Well-quasi-orderings on word languages

Nathan Lhote[1] [0000−0003−3303−5368], Aliaume Lopez[3] [0000−0002−4205−327X], and Lia Schütze[2] [0000−0003−4002−5491]

[1] Aix-Marseille University
[2] Max Planck Institute for Software Systems
[3] INP Bordeaux, LaBRI, CNRS

Abstract. The set of finite words over a well-quasi-ordered set is itself well-quasi-ordered. This seminal result by Higman is a cornerstone of the theory of well-quasi-orderings and has found numerous applications in computer science. However, this result is based on a specific choice of ordering on words, the (scattered) subword ordering. In this paper, we describe to what extent other natural orderings (prefix, suffix, and infix) on words can be used to derive Higman-like theorems. More specifically, we are interested in characterizing *languages* of words that are well-quasi-ordered under these orderings, and explore their properties and connections with other language theoretic notions. We furthermore give decision procedures when the languages are given by various computational models such as automata, context-free grammars, and automatic structures.

This document uses knowledge: a notion points to its *definition*.

1 Introduction

A *well-quasi-ordered* set is a set X equipped with a quasi-order $\preceq$ such that every infinite sequence $(x_n)_{n\in\mathbb{N}}$ of elements taken in X contains an increasing pair $x_i \preceq x_j$ with $i < j$. Well-quasi-orderings serve as a core combinatorial tool powering many termination arguments, and were successfully applied to the verification of infinite state transition systems [2,1]. One of the appealing properties of well-quasi-orderings is that they are closed under many operations, such as taking products, finite unions, and finite powerset constructions [13]. Perhaps more surprisingly, the class of well-quasi-ordered sets is also stable under the operation of taking finite words and finite trees labeled by elements of a well-quasi-ordered set [20,23].

Note that in the case of finite words and finite trees, the precise choice of ordering is crucial to ensure that the resulting structure is well-quasi-ordered. The celebrated result of Higman states that the set of finite words over a well-quasi-ordered alphabet $(X, \preceq)$ is well-quasi-ordered by the so-called subword embedding relation [20]. Let us recall that the subword relation for words over $(X, \preceq)$ is defined as follows: a word u is a *subword* of a word v, written $u \leq^* v$, if there exists an increasing function $f\colon \{1,\ldots,|u|\} \to \{1,\ldots,|v|\}$ such that $u_i \preceq v_{f(i)}$ for all $i \in \{1,\ldots,|u|\}$.

However, there are many other natural orderings on words that could be considered in the context of well-quasi-orderings, even in the simplified setting of a finite alphabet Σ equipped with the equality relation. In this setting, the three alternatives we consider are the *prefix relation* ($u \sqsubseteq_{\mathsf{pref}} v$ if there exists w with $uw = v$), the *suffix relation* ($u \sqsubseteq_{\mathsf{suff}} v$ if there exists w such that $wu = v$), and the *infix relation* ($u \sqsubseteq_{\mathsf{infix}} v$ if there exist w_1, w_2 such that $w_1uw_2 = v$). Note that these three relations straightforwardly generalize to infinite quasi-ordered alphabets. Unfortunately, it is easy to see that none of these relations yield well-quasi-ordered sets as soon as the alphabet contains two distinct letters: for instance, the infinite sequence of words $(ab^na)_{n\in\mathbb{N}}$ is well-quasi-ordered by the subword relation but by neither the prefix relation, nor the suffix relation, nor the infix relation.

While this dooms well-quasi-orderedness of these relations in the general case, there may be *subsets* of Σ^* which are well-quasi-ordered by these relations. As a simple example, take the case of finite sets of (finite) words which are all well-quasi-ordered regardless of the ordering considered. This raises the question of characterizing exactly which subsets $L \subseteq \Sigma^*$ are well-quasi-ordered with respect to the prefix relation (respectively, the suffix relation or the infix relation), and designing suitable decision procedures.

Let us argue that these decision procedures fit a larger picture in the research area of well-quasi-orderings. Indeed, there have been recent breakthroughs in deciding whether a given order is a well-quasi-order, for instance in the context of the verification of infinite state transition systems [19] or in the context of logic [7]. In the graph theory community, recent works have studied classes of graphs that are well-quasi-ordered by the induced subgraph relation using similar language theoretic techniques [12,27,6]. Furthermore, a previous work by Kuske shows that any *reasonable*[4] partially ordered set $(X, \leq)$ can be embedded into $\{a,b\}^*$ with the infix relation [25, Lemma 5.1]. Phrased differently, one can encode a large class of partially ordered sets as subsets of $\{a,b\}^*$. As a consequence, the following decision problem provides a reasonable abstract framework for deciding whether a given partially ordered set is well-quasi-ordered: given a language $L \subseteq \Sigma^*$, decide whether L is well-quasi-ordered by the infix relation.

The runtime of an algorithm based on well-quasi-orderings is deeply related to the "complexity" of the underlying quasi-order $(X, \leq)$ [31]. One way to measure this complexity is to consider its so-called ordinal invariants: for instance, the maximal order type (or m.o.t., $\mathfrak{o}(X)$), originally defined by De Jongh and Parikh [21], is the order type of the maximal linearization of a well-quasi-ordered set. In the case of a finite set, the m.o.t. is precisely the size of the set. Better runtime bounds were obtained by considering two other parameters [32]: the ordinal height ($\mathfrak{h}(X)$)[30], and the ordinal width ($\mathfrak{w}(X)$) [26]. Therefore, when characterizing well-quasi-ordered languages, we will also be interested in deriving upper bounds on their ordinal invariants. This analysis also allows us to better compare the

[4] This will be made precise in Lemma 7.

well-quasi-orderings. We refer to Section 2 for a more detailed introduction to these parameters and ordinal computations in general.

Contributions We focus on languages over a finite alphabet Σ. In this setting, we first characterize languages that are well-quasi-ordered by the prefix relation (and symmetrically, by the suffix relation), and derive tight bounds on their ordinal invariants. These generic results are then used to devise a decision procedure for checking whether a language is well-quasi-ordered by the prefix relation, provided the language is given as input as a finite automaton (Corollary 4). A summary of these results can be found in Figure 1.

L	**Characterisation**	$\mathfrak{w}(L)$	$\mathfrak{o}(L)$
arbitrary	Theorem 5: finite unions of chains	$< \omega$	$< \omega^2$
regular	Corollary 4: finite unions of regular chains	$< \omega$	$< \omega^2$

Fig. 1: Summary of results for the prefix relation (and symmetrically, for the suffix relation).

We then turn our attention to the infix relation. In this case, we notice that Lemma 5.1 from [25] implies that there are well-quasi-ordered languages for the infix relation that have arbitrarily large ordinal invariants (except for the ordinal height, which is always at most ω). Therefore, we focus on two natural semantic restrictions on languages: on the one hand, we consider bounded languages, that is, languages included in some $w_1^* \cdots w_k^*$ for some finite choice of words $w_1, \ldots, w_k$; on the other hand, we consider downward closed languages, that is, languages closed under taking infixes. In both cases, we provide a very precise characterization of well-quasi-ordered languages by the infix relation, and derive tight bounds on their ordinal invariants. These results are summarized in Figure 2. We furthermore notice that for downward closed languages that are well-quasi-ordered by the infix relation, being bounded is the same as being regular (Lemma 33), and that a bounded language is well-quasi-ordered by the infix relation if and only if its downwards closure is well-quasi-ordered by the infix relation (Corollary 15). This shows that, for bounded languages, being well-quasi-ordered implies that their downwards closure is a regular language, which is a weakening of the usual result that the downwards closure of *any language* for the scattered subword relation is always a regular language.

Turning our attention to decision procedures, we consider two computational models respectively tailored to downward closed languages and to bounded languages. For downward closed languages, we consider a model based on representations of infinite words (Section 5.2), for which we provide a decision procedure (Theorem 27). The model used to represent these infinite words is based on automatic sequences and morphic sequences [11], which are well-studied in the context of symbolic dynamics. For bounded languages, we consider the

L	**Characterisation**	$\mathfrak{w}(L)$	$\mathfrak{o}(L)$
arbitrary	Lemma 7: countable well-quasi orders with finite initial segments	$< \omega_1$	$< \omega_1$
bounded	Theorem 8: finite union of products of chains for the prefix and suffix relations	$< \omega^2$	$< \omega^3$
downward closed	Theorem 20: finite union of infixes of ultimately uniformly recurrent words	$< \omega^2$	$< \omega^3$

Fig. 2: Summary of results for the infix relation, the bounds on $\mathfrak{w}(L)$ and $\mathfrak{o}(L)$ are tight, and respectively proven in Corollary 14 and Corollary 26.

model of amalgamation systems [5], which is an abstract computational model that encompasses many classical ones, such as finite automata, context-free grammars, and Petri nets [5]. We show that if a language recognized by an amalgamation system is well-quasi-ordered by the infix relation, then it is a bounded language (Theorem 29), and is therefore regular. Furthermore, we show that we can decide whether a given language recognized by an amalgamation system is well-quasi-ordered by the infix relation (Theorem 30). We defer the introduction of amalgamation systems to Section 6.1.

