Well-quasi-orderings on word languages main 81b14ea82bdf369ba5dbe8de29fa418756900125 2025-10-15 17:43:32 +0200

Anonymized for review

Anonymized for review

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 decition procedures when the languages are given by various comupational models such as automata, context-free grammars, and automatic structures.

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 was successfully applied to the verification of infinite state transition systems [**?**,**?**]. 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 [**?**]. Perhaps more surprisingly, the class of well-quasi-ordered sets is also stable under the operation of taking finite words and finite trees labelled by elements of a well-quasi-ordered set [**?**,**?**].

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 an ordered alphabet $(X, \preceq)$ is well-quasi-ordered by the so-called subword embedding relation [**?**]. 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 exists w_1, w_2 such that $w_1 u w_2 = v$). Note that these three relations straightforwardly generalize to infinite quasi-ordered alphabets. Unfortunately, it is easy to see that none of these constructions are well-quasi-ordered as soon as the alphabet contains two distinct letters: for instance, the infinite sequence $ab^n a$ 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 has 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 [**?**] or in the context of logic [**?**]. 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 [**?**,**?**,**?**]. Furthermore, a previous work by Kuske shows that any *reasonable*[1] partially ordered set $(X, \leq)$ can be embedded into $\{a, b\}^*$ with the infix relation [**?**, 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 [**?**]. One way to measure this complexity is to consider its so-called ordinal invariants: for instance, the maximal order type (or m.o.t.), originally defined by De Jongh and Parikh [**?**], 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 [**?**]: the ordinal height introduced by Schmidt [**?**], and the ordinal width of Kříž and Thomas [**?**]. 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 well-quasi-orderings. We refer to **??** for a more detailed introduction to these parameters and ordinal computations in general.

[1] This will be made precise in **??**.

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 (**??**). A summary of these results can be found in **??**.

L	**Characterisation**	$\mathfrak{w}(L)$	$\mathfrak{o}(L)$
arbitrary	**??**: finite unions of chains	$< \omega$	$< \omega^2$
regular	**??**: 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 [**?**] imply 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 downwards 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 **??**. We furthermore notice that for downwards closed languages that are well-quasi-ordered by the infix relation, being bounded is the same as being regular (**??**), 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 (**??**). 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 downwards closed languages and to bounded languages. For downwards closed languages, we consider a model based on representations of infinite words (**??**), for which we provide a decision procedure (**??**). The model used to represent these infinite words is based on automatic sequences and morphic sequences [**?**], which are well-studied in the context of symbolic dynamics. For bounded languages, we consider the model of amalgamation systems [**?**], which is an abstract computational model that encompasses many classical ones, such as finite automata, context-free grammars, and Petri nets [**?**]. We show that if a language recognized by an amalgamation system is

L	**Characterisation**	$\mathfrak{w}(L)$	$\mathfrak{o}(L)$
arbitrary	**??**: countable well-quasi orders with finite initial segments	$< \omega_1$	$< \omega_1$
bounded	**??**: finite union of products of chains for the prefix and suffix relations	$< \omega^2$	$< \omega^3$
downwards closed	**??**: 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 **??** and **??**.

well-quasi-ordered by the infix relation, then it is a bounded language (**??**), 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 (**??**). We defer the introduction of amalgamation systems to **??**.

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* where already analysed by Bucher, Ehrenfeucht, and Haussler [**?**], and were later extended by D'Alessandro and Varricchio [**?**,**?**]. However, in all those cases the orderings are *multiplicative*, that is, if $u_1 \preceq v_1$ and $u_2 \preceq v_2$ then $u_1u_2 \preceq v_1v_2$. This assumption does not hold for the prefix, suffix, and infix relations.

A similar question was studied by Atminas, Lozin, and Moshkov [**?**], in the hope of finding characterizations of classes of *finite graphs* that are well-quasi-ordered by the *induced subgraph relation* [**?**, 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* [**?**, 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 [**?**, Theorem 1]. Because it is easy to transform a list of forbidden factors into a regular language [**?**, 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 **??** the necessary background on well-quasi-orders and ordinal invariants. In **??**, which is relatively self-contained, we study the prefix relation and prove in **??** the characterization of well-quasi-ordered languages by the prefix relation. In **??**, we obtain the infix analogue of **??** specifically for bounded languages (**??**). In **??**, we study the downwards closed languages, characterize them using a notion of ultimately uniformly recurrent words borrowed from symbolic dynamics (**??**), and compute bounds on their ordinal invariants in **??**. Finally, we generalize these results to all amalgamation systems in **??** in (**??**), 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 (**??**).

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 [**?**] 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 now us 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 [**?**] 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, 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 *downwards 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 [?] and Krivine [?, 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 $(T_i)_{n\in\mathbb{N}}$ then $\alpha_T = \sup\{\alpha_{T_i} + 1 \mid i \in \mathbb{N}\}$. 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 rank of these trees are called respectively the *maximal order type* of X written $\mathfrak{o}(X)$ [?], the *ordinal height* of X written $\mathfrak{h}(X)$ [?], and the *ordinal width* of X written $\mathfrak{w}(X)$ [?]. 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 [?] for a detail 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 [?], and that was recalled in [?, 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 [?, 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 $a^k \sqsubseteq_{\mathsf{pref}} a^k b$ 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 **??** 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 [**?**, Theorem 1.1]. This result will follow from our general decidability result (**??**) 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 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, maximal order type of L 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 **??** 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 $x = y$. A simplified version of the embedding defined in **??** is illustrated for the subword relation in **??** page **??**.

Lemma 7. *[?, 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 **??**, we cannot replay proofs of **??**, 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.

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 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 **??**.

In order to prove the (difficult) left-to-right implication of **??**, 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 **??**). 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 $L \triangleq \{\prod_{i=0}^{n} w_i \mid i \in \mathbb{N}\}$. 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 simultaneously.*

The following combinatorial **??** 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 the upcoming **????**. It is fundamentally based on a classical result of combinatorics on words (**??**) that we recall here for the sake of completeness.

Lemma 11 ([?, 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 arbitrarily large values of $\boldsymbol{n} \in \mathbb{N}^k$. Then, there exists $x, y \in \Sigma^+$ 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:*

1. *$w[\boldsymbol{n}] \in u_1 \,\mathsf{P}{\downarrow}(x)$,*
2. *$w[\boldsymbol{n}] \in \mathsf{P}{\downarrow}(x) u_{k+1}$,*
3. *$w[\boldsymbol{n}] \in \mathsf{P}{\downarrow}(x) u_i \,\mathsf{P}{\downarrow}(y)$ 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:*

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

*Proof (Proof of **??** as stated on page **??**).* We apply **??**, and conclude because $\mathsf{P}{\downarrow}(x)$ is a finite union of chains for the prefix, suffix and infix relations (**??**). ▷ Back to p.??

