Master's Thesis

On Equivalence Classes of Simon's Congruence and Shuffles

Jonas Höfer

Kiel, September 2022

Department of Computer Science
Kiel University

Supervisor Prof. Dr. Dirk Nowotka
Dr. Pamela Fleischmann

by Jonas Höfer
On Equivalence Classes of Simon's Congruence and Shuffles
Master's Thesis, Kiel University, 2022

Typeset with KOMA-Script and LaTeX

Abstract

A word is a *subword*, *subsequence*, *scattered subword*, or *scattered factor* of a word w, if it is obtained by deleting any number of letters from w while preserving the order of the remaining letters. A word w is called *ℓ-universal* if all words of length ℓ are subwords of w. Two words are called k Simon congruent if they share all subwords of up to length k. This relation was introduced by Simon (1972) and is a congruence relation of finite index. The exact structure of the classes and an exact description of the index depending on $|\Sigma|$ and k are open questions posed by Sakarovitch and Simon (1997). In this thesis, we investigate the subword structures in words and in particular the structure of the congruence classes of this relation.

We continue a line of research by Fleischmann et al. (2022) where we characterize classes with a fixed number of absent subwords which are of minimal length. Furthermore, we investigate the $\alpha\beta$-factorization introduced in the same paper. For this factorization, we show a number of necessary properties of the factors, and derive results which characterize words in terms of their factors. Among them, we prove a theorem characterizing the equivalence of ℓ-universal words in terms of 1-universal words. Furthermore, we apply these results to binary words. In this special case, we obtain a full characterization of the classes and a formula for the index of the congruence.

The *shuffle* of two words u, v is the set $u \sqcup\!\sqcup v$ of all words obtained by interleaving u and v while preserving the order of the letters with respect to u and v. This operation extends naturally to sets, by shuffling all pairs of elements and unioning the results. Subwords and shuffles are linked by being quasi-inverse to each other. In this thesis, we focus mostly on so-called *shuffle squares* and *reverse shuffle squares*. These are words arising by shuffling a word v with itself or its reverse respectively. The word v is known as *shuffle root* or *reverse shuffle root* respectively and in general not unique. These words where already investigated by Henshall, Rampersad, and Shallit (2012).

We contrast the shuffle of a set with itself $L \sqcup\!\sqcup L$ and the union of all shuffles of its element with themselves $\bigcup_{v \in L} v \sqcup\!\sqcup v$. We show that these sets only coincide if the considered set L was a singleton or empty, that is, if they coincide trivially by definition. Moreover, we also show a similar result for reverse shuffle squares. Furthermore, we investigate the complexity of the language of all shuffle squares. We conjecture that the language is not an indexed language in the sense of Aho (1968) and start an investigation, considering two necessary conditions for a language to be indexed. Both yield negative results but one of them provides a more complex combinatorial property of the language of shuffle squares. Lastly, we show that a set of all words with a common shuffle root determines this root uniquely, that is, that the mapping $v \mapsto v \sqcup\!\sqcup v$ is injective.

Zusammenfassung

Ein *Teilwort*, eine *Teilfolge* oder ein *gestreuter Faktor* eines Wortes w, ist ein Wort, welches man durch Löschen von Buchstaben von w erhält. Ein Wort w heißt ℓ-universal, wenn alle Wörter der Länge ℓ Teilworte von w sind. Zwei Wörter u, v sind k Simon kongruent, wenn sie dieselben Teilwörter der Länge n kleiner gleich k besitzen. Diese Relation wurde von Simon (1972) eingeführt und ist von endlichem Index. In dieser Arbeit untersuchen wir die Kongruenzklassen dieser Relation.

Wir führen den Ansatz von Fleischmann et al. (2022) fort und untersuchen Klassen von Wörtern mit einer festen Anzahl an Teilworten der Länge k. Des Weiteren untersuchen wir die von Fleischmann et al. (2022) eingeführte $\alpha\beta$-Faktorisierung. Wir zeigen eine Reihe von nötigen Eigenschaften für die einzelnen Faktoren der Faktorisierung und in Spezialfällen charakterisieren wir Klassen von Wort durch Eigenschaften ihrer Faktoren. Unter anderem Charakterisieren wir die Äquivalenz von ℓ-universaler Wörter durch die Äquivalenz ihrer 1-universellen Faktoren. Außerdem verwenden wir diese Resultate im Spezialfall binärer Wörter. Für diesen Spezialfall zeigen wir eine vollständige Charakterisierung der Kongruenzklassen und berechnen den Index der Relation.

