

1.3 Content of the thesis

In this part, I will introduce the notions used in the next chapters of this thesis and give the main results of each chapter in this work; I will keep this relatively short since each chapter starts with its own introduction which redefines the needed notions and introduces the topic more precisely.

1.3.1 Generalized Lyndon words

This chapter presents some results obtained together with Luca Zamboni and published in *Theoretical Computer Science* [PZ19b] on factorizations of infinite words as non-increasing products of some generalization of Lyndon words. The article on which this chapter is based solved a question asked by Dolce, Restivo and Reutenauer who defined this generalization and proved the existence of such a factorization in the case of finite words. This chapter also extends a bit the generalization of Lyndon words, considering Reutenauer et al.'s generalization as a particular case. The aim of this chapter is to prove the following theorem:

Theorem 1.3.1. *Each infinite word $x \in \mathbb{A}^{\mathbb{N}}$ admits precisely one ω-Lyndon factorization.*

In combinatorics on words, it is often convenient to order the words we are considering. The most common way to do so is to first define an order on the alphabet, then define the *lexicographical order* on $\mathbb{A}^+$ the following way: for u and v in $\mathbb{A}^+$, we say $u < v$ when u is a proper prefix of v or if there exists a word - possibly empty - w and two letters a and b in $\mathbb{A}$ verifying $a < b$ such that wa is a prefix of u and wb is a prefix of v. This order is the order used to order words in a dictionary for a natural language.

Definition 1.3.2. A word $w \in \mathbb{A}^+ \cup \mathbb{A}^*$ is said to be Lyndon if w is smaller than all of its strict nonempty suffixes.

Those words have numerous applications (and they were introduced as a tool in algebra). One of their applications is of particular interest to us: they provide a factorization of $\mathbb{A}^+$ and $\mathbb{A}^*$.

Theorem 1.3.3 (Lyndon)**.** *Let $w \in \mathbb{A}^+$. Then w admits a unique non-increasing factorization as product of Lyndon words, i.e.*

$$w = w_1 w_2 \cdots w_n \textit{ where } w_1 \geq \cdots \geq w_n \textit{ and each } w_i \textit{ is Lyndon}.$$

Theorem 1.3.4 (Factorization of infinite words as non-increasing product of Lyndon words [SMDS94])**.** *Let $w \in \mathbb{A}^{\mathbb{N}}$. There exists a unique factorization:*

$$w = w_1 w_2 \cdots w_n s$$

$$\textit{with } \forall i,\ w_i \textit{ finite Lyndon word},\ s \textit{ infinite Lyndon word and } w_1 \geq w_2 \geq \cdots \geq w_n > s$$

or

$$w = \prod_{i=1}^{\infty} w_i \textit{ with } \forall i,\ w_i \textit{ finite Lyndon word and } w_1 \geq w_2 \geq \cdots .$$

In different articles, Reutenauer [Reu05], then Dolce, Restivo and Reutenauer [DRR18, DRR19] describe a new set of lexicographical-like orders respective to which they define *generalized Lyndon words*. The orders they consider are such that the comparison between two letters depends on the position those letters occupy in the word. This allows them to define a broader class of Lyndon words.

In that context, they were able to prove that Theorem 1.3.3 still holds, and asked whether Theorem 1.3.4 still holds. We answered this question in the affirmative with Luca Zamboni, and even improved a bit by generalizing further the class of lexicographical-like orders for which the statement holds.

1.3.2 Open and closed complexity

This chapter of the thesis presents an analogous of the celebrated Morse-Hedlund theorem in the case of open and closed complexity functions we found together with Olga Parshina.

On a finite alphabet, every word w, finite or infinite, satisfies the following property:

$$\forall n \in \mathbb{N},\ |\operatorname{Fact}_n(w)| < +\infty.$$

This allows us to define, for any given word w, the *complexity function* of w:

Definition 1.3.5. For a given word w, we denote by p_w its complexity function defined by

$$p_w : \begin{cases} \mathbb{N} & \to \mathbb{N} \\ n & \to \operatorname{Card}(\operatorname{Fact}_n(w)) \end{cases}.$$

The Morse-Hedlund theorem then provides a link between aperiodicity and complexity function for a given word:

Theorem 1.3.6 (Morse-Hedlund[MH38]). *An infinite word w is aperiodic if and only if its complexity function is unbounded. If w is aperiodic, we have, for any integer n, $p_w(n) \geq n+1$.*

This theorem has many variants in other settings in combinatorics on words, we were interested in trying to prove it in the case of the complexity functions pertaining to a well studied local regularity notion and its counterpart, *open* and *closed words.*

Definition 1.3.7. Given $u, w \in \mathbb{A}^+$ with $|u| < |w|$, we say u is a *border* of w if u is both a prefix and a suffix of w. We say $w \in \mathbb{A}^+$ is *closed* if either $w \in \mathbb{A}$ or w admits a border u which occurs precisely twice in w. Otherwise w is said to be *open*. Thus $w \in \mathbb{A}^+$ is closed if either $w \in \mathbb{A}$ or if its longest border u occurs exactly twice in w, i.e., u has no internal occurrences in w. The longest border of a closed word is called its *frontier*.

The notion of closed word has also been intensively studied under the name *complete first return* by researchers working in symbolic dynamics.

Like in the classical case, for a word w, we can define its *open complexity function* Op_w and its *closed complexity function* Cl_w that count the number of open, or closed, factor of each length, respectively.

Then, it is possible to prove an equivalent of Morse and Hedlund's theorem in the setting of open and closed complexity functions. That question was looked at by other mathematicians since the situation is less clear than in the classical complexity function case. In fact, in the classical case, the complexity function is increasing on $\mathbb{N}$, which is not the case for the two complexity functions we defined, and that made the problem harder to solve.

The main result of this chapter is the following theorem:

Theorem 1.3.8. *Let $x \in \mathbb{A}^{\mathbb{N}}$ be a right-infinite word over a finite alphabet $\mathbb{A}$. The following are equivalent:*

1. *x is aperiodic;*

2. $\limsup_{n \to +\infty} \mathrm{Cl}_x(n) = +\infty$*;*

3. $\liminf_{n \to +\infty} \mathrm{Op}_x(n) = +\infty$.

1.3.3 Antipowers in infinite words

In this chapter, we discuss the notion of antipower introduced recently by Fici, Restivo, Silva and Zamboni. Some of the results in this chapter were the subject of a joint article with Gabriele Fici and Manuel Silva published in *Advances in Applied Mathematics* [FPS19].

The first regularities that were studied in combinatorics on words were *powers*; in fact, the first article in combinatorics on words by Axel Thue in 1906 was about infinite words avoiding squares on a 3-letters alphabet.

The counterpart of this notion, introduced significantly more recently, is the notion of *antipower*.

Definition 1.3.9. A word of length kn, for k and n integers, is a (k, n)*-antipower*, or antipower of order k and length n, if it is concatenation of k pairwise distinct blocks of length n. Namely, $u = u_1 \cdots u_{kn}$ is a (k, n)-antipower if $u_{in+1} \cdots u_{(i+1)n} \neq u_{jn+1} \cdots u_{(j+1)n}$ for every i and j. Again, we will often write k-antipower, without giving n.

Example 1.3.10. The word $abba$ is a 2-antipower but not a 4-antipower.

The authors of the first paper on antipowers, as Axel Thue in his two papers, were looking at a type of question that would be describe, in today's words, as the intersection of Ramsey theory and combinatorics on words. Ramsey theory is the area of mathematics that looks at combinatorial structures and try to tell whether some regularities will arise if the combinatorial structure is large enough. Axel Thue proved that squares (i.e. powers of order 2) are unavoidable regularities on a binary alphabet but are avoidable on a ternary alphabet.

Fici, Restivo, Silva and Zamboni, on the other hand, proved that it is not possible to avoid both powers and antipowers on infinite words:

Theorem 1.3.11. *Every infinite word contains powers of any order or antipowers of any order.*

They also proved a version of this theorem for finite words:

Theorem 1.3.12 (Theorem 14 in [FRSZ18]). *For all integers $l > 1$ and $k > 1$ there exists $N = N(l, k)$ such that every word of length N contains a l-power or a k-antipower. Furthermore, for $k > 2$, one has $k^2 - 1 \leq N(k, k) \leq k^3\binom{k}{2}$.*

The results I present in this chapter are of two types: first, I give a better bound than the one in Theorem 1.3.12 for some class of words, then I study an extension of the definition of antipowers in the abelian setting.

Recall that an uniform morphism is a morphism of constant length on letters. Using the results on recognizability of uniform morphisms, mainly some results of Mossé, I was able to prove the following:

Theorem 1.3.13. *If σ is primitive and m-uniform morphism, with an aperiodic fixed point x, there exists a constant $C = C(\sigma)$ such that: $\forall n, k \in \mathbb{N}$, x contains a k-antipower with block length at most Ck starting at position n.*

In the second part of the chapter, I studied different versions of *abelian antipowers*.

The first notion of abelian antipower was introduced in a joint work with Gabriele Fici and Manuel Silva.

We first exhibited a well-studied word, the *Cantor* or *Sierpiǹski* word, that does not contain abelian antipowers of arbitrarily large order (although it contains abelian powers of every order):

Theorem 1.3.14. *The Sierpiǹski word s does not contain 11–antipowers, hence it does not contain abelian 11–antipowers.*

Remark 2. This result was improved in Riasat's thesis [Ria19], who proved that the optimal bound is 10.

Then, we proved that some class of words, the *paperfolding words*, known to contain abelian powers of every order, also contain abelian antipowers of every order:

Theorem 1.3.15. *Every paperfolding word $\boldsymbol{f}$ contains abelian m-antipowers for every $m \geq 2$.*

In order to complete the picture, I also exhibit a word that contain abelian antipowers of any order but no abelian squares: the *Zimin word*.

Theorem 1.3.16. *The Zimin word contains abelian antipowers of any order and no abelian square.*

Finally, I discuss a possible generalization of powers and antipowers where the focus is on the set of factors of a word, *total-abelian powers and antipowers*, for which I proved the following:

Theorem 1.3.17. *On a binary alphabet, every infinite word w contains total-abelian k-powers for every k in $\mathbb{N}$. Moreover, if the abelian complexity of w is not bounded, then w contains total-abelian k-antipowers for every k in $\mathbb{N}$.*

The initial aim was to prove an analogue of Theorem 1.3.11 in the abelian setting, I hope this problem will be solved soon!

Chapter 2

ω-Lyndon words

Contents

2.1 Classical Lyndon words

In this section we give a brief introduction to the basic properties of Lyndon words. A fundamental result is that every finite and infinite word may be written uniquely as a non-increasing concatenation of Lyndon words. This section does not contain new results.

2.1.1 Introduction

Lyndon words were first introduced by A. Shirshov [Shi53] under the name *regular words* in the study of some Lie algebras and were later independently studied by R. Lyndon [Lyn54], again as a tool related to generators of free Lie's algebras. They have been extensively studied since, having numerous applications in different fields of mathematics. Besides from the previously quoted link to free Lie algebras, in another application in algebra, D. Radford [Rad79] shows that the algebra of polynomials on Lyndon words with rational coefficients is a shuffle algebra. Lyndon words can also be used to construct variants of the Burrows-Wheeler transform, and have a wide range of applications in combinatorics on words. Of particular interest to us is the fact that Lyndon words provide a factorization of the free monoid [Lot97, CFL58, Duv83]. The extension of this property to a new class of Lyndon-like words is the main subject of this chapter.

2.1.2 Finite Lyndon words

2.1.2.1 Definitions

Throughout this chapter, we will consider an ordered alphabet $\mathbb{A}$. Recall the *lexicographical order* on $\mathbb{A}^+$ we defined in the introduction: for u and v in $\mathbb{A}^+$, we say $u < v$ when u is a proper prefix of v or if there exists a word - possibly empty - w and two letters a and b in $\mathbb{A}$ verifying $a < b$ such that wa is a prefix of u and wb is a prefix of v. This order is the order of the dictionary.

Example 2.1.1. If $\mathbb{A}$ is the Roman alphabet with its natural order, one has

$$three < two.$$

If $\mathbb{A} = \{0, 1\}$ with $0 < 1$ one has

$$101 < 11 < 111.$$

Remark 3. I will often use the following properties verified by the lexicographical order:

$$\forall u, v, w \in \mathbb{A}^+, u < v \Leftrightarrow wu < wv, \tag{2.1}$$

$$\forall u, v \in \mathbb{A}^+, w, z \in \mathbb{A}^*, v \notin u\mathbb{A}^+, u < v \Rightarrow uw < vz, \tag{2.2}$$

Recall that, for $n \in \mathbb{N}$, $u = u_1 \cdots u_n$ and $v = v_1 \cdots v_n$, we say that u and v are *conjugates* if there exists i such that $u_i \cdots u_n u_1 \cdots u_{i-1} = v$. This is an equivalence relation and the corresponding equivalence classes are called *conjugacy classes*.

Example 2.1.2. 100 and 001 are conjugates, and $\{abba,\ aabb,\ baab,\ bbaa\}$ is the conjugacy class of $abba$.

Lyndon words were introduced by R. Lyndon and A. Shirshov in the following way:

Definition 2.1.3. A finite word u is said to be *Lyndon* if for any non-trivial conjugate v of u, one has $u < v$.

Remark 4. Since pure powers are self-conjugates, an equivalent definition is that Lyndon words are the words that are the smallest in their conjugacy class and primitive.

Example 2.1.4. In our previous examples, taking the natural orders $0 < 1$ and $a < b$, 001 and $aabb$ are Lyndon words. With $a < p$, there is no Lyndon word in the conjugacy class of $papa$.

Example 2.1.5. The Lyndon words of length 5 on $\{0, 1\}$ are:

$$00001, 00011, 00111, 01111, 00101, 01011.$$

Remark 5. An equivalent definition of Lyndon words was given in the introduction:

Definition 2.1.6. A word $w \in \mathbb{A}^+$ is said to be Lyndon if w is smaller than any of its strict suffixes.

Proof. Let w be a word smaller than each of its proper suffixes, and write $w = uv$ with u and v nonempty. Then $w < v < vu$ hence w is Lyndon. Conversely, let w be a Lyndon word and write $w = uv$ with u and v nonempty. Since $w < vu$, either $w < v$ or v is a prefix of w. In the latter case, $w = uv = vz$ for some nonempty z. Then, since w is Lyndon we have $vz < vu$ and, applying (2.1), $z < u$. Then, using (2.2), we get $zv < uv$, contradicting the fact w is Lyndon. □

Remark 6. Using this definition, it is easy to see that a finite Lyndon word must be unbordered, as a prefix of a word is always smaller than the word itself.

Remark 7. Sometimes, Lyndon words are defined as words w for which there exists an ordering of $\mathbb{A}$ relative to which w is smaller than any of its proper suffixes, so ab and ba would both be Lyndon words, only not respectively to the same order. This is not the convention chosen in this thesis.

2.1.2.2 Factorization theorem for finite Lyndon

Probably one of the most important result on Lyndon words is the fact that they give a factorization of the free monoid. This theorem is a combinatorial equivalent of the famous Poincaré-Birkhoff-Witt theorem [Bir37]. It is also important in computer science, as it can be computed in linear or even logarithmic time (see [Duv83, AC95]), and has various applications (see, for instance, [CR20]). I will recall here some results and some proofs, as I want to highlight what doesn't work in a more general setting. The interested reader can see [Lot97] for more precision.

Proposition 2.1.7. *Let $w \in \mathbb{A}^+$ with $|w| \geq 2$. Then w is Lyndon if and only if there exist u, v Lyndon words with $u < v$ and $w = uv$.*

Proof. Let u and v be Lyndon with $u < v$. Let s be a strict suffix of uv. First suppose s is longer than v, hence $s = s'v$. Then we have $u < s'$ and hence $uv < s$ using (2.2). Now, if s is a suffix of v, we need to distinguish two cases: let us first suppose that $s \notin u\mathbb{A}^*$. Then, $u < v < s$ so using (2.2) gives $uv < s$. Finally, if $s = uz$, the inequality $s < uv$ leads to $z < v$ by (2.1), which contradicts the fact that v is Lyndon.

Let w be a Lyndon word with $|w| \geq 2$. Let v be the smallest proper suffix of w. It is Lyndon: by definition, it is strictly smaller than its suffixes. Write $w = uv$. Since $u < w < v$ we have $u < v$. Moreover, u is Lyndon: let s be a proper suffix of u such that $s < u$. Then, s is a prefix of w, and $v < sv$ leads to v prefix of sv, then smaller than w, a contradiction. □

Example 2.1.8. 00011 and 00101 are both Lyndon, and $00011 < 00101$ so using Proposition 2.1.7 we can say that 0001100101 is Lyndon as well.

Theorem 2.1.9 (Lyndon). *Let $w \in \mathbb{A}^+$. Then w admits a unique non-increasing factorization as product of Lyndon words, i.e.*

$$w = w_1 w_2 \cdots w_n \textit{ where } w_1 \geq \cdots \geq w_n \textit{ and each } w_i \textit{ is Lyndon.}$$

Proof. To prove the existence, we just consider the following algorithm: at each step, take the smallest suffix (not necessary proper) of w. Every element is a Lyndon word since, by definition, it is smaller than its suffixes. And this process also gives the non-increasing property of the factorization: if $w_i < w_{i+1}$, by Proposition 2.1.7 we have $w_i w_{i+1}$ Lyndon hence $w_i w_{i+1} < w_{i+1}$, contradicting the definition of w_{i+1}.

This decomposition is unique: suppose w admits two decomposition, $w = w_1 \cdots w_n = w'_1 \cdots w'_m$. Then $w_1 = w'_1$. Indeed, suppose $|w_1| > |w'_1|$. Then $w_1 = w'_1 w'_2 \cdots w'_i u$ with u nonempty prefix (not necessarily strict) of w'_{i+1}. Since w_1 is Lyndon, we get $w_1 < u \leq w'_{i+1} \leq w'_1 < w_1$. By induction, both decomposition are identical. □

Example 2.1.10. Here are some examples of factorizations given by this algorithm:
$010 = (01)(0), 00 = (0)(0), 01001011 = (01)(001011)$.

A natural question that arises is whether this result can be extended to the set of infinite words over $\mathbb{A}$. For some words, it just seems to work perfectly fine: take, for instance, $1010101 \cdots = (10)^\omega$. We can write $(10)^\omega = (1) \cdot (01) \cdot (01) \cdots$ and it is possible to prove that this decomposition is unique. Meanwhile, for other words there is no such decomposition: the word $011111 \cdots$ provides such an example. In order to get the same result for infinite words, we need the notion of infinite Lyndon words.

2.1.3 Infinite Lyndon words

2.1.3.1 Definitions

In a 1993 paper, R. Siromoney, L. Mathew, V.R. Dare and K.G. Subramanian [SMDS94] introduced the notion of infinite Lyndon words and showed that some properties of Lyndon words could be extended to this new setting, one of them being the unique factorization property.

It is possible to extend the lexicographic order to $\mathbb{A}^+ \cup \mathbb{A}^\mathbb{N}$ the following way: for x and y infinite words, with $x \neq y$, we say $x < y$ if $x = uaz$ and $y = ubz'$ with $u \in \mathbb{A}^*$ and $a, b \in \mathbb{A},\ a < b$. We say $v < x$ with $v \in \mathbb{A}^+$ when $v \leq x_1 \cdots x_{|v|}$ and $x < v$ otherwise.

Example 2.1.11.

$$01^\omega < 1^\omega$$

$$01 < 01^\omega < 1$$

Let us recall this definition we introduced in the introduction:

Definition 2.1.12. An infinite word $w \in \mathbb{A}^\mathbb{N}$ is said to be *infinite Lyndon* (or just *Lyndon*) if it is strictly smaller than each of its strict suffixes.

Example 2.1.13. The word 01^ω is an infinite Lyndon word, as is the word $010110111 \cdots$ where there are n 1s between the n-th 0 and the $(n+1)$-th 0.

Remark 8. This definition is the natural extension of Definition 2.1.6 in the case of infinite words.

Remark 9. An infinite Lyndon word does not admit a *prefixal factorization*, i.e. a factorisation where every element is a prefix. Indeed, if $w = w_1 w_2 \cdots$ where the w_i are prefixes of w and w is infinite Lyndon, then $w < w_2 w_3 \cdots$ and $w = w_2 z$ gives $w < z < w_3 w_4 \cdots$ so $z = w_3 z'$. Repeating this, we get $w = w_2 w_3 \cdots < w$, a contradiction.

This definition is equivalent to the following one [SMDS94]:

Proposition 2.1.14. *An infinite word $w \in \mathbb{A}^{\mathbb{N}}$ is Lyndon if and only if it is the limit of a sequence of Lyndon words, which is if w has infinitely many Lyndon prefixes.*

Proof. Suppose that w is not Lyndon. That means w admits a suffix y with $y < w$. Hence, there exists an integer n such that $y_1 \cdots y_n < w_1 \cdots w_n$. But then, any prefix of w longer than $w_1 \cdots y_n$ cannot be Lyndon, hence w has finitely many Lyndon prefixes.

Suppose w is Lyndon. Then w admits infinitely many Lyndon prefixes. Indeed, if u is a prefix of w which is not Lyndon, u admits a border v: let $u = u'v$ with $u > v$. Then v is a prefix of w, since $w < vz$ where $w = u'vz$. We know from Remark 9 that w does not admit a prefixal factorization, so applying Proposition 3 in [dZ16] we get that w admits infinitely many unbordered, hence Lyndon, prefixes. □

2.1.3.2 Factorization theorem for infinite words

With this new definition, some of the previous results can be extended:

Proposition 2.1.15 ([SMDS94]). *Let $u \in \mathbb{A}^+$ be a finite Lyndon word and let $v \in \mathbb{A}^{\mathbb{N}}$ be an infinite Lyndon word. Then uv is an infinite Lyndon word if and only if $u < v$.*

Proof. Using Proposition 2.1.7 and Proposition 2.1.14, this is immediate. □

Theorem 2.1.16 (Factorization of infinite words as non-increasing product of Lyndon words [SMDS94]). *Let $w \in \mathbb{A}^{\mathbb{N}}$. There exists a unique factorization:*

$$w = w_1 w_2 \cdots w_n s$$

with $\forall i,\ w_i$ *finite Lyndon word,* s *infinite Lyndon word and* $w_1 \geq w_2 \geq \cdots \geq w_n > s$

or

$$w = \prod_{i=1}^{\infty} w_i \text{ with } \forall i,\ w_i \text{ finite Lyndon word and } w_1 \geq w_2 \geq \cdots .$$

Proof. Let $w \in \mathbb{A}^{\mathbb{N}}$. First, we exhibit such a decomposition. Consider the prefixes of w. If an infinite number of those are finite Lyndon words, w is infinite Lyndon and so we get a factorization. If the set of Lyndon prefixes of w is finite, let w_1 be a Lyndon prefix of w appearing infinitely often in the finite Lyndon factorization of the prefixes of w. The (infinite) set of prefixes whose factorization starts with w_1 will be denoted $P_1(w)$. We also let $w = w_1 w_1'$. Again, there are two possibilities: either w_1' is an infinite Lyndon word or it admits a finite number of Lyndon prefixes. In the second case, we can look at the finite factorizations of the elements of $P_1(w)$, and like previously construct a w_2 and a set $P_2(w)$, infinite subset of the prefixes of w whose elements admit a finite Lyndon factorization starting with $w_1 w_2$, and we can iterate this process. If w_1' is infinite Lyndon, all we need to do is to prove $w_1 > w_1'$. If this is not the case, we can apply Proposition 2.1.15 to get that w already was a infinite Lyndon word, a contradiction.

Now we prove the uniqueness of such a factorization. Like in the finite case, we consider two factorizations $w = w_1 w_2 \cdots = w_1' \cdots$ with $w_1' \in \mathbb{A}^{\mathbb{N}}$ or $|w_1'| \geq |w_1|$, and proving the equality $w_1 = w_1'$ is enough to get, by induction, that both factorizations are the same. Let us then suppose that $w_1' \neq w_1$. Let then u be a Lyndon prefix of w longer than w_1 (if w_1' is finite we can choose $u = w_1'$; else w has an infinite number of Lyndon prefixes so we just pick one longer than w_1). Write $u = w_1 \cdots w_i z$ where z is a prefix of w_{i+1}. We have $w_1 < u < z \leq w_1$, a contradiction. □

2.2 Generalizations of Lyndon words

In this section we consider some generalizations of the lexicographic order which in turn give rise to a generalized notion of Lyndon words. We will show that many of the results in the previous section extend to this more general setting.