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 less than ω^2, its ordinal height is at most ω, and its maximal order type less than ω^3. Furthermore, those three bounds are tight.*

Proof. Upper bounds are a direct consequence of **??**, and the tightness is witnessed by the languages: $L_k \triangleq \bigcup_{i=2}^{k+1}(ab^i a)^*(ba^i b)^*$, that are bounded languages of $\{a,b\}^*$, well-quasi-ordered by the infix relation, and have ordinal width, ordinal height and maximal order type respectively equal to $\omega \cdot k$, ω and $\omega^2 \cdot k$.

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 downwards closed. Indeed, the **??** crucially relies on constructing languages that are *not* downwards closed, and we have shown in **??** that the downwards closure of a well-quasi-ordered language is not necessarily well-quasi-ordered.

5.1 Characterization of Well-Quasi-Ordered Downwards Closed Languages

An immediate consequence of **??** 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 **??**.

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.*

The **??** 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 downwards 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 $k \geq 1$.

To strengthen **??**, 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 [**?**].

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 downwards closed for the infix relation, well-quasi-ordered for the infix relation, but is not bounded.*

Proof. By construction $I_{\mathbf{t}}$ is downwards closed for the infix relation, and by **??**, 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 downwards 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 measure, 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 **??**, using for instance the results of [**?**] or [**?**, 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 [**?**], 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 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 **??** that the status of the Thue-Morse sequence is actually representative of downwards 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.

Theorem 20. *Let L be a well-quasi-ordered language for the infix relation that is downwards 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 $L = \bigcup_{i=1}^{n} \mathsf{Infixes}(w_i)$.*

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

Corollary 21. *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.

To connect infixes of a (bi)-infinite word to downwards 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 downwards 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-infinte word $w \in \Sigma^* \cup \Sigma^{\mathbb{N}} \cup \Sigma^{\mathbb{Z}}$ (**??**).

Lemma 22. *Let $L \subseteq \Sigma^*$ be an order ideal for a well-quasi-ordered infix relation. Then L is the set of finite infixes of a finite, infinite or bi-infinite word w.*

Lemma 23. *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 24. *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 **??**.

*Proof (Proof of **??** as stated on page **??**).* 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 **??**.

Conversely, let us consider a well-quasi-ordered language L that is downwards 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 **??**, 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 **??**, and conclude that w_i is ultimately uniformly recurrent.

▷ Back to p.??

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 **??** naturally follow (the proof is in **??** page **??**).

Lemma 25. *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 26. *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_+ and w_- are two ultimately uniformly recurrent words. In this case, the ordinal width of the set of finite infixes of w is less than $\omega \cdot 3$, and this bound is 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 **??**. 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 than one can decide one can decide if an automatic sequence is ultimately uniformly recurrent, a proof of this folklore result can be found in the appendix at **??**. 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.*

Proof. It is easy to see that L is well-quasi-ordered for the infix relation if and only if for every triple (w_i^-, a_i, w_i^+) in the sequence representation of L, the (potentially bi-infinite) word $\mathsf{reversed}(w_i^-)a_i w_i^+$ defines a well-quasi-ordered language. By **??**, this is the case if and only if both w_i^- and w_i^+ are ultimately uniformly recurrent. Since one can decide whether an automatic sequence is ultimately uniformly recurrent using **??**, we conclude the proof.

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 [?] 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 [?, 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 downwards 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.

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 **??**, and in particular **??**, 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 language, or the class of all context-free languages. Such classes will be called effective amalgamative classes (**??**). 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 [**?**] that will be useful for the proof of **??**. The notion of amalgamation system is tailored to produce *pumping arguments*, which is exactly what our **??** 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 a 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 $f\colon u \to v$ between two words u 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 $1 \le i \le k-1$, $\mathsf{G}_k^f = v_{[f(k)+1} \cdots v_{|v|}$, 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 **??** 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*, $\mathsf{can}\colon R \to (\Sigma \uplus \{\#\})^*$ 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* **??** *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 [**?**, Theorem 5.3], reachability and coverability languages of VASS [**?**, Theorem 5.5], and context-free languages [**?**, 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 **??** uses **??** in its proof, and our **??** 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* $\rho \trianglelefteq_g \sigma$*.*

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 downwards 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 ????, 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 [?, Chapter 5, page 64].

Under these assumptions, one can transform the inclusion test of ?? of ?? into an effective procedure, using pumping arguments from [?, Section 4.2], which, in turn, allows us to prove ??. 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, downwards 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 ??, ??, and ?? 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.

Monoid equations It could be interesting to understand which monoids M recognize languages that are well-quasi-ordered by the infix, prefix or suffix relations. This research direction is connected to finding which classes of graphs of *bounded clique-width* are well-quasi-ordered with respect to the *induced subgraph relation*, as shown in [?], and recently revisited in [?].

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 [**?**], 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.

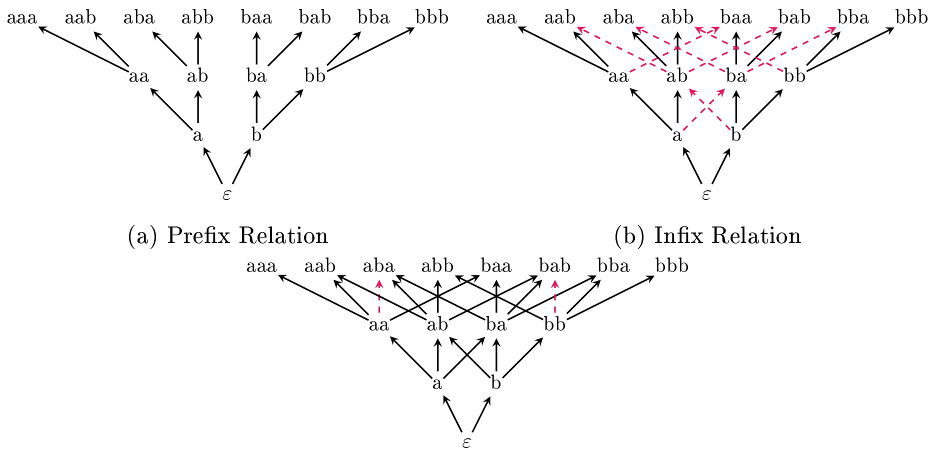
Factor Complexity Let us conclude this section with a few remarks on the notion of factor complexity of languages. Recall that the *factor complexity* of a language $L \subseteq \Sigma^*$ is the function $f_L : \mathbb{N} \to \mathbb{N}$ such that $f_L(n)$ is the number of distinct words of size n in 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 quadratic 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 [**?**, Theorem 24], therefore the factor complexity of a language given by sequence representation using morphic sequences is at most quadratic.

