

# Well-quasi-orderings on word languages

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 decision procedures when the languages are given by various computational 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 [2,1]. One of the appealing properties of well-quasi-orderings is that they are closed under many operations, such as taking products, finite unions, and finite powerset constructions [13]. Perhaps more surprisingly, the class of well-quasi-ordered sets is also stable under the operation of taking finite words and finite trees labeled by elements of a well-quasi-ordered set [20,23].

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

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  $\Sigma$  equipped with the equality relation. In this setting, the three alternatives we consider are the *prefix relation* ( $u \sqsubseteq_{\text{pref}} v$  if there exists  $w$  with

40  $uw = v$ ), the *suffix relation* ( $u \sqsubseteq_{\text{suffix}} v$  if there exists  $w$  such that  $wu = v$ ), and  
 41 the *infix relation* ( $u \sqsubseteq_{\text{infix}} v$  if there exists  $w_1, w_2$  such that  $w_1uw_2 = v$ ). Note  
 42 that these three relations straightforwardly generalize to infinite quasi-ordered  
 43 alphabets. Unfortunately, it is easy to see that none of these relations yield  
 44 well-quasi-ordered sets as soon as the alphabet contains two distinct letters: for  
 45 instance, the infinite sequence of words  $(ab^na)_{n \in \mathbb{N}}$  is well-quasi-ordered by the  
 46 subword relation but by neither the prefix relation, nor the suffix relation, nor  
 47 the infix relation.

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

55 Let us argue that these decision procedures fit a larger picture in the research  
 56 area of well-quasi-orderings. Indeed, there have been recent breakthroughs in  
 57 deciding whether a given order is a well-quasi-order, for instance in the context of  
 58 the verification of infinite state transition systems [19] or in the context of logic [7].  
 59 In the graph theory community, recent works have studied classes of graphs that  
 60 are well-quasi-ordered by the induced subgraph relation using similar language  
 61 theoretic techniques [12,27,6]. Furthermore, a previous work by Kuske shows  
 62 that any *reasonable*<sup>1</sup> partially ordered set  $(X, \leq)$  can be embedded into  $\{a, b\}^*$   
 63 with the infix relation [25, Lemma 5.1]. Phrased differently, one can encode a  
 64 large class of partially ordered sets as subsets of  $\{a, b\}^*$ . As a consequence, the  
 65 following decision problem provides a reasonable abstract framework for deciding  
 66 whether a given partially ordered set is well-quasi-ordered: given a language  
 67  $L \subseteq \Sigma^*$ , decide whether  $L$  is well-quasi-ordered by the infix relation.

68 The runtime of an algorithm based on well-quasi-orderings is deeply related  
 69 to the “complexity” of the underlying quasi-order [31]. One way to measure this  
 70 complexity is to consider its so-called ordinal invariants: for instance, the maximal  
 71 order type (or m.o.t.), originally defined by De Jongh and Parikh [21], is the  
 72 order type of the maximal linearization of a well-quasi-ordered set. In the case of  
 73 a finite set, the m.o.t. is precisely the size of the set. Better runtime bounds were  
 74 obtained by considering two other parameters [32]: the ordinal height introduced  
 75 by Schmidt [30], and the ordinal width of Krřř and Thomas [26]. Therefore,  
 76 when characterizing well-quasi-ordered languages, we will also be interested in  
 77 deriving upper bounds on their ordinal invariants. This analysis also allows us to  
 78 better compare the well-quasi-orderings. We refer to Section 2 for a more detailed  
 79 introduction to these parameters and ordinal computations in general.

80 *Contributions* We focus on languages over a finite alphabet  $\Sigma$ . In this setting, we  
 81 first characterize languages that are well-quasi-ordered by the prefix relation (and  
 82 symmetrically, by the suffix relation), and derive tight bounds on their ordinal

<sup>1</sup> This will be made precise in Lemma 7.

83 invariants. These generic results are then used to devise a decision procedure for  
 84 checking whether a language is well-quasi-ordered by the prefix relation, provided  
 85 the language is given as input as a finite automaton (Corollary 4). A summary  
 86 of these results can be found in Figure 1.

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

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

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

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

Fig. 2: Summary of results for the infix relation, the bounds on  $\mathfrak{w}(L)$  and  $\mathfrak{o}(L)$  are tight, and respectively proven in Corollary 14 and Corollary 21.

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 (Section 5.2), for which we provide a decision procedure (Theorem 27). The model used to represent these infinite words is based on automatic sequences and morphic sequences [11], which are well-studied in the context of symbolic dynamics. For bounded languages, we consider the model of amalgamation systems [5], which is an abstract computational model that encompasses many classical ones, such as finite automata, context-free grammars, and Petri nets [5]. We show that if a language recognized by an amalgamation system is well-quasi-ordered by the infix relation, then it is a bounded language (Theorem 29), and is therefore regular. Furthermore, we show that we can decide whether a given language recognized by an amalgamation system is well-quasi-ordered by the infix relation (Theorem 30). We defer the introduction of amalgamation systems to Section 6.1.

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

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

*Outline* We introduce in Section 2 the necessary background on well-quasi-orders and ordinal invariants. In Section 3, which is relatively self-contained, we study the prefix relation and prove in Theorem 5 the characterization of well-quasi-ordered languages by the prefix relation. In Section 4, we obtain the infix analogue of Theorem 5 specifically for bounded languages (Theorem 8). In Section 5, we study

the downwards closed languages, characterize them using a notion of ultimately uniformly recurrent words borrowed from symbolic dynamics (Theorem 20), and compute bounds on their ordinal invariants in Corollary 21. Finally, we generalize these results to all amalgamation systems in Section 6 in (Theorem 29), and provide a decision procedure for checking whether a language is well-quasi-ordered by the infix relation (resp. prefix and suffix) in this context (Theorem 30).

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

## 2 Preliminaries

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

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

The prefix relation (resp. the suffix relation and the infix relation) on  $\Sigma^*$  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  $\alpha, \beta, \gamma$  to denote ordinals, and use  $\omega$  to denote the first infinite ordinal, i.e., the set of natural numbers with the usual ordering. We also use  $\omega_1$  to denote the first *uncountable* ordinal. We only assume superficial familiarity with ordinal arithmetic, and refer to the books of Kunen [24] and Krivine [22, Chapter II] for a detailed introduction to this domain. Given a tree  $T$  whose branches are all finite we can define an ordinal  $\alpha_T$  inductively as follows: if  $T$  is a leaf then  $\alpha_T = 0$ , if  $T$  has children  $(T_i)_{i \in \mathbb{N}}$  then  $\alpha_T = \sup\{\alpha_{T_i} + 1 \mid i \in \mathbb{N}\}$ . We say that  $\alpha_T$  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  $\mathfrak{o}(X)$  [21], the *ordinal height* of  $X$  written  $\mathfrak{h}(X)$  [30], and the *ordinal width* of  $X$  written  $\mathfrak{w}(X)$  [26]. These three parameters are called the *ordinal invariants* of a well-quasi-ordered set  $X$ . As an example, for  $(\mathbb{N}, \leq)$ , all bad sequences are descending and antichains have size at most 1. In fact,  $(\mathbb{N}, \leq)$  is itself an ordinal, namely  $\omega$ . Hence it is its own maximal order type and ordinal height, and its ordinal width is 1. We refer to the survey of [15] for a detailed discussion on these concepts and their computation on specific classes of well-quasi-ordered sets.

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

### 3 Prefixes and Suffixes

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

The next remark we make is that  $\Sigma^*$  is not well-quasi-ordered by the prefix relation as soon as  $\Sigma$  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  $\Sigma$  is the infinite tree  $T$  whose nodes are the words of  $\Sigma^*$ , and such that the children of a word  $w$  are the words  $wa$  for all  $a \in \Sigma$ .

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

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

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

244 **Lemma 3.** Let  $L \subseteq \Sigma^*$  be a language. Then,  $L$  contains an infinite antichain if  
 245 and only if there exists an antichain branch for  $L$ . ▷ Proven p.23

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

251 **Corollary 4.** If  $L$  is regular, then the existence of an infinite antichain is  
 252 decidable. ▷ Proven p.23

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

256 **Theorem 5.** A language  $L \subseteq \Sigma^*$  is well-quasi-ordered by the prefix relation if  
 257 and only if  $L$  is a union of chains. ▷ Proven p.23

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

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

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



## 4 Infixes and Bounded Languages

In this section, we study languages equipped with the infix relation. As opposed to the prefix and suffix relations, the infix relation can lead to very complicated well-quasi-ordered languages. Formally, the upcoming Lemma 7 due to Kuske shows that *any* countable partial-ordering with finite initial segments can be embedded into the infix relation of a language. To make the former statement precise, let us recall that an *order embedding* from a quasi-ordered set  $(X, \preceq)$  into a quasi-ordered set  $(Y, \preceq')$  is a function  $f: X \rightarrow Y$  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 Lemma 7 is illustrated for the subword relation in Figure 5 page 25.