2.2.1 Generalized Lyndon words

2.2.1.1 Definition

In a 2005 article [Reu05] C. Reutenauer introduced a generalization $\prec_{\mathbb{N}}$ of the lexicographical order $<$. This order was studied again recently by C. Reutenauer, F. Dolce and A. Restivo [DRR18]. To define this new order, they consider a sequence $(<_n)_{n\in\mathbb{N}}$ of total orders on $\mathbb{A}$. This induces a total lexicographic-like order $\preceq_{\mathbb{N}}$ on $\mathbb{A}^{\mathbb{N}}$ defined by $x \preceq y$ if and only if either $x = y$ or if $x = uax'$ and $y = uby'$ for some $u \in \mathbb{A}^*$, $a, b \in \mathbb{A}$, $x', y' \in \mathbb{A}^{\mathbb{N}}$ and $a <_{|u|+1} b$. If $x \neq y$ and $x \preceq_{\mathbb{N}} y$ we write $x \prec_{\mathbb{N}} y$. This means that the order between two letters depends on the position those letters occupy in the word.

Example 2.2.1. Take $\mathbb{A} = \{a, b\}$, and for any non-negative integer k let $a <_{2k+1} b$ and $b <_{2k+2} a$, then $(ab)^\omega \prec_{\mathbb{N}} (aa)^\omega \prec_{\mathbb{N}} (ba)^\omega$.

Remark 10. As the previous example shows, this definition allows orderings that were not possible with the classical lexicographical order. Meanwhile, taking $<_n=<$ for any n gives the lexicographic order, so this new definition can be seen as a generalization of the lexicographic order.

This total order on infinite words in turn defines a relation on $\mathbb{A}^*$: for u and v in $\mathbb{A}^+$ the authors defined $u \preceq_{\mathbb{N}} v$ when $u^\omega \preceq_{\mathbb{N}} v^\omega$, and ε is chosen to be smaller than any nonempty word. This relation is not an order, since it is not anti-symmetric: $a \preceq_{\mathbb{N}} aa$ and $aa \preceq_{\mathbb{N}} a$ yet $a \neq aa$. If we decide that powers of the same word are not comparable, we get a partial order. For that reason, this relation will be referred to as an order, although it isn't technically one.

The authors then define generalized Lyndon words as the words strictly smaller than their proper nonempty suffixes with respect to this new lexicographic-like order $\prec_{\mathbb{N}}$:

Definition 2.2.2. A word $w \in \mathbb{A}^+$ is called generalized Lyndon if $w^\omega \prec_{\mathbb{N}} v^\omega$ for each proper suffix v of w.

Example 2.2.3. Taking the alternative orders of Example 2.2.1, we get that aba is a generalized Lyndon word; this example further shows generalized Lyndon words need not to be unbordered.

Remark 11. Many fundamental properties of usual Lyndon words no longer hold for generalized Lyndon words. First of all, every primitive finite word u and every non-periodic infinite word w is generalized Lyndon relative to some total order $\preceq_{\mathbb{N}}$ on $\mathbb{A}^{\mathbb{N}}$. Indeed, it suffices to look at u^ω (or w) and define $<_n$ such that the smallest element of $\mathbb{A}$ is $(u^\omega)_n$ (or w_n). As a consequence, a finite generalized Lyndon word need not be unbordered. Or an infinite generalized Lyndon word x may be a product of prefixes of x. Or if $u, v \in \mathbb{A}^+$ are generalized Lyndon and $u^\omega \prec_\mathbb{N} v^\omega$ it need not be the case that uv is generalized Lyndon. For example, let $\mathbb{A} = \{a, b\}$ be ordered by $a < b$. Consider the total order $\preceq_\mathbb{N}$ on $\mathbb{A}^\mathbb{N}$ defined as previously by the alternating order of Example 2.2.1. Then $u = abba$ is a generalized Lyndon word, as is $v = b$ and $u^\omega \prec_\mathbb{N} v^\omega$, yet uv is not generalized Lyndon.

2.2.1.2 Factorization theorem

In [Reu05], and then with a simpler and more combinatorial proof in [DRR18], the authors showed that generalized Lyndon words preserve the factorization property for finite words, i.e. Theorem 2.1.9:

Theorem 2.2.4. *Let $(<_n)_{n\in\mathbb{N}}$ be sequence of total orders on $\mathbb{A}$ and $\prec_\mathbb{N}$ be the total order they induce. Let u be a nonempty finite word in $\mathbb{A}^+$. Then there exist a unique n and n generalized Lyndon words $w_1, \cdots, w_n$, such that:*

$$w_1 \succeq_\mathbb{N} \cdots \succeq_\mathbb{N} w_n, \textit{ and } w = w_1 w_2 \cdots w_n.$$

Remark 12. This theorem is significantly more complicated to prove than Theorem 2.1.9, since the two properties (2.1) and (2.2) are no longer verified. For (2.1) it is easy to see by taking the alternating order, and for (2.2), for example, with $a <_4 b$ and $b <_5 a$ we have $ab \prec_\mathbb{N} aba$ but $ab.(aab)^\omega \prec_\mathbb{N} aba.a^\omega$.

As in the classical case, it is possible to define infinite generalized Lyndon words:

Definition 2.2.5. An infinite word $x \in \mathbb{A}^\mathbb{N}$ is called generalized Lyndon if $x \prec_\mathbb{N} y$ for each proper suffix y of x.

It is then natural to try to extend Theorem 2.1.16. That question was asked by the authors in [DRR18]:

Question 2.2.6 (Open Problem 2 in [DRR18])**.** *Prove that each infinite word can be factorized in a unique way as a non-increasing product of finite and infinite generalized Lyndon words.*

2.2.1.3 Comparing infinite words instead of finite words

We saw that the relation Dolce, Restivo and Reutenauer defined is not a total order. Alternatively, a family of total orders $(<_n)_{n\geq 1}$ on $\mathbb{A}$ defines an order $\leq_\mathbb{N}$ on $\mathbb{A}^+$ by $u \leq_\mathbb{N} v$ if and only if either u is a prefix of v or if $u = wau'$ and $v = wbv'$ for some $w \in \mathbb{A}^*$, $a, b \in \mathbb{A}$ and $u', v' \in \mathbb{A}^*$ and $a <_{|w|+1} b$. This relation is a total order. It is also possible possible to define a generalization of Lyndon words respectively to this order, as words strictly smaller tha their suffixes.

It might be strange to choose to consider a partial infinite order to compare finite words, and it could seem more simple to take that total finite order to do so. However, the order defined by infinite words seems to be the good definition to consider to preserve the factorization theorem on finite Lyndon words. Take, for example the sequence of alternating orders from 2.2.1. It is then easy to see that aab doesn't admit a generalized Lyndon non-increasing factorization like in 2.1.9 with respect to $<_{\mathbb{N}}$. Indeed, for example aab is not generalized Lyndon since $ab <_{\mathbb{N}} aab$. Then, the first term of a generalized Lyndon factorization would be a; but $a <_{\mathbb{N}} ab$ and $a <_{\mathbb{N}} b$, so the factorization cannot be non-increasing.

Moreover, this use of an infinite order to compare finite words and study Lyndon-type words is not unique in the literature.

In [DRR19], Dolce, Restivo and Reutenauer considered a new version $<_\omega$ of the lexicographical order defined by comparisons on infinite words: for u and v finite words, $u \leq_\omega v$ when $u^\omega \leq v^\omega$. For infinite words, $<_\omega$ is the lexicographical order.

This order differs from the classical lexicographical order:

Example 2.2.7. Suppose $a < b$. Then $ab < aba$ and $aba <_\omega ab$.

However, for finite words u and v of same length, the orders coincide:

$$u < v \Leftrightarrow u <_\omega v.$$

This induces that Lyndon words, defined as smaller than any of their conjugates, are the same respectively to both orders. Dolce, Restivo and Reutenauer proved, amongst other results, that Lyndon words are exactly the words strictly smaller than their proper suffixes relatively to this infinite order $<_\omega$. This justify, in a way, that considering generalized Lyndon with respect to an infinite order can be seen as a proper generalization of Lyndon words.

2.2.2 ω-Lyndon words

The rest of this chapter is mainly a detailed version of a published article written jointly with Luca Zamboni [PZ19b].

2.2.2.1 Definition

We address Question 2.2.6 while considering yet another generalization of Lyndon words. We remarked that their orders preserve an important property of the lexicographical order: as soon as the order between two words of the same size is fixed, one can add any suffix to those, the order will not change. More precisely, the following is true:

$$\forall n \in \mathbb{N} \text{ and } u, v \in \mathbb{A}^n, \text{ if } u \prec_{\mathbb{N}} v \text{ then } ux \prec_{\mathbb{N}} vy \text{ for all } x, y \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^*. \tag{2.3}$$

This condition seems to be worth preserving. For example, it would be hard to search a word in a dictionary not matching this requirement. With this condition, you are sure that all the words starting with a prefix u are going to be grouped in your dictionary.

For the rest of this chapter, let us fix once and for all a total order $\preceq$ on $\mathbb{A}^{\mathbb{N}}$ verifying the lexicographic-like condition (2.3). This is the setting on which we are defining and studying Lyndon-like words in [PZ19b]. This setting, as we said, encompass the orders studied in [DRR18].

However the two settings are not equivalent. In fact, in the context of the generalised lexicographic order in [DRR18], if, for example, $bb \prec_{\mathbb{N}} ba$ then it would mean that $b <_2 a$ and hence $ab \prec_{\mathbb{N}} aa$. This is no longer true in the setting (2.3) as one may have $bb \prec ba$ and $aa \prec ab$. To define an order $\prec$ only satisfying (2.3) is equivalent than defining a sequence $(<_u)_{u \in \mathbb{A}^*}$ of total orders on $\mathbb{A}$ like in the setting of [DRR18] where the order between two letters a and b is given by what appears before them: $uav \prec ubw \Leftrightarrow a <_u b$.

Let us begin by some important remarks:

Remark 13. The following two properties will be useful in the subsequent proofs; the situation here works as in the classical case.

1. If $x, y \in \mathbb{A}^{\mathbb{N}}$ and $x \preceq y$ then for each prefix u of x and v of y with $|u| = |v|$ one has $u^\omega \preceq v^\omega$ with equality if and only if $u = v$.
2. If $u^\omega \preceq v^\omega$, and neither u nor v is a prefix of the other, then $u^\omega \prec v^\omega$. In particular, if $u^\omega \preceq v^\omega$ and u and v are primitive, then either $u = v$ or $u^\omega \prec v^\omega$.

We begin with the following lemma which is analogous to Lemma 13 in [DRR18]. We omit the proof as it is identical to that of Lemma 13 in [DRR18].

Lemma 2.2.8. *For each $u, v \in \mathbb{A}^+$ and $\star \in \{=, \prec, \succ\}$ the following are equivalent:*

1. $u^\omega \star v^\omega$;
2. $(uv)^\omega \star v^\omega$;
3. $u^\omega \star (vu)^\omega$;
4. $(uv)^\omega \star (vu)^\omega$;

We remark that a slightly modified version of the above lemma also applies in case one of u and v is infinite and the other finite: For example if $u \in \mathbb{A}^+$ and $v \in \mathbb{A}^{\mathbb{N}}$ then $u^\omega \star v$ if and only if $uv \star v$.

We now introduce the definition of ω-Lyndon words. We first give a definition for infinite words. It will be useful to extend this notion also to finite words, however given that the order is defined only on infinite words, we shall be required to pass to infinite words by associating to each finite word w its (periodic) infinite counterpart w^ω. Following [DRR18]:

Definition 2.2.9. An infinite word $x \in \mathbb{A}^{\mathbb{N}}$ is called ω-Lyndon if $x \prec y$ for each proper suffix y of x. A word $w \in \mathbb{A}^+$ is called ω-Lyndon if $w^\omega \prec v^\omega$ for each proper suffix v of w. We let $\mathscr{L}_\omega$ denote the set of all ω-Lyndon words (finite and infinite) relative to $\preceq$.

Remark 14. We note that $\mathbb{A} \subseteq \mathscr{L}_\omega$. If $w \in \mathbb{A}^+$ is ω-Lyndon, then w is primitive and similarly if $x \in \mathbb{A}^{\mathbb{N}}$ is ω-Lyndon, then x is not periodic. It follows from Lemma 2.2.8 that $w \in \mathbb{A}^+$ is ω-Lyndon if and only if for all factorizations $w = uv$ with $u, v \in \mathbb{A}^+$ we have $u^\omega \prec v^\omega$ (see Theorem 14 in [DRR18]). This in turn implies that if $w \in \mathscr{L}_\omega$, then for each prefix u of w and each factor v of w with $|u| = |v|$, either $u = v$ or $u^\omega \prec v^\omega$. In fact, suppose $u \neq v$ and let z be a suffix of w beginning in v. Then if $w \in \mathbb{A}^+$ we have that $w^\omega \prec z^\omega$ and hence $u^\omega \prec v^\omega$. If $w \in \mathbb{A}^{\mathbb{N}}$, then $w \prec z$ and hence $u^\omega \prec v^\omega$.

As for Lemma 2.2.8, we can extend the factorization theorem for finite words to our setting. Again, we omit the proof as it is identical to that of Theorem 16 in [DRR18].

Proposition 2.2.10. *Each $w \in \mathbb{A}^+$ admits a unique factorization $w = l_1 l_2 \cdots l_k$ with $l_i \in \mathscr{L}_\omega$ and $l_1^\omega \succeq l_2^\omega \succeq \cdots \succeq l_k^\omega$.*

Now we come to the factorization of infinite words.

2.2.2.2 Factorization of infinite words

Definition 2.2.11. For $x \in \mathbb{A}^{\mathbb{N}}$ we say x admits an infinite ω-Lyndon factorization if $x = \prod_{i=1}^{\infty} l_i$ with each $l_i \in \mathscr{L}_\omega \cap \mathbb{A}^+$ and $l_1^\omega \succeq l_2^\omega \succeq l_3^\omega \succeq \cdots$. We say x admits a finite ω-Lyndon factorization if $x = l_1 l_2 \dots l_k$ with $l_i \in \mathscr{L}_\omega \cap \mathbb{A}^+$, $l_k \in \mathscr{L}_\omega \cap \mathbb{A}^{\mathbb{N}}$ and $l_1^\omega \succeq l_2^\omega \succeq \cdots \succeq l_{k-1}^\omega \succ l_k$.

Remark 15. Because of the fact that if u and v are Lyndon in the usual sense and $u < v$ then uv is Lyndon, it follows that the factorization of a finite word w as a non-increasing product of Lyndon words is also the shortest factorization of w as a product of Lyndon words. This is no longer true in general for ω-Lyndon words. For example, relative to the total order $\preceq$ defined in Remark 11, the word $w = ababab$ is the product of $ababa$ and b, both of which are ω-Lyndon, yet the ω-Lyndon factorization of w has length three and is given by $w = (ab)(ab)(ab)$.

The following lemma constitutes a generalization of a characterization of infinite Lyndon words given in Proposition 2.1.14:

Lemma 2.2.12. *Let $x \in \mathbb{A}^{\mathbb{N}}$. Then $x \notin \mathscr{L}_\omega$ if and only if either $x = l^\omega$ for some $l \in \mathscr{L}_\omega$ or only a finite number of prefixes of x are members of $\mathscr{L}_\omega$.*

Proof. Assume $x \notin \mathscr{L}_\omega$ and pick a proper suffix y of x with $y \preceq x$.

If $y = x$, then $x = u^\omega$ for some primitive word $u \in \mathbb{A}^+$. As u is primitive, for each nontrivial factorization $u = u_1 u_2$ one has $u_2 u_1 \neq u$ and hence $(u_2 u_1)^\omega \neq u^\omega$. If $u \in \mathscr{L}_\omega$ we are done. If $u \notin \mathscr{L}_\omega$, pick a factorization $u = u_1 u_2$ with $(u_2 u_1)^\omega \prec u^\omega$. Then by Remark 14, if v is a prefix of x with $|v| \geq 2|u|$ then $v \notin \mathscr{L}_\omega$. Hence x has less than $2|u|$ ω-Lyndon prefixes, and we are done.

If $y \prec x$, pick a prefix v of y and a prefix u of x with $|u| = |v|$ and $v^\omega \prec u^\omega$. Then again by Remark 14 any prefix of x containing v as a factor cannot belong to $\mathscr{L}_\omega$.

For the converse, we note that if x is periodic then $x \notin \mathscr{L}_\omega$. So assume x is not periodic and only a finite number of prefixes of x belong to $\mathscr{L}_\omega$. For $n \in \mathbb{N}$, let $x[n]$ denote the prefix of x of length n, and let $l(n)$ denote the length of a longest ω-Lyndon word occurring in the ω-Lyndon factorization of $x[n]$ (see Proposition 2.2.10). If $(l(n))_{n\geq 1}$ is unbounded, then pick n such that

1. $l(n)$ is greater than the length of the longest ω-Lyndon prefix of x
2. $l(n)$ is the length of the last ω-Lyndon word in the ω-Lyndon factorization of $x[n]$.

This is always possible, as the factorization of a finite word is unique, hence if $l(n)$ is not the length of the last ω-Lyndon word in the ω-Lyndon factorization of $x[n]$, it means a smaller integer n' (corresponding to the prefix of x product up to the factor of length $l(n)$ of the factorization of $x[n]$) already satisfies the conditions. Then $x[n] = l_1 l_2 \cdots l_k$ with $l_i \in \mathscr{L}_\omega$ and $l_1^\omega \succeq l_2^\omega \succeq \cdots \succeq l_k^\omega$, and l_k is not a prefix of x. By iteration of Lemma 2.2.8, $(l_1 l_2 \cdots l_k)^\omega \succeq l_k^\omega$ and hence $(l_1 l_2 \cdots l_k)^\omega \succ l_k^\omega$. Writing $x = l_1 l_2 \cdots l_{k-1} y$ with $y \in \mathbb{A}^{\mathbb{N}}$, since l_k is a prefix of y but not of $l_1 l_2 \cdots l_k$ we have $x \succ y$ and hence $x \notin \mathscr{L}_\omega$. If $(l(n))_{n\geq 1}$ is bounded then pick a finite set $F \subseteq \mathscr{L}_\omega$ such that all ω-Lyndon words occurring in the ω-Lyndon factorization of $x[n]$ for $n \in \mathbb{N}$ belong to F. Because of the non increasing order condition in the ω-Lyndon factorization, there exist $l_1, l_2, \dots, l_k$ in F with $l_1^\omega \succeq l_2^\omega \succeq \cdots \succeq l_k^\omega$ such that $l_1 l_2 \cdots l_{k-1} l_k^m$ is a prefix of x for every $m \in \mathbb{N}$. Pick m such that l_k^m is not a prefix of x and write $x = l_1 l_2 \cdots l_{k-1} y$ with $y \in \mathbb{A}^{\mathbb{N}}$. Then as l_k^m is a prefix of y but not of $l_1 l_2 \cdots l_k^m$ and $(l_1 l_2 \cdots l_k^m)^\omega \succeq l_k^\omega$ we deduce that $x \succ y$ and hence $x \notin \mathscr{L}_\omega$. □

Now we can prove that every infinite word admits a Lyndon factorization:

Proposition 2.2.13. *Each $x \in \mathbb{A}^{\mathbb{N}}$ admits either an infinite or a finite ω-Lyndon factorization.*

Proof. Let $x \in \mathbb{A}^{\mathbb{N}}$ and assume x does not admit an infinite ω-Lyndon factorization. We will show that x admits a finite ω-Lyndon factorization. The result is immediate in case $x \in \mathscr{L}_\omega$ so we may assume that $x \notin \mathscr{L}_\omega$. For $n \in \mathbb{N}$, let $l_i^{(n)}$ be the i'th ω-Lyndon word occurring in the ω-Lyndon factorization of $x[n]$, where $x[n]$ is the prefix of length n of x. In other words $x[n] = l_1^{(n)} l_2^{(n)} \cdots l_k^{(n)}$. We may take $l_i^{(n)}$ to be the empty word if the ω-Lyndon factorization of $x[n]$ has fewer than i terms.

By Lemma 2.2.12, the set

$$L_1 = \{l_1^{(n)} \,:\, n \in \mathbb{N}\}$$

is finite and hence there exist $l_1 \in \mathscr{L}_\omega$ and an infinite set $A_1 \subseteq \mathbb{N}$ such that for each $n \in A_1$ the ω-Lyndon factorization of $x[n]$ begins in l_1.

Put

$$L_2 = \{l_2^{(n)} \,:\, n \in A_1\}.$$

If L_2 is finite, then we may pick $l_2 \in \mathscr{L}_\omega$ and an infinite subset $A_2 \subseteq A_1$ such that for each $n \in A_2$ the ω-Lyndon factorization of $x[n]$ begins in $l_1 l_2$ and put $L_3 = \{l_3^{(n)} \,:\, n \in A_2\}$. Continuing as above, if each L_k is finite then x would admit an infinite ω-Lyndon factorization contrary to our assumption. And hence, there exists $k \geq 2$ and $l_1, l_2, \ldots, l_{k-1} \in \mathscr{L}_\omega$ and an infinite set $A_{k-1} \subseteq \mathbb{N}$ such that for each $n \in A_{k-1}$ the ω-Lyndon factorization of $x[n]$ begins in $l_1 l_2 \cdots l_{k-1}$ and $L_k = \{l_k^{(n)} \,:\, n \in A_{k-1}\}$ infinite.

Define $l_k \in \mathbb{A}^{\mathbb{N}}$ by $x = l_1 l_2 \ldots l_{k-1} l_k$. We claim $l_k \in \mathscr{L}_\omega$ and $l_{k-1}^\omega \succ l_k$.

Observe that $l_k \neq l_{k-1}^\omega$ for otherwise $x = l_1 \cdots l_{k-2} l_{k-1}^\omega$ is an infinite ω-Lyndon factorization of x. Pick $m \in \mathbb{N}$ such that l_{k-1}^m is not a prefix of l_k and $n \in A_{k-1}$ such that $|l_{k-1}^m| < |l_k^{(n)}|$. Since $l_{k-1}^\omega \succeq (l_k^{(n)})^\omega$ and l_{k-1}^m is not a prefix of $l_k^{(n)}$, it follows that $l_{k-1}^\omega \succ l_k$. It remains to show that $l_k \in \mathscr{L}_\omega$. By Lemma 2.2.12, if $l_k \notin \mathscr{L}_\omega$ then $l_k = u^\omega$ for some $u \in \mathscr{L}_\omega$ and hence $x = l_1 l_2 \cdots l_{k-1} u^\omega$ is an infinite ω-Lyndon factorization, a contradiction. □

We now turn to the question of uniqueness of ω-Lyndon factorizations for infinite words. We begin by establishing uniqueness for words admitting a finite ω-Lyndon factorization.