Das Mischprodukt von zwei Wörtern u und v ist die Menge $u \sqcup\!\sqcup v$ aller Wörter, welche durch Bogenmischen oder ineinanderschieben von u und v entstehen. Insbesondere ändert sich die interne Reihenfolge der Buchstaben von u und v nicht. Das Mischprodukt zweier Sprachen ist die Vereinigung der Mischprodukte von allen Paaren von Elementen der zwei Sprachen. Mischen von Wörtern und Teilworte sind verwandte, quasi-inverse Ideen. In dieser Arbeit betrachten wir insbesondere *Mischproduktquadrate*, also Wörter w mit $w \in v \sqcup\!\sqcup v$ für ein Wort v. v ist eine *Mischwurzel* von w und im Allgemeinen nicht eindeutig. Wörter w mit $w \in v \sqcup\!\sqcup v^R$ für ein Wort v *umgekehrte Mischproduktquadrate*.

Wir betrachten als Erstes eine Reihe kombinatorische Eigenschaften von Mischproduktquadrate. Konkret vergleichen wir für eine Menge L, das Mischprodukt der Mengen mit sich selbst $L \sqcup\!\sqcup L$ mit der Menge aller Mischproduktquadrate ihrer Elemente $\bigcup_{v \in L} v \sqcup\!\sqcup v$. Wir zeigen, dass diese Mengen nur in trivialen Fällen, wenn $|L| \leq 1$ übereinstimmen. Außerdem zeigen wir eine analoge Eigenschaft für umgekehrte Mischproduktquadrate. Des Weiteren betrachten wir die sprachtheoretische Komplexität der Menge alle Mischproduktquadrate. Wir vermuten, dass diese Sprache keine indexierte Sprache nach Aho (1968) ist. Wir betrachten zwei nötige Bedingungen für indexierte Sprachen. Beide liefern keine Antwort bezüglich der Vermutung, aber eine der zwei Bedingungen liefert eine kombinatorische Eigenschaft der Sprach aller Mischproduktquadrate. Ferner zeigen wir, dass die Abbildung $v \mapsto v \sqcup\!\sqcup v$ injektiv ist, also ein Wort durch die Menge all seiner Mischproduktquadrate eindeutig bestimmt ist.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

A *subword*, *subsequence*, *scattered subword* or *scattered factor* of a word w is a word that is obtained by deleting any number of letters from w while preserving the order of the remaining letters. For example, `oiaoi` and `cmbntrcs` are both subwords of `combinatorics`. In contrast to a factor, like `combinat`, a subword is not necessarily continuous. Furthermore, note that a subword v can occur in different ways inside a word w, for example, `ab` occurs in `aab` as $\texttt{a}\underline{\texttt{ab}}$ and $\texttt{aa}\underline{\texttt{b}}$ (visualized by the underlines). We denote that a word v is a subword of some word w by $v \preceq w$ as this symbol is usually used for subsequences in the infinite case.

The *shuffle* of two words u and v is the set of all words obtained by interleaving u and v while preserving the order of the letters with respect to u and v. For example, the shuffle of `ab` and `ca`, is given by $\{\texttt{abca}, \texttt{acab}, \texttt{acba}, \texttt{caab}, \texttt{caba}\}$. Note that this corresponds exactly to all possible stacks of cards obtained by riffle shuffling two stacks of cards. Furthermore, note that the number of elements of a shuffle, depends on the letters of the words because different ways of interleaving two words, may yield the same word, for example, $\overline{\texttt{ca}}\underline{\texttt{ab}} = \overline{\texttt{c}}\underline{\texttt{a}}\overline{\texttt{a}}\underline{\texttt{b}}$. The shuffle of words u, v is usually denoted by $u \sqcup\!\sqcup v$.[1] The shuffle extends naturally to sets of words or *languages* by shuffling all elements pairwise. Thus, it forms an operation on classes of languages similar to, for example, union, intersection and complementation.