Related work The study of alternative well-quasi-ordered relations over finite words is far from new. For instance, orders obtained by so-called *derivation relations* were already analysed by Bucher, Ehrenfeucht, and Haussler [9], and were later extended by D'Alessandro and Varricchio [16,17]. However, in all those cases the orderings are *multiplicative*, that is, if $u_1 \preceq v_1$ and $u_2 \preceq v_2$ then $u_1 u_2 \preceq v_1 v_2$. This assumption does not hold for the prefix, suffix, and infix relations.

A similar question was studied by Atminas, Lozin, and Moshkov [6], in the hope of finding characterizations of classes of *finite graphs* that are well-quasi-ordered by the *induced subgraph relation* [6, Section 7]. In this setting, it is common to refer to classes of graphs via a list of *forbidden patterns*, which are finite graphs that cannot be found as induced subgraphs in the class. Applying this reasoning to finite words with the infix relation, they provide an efficient decision procedure for checking whether a language $L \subseteq \Sigma^*$ is well-quasi-ordered by the infix relation whenever said language is given as input via a list of *forbidden factors* [6, Theorem 1, Theorem 2]. The key construction of their paper is to study languages L that are *regular* (recognized by some finite deterministic automata), for which they can decide whether L is well-quasi-ordered by the infix relation [6, Theorem 1]. Because it is easy to transform a list of forbidden factors into a regular language [6, Theorem 1], this yields the desired decision procedure. Our work extends this result in several ways: first, we also consider the prefix relation and the suffix relation, then we consider non-regular languages, and finally, we provide very precise descriptions of the well-quasi-ordered languages, as well as tight bounds on their ordinal invariants.

Outline We introduce in Section 2 the necessary background on well-quasi-orders and ordinal invariants. In Section 3, which is relatively self-contained, we study the prefix relation and prove in Theorem 5 the characterization of well-quasi-ordered languages by the prefix relation. In Section 4, we obtain the infix analogue of Theorem 5 specifically for bounded languages (Theorem 8). In Section 5, we study the downward closed languages, characterize them using a notion of ultimately uniformly recurrent words borrowed from symbolic dynamics (Theorem 20), and compute bounds on their ordinal invariants in Corollary 26. Finally, we generalize these results to all amalgamation systems in Section 6 in (Theorem 29), and provide a decision procedure for checking whether a language is well-quasi-ordered by the infix relation (resp. prefix and suffix) in this context (Theorem 30).

Acknowledgements We would like to thank participants of the 2024 edition of Autobóz for their helpful comments and discussions. We would also like to thank Vincent Jugé for his pointers on word combinatorics.

2 Preliminaries

Finite words. In this paper, we use upper Greek letters Σ, Γ to denote finite alphabets, Σ^* to denote the set of finite words over Σ, and ε for the empty word in Σ^*. In order to give some intuition on the decision problems, we will sometimes use the notion of *finite automata*, *regular languages*, and Monadic Second Order logic (MSO) over finite words, and assume the reader to be familiar with them. We refer to the textbook of [33] for a detailed introduction. However, we will require no prior knowledge on word combinatorics.

Orderings and Well-Quasi-Orderings. A *quasi-order* is a reflexive and transitive binary relation, it is a *partial order* if it is furthermore antisymmetric. A *total order* is a partial order where any two elements are comparable. Let us now introduce some notations for well-quasi-orders. A sequence $(x_i)_{n\in\mathbb{N}}$ in a set X is *good* if there exist $i < j$ such that $x_i \leq x_j$. It is *bad* otherwise. Therefore, a well-quasi-ordered set is a set where every infinite sequence is good. A *decreasing sequence* is a sequence $(x_i)_{n\in\mathbb{N}}$ such that $x_{i+1} < x_i$ for all i, a *chain* is a sequence such that $x_i \leq x_{i+1}$ for all i, and an *antichain* is a set of pairwise incomparable elements. An equivalent definition of a well-quasi-ordered set is that it contains no infinite decreasing sequences, nor infinite antichains. We refer to [13] for a detailed survey on well-quasi-orders.

The prefix relation (resp. the suffix relation and the infix relation) on Σ^* are always *well-founded*, i.e., there are no infinite decreasing sequences for this ordering. In particular, for a language $L \subseteq \Sigma^*$ to be well-quasi-ordered with respect to one of these orderings, it suffices to prove that it contains no infinite antichain.

A useful operation on quasi-ordered sets is to compute the *upwards closure* of a set S for a relation $\preceq$, which is defined as $\uparrow_{\preceq} S \triangleq \{y \in \Sigma^* \mid \exists x \in S. x \preceq y\}$. In this paper, we will also use the symmetric notion of *downwards closure*:

$\downarrow_{\preceq} S \triangleq \{y \in \Sigma^* \mid \exists x \in S. y \preceq x\}$. Abusing notations, we will write $\uparrow w$ and $\downarrow w$ for the upwards and downwards closure of a single element w, omitting the ordering relation when it is clear from the context. A set S is called *downward closed* if $\downarrow S = S$.

Ordinal Invariants. An *ordinal* is a well-founded totally ordered set. We use α, β, γ to denote ordinals, and use ω to denote the first infinite ordinal, i.e., the set of natural numbers with the usual ordering. We also use ω_1 to denote the first *uncountable* ordinal. We only assume superficial familiarity with ordinal arithmetic, and refer to the books of Kunen [24] and Krivine [22, Chapter II] for a detailed introduction to this domain. Given a tree T whose branches are all finite we can define an ordinal α_T inductively as follows: if T is a leaf then $\alpha_T = 0$, if T has children $(T_i)_{i \in I}$ then $\alpha_T = \sup\{\alpha_{T_i} + 1 \mid i \in I\}$. We say that α_T is the *rank* of T.

Let $(X, \leq)$ be a well-quasi-ordered set. One can define three well-founded trees from X: the tree of bad sequences, the tree of decreasing sequences, and the tree of antichains. The nodes of these trees are respectively the bad sequences, the decreasing sequences, and the antichains of X, and the ancestor relation is the prefix relation on sequences (or subset relation on antichains). The rank of these trees are called respectively the *maximal order type* of X written $\mathfrak{o}(X)$ [21], the *ordinal height* of X written $\mathfrak{h}(X)$ [30], and the *ordinal width* of X written $\mathfrak{w}(X)$ [26]. These three parameters are called the *ordinal invariants* of a well-quasi-ordered set X. As an example, for $(\mathbb{N}, \leq)$, all bad sequences are descending and antichains have size at most 1. In fact, $(\mathbb{N}, \leq)$ is itself an ordinal, namely ω. Hence it is its own maximal order type and ordinal height, and its ordinal width is 1. We refer to the survey of [15] for a detailed discussion on these concepts and their computation on specific classes of well-quasi-ordered sets.

We will use the following inequality between ordinal invariants, due to [26], and that was recalled in [15, Theorem 3.8]: $\mathfrak{o}(X) \leq \mathfrak{h}(X) \otimes \mathfrak{w}(X)$, where $\otimes$ is the *commutative ordinal product*, also known as the *Hessenberg product*. We will not recall the definition of this product here, and refer to [15, Section 3.5] for a detailed introduction to this concept. The only equalities we will use are $\omega \otimes \omega = \omega^2$ and $\omega^2 \otimes \omega = \omega^3$.

3 Prefixes and Suffixes

In this section, we study the well-quasi-ordering of languages under the prefix relation. Let us immediately remark that the map $u \mapsto u^R$ that reverses a word is an order-bijection between $(X^*, \sqsubseteq_{\mathsf{pref}})$ and $(X^*, \sqsubseteq_{\mathsf{suff}})$, that is, $u \sqsubseteq_{\mathsf{pref}} v$ if and only if $u^R \sqsubseteq_{\mathsf{suff}} v^R$. Therefore, we will focus on the prefix relation in the rest of this section, as $(L, \sqsubseteq_{\mathsf{pref}})$ is well-quasi-ordered if and only if $(L^R, \sqsubseteq_{\mathsf{suff}})$ is.