However, there are downwards 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 [**?**, p. 499], and has exponential factor complexity [**?**, 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.

A Proofs for Section ??



(a) Prefix Relation (b) Infix Relation

(c) Subword Relation

Fig. 3: A simple representation of the subword relation, prefix relation, and infix relation, on the alphabet $\{a, b\}$ for words of length at most 3. The figures are Hasse Diagrams, representing the successor relation of the order. Furthermore, we highlight in dashed red relations that are added when moving from the prefix relation to the infix one, and to the infix relation to the subword one.

B Proofs for Section ??

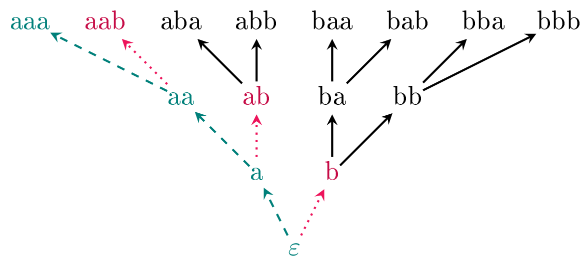


Fig. 4: An antichain branch for the language a^*b, represented in the tree of prefixes over the alphabet $\{a, b\}$. The branch is represented with dashed lines in turquoise, and the antichain is represented in dotted lines in blood-red.

Proof (Proof of **??** *as stated on page* **??***).* Assume that L contains an antichain branch. Let us construct an infinite antichain as follows. We start with a set $A_0 \triangleq \emptyset$ and a node v_0 at the root of the tree. At step i, we consider a word w_i such that v_i is a prefix of w_i, and $w_i \in L \setminus B$, which exists by definition of antichain branches. We then set $A_{i+1} \triangleq A_i \cup \{w_i\}$. To compute v_{i+1}, we consider the largest prefix of w_i that belongs to B, and set v_{i+1} to be the successor of this prefix in B. By an immediate induction, we conclude that for all $i \in \mathbb{N}$, A_i is an antichain, and that v_i is a node in the antichain branch B such that v_i is not a prefix of any word in A_i.

Conversely, assume that L contains an infinite antichain A. Let us construct an antichain branch. Let us consider the subtree of the tree of prefixes that consists in words that are prefixes of words in A. This subtree is infinite, and by König's lemma, it contains an infinite branch. By definition this is an antichain branch.

▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).* If L is regular, then it is MSO-definable, and there exists a formula $\varphi(x)$ in MSO that selects nodes of the tree of prefixes that belong to L. Now, to decide whether there exists an antichain branch for L, we can simply check whether the following formula is satisfied:

$$\exists B.B \text{ is a branch } \wedge \forall x \in B, \exists y.y \text{ is a child of } x \wedge \varphi(y) \wedge y \notin B \quad .$$

Because the above formula is an MSO-formula over the infinite Σ-branching tree, whether it is satisfied is decidable as an easy consequence of the decidability of MSO over infinite binary trees [**?**, Theorem 1.1].

▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).* Assume that L is a finite union of chains. Because the prefix relation is well-founded, and that finite unions of chains have finite antichains, we conclude that L is well-quasi-ordered.

Conversely, assume that L is well-quasi-ordered by the prefix relation. Let us define S_{split} the set of words $w \in \Sigma^*$ such that there exists two words wu and wv both in L that are not comparable for the prefix relation. Let $S = S_{\text{split}} \cup \min_{\sqsubseteq_{\mathsf{pref}}} L$ Assume by contradiction that S is infinite. Then, S equipped with the prefix relation is an infinite tree with finite branching, and therefore contains an infinite branch, which is by definition an antichain branch for L. This contradicts the assumption that L is well-quasi-ordered.

Now, let w be a maximal element for the prefix ordering in S. The upward closure of w in L, $(\uparrow_{\sqsubseteq_{\mathsf{pref}}} w) \cap L$, must be a finite union of chains. Otherwise at least two of the chains would share a common prefix in $w\Sigma$, contradicting the maximality of w.

In particular, letting $S_{\max}$ be the set of all maximal elements of S, we conclude that

$$L \subseteq S \cup \bigcup_{w \in S_{\max}} (\uparrow_{\sqsubseteq_{\mathsf{pref}}} w) \cap L \quad .$$

▷ Back to p.??

Hence, that L is finite union of chains.

Proof (Proof of **??** *as stated on page* **??***).* If L is regular, then it is MSO-definable, and there exists a formula $\varphi(x)$ in MSO that selects nodes of the tree of prefixes that belong to L. Now, to decide whether there exists an antichain branch for L, we can simply check whether the following formula is satisfied:

$$\exists B.B \text{ is a branch } \wedge \forall x \in B, \exists y.y \text{ is a child of } x \wedge \varphi(y) \wedge y \notin B \quad .$$

Because the above formula is an MSO-formula over the infinite Σ-branching tree, whether it is satisfied is decidable as an easy consequence of the decidability of MSO over infinite binary trees [**?**, Theorem 1.1].

▷ Back to p.??

C Proofs for Section ??

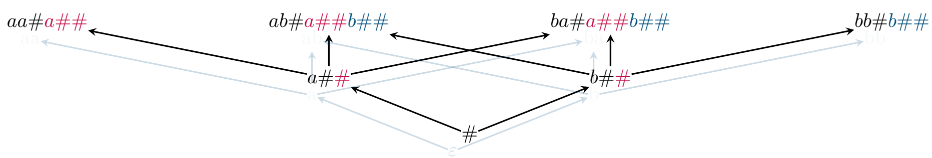


Fig. 5: Representation of the subword relation for $\{a, b\}^*$ inside the infix relation for $\{a, b, \#\}^*$ using a simplified version of **??**, restricted to words of length at most 3.

*Proof (Proof of **??** as stated on page **??**).* Let $x \in \Sigma^+$ be a word, and let P_x be the (finite) set of all prefixes of x, and S_x be the (finite) set of all suffixes of x. Assume that $w \in \mathsf{P}{\downarrow}(x)$, then $w = ux^p v$ for some $u \in S_x$, $v \in P_x$, and $p \in \mathbb{N}$. We have proven that

$$\mathsf{P}{\downarrow}(x) \subseteq \bigcup_{u \in P_x} \bigcup_{v \in S_x} ux^* v \quad .$$

Let us now demonstrate that for all $(u, v) \in S_x \times P_x$, the language ux^*v is a chain for the infix, suffix and prefix relations. To that end, let $(u, v) \in S_x \times P_x$ and $\ell, k \in \mathbb{N}$ be such that $\ell < k$, let us prove that $ux^\ell v \sqsubseteq_{\mathsf{infix}} ux^k v$. Because $v \sqsubseteq_{\mathsf{pref}} x$, we know that there exists w such that $vw = x$. In particular, $ux^\ell vw = ux^{\ell+1}$, and because $\ell < k$, we conclude that $ux^{\ell+1} \sqsubseteq_{\mathsf{pref}} ux^k v$. By transitivity, $ux^\ell v \sqsubseteq_{\mathsf{pref}} ux^k v$, and *a fortiori*, $ux^\ell v \sqsubseteq_{\mathsf{infix}} ux^k v$. Similarly, because $u \sqsubseteq_{\mathsf{suff}} x$, there exists w such that $wu = x$, and we conclude that $ux^\ell v \sqsubseteq_{\mathsf{suff}} wux^\ell v = x^{\ell+1} v \sqsubseteq_{\mathsf{suff}} ux^k v$.