Lemma 2.2.14. *Let $x \in \mathbb{A}^{\mathbb{N}}$ and $u_1 u_2 \cdots u_k$ be a prefix of x such that $u_1^\omega \succeq u_2^\omega \succeq \cdots \succeq u_k^\omega$. If $x \in \mathscr{L}_\omega$ then each u_i is a prefix of x.*

Proof. By iteration of Lemma 2.2.8, for $1 \leq i \leq k$ we have that $(u_1 \cdots u_i)^\omega \succeq u_i^\omega$. Let v_i denote the prefix of x of length $|u_i|$. If $v_i \neq u_i$ then $v_i^\omega \succ u_i^\omega$ contradicting that $x \in \mathscr{L}_\omega$. □

Lemma 2.2.15. *Let $x \in \mathbb{A}^{\mathbb{N}}$ and $k \geq 3$. Assume $u_1 u_2 \cdots u_k$ is a prefix of x such that $u_1^\omega \succeq u_2^\omega \succeq \cdots \succeq u_k^\omega$. If $x \in \mathscr{L}_\omega$, then either $|u_1 \cdots u_{k-2}| \leq |u_k|$ or $u_1 \cdots u_{k-2} u_k$ is a prefix of x.*

Proof. Assume $|u_1 \cdots u_{k-2}| > |u_k|$. By Lemma 2.2.14, we have that u_k is a prefix of x and hence a prefix of $u_1 \cdots u_{k-2}$. By Lemma 2.2.8 we have that $(u_1 \cdots u_{k-2})^\omega \succeq u_{k-2}^\omega \succeq u_{k-1}^\omega$ and hence $(u_1 \cdots u_{k-2} u_{k-1})^\omega \succeq (u_{k-1} u_1 \cdots u_{k-2})^\omega = (u_{k-1} u_k v)^\omega$. Since $x \in \mathscr{L}_\omega$ it follows that $u_{k-1} u_k$ is a prefix of x and hence u_k is a prefix of $u_{k-1} u_k$. Thus $u_1 \cdots u_{k-2} u_k$ is a prefix of x. □

Lemma 2.2.16. *Let $(u_i)_{i\in\mathbb{N}}$ be a sequence in $\mathbb{A}^+$ with $u_1^\omega \succ u_2^\omega \succ \cdots$. Then $\prod_{i=1}^{\infty} u_i \notin \mathscr{L}_\omega$.*

Proof. Put $x = u_1u_2\cdots$ and suppose to the contrary that $x \in \mathscr{L}_\omega$. We will show that $|u_k| < |u_1|$ for each $k \geq 3$ and hence the sequence $(u_i)_{i\in\mathbb{N}}$ is ultimately constant, a contradiction. To see that $|u_k| < |u_1|$ for each $k \geq 3$, suppose to the contrary that $|u_k| \geq |u_1|$ for some $k \geq 3$. By iteration of Lemma 2.2.15, there exists $2 \leq j \leq k-1$ such that $u_1\cdots u_ju_k$ is a prefix of x and $|u_1\cdots u_{j-1}| \leq |u_k|$. By Lemma 2.2.14 we have that u_k is a prefix of x and hence $u_1\cdots u_{j-1}$ is a prefix of u_k. As $(u_1\cdots u_{j-1})^\omega \succ u_j^\omega$ we have that $(u_1\cdots u_{j-1}u_j)^\omega \succ (u_ju_1\cdots u_{j-1})^\omega$ and hence $u_1\cdots u_{j-1}u_j \neq u_ju_1\cdots u_{j-1}$. Since $u_1\cdots u_{j-1}$ is a prefix of u_k it follows that the suffix of x beginning in u_ju_k is smaller than x contradicting that $x \in \mathscr{L}_\omega$. □

Lemma 2.2.17. *Let $x \in \mathbb{A}^\mathbb{N}$. If $x = v_1v_2v_3\cdots$ with $v_1^\omega \succeq v_2^\omega \succeq \cdots$ then $x \notin \mathscr{L}_\omega$.*

Proof. Assume to the contrary that $x \in \mathscr{L}_\omega$. Without loss of generality we may assume that each v_i is primitive. We claim $(v_i)_{i\geq 1}$ is ultimately periodic. In fact, if the sequence $(v_i)_{i\geq 1}$ is not ultimately periodic, then by concatenating together the consecutive terms of the sequence which are equal, we may write $x = u_1u_2\cdots$ with $u_1^\omega \succ u_2^\omega \succ \cdots$ in contradiction with Lemma 2.2.16. As $x \in \mathscr{L}_\omega$ and hence not periodic, write $x = v_1\cdots v_kv_{k+1}^\omega$ with $v_1^\omega \succeq \cdots \succeq v_k^\omega \succ v_{k+1}^\omega$ and pick m such that v_{k+1}^m is not a prefix of x. As $(v_1\cdots v_kv_{k+1}^m)^\omega \succ v_{k+1}^\omega$, it follows that the suffix $v_{k+1}^\omega \prec x$ contradicting that $x \in \mathscr{L}_\omega$. □

Lemma 2.2.18. *If x admits an infinite ω-Lyndon factorization, then no suffix of x belongs to $\mathscr{L}_\omega$. In particular x does not admit a finite ω-Lyndon factorization.*

Proof. Suppose x admits an infinite ω-Lyndon factorization $x = l_1l_2l_3\cdots$. Then any suffix y of x may be written as $y = s_il_{i+1}l_{i+2}\cdots$ with $i \geq 1$ and s_i a suffix of l_i. Since $s_i^\omega \succeq l_i^\omega$ it follows that $s_i^\omega \succeq l_{i+1}^\omega \succeq \cdots$. By Lemma 2.2.17, it follows that $y \notin \mathscr{L}_\omega$. □

We can then prove the uniqueness of the ω-Lyndon factorization in the case of finite factorization:

Corollary 2.2.19. *If an infinite word $x \in \mathbb{A}^\mathbb{N}$ admits a finite ω-Lyndon factorization $x = l_1l_2\cdots l_k$, then it is the unique ω-Lyndon factorization of x.*

Proof. It follows from Lemma 2.2.18 that x does not admit an infinite ω-Lyndon factorization. It remains to show that x admits no other finite ω-Lyndon factorization. For this purpose, write $x = ul_k$ with $u \in \mathbb{A}^*$ and $l_k \in \mathscr{L}_\omega$ and observe that if $v \in \mathbb{A}^+$ is any suffix of u, then (by iteration of Lemma 2.2.8) $l_k \preceq vl_k$. In other words, $vl_k \notin \mathscr{L}_\omega$ and hence l_k is necessarily the first ω-Lyndon suffix of x. Uniqueness now follows from Proposition 2.2.10. □

We next prove uniqueness of ω-Lyndon factorizations for those infinite words x not admitting a finite ω-Lyndon factorization. We first consider the case that x is ultimately periodic:

Lemma 2.2.20. *Assume $x \in \mathbb{A}^\mathbb{N}$ is ultimately periodic. Then x admits a unique ω-Lyndon factorization.*

Proof. By Corollary 2.2.19 we may suppose that x does not admit a finite ω-Lyndon factorization. Using Proposition 2.2.13 let $x = l_1l_2\cdots$ be an infinite ω-Lyndon factorization of x. We claim the sequence $(l_i)_{i\in\mathbb{N}}$ is ultimately constant.

As the sequence $(l_i)_{i\in\mathbb{N}}$ is decreasing, it suffices to show that $\liminf_{i\to\infty}|l_i| < +\infty$. So pick a suffix x' of x and an infinite set $I \subseteq \mathbb{N}$ such that l_i is a prefix of x' for each $i \in I$. Since infinitely many l_i start in the periodic part of x, it is possible to find such x'. Applying lemma 2.2.18, we know that $x' \notin \mathscr{L}_\omega$. Using lemma 2.2.12 it follows that either $x' = l^\omega$ for some $l \in \mathscr{L}_\omega$ or x' has only finitely many ω-Lyndon prefixes. In the latter case $\{|l_i| : i \in I\}$ is clearly bounded. In the former case pick $j < k$ in I such that $x' = \prod_{i\geq j} l_i$ and $\min\{|l_j|, |l_k|\} \geq 2|l|$. Then $w = l_j \cdots l_{k-1} = l^r$ for some $r \in \mathbb{N}$. Indeed, the contrary would imply that l^2 is an internal factor of l^3; but ω-Lyndon words are primitive. That contradicts Proposition 2.2.10.

Having proved that any infinite ω-Lyndon factorization of x is ultimately constant, uniqueness of the factorization now follows. In fact, suppose $x = l'_1 l'_2 \cdots$ is another ω-Lyndon factorization with $l'_i = l'$ for all i greater than some k' and $l' \in \mathscr{L}_\omega$. Then since l and l' are each primitive, it follows that $|l| = |l'|$ whence $l = l'$ and the two factorizations must ultimately synchronise, i.e., $l_i = l'_i$ for all sufficiently large i. The rest now follows from Proposition 2.2.10. □

Definition 2.2.21. A factor $u \in \mathbb{A}^+$ of an infinite word x is said to be minimal in x if $u^\omega \preceq v^\omega$ for all factors v of x with $|v| = |u|$.

We note that, if u is a minimal factor of x, then so is every prefix of u. The following lemma will be applied to show that any infinite aperiodic word x admits at most one infinite ω-Lyndon factorization, and how to construct it.

Lemma 2.2.22. *Assume $x \in \mathbb{A}^\mathbb{N}$ and $u \in \mathbb{A}^+$ is a minimal factor of x. Let $w \in \mathbb{A}^*$ be the longest prefix of x preceding the first occurrence of u in x. Assume x admits an infinite ω-Lyndon factorization $x = l_1 l_2 l_3 \cdots$ with $\limsup_{i\to\infty}|l_i| = +\infty$. Then either $w = \varepsilon$ or $w = l_1 \cdots l_k$ for some $k \in \mathbb{N}$.*

Proof. Put $n = |u|$ and write $u = u_1 u_2 \cdots u_n$. Also write $x = wux'$ with $x' \in \mathbb{A}^\mathbb{N}$; by assumption wu contains exactly one occurrence of u. Assume $w \neq \varepsilon$ and let k be the least positive integer such that $\sum_{i=1}^k |l_i| \geq |w|$. We must show that $\sum_{i=1}^k |l_i| = |w|$. Suppose to the contrary that $\sum_{i=1}^k |l_i| > |w|$. We first note that u cannot be fully contained inside l_k for otherwise, if v denotes the prefix of l_k with $|v| = |u|$, then as $v \neq u$ and $l_k \in \mathscr{L}_\omega$ we have $v^\omega \prec u^\omega$ which contradicts that u is minimal. Thus we may write $l_k = zu_1 \cdots u_p$ for some $z \neq \varepsilon$ and $p < n$. Let $r = \min\{|l_i| : i \geq k+1\}$ and pick $j \geq k+1$ with $|l_j| = r$. Also pick $j' > j$ such that $|l_{j'}| \geq p + r$.

Case 1: $n \geq p + r$

By definition of r it follows that $u_{p+1}...u_{p+r}$ is a prefix of l_{k+1}. We first claim that

$$u_{p+1}...u_{p+r} = u_1 \cdots u_r = (u_1 \cdots u_p)^{\frac{r}{p}} \tag{2.4}$$

where $(u_1 \cdots u_p)^{\frac{r}{p}}$ denotes the prefix of length r of $(u_1 \cdots u_p)^\omega$. In fact, as u is a minimal factor of x we have that $(u_1 \cdots u_r)^\omega \preceq (u_{p+1} \cdots u_{p+r})^\omega$. Furthermore since $l_{k+1}^\omega \preceq l_k^\omega \prec (u_1 \cdots u_p)^\omega$ and $u_{p+1} \cdots u_{p+r}$ is a prefix of l_{k+1} it follows that $(u_{p+1} \cdots u_{p+r})^\omega \preceq ((u_1 \cdots u_p)^{\frac{r}{p}})^\omega$. Combining we get $(u_1 \cdots u_r)^\omega \preceq (u_{p+1} \cdots u_{p+r})^\omega \preceq ((u_1 \cdots u_p)^{\frac{r}{p}})^\omega$ from which (2.4) follows.

We also claim that

$$l_j = u_1 \cdots u_r. \tag{2.5}$$

Indeed, since u is a minimal factor of x we have $(u_1 \cdots u_r)^\omega \preceq l_j^\omega$. On the other hand $l_j^\omega \preceq l_{k+1}^\omega$ and so by taking the prefix of length r of both words we obtain $l_j^\omega \preceq (u_{p+1} \cdots u_{p+r})^\omega$. So combining and using (2.4) we deduce that $(u_1 \cdots u_r)^\omega \preceq l_j^\omega \preceq (u_1 \cdots u_r)^\omega$ from which (2.5) follows.

Thus we have

$$(u_1 \cdots u_r)^\omega = l_{j'}^\omega \preceq l_k^\omega \prec (u_1 \cdots u_p)^\omega$$

and hence by Lemma 2.2.8

$$(u_1 \cdots u_r u_1 \cdots u_p)^\omega \prec (u_1 \cdots u_p u_1 \cdots u_r)^\omega. \tag{2.6}$$

Using the fact $u_1 \cdots u_p u_{p+1} \cdots u_{p+r}$ is a minimal factor of x together with (2.4) and (2.5) gives

$$(u_1 \cdots u_p u_1 \cdots u_r)^\omega = (u_1 \cdots u_{p+r})^\omega \preceq (l_{j'}[p+r])^\omega \preceq (l_j^\omega[p+r])^\omega = ((u_1 \cdots u_r)^{\frac{p+r}{r}})^\omega.$$

Together with (2.6) gives

$$(u_1 \cdots u_r u_1 \cdots u_p)^\omega \prec (u_1 \cdots u_p u_1 \cdots u_r)^\omega \preceq ((u_1 \cdots u_r)^{\frac{p+r}{r}})^\omega. \tag{2.7}$$

It follows from (2.7) that $u_1 \cdots u_p = (u_1 \cdots u_r)^{\frac{p}{r}}$ and hence

$$u_1 \cdots u_r u_1 \cdots u_p = (u_1 \cdots u_r)^{\frac{p+r}{r}}$$

which by (2.7) gives $((u_1 \cdots u_r)^{\frac{p+r}{r}})^\omega \prec ((u_1 \cdots u_r)^{\frac{p+r}{r}})^\omega$, a contradiction.

Case 2: $n < p + r$**.**

In this case $u_{p+1} \cdots u_n$ is a prefix of l_{k+1} and the same arguments used to prove (2.4) shows that

$$u_1 \cdots u_n = (u_1 \cdots u_p)^{\frac{n}{p}}. \tag{2.8}$$

As $l_k^\omega = (zu_1 \cdots u_p)^\omega \prec (u_1 \cdots u_p)^\omega$, it follows that $|l_k| = |z| + p < n$ for otherwise $u = u_1 \cdots u_n$ would be a prefix of l_k which would imply an earlier occurrence of u in x. Thus

$$zu_1 \cdots u_p = (u_1 \cdots u_p)^{\frac{|z|+p}{p}} = (u_1 \cdots u_p)^a u_1 \cdots u_q \tag{2.9}$$

for some choice of integers a, q and as l_k is primitive we have that $1 \leq q < p$.

Finally, we have $z = (u_1 \cdots u_p)^{a-1} u_1 \cdots u_q$ and hence

$$u_1 \cdots u_p u_1 \cdots u_q = u_1 \cdots u_q u_1 \cdots u_p \tag{2.10}$$

from which it follows that l_k is not primitive, a contradiction. □

Proposition 2.2.23. *Let $x \in \mathbb{A}^\mathbb{N}$ be an aperiodic infinite word and $x = l_1 l_2 l_3 \cdots = l'_1 l'_2 l'_3 \cdots$ two infinite ω-Lyndon factorizations of x. Then $l_i = l'_i$ for each $i \in \mathbb{N}$.*

Proof. Suppose to the contrary that $l_i \neq l'_i$ for some $i \in \mathbb{N}$. Short of replacing x by some suffix of x, we may assume that $l_1 \neq l'_1$. By Lemma 2.2.18 it follows that $x \notin \mathscr{L}_\omega$ and hence x contains a minimal factor u which is not a prefix of x. Let $w \in \mathbb{A}^+$ denote the prefix of x which precedes the first occurrence of u in x. As x is aperiodic it follows that $\limsup_{i\to\infty} |l_i| = \limsup_{i\to\infty} |l'_i| = +\infty$. By Lemma 2.2.22 it follows that there exist $k, k' \in \mathbb{N}$ such that $w = l_1 \ldots l_k = l'_1 \cdots l'_{k'}$ contradicting Proposition 2.2.10. □

2.2.2.3 Conclusion

Putting together the finite and infinite cases, we get the following theorem, which entirely answers to the question asked in [DRR18]:

Theorem 2.2.24. *Each infinite word $x \in \mathbb{A}^{\mathbb{N}}$ admits precisely one ω-Lyndon factorization.*

Proof. Existence follows from Proposition 2.2.13. For uniqueness, if x admits a finite ω-Lyndon factorization, then uniqueness follows from Corollary 2.2.19. So we may suppose that x admits only infinite ω- Lyndon factorizations. If x is ultimately periodic uniqueness follows from Lemma 2.2.20 while if x is aperiodic uniqueness follows from Proposition 2.2.23. □

As was already stressed, our setting was a generalization of the setting in [DRR18]. Interestingly, in [BW20] Amanda Burcroff and Eric Winsor answered the question in [DRR18] as well, using a different generalization and other tools. In fact, they kept the notion of order defined by a sequence of total orders $<_n$ on $\mathbb{A}$, but unlike in the original setting or in our setting, the alphabet $\mathbb{A}$ is not supposed to be finite. Since we achieved our results independently and approximately at the same time we decided to publish in the same issue of *Theoretical Computer Science*.

Chapter 3

Open and closed complexity

The results of this chapter are the subject of an article written jointly with Olga Parshina, submitted for publication to the Bulletin of the London Mathematical Society [PP20].

Contents

3.1 Introduction

3.1.1 Morse and Hedlund theorem

A fundamental problem in many areas of mathematics is to describe local constraints that imply global regularities. An example of this local to global phenomena is found in the study of periodicity in the framework of symbolic dynamics. The factor complexity function p_x, first introduced by G.A. Hedlund and M. Morse in their 1938 seminal paper on symbolic dynamics [MH38], counts the number of distinct blocks (or *factors*) of each length occurring in an infinite word $x = x_1x_2x_3\cdots$ over a finite set $\mathbb{A}$.

Example 3.1.1. $p_{(ab)^\omega}(1) = 2$ and the corresponding factors are a and b; $p_{(ab)^\omega}(2) = 2$ and the corresponding factors are ab and ba, while $p_{a(ab)^\omega}(2) = 3$ and the corresponding factors are $aa,\ ab$ and ba

They proved that each aperiodic infinite word contains at least $n+1$ distinct factors of each length n, and hence in particular the sequence $(p_x(n))_{n\in\mathbb{N}}$ is unbounded:

Theorem 3.1.2 (Morse-Hedlund [MH38]). *An infinite word w is aperiodic if and only if its complexity function is unbounded. If w is aperiodic, we have, for any integer n, $p_w(n) \geq n+1$.*

Here I will recall a proof of this theorem, so that we can see in the latter setting what are the differences in the proofs.

Proof. We first need to prove the following essential lemma:

Lemma 3.1.3. *For an aperiodic word w, the complexity function $p_w(n)$ is strictly increasing.*

Proof. Let w be an infinite word. Every block of length n of w admits at least one right extension since w is infinite, and two blocks of length $n+1$ differing on the n first letters are different. This ensure that the complexity function is non-decreasing. If we suppose that there exists an n such that $p_w(n) = p_w(n+1)$, w is ultimately periodic. Indeed, $p_w(n) = p_w(n+1)$ implies that every factor $u_1 \cdots u_n$ of length n of w admits a unique right extension $u_1 \cdots u_n u_{n+1}$, meaning that each time $u_1 \cdots u_n$ appears in w, we know for sure that the next letter is u_{n+1}. But then, this right extension admits a unique right extension $u_1 \cdots u_n u_{n+1} u_{n+2}$ as well, since it ends with a factor of length n. Hence, repeating this process, there is at each step a unique way to extend it to the right. After a certain number of extension, as there is a finite number of factors of length n, the suffix $u_p \cdots u_{p+n}$ of our word has another occurrence $u_k \cdots u_{k+n}$ previously in our factor. But then, by uniqueness of the right extensions, we have $u_{k+j} = u_{p+j}$ for any $j \in \mathbb{N}$. Hence, w is ultimately periodic with period $p-k$. □

Now for an ultimately periodic word w it is really easy to see that the complexity function is bounded: writing $w = uv^\omega$ the number of blocks of length n is smaller than $|u| + |v|$, since there is at most $|v|$ distinct blocks of length n starting in v^ω. □

Morse and Hedlund further showed that an infinite word $x \in \mathbb{A}^\mathbb{N}$ has exactly $n+1$ distinct factors of each length n if and only if x is binary, aperiodic and balanced, i.e., x is a Sturmian word (see [MH40]). Sturmian words are aperiodic words of lowest factor complexity and they arise naturally in different areas of mathematics including combinatorics, algebra, number theory, ergodic theory, dynamical systems and differential equations.

There are numerous variations and extensions of the Morse-Hedlund theorem associated with other complexity functions defined on infinite words $x \in \mathbb{A}^\mathbb{N}$ including *Abelian complexity* [CH73, RSZ11], which counts the number of distinct Abelian classes of words of each length occurring in x, or *palindrome complexity* [ABCD03] counting the number of distinct palindromes of each length occurring in x, or *cyclic complexity* [CFSZ17] counting the number of conjugacy classes of factors of each length in x. As in the case of the Morse and Hedlund theorem, in most cases these different complexity functions may be used to characterise aperiodicity in words.

In this chapter we investigate two new and complementary complexity functions defined on infinite words, and their relation to aperiodicity.

3.1.2 Open and closed words

Definition 3.1.4. Given $u, w \in \mathbb{A}^+$ with $|u| < |w|$, we say u is a *border* of w if u is both a prefix and a suffix of w. We say $w \in \mathbb{A}^+$ is *closed* if either $w \in \mathbb{A}$ or w admits a border u which occurs precisely twice in w. Otherwise w is said to be *open*. Thus $w \in \mathbb{A}^+$ is closed if either $w \in \mathbb{A}$ or if its longest border u occurs exactly twice in w, i.e., u has no internal occurrences in w. The longest border of a closed word is called its *frontier*.

This terminology was first introduced by G. Fici in [Fic11].

Example 3.1.5. The word $w = abaaaab$ is closed (with frontier ab) while $aabab$ and $aabaaa$ are both open.

We note that every closed word $w \in \mathbb{A}^+$ either belongs to $\mathbb{A}$ or may be written in the form $w = uv = vu'$ for some choice of $u, u', v \in \mathbb{A}^+$, and moreover w has no other occurrences of v other than the two witnessed by the above factorizations. Thus in the language of symbolic dynamics, a closed factor $w \in \mathbb{A}^+ \setminus \mathbb{A}$ of an infinite word $x \in \mathbb{A}^{\mathbb{N}}$ is called a *complete first return to v in x* and the factor u is called a *return word* or *first return* to v in x.

Return words constitute a powerful tool in the study of symbolic dynamical systems. For example, they play an important role in the theory of substitution dynamical systems. Return words were used by F. Durand [Dur98] and independently by C. Holton and L.Q. Zamboni in [HZ98] to define so-called *derived words* and *derived substitutions* both of which may be used to characterise infinite words generated by primitive substitutions. An analogous characterisation was later discovered by N. Priebe [Pri00] in the framework of bi-dimensional tilings using the notion of derived tilings involving Voronoï cells. In [DHS99], Durand et al. derived a simple algorithm using return words for computing the dimension group of minimal Cantor systems arising from primitive substitutions. A slightly different notion of return words was used by S. Ferenczi, C. Mauduit and A. Nogueira [FMN96] to compute the eigenvalues of the dynamical system associated with a primitive substitution. Return words were an essential tool used by the authors in [HVZ16] to give a partial answer to a question posed by A. Hof, O. Knill and B. Simon in [HKS95] on a sufficient combinatorial criterion on the subshift Ω of the potential of a discrete Schrödinger operator which guarantees purely singular continuous spectrum on a generic subset of Ω.

There are many other examples of the use of return words in the study of more general symbolic dynamical systems. In [Vui01], L. Vuillon showed that an infinite binary word x is Sturmian if and only if each factor of x admits exactly two first returns in x. We observe that a recurrent word $x \in \mathbb{A}^{\mathbb{N}}$ containing a factor v having only one first return in x is necessarily ultimately periodic, i.e., $x = u'u^\omega$ where u is the unique first return to v in x. Words having exactly k first returns to each factor for $k \geq 3$ have also been extensively studied (see, for example, [BPS06]) and include the symbolic coding of orbits under a k-interval exchange transformation [KS67] as well as Arnoux-Rauzy words [AR91] on a k-letter alphabet. Finally, there has been much recent interest in open and closed words in the framework of combinatorics on words and we refer the interested reader to the nice survey article [Fic17] by G. Fici.

We now define two new complexity functions, the *closed complexity* and the *open complexity*, which will be the main objects studied in this chapter.

3.1.3 Open and closed complexity functions

Definition 3.1.6. For an infinite word $x \in \mathbb{A}^{\mathbb{N}}$ we define the *closed complexity function* Cl_x which associates to each $n \in \mathbb{A}^{\mathbb{N}}$, the number of closed factors of x of length n. Similarly, we define the *open complexity function* Op_x which associates to each $n \in \mathbb{A}^{\mathbb{N}}$, the number of open factors of x of length n.

Example 3.1.7. $\mathrm{Cl}_{(abba)^\omega}(2) = 2$ since aa and bb are both closed while $\mathrm{Cl}_{(abba)^\omega}(3) = 0$. $\mathrm{Op}_{(abba)^\omega}(2) = 2$ since ab and ba are open; $\mathrm{Op}_{(abba)^\omega}(3) = 4$.

Given an infinite word $x \in \mathbb{A}^{\mathbb{N}}$, we are interested in the asymptotic behaviour of the complexity functions Cl_x and Op_x and their relationship to periodicity. As every finite word

$w \in \mathbb{A}^+$ is either open or closed, one has that $p_x(n) = \mathrm{Op}_x(n) + \mathrm{Cl}_x(n)$ for each $n \in \mathbb{N}$. Thus if x is aperiodic, then it follows by the Morse and Hedlund theorem that at least one of the two sequences $(\mathrm{Op}_x(n))_{n\in\mathbb{N}}$, $(\mathrm{Cl}_x(n))_{n\in\mathbb{N}}$ is unbounded. For instance, in [PZ19a] O. Parshina and L.Q. Zamboni obtained explicit formulae for the closed and open complexity functions for Arnoux-Rauzy words on a k-letter alphabet (and hence in particular Sturmian words). They also showed that $\liminf \mathrm{Cl}_x(n) = +\infty$ when x is an Arnoux-Rauzy word. However, for a general aperiodic word, the $\liminf \mathrm{Cl}_x(n)$ may be finite, and in fact in [SS16], L. Schaeffer and J. Shallit proved that for the regular paperfolding word one has that $\liminf \mathrm{Cl}_x(n) = 0$, which is somewhat surprising. More generally, they showed that in the case of automatic sequences, the property of being closed is expressible in first-order logic, which allows them to compute the closed complexity for various well known infinite words including the Thue-Morse word, the Rudin-Shapiro word, the ordinary paperfolding word and the period-doubling word (for definition of those words, see, for instance, [AS03]).