Subwords and shuffles are linked by being quasi-inverse operations to each other. For all words $w \in u \sqcup\!\sqcup v$ we have $u, v \preceq w$. Furthermore, the set of all words that contain v as a subword is exactly given by $v \sqcup\!\sqcup \Sigma^*$. The subword order appears in many contexts. One of the more well-known is a lemma by Higman [17], which states if $(A, \leq)$ is well-quasi ordered,[2] then so is $(A^*, \leq^*)$ where $u \leq^* w$ if there exists a subword $v \preceq w$ of the same length as u, such that u is component-wise smaller than v with respect to $\leq$. This lemma implies that there does not exist an infinite set of words, where all elements are pairwise incomparable with respect to the subword order. Both subwords and shuffles arise in combinatorics on words [22] and formal language theory [29, 28]. In this thesis, we focus on the congruence relation $\sim_k$ for $k \in \mathbb{N}_0$ which is known as Simon's congruence and also

[1] This symbol is similar to the Cyrillic letter ш *sha*.

[2] A well-quasi order is a preorder where no infinite descending chains exist, that is, for every sequence $(s_i)_{i \in \mathbb{N}}$ exist $i < j$ such that $s_i \leq s_j$. Equivalently, a preorder is well-quasi if no infinite *antichains* (subsets of incomparable elements) exist.

appears in this context. For two words, we have $u \sim_k v$, if and only if, they share all subwords of up to length k. Unions of the congruence classes of this relation are used to form the *piecewise testable languages*, a subclass of the regular languages which was first studied by Simon [28]. Equivalently, this class of languages can be defined as all languages that are boolean combinations of languages of the form $v \sqcup\!\sqcup \Sigma^*$ [22, Proposition 6.2.5]. A long-standing question, posed by Sakarovitch and Simon [22, Chapter 6], is the exact structure of the congruence classes of $\sim_k$ and the index of congruence relation itself. Two existing results include a characterization of the congruence in terms of a special common upper bound of two words [29, Lemma 6], as well as, a characterization of the (not unique) minimal elements of the congruence classes [22, Theorem 6.2.9] [28, 8]. The index of the relation was described asymptotically by Karandikar, Kufleitner, and Schnoebelen [20] but currently no exact formula is known. Fleischmann et al. [10] investigate the class of Simon's congruence separated by the number of absent subwords, characterize the classes for arbitrary alphabets for some fixed numbers of absent subwords and give explicit formulas for these subsets. Linked to Simon's congruence is the notion of *scattered factor* or *subword universality* [2]. A word w is called *ℓ-universal* if it has all subwords of length ℓ. Barker et al. [2] and Fleischmann, Germann, and Nowotka [9] study the universality of words, as well as, how the universality of a word changes when considering repetitions of a word.

The most common algorithmic problems regarding Simon's congruence are testing whether two words u, v are congruent for a fixed k and the optimization problem of finding the largest k such that they are congruent. The former was approached by finding the (lexicographical least element of the) minimal elements of the congruence classes of u and v. This was done by Fleischer and Kufleitner [8] and improved by Barker et al. [2]. The latter was approached in the binary case by Hébrard [15], using the characterization above by Simon [30], and was recently solved in linear time using a new approach by Gawrychowski et al. [12].

In this thesis, regarding the shuffle operation, we focus mostly on so-called *shuffle squares* and *reverse shuffle squares*. These are words w such that there exists a word v with $w \in v \sqcup\!\sqcup v$ and $w \in v \sqcup\!\sqcup v^R$ respectively. This operation is also studied from a language theoretic point of view by Henshall, Rampersad, and Shallit [16]. They show that even for a regular language L, the language $\bigcup_{v \in L} v \sqcup\!\sqcup v$ does not need to be context free. Shuffle squares are also studied from an algorithmic point of view by Buss and Soltys [6] and Bulteau and Vialette [5]. They show that deciding whether a word w is a shuffle square is NP-hard, even in the case of binary words.