▷ Back to p.??

*Proof (Proof of **??** as stated on page **??**).* Note that the result is obvious if $k = 0$, and therefore we assume $k \geq 1$ in the following proof.

Let us construct a sequence of words $(w_i)_{i \in \mathbb{N}}$, where $w_i \triangleq w[\boldsymbol{n_i}]$ for some well-chosen indices $\boldsymbol{n_i} \in \mathbb{N}^k$. The goal being that if $w[\boldsymbol{n_i}]$ is an infix of $w[\boldsymbol{n_j}]$, then it can intersect at most *two* iterated words, with an intersection that is long enough to successfully apply **??**. In order to achieve this, let us first define s as the maximal size of a word v_i $(1 \leq i \leq k)$ and u_j $(1 \leq j \leq k+1)$. Then, we consider $\boldsymbol{n_0} \in \mathbb{N}^k$ such that $\boldsymbol{n_0}$ has all its components greater than $s!$ and such that $w[\boldsymbol{n_0}]$ belongs to L. Then, we inductively define $\boldsymbol{n_{i+1}}$ as the smallest vector of numbers greater than $\boldsymbol{n_i}$, such that $w[\boldsymbol{n_{i+1}}]$ belongs to L, and with $\boldsymbol{n_i}$ having all components greater than $2\,|w[\boldsymbol{n_i}]|$.

Let us assume that $k \geq 2$ in the following proof for symmetry purposes, and argue later on that when $k = 1$ the same argument goes through. Because L is well-quasi-ordered by the infix relation, there exists $i < j$ such that $w[\boldsymbol{n_i}]$ is an infix of $w[\boldsymbol{n_j}]$. Now, because of the chosen values for $\boldsymbol{n_j}$, there exists $1 \leq \ell \leq k-1$ such that one of the three following equations holds:

- $w[\boldsymbol{n_i}] \sqsubseteq_{\mathsf{infix}} v_\ell^{n_{j,\ell}} u_{\ell+1} v_{\ell+1}^{n_{j,\ell+1}}$,
- $w[\boldsymbol{n_i}] \sqsubseteq_{\mathsf{infix}} u_\ell v_\ell^{n_{j,\ell}}$,
- $w[\boldsymbol{n_i}] \sqsubseteq_{\mathsf{infix}} v_\ell^{n_{j,\ell}} u_{\ell+1}$.

In the sake of simplicity, we will only consider one of the three cases, namely $w[\boldsymbol{n_i}] \sqsubseteq_{\mathsf{infix}} v_\ell^{n_{j,\ell}} u_{\ell+1}$, the other two being similar. Because the lengths used in $\boldsymbol{n_i}$ are all sufficiently large, we know that for every k, $v_k^{n_{i,k}}$ is an infix of a v_ℓ^p for some non-zero p. Therefore, we can apply **??** to conclude that there exists a word x such that every v_k is a power of a conjugate of x (a cyclic shift of x), and v_ℓ is a power of x. We can therefore rewrite $w[\boldsymbol{n_i}]$ as $u_1(\sigma_1(x))^{n_{i,1}} u_2 \cdots$, where σ_k is some conjugacy operation (cyclic shift). Now, in order for $w[\boldsymbol{n_i}]$ to be an infix of $x^{p \times n_{j,\ell}} u_{\ell+1}$, we must conclude that all the u_k's are suffixes or prefixes of x, and that they align properly with the $\sigma_k(x)$'s to form an infix of some power of x, except for the last one. In particular, $w[\boldsymbol{n_i}] \in \mathsf{P}{\downarrow}(x) u_{\ell+1}$, but also, every other choice of $\boldsymbol{n}$ will lead to a word in $\mathsf{P}{\downarrow}(x) u_{\ell+1}$, because the alignment constraints are stable under pumping.

In the case of two iterated words, the reasoning is similar, distinguishing between the v_i's that are occurring before and after the junction of the two iterated words.

When $k = 1$, the situation is a bit more specific since we only have two cases: either $w_i \sqsubseteq_{\mathsf{infix}} u_1 v_1^{n_j}$ or $w_i \sqsubseteq_{\mathsf{infix}} v_1^{n_j} u_2$, and we conclude with an identical reasoning. ▷ Back to p.??

*Proof (Proof of **??** as stated on page **??**).* Let $w_1, \ldots, w_n$ be such that $L \subseteq w_1^* \cdots w_n^*$. Let us define $m \triangleq \max\{|w_i| \mid 1 \leq i \leq n\}$

Let $w[\boldsymbol{k}] \triangleq w_1^{k_1} \cdots w_n^{k_n}$ be a map from $\mathbb{N}^k$ to Σ^*. We are interested in the intersection of the image of w with L. Let us assume for instance that for all $\boldsymbol{k} \in \mathbb{N}^n$, there exists $\boldsymbol{\ell} \geq \boldsymbol{k}$ such that $w[\boldsymbol{\ell}] \in L$. Then, leveraging **??**, we conclude that there exists x, y of size at most $\max\{|w_i| \mid 1 \leq i \leq n\}$ such that $w[\boldsymbol{k}] \in \mathsf{P}{\downarrow}(x) \cup \mathsf{P}{\downarrow}(x)\,\mathsf{P}{\downarrow}(y)$, and we conclude that $L \subseteq \mathsf{P}{\downarrow}(x) \cup \mathsf{P}{\downarrow}(x)\,\mathsf{P}{\downarrow}(y)$.

Now, it may be the case that one cannot simultaneously assume that two component of the vector $\boldsymbol{k}$ are unbounded. In general, given a set $S \subseteq \{1, \ldots, n\}$ of indices, we say that S is admissible if there exists a bound N_0 such that for all $\boldsymbol{b} \in \mathbb{N}^S$, there exists a vector $\boldsymbol{k} \in \mathbb{N}^n$, such that $\boldsymbol{k}$ is greater than $\boldsymbol{b}$ on the S components, and the other components are below the bound N_0. The language of an admissible set S is the set of words obtained by repeating w_i at most N_0 times if it is not in S ($w_i^{\leq N_0}$) and arbitrarily many times otherwise (w_i^*). Note that $L \subseteq \bigcup_{S \text{ admissible}} L(S)$.