The next remark we make is that Σ^* is not well-quasi-ordered by the prefix relation as soon as Σ contains two distinct letters a and b. As an example of infinite antichain, we can consider the set of words $a^n b$ for $n \in \mathbb{N}$. As mentioned in the introduction, there are however some languages that are well-quasi-ordered by the prefix relation. A simple example being the (regular) language $a^* \subseteq \{a,b\}^*$, which is order-isomorphic to natural numbers with their usual orderings $(\mathbb{N}, \leq)$.

In order to characterize the existence of infinite antichains for the prefix relation, we will introduce the following tree.

Definition 1. *The tree of prefixes over a finite alphabet Σ is the infinite tree T whose nodes are the words of Σ^*, and such that the children of a word w are the words wa for all $a \in \Sigma$.*

We will use this tree of prefixes to find simple witnesses of the existence of infinite antichains in the prefix relation for a given language L, namely by introducing antichain branches.

Definition 2. *An antichain branch for a language L is an infinite branch B of the tree of prefixes such that from every point of the branch, one can reach a word in $L \setminus B$. Formally: $\forall u \in B, \exists v \in \Sigma^*, uv \in L \setminus B$.*

Let us illustrate the notion of antichain branch over the alphabet $\Sigma = \{a,b\}$, and the language $L = a^*b$. In this case, the set a^* (which is a branch of the tree of prefixes) is an antichain branch for L. This holds because for any a^k, the word $a^k b$ belongs to $L \setminus a^*$. In general, the existence of an antichain branch for a language L implies that L contains an infinite antichain, and because the alphabet Σ is assumed to be finite, one can leverage the fact that the tree of prefixes is finitely branching to prove that the converse holds as well.

Lemma 3. *Let $L \subseteq \Sigma^*$ be a language. Then, L contains an infinite antichain if and only if there exists an antichain branch for L.*

One immediate application of Lemma 3 is that antichain branches can be described inside the tree of prefixes by a monadic second order formula (MSO-formula), allowing us to leverage the decidability of MSO over infinite binary trees [29, Theorem 1.1]. This result will follow from our general decidability result (Theorem 30) but is worth stating on its own for its simplicity.

Corollary 4. *If L is regular, then the existence of an infinite antichain is decidable.*

Let us now go further and fully characterize languages L such that the prefix relation is well-quasi-ordered, without any restriction on the decidability of L itself.

Theorem 5. *A language $L \subseteq \Sigma^*$ is well-quasi-ordered by the prefix relation if and only if L is a finite union of chains.*

As an immediate consequence, we have a very fine-grained understanding of the ordinal invariants of such well-quasi-ordered languages, which can be leveraged in bounding the complexity of algorithms working on such languages.

Corollary 6. *Let $L \subseteq \Sigma^*$ be a language that is well-quasi-ordered by the prefix relation. Then, the maximal order type of L is strictly smaller than ω^2, the ordinal height of L is at most ω, and its ordinal width is finite. Furthermore, these bounds are tight.*

Proof. The upper bounds follow from the fact that L is a finite union of chains. The tightness can be obtained by considering the languages $L_k \triangleq \bigcup_{i=0}^{k-1} a^i b^*$ for $k \in \mathbb{N}$, which are well-quasi-ordered by the prefix relation (as they are finite unions of chains), and satisfy that $\mathfrak{w}(L_k) = k$, $\mathfrak{h}(L_k) = \omega$, and therefore $\mathfrak{o}(L_k) = k \cdot \omega$. □

4 Infixes and Bounded Languages

In this section, we study languages equipped with the infix relation. As opposed to the prefix and suffix relations, the infix relation can lead to very complicated well-quasi-ordered languages. Formally, the upcoming Lemma 7 due to Kuske shows that *any* countable partial-ordering with finite initial segments can be embedded into the infix relation of a language. To make the former statement precise, let us recall that an *order embedding* from a quasi-ordered set $(X, \preceq)$ into a quasi-ordered set $(Y, \preceq')$ is a function $f : X \to Y$ such that for all $x, y \in X$, $x \preceq y$ if and only if $f(x) \preceq' f(y)$. When such an embedding exists, we say that X *embeds into* Y. Recall that a quasi-ordered set $(X, \preceq)$ is a partial ordering whenever the relation $\preceq$ is antisymmetric, that is $x \preceq y$ and $y \preceq x$ implies $x = y$.

Lemma 7. *[25, Lemma 5.1]* Let $(X, \preceq)$ be a partially ordered set, and Σ be an alphabet with at least two letters. Then the following are equivalent:

1. X embeds into $(\Sigma^*, \sqsubseteq_{\mathsf{infix}})$,
2. X is countable, and for every $x \in X$, its downwards closure $\downarrow_{\preceq} x$ is finite (that is, $(X, \preceq)$ has *finite initial segments*).

As a consequence of Lemma 7, we cannot replay proofs of Section 3, and will actually need to leverage some regularity of the languages to obtain a characterization of well-quasi-ordered languages under the infix relation. This regularity will be imposed through the notion of *bounded languages*, i.e., languages $L \subseteq \Sigma^*$ such that there exists words $w_1, \ldots, w_n$ satisfying $L \subseteq w_1^* \cdots w_n^*$. Let us now state the main theorem of this section.

▷ Proven p.10

Theorem 8. *Let L be a bounded language of Σ^*. Then, L is a well-quasi-order when endowed with the infix relation if and only if it is included in a finite union of products $S_i \cdot P_i$ where S_i is a chain for the suffix relation, and P_i is a chain for the prefix relation, for all $1 \leq i \leq n$.*

Let us first remark that if S is a chain for the suffix relation and P is a chain for the prefix relation, then SP is well-quasi-ordered for the infix relation. This proves the (easy) right-to-left implication of Theorem 8.

In order to prove the (difficult) left-to-right implication of Theorem 8, we will rely heavily on the combinatorics of periodic words. Let us use a slightly non-standard notation by saying that a non-empty word $w \in \Sigma^+$ is *periodic* with period $x \in \Sigma^*$ if there exists a $p \in \mathbb{N}$ such that $w \sqsubseteq_{\mathsf{infix}} x^p$. The *periodic length* of a word u is the minimal length of a period x of u.

The reason why periodic words built using a given period $x \in \Sigma^+$ are interesting for the infix relation is that they naturally create chains for the prefix and suffix relations. Indeed, if $x \in \Sigma^+$ is a finite word, then $\{x^p \mid p \in \mathbb{N}\}$ is a chain for the infix relation. Note that in general, the downwards closure of a chain is *not* a chain (see Remark 9). However, for the chains generated using periodic words, the downwards closure $\downarrow_{\sqsubseteq_{\mathsf{infix}}} \{x^p \mid p \in \mathbb{N}\}$ is a *finite union* of chains. Because this set will appear in bigger equations, we introduce the shorter notation $\mathsf{P}{\downarrow}(x)$ for the set of infixes of words of the form x^p, where $p \in \mathbb{N}$.

Remark 9. Let $(X, \preceq)$ be a quasi-ordered set, and $L \subseteq X$ be such that $(L, \preceq)$ is well-quasi-ordered. It is not true in general that $(\downarrow L, \preceq)$ is well-quasi-ordered. In the case of $(\Sigma^*, \sqsubseteq_{\mathsf{infix}})$ a typical example is to start from an infinite antichain A, together with an enumeration $(w_i)_{i \in \mathbb{N}}$ of A, and build the language $L \triangleq \{\prod_{i=0}^{n} w_i \mid n \in \mathbb{N}\}$. By definition, L is a chain for the infix ordering, hence well-quasi-ordered. However, $\downarrow_{\sqsubseteq_{\mathsf{infix}}} L$ contains A, and is therefore not well-quasi-ordered.

Lemma 10. *Let $x \in \Sigma^+$ be a word. Then $\mathsf{P}{\downarrow}(x)$ is a finite union of chains for the infix, prefix and suffix relations.*

The following combinatorial Lemma 12 connects the property of being well-quasi-ordered to a property of the periodic lengths of words in a language, based on the assumption that some factors can be iterated. It is the core result that powers the analysis done in Theorems 8 and 29. It is fundamentally based on a classical result of combinatorics on words (Lemma 11) that we recall here for the sake of completeness.