Our Contribution We show a number of combinatorial results regarding shuffle squares and reverse shuffle squares. In particular, we show that arbitrary (reverse) shuffles of a set L with itself, $L \sqcup\!\sqcup L$, only coincide with the set of its (reverse) shuffle squares $\bigcup_{v \in L} v \sqcup\!\sqcup v$ in the trivial cases. Furthermore, we continue a language theoretic investigation of the language of all shuffle squares. We conjecture that the language of shuffle squares is not an indexed language and consider two necessary conditions. Both yield negative results regarding the language theoretic question but yield properties of the language of shuffle

squares.

We study the $\alpha\beta$-factorization which was used by Fleischmann et al. [10], We continue their approach and characterize the classes of words with 2 and 3 absent subwords respectively. Furthermore, we investigate the $\alpha\beta$-factorization as an object of independent interest and give necessary and sufficient conditions for the equivalence of word in terms of their single factors. We use these results to characterize the classes of binary words and calculate the index in this special case. Lastly, we show that the equivalence of words can be reduced to their equivalence of 1-universal words by decomposing them into $\alpha\beta\alpha$-triple.

Structure of the Work In Chapter 2 we establish a common notation and introduce subwords and shuffles. The main part of the thesis is split into two chapters focusing on shuffles and classes of Simon's congruence respectively.

In Chapter 3, we focus on the results regarding shuffles and in particular sets of shuffle squares. After some preliminary results in Section 3.1, we start with our results regarding (reverse) shuffle squares of sets in Section 3.2. Afterwards, we show the language theoretic motivated results about sets of shuffle squares in Sections 3.3 and 3.4.

In Chapter 4, we establish our results on equivalence classes of Simon's congruence. The chapter is organized into the three classes of results. First in Section 4.1, we introduce the $\alpha\beta$-factorization and use this factorization to give our results regarding words with a fixed number of subwords. Second, in Sections 4.2 and 4.3, we give some general results about the $\alpha\beta$-factorization and use these to characterize classes in the case of binary alphabets and calculate the index of the congruence in this special case. Third, in Section 4.4, we show our result on $\alpha\beta$-factorization which characterizes classes of ℓ-universal words in terms of their $\alpha\beta\alpha$-factors, that is, 1-universal words.

In Chapter 5, we conclude and give ideas for further research and open questions.

Chapter 2

Preliminaries

In this chapter we cover some preliminary definitions and notations, regarding combinatorics on words in general, as well as, subwords and shuffles in particular. For a more detailed introduction we refer the reader to the monograph by Lothaire [22].

2.1 General Notation

We use standard set theoretic notation and denote subsets by $\subseteq$ and proper subsets by $\subset$. Furthermore, we mark that a union is disjoint by writing $\sqcup$ instead of $\cup$. Let A be a set, then we denote its power set by 2^A and the set of its subsets of cardinality n by $\binom{A}{n}$. If the referenced superset is clear from context, we denote the complement of $A \subseteq U$ by $A^{\complement}$.

Let $\mathbb{N}$ be the set of natural numbers starting with 1 and let $\mathbb{N}_0 \coloneqq \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}_0$ define $[n] \coloneqq \{1 \dots n\}$ and $[n]_0 \coloneqq [n] \cup \{0\}$. If $j < i$ then we let $\{i, \dots, j\} = \emptyset$ and thus if $n = 0$ we have $[n] = \emptyset$.

Definition 2.1. An *alphabet* is a finite set whose elements are called *letters*. We denote by Σ_k for $k \in \mathbb{N}_0$ the alphabet with exactly k letters. For the set Σ_2, we denote for each $\mathtt{x} \in \Sigma_2$ the unique second element by $\overline{\mathtt{x}}$. Denote by $\Sigma^* \coloneqq \bigcup_{n=0}^{\infty} \Sigma^n$ the set of *finite sequences* or *(finite) words* over Σ. We denote elements of Σ^* by $a_1 a_2 \dots a_n$ instead of $(a_1, a_2, \dots, a_n)$ for $a_i \in \Sigma$ for all $i \in [n]$, identify Σ and the subset of one element sequences, and denote the unique sequence of length 0 by ε. Furthermore, define $\Sigma^+ \coloneqq \Sigma^* \setminus \{\varepsilon\}$ as the set of *non-empty words* over Σ. If $\Sigma = \{\sigma\}$ is a singleton set, we may omit the braces and write σ^* and σ^+ instead of $\{\sigma\}^*$ and $\{\sigma\}^+$ respectively. The operation of *concatenation*

$$\cdot : \Sigma^* \times \Sigma^* \to \Sigma^*, (u, v) \mapsto u \cdot v \coloneqq uv$$

is a binary, associative operation on Σ^*.