Now, admissible languages are ready to be pumped according to **??**. For every admissible language, the size of a word that is not iterated is at most $N_0 \times m$ by definition, and we conclude that:

$$L \subseteq \bigcup_{x,y \in \Sigma^{\leq n}} \bigcup_{u \in \Sigma^{\leq m \times N_0}} \mathsf{P}{\downarrow}(x) u \,\mathsf{P}{\downarrow}(y) \cup \mathsf{P}{\downarrow}(x) u \cup u \,\mathsf{P}{\downarrow}(x) \quad . \tag{1}$$

▷ Back to p.??

D Proofs for Section ??

Proof (Proof of **??** *as stated on page* **??***).* Because $L \subseteq {\downarrow_{\sqsubseteq_{\mathsf{infix}}}} L$, the right-to-left implication is trivial. For the left-to-right implication, let us assume that L is a well-quasi-ordered language for the infix relation. Then L is included in a finite union of products of chains for the prefix and suffix relations thanks to **??**:

$$L \subseteq \bigcup_{i=1}^{n} S_i \cdot P_i \quad .$$

Remark that if S_i is a chain for the suffix relation and P_i is a chain for the prefix relation, then

$${\downarrow_{\sqsubseteq_{\mathsf{infix}}}} (S_i \cdot P_i) = ({\downarrow_{\sqsubseteq_{\mathsf{suff}}}} S_i) \cdot ({\downarrow_{\sqsubseteq_{\mathsf{pref}}}} P_i) \quad .$$

Indeed, any infix of a word in $S_i P_i$ can be split into a suffix of a word in S_i and a prefix of a word in P_i. Conversely, any such concatenations are infixes of a word in $S_i P_i$.

As a consequence, we conclude that ${\downarrow_{\sqsubseteq_{\mathsf{infix}}}} L$ is itself included in a finite union of products of chains. Furthermore, by definition of bounded languages, ${\downarrow_{\sqsubseteq_{\mathsf{infix}}}} L$ is also a bounded language. Hence, it is well-quasi-ordered by the infix relation via **??**.

▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).* Let us assume that L is infinite. The case when it is finite is similar, but will result in a finite word.

aliaume: do we need wqo here? the proof should go through without it: the sequence u_i is already increasing for infix

Because the alphabet Σ is finite, we can enumerate the words of L as $(w_i)_{i \in \mathbb{N}}$. From $(w_i)_{i \in \mathbb{N}}$, we construct a sequence $(u_i)_{i \in \mathbb{N}}$ by induction as follows: $u_0 = w_0$, and u_{i+1} is a word that contains u_i and w_i, which exists in L because L is directed. Since L is well-quasi-ordered, one can extract an infinite set of indices $I \subseteq \mathbb{N}$ such that $u_i \sqsubseteq_{\mathsf{infix}} u_j$ for all $i \leq j \in I$.

We can build a word w as the limit of the sequence $(u_i)_{i \in I}$. This word is infinite or bi-infinite, and contains as infixes all the words u_i for $i \in I$. Because every word of L is an infix of every u_i for a large enough I, one concludes that L is contained in the set of finite infixes of w. Conversely, every finite infix of w is an infix of some u_i by definition of the limit construction, hence belongs to L since $u_i \in L$ and L is downwards closed.

▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).*

Assume that w is ultimately uniformly recurrent. Consider a sequence of words $(w_i)_{i \in \mathbb{N}}$ that are finite infixes of w. Because w is ultimately uniformly recurrent, there exists a bound N_0 such that $w_{\geq N_0}$ is uniformly recurrent. Let $i < N_0$, we claim that, without loss of generality, only finitely many words in the sequence $(w_i)_{i \in \mathbb{N}}$ can be found starting at the position i in w. Indeed, if it is not the case, then we have an infinite subsequence of words that are all comparable for the infix relation, and therefore a good sequence, because the infix relation is well-founded. We can therefore assume that all words in the sequence $(w_i)_{i \in \mathbb{N}}$ are such that they start at a position $i \geq N_0$. But then they are all finite infixes of $w_{\geq N_0}$, which is a uniformly recurrent word, whose set of finite infixes is well-quasi-ordered (**??**).

Conversely, assume that the set of finite infixes of w is well-quasi-ordered. Let us write $\mathsf{Rec}(w)$ the set of finite infixes of w that appear infinitely often. We can similarly define $\mathsf{Rec}(w_{\geq i})$ for any (infinite) suffix of w. The sequence $R_i \triangleq \mathsf{Rec}(w_{\geq i})$ is a descending sequence of downwards closed sets of finite words, included in the set of finite infixes of w by definition. Because the latter is well-quasi-ordered, there exists an $N_0 \in \mathbb{N}$, such that $\bigcap_{i \in \mathbb{N}} R_i = R_{N_0}$. Now, consider $v \triangleq w_{\geq N_0}$. By construction, every finite infix of v appears infinitely often in v. Given some finite infix $u \sqsubseteq_{\mathsf{infix}} v$, we there exists a bound N_u on the distance between two consecutive occurrences of u in v. Indeed, if it is not the case, then there exists an infinite sequence $(ux_iu)_{i \in \mathbb{N}}$ of infixes of v, such that x_i is a word of size $\geq i$ and no shorter word uyu is an infix of ux_iu. Because the finite infixes of w (hence, of v) are well-quasi-ordered, one can extract an infinite set of indices $I \subseteq \mathbb{N}$ such that $ux_iu \sqsubseteq_{\mathsf{infix}} ux_ju$ for all $i \leq j \in I$. In particular, $ux_iu \sqsubseteq_{\mathsf{infix}} ux_ju$ for some $j > |x_i|$, which contradicts the fact that ux_ju coded two consecutive occurrences of u in v.

We have shown that for every finite infix u of v, there exists a bound N_u such that every two occurrences of u in v start at distance at most N_u. In particular, there exists a bound M_u such that every infix of v of size at least M_u contains u. We have proven that v is uniformly recurrent, hence that w is ultimately uniformly recurrent. ▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).* Given a bi-infinite word $w \in \Sigma^{\mathbb{Z}}$, we can consider $w_+ \in \Sigma^{\mathbb{N}}$ and $w_- \in \Sigma^{\mathbb{N}}$ the two infinite words obtained as follows: for all $i \in \mathbb{N}$, $(w_+)_i = w(i)$ and $(w_-)_i = w(-i)$. Note that the two share the letter at position 0.