Lemma 11 ([18, Theorem 1]). *Let $u, v \in \Sigma^+$ be two words and $n = \gcd(|u|, |v|)$. If there exists $p, q \in \mathbb{N}$ such that u^p and v^q have a common prefix of length at least $|uv| - n$, then there exists $z \in \Sigma^+$ such that u and v are powers of z, and in particular z has length at most $\min\{|u|, |v|\}$.*

Lemma 12. *Let $L \subseteq \Sigma^*$ be a language that is well-quasi-ordered by the infix relation. Let $k \in \mathbb{N}$, $u_1, \cdots, u_{k+1} \in \Sigma^*$, and $v_1, \cdots, v_k \in \Sigma^+$ be such that $w[\boldsymbol{n}] \triangleq (\prod_{i=1}^{k} u_i v_i^{n_i}) u_{k+1}$ belongs to L for vectors $\boldsymbol{n} \in \mathbb{N}^k$ with all coordinates arbitrarily large. Then, there exist $x, y \in \Sigma^+$ of size at most $\max\{|v_i| \mid 1 \leq i \leq k\}$ such that for all $\boldsymbol{n} \in \mathbb{N}^k$ one of the following holds: $w[\boldsymbol{n}] \in \mathsf{P}{\downarrow}(x) u_i \mathsf{P}{\downarrow}(y)$ for some $1 \leq i \leq k+1$.*

Lemma 13. *Let $L \subseteq \Sigma^*$ be a bounded language that is well-quasi-ordered by the infix relation. Then, there exists a finite subset $E \subseteq (\Sigma^*)^3$, such that:*

$$L \subseteq \bigcup_{(x,u,y) \in E} \mathsf{P}{\downarrow}(x) u \, \mathsf{P}{\downarrow}(y) \quad .$$

▷ Back to p.8

Proof (Proof of Theorem 8 as stated on page 8). We apply Lemma 13, and conclude because $\mathsf{P}{\downarrow}(x)$ is a finite union of chains for the prefix, suffix and infix relations (Lemma 10). ☐

Corollary 14. *Let L be a bounded language of Σ^* that is well-quasi-ordered by the infix relation. Then, the ordinal width of L is less than ω^2, its ordinal height is at most ω, and its maximal order type is less than ω^3. Furthermore, those three bounds are tight.*

Proof. Upper bounds are a direct consequence of Theorem 8. For the tightness, remark that the ordinal width, ordinal height and maximal order type of the language $a * b*$ are respectively ω, ω and ω^2. Thus, by using finite unions of languages of the form $a_i * b_i*$, one can reach the desired bounds $\omega \cdot k$, ω and $\omega^2 \cdot k$.[5] ☐

5 Infixes and Downwards Closed Languages

Let us now discuss another classical restriction that can be imposed on languages when studying well-quasi-orders, that of being downward closed. Indeed, Lemma 7 crucially relies on constructing languages that are *not* downward closed, and we have shown in Remark 9 that the downwards closure of a well-quasi-ordered language is not necessarily well-quasi-ordered.

5.1 Characterization of Well-Quasi-Ordered Downward Closed Languages

An immediate consequence of Theorem 8 is that if L is a bounded language, then considering L or its downwards closure ${\downarrow_{\sqsubseteq_{\mathsf{infix}}}} L$ is equivalent with respect to being well-quasi-ordered by the infix relation, as opposed to the general case illustrated in Remark 9.

Corollary 15. *Let L be a bounded language of Σ^*. Then, L is a well-quasi-order when endowed with the infix relation if and only if ${\downarrow_{\sqsubseteq_{\mathsf{infix}}}} L$ is.*

Corollary 15 is reminiscent of a similar result for the subword embedding, stipulating that for any language $L \subseteq \Sigma^*$, the downwards closure ${\downarrow_{\leq^*}} L$ is described using finitely many excluded subwords, hence is regular. However, this is not the case for the infix relation, even with bounded languages, as we will now illustrate with the following example.

Example 16. Let $L \triangleq a^*b^* \cup b^*a^*$. This language is bounded, is downward closed for the infix relation, is well-quasi-ordered for the infix relation, but is characterized by an *infinite* number of excluded infixes, respectively of the form ab^ka and ba^kb where $k \geq 1$.

[5] Note that one can encode these languages using a binary alphabet.

To strengthen Example 16, we will leverage the *Thue-Morse sequence* $\mathbf{t} \in \{0,1\}^{\mathbb{N}}$, which we will use as a black-box for its two main characteristics: it is cube-free and uniformly recurrent. Being *cube-free* means that no (finite) word of the form uuu is an infix of $\mathbf{t}$, and being *uniformly recurrent* means that for every word u that is an infix of $\mathbf{t}$, there exists $k \geq 1$ such that u occurs as an infix of every k-sized infix $v \sqsubseteq_{\mathsf{infix}} \mathbf{t}$. We refer the reader to a nice survey of Allouche and Shallit for more information on this sequence and its properties [4]. We are using the notations $\Sigma^{\mathbb{N}}$ for the set of infinite words over the alphabet Σ, and will be using $\Sigma^{\mathbb{Z}}$ for the set of bi-infinite words over Σ.

Theorem 17. *Let $w \in \Sigma^{\mathbb{N}}$ be a uniformly recurrent word. Then, the set of finite infixes of w is well-quasi-ordered for the infix relation.*

Proof. Let L be the set of finite infixes of w. Consider a sequence $(u_i)_{i\in\mathbb{N}}$ of words in L. Without loss of generality, we may consider a subsequence such that $|u_i| < |u_{i+1}|$ for all $i \in \mathbb{N}$. Because $\mathbf{t}$ is uniformly recurrent, there exists $k \geq 1$ such that u_1 is an infix of every word v of size at least k. In particular, u_1 is an infix of u_k, hence the sequence $(u_i)_{i\in\mathbb{N}}$ is good. ☐

Lemma 18. *The language $I_{\mathbf{t}}$ of infixes of the Thue-Morse sequence is downward closed for the infix relation, well-quasi-ordered for the infix relation, but is not bounded.*

Proof. By construction $I_{\mathbf{t}}$ is downward closed for the infix relation, and by Theorem 17, it is well-quasi-ordered.

Assume by contradiction that $I_{\mathbf{t}}$ is bounded. In this case, there exist words $w_1, \ldots, w_k \in \Sigma^*$ such that $I_{\mathbf{t}} \subseteq w_1^* \cdots w_k^*$. Since $I_{\mathbf{t}}$ is infinite and downward closed, there exists a word $u \in I_{\mathbf{t}}$ such that $u = w_i^3$ for some $1 \leq i \leq k$. This is a contradiction, because $u \sqsubseteq_{\mathsf{infix}} \mathbf{t}$, which is cube-free. ☐

One may refine our analysis of the Thue-Morse sequence to obtain precise bounds on the ordinal invariants of its language of infixes.

Lemma 19. *Under $\sqsubseteq_{\mathsf{infix}}$, the maximal order type of $I_{\mathbf{t}}$ is ω, the ordinal height of $I_{\mathbf{t}}$ is ω, the ordinal width of $I_{\mathbf{t}}$ is ω.*

Proof. We first show that ω is an upper bound for each of these measures, before showing that the bounds are tight.

Let us prove that these are upper bounds for the ordinal invariants of $I_{\mathbf{t}}$. The bound of the ordinal height holds for any language L, as the length of a decreasing sequence of words is bounded by the length of its first element. For the maximal order type, we remark that the uniform recurrence of $\mathbf{t}$ means that the maximal length of a bad sequence is determined by its first element, hence that it is at most ω. Finally, because the ordinal width is at most the maximal order type (as per Section 2, using for instance the results of [26] or [15, Theorem 3.8] stating $\mathfrak{o}(X) \leq \mathfrak{h}(X) \otimes \mathfrak{w}(X)$): we conclude that the ordinal width is also at most ω.