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

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

As a consequence of Lemma 7, we cannot replay proofs of Section 3, and will actually need to leverage some regularity of the languages to obtain a characterization of well-quasi-ordered languages under the infix relation. This regularity will be imposed through the notion of *bounded languages*, i.e., languages  $L \subseteq \Sigma^*$  such that there exists words  $w_1, \dots, w_n$  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  $\Sigma^*$ . Then,  $L$  is a well-quasi-order when endowed with the infix relation if and only if it is included in a finite union of products  $S_i \cdot P_i$  where  $S_i$  is a chain for the suffix relation, and  $P_i$  is a chain for the prefix relation, for all  $1 \leq i \leq n$ .*

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

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

The reason why periodic words built using a given period  $x \in \Sigma^+$  are interesting for the infix relation is that they naturally create chains for the prefix and suffix relations. Indeed, if  $x \in \Sigma^+$  is a finite word, then  $\{x^p \mid p \in \mathbb{N}\}$  is a chain for the infix relation. Note that in general, the downwards closure of a chain is *not* a chain (see Remark 9). However, for the chains generated using periodic words, the downwards closure  $\downarrow_{\sqsubseteq_{\text{infix}}} \{x^p \mid p \in \mathbb{N}\}$  is a *finite union* of



311 chains. Because this set will appear in bigger equations, we introduce the shorter  
 312 notation  $\mathbf{P}\downarrow(x)$  for the set of infixes of words of the form  $x^p$ , where  $p \in \mathbb{N}$ .

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

320 **Lemma 10.** *Let  $x \in \Sigma^+$  be a word. Then  $\mathbf{P}\downarrow(x)$  is a finite union of chains for* ▷ Proven p.25  
 321 *the infix, prefix and suffix relations simultaneously.*

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

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

332 **Lemma 12.** *Let  $L \subseteq \Sigma^*$  be a language that is well-quasi-ordered by the infix* ▷ Proven p.25  
 333 *relation. Let  $k \in \mathbb{N}$ ,  $u_1, \dots, u_{k+1} \in \Sigma^*$ , and  $v_1, \dots, v_k \in \Sigma^+$  be such that*  
 334  *$w[\mathbf{n}] \triangleq (\prod_{i=1}^k u_i v_i^{n_i}) u_{k+1}$  belongs to  $L$  for arbitrarily large values of  $\mathbf{n} \in \mathbb{N}^k$ .*  
 335 *Then, there exists  $x, y \in \Sigma^+$  of size at most  $\max\{|v_i| \mid 1 \leq i \leq k\}$  such that for*  
 336 *all  $\mathbf{n} \in \mathbb{N}^k$  one of the following holds:*

- 337 1.  $w[\mathbf{n}] \in u_1 \mathbf{P}\downarrow(x)$ ,
- 338 2.  $w[\mathbf{n}] \in \mathbf{P}\downarrow(x) u_{k+1}$ ,
- 339 3.  $w[\mathbf{n}] \in \mathbf{P}\downarrow(x) u_i \mathbf{P}\downarrow(y)$  for some  $1 \leq i \leq k+1$ .

340 **Lemma 13.** *Let  $L \subseteq \Sigma^*$  be a bounded language that is well-quasi-ordered by the* ▷ Proven p.26  
 341 *infix relation. Then, there exists a finite subset  $E \subseteq (\Sigma^*)^3$ , such that:*

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

342 *Proof (Proof of Theorem 8 as stated on page 8).* We apply Lemma 13, and  
 343 conclude because  $\mathbf{P}\downarrow(x)$  is a finite union of chains for the prefix, suffix and infix  
 344 relations (Lemma 10). ▷ Back to p.8

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

*Proof.* Upper bounds are a direct consequence of Theorem 8, 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  $\omega \cdot k$ ,  $\omega$  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 Lemma 7 crucially relies on constructing languages that are *not* downwards closed, and we have shown in Remark 9 that the downwards closure of a well-quasi-ordered language is not necessarily well-quasi-ordered.

### 5.1 Characterization of Well-Quasi-Ordered Downwards Closed Languages

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

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

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

*Example 16.* Let  $L \triangleq a^* b^* \cup b^* a^*$ . This language is bounded, is 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  $ab^k a$  and  $ba^k b$  where  $k \geq 1$ .

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

**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 Theorem 17, it is well-quasi-ordered.

Assume by contradiction that  $I_{\mathbf{t}}$  is bounded. In this case, there exist words  $w_1, \dots, w_k \in \Sigma^*$  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  $u \sqsubseteq_{\text{infix}} \mathbf{t}$ , 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  $\sqsubseteq_{\text{infix}}$ , the maximal order type of  $I_{\mathbf{t}}$  is  $\omega$ , the ordinal height of  $I_{\mathbf{t}}$  is  $\omega$ , the ordinal width of  $I_{\mathbf{t}}$  is  $\omega$ .*

*Proof.* We first show that  $\omega$  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  $\omega$ . Finally, because the ordinal width is at most the maximal order type (as per Section 2, using for instance the results of [26] or [15, Theorem 3.8] stating  $\mathfrak{o}(X) \leq \mathfrak{h}(X) \otimes \mathfrak{w}(X)$ ): we conclude that the ordinal width is also at most  $\omega$ .

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

We prove in the upcoming Theorem 20 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  $\text{Infixes}(w)$  for the set of finite

infixes of a (possibly infinite or bi-infinite) word  $w \in \Sigma^* \cup \Sigma^{\mathbb{N}} \cup \Sigma^{\mathbb{Z}}$ . We say that an infinite word  $w \in \Sigma^{\mathbb{N}}$  is *ultimately uniformly recurrent* if there exists a bound  $N_0 \in \mathbb{N}$  such that  $w_{\geq N_0}$  is uniformly recurrent. We extend this notion to finite words by considering that they all are ultimately uniformly recurrent, and to bi-infinite words by considering that they are ultimately uniformly recurrent if and only if both their left-infinite and right-infinite parts are.

▷ Proven p.12

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

Thanks to Theorem 20, and by analysing the ordinal invariants of infixes of an ultimately uniformly recurrent infinite word  $w$  (Lemma 23), we conclude that the ordinal invariants of a well-quasi-ordered downwards closed language are relatively small.

▷ Proven p.30

**Corollary 21.** *Let  $L$  be a well-quasi-ordered downwards closed language for the infix relation. Then, the maximal order type of  $L$  is strictly less than  $\omega^3$ , its ordinal height is at most  $\omega$ , and its ordinal width is at most  $\omega^2$ .*

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-infinite word  $w \in \Sigma^* \cup \Sigma^{\mathbb{N}} \cup \Sigma^{\mathbb{Z}}$  (Lemma 22).

▷ Proven p.27

**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  $w$ .*

▷ Proven p.27

**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.*

▷ Proven p.28

**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 Theorem 20.

*Proof (Proof of Theorem 20 as stated on page 12).* It is clear that the set of finite infixes of a finite, infinite or bi-infinite ultimately uniformly recurrent word is well-quasi-ordered for the infix relation thanks to Lemma 23.

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  $L = \bigcup_{i=1}^n L_i$ .

467 For every such ideal  $L_i$ , we can apply Lemma 22, and conclude that  $L_i$  is  
 468 the set of finite infixes of a finite, infinite or bi-infinite word  $w_i$ . Because the  
 469 languages  $L_i$  are well-quasi-ordered, we can apply Lemma 23, and conclude that  
 470  $w_i$  is ultimately uniformly recurrent. ▷ Back to p.12

471 Finally, we comment on the ordinal invariants of the set of finite infixes  
 472 of an ultimately uniformly recurrent infinite word, from which the bounds of  
 473 Corollary 21 naturally follow (the proof is in Section D page 30).

474 **Lemma 25.** *Let  $w \in \Sigma^{\mathbb{N}}$  be an ultimately uniformly recurrent word. Then, the*  
 475 *set of finite infixes of  $w$  has ordinal width less than  $\omega \cdot 2$ . Furthermore, this bound*  
 476 *is tight.* ▷ Proven p.29

477 **Lemma 26.** *Let  $w \in \Sigma^{\mathbb{Z}}$  be a bi-infinite word. Then, the set of finite infixes of*  
 478  *$w$  is well-quasi-ordered for the infix relation if and only if  $w_+$  and  $w_-$  are two*  
 479 *ultimately uniformly recurrent words. In this case, the ordinal width of the set of*  
 480 *finite infixes of  $w$  is less than  $\omega \cdot 3$ , and this bound is tight.* ▷ Proven p.29

## 481 5.2 Decision Procedures

482 As we have demonstrated, infinite (or bi-infinite words) can be used to represent  
 483 languages that are well-quasi-ordered for the infix relation by considering their  
 484 set of finite infixes. Let us formalise the representation of languages by sets of  
 485 bi-infinite words that we will use in this section, following the characterization  
 486 of Lemma 22. A *sequence representation* of a language  $L \subseteq \Sigma^*$  is a finite set of  
 487 triples  $(w_i^-, a_i, w_i^+)$  <sub>$1 \leq i \leq n$</sub>  where  $w_i^-, w_i^+ \in \Sigma^{\mathbb{N}} \cup \Sigma^*$  are two potentially infinite  
 488 words, and  $a_i \in \Sigma$  is a letter, such that

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

489 Given an effective representation of sequences, one obtains an effective rep-  
 490 resentation of languages via sequence representations. In this section, we will  
 491 rely on definitions originating from the area of symbolic dynamics, that precisely  
 492 study infinite words whose generation follows from a finitely described process.  
 493 However, we will not assume that the reader is familiar with this domain, and  
 494 we will use as black-boxes key results from this area.