Assume that w_+ and w_- are ultimately uniformly recurrent. Let us write $\mathsf{Infixes}(w)$ the set of finite infixes of w. Consider an infinite sequence of words $(u_i)_{i \in \mathbb{N}}$ in $\mathsf{Infixes}(w)$. If there is an infinite subsequence of words that are all in $\mathsf{Infixes}(w_+)$, then there exists an increasing pair of indices $i < j$ such that $u_i \sqsubseteq_{\mathsf{infix}} u_j$ because **??** applies to w_+. Similarly, if there is an infinite subsequence of words that are all in $\mathsf{Infixes}(w_-)$, then there exists an increasing pair of indices $i < j$ such that $u_i \sqsubseteq_{\mathsf{infix}} u_j$ because **??** applies to w_- (and the infix relation is compatible with mirroring). Otherwise, one can assume without loss of generality that all words in the sequence have a starting position in w_- and an ending position in w_+. In this case, let us write $(k_i, l_i) \in \mathbb{N}^2$ the pair of indices such that u_i is the infix of w that starts at position $-k_i$ of w (i.e., k_i of w_-) and ends at position l_i of w (i.e., l_i of w_+). Because $\mathbb{N}^2$ is a well-quasi-ordering with the product ordering, there exists $i < j$ such that $k_i \leq k_j$ and $l_i \leq l_j$, in particular, $u_i \sqsubseteq_{\mathsf{infix}} u_j$. We have proven that every infinite sequence of words in $\mathsf{Infixes}(w)$ is good, hence $\mathsf{Infixes}(w)$ is well-quasi-ordered.

Conversely, assume that $\mathsf{Infixes}(w)$ is well-quasi-ordered. In particular, the subset $\mathsf{Infixes}(w_+) \subseteq \mathsf{Infixes}(w)$ is well-quasi-ordered. Similarly, $\mathsf{Infixes}(w_-)$ is well-quasi-ordered because the infix relation is compatible with mirroring. Applying **??**, we conclude that both are ultimately uniformly recurrent words. ▷ Back to p.??

*Proof (Proof of **??** as stated on page **??**).* Given a bi-infinite word $w \in \Sigma^{\mathbb{Z}}$, recall that we can consider $w_+ \in \Sigma^{\mathbb{N}}$ and $w_- \in \Sigma^{\mathbb{N}}$ the two infinite words obtained as follows: for all $i \in \mathbb{N}$, $(w_+)_i = w(i)$ and $(w_-)_i = w(-i)$. Note that the two share the letter at position 0.

To obtain the upper bound of $\omega \cdot 3$, we can consider the same argument as for **??**. We let N_0 be such that $w_{\geq N_0}$ and $(w_-)_{\geq N_0}$ are uniformly recurrent words. In any sequence of incomparable elements of $\mathsf{Infixes}(w)$, there are less than N_0^2 elements that are found in $(w_{<N_0})_{>-N_0}$. Then, one has to pick a finite infix in either $w_{\geq N_0}$ or $w_{\leq -N_0}$. Because of **??**, any sequence of incomparable elements of these two infinite words has length bounded based on the choice of the first element of that sequence. This means that the ordinal width of $\mathsf{Infixes}(w)$ is at most $\omega + \omega + N_0^2$. We conclude that $\mathfrak{w}(\mathsf{Infixes}(w)) < \omega \cdot 3$.

Let us briefly argue that the bound is tight. Indeed, one can construct a bi-infinite word w by concatenating a reversed Thue-Morse sequence on a binary alphabet $\{a, b\}$, a finite antichain of arbitrarily large size over a distinct alphabet $\{c, d\}$, and then a Thue-Morse sequence on a binary alphabet $\{e, f\}$. The ordinal width of the set of infixes of w is then at least $\omega \cdot 2 + K$, where K is the size of the chosen antichain, following the same argument as in the proof of **??**, using **??**. ▷ Back to p.??

Lemma 35. *Given an automatic sequence $w \in \Sigma^{\mathbb{N}}$, one can decide whether it is ultimately uniformly recurrent.*

*Proof (Proof of **??** as stated on page **??**).* We can rewrite this as a question on the automatic sequence w as follows:

$$
\begin{aligned}
&\exists N_0, && \text{ultimately} \\
&\forall i_s \geq N_0, && \text{for every infix (start) } u \\
&\forall i_e > i_s, && \text{for every infix (end) } u \\
&\exists k \geq 1, && \text{there exists a bound} \\
&\forall j_s \geq N_0, && \text{for every other infix (start) } v \\
&\forall j_e \geq j_s + k, && \text{of size at least } k \\
&\exists l \geq 0, && \text{there exists a position in } v \\
&\forall 0 \leq m < i_e - i_s, && \text{where } u \text{ can be found} \\
&j_s + m + l < j_e \land w(i_s + m) = w(j_s + m + l) \quad .
\end{aligned}
$$

Because w is computable by a finite automaton, one can reduce the above formula to a regular language, for which it suffices to check emptiness, which is decidable. ▷ Back to p.??

*Proof (Proof of **??** as stated on page **??**).* Let N_0 be a bound such that $w_{\geq N_0}$ is uniformly recurrent. Let us write $\mathsf{Infixes}(w)$ the set of finite infixes of w. We prove that $\mathfrak{w}(\mathsf{Infixes}(w)) \leq \omega + N_0$. Let $u_1 \sqsubseteq_{\mathsf{infix}} w$ be a finite word.

If u_1 is an infix of $w_{\geq N_0}$, then there exists $k \geq 1$ such that u_1 is an infix of every word of size at least k. In particular, there is finite bound on the length of every sequence of incomparable elements starting with u_1. We conclude in particular that $\mathsf{Infixes}(w) \setminus \uparrow u_1$ has a finite ordinal width.

Otherwise, u_1 can only be found *before* N_0. In this case, we consider a second element of a bad sequence $u_2 \sqsubseteq_{\mathsf{infix}} w$, which is incomparable with u_1 for the infix relation. If u_2 is an infix of $w_{\geq N_0}$, then we can conclude as before. Otherwise, notice that u_1 and u_2 cannot start at the same position in w (because they are incomparable). Continuing this argument, we conclude that there are at most N_0 elements starting before N_0 at the start of any sequence of incomparable elements in $\mathsf{Infixes}(w)$. We conclude that $\mathfrak{w}(\mathsf{Infixes}(w)) \leq \omega + N_0$.