Now, let us prove that these bounds are tight. It is clear that $\mathfrak{h}(I_{\mathbf{t}}) = \omega$: given any number $n \in \mathbb{N}$, one can construct a decreasing sequence of words in $I_{\mathbf{t}}$ of length n, for instance by considering the first n prefixes of the Thue-Morse sequence by decreasing size. Let us now prove that $\mathfrak{w}(I_{\mathbf{t}}) = \omega$. To that end, we can leverage the fact that the number of infixes of size n in $I_{\mathbf{t}}$ is bounded below by a non-constant affine function in n [34], and that two words of length n are comparable for the infix relation if and only if they are equal. Hence, there cannot be a finite bound on the size of an antichain in $I_{\mathbf{t}}$, and we conclude that $\mathfrak{w}(I_{\mathbf{t}}) = \omega$. Finally, because the ordinal width is at most the maximal order type, we conclude that the maximal order type of $I_{\mathbf{t}}$ is also ω. ☐

We prove in the upcoming Theorem 20 that the status of the Thue-Morse sequence is actually representative of downward closed languages for the infix relation. To that end, let us introduce the notation $\mathsf{Infixes}(w)$ for the set of finite infixes of a (possibly infinite or bi-infinite) word $w \in \Sigma^* \cup \Sigma^{\mathbb{N}} \cup \Sigma^{\mathbb{Z}}$. We say that an infinite word $w \in \Sigma^{\mathbb{N}}$ is *ultimately uniformly recurrent* if there exists a bound $N_0 \in \mathbb{N}$ such that $w_{\geq N_0}$ is uniformly recurrent. We extend this notion to finite words by considering that they all are ultimately uniformly recurrent, and to bi-infinite words by considering that they are ultimately uniformly recurrent if and only if both their left-infinite and right-infinite parts are.

▷ Proven p.12

Theorem 20. *Let L be a well-quasi-ordered language for the infix relation that is downward closed. Then, there exist finitely many ultimately uniformly recurrent words $w_1, \ldots, w_n \in \Sigma^* \cup \Sigma^{\mathbb{N}} \cup \Sigma^{\mathbb{Z}}$ such that $L = \bigcup_{i=1}^{n} \mathsf{Infixes}(w_i)$.*

To connect infixes of a (bi)-infinite word to downward closed languages, a useful notion is that of directed sets. A subset $I \subseteq X$ is *directed* if, for every $x, y \in I$, there exists $z \in I$ such that $x \leq z$ and $y \leq z$. Given a well-quasi-order $(X, \leq)$, one can always decompose X into a finite union of *order ideals*, that is, non-empty sets $I \subseteq X$ that are downward closed and directed for the relation $\leq$. In our case, a well-quasi-ordered order ideal for the infix relation is the set of finite infixes of a finite, infinite, or bi-infinite word $w \in \Sigma^* \cup \Sigma^{\mathbb{N}} \cup \Sigma^{\mathbb{Z}}$ (Lemma 21).

Lemma 21. *Let $L \subseteq \Sigma^*$ be an order ideal for the infix relation. Then L is the set of finite infixes of a finite, infinite or bi-infinite word w.*

Lemma 22. *Let $w \in \Sigma^{\mathbb{N}}$ be an infinite word. Then, the set of finite infixes of w is well-quasi-ordered for the infix relation if and only if w is ultimately uniformly recurrent.*

Lemma 23. *Let $w \in \Sigma^{\mathbb{Z}}$ be a bi-infinite word. Then, the set of finite infixes of w is well-quasi-ordered for the infix relation if and only if w is ultimately uniformly recurrent as a bi-infinite word.*

We are now ready to conclude the proof of Theorem 20.

Proof (Proof of Theorem 20 as stated on page 12). It is clear that the set of finite infixes of a finite, infinite or bi-infinite ultimately uniformly recurrent word is well-quasi-ordered for the infix relation thanks to Lemma 22.

Conversely, let us consider a well-quasi-ordered language L that is downward closed for the infix relation. Because it is a well-quasi-ordered set, it can be written as a finite union of order ideals $L = \bigcup_{i=1}^{n} L_i$.

For every such ideal L_i, we can apply Lemma 21, and conclude that L_i is the set of finite infixes of a finite, infinite or bi-infinite word w_i. Because the languages L_i are well-quasi-ordered, we can apply Lemma 22, and conclude that w_i is ultimately uniformly recurrent. ☐ ▷ Back to p.12

Finally, we comment on the ordinal invariants of the set of finite infixes of an ultimately uniformly recurrent infinite word, from which the bounds of Corollary 26 naturally follow.

Lemma 24. *Let $w \in \Sigma^{\mathbb{N}}$ be an ultimately uniformly recurrent word. Then, the set of finite infixes of w has ordinal width less than $\omega \cdot 2$. Furthermore, this bound is tight.*

Lemma 25. *Let $w \in \Sigma^{\mathbb{Z}}$ be an ultimately uniformly recurrent bi-infinite word. Then, the ordinal width of the set of finite infixes of w is less than $\omega \cdot 3$, and this bound is tight.*

Thanks to Theorem 20, and by analysing the ordinal invariants of infixes of an ultimately uniformly recurrent infinite word w (Lemma 22), we conclude that the ordinal invariants of a well-quasi-ordered downward closed language are relatively small.

Corollary 26. *Let L be a well-quasi-ordered downward closed language for the infix relation. Then, the maximal order type of L is strictly less than ω^3, its ordinal height is at most ω, and its ordinal width is at most ω^2.*

Furthermore, those bounds are tight.

5.2 Decision Procedures

As we have demonstrated, infinite (or bi-infinite words) can be used to represent languages that are well-quasi-ordered for the infix relation by considering their set of finite infixes. Let us formalise the representation of languages by sets of bi-infinite words that we will use in this section, following the characterization of Lemma 21. A *sequence representation* of a language $L \subseteq \Sigma^*$ is a finite set of triples $(w_i^-, a_i, w_i^+)_{1 \leq i \leq n}$ where $w_i^-, w_i^+ \in \Sigma^{\mathbb{N}} \cup \Sigma^*$ are two potentially infinite words, and $a_i \in \Sigma$ is a letter, such that

$$L = \bigcup_{i=1}^{n} \mathsf{Infixes}(\mathsf{reversed}(w_i^-) a_i w_i^+) \quad .$$

Given an effective representation of sequences, one obtains an effective representation of languages via sequence representations. In this section, we will rely on definitions originating from the area of symbolic dynamics, that precisely study infinite words whose generation follows from a finitely described process.

However, we will not assume that the reader is familiar with this domain, and we will use as black-boxes key results from this area.

A first model that one can use to represent infinite words is the model of *automatic sequences*. In this case, the infinite word w is described by a finite state automaton, that can compute the i-th letter of the word w given as input the number i written in some base $b \in \mathbb{N}$. An example of such a sequence is the Thue-Morse sequence that can be described by a finite automaton using a binary representation of the indices. The good algorithmic properties of automatic sequences come from the fact that a Presburger definable property that uses letters of the sequence can be (trivially) translated into a finite automaton that reads the base b representation of the free variables (that are indices of the sequence). In particular, it follows that one can decide if an automatic sequence is ultimately uniformly recurrent. Based on this, we now prove:

Theorem 27. *Given a sequence representation of a language $L \subseteq \Sigma^*$ where all infinite words are automatic sequences, one can decide whether L is well-quasi-ordered for the infix relation.*

In fact, automatic sequences are part of a larger family of sequences studied in symbolic dynamics, called morphic sequences. Let us first recall that a *morphism* is a function $f\colon \Sigma^* \to \Gamma^*$ such that for every $u, v \in \Sigma^*$, $f(uv) = f(u)f(v)$. A *morphic sequence* w is an infinite word obtained by iterating a morphism $f\colon \Sigma^* \to \Sigma^*$ on a letter $a \in \Sigma$ such that $f(a)$ starts with a, and then applying a homomorphism $h\colon \Sigma^* \to \Gamma^*$. The infinite word $f^\omega(a)$ is the limit of the sequence $(f^n(a))_{n\in\mathbb{N}}$, which is well-defined because $f(a)$ starts with a, and the morphic sequence is $w \triangleq h(f^\omega(a))$.

Every automatic sequence is a morphic sequence, but not the other way around. We refer the reader to a short survey of [3] for more details on the possible variations on the definition of morphic sequences and their relationships. It was relatively recently proven that one can decide whether a morphic sequence is uniformly recurrent [14, Theorem 1]. We were not able to find in the literature whether one can decide ultimate uniform recurrence, but conjecture that it is the case, which would allow us to decide whether a language represented by morphic sequences is well-quasi-ordered for the infix relation.

Conjecture 28. Given a morphic sequence $w \in \Sigma^{\mathbb{N}}$, one can decide whether it is ultimately uniformly recurrent.

6 Infixes and Amalgamation Systems

In the previous section, we have represented languages that are downward closed by the infix relation as infixes of infinite words. However, there are many other natural ways to represent languages, such as finite automata or context-free grammars. In this section, we are going to show that our results on bounded languages can be applied to a large class of systems, called amalgamation systems, that includes as particular examples finite automata and context-free grammars.