One essential difference between the usual factor complexity on one hand, and the open and closed complexities on the other, is that the latter complexities are not in general monotone (e.g. see [PZ19a]). Nevertheless, we were able to prove a refinement of the Morse-Hedlund theorem that may be stated as follows, and constitutes the main result of this chapter:

Theorem 3.1.8. *Let $x \in \mathbb{A}^{\mathbb{N}}$ be a right-infinite word over a finite alphabet $\mathbb{A}$. The following are equivalent:*

1. *x is aperiodic;*
2. $\limsup\limits_{n\to+\infty} \mathrm{Cl}_x(n) = +\infty$*;*
3. $\liminf\limits_{n\to+\infty} \mathrm{Op}_x(n) = +\infty$.

In particular, both complexity functions are unbounded if x is aperiodic. Actually we prove something slightly more general, for which we need the following definition:

Definition 3.1.9. A subset S of $\mathbb{N}$ is *syndetic* if there exists a positive integer d such that $S \cap \{n, n+d\} \neq \emptyset$ for every $n \in \mathbb{N}$.

We prove that condition 2. can be replaced by $\limsup\limits_{n\in S} \mathrm{Cl}_x(n) = +\infty$, where S is any syndetic subset of $\mathbb{N}$. Of course, that conditions 2. and 3. each imply 1. is an immediate consequence of the Morse and Hedlund theorem. Since the limit inferior of the closed complexity of an aperiodic infinite word may be finite (as in the case of the regular paperfolding word) as it may be infinite (in the case of Sturmian words), we cannot hope to characterise periodicity in terms of $\liminf \mathrm{Cl}_x(n)$. Finally, it is necessary to assume the finiteness of the underlying alphabet, otherwise taking $x = 1234567\cdots \in \mathbb{N}^{\mathbb{N}}$, we see that x contains no closed factors of length greater than one.

3.1.4 Topological tools and graphs

We give and recall here some definitions and results unrelated to close and open words that will be needed in the rest of the chapter.

An operator often used in symbolic dynamics is the *shift operator* T. That operator has many uses in combinatorics on words. It is defined on infinite words as follows:

$$\forall x \in \mathbb{A}^{\mathbb{N}}, x = x_0x_1x_2\cdots, T(x) = x_1x_2x_3\cdots$$

For $x \in \mathbb{A}^{\mathbb{N}}$, we let $\Omega(x)$ denote the *shift orbit closure* of x, i.e. the closure in $\mathbb{A}^{\mathbb{N}}$ of the set

$$\{T^n(x) \mid n \in \mathbb{N}\}.$$

The following result (Theorem 1.5.11 in [Lot02]) will be useful in a later proof:

Theorem 3.1.10. *For every infinite word x over a finite alphabet, $\Omega(x)$ contains at least one uniformly recurrent element.*

For $x \in \mathbb{A}^{\mathbb{N}}$ and $n \in \mathbb{N}$, the *Rauzy graph* of order n of $x \in \mathbb{A}^{\mathbb{N}}$ is the directed graph whose set of vertices (resp. edges) consists of all factors of x of length n (resp. $n+1$). There is a directed edge from u to v labeled w if u is a prefix of w and v a suffix of w. A *path* of length k in a graph is an alternating sequence of vertices and edges $v_1, e_1, v_2, e_2, v_3, \ldots, v_k, e_k, v_{k+1}$ which begins and ends with a vertex and where each e_i is a directed edge from v_i to v_{i+1}. The *distance* between two vertices in a Rauzy graph is the length of the shortest path between them.

3.2 Words with finite $\liminf(\mathrm{Op}_x(n))_{n\in\mathbb{N}}$ are ultimately periodic.

The next two propositions as well as Corollary 3.2.2 also hold in case $\mathbb{A}$ is infinite.

Proposition 3.2.1. *Let $x \in \mathbb{A}^{\mathbb{N}}$ and $N \in \mathbb{N}$. Let w_1 and w_2 be two factors of x, such that there is a path of length i from w_1 to w_2 in the Rauzy graph of order N of x. Suppose w_1 and w_2 are closed with frontiers u_1 and u_2 respectively. Then $||u_1| - |u_2|| < i$.*
In particular, if $i = 1$ the frontiers are of the same length: $|u_1| = |u_2|$.

Proof. The situation is as illustrated on the Figure 3.1.

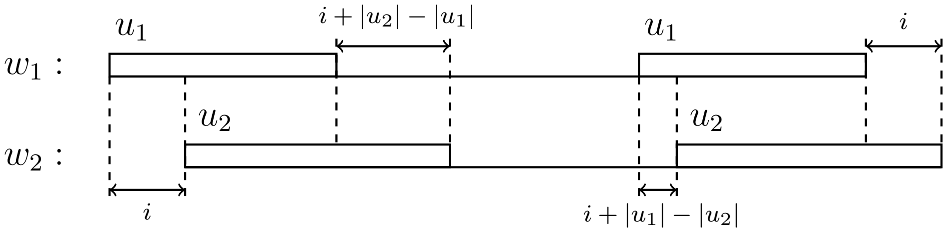


Figure 3.1: Factors w_1 and w_2.

Since w_2 is closed, u_2 cannot be a factor of u_1. Hence $i + |u_2| - |u_1| > 0$. Since w_1 is closed, u_1 cannot be a factor of u_2. Hence $i + |u_1| - |u_2| > 0$. The result follows. □

Corollary 3.2.2. *Let w_1, w_2, u_1, u_2 be as in Proposition 3.2.1. If there exists a path between w_1 and w_2 in the Rauzy graph consisting of only closed factors, then $|u_1| = |u_2|$. Thus, if there exists a path between w_1 and w_2 with n distinct open factors, $||u_1| - |u_2|| \leq n$.*

Proposition 3.2.3. *Let $x \in \mathbb{A}^{\mathbb{N}}$. For every $j > 1$, every vertex in the Rauzy graph of x of order j has at most one closed predecessor and one closed successor.*

Proof. Let w be a word of length $j-1$ and consider bw and cw to be both closed with $b, c \in \mathbb{A}$, $b \neq c$. Then, labelling u the frontier of bw, and v the frontier of cw, one has $|u| \neq |v|$, since u and v are both suffixes of w but do not start with the same letter. Suppose, without loss of generality, that $|u| < |v|$. This means u is a proper suffix of v, hence appears in w as a proper suffix of the first occurrence of v in cw. This leads to at least three occurrences of u in bw, which is then not closed. Symmetrically, there is at most one letter $b' \in \mathbb{A}$ such that wb' is closed.

□

Theorem 3.2.4. *Let x be an infinite word over a finite alphabet $\mathbb{A}$. Let $k \in \mathbb{N}$ be such that* $\liminf \mathrm{Op}_x(n) = k$*. Then x is ultimately periodic.*

In order to prove Theorem 3.2.4, we will start by proving some lemmas, where k, x and $\mathbb{A}$ are defined as in the theorem statement.

Lemma 3.2.5. *Suppose that x is aperiodic. Let $N > 11k+2$ be such that $\mathrm{Op}_x(N) = k$. Then $u^N \notin \mathrm{Fact}(x)$ for any choice of u with $|u| < 2k$.*

Proof. Suppose to the contrary that $u^N \in \mathrm{Fact}(x)$ for some primitive word $u \in \mathbb{A}^+$ with $|u| \leq 2k$. Since x is aperiodic, up to considering a cyclic rotation of u there exists $a \in \mathbb{A}$ such that $u^{\frac{N-1}{|u|}}a$ is a factor of x with $u^{\frac{N-1}{|u|}}a \neq u^{\frac{N}{|u|}}$. This factor is open: if not, its frontier has length at least $N-1-|u| > 3|u|+1$, which implies that u occurs internally in uu contradicting the fact that u is primitive (see Figure 2).

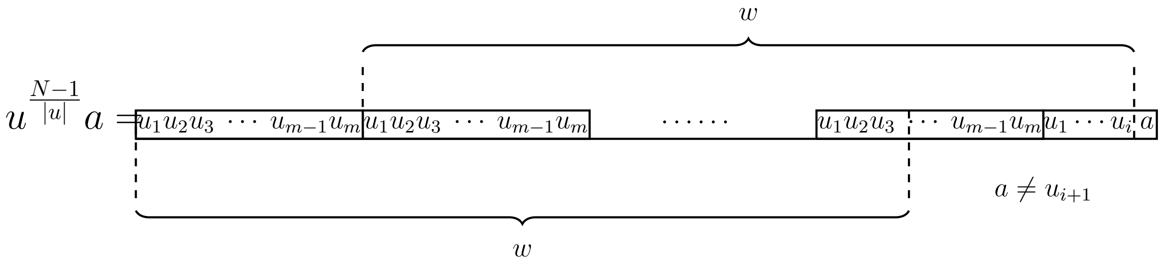

w
$u^{\frac{N-1}{|u|}}a =$
$u_1u_2u_3 \cdots u_{m-1}u_m$
$u_1u_2u_3 \cdots u_{m-1}u_m$
$\cdots\cdots$
$u_1u_2u_3 \cdots u_{m-1}u_m$
$u_1 \cdots u_i$
a
$a \neq u_{i+1}$
w

Figure 3.2: The frontier should be longer than $w = u^{\frac{N-1}{|u|}-1}$.

Let us consider, for $j \leq k$, a factor $u_{j+1}\cdots u_{|u|}u^{\frac{N-1}{|u|}-1}ab_1\cdots b_j$, which is a successor of $u^{\frac{N-1}{|u|}}a$ at distance at most k in the Rauzy graph of order N of x. Again, this factor is open: otherwise the length of its frontier would be at least $N-1-|u|-j > 3|u|+1$, and u would be a factor of uu. Besides, those factors are pairwise distinct, since an equality between two of them would imply that u is an internal factor of uu. This produces at least $k+1$ distinct open factors of length N, thereby contradicting our initial assumption on N. □

Lemma 3.2.6. *Let $j \in \mathbb{N}$ be such that $\mathrm{Op}_x(j) = k$. Let u and v be two closed factors of length j whose frontiers are of length r and p respectively. Then $|p-r| \leq k$.*

Proof. Consider the Rauzy graph of x of order j. By Corollary 3.2.2, it is enough to count the number of distinct open factors on a path between u and v to know the bound on $|p-r|$. There can be at most k of them, so $|p-r| \leq k$.

□

Lemma 3.2.7. *Suppose x is aperiodic. Let $m \in \mathbb{N}$, $t = |\mathbb{A}|$, and $N \geq k(t^m + m + 2)$ such that $\mathrm{Op}_x(N) = k$. Then the frontier of any closed factor of length N is longer than m.*

Proof. Let N be as above. Since x is aperiodic, it contains at least $N+1$ different factors of length N.

By Proposition 3.2.3, there exists a factor such that the shortest path in the Rauzy graph between it and an open factor is of length $\frac{N+1-k}{k}$. By Corollary 3.2.2, all closed words on this path have frontiers of the same length.

Let us suppose that this common frontier length is smaller than m. There are at most $t^m < \frac{N+1-k}{k}$ such frontiers, so by the pigeon hole principle two of those factors have the same frontier with their distance in the Rauzy graph being less than $t^m + 1$. Since this frontier cannot occur internally, the distance between those factors is at least $N - m$; hence $N - m < t^m + 1$, contradicting the definition of N. □

Proof of Theorem 3.2.4. Let $m = (11k+3)k + 2k$. In this case if a word of length at least $m - k$ overlaps itself with distance less than k, then it contains a power of exponent $11k+3$ with root shorter than k. Let $N > k(t^m + m + 2)$ be such that $\mathrm{Op}_x(N) = k$. Consider a right special factor $w = w_1 \cdots w_N$ of x (which exists since x is aperiodic). By Proposition 3.2.3, there exists $i \leq k$ such that $w_a = w_{i+1} \cdots w_N a y_1 \cdots y_{i-1}$ and $w_b = w_{i+1} \cdots w_N b z_1 \cdots z_{i-1}$ with $a \neq b \in \mathbb{A}$ are both closed factors of x. See Figure 3.3: at each step before the rightmost one, either on top, bottom, or both paths, there must be an open factor, and each open factor can only appear once.

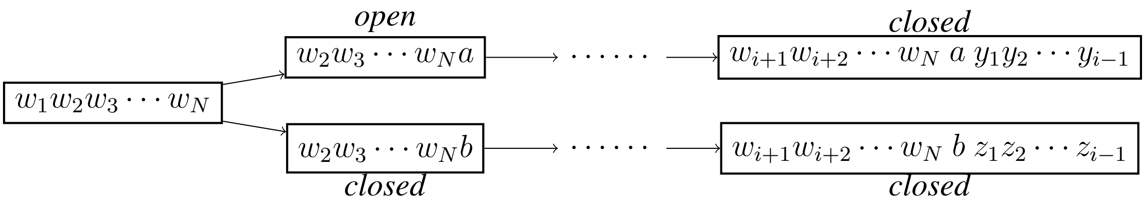

open
closed
$w_2w_3\cdots w_N a$
$w_{i+1}w_{i+2}\cdots w_N\ a\ y_1y_2\cdots y_{i-1}$
$w_1w_2w_3\cdots w_N$
$w_2w_3\cdots w_N b$
$w_{i+1}w_{i+2}\cdots w_N\ b\ z_1z_2\cdots z_{i-1}$
closed
closed
......
......

Figure 3.3: The sequence of open and closed factors in the Rauzy graph of order N.

Let us denote the frontiers of w_a and w_b by u and v respectively. For the illustration of the following reasoning see Figure 3.4. Applying Lemma 3.2.6, we get $||u| - |v|| \leq k$. Since both u and v are longer than k and $a \neq b$, they cannot be equal. This implies $|u| \neq |v|$ since w_a and w_b have a long common prefix. Suppose, without loss of generality, that $|u| < |v|$. Lemma 3.2.7 gives $m < |v|$. Let u' and v' be prefixes of u and v such that $u = u' a y_1 \cdots y_{i-1}$ and $v = v' b z_1 \cdots z_{i-1}$. Then, $|v'| > m - k$ and u' is a prefix and a suffix of v'. Hence v' overlaps itself with a difference less than k, what contradicts Lemma 3.2.5.

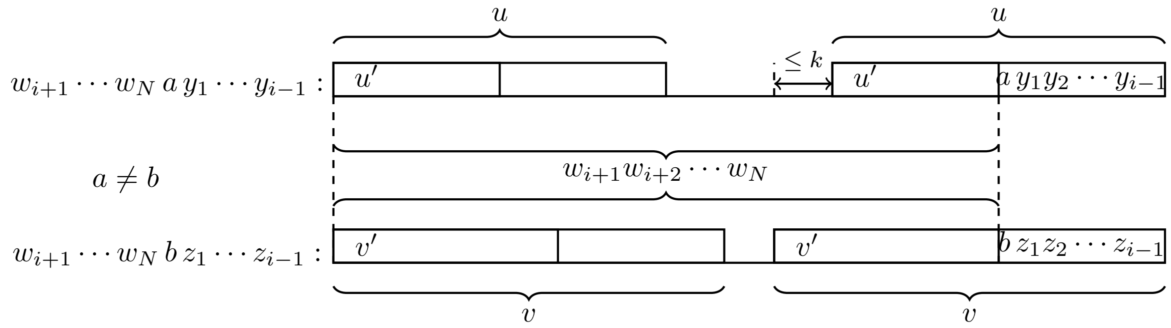

u
u
$w_{i+1}\cdots w_N\, a\, y_1\cdots y_{i-1}:$
u'
$\leq k$
u'
$a\, y_1 y_2\cdots y_{i-1}$
$a\neq b$
$w_{i+1}w_{i+2}\cdots w_N$
$w_{i+1}\cdots w_N\, b\, z_1\cdots z_{i-1}:$
v'
v'
$b\, z_1 z_2\cdots z_{i-1}$
v
v

Figure 3.4: v' overlaps itself with difference smaller than k.

□

3.3 Words with bounded closed complexity are ultimately periodic

The goal of this section is to prove the following theorem, which gives a characterisation of ultimately periodic words in terms of closed complexity.

Theorem 3.3.1. *Let $x \in \mathbb{A}^{\mathbb{N}}$ be such that there exist a positive integer d and a syndetic subset $S \subseteq \mathbb{N}$ with gaps smaller than d on which the closed complexity of x is bounded, i.e. there exists $k \in \mathbb{N}$ such that $\mathrm{Cl}_x(n) < k$ for every $n \in S$. Then x is ultimately periodic.*

In what follows, x, k and S are defined as in the theorem.

The following lemma states that every recurrent factor is close to being right (or left) special.

Lemma 3.3.2. *Let x be aperiodic. At least one of the two following assertions holds:*

1. $\forall u \in \mathrm{RecFact}(x), \exists (p \leq k, a_1, \ldots, a_p, a, b) \in \mathbb{N} \times \mathbb{A}^{p+2}, a \neq b$, *such that*

$$aa_1\cdots a_p u \in \mathrm{RecFact}(x) \quad \textit{and} \quad ba_1\cdots a_p u \in \mathrm{RecFact}(x);$$

2. $\forall u \in \mathrm{RecFact}(x), \exists (p \leq k+d, a_1, \cdots, a_p, a, b) \in \mathbb{N} \times \mathbb{A}^{p+2}, a \neq b$, *such that*

$$ua_1\cdots a_p a \in \mathrm{RecFact}(x) \quad \textit{and} \quad ua_1\cdots a_p b \in \mathrm{RecFact}(x).$$

Proof. Let us suppose to the contrary that neither of the above two assertions holds. Let $u \in \mathrm{RecFact}(x)$ and $a_1, \cdots, a_k$ be such that $a_1 \ldots a_k u$ is the only recurrent left extension of u. Also, let $v \in \mathrm{RecFact}(x)$ and $b_1, \cdots, b_{k+d}$ be such that $vb_1\cdots b_{k+d}$ is the only recurrent right extension of v. Up to considering a suffix y of x, we can assume that every occurrence of u is preceded by $a_1\cdots a_k$ and every occurrence of v is followed by $b_1\cdots b_{k+d}$. Let us consider a word w such that $uwv \in \mathrm{RecFact}(x)$, and a complete first return w' to this factor in y. We know that this factor is preceded by $a_1\cdots a_k$ and is followed by $b_1\cdots b_{k+d}$. The situation is the following:

$$\begin{aligned} w' &= uwvb_1 \cdots b_{k+d} \cdots a_1 \cdots a_k uwv, \\ x &= \cdots a_1 \cdots a_k w' b_1 \cdots b_{k+d} \cdots . \end{aligned}$$

Let $d' \leq d$ be such that $|w'| + k + d' \in S$.

Then we can find k closed factors of x of length $|w'| + k + d'$, contradicting the definition of S (see Figure 3.5). Indeed, for every i with $0 < i \leq k$, the factor $a_i \cdots a_k w' b_1 \cdots b_{d'+i-1}$ is closed, since there is no internal occurrence of uwv in w'.

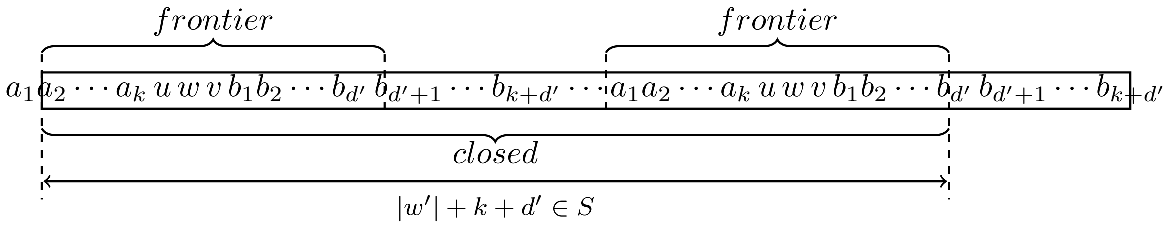

$frontier$
$frontier$
$a_1 a_2 \cdots a_k\, u\, w\, v\, b_1 b_2 \cdots b_{d'}\, b_{d'+1} \cdots b_{k+d'} \cdots a_1 a_2 \cdots a_k\, u\, w\, v\, b_1 b_2 \cdots b_{d'}\, b_{d'+1} \cdots b_{k+d'}$
$closed$
$|w'| + k + d' \in S$

Figure 3.5: k closed factors of x.

Let us show that all these factors are pairwise distinct. By contradiction, suppose that $i < i'$ are such that $a_i \cdots a_k w' b_1 \cdots b_{d'+i-1} = a_i' \cdots a_k w' b_1 \cdots b_{d'+i'-1}$.

In particular, $a_i \cdots a_k uwvb_1 \cdots b_{d'+i} = a_i \cdots a_{i'-1} a_i \cdots a_k uwvb_1 \cdots b_{d'+i-i'}$ which extends uniquely in x to $a_i \cdots a_{i'-1} a_i \cdots a_k uwvb_1 \cdots b_{d'+i} = (a_i \cdots a_{i'-1})^2 a_i \cdots a_k uwvb_1 \cdots b_{d'+i-i'}$, and so x is ultimately periodic with period $a_i \cdots a_{i'-1}$. □

Corollary 3.3.3. *Let x and S be defined as in Theorem 3.3.1. If x is uniformly recurrent, then it is periodic.*

Proof. Let us suppose that x is aperiodic (a uniformly recurrent word that is ultimately periodic is periodic). By Lemma 3.3.2, either every factor is close to being left special or every factor is close to being right special. Without loss of generality we can assume that we are in the second case. The idea of the proof is the following: using Lemma 3.3.2, we can produce a factor u of x such that if there is an overlap $uw = vu$, then $|v| > k + d$. Then using the same lemma we can construct an arbitrary long factor of x that does not contain u, contradicting the uniform recurrence of x.

Let us begin with considering a recurrent factor u of x. Applying the branching process from Lemma 3.3.2, we can extend u in a way that if u overlaps itself with $uw = vu$ for some factors v, w, then $|v| > 1$: at the first branching point a_p, where $u_1 \cdots u_n a_1 \cdots a_p$ is a recurrent extension of u, it is sufficient to take $a_{p+1} \neq a_p$ (or $a_1 \neq u_n$ if u is right special).

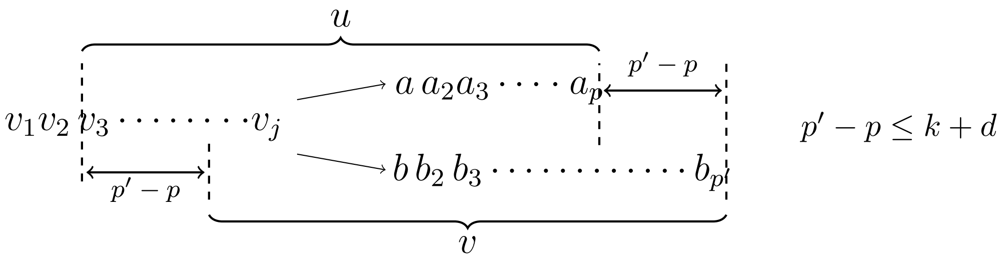
Applying the same process to $u^{(1)} = u_1 \cdots u_n a_1 \cdots a_{p+1}$ we obtain the factor

$$u_1^{(1)} \cdots u_{n+p+1}^{(1)} a_1' \cdots a_{p'}' .$$

Now we chose $a_{p'+1}' \neq a_{p'-1}'$ (or $a_{p'+1}' \neq u_{n+p+p'-1}^{(1)}$ if $p' \leq 2$) and we set $u^{(2)} = u_1^{(1)} \cdots u_{n+p+1}^{(1)} a_1' \cdots a_{p'+1}'$. Thus, if there exist factors v, w such that $u^{(2)} w = vu^{(2)}$, then

$|v| > 2$. We apply recursively the same reasoning $k+d$ times and get a recurrent factor $u^{(k+d)}$ that satisfies $(u^{(k+d)}w = vu^{(k+d)}) \Rightarrow (|v| > k+d)$. For simplicity of notation, we will denote this factor by u in the rest of the proof. Let us note, that to implement this construction, we only need the fact that every recurrent factor admits a right special extension, and so this can be done in any aperiodic word.