495 A first model that one can use to represent infinite words is the model of  
 496 *automatic sequences*. In this case, the infinite word  $w$  is described by a finite  
 497 state automaton, that can compute the  $i$ -th letter of the word  $w$  given as input  
 498 the number  $i$  written in some base  $b \in \mathbb{N}$ . An example of such a sequence is  
 499 the Thue-Morse sequence that can be described by a finite automaton using a  
 500 binary representation of the indices. The good algorithmic properties of automatic  
 501 sequences come from the fact that a Presburger definable property that uses  
 502 letters of the sequence can be (trivially) translated into a finite automaton that  
 503 reads the base  $b$  representation of the free variables (that are indices of the

sequence). In particular, it follows that one can decide if an automatic sequence is ultimately uniformly recurrent, a proof of this folklore result can be found in the appendix at Lemma 35. 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  $\text{reversed}(w_i^-)a_iw_i^+$  defines a well-quasi-ordered language. By Lemma 26, 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 Lemma 35, 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  $f: \Sigma^* \rightarrow \Gamma^*$  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  $f: \Sigma^* \rightarrow \Sigma^*$  on a letter  $a \in \Sigma$  such that  $f(a)$  starts with  $a$ , and then applying a homomorphism  $h: \Sigma^* \rightarrow \Gamma^*$ . The infinite word  $f^\omega(a)$  is the limit of the sequence  $(f^n(a))_{n \in \mathbb{N}}$ , which is well-defined because  $f(a)$  starts with  $a$ , and the morphic sequence is  $w \triangleq h(f^\omega(a))$ .

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

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

## 6 Infixes and Amalgamation Systems

In the previous section, we have represented languages that are 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 Section 4, and in particular Theorem 8, can safely be applied to amalgamation systems.

**Theorem 29.** *Let  $L \subseteq \Sigma^*$  be a language recognized by an amalgamation system. If  $L$  is well-quasi-ordered by the infix relation then  $L$  is bounded.* ▷ Proven p.33

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* (Section 6.1). In the following theorem, we prove that under such assumptions, testing well-quasi-ordering is inter-reducible to testing whether a language of the class is empty, which is usually the simplest problem for a computational model.

**Theorem 30.** *Let  $\mathcal{C}$  be an effective amalgamative class of languages. Then the following are equivalent:* ▷ Proven p.33

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

## 6.1 Amalgamation Systems

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

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

Given a subword embedding  $f: u \rightarrow v$  between two words  $u$  and  $v$ , there exists a unique decomposition  $v = G_0^f u_1 G_1^f \cdots G_{k-1}^f u_k G_k^f$  where  $G_i^f = v_{f(i)+1} \cdots v_{f(i+1)-1}$  for all  $1 \leq i \leq k-1$ ,  $G_k^f = v_{f(k)+1} \cdots v_{|v|}$ , and  $G_0^f = v_1 \cdots v_{f(1)-1}$ . We say that



$G_i^f$  is the  $i$ -th *gap word* of  $f$ . We encourage the reader to look at Figure 6 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  $(\Sigma, R, \text{can}, E)$  where  $\Sigma$  is a finite alphabet,  $R$  is a set of so-called runs,  $\text{can}: R \rightarrow (\Sigma \uplus \{\#\})^*$  is a function computing a *canonical decomposition* of a run, and  $E$  describes the so-called *admissible embeddings* between runs: If  $\rho$  and  $\sigma$  are runs from  $R$ , then  $E(\rho, \sigma)$  is a subset of the subword embeddings between  $\text{can}(\rho)$  and  $\text{can}(\sigma)$ . 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  $i \in [0, |\text{can}(r)|]$ , the *gap language* of  $r$  at position  $i$  is  $L_i^r \triangleq \{G_i^f \mid \exists s \in R. \exists f \in E(r, s)\}$ . An amalgamation system furthermore satisfies the following properties:

1.  $(R, E)$  Forms a Category. For all  $\rho, \sigma, \tau \in R$ ,  $\text{id} \in E(\rho, \rho)$ , 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  $\rho_0, \rho_1, \rho_2$  be runs with  $\rho_0 \trianglelefteq_f \rho_1$  and  $\rho_0 \trianglelefteq_g \rho_2$ . Then for all  $0 \leq i \leq |\text{can}(\rho_0)|$ , 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  $G_j^f G_j^g$  or  $G_j^h = G_j^g G_j^f$ . Specifically, for  $i$  we may fix  $G_i^h = G_i^f G_i^g$ . We refer to Figure 7 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.  $\text{yield}(\rho) = \pi_\Sigma(\rho)$ . The language recognized by an amalgamation system is  $\text{yield}(R)$ .

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

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

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

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

For any  $0 \leq \ell \leq n$ , we have  $G_\ell^f \sqsubseteq_{\text{infix}} G_\ell^g$  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!

▷ Proven p.33

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

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

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

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

641 Under these assumptions, one can transform the inclusion test of Equation (1)  
 642 of Theorem 8 into an effective procedure, using pumping arguments from [5,  
 643 Section 4.2], which, in turn, allows us to prove Theorem 30. Since the class  $\mathcal{C}_{\text{aut}}$   
 644 of regular languages and the class  $\mathcal{C}_{\text{cfg}}$  of context-free languages are examples of  
 645 effective amalgamative classes, the following corollary is immediate.

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

## 649 7 Conclusion

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

664 *Towards infinite alphabets* In this paper, we restricted our attention to *finite*  
 665 alphabets, having in mind the application to regular languages. However, the  
 666 conclusions of Theorem 8, Corollary 21, and Theorem 5 could be conjectured  
 667 to hold in the case of infinite alphabets (themselves equipped with a well-quasi-  
 668 ordering). This would require new techniques, as the finiteness of the alphabet is  
 669 crucial to all of our positive results.

670 *Monoid equations* It could be interesting to understand which monoids  $M$   
 671 recognize languages that are well-quasi-ordered by the infix, prefix or suffix  
 672 relations. This research direction is connected to finding which classes of graphs  
 673 of *bounded clique-width* are well-quasi-ordered with respect to the *induced subgraph*  
 674 *relation*, as shown in [12], and recently revisited in [27].

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

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

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

693 In some sense, our results confirm this intuition in the case of languages  
 694 described by a simple computational model. For languages recognized by amalga-  
 695 mation systems, being well-quasi-ordered implies being a bounded language, and  
 696 therefore being included in some finite union of languages of the form  $w_1^*w_2w_3^*$ .  
 697 Hence, these languages have at most a linear factor complexity. This is also the  
 698 case for languages described as the infixes of a finite set of pairs of morphic  
 699 sequences. Indeed, the factor complexity of a morphic sequence that is uniformly  
 700 recurrent is linear [28, Theorem 24], therefore the factor complexity of a language  
 701 given by sequence representation using morphic sequences is at most linear.

702 However, there are downwards closed languages that are well-quasi-ordered for  
 703 the infix relation but have an exponential factor complexity: the  $(5, 3)$ -Toeplitz  
 704 word is uniformly recurrent [11, p. 499], and has exponential factor complexity  
 705 [11, Theorem 5]. This shows that our computational models somehow fail to  
 706 capture vast classes of well-quasi-ordered languages with a high factor complexity.

It would be interesting to understand which new proof techniques would be required to obtain decidability for these languages.

To conclude on a positive note for the infix relation, our results show that for downwards closed and well-quasi-ordered languages, there is a strong connection between the factor complexity and the ordinal width: it is the same to have bounded factor complexity and finite ordinal width. A short proof can be found in appendix (Lemma 37).