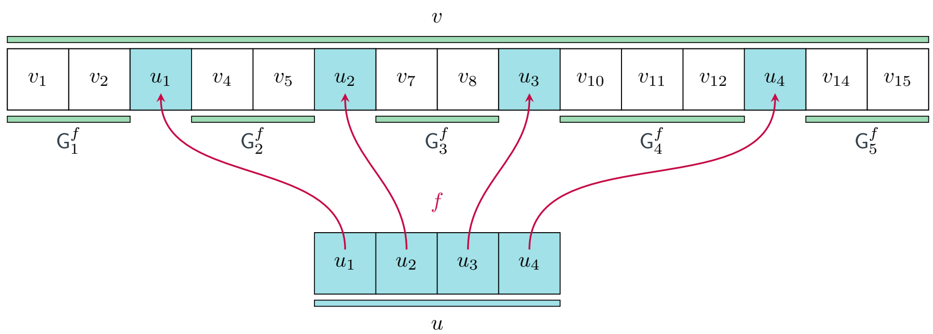

Fig. 3: The gap words resulting from a subword embedding between two finite words.

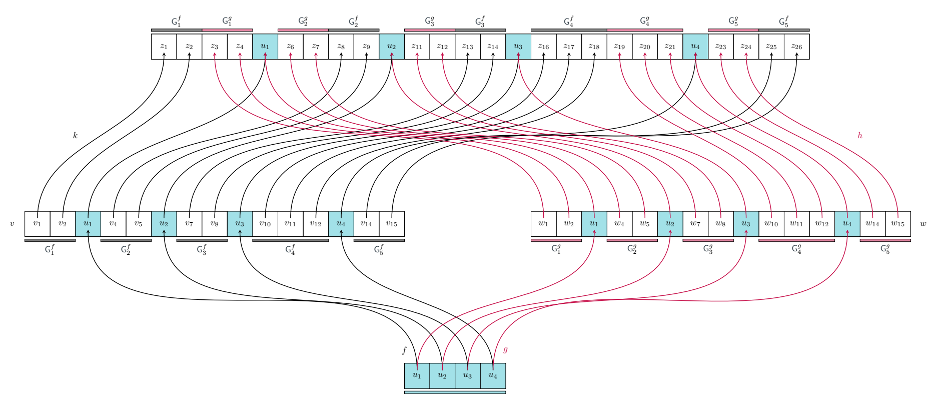

Fig. 4: We illustrate how embeddings f and g between runs of an amalgamation system can be glued together, seen on their canonical decomposition.

Our first result, of theoretical nature, is that amalgamation systems cannot define well-quasi-ordered languages that are not bounded. This implies that all the results of Section 4, and in particular Theorem 8, can safely be applied to amalgamation systems.

Theorem 29. *Let $L \subseteq \Sigma^*$ be a language recognized by an amalgamation system. If L is well-quasi-ordered by the infix relation then L is bounded.*

Our second focus is of practical nature: we want to give a decision procedure for being well-quasi-ordered. This will require us to introduce *effectiveness assumptions* on the amalgamation systems. While most of them will be innocuous, an important consequence is that we have to consider *classes of languages* rather than individual ones, for instance: the class of all regular languages, or the class of all context-free languages. Such classes will be called effective amalgamative classes (Section 6.1). In the following theorem, we prove that under such assumptions, testing well-quasi-ordering is inter-reducible to testing whether a language of the class is empty, which is usually the simplest problem for a computational model.

Theorem 30. *Let $\mathcal{C}$ be an effective amalgamative class of languages. Then the following are equivalent:*

1. *Well-quasi-orderedness of the infix relation is decidable for languages in $\mathcal{C}$.*
2. *Well-quasi-orderedness of the prefix relation is decidable for languages in $\mathcal{C}$.*
3. *Emptiness is decidable for languages in $\mathcal{C}$.*

6.1 Amalgamation Systems

Let us now formally introduce the notion of amalgamation systems, and recall some results from [5] that will be useful for the proof of Theorem 29. The notion of amalgamation system is tailored to produce *pumping arguments*, which is exactly what our Lemma 12 talks about. At the core of a pumping argument, there is a notion of a *run*, which could for instance be a sequence of transitions taken in a finite state automaton. Continuing on the analogy with finite automata, there is a natural ordering between runs, i.e., a run is smaller than another one if one can "delete" loops of the larger run to obtain the other. Typical pumping arguments then rely on the fact that *minimal* runs are of finite size, and that all other runs are obtained by "gluing" loops to minimal runs. Generalizing this notion yields the notion of amalgamation systems.

Let us recall that over an alphabet $(\Sigma, =)$ a subword embedding between two words $u \in \Sigma^*$ and $v \in \Sigma^*$ is an order-preserving function $\rho\colon [1, |u|] \to [1, |v|]$ such that $u_i = v_{\rho(i)}$ for all $i \in [1, |u|]$. We write $\mathsf{Hom}^*(u, v)$ the set of all subword embeddings between u and v. It may be useful to notice that the set of finite words over Σ forms a category when we consider subword embeddings as morphisms, which is a fancy way to state that $\mathrm{id} \in \mathsf{Hom}^*(u, u)$ and that $f \circ g \in \mathsf{Hom}^*(u, w)$ whenever $g \in \mathsf{Hom}^*(u, v)$ and $f \in \mathsf{Hom}^*(v, w)$, for any choice of words $u, v, w \in \Sigma^*$.

Given a subword embedding $f: u \to v$ between two words $u = u_1 \cdots u_k$ and v, there exists a unique decomposition $v = \mathsf{G}_0^f\, u_1\, \mathsf{G}_1^f \cdots \mathsf{G}_{k-1}^f\, u_k\, \mathsf{G}_k^f$ where $\mathsf{G}_i^f = v_{f(i)+1} \cdots v_{f(i+1)-1}$ for all $1 \leq i \leq k-1$, $\mathsf{G}_k^f = v_{f(k)+1} \cdots v_{|v|}$, and $\mathsf{G}_0^f = v_1 \cdots v_{f(1)-1}$. We say that G_i^f is the i-th *gap word* of f. We encourage the reader to look at Figure 3 to see an example of the gap words resulting from a subword embedding between two words. These gap words will be useful to describe how and where runs of a system (described by words) can be combined.

Definition 31. *An* amalgamation system *is a tuple* $(\Sigma, R, \mathsf{can}, E)$ *where* Σ *is a finite alphabet,* R *is a set of so-called* runs, $\mathsf{can}: R \to (\Sigma \uplus \{\#\})^*$ *is a function computing a* canonical decomposition *of a run, and* E *describes the so-called* admissible embeddings *between runs: If* ρ *and* σ *are runs from* R*, then* $E(\rho, \sigma)$ *is a subset of the subword embeddings between* $\mathsf{can}(\rho)$ *and* $\mathsf{can}(\sigma)$*. We write* $\rho \trianglelefteq \sigma$ *if* $E(\rho, \sigma)$ *is non-empty. If we want to refer to a specific embedding* $f \in E(\rho, \sigma)$*, we also write* $\rho \trianglelefteq_f \sigma$*. Given a run* $r \in R$*, and* $i \in [0, |\mathsf{can}(r)|]$*, the* gap language *of* r *at position* i *is* $\mathsf{L}_i^r \triangleq \{\mathsf{G}_i^f \mid \exists s \in R. \exists f \in E(r, s)\}$*. An amalgamation system furthermore satisfies the following properties:*

1. (R, E) Forms a Category. *For all* $\rho, \sigma, \tau \in R$*,* $\mathrm{id} \in E(\rho, \rho)$*, and whenever* $f \in E(\rho, \sigma)$ *and* $g \in E(\sigma, \tau)$*, then* $g \circ f \in E(\rho, \tau)$*.*
2. Well-Quasi-Ordered System. $(R, \trianglelefteq)$ *is a well-quasi-ordered set.*
3. Concatenative Amalgamation. *Let* ρ_0, ρ_1, ρ_2 *be runs with* $\rho_0 \trianglelefteq_f \rho_1$ *and* $\rho_0 \trianglelefteq_g \rho_2$*. Then for all* $0 \leq i \leq |\mathsf{can}(\rho_0)|$*, there exists a run* $\rho_3 \in R$ *and embeddings* $\rho_1 \trianglelefteq_{g'} \rho_3$ *and* $\rho_2 \trianglelefteq_{f'} \rho_3$ *satisfying two conditions: (a)* $g' \circ f = f' \circ g$ *(we write* h *for this composition) and (b) for every* $0 \leq j \leq |\rho_0|$*, the gap word* G_j^h *is either* $\mathsf{G}_j^f\, \mathsf{G}_j^g$ *or* $\mathsf{G}_j^h = \mathsf{G}_j^g\, \mathsf{G}_j^f$*. Specifically, for* i *we may fix* $\mathsf{G}_i^h = \mathsf{G}_i^f\, \mathsf{G}_i^g$*. We refer to Figure 4 for an illustration of this property.*

The yield *of a run is obtained by projecting away the separator symbol* $\#$ *from the canonical decomposition, i.e.* $\mathsf{yield}(\rho) = \pi_\Sigma(\rho)$*. The language recognized by an amalgamation system is* $\mathsf{yield}(R)$*.*

We say a language L *is an* amalgamation language *if there exists an amalgamation system recognizing it.*

Intuitively, the definition of an amalgamation system allows the comparison of runs, and the proper "gluing" of runs together to obtain new runs. A number of well-known language classes can be seen to be recognized by amalgamation systems, e.g., regular languages [5, Theorem 5.3], reachability and coverability languages of VASS [5, Theorem 5.5], and context-free languages [5, Theorem 5.10].