Let us now justify that this bound is tight. The Thue-Morse sequence over a binary alphabet $\{a, b\}$ has ordinal width ω from **??**. Given a number $N_0 \in \mathbb{N}$, one can construct an arbitrarily long antichain of words for the infix relation by using a new letter c. When concatenating this (finite) antichain as a prefix of the Thue-Morse sequence, one obtains a new (infinite) word w. It is clear that the ordinal width of $\mathsf{Infixes}(w)$ is now at least $\omega + N_0$. ▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).* It is always true that the ordinal height of a language over a finite alphabet is at most ω. Let us now consider a well-quasi-ordered language L that is downwards closed for the infix relation. Applying **??**, we can write $L = \bigcup_{i=1}^{n} L_i$ where each L_i is the set of finite infixes of a finite, infinite or bi-infinite ultimately uniformly recurrent word w_i. We can then directly conclude that $\mathfrak{w}(L_i)$ less than ω (in the case of a finite word), less than $\omega \cdot 2$ (in the case of an infinite word thanks to **??**), or less than $3 \cdot \omega$ (in the case of a bi-infinite word, thanks to **??**). In any case, we have the bound $\mathfrak{w}(L_i) < \omega \cdot 3$.

Now, $\mathfrak{w}(L) \leq \sum_{i=1}^{n} \mathfrak{w}(L_i) < \omega \cdot 3 < \omega^2$. Finally, the inequality $\mathfrak{o}(L) \leq \mathfrak{w}(L) \otimes \mathfrak{h}(L) < \omega \otimes \omega^2 = \omega^3$ allows us to conclude.

The tightness of the bounds is a direct consequence of **??**, and by considering a finite union of these examples over disjoint alphabets (or even, by considering a binary alphabet and using unambiguous codes to separate the different components). ▷ Back to p.??

E Proofs for Section ??

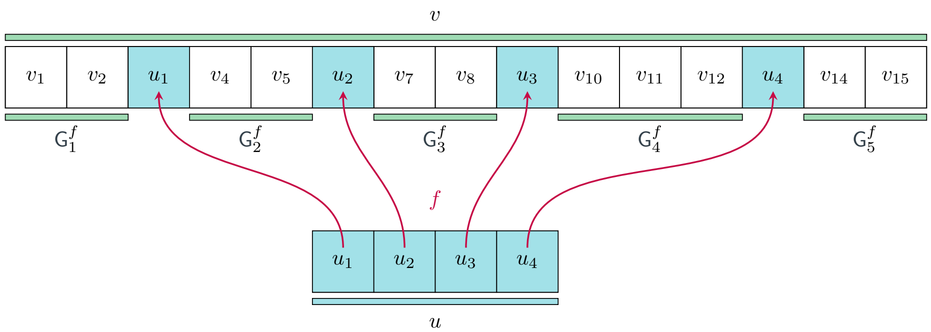


Fig. 6: The gap words resulting from a subword embedding between two finite words.

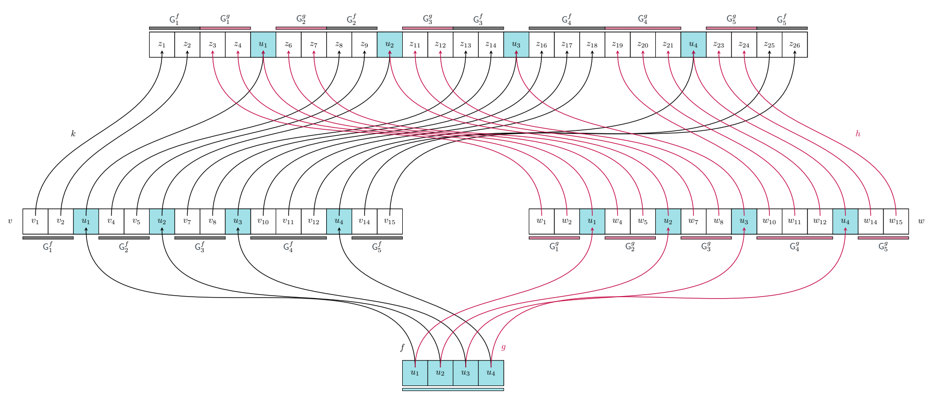


Fig. 7: We illustrate how embeddings f and g between runs of an amalgamation system can be glued together, seen on their canonical decomposition.

For this paper to be self-contained, we will also recall how runs of a finite state automaton can be understood as an amalgamation system.

Example 36 ([?, Section 3.2]). Let $A = (Q, \delta, q_0, F)$ be a finite state automaton over a finite alphabet Σ. Let Δ be the set of transitions $(q_1, a, q_2) \in Q \times \Sigma \times Q$, and $R \subseteq \Delta^*$ be the set of words over transitions that start with the initial state q_0, end in a final state $q_f \in F$, and such that the end state of a letter is the start state of the following one. The canonical decomposition can is defined as a morphism from Δ^* to Σ^* that maps (q,a,p) to a. Because of the one-to-one correspondence of steps of a run ρ and letters in its canonical decomposition, we may treat the two interchangeably. Finally, given two runs ρ and σ of the automaton, we say that an embedding $f \in \mathsf{Hom}^*(\mathsf{can}(\rho), \mathsf{can}(\sigma))$ belongs to $E(\rho,\sigma)$ when f is also defining an embedding from ρ to σ as words in Δ^*.

The system $(\Sigma, R, E, \mathsf{can})$ is an amalgamation system, whose language is precisely the language of words recognized by the automaton A.

Proof. By definition, the embeddings inside $E(\rho,\sigma)$ are in of $\mathsf{Hom}^*(\mathsf{can}(\rho), \mathsf{can}(\sigma))$, and they compose properly. Because $\Delta = Q \times \Sigma \times Q$ is finite, it is a well-quasi-ordering when equipped with the equality relation, and we conclude that Δ^* with $\leq^*$ is a well-quasi-order according to Higman's Lemma [**?**].

Let us now move to proving that the system satisfies the amalgamation property. Given three runs $\rho, \sigma, \tau \in R$, and two embeddings $f \in E(\rho,\sigma)$ and $g \in E(\rho,\tau)$, we want to construct an amalgamated run $\sigma \vee \tau$. Because letters in the run ρ respect the transitions of the automaton (i.e., if the letter i ends in state q, then the letter $i+1$ starts in state q), then the gap word at position i starts in state q and ends in state q too. This means that for both embeddings f and g, the gap words are read by the automaton by looping on a state. In particular, these loops can be taken in any order and continue to represent a valid run. That is, we can even select the order of concatenation in the amalgamation for *all* $0 \leq i \leq |\mathsf{can}(\rho)|$ and not just for one separately.

We conclude by remarking that the language of this amalgamation system is the set of $\mathsf{yield}(R)$, because R is the set of valid runs of the automaton, and $\mathsf{yield}(\rho)$ is the word read along a run ρ.

Proof (Proof of **??** *as stated on page* **??***).* Write u for G^f_ℓ and v for G^g_ℓ. We may assume that both u and v are non-empty, as otherwise the lemma holds trivially. Then, for all $k \in \mathbb{N}$, there exists a run with canonical decomposition

$$w_k = L_0 a_1 \cdots a_n L_n,$$

where $L_i \in \{vvu^k, vu^kv, u^kvv\}$ and specifically $L_\ell = vu^kv$.