## References

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

- 754 12. Daligault, J., Rao, M., Thomassé, S.: Well-quasi-order of relabel functions. *Order*  
755 **27**(3), 301–315 (September 2010). <https://doi.org/10.1007/s11083-010-9174-0>,  
756 <http://dx.doi.org/10.1007/s11083-010-9174-0>
- 757 13. Demeri, S., Finkel, A., Goubault-Larrecq, J., Schmitz, S., Schnoebelen, P.: Algorithmic aspects of wqo theory (mpri course) (2012), <https://cel.archives-ouvertes.fr/cel-00727025>, course notes
- 760 14. Durand, F.: Decidability of uniform recurrence of morphic sequences. *International Journal of Foundations of Computer Science* **24**(01), 123–146 (2013). <https://doi.org/10.1142/S0129054113500032>
- 763 15. Džamonja, M., Schmitz, S., Schnoebelen, P.: On Ordinal Invariants in Well Quasi Orders and Finite Antichain Orders, pp. 29–54. Springer International Publishing, Cham (2020). [https://doi.org/10.1007/978-3-030-30229-0\\_2](https://doi.org/10.1007/978-3-030-30229-0_2), [https://doi.org/10.1007/978-3-030-30229-0\\_2](https://doi.org/10.1007/978-3-030-30229-0_2)
- 767 16. D’Alessandro, F., Varricchio, S.: On well quasi-orders on languages, pp. 230–241. Springer Berlin Heidelberg (2003). [https://doi.org/10.1007/3-540-45007-6\\_18](https://doi.org/10.1007/3-540-45007-6_18)
- 769 17. D’Alessandro, F., Varricchio, S.: Well quasi-orders, unavoidable sets, and derivation systems. *RAIRO - Theoretical Informatics and Applications* **40**(3), 407–426 (Jul 2006). <https://doi.org/10.1051/ita:2006019>, <http://dx.doi.org/10.1051/ita:2006019>
- 773 18. Fine, N.J., Wilf, H.S.: Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society* **16**(1), 109–114 (Feb 1965). <https://doi.org/10.1090/s0002-9939-1965-0174934-9>, <http://dx.doi.org/10.1090/s0002-9939-1965-0174934-9>
- 777 19. Finkel, A., Gupta, E.: The well structured problem for presburger counter machines. In: Chattopadhyay, A., Gastin, P. (eds.) 39th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2019, December 11–13, 2019, Bombay, India. *LIPIcs*, vol. 150, pp. 41:1–41:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2019). <https://doi.org/10.4230/LIPICS.FSTTCS.2019.41>, <https://doi.org/10.4230/LIPICS.FSTTCS.2019.41>
- 783 20. Higman, G.: Ordering by divisibility in abstract algebras. *Proceedings of the London Mathematical Society* **3**, 326–336 (1952). <https://doi.org/10.1112/plms/s3-2.1.326>
- 786 21. de Jongh, D., Parikh, R.: Well-partial orderings and hierarchies. *Indagationes Mathematicae (Proceedings)* **80**(3), 195–207 (1977). [https://doi.org/10.1016/1385-7258\(77\)90067-1](https://doi.org/10.1016/1385-7258(77)90067-1), [http://dx.doi.org/10.1016/1385-7258\(77\)90067-1](http://dx.doi.org/10.1016/1385-7258(77)90067-1)
- 789 22. Krivine, J.L.: Introduction to Axiomatic Set Theory. Springer Netherlands (1971). <https://doi.org/10.1007/978-94-010-3144-8>
- 791 23. Kruskal, J.B.: The theory of well-quasi-ordering: A frequently discovered concept. *Journal of Combinatorial Theory, Series A* **13**(3), 297–305 (Nov 1972). [https://doi.org/10.1016/0097-3165\(72\)90063-5](https://doi.org/10.1016/0097-3165(72)90063-5), [http://dx.doi.org/10.1016/0097-3165\(72\)90063-5](http://dx.doi.org/10.1016/0097-3165(72)90063-5)
- 795 24. Kunen, K.: Set Theory. Elsevier (1980). [https://doi.org/10.1016/s0049-237x\(08\)x7037-5](https://doi.org/10.1016/s0049-237x(08)x7037-5), [http://dx.doi.org/10.1016/s0049-237x\(08\)x7037-5](http://dx.doi.org/10.1016/s0049-237x(08)x7037-5)
- 797 25. Kuske, D.: Theories of orders on the set of words. *RAIRO Theor. Informatics Appl.* **40**(1), 53–74 (2006). <https://doi.org/10.1051/ITA:2005039>, <https://doi.org/10.1051/ita:2005039>
- 800 26. Kříž, I., Thomas, R.: Ordinal Types in Ramsey Theory and Well-Partial-Ordering Theory, p. 57–95. Springer Berlin Heidelberg (1990). [https://doi.org/10.1007/978-3-642-72905-8\\_7](https://doi.org/10.1007/978-3-642-72905-8_7), [http://dx.doi.org/10.1007/978-3-642-72905-8\\_7](http://dx.doi.org/10.1007/978-3-642-72905-8_7)

- 803 27. Lopez, A.: Labelled Well Quasi Ordered Classes of Bounded Linear Clique-Width.  
804 In: Gawrychowski, P., Mazowiecki, F., Skrzypczak, M. (eds.) 50th International  
805 Symposium on Mathematical Foundations of Computer Science (MFCS 2025).  
806 Leibniz International Proceedings in Informatics (LIPIcs), vol. 345, pp. 70:1–70:17.  
807 Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2025).  
808 <https://doi.org/10.4230/LIPIcs.MFCS.2025.70>, [https://drops.dagstuhl.de/](https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.MFCS.2025.70)  
809 [entities/document/10.4230/LIPIcs.MFCS.2025.70](https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.MFCS.2025.70)
- 810 28. Nicolas, F., Pritykin, Y.: On uniformly recurrent morphic sequences. Inter-  
811 national Journal of Foundations of Computer Science **20**(05), 919–940 (Oct  
812 2009). <https://doi.org/10.1142/s0129054109006966>, [http://dx.doi.org/10.](http://dx.doi.org/10.1142/S0129054109006966)  
813 [1142/S0129054109006966](http://dx.doi.org/10.1142/S0129054109006966)
- 814 29. Rabin, M.O.: Decidability of second-order theories and automata on infinite trees.  
815 Transactions of the American Mathematical Society **141**, 1–35 (1969), [http://www.](http://www.jstor.org/stable/1995086)  
816 [jstor.org/stable/1995086](http://www.jstor.org/stable/1995086)
- 817 30. Schmidt, D.: The relation between the height of a well-founded partial ordering and  
818 the order types of its chains and antichains. Journal of Combinatorial Theory, Series  
819 B **31**(2), 183–189 (Oct 1981). [https://doi.org/10.1016/s0095-8956\(81\)80023-8](https://doi.org/10.1016/s0095-8956(81)80023-8),  
820 [http://dx.doi.org/10.1016/S0095-8956\(81\)80023-8](http://dx.doi.org/10.1016/S0095-8956(81)80023-8)
- 821 31. Schmitz, S.: Algorithmic Complexity of Well-Quasi-Orders. Habilitation à diriger  
822 des recherches, École normale supérieure Paris-Saclay (Nov 2017), [https://theses.](https://theses.hal.science/tel-01663266)  
823 [hal.science/tel-01663266](https://theses.hal.science/tel-01663266)
- 824 32. Schmitz, S.: The parametric complexity of lossy counter machines. In: Baier, C.,  
825 Chatzigiannakis, I., Flocchini, P., Leonardi, S. (eds.) 46th International Colloquium  
826 on Automata, Languages, and Programming (ICALP 2019). Leibniz International  
827 Proceedings in Informatics (LIPIcs), vol. 132, pp. 129:1–129:15. Schloss Dagstuhl  
828 – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2019). [https://doi.](https://doi.org/10.4230/LIPIcs.ICALP.2019.129)  
829 [org/10.4230/LIPIcs.ICALP.2019.129](https://doi.org/10.4230/LIPIcs.ICALP.2019.129), [https://drops.dagstuhl.de/entities/](https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2019.129)  
830 [document/10.4230/LIPIcs.ICALP.2019.129](https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2019.129)
- 831 33. Thomas, W.: Languages, automata, and logic. In: Rozenberg, G., Salomaa, A. (eds.)  
832 Handbook of formal languages, pp. 389–455. Springer (1997). [https://doi.org/](https://doi.org/10.1007/978-3-642-59136-5)  
833 [10.1007/978-3-642-59136-5](https://doi.org/10.1007/978-3-642-59136-5)
- 834 34. Tromp, J., Shallit, J.: Subword complexity of a generalized thue-morse word. Infor-  
835 mation Processing Letters **54**(6), 313–316 (Jun 1995). [https://doi.org/10.1016/](https://doi.org/10.1016/0020-0190(95)00074-m)  
836 [0020-0190\(95\)00074-m](https://doi.org/10.1016/0020-0190(95)00074-m), [http://dx.doi.org/10.1016/0020-0190\(95\)00074-M](http://dx.doi.org/10.1016/0020-0190(95)00074-M)