We can now show a simple lemma that illuminates much of the structure of amalgamation systems whose language is well-quasi-ordered by $\sqsubseteq_{\mathsf{infix}}$. Note that Lemma 32 uses Lemma 12 in its proof, and our Theorem 29 follows from it.

Lemma 32. *Let* L *by an amalgamation language recognized by* $(\Sigma, R, E, \mathsf{can})$ *that is well-quasi-ordered by* $\sqsubseteq_{\mathsf{infix}}$*. Let* ρ *be a run with* $\rho = a_1 \cdots a_n$*, and let* σ, τ *be runs with* $\rho \trianglelefteq_f \sigma$ *and* $\rho \trianglelefteq_g \tau$*.*

For any $0 \leq \ell \leq n$*, we have* $\mathsf{G}_\ell^f \sqsubseteq_{\mathsf{infix}} \mathsf{G}_\ell^g$ *or vice versa.*

If we additionally assume that such a language is closed under taking infixes, we obtain an even stronger structure: All such languages are regular!

Lemma 33. *Let $L \subseteq \Sigma^*$ be a downward closed language for the infix relation that is well-quasi-ordered. Then, the following are equivalent:*

(i) *L is a regular language,*
(ii) *L is recognized by* some *amalgamation system,*
(iii) *L is a bounded language,*
(iv) *There exists a finite set $E \subseteq (\Sigma^*)^3$ such that $L = \bigcup_{(x,u,y)\in E} \mathsf{P}{\downarrow}(x) u \, \mathsf{P}{\downarrow}(y)$.*

Combining Lemmas 18 and 33, we can conclude that the collection of infixes of the Thue-Morse sequence cannot be recognized by *any* amalgamation system.

To construct a decision procedure for well-quasi-orderedness under $\sqsubseteq_{\mathsf{infix}}$, we need our amalgamation systems to satisfy certain *effectiveness assumptions*. We require that for an amalgamation system $(\Sigma, R, E, \mathsf{can})$, R is recursively enumerable, the function $\mathsf{can}(\cdot)$ is computable, and for any two runs $\rho, \sigma \in R$, the set $E(\rho, \sigma)$ is computable. Additionally, we require the class to be effectively closed under rational transductions [8, Chapter 5, page 64].

Under these assumptions, one can improve on Theorem 8 into an effective procedure, using pumping arguments from [5, Section 4.2], which, in turn, allows us to prove Theorem 30. Since the class $\mathcal{C}_{\mathrm{aut}}$ of regular languages and the class $\mathcal{C}_{\mathrm{cfg}}$ of context-free languages are examples of effective amalgamative classes, the following corollary is immediate.

Corollary 34. *Let $\mathcal{C} \in \{\mathcal{C}_{aut}, \mathcal{C}_{cfg}\}$. It is decidable whether a language in $\mathcal{C}$ is well-quasi-ordered by the infix relation. Furthermore, whenever it is well-quasi-ordered by the infix relation, it is a bounded language.*

7 Conclusion

We have described the landscapes of well-quasi-ordered languages for the natural orderings on finite words: prefix, suffix, and infix relations. While the prefix and suffix relation exhibit very simple behaviours, the infix relation can encode many complex quasi-orders (and even simulate the subword ordering). In the case of languages that are described by simple computational models, or languages that are "structurally simple" (bounded languages, downward closed languages), we showed that only very simple well-quasi-orders can be obtained: they are essentially isomorphic to disjoint unions of copies of finite sets, $(\mathbb{N}, \leq)$, and $(\mathbb{N}^2, \leq)$. Finally, under effectiveness assumptions on the language (such as being recognized by an amalgamation system, or being the set of infixes of an automatic sequence), we proved the decidability of being well-quasi-ordered for the infix relation. We believe that these very encouraging results pave the way for further research on deciding which sets are well-quasi-ordered for other orderings. Let us now discuss some possible research directions and remarks.

Towards infinite alphabets In this paper, we restricted our attention to *finite* alphabets, having in mind the application to regular languages. However, the conclusions of Theorem 8, Corollary 26, and Theorem 5 could be conjectured to hold in the case of infinite alphabets (themselves equipped with a well-quasi-ordering). This would require new techniques, as the finiteness of the alphabet is crucial to all of our positive results.

Lexicographic orderings There is another natural ordering on words, the *lexicographic ordering*, which does not fit well in our current framework because it is always of ordinal width 1. However, the order-type of the lexicographic ordering over regular languages has already been investigated in the context of infinite words [10], and it would be interesting to see if one can extend these results to decide whether such an ordering is well-founded for languages recognized by amalgamation systems.

Factor Complexity Let us conclude this section with a few remarks on the notion of factor complexity of languages. Recall that the *factor complexity* of a language $L \subseteq \Sigma^*$ is the function $f_L : \mathbb{N} \to \mathbb{N}$ such that $f_L(n)$ is the number of distinct words of size n in L. We extend the notion of factor complexity to finite, infinite, and bi-infinite words as the factor complexity of their set of finite infixes. For the prefix relation and the suffix relation, all well-quasi-ordered languages have a bounded factor complexity, since they are finite unions of chains.

While there clearly are languages with low factor complexity that are not well-quasi-ordered for the infix relation, such as the language $L \triangleq \downarrow ab^*a$; one would expect that languages that are well-quasi-ordered for the infix relation would have a low factor complexity.

In some sense, our results confirm this intuition in the case of languages described by a simple computational model. For languages recognized by amalgamation systems, being well-quasi-ordered implies being a bounded language, and therefore being included in some finite union of languages of the form $w_1^* w_2 w_3^*$. Hence, these languages have at most a linear factor complexity. This is also the case for languages described as the infixes of a finite set of pairs of morphic sequences. Indeed, the factor complexity of a morphic sequence that is uniformly recurrent is linear [28, Theorem 24], therefore the factor complexity of a language given by sequence representation using morphic sequences is at most linear.

However, there are downward closed languages that are well-quasi-ordered for the infix relation but have an exponential factor complexity: the $(5,3)$-Toeplitz word is uniformly recurrent [11, p. 499], and has exponential factor complexity [11, Theorem 5]. This shows that our computational models somehow fail to capture vast classes of well-quasi-ordered languages with a high factor complexity. It would be interesting to understand which new proof techniques would be required to obtain decidability for these languages.

To conclude on a positive note for the infix relation, our results show that for downward closed and well-quasi-ordered languages, there is a strong connection between the factor complexity and the ordinal width: it is the same to have bounded factor complexity and finite ordinal width.