Since x is uniformly recurrent, there exists $n \in \mathbb{N}$ such that every factor of x of length n contains u. Let us construct a factor contradicting this. We start with $u_1 \cdots u_{|u|-(k+d)}$ and go to the next branching point given by Lemma 3.3.2, that is $u_1 \cdots u_{|u|-(k+d)+p}$. At this point we choose a letter that differs from $u_{|u|-(k+d)+p+1}$. This ensures that u does not occur before the next branching point (see Figure 3.6).



u
$p' - p$
$a\, a_2 a_3 \cdots\cdots a_p$
$v_1 v_2 v_3 \cdots\cdots\cdots v_j$
$p' - p \leq k + d$
$b\, b_2\, b_3 \cdots\cdots\cdots\cdots b_p$
$p' - p$
v

Figure 3.6: $u = v$ would produce an overlap of u.

This reasoning can be applied to construct our factor recursively: at each branching point, knowing that u does not appear before we can choose a branch such that u will not occur after adding of any $k+d$ letters to the right. Indeed, if it was not the case it would mean u appears in both branches shown in Figure 3.6. This allows us to construct, in at most n steps, a factor longer than n that does not contain u. □

The following lemma states that every periodic word in the subshift of x has a short period.

Lemma 3.3.4. *Let u be a primitive word in* $\mathrm{Fact}(x)$ *such that, for every $n \in \mathbb{N}$, $u^n \in$* $\mathrm{Fact}(x)$*. Then $|u| < k$.*

Proof. Let $p < |u|$ and $n > 2$ be such that $p + n|u| \in S$. Let us denote the i-th rotation of $u = u_1 \cdots u_{|u|}$ by $r^i(u) = u_i \cdots u_{|u|} u_1 \cdots u_{i-1}$. Since u is primitive, so is $r^i(u)$, hence each $(r^i(u))^n u_i \cdots u_{i+p}$ is closed with frontier $(r^i(u))^{n-1} u_i \cdots u_{i+p}$ (otherwise if the frontier occurred inside, then $r^i(u)$ would be an internal factor of $(r^i(u))^2$, contradicting primitivity of $r^i(u)$).

All the rotations $r^i(u)$ are pairwise distinct, and so are all closed factors $(r^i(u))^n u_i \cdots u_{i+p}$, $i = 1, 2, 3, \ldots, |u|$. By Theorem 3.3.1, $\mathrm{Cl}_x(n) < k$ and thus $|u| < k$. □

Proof of Theorem 3.3.1. Let x and S be as in the statement of the theorem. Let us suppose that x is aperiodic and consider the set $P_x = \{u \in \mathrm{Fact}(x) \,|\, \forall n \in \mathbb{N}, u^n \in \mathrm{Fact}(x)\}$. By Lemma 3.3.4, P_x is finite. According to Lemma 3.3.2, there exists $N \in \mathbb{N}$ such that we can produce an infinite word y in the subshift $\Omega(x)$ that does not contain u^N as a factor for any $u \in P_x$. Moreover, since $y \in \Omega(x)$, it verifies Lemmas 3.3.2 and 3.3.4, and thus $P_y = \emptyset$. Let z be any uniformly recurrent word in $\Omega(y)$. By Corollary 3.3.3 the word z is periodic and can be represented as $z = u^\omega$ for some finite word u. Then $u \in P_y$, what contradicts $P_y = \emptyset$. Hence x is ultimately periodic. □

3.4 Concluding remarks and open questions

Combining Theorem 3.2.4, Theorem 3.3.1 and the fact that $\mathrm{Op}_x(n)+\mathrm{Cl}_x(n) = p_x(n)$ for every $n \in \mathbb{N}$, we obtain the following result:

Theorem 3.4.1. *Let x be an infinite word over a finite alphabet $\mathbb{A}$. The following are equivalent:*

1. *x is aperiodic;*
2. *$\forall S \subseteq \mathbb{N}$ syndetic, $\limsup\limits_{n\in S} \mathrm{Cl}_x(n) = +\infty$;*
3. *$\liminf\limits_{n\to+\infty} \mathrm{Op}_x(n) = +\infty$.*

Since the factor complexity of an aperiodic word is a strictly increasing function, $\liminf p_x(n) = +\infty$ is equivalent to p_x being unbounded. However, it is not always the case for open and closed complexity functions (ex. see [PZ19a]). Even though the result we obtained in terms of open complexity is as strong as Morse-Hedlund theorem since it is expressed in terms of $\liminf \mathrm{Op}_x(n)$, the characterisation in terms of closed complexity cannot be improved to the same setting. In fact, it is already known that some aperiodic words can have $\liminf \mathrm{Cl}_x(n) < +\infty$. For example, L. Schaeffer and J. Shallit showed in [SS16] that $\liminf \mathrm{Cl}_x(n) = 0$ when x is the paperfolding word. It is even possible for pure morphic words to have finite limit inferior for the closed complexity: For example, for the celebrated Cantor word c, also sometimes referred to as the Sierpinski word, one verifies that $\liminf \mathrm{Cl}_c(n) = 1$, and this value is attained for $n = 7 \cdot 3^k + 1$, for any k. The proof of this result can be easily obtained with case by case study, and thus is omitted; however it leads to the following question:

Question 3.4.2. *Is it possible to find, for any $k \in \mathbb{N}$, an aperiodic pure morphic word x such that $\liminf \mathrm{Cl}_x(n) = k$?*

Although it is not possible to have the equivalence $\liminf \mathrm{Cl}_x(n)$ bounded $\Longleftrightarrow$ x ultimately periodic, it might still be possible to obtain something stronger than our theorem: we already improved the first version of the theorem to the setting of syndetic sets, but it may be possible to get the same result in the case of piecewise syndetic sets. A set S is piecewise syndetic if it is the intersection of a syndetic set and a thick set, hence if there are arbitrarily long intervals where it has bounded gaps.

Question 3.4.3. *Is it true that, for any aperiodic word x and any piecewise syndetic set S, $\limsup\limits_{n\in S} \mathrm{Cl}_x(n) = +\infty$?*

Chapter 4

Antipowers in infinite words

Contents

4.1 Introduction

4.1.1 Unavoidable regularities

In combinatorics, an *unavoidable regularity* is a property P such that every sufficiently large structure satisfies P. In term of combinatorics on words, the definition would be the following:

Definition 4.1.1. An *unavoidable regularity* is a property P such that it is not possible to construct arbitrarily long words not satisfying P.

The study of unavoidable regularities, known as Ramsey theory, has been an important research field of combinatorics and theoretical computer science for the past century. The aim of this theory was mainly to study the regular pattern of all-equal elements that arise in any large enough structure. Amongst other important results, the works of F. Ramsey [Ram30] in 1930 have multiple applications in different fields of mathematics. One older and very important result of Ramsey theory is the following, known as Van der Waerden's theorem [van27]:

Theorem 4.1.2. *For every finite colouring c of $\mathbb{N}$, there exist monochromatic arithmetic progressions of any length, id est $\forall I \subset \mathbb{N}, |I| < +\infty$ and $c : \mathbb{N} \to I$, for every $N \in \mathbb{N}$, the following is verified:*

$$\exists (d,p) \in \mathbb{N}^2, c(p) = c(p+d) = \cdots = c(p+Nd).$$

In terms of combinatorics of words, this means that any infinite word over a finite alphabet contains arithmetic progressions of any length of constant letter: let $\mathbb{A}$ be a finite alphabet, then any infinite word w over $\mathbb{A}$ satisfies the following property:

$$\forall N \in \mathbb{N}, \exists (d,p) \in \mathbb{N}^2, w_p = w_{p+d} = \cdots = w_{p+Nd}.$$

Van der Waerden's theorem for infinite words is a corollary from Van der Waerden's theorem for finite words [van27]:

Theorem 4.1.3. *Let n and k be natural integers. Then, there exists an integer $W(r,k)$ such that, for every $c : \mathbb{N} \to \{1, r\}$ colouring of the integers with r colors, the set $\{1, 2, \cdots, W(r,k)\}$ contains a monochromatic arithmetic progression of length k.*

Again, this result can be expressed in terms of combinatorics on words: let $\mathbb{A}$ be a finite alphabet with $|\mathbb{A}| = r$, then for every $k \in \mathbb{N}$, there exists an integer $W(r,k)$ such that every word u on $\mathbb{A}$ of length at least $W(r,k)$ satisfies the following property:

$$\text{There exists two integers } p \text{ and } d \text{ such that } u_p = u_{p+d} = \cdots = u_{p+kd}.$$

In fact, in the field of combinatorics on words, the question of unavoidable regularities has been extensively studied. It is even at the foundation of this branch of mathematics, since the founding articles from A. Thue [Thu06, Thu12] dealt with the avoidability of $powers$ in infinite words, already introduced in Introduction. I recall this definition here:

Definition 4.1.4. A word of length kn, for k and n integers, is a (k,n)-*power*, or power of order k and length of the blocks n, if it is concatenation of k identical blocks of length n. Namely, $u = u_1 \cdots u_{kn}$ is a (k,n)-power if $u_1 \cdots u_n = u_{pn+1} \cdots u_{(p+1)n}$ for any p. Most of the time, we don't precise n, and only talk about k-powers.

Example 4.1.5. $ababab$ is a $(3,2)$-power, or more simply put, a $cube$. $abbabb$ is a square, and aba is not a power (and does not contain any square).

Remark 16. For $k \geq 3$, any (k,n)-power contains a $k-1$-power as strict factor, since every factor of length $(k-1)n$ is a $k-1$-power.

Many questions pertaining to powers have been studied in words since the founding work of A. Thue. His original question was the following: is it possible, on an alphabet $\mathbb{A}$ with n letters, to construct an infinite word avoiding squares?

On a two-letter alphabet, the answer to this question is no, since every word of length greater than 3 must contain a square. But A. Thue showed that it is possible to get something close to that: he proved that the celebrated Prouhet-Thue-Morse word is *overlap-free*, where an overlap is a word of the form $xuxux$ with x a letter and u a word.

Recall that a *morphism* σ over $\mathbb{A}$ is a map $\sigma : \mathbb{A}^* \to \mathbb{B}^*$ where $\mathbb{B}$ is another alphabet such that $\sigma(ww') = \sigma(w)\sigma(w')$. A *substitution* σ over $\mathbb{A}$ is a morphism where $\mathbb{A} = \mathbb{B}$.

Theorem 4.1.6. *The Prouhet-Thue-Morse word*

$$t = 0110100110010110100101100110100110010110011010\cdots$$

which is the fixed point of the morphism $\mu(0) = 01, \mu(1) = 10$ *starting with 0, is overlap-free.*

The demonstration of this theorem uses the fact that, given a long enough factor u of t, there is only one possible slicing of u coinciding with the images of letters by μ in $t = \mu^{\infty}(0)$. That property is called recognizability, it will be used latter in this chapter.

Using this result, A. Thue was able to construct a word on a three-letter alphabet that avoided squares:

Theorem 4.1.7 (Thue, 1906)**.** *The word*

$$w = 2102012101202102012021010121\cdots$$

obtained as the fixed point of the substitution $0 \mapsto 1, 1 \mapsto 20, 2 \mapsto 210$*, does not contain any square, that is, a pattern of the form* xx*.*

Proof. The word w is also given by the number of 1s between two consecutive 0s in t. To see this, we need to compute the images of 0110, 010 and 00 under μ. We have $\mu(0110) = 01101001$ so the image of 2 would be 210, while $\mu(010) = 011001$ so the image of 1 is 20, and $\mu(00) = 0101$, so the image of 0 would be 1.

Now let us prove that w does not contain squares: suppose that uu is a factor of w with $u = u_1 \cdots u_n$ a non-empty word. Then $01^{u_1}01^{u_2}0 \cdots 1^{u_n}01^{u_1}01^{u_2}0 \cdots 1^{u_n}0$ is a factor of t, but it is also an overlap and we know t is overlap-free. So w is square free. □

Since Thue's results, the intersection between Ramsey theory and combinatorics on words has been studied intensively [WZ18a, WZ18b, PZ15, BPZ15, dLPZ14, dLPZ13, BHPZ13, BZ13, dZ16, dLZ16].

4.1.2 Antipowers

More recently, a branch of Ramsey theory has developed, based on the work of Erdős, Simonovits and Sós, often called anti-Ramsey theory, studying the avoidability of pattern of *all-distinct* elements. One such example is the study of *rainbows*, i.e. subgraphs of an edge-colored graph whose edges have all different colors - see [FMO10] for a survey. An anti-Ramsey notion has been introduced recently in combinatorics on words by Fici, Restivo, Silva and Zamboni [FRSZ18], the notion of *antipowers*.

Definition 4.1.8. A word of length kn, for k and n integers, is a (k, n)-*antipower*, or antipower of order k and length n, if it is concatenation of k pairwise distinct blocks of length n. Namely, $u = u_1 \cdots u_{kn}$ is a (k, n)-antipower if $u_{in+1} \cdots u_{(i+1)n} \neq u_{jn+1} \cdots u_{(j+1)n}$ for any i and j. Again, we will often write k-antipower, without giving n.

This definition is the analogous of the definition of powers when one is looking for *diversity* instead of all-equal objects. Since powers have been extensively studied, it seems really natural to introduce this definition.

Example 4.1.9. $abba$ is a $(2, 2)$-antipower, but not a $(4, 1)$-antipower.

Remark 17. There are a lot of antipowers: let us, for example, consider the words of length 4 on a binary alphabet. Up to a permutation of letters, those are the following eight words:

$$aaaa,\ aaab,\ aaba,\ aabb,\ abaa,\ abab,\ abba,\ abbb$$

and of those six are 2-antipowers:

$$aaab,\ aaba,\ aabb,\ abaa,\ abba,\ abbb.$$

It then seems that antipowers would be pretty hard to avoid, but it is still possible to construct infinite words that avoid them: take, for example,

$$aaaaa\cdots = a^{\omega}.$$

This clearly avoids every possible antipower.

However, in their original article [FRSZ18], Fici et al. were able to prove the following anti-Ramsey result:

Theorem 4.1.10. *Every infinite word contains powers of any order or antipowers of any order.*

Remark 18. It is not surprising that if we want to be sure that a word contain antipowers of any order we suppose it does not contain long powers; as our example showed, long powers are a place where there are no antipowers.

Like often in Ramsey theory, the result Fici et al. proved admits a stronger but less easy-to-state formulation. The stronger version of their theorem provides a bound $N(l,k)$, given an alphabet size and two integers l and k greater than 1, such that every word of length at least $N(l,k)$ contains a l-power or a k-antipower:

Theorem 4.1.11 (Theorem 14 in [FRSZ18])**.** *For all integers $l > 1$ and $k > 1$ there exists $N = N(l,k)$ such that every word of length N contains a l-power or a k-antipower. Furthermore, for $k > 2$, one has $k^2 - 1 \leq N(k,k) \leq k^3\binom{k}{2}$.*

Proof. The proof we will give here is very close from the one in [FRSZ18], but for simplicity we don't give the best possible bound. What we will be interested in is the difficulty to extend this proof to other settings, which will be possible to discuss with this formulation.

Let w be an infinite word avoiding k-antipowers. Consider its prefix $u_{M,1}u_{M,2}\cdots u_{M,k}$ where $|u_{M,i}| = M$ with M a (big) integer we will determine later. Since w does not contain k-antipowers, we can find i and j such that $u_{M,i} = u_{M,j}$.

Now consider the prefix of w of length $(M+1)k$: $u_{M+1,1}u_{M+1,2}\cdots u_{M+1,k}$ where $|u_{M+1,i}| = M+1$. Again, w avoids k-antipowers, hence we can find i', j' such that $u_{M+1,i'} = u_{M+1,j'}$. There are less than k^2-1 couple (i,j) with $1 \leq i < j \leq k$, so iterating this process, in less than k^2 step, we will have a couple i,j such that $u_{M+p,i} = u_{M+p,j}$ and $u_{M+q,i} = u_{M+q,j}$.

If M is taken to be large enough with respect to k, this will create a long factor overlapping itself, and hence containing a power (this type of argument has already been used in our chapter on open and closed complexity). See the graph below for a graphic explanation:

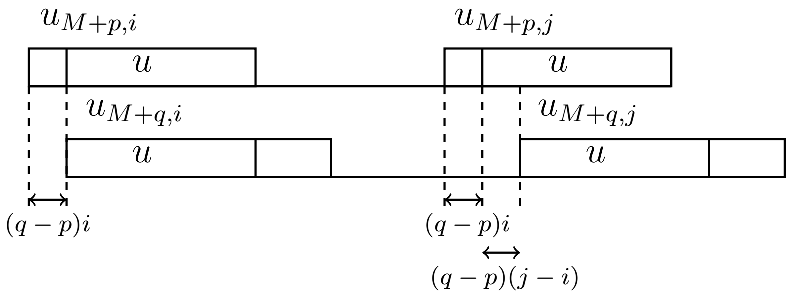


Figure 4.1: w contains a large power.

The difference between the starting index of $u_{M+p,i}$ and $u_{M+q,i}$ is

$$(M+q)i - (M+p)i = (q-p)i$$

and the difference between the starting point of $u_{M+p,j}$ and $u_{M+q,j}$ is

$$(M+q)j - (M+p)j = (q-p)j.$$

Hence, labeling u the prefix of $u_{M+q,i}$ which is a suffix of $u_{M+p,i}$, we have that u overlaps itself with difference $(q-p)(j-i)$, and that leads to the construction of a long power: $|u| = M + p - (q-p)(j-i) \geq M - kk^2 = M - k^3$ so for $M > Rk^3 + k^3$ we have that u starts with a R-power at least. □

These results attracted a lot of interest ([Gae18, Bur18, Ria19, BFP18, Def17, FPS19, Pos19, Gar19, FRS20, KRR$^+$19, Nar17]) that studied mainly four different type of questions. Like often in Ramsey theory, one of the subject of interest was to improve the bounds. The first contribution I will present dealt about this. Another topic investigated was the possible generalizations of antipowers to more general settings. The main subject of this thesis was initially related to this question. Finally, some articles studied antipowers in some well known words, and some tried to find efficient algorithms to compute them.

4.2 Improving the bounds for uniform morphic words

This section contains new results which are the subject to an article on arXiv: [Pos19].

We will first start with results pending to the bound $N(k,l)$ of Theorem 4.1.11. There has been a lot of results on this field since Fici et al. published their article (see [FRS20, Nar17, Gae18, Gar19, Pos19, Bur18, Def17]).

In one of those contributions, A. Berger and C. Defant [BD20] studied the block length of antipowers arising in morphic words. The main result in this part answers a question they left open.

4.2.1 Introduction

We begin by recalling some basic notions needed to describe the problem.

Definition 4.2.1. A *uniform* morphism σ is a morphism of constant length over letters: $\forall\, a, b \in \mathbb{A},\ |\sigma(a)| = |\sigma(b)|$. If $|\sigma(a)| = m$ then we say σ is a m-uniform morphism.

Remark 19. In most articles pertaining to recognizability the term substitution is favored to talk about morphisms from $\mathbb{A}$ to $\mathbb{A}$. In [BD20] the term morphism is used instead of substitution. We indifferently use both.

Definition 4.2.2. Let $\mathbb{A} = \{a_1, \cdots, a_r\}$. A morphism σ over $\mathbb{A}$ is said to be *primitive* if:

$$\exists n,\ \forall i,\ \forall j,\ a_j \text{ occurs in } \sigma^n(a_i).$$

For a morphism σ, $x \in \mathbb{A}^{\mathbb{N}}$ is called a *fixed point* if $x = \sigma(x)$. Recall from the previous chapter that the *shift orbit closure* $\Omega(x)$ is the closure under the natural topology on $\mathbb{A}^{\mathbb{N}}$ of the orbit of x under the shift operator $\tau : a_1a_2\cdots \to a_2a_3\cdots$. If σ is primitive, it is easy to see that $\Omega(x) = \Omega(y)$ for any x and y fixed points of σ. Hence we can define $\Omega(\sigma) = \Omega(x)$ in this case.

Remark 20. For a word x, having $\Omega(x) = \Omega(y)$ for every $y \in \Omega(x)$ is equivalent to

Up to changing the definitions slightly, Theorem 5 in [BD20] can be formulated the following way:

Theorem 4.2.3. *If w is aperiodic, fixed point of a primitive binary uniform morphism, then there is a constant $C = C(w)$ such that $\forall n,\ k \in \mathbb{N},$ w contains a k-antipower with blocks of length at most Ck beginning at its n^{th} position.*

This is a very big improvement on the bound given by Fici et al., which is $k^3\binom{k}{2}$.

Berger and Defant asked to what extent are these results generalisable to a broader class of morphic words, and in particular if it was still true without the binary condition. Using the notion of recognisability first introduced by B. Mossé in [Mos96], we answer that question by showing that their result extends to fixed points of uniform primitive morphisms on arbitrary finite alphabets:

Theorem 4.2.4. *If σ is a primitive m-uniform morphism over a finite alphabet $\mathbb{A}$, with an aperiodic fixed point x, there exists a constant $C = C(\sigma)$ such that: $\forall y \in \Omega(\sigma),\ \forall n, k \in \mathbb{N},$ y contains a k-antipower with block length at most Ck starting at position n.*

This result was discovered independently by S. Garg [Gar19]. The proof he used is totally different.

4.2.2 Recognizability

Definition 4.2.5. A m-uniform primitive morphism σ is said to be *recognizable* if $\exists N \in \mathbb{N}$ such that $\forall y \in \Omega(\sigma), \forall w \in \mathbb{A}^+, \sigma(y)_{[\alpha,\alpha+|w|-1]} = \sigma(y)_{[\beta,\beta+|w|-1]} = w$ with $|w| \geq N$ and $\alpha \equiv 0 \pmod m$ then $\beta \equiv 0 \pmod m$. N is refered to as *recognizability constant* of σ in this chapter.

This notion is fundamental in the theory of primitive substitutions. In fact, as I already mentioned, it was used by Axel Thue on his proof of the overlap-freeness of the Prouhet-Thue-Morse word t. Indeed, take any factor of length 3 of t. It will contain 0 as a factor. Hence, any factor of length at least $18 = 4 \cdot 3 + 3 + 3$ contain the image under μ^2 of a factor of length 3, so it contains $\mu^2(0) = 0110$. But, whenever you see 11, there is only one possibility: the first 1 is the end of the image under μ of a 0 and the second 1 is the image under μ of a 1. This means, as soon as the factor you are considering is of length at least 18, you can be sure that if you are considering two occurences of it, their distance will be a multiple of 2 since the slicing of their 11 must be the same.

It is always easy, given a substitution, to substitute, i.e. to apply it to any word; but it is far less obvious to desubstitute, which is, given a word u and a morphism σ, to find the shortest v such that u is a factor of $\sigma(v)$. In fact, it might even be possible that v is not unique. The substitution theory provides us with cases where we can tell for sure that it is possible to find a unique v. The notion of recognizability introduced on Definition 4.2.5 can be extended to non-uniform morphisms, with this idea of being able to desubstitute for factors long enough.

It is important to note that not every morphism is recognizable; for example, the morphism ϕ on $\{a, b\}$ defined by $\phi(a) = aba$ and $\phi(b) = ba$ is not recognizable.

Remark 21. Let σ be a m-uniform primitive recognizable morphism and N given by Definition 4.2.5. Then $\forall y \in \Omega(\sigma), \forall w \in \mathbb{A}^+, \sigma(y)_{[\alpha,\alpha+|w|-1]} = \sigma(y)_{[\beta,\beta+|w|-1]} = w$ with $|w| \geq N + m$ gives $\alpha \equiv \beta \pmod m$.

Proof. Denote $h = \beta - \alpha$ and $w' = \sigma(y)_{[m\lceil \frac{\alpha}{m}\rceil,\alpha+|w|-1]}$. Then $|w'| \geq N$ and

$$w' = \sigma(y)_{[h+m\lceil \frac{\alpha}{m}\rceil,h+\alpha+|w|-1]}.$$

By definition, this implies $h \equiv 0 \pmod m$ hence $\alpha \equiv \beta \pmod m$. □

Remark 22. If $r = |\mathbb{A}| = 2$, an aperiodic word w that is a fixed point of a m-uniform morphism σ is uniformly recurrent if and only if σ is primitive.

Proof. Without loss of generality, let us fix $\mathbb{A} = \{0, 1\}$ with $w = \sigma^\infty(0)$.

Let us first suppose σ is primitive. It is easy to see that the n in Definition 4.2.2 is 2 or 1. Let then x be a factor of w. There exists k such that $x \in \text{Fact}(\sigma^k(0))$. Every factor of w of length at least $2m^{k+2}$ contains, for some $a \in \mathbb{A}$, $\sigma^{k+2}(a)$, so it contains $\sigma^k(0)$ hence x, and so w is uniformly recurrent.