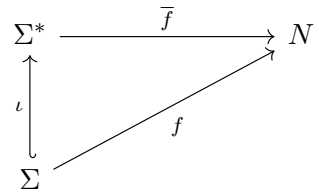
Let $w = xyz$ for $x, y, z \in \Sigma^*$. Then y is called an *infix* or *factor* of w. If $x = \varepsilon$ or $z = \varepsilon$ then y is a *prefix* or *suffix* of w respectively. If x is a factor, prefix or suffix of w and $x \neq w$, then x is called a *proper factor, proper prefix*, or *proper suffix* respectively. Denote by $\mathrm{Fact}(w)$ the set of all factors of w, by $\mathrm{Pref}(w)$ the set of all prefixes of w, and by $\mathrm{Suff}(w)$ the set of all suffixes of w. For a prefix x of $w = xy$ define $x^{-1}w = y$. Given $L \subseteq \Sigma^*$ and $w \in \Sigma^*$ define *the left quotient of L by w* as $w^{-1}L = \{u \mid wu \in L\}$. Define

wx^{-1} and the *right quotient of L by w*, Lw^{-1} analogously. Note that $x^{-1}w$ and wx^{-1} are only defined if x is a prefix or suffix of w respectively.

Let $w = w_1 w_2 \cdots w_n \in \Sigma^*$ with $w_i \in \Sigma$ for all $i \in [n]$. Then define $w^R \coloneqq w_n w_{n-1} \cdots w_2 w_1$. Furthermore, for $L \subseteq \Sigma^*$ define $L^R \coloneqq \{w^R \mid w \in L\}$.

Remark 2.1.1. The set of words Σ^* is a monoid under concatenation with neutral element ε. In particular, $(\Sigma^*, \cdot, \varepsilon)$ is a *free monoid over* Σ (under the inclusion mapping $\iota_\Sigma^{\Sigma^*} : \Sigma \hookrightarrow \Sigma^*$). This means that there exists a bijection $\bar{\cdot} : N^\Sigma \cong \mathrm{Hom}(\Sigma^*, N)$, between the set of mappings out of Σ into the underlying set N of a monoid $(N, \star, e_N)$ and the set of monoid homomorphisms out of $(\Sigma^*, \cdot, \varepsilon)$ into the monoid $(N, \star, e_N)$ such that the following diagram commutes.

$$(\Sigma, \cdot, \varepsilon) \dashrightarrow (N, \star, e_N)$$



The isomorphism is given by

$$\bar{\cdot} : \Sigma^N \to \mathrm{Hom}(\Sigma^*, N), f \mapsto (a_1 a_2 \cdots a_n \mapsto f(a_1) \star f(a_2) \star \cdots \star f(a_n)).$$

In particular, each monoid homomorphism out of the monoid Σ^* is completely specified by a function out of Σ into a monoid. We identify f and $\overline{f}$. Most functions out of Σ^* are defined this way, and we will implicitly use the fact that they are homomorphisms.

Definition 2.2. For a subset $\Omega \subseteq \Sigma$ define $\pi_\Omega : \Sigma^* \to \Omega^*$, the *projection onto* Ω, as the unique extension of

$$\pi_\Omega : \Sigma \to \Omega^*, \sigma \mapsto \pi_\Omega(\sigma) \coloneqq \begin{cases} \sigma & \sigma \in \Omega, \\ \varepsilon & \sigma \notin \Omega. \end{cases}$$

If $\Omega = \{\mathtt{a}\}$ is a singleton set, we write $\pi_\mathtt{a}$ instead of $\pi_{\{\mathtt{a}\}}$.