## A Proofs for Section 1

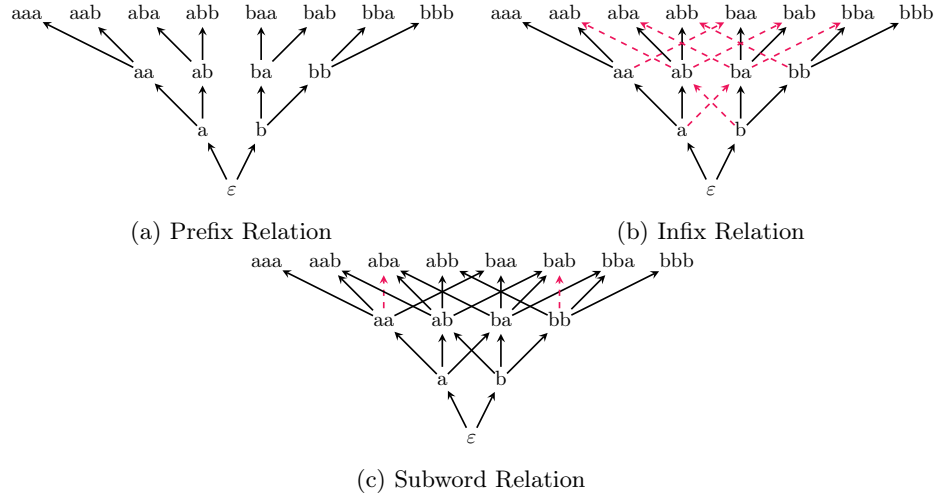


Fig. 3: A simple representation of the subword relation, prefix relation, and infix relation, on the alphabet  $\{a, b\}$  for words of length at most 3. The figures are Hasse Diagrams, representing the successor relation of the order. Furthermore, we highlight in dashed red relations that are added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



838 **B Proofs for Section 3**

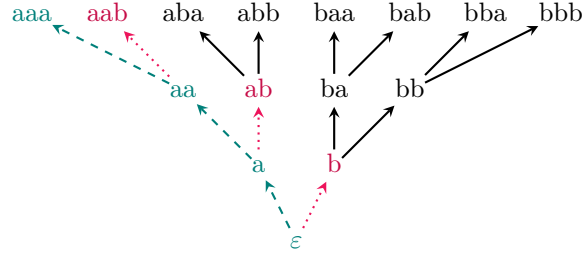


Fig. 4: An antichain branch for the language  $a^*b$ , represented in the tree of prefixes over the alphabet  $\{a, b\}$ . The branch is represented with dashed lines in turquoise, and the antichain is represented in dotted lines in blood-red.

839 *Proof (Proof of Lemma 3 as stated on page 7).* Assume that  $L$  contains an  
 840 antichain branch. Let us construct an infinite antichain as follows. We start with  
 841 a set  $A_0 \triangleq \emptyset$  and a node  $v_0$  at the root of the tree. At step  $i$ , we consider a word  
 842  $w_i$  such that  $v_i$  is a prefix of  $w_i$ , and  $w_i \in L \setminus B$ , which exists by definition of  
 843 antichain branches. We then set  $A_{i+1} \triangleq A_i \cup \{w_i\}$ . To compute  $v_{i+1}$ , we consider  
 844 the largest prefix of  $w_i$  that belongs to  $B$ , and set  $v_{i+1}$  to be the successor of this  
 845 prefix in  $B$ . By an immediate induction, we conclude that for all  $i \in \mathbb{N}$ ,  $A_i$  is an  
 846 antichain, and that  $v_i$  is a node in the antichain branch  $B$  such that  $v_i$  is not a  
 847 prefix of any word in  $A_i$ .

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

▷ Back to p.7

853 *Proof (Proof of Corollary 4 as stated on page 7).* If  $L$  is regular, then it is  
 854 MSO-definable, and there exists a formula  $\varphi(x)$  in MSO that selects nodes of  
 855 the tree of prefixes that belong to  $L$ . Now, to decide whether there exists an  
 856 antichain branch for  $L$ , we can simply check whether the following formula is  
 857 satisfied:

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

858 Because the above formula is an MSO-formula over the infinite  $\Sigma$ -branching tree,  
 859 whether it is satisfied is decidable as an easy consequence of the decidability of  
 860 MSO over infinite binary trees [29, Theorem 1.1].

▷ Back to p.7

861 *Proof (Proof of Theorem 5 as stated on page 7).* Assume that  $L$  is a finite  
 862 union of chains. Because the prefix relation is well-founded, and that finite unions  
 863 of chains have finite antichains, we conclude that  $L$  is well-quasi-ordered.

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

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

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

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

877 Hence,  $L$  is a finite union of chains.

878 **C Proofs for Section 4**

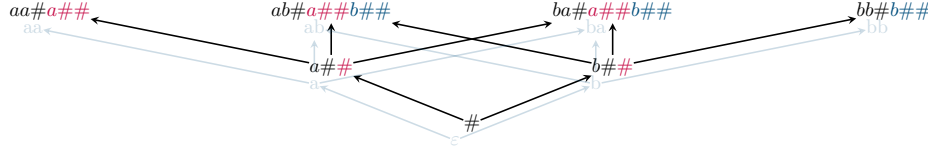


Fig. 5: Representation of the subword relation for  $\{a, b\}^*$  inside the infix relation for  $\{a, b, \#\}^*$  using a simplified version of Lemma 7, restricted to words of length at most 3.

879 *Proof (Proof of Lemma 10 as stated on page 9).* Let  $x \in \Sigma^+$  be a word, and let  
 880  $P_x$  be the (finite) set of all prefixes of  $x$ , and  $S_x$  be the (finite) set of all suffixes  
 881 of  $x$ . Assume that  $w \in \text{Pl}(x)$ , then  $w = ux^pv$  for some  $u \in S_x$ ,  $v \in P_x$ , and  $p \in \mathbb{N}$ .  
 882 We have proven that

$$\text{Pl}(x) \subseteq \bigcup_{u \in P_x} \bigcup_{v \in S_x} ux^*v.$$

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

▷ Back to p.9

890 *Proof (Proof of Lemma 12 as stated on page 9).* Note that the result is obvious  
 891 if  $k = 0$ , and therefore we assume  $k \geq 1$  in the following proof.

892 Let us construct a sequence of words  $(w_i)_{i \in \mathbb{N}}$ , where  $w_i \triangleq w[\mathbf{n}_i]$  for some  
 893 well-chosen indices  $\mathbf{n}_i \in \mathbb{N}^k$ . The goal being that if  $w[\mathbf{n}_i]$  is an infix of  $w[\mathbf{n}_j]$ ,  
 894 then it can intersect at most *two* iterated words, with an intersection that is  
 895 long enough to successfully apply Lemma 11. In order to achieve this, let us first  
 896 define  $s$  as the maximal size of a word  $v_i$  ( $1 \leq i \leq k$ ) and  $u_j$  ( $1 \leq j \leq k+1$ ).  
 897 Then, we consider  $\mathbf{n}_0 \in \mathbb{N}^k$  such that  $\mathbf{n}_0$  has all its components greater than  
 898  $s!$  and such that  $w[\mathbf{n}_0]$  belongs to  $L$ . Then, we inductively define  $\mathbf{n}_{i+1}$  as the  
 899 smallest vector of numbers greater than  $\mathbf{n}_i$ , such that  $w[\mathbf{n}_{i+1}]$  belongs to  $L$ , and  
 900 with  $\mathbf{n}_i$  having all components greater than  $2|w[\mathbf{n}_i]|$ .

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

$$\begin{aligned}
906 \quad & - w[\mathbf{n}_i] \sqsubseteq_{\text{infix}} v_\ell^{n_{j,\ell}} u_{\ell+1} v_{\ell+1}^{n_{j,\ell+1}}, \\
907 \quad & - w[\mathbf{n}_i] \sqsubseteq_{\text{infix}} u_\ell v_\ell^{n_{j,\ell}}, \\
908 \quad & - w[\mathbf{n}_i] \sqsubseteq_{\text{infix}} v_\ell^{n_{j,\ell}} u_{\ell+1}.
\end{aligned}$$

909 In the sake of simplicity, we will only consider one of the three cases, namely  
910  $w[\mathbf{n}_i] \sqsubseteq_{\text{infix}} v_\ell^{n_{j,\ell}} u_{\ell+1}$ , the other two being similar. Because the lengths used in  
911  $\mathbf{n}_i$  are all sufficiently large, we know that for every  $k$ ,  $v_k^{n_{i,k}}$  is an infix of a  $v_\ell^p$   
912 for some non-zero  $p$ . Therefore, we can apply Lemma 11 to conclude that there  
913 exists a word  $x$  such that every  $v_k$  is a power of a conjugate of  $x$  (a cyclic shift of  
914  $x$ ), and  $v_\ell$  is a power of  $x$ . We can therefore rewrite  $w[\mathbf{n}_i]$  as  $u_1(\sigma_1(x))^{n_{i,1}} u_2 \cdots$ ,  
915 where  $\sigma_k$  is some conjugacy operation (cyclic shift). Now, in order for  $w[\mathbf{n}_i]$  to be  
916 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  
917 of  $x$ , and that they align properly with the  $\sigma_k(x)$ 's to form an infix of some  
918 power of  $x$ , except for the last one. In particular,  $w[\mathbf{n}_i] \in \text{P}\downarrow(x)u_{\ell+1}$ , but also,  
919 every other choice of  $\mathbf{n}$  will lead to a word in  $\text{P}\downarrow(x)u_{\ell+1}$ , because the alignment  
920 constraints are stable under pumping.