Let us now suppose σ is not primitive. Since σ is m-uniform, w has to be eventually periodic. Indeed, if $\sigma(0) = 0^m$, then $w = 0^\infty$. If not and $\sigma(1) = 1^m$, then w contains arbitrary long runs of consecutive 1 ($1^\infty \in \Omega(w)$), hence arbitrary long factors do not contain the factor 0. The only option left is $\sigma(0) = 1^m$ and $\sigma(1) = 0^m$. But this leads to no fixed point ($\sigma(0)$ must start with a 0). □

We will now give some well-known results on substitutions and recognizability that we will need later:

Theorem 4.2.6 (Corollary 3.2 in [Mos96])**.** *Let σ be a primitive m-uniform substitution and let x be aperiodic such that $\sigma(x) = x$. Then σ is recognizable.*

We will also use the following proposition from [HZ99] (actually what is proved is somewhat stronger, but we only need this formulation):

Proposition 4.2.7. *If σ is a m-uniform morphism, and x is aperiodic with $x = \sigma(x)$ and $x_0 = a$, then $\exists N_1 \in \mathbb{N}$ such that $l \geq N_1$ implies that each occurrence $\sigma^l(a)$ in x is the image under σ of an occurrence of $\sigma^{l-1}(a)$ in x.*

4.2.3 Main Part

The goal of this part is to prove Therorem 4.2.4.

We will first give a lemma that is easily deduced from Proposition 4.2.7:

Lemma 4.2.8. *Let σ be a m-uniform morphism, and x aperiodic with $x = \sigma(x)$ and $x_0 = a$. Let $N_1 \in \mathbb{N}$ be given by Proposition 4.2.7. Then for every $r,\ l \in \mathbb{N}$, each occurrence $\sigma^{l+N_1+r}(a)$ in x is the image under σ^l of an occurrence of $\sigma^{N_1+r}(a)$ in x.*

Proof. By induction on l. Base case is just the result of Proposition 4.2.7. Let $l \in \mathbb{N}$ be fixed. Suppose the result holds for l and let $x_{[\alpha,\alpha+m^{l+1+N_1+r}-1]} = \sigma^{l+1+N_1+r}(a)$. Then by the recurrence hypothesis, $\alpha = m^l\alpha'$ and $x_{[\alpha',\alpha'+m^{1+N_1+r}-1]} = \sigma^{1+N_1+r}(a)$. But now we can apply Proposition 4.2.7: $\alpha' = m\alpha''$ and $x_{[\alpha'',\alpha''+m^{N_1+r}-1]} = \sigma^{N_1+r}(a)$. So $x_{[\alpha,\alpha+m^{l+1+N_1+r}-1]} = \sigma^{l+1+N_1+r}(a)$ is the image under σ^{l+1} of an occurrence of $\sigma^{N_1+r}(a)$ in x. □

We can now prove the following lemma:

Lemma 4.2.9. *If σ is a primitive m-uniform morphism, and x is aperiodic with $x = \sigma(x)$, $\exists N' \in \mathbb{N}$ such that $\forall i \in \mathbb{N}$, σ^i is recognizable with a recognizability constant less or equal to m^iN'.*

Proof. Let then σ be an aperiodic m-uniform morphism, and $x = \sigma(x)$ with $x_0 = a$. Let N_1 be given by Proposition 4.2.7, and let $y \in \Omega(\sigma)$.

By Theorem 4.2.6, σ is recognizable. Let N be a recognizability constant of σ and let $r \in \mathbb{N}$ such that $m^{N_1+r} \geq N + m$. The prefix of x of length m^{N_1+r} is $p = \sigma^{N_1+r}(a)$. Since σ is primitive and uniformly recurrent, $\exists M \in \mathbb{N}$ such that every factor of x, hence of y, of length at least M contains p. I then claim that $N' = 2M$ has the required property.

Indeed, let w with $|w| \geq N'm^i$ be fixed. We show the following:

$$\sigma^i(y)_{[\alpha,\alpha+|w|-1]} = \sigma^i(y)_{[\beta,\beta+|w|-1]} = w \Rightarrow \beta \equiv \alpha \pmod{m^i} \quad (1).$$

Let α, β be as in (1) and $h = \beta - \alpha$. So:

$$\sigma^i(y)_{[\alpha,\alpha+|w|-1]} = \sigma^i(y)_{[\alpha+h,\alpha+h+|w|-1]} = w.$$

By $y \in \Omega(\sigma)$ we get an α' with

$$x_{[\alpha',\alpha'+|w|-1]} = x_{[\alpha'+h,\alpha'+h+|w|-1]} = w. \quad (2)$$

Since $|w| \geq N'm^i = 2Mm^i$, there exists $\gamma \geq 0, w' \in \mathrm{Fact}(w)$ with
$w' = x_{[\gamma m^i,(\gamma+M)m^i-1]} = x_{[\gamma m^i+h,(\gamma+M)m^i+h-1]}$. Let $z = x_{[\gamma,\gamma+M-1]}$ so $w' = \sigma^i(z)$.
Since $|z| = M$, $\exists\gamma'$ with $\gamma \leq \gamma' < \gamma' + m^{N_1+r} - 1 \leq \gamma + M - 1$ such that:

$$x_{[\gamma',\gamma'+m^{N_1+r}-1]} = \sigma^{N_1+r}(a).$$

Applying σ^i to x gives $x_{[\gamma'm^i,(\gamma'+m^{N_1+r})m^i-1]} = \sigma^{N_1+r+i}(a)$, and by (2),

$$x_{[\gamma'm^i,(\gamma'+m^{N_1+r})m^i-1]} = x_{[\gamma'm^i+h,(\gamma'+m^{N_1+r})m^i+h-1]} = \sigma^{N_1+r+i}(a).$$

Using Lemma 4.2.8, this implies $h \equiv 0 \pmod{m^i}$: $x_{[\gamma'm^i+h,(\gamma'+m^{N_1+r})m^i+h-1]}$ is the σ^i image of $x_{[\delta,\delta+m^{N_1+r}]} = \sigma^{N_1+r}(a)$, hence

$$\gamma'm^i + h = \delta m^i \text{ and so } h \equiv 0 \pmod{m^i}.$$

□

Proof of theorem 4.2.4. Let σ be aperiodic, primitive and m-uniform and let $C(\sigma) = (N'+1)m$ where N' is the constant given by Lemma 4.2.9. Let then $k, n \in \mathbb{N}$ and $y \in \Omega(\sigma)$ be fixed.

Let $i \in \mathbb{N}$ be such that $m^{i-1} \leq k < m^i$. Then consider the k consecutive blocks of length $N'm^i + 1$ starting at position n: the block number s is then $y_{[n+s(N'm^i+1),n+(s+1)(N'm^i+1)-1]}$. We have $kC(\sigma) \geq (N'+1)m^i \geq N'm^i + 1$. Moreover, using Lemma 4.2.9, we get that two of these blocks, say blocks s and t, are equal implies the difference between their starting indices is 0 modulo m^i:

$$y_{[n+s(N'm^i+1),n+(s+1)(N'm^i+1)-1]} = y_{[n+t(N'm^i+1),n+(t+1)(N'm^i+1)-1]} \tag{4.1}$$

$$\Rightarrow n + s(N'm^i + 1) \equiv n + t(N'm^i + 1) \pmod{m^i} \tag{4.2}$$

$$\Rightarrow s \equiv t \pmod{m^i}. \tag{4.3}$$

But since k is smaller than m^i this implies $s = t$, which completes the proof. □

4.2.4 Conclusion

Using the theory of recognizability, we have been able to improve Berger and Defant's result to the class of aperiodic fixed points of primitive and m-uniform morphisms. The same result was obtained in [Gar19] without the use of the known theorems on recognizability. The two following examples show that the conditions aperiodic and primitive are tight. Let $\sigma : \begin{cases} 0 \to 01 \\ 1 \to 01 \end{cases}$. Since $(01)^\infty$, which is not aperiodic, does only contain two factors of each length, it cannot contain k-antipowers for k greater than 2. Let then $\sigma : \begin{cases} 0 \to 010 \\ 1 \to 111 \end{cases}$. This substitution is not primitive, and give rise to the celebrated Cantor word $c = \sigma^\infty(0)$. Since c contains arbitrary long runs of consecutive 1, it is clear that Theorem 4.2.4 doesn't apply here.

On the other hand, this result might be extendable to the class of recognizable substitutions, alas, I was not successful in finding an equivalent to Lemma 4.2.9; this seems to be the key to extend this result.

One non-trivial part of this result is that we are sure to find antipowers of any order at every position in w, aperiodic fixed point of a primitive uniform morphism. This result is not trivial, since some words containing antipowers of order k don't have prefixes that are k-antipowers:

Theorem 4.2.10. *Let k be a positive integer, $k > 4$. On every alphabet with at least two letters, there exist (recurrent) words w containing k-antipowers such that, for every k, w doesn't start with a k-antipower of any length.*

Proof. On k letters, it is easy to see that this is true even for $k > 2$: take the word $aabc \cdots ka^\omega$. This word contains a k-antipower, $abc \cdots k$ but never starts with a k-antipower as, calling i the length of the blocks, any k-tuple starting at the beginning would contain at least two blocks a^i. This result can be extended to two-letter alphabet if $k > 3$: consider the word

$$0^{k^2}0^k0^{k-1}10^{k-2}1^2 \cdots 01^{k-1}1^k0^\omega = 0^{k^2}(\prod_{i=0}^{k} 0^{k-i}1^i)0^\omega.$$

It contains a k-antipower of block of length k but no k-antipower starting at the beginning of the word.

In fact, it is possible to improve this result further for $k > 4$, by taking a word that is recurrent: consider the sequence of words defined by $u_0 = u$ and for each n, $u_{n+1} = u_n 0^{k_{n+1}} u_n$ and take their limit

$$w = u0^{k_1}u0^{k_2}u0^{k_1}u0^{k_3}u0^{k_1}u0^{k_2}u0^{k_3}\cdots$$

where $u = 0^{k^2}0^k0^{k-1}10^{k-2}1^2\cdots01^{k-1}1^k$ and k_i is greater than k times what is preceding in the word.

This word is clearly recurrent since after a long enough run of 0s we put a copy of what was before. Moreover, it contains a k-antipower since u contains one. But it contains no k-antipower starting at the beginning of the word. Indeed, suppose $v_1 \cdots v_k$ is such an antipower. Since u starts with o^{k^2} we have $|v_1| > \frac{k^2}{2}$. Let j be minimal such that v_1 ends before the start of the first run of 0 of length k_j. If v_1 ends after the run of 0 of length k_{j-1}, then $v_3 = v_4 = 0^{|v_1|}$. If $v_2 \neq 0^{|v_1|}$ then $v_4 = v_5 = 0^{|v_1|}$. Finally, if $v_2 = 0^{|v_1|}$ and $v_3 \neq 0^{|v_1|}$ we have $v_5 = 0^{|v_1|}$ and in no case this is an antipower.

To make it clearer, this is what the situation looks like, with $a = u0^{k_1}u\cdots0^{k_{j-2}}u$:

$$\omega = a0^{k_{j-1}}a0^{k_j}\cdots$$

The first case corresponds to $v_1 = a0^{k_{j-1}}*$, the second to $v_1 = a0\cdots0$ and $v_2 = 0\cdots0*$ and $v_3 = *^{-1}a0\cdots0$, the last case to $v_1 = a0\cdots0$ and $v_2 = 0\cdots0$ and $v_3 = 0\cdots0*$ and $v_4 = *^{-1}a0\cdots0$. □

4.3 Abelian powers and antipowers

In this part I will present the notion of abelian-antipowers. This is a new direction which merits to be explored.

4.3.1 Abelian powers in infinite words

Many of the classical definitions in combinatorics on words (e.g., period, power, factor complexity, etc.) have a counterpart in the abelian setting, though they may not enjoy the same properties. One important definition in abelian combinatorics on words is that of $Parikh\ vectors$.

Definition 4.3.1. The $Parikh\ vector$ $P(w)$ of a word w over a finite ordered alphabet $\mathbb{A} = \{a_1, a_2, \ldots, a_{|\mathbb{A}|}\}$ is the vector whose i-th component is equal to the number of occurrences of the letter a_i in w, $1 \leq i \leq |\mathbb{A}|$.

Example 4.3.2. The Parikh vector of $w = abbca$ over $\mathbb{A} = \{a, b, c\}$ is $P(w) = (2, 2, 1)$.

This notion is at the basis of the abelian combinatorics on words, where two words are considered equivalent if and only if they have the same Parikh vector:

Definition 4.3.3. For u ad v finite words, we say u is $abelian\ equivalent$ to v, and write $u \sim_{ab} v$, if $P(u) = P(v)$.

The first question asked (and answered) in abelian combinatorics on words were about the avoidability of abelian powers in long or infinite words (see, for example, [Dek79, Ker92]).

Definition 4.3.4. A word $w = u_1 u_2 \cdots u_n$ where for all i, $|u_1| = |u_i|$ is said to be an *abelian* n*-power* if for all i we have $u_i \sim_{ab} u_1$.

Example 4.3.5. $abba = ab \cdot ba$ is an abelian 2-power.

Remark 23. It is easy to see that a k-power is an abelian k-power; however as the previous example show, the converse is not true. In fact, there are significantly more abelian powers than powers, and the question of avoidability of abelian powers remained open for a long time before Evdokimov, Pleasants, Justin, Dekking and Keranen solved it [Evd68, Ple70, Jus72, Dek79, Ker92].

The classical notion of factor complexity (the function that counts the number of distinct factors of length n of a word, for every n) can be generalized by considering the *abelian factor complexity* $ab_w : \mathbb{N} \to \mathbb{N}$ (or *abelian complexity* for short), that is the function that counts the number of distinct Parikh vectors of factors of length n, for every n.

Remark 24. Let w be an infinite word, u and v be factors of w with $|u| = |v| = n$ and a a letter. Then, with $|u|_a$ the number of a in u:

$$ab_w(n) \geq ||u|_a - |v|_a|.$$

Proof. This comes from the fact that when you shift a factor of size n in w this factor's number of a will move by at most 1. If we suppose, without loss of generality, that u appears before v in w, we get the result by shifting from u to v, with every different number of as leading to a new Parikh vector. □

Morse and Hedlund [MH38] proved, as we discussed largely in the previous chapter, that an infinite word is aperiodic if and only if its factor complexity is unbounded. This characterization does not hold in the case of the abelian complexity, as there exist aperiodic words with bounded abelian complexity. For example, the Prouhet-Thue-Morse word has abelian complexity bounded by 3, yet it is aperiodic.

Still, knowing the abelian complexity of a word gives informations on the word: using Van der Waerden's theorem, Richomme et al. [RSZ11] proved that if an infinite word has bounded abelian complexity, then it contains abelian powers of every order:

Theorem 4.3.6. *[RSZ11] Let $\mathbb{A}$ be a finite alphabet $\{a_0, \cdots, a_n\}$, let M be an integer and let w be a word with abelian complexity lower than M. Then, for every positive integer k, w contains a k abelian power.*

Proof. Consider the colouring $c_0 : \mathbb{A} \to \{0, 1\}$ defined by

$$c_0 : a_i \to \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}.$$

Consider the finite colouring of the integers $s_0 : \mathbb{N} \to [[0, M]]$ defined by

$$s_0(n) \equiv \sum_{j=1}^{n} c_0(w_j) \pmod{M+1}.$$

By Van der Waerden theorem, for every integer $C_0(k)$, there exists an arithmetic progression $\{w_p, w_{p+d}, \cdots, w_{p+rd}\}$ of length at least $C_0(k)$ which is monochromatic. We then have, for every factor $u = w_{i+1} \cdots w_j$ of w,

$$|u|_{a_0} \equiv s(j) - s(i) \pmod{M+1}.$$

In particular, labelling $u_i = w_{p+(d-1)i+1} \cdots w_{p+di}$, we have that

$$\forall i \in [[1,r]] \qquad |u_i| = |u_1| \text{ and } |u_i|_{a_0} \equiv |u_1|_{a_0} \pmod{M+1}.$$

But since the abelian complexity is at most M this gives, thanks to Remark 24,

$$\forall i \in [[1,r]] \quad |u_i|_{a_0} = |u_1|_{a_0}.$$

Let us take $C_0(k) = W(C_1(k), 2)$ for an integer $C_1(k)$ that we will define.

Now we do the same with a new colouring $c_1 : \mathbb{A} \to \{0,1\}$ defined by

$$c_1 : a_i \to \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise.} \end{cases}.$$

Then, as previously, let s_1 be defined with respect to c_1 as s_0 for c_0 (hence it is now counting the number of a_1):

$$s_1(n) \equiv \sum_{j=1}^{n} c_1(w_j) \pmod{M+1}.$$

By Van den Waerden, we can find a monochromatic arithmetic progression of length $C_1(k)$ in $\{p, p+d, \cdots, p+rd\}$. Label this progression $\{p', p'+d', \cdots, p'+C_1(k)d'\}$. For the same reasons we get $C_1(k) - 1$ consecutive factors which now have same number of a_1 and a_0:

$$\begin{aligned} |w_{p'+id'+1} \cdots w_{p'+(i+1)d'}|_{a_1} &= s_1(p'+(i+1)d') - s_1(p'+id'+1) \\ &= s_1(p'+d') - s_1(p'+1) \end{aligned}$$

and

$$|w_{p'+id'+1} \cdots w_{p'+(i+1)d'}|_{a_0} = \frac{d'}{d}|u_1|_{a_0}.$$

Taking $C_1(k) = W(C_2(k), 2)$ and iterating this, we finally get, with $C_n(k) = k+1$, k consecutive factors of w with the same number of each letter, hence the same Parikh vector. This is an abelian k-power. □

However, this is not a characterization of words with bounded abelian complexity. Indeed, Štěpán Holub [Hol13] proved that all paperfolding words contain abelian powers of every order, and paperfolding words have unbounded abelian complexity (a property that by the way follows from the main result of section 4.3.4). The class of paperfolding words therefore constitutes an interesting example, as they are uniformly recurrent (every factor appears infinitely often and with bounded gaps) aperiodic words with linear factor complexity.

4.3.2 Abelian antipowers

Like for powers, it seems natural to extend the notion of an antipower to the abelian setting.

Definition 4.3.7. An *abelian antipower of order k*, or simply an *abelian k-antipower*, is a concatenation of k consecutive words of the same length having pairwise distinct Parikh vectors.

Example 4.3.8. The word

$$aabaaabbbabb = aab \cdot aaa \cdot bbb \cdot abb$$

is an abelian 4–antipower, while the word

$$abba = ab \cdot ba$$

is a 2-antipower but not an abelian 2-antipower.

Notice that an abelian k-antipower is a k-antipower but, as the previous example show, the converse does not necessarily hold (which is dual to the fact that a k–power is an abelian k–power but the converse does not necessarily hold).

So the situation is the following: for any integer k, there are less abelian k-antipowers, but more abelian k-powers.

It is possible then that an analogue of Theorem 4.1.11 may still hold in the case of abelian antipowers. Unfortunately, the proof of Theorem 4.1.11 does not generalize to the abelian setting: recall that we used the fact that a factor u was overlapping itself, which implied that it contained a long power. In the case of the abelian setting, we would only have $u_{M+p,i} \sim_{ab} u_{M+p,j}$, and that doesn't allow us to construct the previous factor u: in fact, it is not even sure that the suffixes of $u_{M+p,i}$ and $u_{M+p,j}$ would have the same Parikh vectors. Anyway, the examples of infinite words I considered all seemed to satisfy an abelian version of Theorem 4.1.11, so we tried to answer the following question:

Problem 1. Does every infinite word contain abelian powers of every order or abelian antipowers of every order?

Clearly, if a word has bounded abelian complexity, then it cannot contain abelian antipowers of every order. However, a word can avoid large abelian antipowers even if its abelian complexity is unbounded. Indeed, in [FRSZ18], an example is shown of an aperiodic recurrent word avoiding 6-antipowers (and therefore avoiding abelian 6-antipowers), and from the construction it can easily be verified that the abelian complexity of this word is unbounded.

A similar situation can be illustrated with the well-known Sierpiński word.

4.3.3 Sierpiński Word

This result was presented in a published article written jointly with Gabriele Fici and Manuel Silva [FPS19].

Recall that the Sierpiński word (also known as Cantor word) s is the fixed point starting with a of the substitution

$$\begin{aligned} \sigma : a &\to aba \\ b &\to bbb \end{aligned}$$

so that the word s begins as follows:

$$ababbbababbbbbbbbbababbbabab^{27}a\cdots$$

Therefore, s can be obtained as the limit, for $n \to \infty$, of the sequence of words $(s_n)_{n\geq 0}$ defined by: $s_0 = a$, $s_{n+1} = s_n b^{3^n} s_n$ for $n \geq 1$. Notice that for every n one has $|s_n| = 3^n$.

We will show that the abelian complexity of s is unbounded, but nonetheless the following property is verified:

Theorem 4.3.9. *The Sierpiński word s does not contain 11–antipowers, hence it does not contain abelian 11–antipowers.*

Remark 25. As mentionned in the introduction, this bound was latter improved to 10 by Riasat [Ria19].

Blanchet-Sadri, Fox and Rampersad [BSFR14] characterized the asymptotic behavior of the abelian complexity of a morphic word. In the following proposition, we give the precise bounds of the abelian complexity of the Sierpiński word.

Proposition 4.3.10. *The abelian complexity ab_s of the Sierpiński word verifies $ab_s(n) = \Theta(n^{\log_3 2})$.*

Proof. The Sierpiński word s is prefix normal with respect to the letter a (see [FL11, BFL+17] for the definition of prefix normal word), that is, for each length n, no factor of s of length n contains more occurrences of the letter a than the prefix of length n. Since s contains arbitrarily long blocks of bs, the number of distinct Parikh vectors of factors of s of a given length n is given by 1 plus the number of as in the prefix of length n. It is easy to see that the values of n for which the proportion of a's is maximal in a prefix of length n are of the form $n = 3^k$, while those for which the proportion of a's is minimal are of the form $n = 2 \cdot 3^k$, and in both cases the prefix of length n contains 2^k as. With a standard algebraic manipulation, this gives

$$\frac{n^{\log_3 2}}{2^{\log_3 2}} \leq ab_s(n) \leq n^{\log_3 2}.$$

□

Proof of Theorem 4.3.9. Suppose that s contains an 11–antipower $u = u_1u_2\cdots u_{11}$, of length $11m$. Let us then consider the first occurrence of u in s. Let n be the smallest integer such that u occurs in $s_{n+1}b^{3^{n+1}}$ but not in $s_nb^{3^n}$.

Let us first suppose that no u_i is equal to b^m for some i. Then $u_1\cdots u_{10}$ is a factor of $s_{n+1} = s_nb^{3^n}s_n$, so $10m < 3^{n+1}$ hence $m < 3^{n-1}$. Then, by minimality of n, there are only two possible cases: either u_1 starts before the block b^{3^n}, or u_1 starts in the block b^{3^n} and ends in s_n.

In the first case, by minimality of n, u ends after the block b^{3^n}, and since no u_i equals b^m, we get $2m > 3^n$, which is in contradiction with $m < 3^{n-1}$.

If u_1 starts in the block b^{3^n} and ends in s_n, $u_2\cdots u_{10}$ is a factor of $s_n = s_{n-1}b^{3^{n-1}}s_{n-1}$ and so $9m < 3^n$ hence $m < 3^{n-2}$. By minimality of n, u_{11} ends after the block $b^{3^{n-1}}$. Again, since no u_i equals b^m, we get $2m > 3^{n-1}$, which is in contradiction with $m < 3^{n-2}$.

Let us then suppose that $u_{11} = b^m$, so that $u_1\cdots u_9$ is a factor of s_{n+1}. The same reasoning as before holds, since $(9m < 3^{n+1}) \Rightarrow (m < 3^{n-1})$ and $(8m < 3^n) \Rightarrow (2m < 3^{n-1})$.

If $u_1 = b^m$, $u_2 \cdots u_{10}$ is a factor of s_n with no $u_i = b^m$ and we can again apply the same reasoning.

Finally, suppose that $u_i = b^m$ with $i \neq 1$ and $i \neq 11$. Hence, $u_1 \cdots u_{10}$ is a factor of $s_{n+1} = s_n b^{3^n} s_n$, and $10m < 3^{n+1}$. If u_1 starts before the block b^{3^n} (and u ends after by minimality of n), we get $3m > 3^n$ since otherwise u would contain two blocks b^m, and this contradicts $10m < 3^{n+1}$. If u_1 does not start before the block b^{3^n}, then by minimality of n it starts in this block, so $u_2 \cdots u_{10}$ is a factor of $s_n = s_{n-1} b^{3^{n-1}} s_{n-1}$ which ends after the block $b^{3^{n-1}}$, again by minimality of n. This shows that $9m < 3^n$, and at the same time $3m > 3^{n-1}$, which produces a contradiction. □

4.3.4 Paperfolding Words

These results were presented in a published article written jointly with Gabriele Fici and Manuel Silva [FPS19].