For $w \in \Sigma^*$, denote the set of all letters occurring in w or the *alphabet of* w by $\mathrm{alph}(w)$. Formally, define $\mathrm{alph}(w)$ as the unique homomorphism into $(2^\Sigma, \cup, \emptyset)$ mapping each letter $\sigma \in \Sigma$ to the singleton set containing it.

We denote the number of letters in or the *length* of a word $w \in \Sigma^*$ by $|w|$. More formally, define $|\cdot| : \Sigma^* \to \mathbb{N}_0$ as the morphism into the monoid $(\mathbb{N}_0, +, 0)$, mapping each letter to 1. For some $k \in \mathbb{N}_0$ the set of words over Σ with length less than k, less than or equal to k, greater than k, and greater than or equal to k by $\Sigma^{<k}, \Sigma^{\leq k}, \Sigma^{>k}$ and $\Sigma^{\geq k}$ respectively. For $\ell \in \mathbb{N}_0$ and $w \in \Sigma^*$ define by $\mathrm{pref}_\ell(w)$ the unique longest element of $\mathrm{Pref}(w) \cap \Sigma^{\leq \ell}$. Define $\mathrm{suff}_\ell(w)$ analogously.

Let $w \in \Sigma^*$ and $\mathtt{a} \in \Sigma$, then denote the number of occurrences of $\mathtt{a}$ in w by $|w|_\mathtt{a} \coloneqq |\pi_\mathtt{a}(w)|$. Fix a total order $\leq$ on the elements of Σ such that $\sigma_1 < \sigma_2 < \cdots < \sigma_{|\Sigma|}$ for $\sigma_i \in \Sigma$ for all $i \in [|\Sigma|]$. Define the *Parikh mapping* [25, Definition 13] as the unique extension of

$$\mathbf{p} : \Sigma \to \mathbb{N}_0^{|\Sigma|}, \sigma_i \mapsto \mathbf{p}(\sigma_i) \coloneqq \mathbf{e}_i,$$

where $\mathbf{e}_i$ denotes the i^{th} unit vector.[1] Furthermore, for some $w \in \Sigma^*$ call $\mathbf{p}(w)$ the *Parikh vector of* w. Note that $\mathbf{p}(w) = (|w|_{\sigma_1}, |w|_{\sigma_2}, \ldots, |w|_{\sigma_{|\Sigma|}})^\intercal$ where $\cdot^\intercal$ denotes the transposition.

2.2 Subwords

Definition 2.3. Let $w \in \Sigma^*$ and $i \in [|w|]$. Then $w[i]$ denotes the i^{th} letter of w. Define $w[\emptyset] \coloneqq \varepsilon$. Let $I \subseteq [|w|]$ be a non-empty subset of indices of w and define inductively $w[I] \coloneqq w[\min I] \cdot w[I \setminus \{\min I\}]$. When writing a set in the brackets we may omit the braces, for example, for $i, j \in \mathbb{N}$ we may write $w[i..j]$ instead of $w[\{i \ldots j\}]$. Furthermore, when writing a set of indices as a family, $I = \{i_k\}_{k \in [|I|]}$, we always assume that the indexation is ascending unless stated otherwise.

Definition 2.4 (Subword)**.** Let $w \in \Sigma^*$, $w_0, w_1, w_2, \ldots, w_m \in \Sigma^*$ and $x_1, x_2, \ldots, x_m \in \Sigma$ such that $w = w_0 x_1 w_1 x_2 \ldots x_m w_m$, then $x_1 x_2 \ldots x_m$ is called a *subword* of w. This is equivalent to the existence of a set $O \subseteq [|w|]$ such that $v = w[O]$. We call O an *occurrence of v in w*.

We denote the set of all subwords of w with $\mathrm{SubWords}(w)$ and write $x \preceq w$ for $x \in \mathrm{SubWords}(w)$. If $x \preceq w$ and $x \neq w$ we call x a *proper subword of* w and write $x \prec w$. Note that ε is a proper subword of every element of Σ^+. Furthermore, for $k \in \mathbb{N}_0$ we define $\mathrm{SubWords}_k(w) \coloneqq \mathrm{SubWords}(w) \cap \Sigma^k$ and $\mathrm{SubWords}_{\leq k}(w) \coloneqq \mathrm{SubWords}(w) \cap \Sigma^{\leq k}$. Two words $w, w' \in \Sigma^*$ are called k *Simon congruent*, denoted $w \sim_k w'$, for some $k \in \mathbb{N}_0$ if $\mathrm{SubWords}_{\leq k}(w) = \mathrm{SubWords}_{\leq k}(w')$.