From **??**, we may conclude that there are a finite number of words x, y, and w such that each w_k is contained in a language $\mathsf{P}{\downarrow}(x) w \, \mathsf{P}{\downarrow}(y)$.

As there is an infinite number of words w_k, we may fix x, y, and w and an infinite subset $I \subseteq \mathbb{N}$ such that $\{w_i \mid i \in I\} \subseteq \mathsf{P}{\downarrow}(x) w \, \mathsf{P}{\downarrow}(y)$. This implies that either for infinitely many $m \in \mathbb{N}$, $u^m v \in \mathsf{P}{\downarrow}(y)$ or for infinitely many m, $vu^m \in \mathsf{P}{\downarrow}(x)$.

In either case, we may conclude that either $u \sqsubseteq_{\mathsf{infix}} v$ or $v \sqsubseteq_{\mathsf{infix}} u$: Let $m, n \in \mathbb{N}$ such that $m < n$ and $u^m v, u^n v \in \mathsf{P}{\downarrow}(y)$ (the case for vu^m and vu^n proceeding analogously). Without loss of generality, assume that $|u^m|$ and $|u^n|$ are multiples of $|y|$. We therefore find $p \sqsubseteq_{\mathsf{pref}} y, s \sqsubseteq_{\mathsf{suff}} y$ such that $u^m, u^n \in sy^*p$, ergo $ps = y$. In other words, we can write $u^m = (sp)^{m'}, u^n = (sp)^{n'}$. As $u^m v \in \mathsf{P}{\downarrow}(y)$, it follows that v is a prefix of some word in $(sp)^*$. Hence either v is a prefix of u or u vice versa. ▷ Back to p.??

Proof (Proof of **??** *as stated on page* **??***).* It is clear that **??** $\Rightarrow$ **??** because regular languages are recognized by finite automata, and finite automata are a particular case of amalgamation systems. The implication **??** $\Rightarrow$ **??** is the content of **??**. The implication **??** $\Rightarrow$ **??** is **??**. Finally, the implication **??** $\Rightarrow$ **??** is simply because a downwards closed language that is a finite union of products of chains is a regular language.

Indeed, assume that L is downwards closed and included in a finite union of sets of the form $\mathsf{P}{\downarrow}(x)u\,\mathsf{P}{\downarrow}(y)$ where x, y, u are possibly empty words. We can assume without loss of generality that for every n, $x^n u y^n$ is in L, otherwise, we have a bound on the maximal n such that $x^n u y^n$ is in L, and we can increase the number of languages in the union, replacing x or y with the empty word as necessary. Let us write $L' \triangleq \bigcup_{i=1}^{k} x_i^* u_i y_i^*$. Then, $L' \subseteq L$ by construction. Furthermore, $L \subseteq {\downarrow} L'$, also by construction. Finally, we conclude that $L = {\downarrow} L'$ because L is downwards closed. Now, because L' is a regular language, and regular languages are closed under downwards closure, we conclude that L is a regular language.

▷ Back to p.**??**

Proof (Proof of **??** *as stated on page* **??***).* Assume that L is well-quasi-ordered by the infix relation, and obtained by an amalgamation system $(\Sigma, R, E, \mathsf{can})$.

Let us consider the set M of minimal runs for the relation $\leq_E$, which is finite because the latter is a well-quasi-ordering. By **??**, we know that for each minimal run $\rho \in M$, each gap language L_i^ρ of ρ is totally ordered by $\sqsubseteq_{\mathsf{infix}}$. Adapting the proof of language boundedness from [**?**, Section 4.2], we may conclude that $\mathsf{L}_i^\rho \subseteq \mathsf{P}{\downarrow}(w)$ for some $w \in \mathsf{L}_i^\rho$. As $\mathsf{P}{\downarrow}(w)$ is language bounded and this property is stable under subsets, concatenation and finite union, we can conclude that L is bounded as well.

▷ Back to p.**??**

Let us briefly recall that a *rational transduction* is a relation $R \subseteq \Sigma^* \times \Gamma^*$ such that there exists a finite state automaton that reads pairs of letters $(a, b) \in (\Sigma \cup \{\epsilon\}) \times (\Gamma \cup \{\epsilon\})$ and recognizes R. A class of languages $\mathcal{C}$ is *closed under rational transductions* if for every $L \in \mathcal{C}$ and every rational transduction R, the language $R(L) \triangleq \{v \in \Gamma^* \mid \exists u \in L, (u, v) \in R\}$ also belongs to $\mathcal{C}$.

Proof (Proof of **??** *as stated on page* **??***).* We first show **??** $\Rightarrow$ **??**. We aim to make the inclusion test of **??** of **??** effective. Let $R(n, m, N_0) \triangleq \bigcup_{x,y \in \Sigma^{\leq n}} \bigcup_{u \in \Sigma^{\leq m \times N_0}} \mathsf{P}{\downarrow}(x)u\,\mathsf{P}{\downarrow}(y) \cup \mathsf{P}{\downarrow}(x)u \cup u\,\mathsf{P}{\downarrow}(x)$. For any concrete values of the bounds n, m, and N_0, this language is regular. The map $L \mapsto L \cap \Sigma^* \setminus R(n, m, N_0)$ is a rational transduction because $\Sigma^* \setminus R(n, m, N_0)$ is regular. Since $\mathcal{C}$ is closed under rational transductions, we can therefore reduce the inclusion to emptiness of this language. However, we need to find these bounds first.

To determine values for n and m, we first test if L is bounded. Since emptiness is decidable, we can apply the algorithm in [**?**, Section 4.2] to decide if L is bounded. If L is bounded, this algorithm yields words $w_1, \dots w_n$ such that $L \subseteq w_1^* \cdots w_n^*$ and therefore yields also the bounds in questions: n is the number of words, and m is the maximal length of a word w_i where $1 \leq i \leq n$. If L is not bounded, then L cannot be well-quasi-ordered by the infix relation because of **??** and we immediately return false.

To determine the value for N_0, we then compute the downward closure (with respect to subwords) of L. This is effective and yields a finite-state automaton. Recall that N_0 is the maximum number of repetitions of a word w_i that can not be iterated arbitrarily often. This value is therefore bounded above by the length of the longest simple path in this automaton.

?? $\Rightarrow$ **??**. We just consider the transduction f that maps every word w to $\#w$ where $\#$ is a fresh symbol. Then, for any language $L \in \mathcal{C}$, L is well-quasi-ordered by prefix if and only if $f(L)$ is well-quasi-ordered by infix.

?? $\Rightarrow$ **??**. We consider the transduction $R \triangleq \Sigma^* \times \{a, b\}^*$. Then for any language $L \in \mathcal{C}$, the image of L through R is well-quasi-ordered by prefix if and only if L is empty.

▷ Back to p.??