An infinite word can contain both abelian powers of every order and abelian antipowers of every order. This is the case, for example, of any word with full factor complexity. However, finding a class of uniformly recurrent words with linear factor complexity satisfying this property seems a more difficult task. Indeed, most of the well-known examples (Thue-Morse, Sturmian words, etc.) have bounded abelian complexity, hence they cannot contain abelian antipowers of every order — whereas, by the aforementioned result of Richomme et al. [RSZ11], they contain abelian powers of every order. Building upon the framework that Štěpán Holub developed to prove that all paperfolding words contain abelian powers of every order [Hol13], we prove in this section that all paperfolding words contain also abelian antipowers of every order.

In what follows, we recall the combinatorial framework for dealing with paperfolding words introduced in [Hol13], although we use the alphabet $\{0, 1\}$ instead of $\{1, -1\}$.

A paperfolding word is the sequence of ridges and valleys obtained by unfolding a sheet of paper which has been folded infinitely many times. At each step, one can fold the paper in two different ways, thus generating uncountably many sequences. It is known that all the paperfolding words are uniformly recurrent and have the same factor complexity $c(n)$, and that $c(n) = 4n$ for $n \geq 7$ [All92]. Madill and Rampersad [MR13] studied the abelian complexity of the regular paperfolding word and proved that it is a 2-regular sequence. The regular paperfolding word

$$\mathbf{p} = 00100110001101100010011100110110 \cdots$$

is the paperfolding word obtained by folding at each step in the same way. It can be defined as a Toeplitz word (see [CK97] for a definition of Toeplitz words) as follows: Consider the infinite periodic word $\gamma = (0?1?)^\omega$, defined over the alphabet $\{0, 1\} \cup \{?\}$. Then define $p_0 = \gamma$ and, for every $n > 0$, p_n as the word obtained from p_{n-1} by replacing the symbols ? with the letters of γ. So,

$$\begin{aligned} p_0 &= 0?1?0?1?0?1?0?1?0?1?0?1?0?1? \cdots , \\ p_1 &= 001?011?001?011?001?011?001? \cdots , \\ p_2 &= 0010011?0011011?0010011?0011 \cdots , \\ p_3 &= 001001100011011?001001110011 \cdots , \end{aligned}$$

etc. Thus, $\mathbf{p} = \lim_{n\to\infty} p_n$, and hence $\mathbf{p}$ does not contain occurrences of the symbol ?.

More generally, one can define a paperfolding word **f** by considering the two infinite periodic words $\gamma = (0?1?)^\omega$ and $\bar{\gamma} = (1?0?)^\omega$. Then, let $\boldsymbol{b} = b_0 b_1 \cdots$ be an infinite word over $\{-1, 1\}$, called *the sequence of instructions*. Define $(\gamma_n)_{n \geq 0}$ where, for every n, $\gamma_n = \gamma$ if $b_n = 1$ or $\gamma_n = \bar{\gamma}$ if $b_n = -1$. The paperfolding word **f** *associated with* $\boldsymbol{b}$ is the limit of the sequence of words f_n defined by $f_0 = \gamma_0$ and, for every $n > 0$, f_n is obtained from f_{n-1} by replacing the symbols ? with the letters of γ_n.

Recall that every positive integer i can be uniquely written as $i = 2^k(2j+1)$, where k is called the *order* of i (a.k.a. the 2-adic valuation of i), and $(2j+1)$ is called the *odd part* of i. One can verify that the previous definition of **f** is equivalent to the following: for every $i = 1, 2, \ldots$ define $w_i = (-1)^j b_k$, where $i = 2^k(2j+1)$. Then $f_i = 0$ if $w_i = 1$ and $f_i = 1$ if $w_i = -1$. This is equivalent to

$$f_i = 1 \quad \text{iff} \quad i \equiv 2^k(2 + b_k) \mod 2^{k+2}.$$

Remark 26. The regular paperfolding word corresponds to the sequence of instructions $\boldsymbol{b} = 1^\omega$.

Definition 4.3.11. Let **f** be a paperfolding word. An occurrence of a letter in **f** at position i is said to be *of order k* if the letter at position i is ? in f_{k-1} and different from ? in f_k. We consider the letters occurring in f_0 as of order 0.

Hence, in a paperfolding word **f** associated with the sequence $\boldsymbol{b} = b_0 b_1 \cdots$, the 1's of order 0 appear at positions $2 + b_0 + 4t$, $t \geq 0$, the 1's of order 1 appear at positions $2(2 + b_1 + 4t)$, $t \geq 0$, and, in general, the 1's of order k appear at positions $2^k(2 + b_k + 4t)$, $t \geq 0$.

Let $\mathbf{f} = f_1 f_2 \cdots$ be a paperfolding word associated with the sequence $\boldsymbol{b} = b_0 b_1 \cdots$. A factor of **f** of length n starting at position $\ell + 1$, denoted by $\mathbf{f}[\ell+1, \ldots, \ell+n]$, contains a number of 1's that is given by the sum, for all $k \geq 0$, of the 1's of order k in the interval $[\ell+1, \ell+n]$. For each k, since the 1's of order k are at distance 2^{k+2} one from another, the number of occurrences of 1's of order k in $\mathbf{f}[\ell+1, \ldots, \ell+n]$ is given by

$$\left\lfloor \frac{n - \ell}{2^{k+2}} \right\rfloor + \epsilon_{k,b_k}(\ell, n),$$

where $\epsilon_{k,b_k}(\ell, n) \in \{0, 1\}$ depends on the sequence $\boldsymbol{b}$ (in fact, b_k determines the positions of the occurrences of the 1's of order k in **f**). We set

$$\Delta(\ell, n) = \sum_{k \geq 0} \epsilon_{k,b_k}(\ell, n)$$

the number of "extra" 1's in $\mathbf{f}[\ell+1, \ldots, \ell+n]$.

For example, in the prefix $\mathbf{p}[1, 14]$ of length 14 of the regular paperfolding word, we know that there are at least $3 = \lfloor \frac{14}{4} \rfloor$ 1's of order 0, $1 = \lfloor \frac{14}{8} \rfloor$ of order 1 and $0 = \lfloor \frac{14}{16} \rfloor$ of order 2. In the interval $[1, 14]$ there are three 1's of order 0 (at positions 3, 7 and 11), two 1's of order 1 (at positions 6 and 14), and one 1 of order 2 (at position 12), so we have in $\mathbf{p}[1, 14]$ no extra 1 of order 0, i.e., $\epsilon_{0,1}(0, 14) = 0$, one extra 1 of order 1, i.e., $\epsilon_{1,1}(0, 14) = 1$ and one extra 1 of order 2, i.e., $\epsilon_{2,1}(0, 14) = 1$, so that $\Delta(0, 14) = 2$.

We set

$$\mathcal{E}_{k,b_k}(\ell, d, m) = (\epsilon_{k,b_k}(\ell, \ell + d), \ldots, \epsilon_{k,b_k}(\ell + (m-1)d, \ell + md))$$

and

$$\Delta(\ell, d, m) = \sum_{k\geq 0} \mathcal{E}_{k,b_k}(\ell, d, m) = (\Delta(\ell, \ell+d), \dots, \Delta(\ell+(m-1)d, \ell+md))\,.$$

The factor of $\mathbf{f}$ of length dm starting at position $\ell+1$ is an abelian m-power if and only if the components of the vector $\Delta(\ell, d, m)$ are all equal, while it is an abelian m-antipower if and only if the components of the vector $\Delta(\ell, d, m)$ are pairwise distinct.

The next result (Lemma 4 of [Hol13]) will be the fundamental ingredient for the construction of abelian antipowers in paperfolding words.

Lemma 4.3.12 (Additivity Lemma)**.** *Let $\ell, \ell' \geq 0$ and $m, d, d' \geq 1$ be integers with ℓ' and d' both even. Let r be such that $2^r > \ell + md$, and for each $k \geq 0$ the following implication holds: if $\mathcal{E}_{k,1}(\ell', d', m) \neq \mathcal{E}_{k,-1}(\ell', d', m)$ then $b_k = b_{k+r}$.*

Then

$$\Delta(\ell, d, m) + \Delta(\ell', d', m) = \Delta(\ell + 2^r\ell', d + 2^r d', m).$$

Using the Additivity Lemma, Holub [Hol13] proved that all paperfolding words contain abelian powers of every order. We will use the Additivity Lemma to prove that all paperfolding words contain abelian antipowers of every order. We start with the regular paperfolding word, then we extend the argument to all paperfolding words.

4.3.4.1 Regular paperfolding word

Let

$$\begin{array}{cccc} \Phi & : & \{0,1\}^2 & \rightarrow & \{x, y, z\} \\ & & 00 & \mapsto & x \\ & & 01 & \mapsto & y \\ & & 10 & \mapsto & y \\ & & 11 & \mapsto & z \end{array}$$

be the morphism that identifies words of length 2 over the alphabet $\{0, 1\}$ that are abelian equivalent. We have the following lemma:

Lemma 4.3.13. *Let $n \geq 3$ be an integer. Let $p = \boldsymbol{p}[\ell+1, \dots, \ell+2^n] = u_1v_1\cdots u_{2^{n-1}}v_{2^{n-1}}$ be a factor of $\boldsymbol{p}$ of length 2^n. Then, no $q < 2^{n-1}$ exists such that*

$$\Phi(p) = \Phi(u_1v_1)\cdots\Phi(u_{2^{n-1}}v_{2^{n-1}}) = \Phi(u_{q+1}v_{q+1})\cdots\Phi(u_{2^{n-1}}v_{2^{n-1}})\Phi(u_1v_1)\cdots\Phi(u_qv_q). \tag{4.4}$$

Proof. First, notice that if q' is the smallest solution of (4.4), then $q'|2^{n-1}$. Indeed, writing $w_i = \Phi(u_iv_i)$, we have

$$\begin{aligned} w_1\cdots w_{2^{n-1}} &= w_1\cdots w_{q'}w_{q'+1}\cdots w_{2^{n-1}} \\ &= w_{q'+1}\cdots w_{2^{n-1}}w_1\cdots w_{q'}, \end{aligned}$$

and since two words commute if and only if they are powers of the same word, there exists a word z and positive integers s and t such that

$$w_1\cdots w_{q'} = z^s \text{ and } w_{q'+1}\cdots w_{2^{n-1}} = z^t.$$

This gives $|z| \cdot (s+t) = 2^{n-1}$ and $|z| \cdot s = q'$. By the minimality of q', we have that $s = 1$ and so $|z| = q'$ divides 2^{n-1}. Thus, $q' = 2^j$ for some integer $j < n$.

By the Toeplitz construction of $\mathbf{p}$, we immediately have that

$$u_1 v_1 \cdots u_{2^{n-1}} v_{2^{n-1}} = a v_1 \overline{a} v_2 a v_3 \overline{a} \cdots \overline{a} v_{2^{n-1}}$$

or

$$u_1 v_1 \cdots u_{2^{n-1}} v_{2^{n-1}} = u_1 a u_2 \overline{a} u_3 a u_4 \overline{a} \cdots u_{2^{n-1}} \overline{a}$$

with $a \in \{0,1\}$ and $\overline{a} = 1 - a$.

Suppose $q' \neq 1$ and $q' \neq 2^{n-1}$. Since q' is even, we have that $\Phi(u_i v_i) = \Phi(u_{i+q'} v_{i+q'})$ implies $u_i v_i = u_{i+q'} v_{i+q'}$. But this cannot be the case, since two consecutive letters of order j occur in $\mathbf{p}$ at distance 2^{j+1}. Since $j \leq n-2$, we have $2^{j+2} \leq 2^n$, so the factor p contains at least two consecutive letters of order j. Suppose that the first of such letters is u_i; then $u_{i+q'}$ is at distance $2q' = 2^{j+1}$, so $u_{i+q'} \neq u_i$, against the hypothesis that q' is a solution of (4.4).

Thus, we must have $q' = 1$ or $q' = 2^{n-1}$. Since $n \geq 3$, $\mathbf{p}[\ell+1, \ldots, \ell+2^n]$ contains two consecutive letters of order 1. Let us first suppose that v_i is a 1 of order 1, u_i is a 1 of order 0 and v_{i+2} is a 0 of order 1. Then, $\Phi(u_i v_i) = \Phi(11) \neq \Phi(10) = \Phi(u_{i+2} v_{i+2})$. The other cases would give $10 u_{i+1} v_{i+1} 11$ with v_i a 0 of order 1 and v_{i+2} a 1 of order 1, $00 u_{i+1} v_{i+1} 01$ and $00 u_{i+1} v_{i+1} 01$ respectively in the case u_i is a 0 of order 0. Similary, we get $10 u_{i+1} v_{i+1} 00$ and $00 u_{i+1} v_{i+1} 10$ if u_i is a 1 of order 1 and u_{i+2} a 0 of order 1 or vice versa, and v_i a 0 of order 0. The cases with v_i a 1 of order 0 are symetric. Every case leads to $\Phi(u_i v_i) \neq \Phi(u_{i+2} v_{i+2})$. This implies $q' \neq 1$ and so $q' = 2^{n-1}$. By minimality of q', the only solution of (4.4) is $q = 2^{n-1}$. □

Theorem 4.3.14. *The regular paperfolding word contains abelian m-antipowers for every $m \geq 2$.*

Proof. The proof is mainly based on the Additivity Lemma. Let $m \geq 2$ be fixed. To prove the result it is sufficient to find a vector $\Delta(s, d, m)$ having pairwise distinct components. Let k be an integer such that $2^k \geq m$. Consider the first factor of length $2^{k+2} - 1$ containing a 1 of order k in the middle; our factor is then of the form

$$w1w'$$

with $|w| = |w'| = 2^{k+1} - 1$. Since for every positive integers i, k', s, we have

$$p_i \text{ of order } k' \Rightarrow p_{i+2^{k'+s}} \text{ of order } k'$$

and

$$p_i \text{ of order } k' \Rightarrow p_{i+2^{k'+2}} = p_i \neq p_{i+2^{k'+1}}$$

we get:

$$p_i \text{ of order } k' \Rightarrow p_{i+2^{k'+2+s}} = p_i \neq p_{i+2^{k'+1}} \tag{4.5}$$

then, up to applying a translation, we can suppose $w = w'$. In fact, since $|w1| = 2^{k+1}$, the equality is true for every letter of order smaller than k by (4.5). Now, take the smallest order $r > k$ of a letter 0 in w or w'. It is the only letter of this order in our factor since two letters of order r are distant of $2^{r+1} > |w1w'|$. If we consider the factor translated by 2^{r+1}, by (4.5) the letters of order smaller than r are the same and the letter we considered becomes a 1. Since the length of $w1w'$ is $2^{k+2} - 1$ and the distance between two letters of order higher than k is at least 2^{k+1}, the factor $w1w'$ contains exactly two letters of order higher than k. Hence, in at most 2 steps we get $w1w$ with every letter of order greater than k being a 1. Writing $\ell+1$ the starting position of an occurrence in **p** of the factor $w1w$, we set $\ell' = \ell$ if ℓ is even or $\ell' = \ell+1$ otherwise. Consider the vectors

$$\Delta(\ell', 2, 2^k), \Delta(\ell'+2, 2, 2^k), \Delta(\ell'+4, 2, 2^k), \Delta(\ell'+6, 2, 2^k), \ldots, \Delta(\ell'+2^{k+1}-2, 2, 2^k).$$

We claim that these vectors are pairwise distinct. By contradiction, if $\Delta(\ell'+2p, 2, 2^k) = \Delta(\ell'+2q, 2, 2^k)$ for some p, q with $p \leq q$, then we have that

$$\Phi(p_{\ell'+2p+1} \cdots p_{\ell'+2p+2^{k+1}}) = \Phi(p_{\ell'+2q+1} \cdots p_{\ell'+2q+2^{k+1}}). \tag{4.6}$$

Since the factor we are considering is $w1w$, we have

$$p_{\ell'+2p+1} \cdots p_{\ell'+2q} = p_{\ell'+2p+1+2^{k+1}} \cdots p_{\ell'+2q+2^{k+1}}$$

and so

$$\Phi(p_{\ell'+2q+1} \cdots p_{\ell'+2q+2^{k+1}}) = \Phi(p_{\ell'+2q+1} \cdots p_{\ell'+2p+2^{k+1}} p_{\ell'+2p+1} \cdots p_{\ell'+2q})$$

but this and (4.6) contradicts Lemma 4.3.13.

Finally, as the vectors are different, we use the Additivity Lemma to obtain a vector whose components are pairwise distinct: applying n times the Additivity Lemma on $\Delta(\ell'+2p, 2, 2^k)$ one can obtain $n\Delta(\ell'+2p, 2, 2^k)$. It then suffices to take a sequence of integers $\alpha_0, \ldots, \alpha_{2^k-1}$ increasing enough to have

$$\Sigma_{i=0}^{2^k-1} \alpha_i \Delta(s'+2i, 2, 2^k),$$

a vector whose components are pairwise distinct. Indeed, labelling a_j the j-th component of this vector and $x_{i,j}$ the j-th component of $\Delta(s'+2i, 2, 2^k)$, we have

$$a_j = a_{j'} \Leftrightarrow \Sigma_{i=0}^{2^k-1} \alpha_i x_{i,j} = \Sigma_{i=0}^{2^k-1} \alpha_i x_{i,j'} \Leftrightarrow \Sigma_{i=0}^{2^k-1} \alpha_i (x_{i,j} - x_{i,j'}) = 0.$$

By "increasing enough", we precisely mean $\alpha_r > \Sigma_{i=0}^{r-1} \alpha_i \sup_{0 \leq q, q' \leq 2^k-1} (x_{i,q} - x_{i,q'})$, so that by decreasing induction we have that for every i, with $0 \leq i \leq 2^k - 1$, one has $x_{i,j} = x_{i,j'}$. In particular, this gives $\Delta(\ell'+2j, 2, 2^k) = \Delta(\ell'+2j', 2, 2^k)$, which implies $j = j'$. Hence, all the components are pairwise distinct and the proof is complete. □

4.3.4.2 All paperfolding words

To generalize the result above to all paperfolding words, one has to take care of the condition $b_i = b_{i+r}$ in the Additivity Lemma.

Lemma 4.3.13 can be modified so that the translation is not by 2 but by 2^u, for any $u > 1$. Let

$$\begin{array}{cccc} \phi & : & \{0,1\}^{2^u} & \to & \mathbb{N} \\ & & a_1 \cdots a_{2^u} & \mapsto & |\{i \mid a_i = 1\}| \end{array}$$

be the morphism that identifies words of length 2^u over $\{0,1\}$ that are abelian equivalent. Then we have the following lemma, analogous to Lemma 4.3.13:

Lemma 4.3.15. *Let $n \geq u+3$ be an integer and let $\boldsymbol{f}$ be a paperfolding word. Every factor $f = \boldsymbol{f}[\ell+1, \ell+2^{n+u-1}] = a_{1,1}a_{1,2}\cdots a_{2^{n-1},2^u-1}a_{2^{n-1},2^u}$ of $\boldsymbol{f}$ of length 2^{n+u-1} satisfies the following property: If q is such that*

$$\begin{aligned}&\phi(f) = \phi(a_{1,1}\cdots a_{1,2^u})\cdots\phi(a_{2^{n-1},1}\cdots a_{2^{n-1},2^u}) = \\ &\phi(a_{q+1,1}\cdots a_{q+1,2^u})\cdots\phi(a_{2^{n-1},1}\cdots a_{2^{n-1},2^u})\phi(a_{1,1}\cdots a_{1,2^u})\cdots\phi(a_{q,1}\cdots a_{q,2^u}),\end{aligned}$$

then $q = 2^{n-1}$.

Proof. The proof of Lemma 4.3.13 mainly applies here; we only need to change the part where we use the Toeplitz construction to justify $j = n-1$. Here, in each 2^u-tuple one can find one letter of order $u-1$ and one letter of higher order. Using (4.5), we then see that $\phi(a_{i,1}\cdots a_{i,2^u})$ is totally determined by the letter of order $u-1$ and the letter of higher order in $a_{i,1}\cdots a_{i,2^u}$. Applying again (4.5) to the letter of order $u-1$, we can apply exactly the same reasoning as in the proof of Lemma 4.3.13 (in a sense, our new ϕ is the previous one modulo the letters of order smaller than $u-1$). □

Now, we can prove the main theorem:

Theorem 4.3.16. *Every paperfolding word $\boldsymbol{f}$ contains abelian m-antipowers for every $m \geq 2$.*

Proof. Let k be an integer such that $2^k \geq m$. As before, we will prove that **f** contains abelian 2^k-antipowers, hence it will contain abelian m-antipowers. Since the alphabet $\{0,1\}$ is finite, there must exist a factor $b_{u-1}\cdots b_{u+k+4}$ of $\boldsymbol{b}$ that occurs infinitely often. As before, let us start with the first block of length $2^{u+k+2}-1$ containing a 1 of order $u+k$ in the middle; our block is then

$$w1w'$$

with $|w| = |w'| = 2^{u+k+1}-1$. As before, in at most two steps, we can have $w = w'$, and the maximum order of a letter appearing in this factor is $u+k+4$. Again, writing ℓ the starting position of an occurrence of this factor, we set $\ell' = \ell$ if ℓ is even or $\ell' = \ell+1$ otherwise. Consider the vectors

$$\Delta(\ell', 2^u, 2^k), \Delta(\ell'+2^u, 2^u, 2^k), \Delta(\ell'+2^{u+1}, 2^u, 2^k), \ldots, \Delta(\ell'+2^{u+k+1}-2^u, 2^u, 2^k).$$

Here again, these vectors are pairwise distinct: if $\Delta(\ell'+2^u p, 2^u, 2^k) = \Delta(\ell'+2^u q, 2^u, 2^k)$, we have that

$$\phi(p_{\ell'+2^u p+1}\cdots p_{\ell'+2^u(p+2^k)}) = \phi(p_{\ell'+2^u q+1}\cdots p_{\ell'+2^u(q+2^k)})$$

and this contradicts Lemma 4.3.15 because, here again, $w = w'$ and so

$$p_{\ell'+2^u p+1}\cdots p_{\ell'+2^u q} = p_{\ell'+2^u(p+2^k)+1}\cdots p_{\ell'+2^u(q+2^k)}.$$

Moreover, $\varepsilon_{i,-1}(\ell'+2^u p, 2^u, 2^k) \neq \varepsilon_{i,1}(\ell'+2^u p, 2^u, 2^k) \Rightarrow u-1 \leq i \leq u+k+4$, using (4.5) and the fact that no letter of order higher than $u+k+4$ appears in the factor $w1w$. So, choosing r such that $2^r > \ell'+2^{u+k+1}-2^u+2^{u+k}$ and $b_{u-1}\cdots b_{u+k+4} = b_{r+u-1}\cdots b_{r+u+k+4}$, we can apply the Additivity Lemma and, as for the regular paperfolding word, construct an abelian 2^k-antipower that occurs as a factor in **f**. □

Remark 27. From Theorem 4.3.16 it follows immediately that every paperfolding word has unbounded abelian complexity.

Remark 28. In [CRSZ11] Cassaigne et al. prove that a word with bounded abelian complexity contains abelian powers of any order. In fact, to apply the proof to a word w they used one only needs the following:

$$\exists N, \forall m, \exists v \in \text{Fact}(w), |v| = m \text{ and } c^{ab}(v) \leq N.$$

Since every paperfolding word is uniformly reccurrent, using the above remark and the fact paper folding words contain abelian powers of any order we see that this condition is not necessary.

4.3.5 Zimin word

This section contains new results.

The $Zimin\ word$ is a famous word (also kown as sesquipower) defined on an infinite alphabet, it is another example where Problem 1 is verified again. For more information about this word, one can consult [Lot02].

Definition 4.3.17. The $Zimin\ word$ **z** can be defined by a Toeplitz process, like paper-folding words. It is the limit of the following process:

$$\begin{aligned}
z_0 &= ????????????????????????????\cdots, \\
z_1 &= 1?1?1?1?1?1?1?1?1?1?1?1?1?1?\cdots, \\
z_2 &= 121?121?121?121?121?121?121?\cdots, \\
z_3 &= 1213121?1213121?1213121?1213\cdots, \\
z_4 &= 121312141213121?121312141213\cdots,
\end{aligned}$$

At n-step, replace every other ? by n (starting with $z_0 = ?^\omega$). One then have $\mathbf{z} = \lim_{n\to+\infty} z_n$.

Remark 29. Alternatively, z can be defined as the fixed point starting with 0 of the morphism $\phi : i \to 1 \cdot (i + 1)$.

We will show that this word avoids abelian squares but contains abelian antipowers of any order.

4.3.5.1 Zimin word avoids abelian squares

Theorem 4.3.18. *The Zimin word* **z** *does not contain abelian squares.*

Proof. Suppose u_1u_2 is an abelian square with $m = |u_1| = |u_2|$ minimal. Its is easy to see that $m \neq 1$. But if $m > 1$, we have $|u_1u_2|_1 = m$ since every other letter is a 1. But we have $|u_1|_1 = |u_2|_1$ hence m is even.