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

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

927 *Proof (Proof of Lemma 13 as stated on page 9).* Let  $w_1, \dots, w_n$  be such that  
928  $L \subseteq w_1^* \cdots w_n^*$ . Let us define  $m \triangleq \max\{|w_i| \mid 1 \leq i \leq n\}$

929 Let  $w[\mathbf{k}] \triangleq w_1^{k_1} \cdots w_n^{k_n}$  be a map from  $\mathbb{N}^k$  to  $\Sigma^*$ . We are interested in the  
930 intersection of the image of  $w$  with  $L$ . Let us assume for instance that for all  
931  $\mathbf{k} \in \mathbb{N}^n$ , there exists  $\ell \geq \mathbf{k}$  such that  $w[\ell] \in L$ . Then, leveraging Lemma 12, we  
932 conclude that there exists  $x, y$  of size at most  $\max\{|w_i| \mid 1 \leq i \leq n\}$  such that  
933  $w[\mathbf{k}] \in \text{P}\downarrow(x) \cup \text{P}\downarrow(x) \text{P}\downarrow(y)$ , and we conclude that  $L \subseteq \text{P}\downarrow(x) \cup \text{P}\downarrow(x) \text{P}\downarrow(y)$ .

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

942 Now, admissible languages are ready to be pumped according to Lemma 12.  
943 For every admissible language, the size of a word that is not iterated is at most  
944  $N_0 \times m$  by definition, and we conclude that:

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

## 946 D Proofs for Section 5

947 *Proof (Proof of Corollary 15 as stated on page 10).* Because  $L \subseteq \downarrow_{\sqsubseteq_{\text{infix}}} L$ , the  
 948 right-to-left implication is trivial. For the left-to-right implication, let us assume  
 949 that  $L$  is a well-quasi-ordered language for the infix relation. Then  $L$  is included  
 950 in a finite union of products of chains for the prefix and suffix relations thanks  
 951 to Theorem 8:

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

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

$$\downarrow_{\sqsubseteq_{\text{infix}}} (S_i \cdot P_i) = (\downarrow_{\sqsubseteq_{\text{suffix}}} S_i) \cdot (\downarrow_{\sqsubseteq_{\text{prefix}}} P_i) \quad .$$

954 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  
 955 prefix of a word in  $P_i$ . Conversely, any such concatenations are infixes of a word  
 956 in  $S_i P_i$ .

957 As a consequence, we conclude that  $\downarrow_{\sqsubseteq_{\text{infix}}} L$  is itself included in a finite union  
 958 of products of chains. Furthermore, by definition of bounded languages,  $\downarrow_{\sqsubseteq_{\text{infix}}} L$   
 959 is also a bounded language. Hence, it is well-quasi-ordered by the infix relation  
 960 via Theorem 8. ▷ Back to p.10

961 *Proof (Proof of Lemma 22 as stated on page 12).* Let us assume that  $L$  is  
 962 infinite. The case when it is finite is similar, but will result in a finite word.

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

968 We can build a word  $w$  as the limit of the sequence  $(u_i)_{i \in I}$ . This word is  
 969 infinite or bi-infinite, and contains as infixes all the words  $u_i$  for  $i \in I$ . Because  
 970 every word of  $L$  is an infix of every  $u_i$  for a large enough  $I$ , one concludes that  $L$   
 971 is contained in the set of finite infixes of  $w$ . Conversely, every finite infix of  $w$  is  
 972 an infix of some  $u_i$  by definition of the limit construction, hence belongs to  $L$   
 973 since  $u_i \in L$  and  $L$  is downwards closed. ▷ Back to p.12

974 *Proof (Proof of Lemma 23 as stated on page 12).*

975 Assume that  $w$  is ultimately uniformly recurrent. Consider a sequence of  
 976 words  $(w_i)_{i \in \mathbb{N}}$  that are finite infixes of  $w$ . Because  $w$  is ultimately uniformly  
 977 recurrent, there exists a bound  $N_0$  such that  $w_{\geq N_0}$  is uniformly recurrent. Let  
 978  $i < N_0$ , we claim that, without loss of generality, only finitely many words in  
 979 the sequence  $(w_i)_{i \in \mathbb{N}}$  can be found starting at the position  $i$  in  $w$ . Indeed, if  
 980 it is not the case, then we have an infinite subsequence of words that are all  
 981 comparable for the infix relation, and therefore a good sequence, because the infix  
 982 relation is well-founded. We can therefore assume that all words in the sequence  
 983  $(w_i)_{i \in \mathbb{N}}$  are such that they start at a position  $i \geq N_0$ . But then they are all finite  
 984 infixes of  $w_{\geq N_0}$ , which is a uniformly recurrent word, whose set of finite infixes is  
 985 well-quasi-ordered (Theorem 17).

Conversely, assume that the set of finite infixes of  $w$  is well-quasi-ordered. Let us write  $\text{Rec}(w)$  the set of finite infixes of  $w$  that appear infinitely often. We can similarly define  $\text{Rec}(w_{\geq i})$  for any (infinite) suffix of  $w$ . The sequence  $R_i \triangleq \text{Rec}(w_{\geq i})$  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  $u \sqsubseteq_{\text{infix}} v$ , 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  $ux_iu \sqsubseteq_{\text{infix}} ux_ju$  for all  $i \leq j \in I$ . In particular,  $ux_iu \sqsubseteq_{\text{infix}} ux_ju$  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.

*Proof (Proof of Lemma 24 as stated on page 12).* 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  $\text{Infixes}(w)$  the set of finite infixes of  $w$ . Consider an infinite sequence of words  $(u_i)_{i \in \mathbb{N}}$  in  $\text{Infixes}(w)$ . If there is an infinite subsequence of words that are all in  $\text{Infixes}(w_+)$ , then there exists an increasing pair of indices  $i < j$  such that  $u_i \sqsubseteq_{\text{infix}} u_j$  because Theorem 17 applies to  $w_+$ . Similarly, if there is an infinite subsequence of words that are all in  $\text{Infixes}(w_-)$ , then there exists an increasing pair of indices  $i < j$  such that  $u_i \sqsubseteq_{\text{infix}} u_j$  because Theorem 17 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,  $u_i \sqsubseteq_{\text{infix}} u_j$ . We have proven that every infinite sequence of words in  $\text{Infixes}(w)$  is good, hence  $\text{Infixes}(w)$  is well-quasi-ordered.

Conversely, assume that  $\text{Infixes}(w)$  is well-quasi-ordered. In particular, the subset  $\text{Infixes}(w_+) \subseteq \text{Infixes}(w)$  is well-quasi-ordered. Similarly,  $\text{Infixes}(w_-)$  is well-quasi-ordered because the infix relation is compatible with mirroring. Applying Lemma 23, we conclude that both are ultimately uniformly recurrent words.

1029 *Proof (Proof of Lemma 25 as stated on page 13).* Let  $N_0$  be a bound such that  
 1030  $w_{\geq N_0}$  is uniformly recurrent. Let us write  $\text{Infixes}(w)$  the set of finite infixes of  $w$ .  
 1031 We prove that  $\mathfrak{w}(\text{Infixes}(w)) \leq \omega + N_0$ . Let  $u_1 \sqsubseteq_{\text{infix}} w$  be a finite word.

1032 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  
 1033 every word of size at least  $k$ . In particular, there is finite bound on the length  
 1034 of every sequence of incomparable elements starting with  $u_1$ . We conclude in  
 1035 particular that  $\text{Infixes}(w) \setminus \uparrow u_1$  has a finite ordinal width.

1036 Otherwise,  $u_1$  can only be found *before*  $N_0$ . In this case, we consider a second  
 1037 element of a bad sequence  $u_2 \sqsubseteq_{\text{infix}} w$ , which is incomparable with  $u_1$  for the infix  
 1038 relation. If  $u_2$  is an infix of  $w_{\geq N_0}$ , then we can conclude as before. Otherwise,  
 1039 notice that  $u_1$  and  $u_2$  cannot start at the same position in  $w$  (because they are  
 1040 incomparable). Continuing this argument, we conclude that there are at most  
 1041  $N_0$  elements starting before  $N_0$  at the start of any sequence of incomparable  
 1042 elements in  $\text{Infixes}(w)$ . We conclude that  $\mathfrak{w}(\text{Infixes}(w)) \leq \omega + N_0$ .