References

1. Abdulla, P.A., Jonsson, B.: Verifying networks of timed processes. In: Proceedings of TACAS'98. vol. 1384, pp. 298–312. Springer (1998). https://doi.org/10.1007/BFb0054179
2. Abdulla, P.A., Čerāns, K., Tsay, B.J., Yih-Kuen: General decidability theorems for infinite-state systems. In: Proceedings of LICS'96. pp. 313–321. IEEE (1996). https://doi.org/10.1109/LICS.1996.561359
3. Allouche, J.P., Cassaigne, J., Shallit, J., Zamboni, L.Q.: A taxonomy of morphic sequences (2017), https://arxiv.org/abs/1711.10807
4. Allouche, J.P., Shallit, J.: The ubiquitous prouhet-thue-morse sequence. Discrete Mathematics and Theoretical Computer Science p. 1–16 (1999). https://doi.org/10.1007/978-1-4471-0551-0_1, http://dx.doi.org/10.1007/978-1-4471-0551-0_1
5. Anand, A., Schmitz, S., Schütze, L., Zetzsche, G.: Verifying unboundedness via amalgamation. In: Proceedings of the 39th Annual ACM/IEEE Symposium on Logic in Computer Science. LICS '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3661814.3662133, https://doi.org/10.1145/3661814.3662133
6. Atminas, A., Lozin, V., Moshkov, M.: Wqo is decidable for factorial languages. Information and Computation **256**, 321–333 (Oct 2017). https://doi.org/10.1016/j.ic.2017.08.001, http://dx.doi.org/10.1016/j.ic.2017.08.001
7. Bergsträßer, P., Ganardi, M., Lin, A.W., Zetzsche, G.: Ramsey quantifiers in linear arithmetics. Proc. ACM Program. Lang. **8**(POPL), 1–32 (2024). https://doi.org/10.1145/3632843, https://doi.org/10.1145/3632843
8. Berstel, J.: Transductions and Context-Free Languages. Vieweg+Teubner Verlag (1979). https://doi.org/10.1007/978-3-663-09367-1, http://dx.doi.org/10.1007/978-3-663-09367-1
9. Bucher, W., Ehrenfeucht, A., Haussler, D.: On total regulators generated by derivation relations. In: International Colloquium on Automata, Languages, and Programming. vol. 40, pp. 71–79 (1985). https://doi.org/10.1016/0304-3975(85)90162-8, https://www.sciencedirect.com/science/article/pii/0304397585901628, eleventh International Colloquium on Automata, Languages and Programming
10. Carton, O., Colcombet, T., Puppis, G.: An algebraic approach to mso-definability on countable linear orderings. The Journal of Symbolic Logic **83**(3), 1147–1189 (2018), https://www.jstor.org/stable/26600366
11. Cassaigne, J., Karhumäki, J.: Toeplitz words, generalized periodicity and periodically iterated morphisms. European Journal of Combinatorics **18**(5), 497–510 (Jul 1997). https://doi.org/10.1006/eujc.1996.0110, http://dx.doi.org/10.1006/eujc.1996.0110
12. Daligault, J., Rao, M., Thomassé, S.: Well-quasi-order of relabel functions. Order **27**(3), 301–315 (September 2010). https://doi.org/10.1007/s11083-010-9174-0, http://dx.doi.org/10.1007/s11083-010-9174-0
13. Demeri, S., Finkel, A., Goubault-Larrecq, J., Schmitz, S., Schnoebelen, P.: Algorithmic aspects of wqo theory (mpri course) (2012), https://cel.archives-ouvertes.fr/cel-00727025, course notes
14. Durand, F.: Decidability of uniform recurrence of morphic sequences. International Journal of Foundations of Computer Science **24**(01), 123–146 (2013). https://doi.org/10.1142/S0129054113500032

15. Džamonja, M., Schmitz, S., Schnoebelen, P.: On Ordinal Invariants in Well Quasi Orders and Finite Antichain Orders, pp. 29–54. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-30229-0_2, https://doi.org/10.1007/978-3-030-30229-0_2
16. D'Alessandro, F., Varricchio, S.: On well quasi-orders on languages, pp. 230–241. Springer Berlin Heidelberg (2003). https://doi.org/10.1007/3-540-45007-6_18
17. D'Alessandro, F., Varricchio, S.: Well quasi-orders, unavoidable sets, and derivation systems. RAIRO - Theoretical Informatics and Applications **40**(3), 407–426 (Jul 2006). https://doi.org/10.1051/ita:2006019, http://dx.doi.org/10.1051/ita:2006019
18. Fine, N.J., Wilf, H.S.: Uniqueness theorems for periodic functions. Proceedings of the American Mathematical Society **16**(1), 109–114 (Feb 1965). https://doi.org/10.1090/s0002-9939-1965-0174934-9, http://dx.doi.org/10.1090/S0002-9939-1965-0174934-9
19. Finkel, A., Gupta, E.: The well structured problem for presburger counter machines. In: Chattopadhyay, A., Gastin, P. (eds.) 39th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2019, December 11-13, 2019, Bombay, India. LIPIcs, vol. 150, pp. 41:1–41:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2019). https://doi.org/10.4230/LIPICS.FSTTCS.2019.41, https://doi.org/10.4230/LIPIcs.FSTTCS.2019.41
20. Higman, G.: Ordering by divisibility in abstract algebras. Proceedings of the London Mathematical Society **3**, 326–336 (1952). https://doi.org/10.1112/plms/s3-2.1.326
21. de Jongh, D., Parikh, R.: Well-partial orderings and hierarchies. Indagationes Mathematicae (Proceedings) **80**(3), 195–207 (1977). https://doi.org/10.1016/1385-7258(77)90067-1, http://dx.doi.org/10.1016/1385-7258(77)90067-1
22. Krivine, J.L.: Introduction to Axiomatic Set Theory. Springer Netherlands (1971). https://doi.org/10.1007/978-94-010-3144-8
23. Kruskal, J.B.: The theory of well-quasi-ordering: A frequently discovered concept. Journal of Combinatorial Theory, Series A **13**(3), 297–305 (Nov 1972). https://doi.org/10.1016/0097-3165(72)90063-5, http://dx.doi.org/10.1016/0097-3165(72)90063-5
24. Kunen, K.: Set Theory. Elsevier (1980). https://doi.org/10.1016/s0049-237x(08)x7037-5, http://dx.doi.org/10.1016/s0049-237x(08)x7037-5
25. Kuske, D.: Theories of orders on the set of words. RAIRO Theor. Informatics Appl. **40**(1), 53–74 (2006). https://doi.org/10.1051/ITA:2005039, https://doi.org/10.1051/ita:2005039
26. Kříž, I., Thomas, R.: Ordinal Types in Ramsey Theory and Well-Partial-Ordering Theory, p. 57–95. Springer Berlin Heidelberg (1990). https://doi.org/10.1007/978-3-642-72905-8_7, http://dx.doi.org/10.1007/978-3-642-72905-8_7
27. Lopez, A.: Labelled Well Quasi Ordered Classes of Bounded Linear Clique-Width. In: Gawrychowski, P., Mazowiecki, F., Skrzypczak, M. (eds.) 50th International Symposium on Mathematical Foundations of Computer Science (MFCS 2025). Leibniz International Proceedings in Informatics (LIPIcs), vol. 345, pp. 70:1–70:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2025). https://doi.org/10.4230/LIPIcs.MFCS.2025.70, https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.MFCS.2025.70
28. Nicolas, F., Pritykin, Y.: On uniformly recurrent morphic sequences. International Journal of Foundations of Computer Science **20**(05), 919–940 (Oct 2009). https://doi.org/10.1142/s0129054109006966, http://dx.doi.org/10.1142/S0129054109006966

29. Rabin, M.O.: Decidability of second-order theories and automata on infinite trees. Transactions of the American Mathematical Society **141**, 1–35 (1969), http://www.jstor.org/stable/1995086
30. Schmidt, D.: The relation between the height of a well-founded partial ordering and the order types of its chains and antichains. Journal of Combinatorial Theory, Series B **31**(2), 183–189 (Oct 1981). https://doi.org/10.1016/s0095-8956(81)80023-8, http://dx.doi.org/10.1016/S0095-8956(81)80023-8
31. Schmitz, S.: Algorithmic Complexity of Well-Quasi-Orders. Habilitation à diriger des recherches, École normale supérieure Paris-Saclay (Nov 2017), https://theses.hal.science/tel-01663266
32. Schmitz, S.: The parametric complexity of lossy counter machines. In: Baier, C., Chatzigiannakis, I., Flocchini, P., Leonardi, S. (eds.) 46th International Colloquium on Automata, Languages, and Programming (ICALP 2019). Leibniz International Proceedings in Informatics (LIPIcs), vol. 132, pp. 129:1–129:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2019). https://doi.org/10.4230/LIPIcs.ICALP.2019.129, https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2019.129
33. Thomas, W.: Languages, automata, and logic. In: Rozenberg, G., Salomaa, A. (eds.) Handbook of formal languages, pp. 389–455. Springer (1997). https://doi.org/10.1007/978-3-642-59136-5
34. Tromp, J., Shallit, J.: Subword complexity of a generalized thue-morse word. Information Processing Letters **54**(6), 313–316 (Jun 1995). https://doi.org/10.1016/0020-0190(95)00074-m, http://dx.doi.org/10.1016/0020-0190(95)00074-M