Hence, up to shifting u_1u_2 to the right, we can suppose u_1 starts with a 1. Now, we can look at $\phi^{-1}(u_1u_2)$. It is easy to check that $\phi^{-1}(u_1u_2)$ is also an abelian square of length strictly smaller than m which is contradictory with our minimality hypothesis. □

4.3.5.2 Zimin word contains abelian antipowers of every order

Theorem 4.3.19. *The Zimin word* **z** *contains abelian* k*-antipowers for every* k *in* $\mathbb{N}$.

Proof. By induction. The result is clearly true for $k = 1, 2, 3$. Suppose the result is true for some $k \in \mathbb{N}$; let us show that this implies **z** contains abelian $(k+1)$-antipowers. Let us write $u_1 \cdots u_k$ with $|u_1| = m$ an abelian k-antipower. First of all, if $P(u_{k+1}) \notin \{P(u_i) | 1 \leq i \leq k\}$ we have an abelian $(k+1)$-antipower. If not, let $j \in [[1, k]]$ be such that $P(u_{k+1}) = P(u_j)$. Let us write n the greatest letter appearing in $u_1 \cdots u_k$. We then know that $\forall i, i' \in [[1, k]], \exists p \leq n$ such that $|u_i|_p \neq |u_{i'}|_p$. Moreover, if $i \neq j$, there exists a $p \leq n$ such that $|u_i|_p \neq |u_{k+1}|_p$.

Now, consider, for a positive integer r that will defined latter, $\phi^r(u_1 \cdots u_{k+1})$, which is a factor of **z**. We have $|\phi^r(u_i)| = 2^r m$ and $\phi(u_i)$ starts with $2^r - 1$ letters of order smaller or equal to $r + 1$(*). Moreover,

$$\forall p \in [[1, n]],\ |\phi^r(u_i)|_{r+p} = |u_i|_p. \tag{4.7}$$

It is easy to see that $\phi^r(u_1 \cdots u_k)$ is an abelian k-antipower. Indeed, let i, i', p be such that $|u_i|_p \neq |u_{i'}|_p$. Then

$$|\phi^r(u_i)|_{r+p} \neq |\phi^r(u_{i'})|_{r+p}. \tag{4.8}$$

The idea is then to take blocks $u'_1 \cdots u'_{k+1}$ just a bit bigger than $2^r m$ so that we keep the same amount of letters of order $r+1, \cdots, r+p$ but we insure that $P(u'_j) \neq P(u'_{k+1})$. Let us then suppose that

$$2^r - 1 > k + 2^k \text{ and } r > k. \tag{4.9}$$

For example we can take $r = k + 2$. Then consider $u'_1 \cdots u'_{k+1}$ the factor starting $2^k - j$ letter after the beginning of $\phi^r(u_1 \cdots u_{k+1})$, with $|u'_i| = 2^r m + 1$ (so the first letter of u'_j is k). It is then easy to check that

$$\forall p \in [[1, k+1]], |\phi^r(u_i)|_{r+p} = |u'_i|_{r+p}. \tag{4.10}$$

thanks to (4.9) and (*). I then claim that $u'_1 \cdots u'_{k+1}$ is an abelian $(k+1)$-antipower. Indeed, since (4.8) and (4.10), $u'_1 \cdots u'_k$ is an abelian k-antipower, and $\forall i \neq j, P(u'_{k+1}) \neq P(u'_i)$. It remains to prove $P(u'_{k+1}) \neq P(u'_j)$. But

$$|u'_{k+1}|_k = |u'_j|_k - 1 = 2^{r-k} m.$$

Indeed, since the letter k has periodicity 2^k, the $2^r m$ last letters of these two factors contain exactly $2^{r-k} m$ occurences of the letter k. The first letter of u'_j is then a k, and since $2^k \nmid (2^r m + 1)(k + 1 - j)$ the first letter of u'_{k+1} is not a k. So $u'_1 \cdots u'_{k+1}$ is an abelian $(k+1)$-antipower, and **z** contains abelian k-antipowers for every positive integer k. □

4.3.6 Another possible generalizations of antipowers

Finally, I would like to present some idea we investigated in our efforts to generalize Theorem 4.1.11. Instead of looking at the abelian setting, which seems to be difficult to solve, we looked at a new version of powers, only based on the set of Parikh vectors of a word. The idea is to loosen the condition on the proximity of the factors giving those Parikh vectors, as the abelian setting loosens the condition on the position of the letters in a word for two words to be equivalent.

Definition 4.3.20. Let w be a word, k a natural integer, $u_1, u_2, \cdots, u_k$ be finite factors of w, P_i their Parikh vectors. We say that $(u_1, u_2, \cdots, u_k)$ is a *total-abelian k-power* if $P_i = iP_1$ for every i. We say that $(u_1, u_2, \cdots, u_k)$ is a *total-abelian k-antipower* if:

- $P_i \neq P_j$ for every (i, j);
- $|u_i| = |u_1|$ for all i;
- there exist $v_2, \cdots, v_k$ factors of w such that: $\forall i \; P(v_i) = \sum\limits_{m=1}^{i} P_m$.

Remark 30. This condition is less strong than the one of abelian powers or antipowers. In fact, an abelian (anti)power is a total-abelian (anti)power. This allows us to say that if Theorem 4.1.11 is true in the abelian setting, it has to be true in this total-abelian setting, or in other words, Problem 1 should be easier to answer in this total-abelian setting.

In this context, we were able to find an answer to Problem 1 for the binary alphabet:

Theorem 4.3.21. *On a binary alphabet $\mathbb{A} = \{0, 1\}$, every infinite word w contains total-abelian k-powers for every k in $\mathbb{N}$. Moreover, if the abelian complexity of w is not bounded, then w contains total-abelian k-antipowers for every k in $\mathbb{N}$.*

Proof. There are two different possibilities: firstly, if the abelian complexity of w is bounded, then w contains k-abelian powers for every k, hence total-abelian k-powers. The remaining case is then: w has unbounded abelian complexity. We will then use the following lemma:

Lemma 4.3.22. *Let w be a word whose abelian complexity is not bounded. Then, for every n in $\mathbb{N}$, w contains two adjacent factors of same size whose number of 0 differs of at least n.*

Proof. Since the abelian complexity is unbounded, it is always possible, for every n, to find two non overlapping factors of same size whose number of 0 differs from at least $2n + 1$. Labeling them u and v, let us write $w = auz_1z_2vb$ with $0 \leq |z_1| - |z_2| \leq 1$. If $|z_1| - |z_2| = 1$, translate the beginning and the end point of v of one. We then have $||u|_1 - |v|_1| \geq 2n$. Then, consider the factor uz_1 and the factor z_2v. Either $||uz_1|_1 - |z_2v|_1| \geq n$ and the announced result is proved for n, or $||uz_1|_1 - |z_2v|_1| < n$ which implies $|z_1|_1 - |z_2|_1 > n$ and the result is still proved for n. □

Let then n be a fixed natural integer, and let us prove that w contains total-abelian n-powers and antipowers. Applying the lemma, it is possible to find u and v two consecutive factors of same size with $||u|_1 - |v|_1| \geq (2n + 1)n!$. Let us denote $p := \lfloor \frac{|u|}{n!} \rfloor$. Taking if necessary u and v shorter we can suppose

$$|u| = pn! \text{ and } ||u|_1 - |v|_1| > 2nn! \tag{4.11}$$

Without loss of generality we can assume $|u|_1 < |v|_1$. For $r \in \mathbb{N}$ we can define $m_r = \frac{r|uv|_1}{|uv|}$ the mean number of 1 in a factor of length r of uv. Since $|v|_1 - |u|_1 > 2nn!$, there exist at least one pair (u', v') of factor of length p respectively of u and v such that

$$m_p - |u'|_1 \geq n \text{ and } |v'|_1 - m_p \geq n.$$

Indeed, on one hand one has

$$|u|_1 + |v|_1 = |uv|_1 = 2n!m_p. \tag{4.12}$$

On the other hand, cutting u and v in n blocks of size p, one gets, if no such u' exists,

$$|u|_1 > n!(m_p - n)$$

and then

$$|u|_1 + |v|_1 > 2n!(m_p - n) + 2nn! = 2n!m_p$$

by (4.11) which contradicts (4.12). Similarily, if no such v' exists one gets

$$|v|_1 < n!(m_p + n)$$

and then

$$|u|_1 + |v|_1 < 2n!(m_p + n) - 2nn! = 2n!m_p.$$

With T the shift operator, for any factor x of w,

$$||x|_1 - |T(x)|_1| \leq 1, \tag{4.13}$$

so there exist $z_1, \cdots, z_n$ factors of uv of length p such that $m_p - |z_i|_1 = i$. Those n factors provide a total-abelian n-antipower: their Parikh vectors are different, so let us prove that it is possible to find factors $z'_1, z'_2, \cdots, z'_n$ with $|z'_i| = ip$ and $im_p - |z'_i|_1 = \frac{i(i+1)}{2}$. Using (4.13), it is enough to show that for any i it is possible to find a factor u'_i with $|u'_i| = ip$ and $im_p - |u'_i|_1 \geq in$ and a factor v'_i with $|v'_i| = ip$ and $im_p - |v'_i|_1 \leq 0$.

For v'_i, it suffices to cut uv in $\frac{2n!}{i}$ blocks of length ip, since the total number of 1 is $2n!m_p$ at least one of the blocks contains at least im_p times 1. For u'_i, we can apply the same reasonning than previously: if no such u'_i exists then by cutting u into blocks of length ip we get

$$|u|_1 > \frac{n!}{i}(im_p - in) = n!(m_p - n)$$

and then

$$|uv|_1 > 2n!(m_p - n) + 2nn! = 2n!m_p.$$

Moreover, using the same u'_i and v'_i, and (4.13), it is easy to construct vectors $z''_1, \cdots, z''_n$ with $|z''_i| = ip$ and $|z''_i|_1 = im_p$, and this is, by definition, a total-abelian n-power. □

4.3.7 Conclusion

While we have not been able to answer Problem 1, for every example we studied, we found the analogy of Theorem 4.1.11 to be verified in the abelian setting. I am very curious about this question, and I will try to use the coming months to keep on trying to solve it.

Bibliography

[ABCD03] Jean-Paul Allouche, Michael Baake, Julien Cassaigne, and David Damanik. Palindrome complexity. *Theoretical Computer Science*, 292(1):9–31, 2003.

[AC95] Alberto Apostolico and Maxime Crochemore. Fast parallel Lyndon factorization with applications. *Mathematical systems theory*, 28:89–108, 1995.

[All92] Jean-Paul Allouche. The number of factors in a paperfolding sequence. *Bulletin of the Australian Mathematical Society*, 46:23–32, 1992.

[AR91] Pierre Arnoux and Gérard Rauzy. Représentation géométrique de suites de complexité $2n + 1$. *Bulletin de la Société Mathématique de France*, 119(2):199–215, 1991.

[AS03] Jean-Paul Allouche and Jeffrey Shallit. *Automatic sequences: Theory, applications, generalizations*. Cambridge University Press, 2003.

[BD20] Aaron Berger and Colin Defant. On anti-powers in aperiodic recurrent words. *Advances in Applied Mathematics*, 121:102104, 2020.

[BFL$^+$17] Peter Burcsi, Gabriele Fici, Zsuzsanna Lipták, Frank Ruskey, and Joe Sawada. On prefix normal words and prefix normal forms. *Theoretical Computer Science*, 659:1–13, 2017.

[BFP18] Golnaz Badkobeh, Gabriele Fici, and Simon J. Puglisi. Algorithms for anti-powers in strings. *Information Processing Letters*, 137:57 – 60, 2018.

[BHPZ13] Michelangelo Bucci, Neil Hindman, Svetlana Puzynina, and Luca Q. Zamboni. On additive properties of sets defined by the Thue-Morse word. *J. Combin. Theory Ser. A*, 120(6):1235–1245, 2013.

[Bir37] Garrett Birkhoff. Representability of lie algebras and lie groups by matrices. *Annals of Mathematics*, 38(2):526–532, 1937.

[BPS06] Lubomira Balkova, Edita Pelantova, and Wolfgang Steiner. Sequences with constant number of return words. *Monatshefte für Mathematik*, 155, 2006.

[BPZ15] Michelangelo Bucci, Svetlana Puzynina, and Luca Q. Zamboni. Central sets generated by uniformly recurrent words. *Ergodic Theory and Dynamical Systems*, 35(3):714–736, 2015.

[BSFR14] Francine Blanchet-Sadri, Nathan Fox, and Narad Rampersad. On the asymptotic abelian complexity of morphic words. *Advances in Applied Mathematics*, 61:46 – 84, 2014.

[Bur18] Amanda Burcroff. (k,)-anti-powers and other patterns in words. *Electronic Journal of Combinatorics*, 25:P4.41, 2018.

[BW20] Amanda Burcroff and Eric Winsor. Generalized lyndon factorizations of infinite words. *Theoretical Computer Science*, 809:30 – 38, 2020.

[BZ13] Marcy Barge and Luca Q. Zamboni. Central sets and substitutive dynamical systems. *Advances in Mathematics*, 248:308–323, 2013.

[CFL58] Kuo-Tsai Chen, Ralph H. Fox, and Roger C. Lyndon. Free Differential Calculus, IV. The Quotient Groups of the Lower Central Series. *Annals of Mathematics*, 68(1):81–95, 1958.

[CFSZ17] Julien Cassaigne, Gabriele Fici, Marinella Sciortino, and Luca Q. Zamboni. Cyclic complexity of words. *Journal of Combinatorial Theory, Series A*, 145:36 – 56, 2017.

[CH73] Ethan Coven and Gustav Hedlund. Sequences with minimal block growth. *Mathematical Systems Theory*, 7:138–153, 1973.

[CK97] Julien Cassaigne and Juhani Karhumäki. Toeplitz words, generalized periodicity and periodically iterated morphisms. *European Journal of Combinatorics*, 18(5):497–510, 1997.

[CR20] Maxime Crochemore and Luís M. S. Russo. Cartesian and Lyndon trees. *Theoretical Computer Science*, 806:1–9, 2020.

[CRSZ11] Julien Cassaigne, Gwenaël Richomme, Kalle Saari, and Luca Q. Zamboni. Avoiding Abelian powers in binary words with bounded Abelian complexity. *International Journal of Foundations of Computer Science*, 22(4):905–920, 2011.

[Def17] Colin Defant. Anti-power prefixes of the thue-morse word. *Electronic Journal of Combinatorics*, 24, 2017.

[Dek79] Michel Dekking. Strongly nonrepetitive sequences and progression-free sets. *Journal of Combinatorial Theory, Series A*, 27(2):181–185, 1979.

[DHS99] Fabien Durand, Bernard Host, and Christian Skau. Substitutional dynamical systems, Bratteli diagrams and dimension groups. *Ergodic Theory and Dynamical Systems*, 19:953–993, 1999.

[dLPZ13] Aldo de Luca, Elena V. Pribavkina, and Luca Q. Zamboni. A coloring problem for Sturmian and episturmian words. In *Developments in language theory*, volume 7907 of *Lecture Notes in Computer Science*, pages 191–202. Springer, Heidelberg, 2013.

[dLPZ14] Aldo de Luca, Elena V. Pribavkina, and Luca Q. Zamboni. A coloring problem for infinite words. *Journal of Combinatorial Theory. Series A*, 125:306–332, 2014.

[dLZ16] Aldo de Luca and Luca Q. Zamboni. On some variations of coloring problems of infinite words. *Journal of Combinatorial Theory. Series A*, 137:166–178, 2016.

[DRR18] Francesco Dolce, Antonio Restivo, and Christophe Reutenauer. On generalized lyndon words. *Theoretical Computer Science*, 777:232–242, 2018.

[DRR19] Francesco Dolce, Antonio Restivo, and Christophe Reutenauer. Some variations on lyndon words. *arXiv:1904.00954*, 2019.

[Dur98] Fabien Durand. A characterization of substitutive sequences using return words. *Discrete Mathematics*, 179:89–101, 1998.

[Duv83] Jean-Pierre Duval. Factorizing words over an ordered alphabet. *Journal of Algorithms*, 4(4):363–381, 1983.

[dZ16] Aldo de Luca and Luca Q. Zamboni. On prefixal factorizations of words. *European Journal of Combinatorics*, 52:59 – 73, 2016.

[Evd68] Alexader A. Evdokimov. Strongly asymmetric sequences generated by a finite number of symbols. *Soviet Math. Dokl.*, 9:536–539, 1968.

[Fic11] Gabriele Fici. A classification of trapezoidal words. In *Proceedings 8th International Conference Words 2011, Prague, Czech Republic, 12-16th September 2011*, volume 63 of *EPTCS*, pages 129–137, 2011.

[Fic17] Gabriele Fici. Open and closed words. *Bulletin of the European Association for Theoretical Computer Science*, 123:140–149, 2017.

[FL11] Gabriele Fici and Zsuzsanna Lipták. On prefix normal words. In *Proc. of the 15th Intern. Conf. on Developments in Language Theory (DLT 2011)*, volume 6795 of *LNCS*, pages 228–238. Springer, 2011.

[FMN96] Sébastien Ferenczi, Christian Mauduit, and Arnaldo Nogueira. Substitution dynamical systems: algebraic characterization of eigenvalues. *Annales scientifiques de l'École Normale Supérieure*, Ser. 4, 29(4):519–533, 1996.

[FMO10] Shinya Fujita, Colton Magnant, and Kenta Ozeki. Rainbow Generalizations of Ramsey Theory: A Survey. *Graphs and Combinatorics*, 26(1):1–30, 2010.

[FPS19] Gabriele Fici, Mickael Postic, and Manuel Silva. Abelian antipowers in infinite words. *Advances in Applied Mathematics*, 108:67 – 78, 2019.

[FRS20] Lukas Fleischer, Samin Riasat, and Jeffrey Shallit. New bounds on antipowers in words. *Information Processing Letters*, page 106021, 2020.

[FRSZ18] Gabriele Fici, Antonio Restivo, Manuel Silva, and Luca Q. Zamboni. Anti-powers in infinite words. *Journal of Combinatorial Theory, Series A*, 157:109–119, 2018.

[Gae18] Marisa Gaetz. Anti-power j-fixes of the thue-morse word. *arXiv:1808.01528*, 2018.

[Gar19] Swapnil Garg. Antipowers in uniform morphic words and the fibonacci word. *arXiv:1907.10816*, 2019.

[HKS95] Albertus Hof, Oliver Knill, and Barry Simon. Singular continuous spectrum for palindromic Schrödinger operators. *Communications in Mathematical Physics*, 174:149–159, 1995.

[Hol13] Stepan Holub. Abelian powers in paper-folding words. *Journal of Combinatorial Theory, Serie A*, 120(4):872–881, 2013.

[HVZ16] Tero Harju, Jetro Vesti, and Luca Q. Zamboni. On a question of Hof, Knill and Simon on palindromic substitutive systems. *Monatshefte für Mathematik*, 179(3):379–388, 2016.

[HZ98] Charles Holton and Luca Q. Zamboni. Geometric realizations of substitutions. *Bulletin de la Société Mathématique de France*, 126(2):149–179, 1998.

[HZ99] Charles Holton and Luca Q. Zamboni. Descendants of primitive substitutions. *Theory of Computing Systems*, 32:133–157, 1999.

[Jus72] Jacques Justin. Characterization of the repetitive commutative semigroups. *Journal of Algebra*, 21(1):87 – 90, 1972.

[Ker92] Veikko Keränen. Abelian squares are avoidable on 4 letters. In *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 41–52, Berlin, Heidelberg, 1992. Springer.

[KRR$^+$19] Tomasz Kociumaka, Jakub Radoszewski, Wojciech Rytter, Juliusz Straszyński, Tomasz Waleń, and Wiktor Zuba. Efficient Representation and Counting of Antipower Factors in Words. In *Language and Automata Theory and Applications*, Lecture Notes in Computer Science, pages 421–433, Cham, 2019. Springer International Publishing.

[KS67] Anatole B. Katok and Anatoly M. Stepin. Approximations in ergodic theory. *Uspekhi Matematicheskikh Nauk*, 22:81–106, 1967. In Russian, translated in *Russian Mathematical Surveys* 22:76–102 , 1967.

[Lot97] M. Lothaire. *Combinatorics on Words, 2nd edition*. Cambridge University Press, 1997.

[Lot02] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002.

[Lyn54] Roger C. Lyndon. On burnside's problem. *Transactions of the American Mathematical Society*, 77(2):202–215, 1954.

[MH38] Marston Morse and Gustav A. Hedlund. Symbolic dynamics. *American Journal of Mathematics*, 60(4):815–866, 1938.

[MH40] Marston Morse and Gustav A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *American Journal of Mathematics*, 62:1–42, 1940.

[Mos96] Brigitte Mossé. Reconnaissabilité des substitutions et complexité des suites automatiques. *Bulletin de la Société Mathématique de France*, 124(2):329–346, 1996.

[MR13] Blake Madill and Narad Rampersad. The abelian complexity of the paperfolding word. *Discrete Mathematics*, 313(7):831–838, 2013.

[Nar17] Shyam Narayanan. Functions on antipower prefix lengths of the thue-morse word, 2017.

[Ple70] Peter A. B. Pleasants. Non-repetitive sequences. *Mathematical Proceedings of the Cambridge Philosophical Society*, 68(2):267–274, 1970.

[Pos19] Mickaël Postic. Anti-powers in primitive uniform substitutions. *arXiv:1908.10627*, 2019.

[PP20] Olga Parshina and Mickaël Postic. Open and closed complexity of infinite words. *arXiv:2005.06254*, 2020.

[Pri00] Natalie Priebe. Towards a characterization of self-similar tilings in terms of derived Voronoï tessellations. *Geometriae Dedicata*, 79:239–265, 2000.

[PZ15] Svetlana Puzynina and Luca Q. Zamboni. Additive properties of sets and substitutive dynamics. In *Mathematics of aperiodic order*, volume 309 of *Progress in Mathematics*, pages 371–403. Birkhäuser/Springer, Basel, 2015.

[PZ19a] Olga Parshina and Luca Q. Zamboni. Open and closed factors in Arnoux-Rauzy words. *Advances in Applied Mathematics*, 107:22–31, 2019.

[PZ19b] Mickaël Postic and Luca Q. Zamboni. Omega-lyndon words. *Theoretical Computer Science*, 809:39–44, 2019.

[Rad79] David Radford. A natural ring basis for the shuffle algebra and an application to group schemes. *Journal of Algebra*, 58:432–454, 1979.

[Ram30] Frank Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, 30:264–286, 1930.

[Reu05] Christophe Reutenauer. Mots de lyndon généralisés. *Seminaire Lotharingien de Combinatoire*, 2005.

[Ria19] Samin Riasat. Powers and anti-powers in binary words. Master's thesis, University of Waterloo, 2019.

[RSZ11] Gwenaël Richomme, Kalle Saari, and Luca Q. Zamboni. Abelian complexity in minimal subshifts. *Journal of the London Mathematical Society*, 83:79–95, 2011.

[Shi53] Anatoly Shirshov. Subalgebras of free lie algebras. *(Russian) Mathematical Sbornik N.S. 33(75)*, pages 441–452, 1953.

[SMDS94] Rani Siromoney, Lisa Mathew, Vincent Rajkumare Dare, and Kumbakonan G. Subramanian. Infinite lyndon words. *Information Processing Letters*, 50(2):101–104, 1994.

[SS16] Luke Schaeffer and Jeffrey Shallit. Closed, rich, privileged, trapezoidal, and balanced words in automatic sequences. *Electronic Journal of Combinatorics*, 23, 2016.

[Thu06] Axel Thue. Uber Unendliche Zeichenreihen. *Norske Vid Selsk. Skr. I Mat-Nat Kl.(Christiana)*, 7:1–22, 1906.

[Thu12] Axel Thue. Uber die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifer, I. Mat. Nat. Kl.*, pages 1–67, 1912.

[van27] Bartel Leendert van der Waerden. Beweis einer Baudetschen Vermutung. *Nieuw Arch. Wiskd., II. Ser.*, 15:212–216, 1927.

[Vui01] Laurent Vuillon. A characterization of Sturmian words by return words. *European Journal of Combinatorics*, 22(2):263–275, 2001.

[WZ18a] Caïus Wojcik and Luca Q. Zamboni. Coloring problems for infinite words. In *Sequences, groups, and number theory*, Trends Math., pages 213–231. Birkhäuser/Springer, Cham, 2018.

[WZ18b] Caïus Wojcik and Luca Q. Zamboni. Monochromatic factorizations of words and periodicity. *Mathematika*, 64(1):115–123, 2018.