1043 Let us now justify that this bound is tight. The Thue-Morse sequence over  
 1044 a binary alphabet  $\{a, b\}$  has ordinal width  $\omega$  from Lemma 19. Given a number  
 1045  $N_0 \in \mathbb{N}$ , one can construct an arbitrarily long antichain of words for the infix  
 1046 relation by using a new letter  $c$ . When concatenating this (finite) antichain as  
 1047 a prefix of the Thue-Morse sequence, one obtains a new (infinite) word  $w$ . It is  
 1048 clear that the ordinal width of  $\text{Infixes}(w)$  is now at least  $\omega + N_0$ .

▷ Back to p.13

1049 *Proof (Proof of Lemma 26 as stated on page 13).* Given a bi-infinite word  
 1050  $w \in \Sigma^{\mathbb{Z}}$ , recall that we can consider  $w_+ \in \Sigma^{\mathbb{N}}$  and  $w_- \in \Sigma^{\mathbb{N}}$  the two infinite  
 1051 words obtained as follows: for all  $i \in \mathbb{N}$ ,  $(w_+)_i = w(i)$  and  $(w_-)_i = w(-i)$ . Note  
 1052 that the two share the letter at position 0.

1053 To obtain the upper bound of  $\omega \cdot 3$ , we can consider the same argument as for  
 1054 Lemma 25. We let  $N_0$  be such that  $w_{\geq N_0}$  and  $(w_-)_{\geq N_0}$  are uniformly recurrent  
 1055 words. In any sequence of incomparable elements of  $\text{Infixes}(w)$ , there are less than  
 1056  $N_0^2$  elements that are found in  $(w_{< N_0})_{> -N_0}$ . Then, one has to pick a finite infix  
 1057 in either  $w_{\geq N_0}$  or  $w_{\leq -N_0}$ . Because of Lemma 25, any sequence of incomparable  
 1058 elements of these two infinite words has length bounded based on the choice of  
 1059 the first element of that sequence. This means that the ordinal width of  $\text{Infixes}(w)$   
 1060 is at most  $\omega + \omega + N_0^2$ . We conclude that  $\mathfrak{w}(\text{Infixes}(w)) < \omega \cdot 3$ .

1061 Let us briefly argue that the bound is tight. Indeed, one can construct a  
 1062 bi-infinite word  $w$  by concatenating a reversed Thue-Morse sequence on a binary  
 1063 alphabet  $\{a, b\}$ , a finite antichain of arbitrarily large size over a distinct alphabet  
 1064  $\{c, d\}$ , and then a Thue-Morse sequence on a binary alphabet  $\{e, f\}$ . The ordinal  
 1065 width of the set of infixes of  $w$  is then at least  $\omega \cdot 2 + K$ , where  $K$  is the size of  
 1066 the chosen antichain, following the same argument as in the proof of Lemma 25,  
 1067 using Lemma 19.

▷ Back to p.13

1068 **Lemma 35.** *Given an automatic sequence  $w \in \Sigma^{\mathbb{N}}$ , one can decide whether it is*  
 1069 *ultimately uniformly recurrent.*

▷ Proven p.30



1070 *Proof (Proof of Lemma 35 as stated on page 29).* We can rewrite this as a  
 1071 question on the automatic sequence  $w$  as follows:

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

1072 Because  $w$  is computable by a finite automaton, one can reduce the above formula  
 1073 to a regular language, for which it suffices to check emptiness, which is decidable.  
 1074

1075 *Proof (Proof of Corollary 21 as stated on page 12).* It is always true that the  
 1076 ordinal height of a language over a finite alphabet is at most  $\omega$ . Let us now  
 1077 consider a well-quasi-ordered language  $L$  that is downwards closed for the infix  
 1078 relation. Applying Theorem 20, we can write  $L = \bigcup_{i=1}^n L_i$  where each  $L_i$  is the  
 1079 set of finite infixes of a finite, infinite or bi-infinite ultimately uniformly recurrent  
 1080 word  $w_i$ . We can then directly conclude that  $\mathfrak{w}(L_i)$  is less than  $\omega$  (in the case of a  
 1081 finite word), less than  $\omega \cdot 2$  (in the case of an infinite word thanks to Lemma 25),  
 1082 or less than  $3 \cdot \omega$  (in the case of a bi-infinite word, thanks to Lemma 26). In any  
 1083 case, we have the bound  $\mathfrak{w}(L_i) < \omega \cdot 3$ .

1084 Now,  $\mathfrak{w}(L) \leq \sum_{i=1}^n \mathfrak{w}(L_i) < \omega \cdot 3 < \omega^2$ . Finally, the inequality  $\mathfrak{o}(L) \leq$   
 1085  $\mathfrak{w}(L) \otimes \mathfrak{h}(L) < \omega \otimes \omega^2 = \omega^3$  allows us to conclude.

1086 The tightness of the bounds is a direct consequence of Lemma 26, and by  
 1087 considering a finite union of these examples over disjoint alphabets (or even,  
 1088 by considering a binary alphabet and using unambiguous codes to separate the  
 1089 different components).

▷ Back to p.29

▷ Back to p.12

1090 **E Proofs for Section 6**

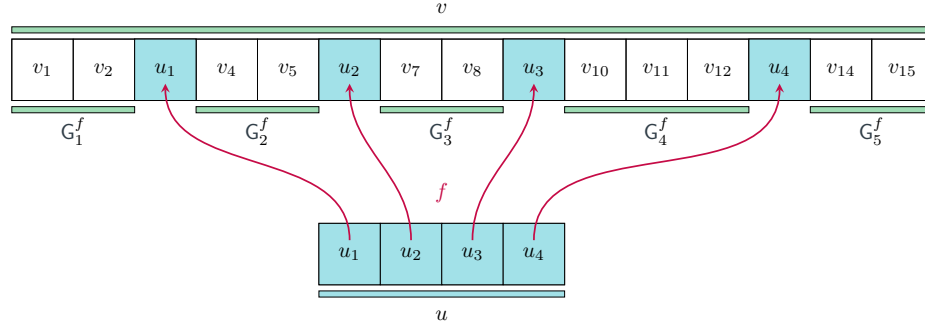


Fig. 6: The gap words resulting from a subword embedding between two finite words.

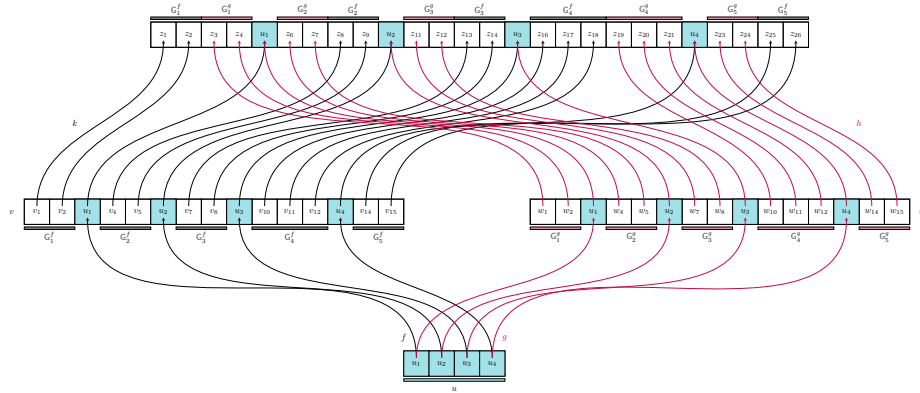


Fig. 7: We illustrate how embeddings  $f$  and  $g$  between runs of an amalgamation system can be glued together, seen on their canonical decomposition.

1091 For this paper to be self-contained, we will also recall how runs of a finite  
 1092 state automaton can be understood as an amalgamation system.

1093 *Example 36 ([5, Section 3.2]).* Let  $A = (Q, \delta, q_0, F)$  be a finite state automaton  
 1094 over a finite alphabet  $\Sigma$ . Let  $\Delta$  be the set of transitions  $(q_1, a, q_2) \in Q \times \Sigma \times Q$ ,  
 1095 and  $R \subseteq \Delta^*$  be the set of words over transitions that start with the initial state  
 1096  $q_0$ , end in a final state  $q_f \in F$ , and such that the end state of a letter is the  
 1097 start state of the following one. The canonical decomposition  $\text{can}$  is defined as

a morphism from  $\Delta^*$  to  $\Sigma^*$  that maps  $(q, a, p)$  to  $a$ . Because of the one-to-one correspondence of steps of a run  $\rho$  and letters in its canonical decomposition, we may treat the two interchangeably. Finally, given two runs  $\rho$  and  $\sigma$  of the automaton, we say that an embedding  $f \in \text{Hom}^*(\text{can}(\rho), \text{can}(\sigma))$  belongs to  $E(\rho, \sigma)$  when  $f$  is also defining an embedding from  $\rho$  to  $\sigma$  as words in  $\Delta^*$ .

The system  $(\Sigma, R, E, \text{can})$  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  $\text{Hom}^*(\text{can}(\rho), \text{can}(\sigma))$ , 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  $\Delta^*$  with  $\leq^*$  is a well-quasi-order according to Higman's Lemma [20].

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  $\rho$  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  $0 \leq i \leq |\text{can}(\rho)|$  and not just for one separately.

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

*Proof (Proof of Lemma 32 as stated on page 16).* Write  $u$  for  $G_\ell^f$  and  $v$  for  $G_\ell^g$ . 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  $L_i \in \{vvu^k, vu^k v, u^k vv\}$  and specifically  $L_\ell = vu^k v$ .

From Lemma 12, 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  $\text{Pl}(x)w\text{Pl}(y)$ .

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  $\{w_i \mid i \in I\} \subseteq \text{Pl}(x)w\text{Pl}(y)$ . This implies that either for infinitely many  $m \in \mathbb{N}$ ,  $u^m v \in \text{Pl}(y)$  or for infinitely many  $m$ ,  $vu^m \in \text{Pl}(x)$ .

In either case, we may conclude that either  $u \sqsubseteq_{\text{infix}} v$  or  $v \sqsubseteq_{\text{infix}} u$ : Let  $m, n \in \mathbb{N}$  such that  $m < n$  and  $u^m v, u^n v \in \text{Pl}(y)$  (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  $p \sqsubseteq_{\text{pref}} y, s \sqsubseteq_{\text{suff}} y$  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  $u^m v \in \text{Pl}(y)$ , it follows that  $v$  is a prefix of some word in  $(sp)^*$ . Hence either  $v$  is a prefix of  $u$  or  $u$  vice versa.

1136 *Proof (Proof of Theorem 29 as stated on page 15).* Assume that  $L$  is well-  
 1137 quasi-ordered by the infix relation, and obtained by an amalgamation system  
 1138  $(\Sigma, R, E, \text{can})$ .

1139 Let us consider the set  $M$  of minimal runs for the relation  $\leq_E$ , which is  
 1140 finite because the latter is a well-quasi-ordering. By Lemma 32, we know that  
 1141 for each minimal run  $\rho \in M$ , each gap language  $L_i^\rho$  of  $\rho$  is totally ordered by  
 1142  $\sqsubseteq_{\text{infix}}$ . Adapting the proof of language boundedness from [5, Section 4.2], we may  
 1143 conclude that  $L_i^\rho \subseteq P\downarrow(w)$  for some  $w \in L_i^\rho$ . As  $P\downarrow(w)$  is language bounded and  
 1144 this property is stable under subsets, concatenation and finite union, we can  
 1145 conclude that  $L$  is bounded as well.

▷ Back to p.15

1146 *Proof (Proof of Lemma 33 as stated on page 17).* It is clear that Item i  $\Rightarrow$   
 1147 Item ii because regular languages are recognized by finite automata, and finite  
 1148 automata are a particular case of amalgamation systems. The implication Item ii  
 1149  $\Rightarrow$  Item iii is the content of Theorem 29. The implication Item iii  $\Rightarrow$  Item iv  
 1150 is Lemma 13. Finally, the implication Item iv  $\Rightarrow$  Item i is simply because a  
 1151 downwards closed language that is a finite union of products of chains is a regular  
 1152 language.

1153 Indeed, assume that  $L$  is downwards closed and included in a finite union  
 1154 of sets of the form  $P\downarrow(x)uP\downarrow(y)$  where  $x, y, u$  are possibly empty words. We can  
 1155 assume without loss of generality that for every  $n$ ,  $x^nuy^n$  is in  $L$ , otherwise, we  
 1156 have a bound on the maximal  $n$  such that  $x^nuy^n$  is in  $L$ , and we can increase  
 1157 the number of languages in the union, replacing  $x$  or  $y$  with the empty word  
 1158 as necessary. Let us write  $L' \triangleq \bigcup_{i=1}^k x_i^* u_i y_i^*$ . Then,  $L' \subseteq L$  by construction.  
 1159 Furthermore,  $L \subseteq \downarrow L'$ , also by construction. Finally, we conclude that  $L = \downarrow L'$   
 1160 because  $L$  is downwards closed. Now, because  $L'$  is a regular language, and  
 1161 regular languages are closed under downwards closure, we conclude that  $L$  is a  
 1162 regular language.

▷ Back to p.17

1163 Let us briefly recall that a *rational transduction* is a relation  $R \subseteq \Sigma^* \times \Gamma^*$   
 1164 such that there exists a finite state automaton that reads pairs of letters  $(a, b) \in$   
 1165  $(\Sigma \cup \{\epsilon\}) \times (\Gamma \cup \{\epsilon\})$  and recognizes  $R$ . A class of languages  $\mathcal{C}$  is *closed under*  
 1166 *rational transductions* if for every  $L \in \mathcal{C}$  and every rational transduction  $R$ , the  
 1167 language  $R(L) \triangleq \{v \in \Gamma^* \mid \exists u \in L, (u, v) \in R\}$  also belongs to  $\mathcal{C}$ .

1168 *Proof (Proof of Theorem 30 as stated on page 15).* We first show Item 3  $\Rightarrow$   
 1169 Item 1. We aim to make the inclusion test of Equation (1) of Theorem 8 effective.  
 1170 Let  $R(n, m, N_0) \triangleq \bigcup_{x, y \in \Sigma^{\leq n}} \bigcup_{u \in \Sigma^{\leq m \times N_0}} P\downarrow(x)uP\downarrow(y) \cup P\downarrow(x)u \cup uP\downarrow(x)$ . For  
 1171 any concrete values of the bounds  $n$ ,  $m$ , and  $N_0$ , this language is regular. The  
 1172 map  $L \mapsto L \cap \Sigma^* \setminus R(n, m, N_0)$  is a rational transduction because  $\Sigma^* \setminus R(n, m, N_0)$   
 1173 is regular. Since  $\mathcal{C}$  is closed under rational transductions, we can therefore reduce  
 1174 the inclusion to emptiness of this language. However, we need to find these  
 1175 bounds first.

1176 To determine values for  $n$  and  $m$ , we first test if  $L$  is bounded. Since emptiness  
 1177 is decidable, we can apply the algorithm in [5, Section 4.2] to decide if  $L$  is bounded.  
 1178 If  $L$  is bounded, this algorithm yields words  $w_1, \dots, w_n$  such that  $L \subseteq w_1^* \cdots w_n^*$   
 1179 and therefore yields also the bounds in questions:  $n$  is the number of words, and

1180  $m$  is the maximal length of a word  $w_i$  where  $1 \leq i \leq n$ . If  $L$  is not bounded, then  
 1181  $L$  cannot be well-quasi-ordered by the infix relation because of Theorem 29 and  
 1182 we immediately return false.

1183 To determine the value for  $N_0$ , we then compute the downward closure (with  
 1184 respect to subwords) of  $L$ . This is effective and yields a finite-state automaton.  
 1185 Recall that  $N_0$  is the maximum number of repetitions of a word  $w_i$  that can not  
 1186 be iterated arbitrarily often. This value is therefore bounded above by the length  
 1187 of the longest simple path in this automaton.

1188 Item 1  $\Rightarrow$  Item 2. We just consider the transduction  $f$  that maps every  
 1189 word  $w$  to  $\#w$  where  $\#$  is a fresh symbol. Then, for any language  $L \in \mathcal{C}$ ,  $L$  is  
 1190 well-quasi-ordered by prefix if and only if  $f(L)$  is well-quasi-ordered by infix.

1191 Item 2  $\Rightarrow$  Item 3. We consider the transduction  $R \triangleq \Sigma^* \times \{a, b\}^*$ . Then for  
 1192 any language  $L \in \mathcal{C}$ , the image of  $L$  through  $R$  is well-quasi-ordered by prefix if  
 1193 and only if  $L$  is empty.

▷ Back to p.15

## 1194 F Proofs for Section 7

1195 **Lemma 37.** *Let  $L$  be a downwards closed language that is well-quasi-ordered by*  
 1196 *the infix relation. Then, the following are equivalent:*

- 1197 1.  $L$  has bounded factor complexity,
- 1198 2.  $L$  has finite ordinal width,
- 1199 3.  $L$  is a finite union of chains,
- 1200 4.  $L$  is a finite union of languages of the form  $\text{Infixes}(w)$  where  $w$  is an ultimately
- 1201 periodic word.

1202 *Proof.* First, Item 3  $\iff$  Item 2 is a standard fact regarding ordinal width.

1203 Then, Item 4  $\Rightarrow$  Item 1 is clear because ultimately periodic words have  
 1204 bounded factor complexity.

1205 In turn, Item 1  $\Rightarrow$  Item 2 is also clear because unbounded factor complexity  
 1206 implies the existence of arbitrarily large antichains.

1207 Finally, Item 2  $\Rightarrow$  Item 4 is a direct consequence of Theorem 20 and the fact  
 1208 that bounded factor complexity implies that the (bi)infinite words describing the  
 1209 language are ultimately periodic.