Remark 2.4.1. $\sim_k$ is by definition an equivalence relation and can be shown to be congruence relation. Therefore, $\Sigma^*/\sim_k$ is a monoid.

The following factorization, originally introduced by Hébrard [15] under the name *arch factorization*, is an important tool for the study of subwords. We mostly follow the vocabulary and notation by Barker et al. [2]. This factorization was also used by Karandikar, Kufleitner, and Schnoebelen [20] under the name *rich factorization*. Furthermore, similar techniques where used in the literature [29, Lemma 3] [22, Proposition 6.2.15]. Figure 2.1 shows the arch factorization of a word.

Definition 2.5 (Arch Factorization [15, 2])**.** Let $w \in \Sigma^*$ and

$$\mathrm{ar}_1(w)\,\mathrm{ar}_2(w) \ldots \mathrm{ar}_n(w)\,\mathrm{re}(w) = w$$

[1] Note that $(\mathbb{N}_0^{|\Sigma|}, +, \mathbf{0})$ is the free abelian monoid on $|\Sigma|$ generators and simultaneously the abelianization of Σ^*.

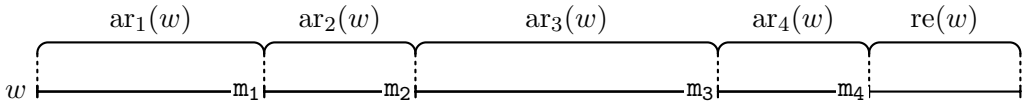


Figure 2.1: Arch Factorization of a word w where $\mathtt{m_i} \coloneqq \mathrm{m}_i(w)$

be the unique factorization satisfying $\mathrm{alph}(\mathrm{ar}_i(w)) = \Sigma$ and $|\,\mathrm{ar}_i(w)|_{\mathrm{m}_i(w)} = 1$ where $\mathrm{m}_i(w) \coloneqq \mathrm{ar}_i(w)[|\,\mathrm{ar}_i(w)|]$ for all $i \in [n]$, and $\mathrm{alph}(\mathrm{re}(w)) \subset \Sigma$. Denote the number of arches by $\iota(w)$ and call a word w with $\iota(w) = m$ *m-universal*. The number $\iota(w)$ is called the *universality index of* w. Furthermore, denote by $\mathrm{m}(w)$ the concatenation of all modi, $\mathrm{m}_1(w)\,\mathrm{m}_2(w)\cdots\mathrm{m}_{\iota(w)}(w)$. The letter $\mathrm{m}_i(w)$ is the *modus of the* i^{th} *arch* and $\mathrm{m}(w)$ is the *modus of* w. If $\mathrm{re}(w) = \varepsilon$ for some $w \in \Sigma^*$, then w is called $\iota(w)$*-perfect universal*.

Remark 2.5.1. The arch factorization of a word $w \in \Sigma^*$ has two important properties. Firstly, since the last letter of an arch is unique, if $\mathrm{m}_1(w)v \preceq w$ for some $v \in \Sigma^*$, then we can conclude that $v \preceq \mathrm{ar}_1(w)^{-1}w$. This allows us to conclude that the suffix of subwords starting with $\mathrm{pref}_\ell(\mathrm{m}(w))$ occurs in the suffix $\left(\prod_{i=1}^{\ell}\mathrm{ar}_i(w)\right)^{-1}\cdot w = \prod_{i=\ell+1}^{\iota(w)}\mathrm{ar}_i(w)\cdot \mathrm{re}(w)$ of w. Because the other direction follows directly from $\preceq$ being compatible with concatenation, we can conclude

$$\prod_{i=1}^{\ell}\mathrm{m}_\ell(w)\cdot v \preceq w \iff v \preceq \left(\prod_{i=1}^{\ell}\mathrm{ar}_i(w)\right)^{-1}\cdot w = \prod_{i=\ell+1}^{\iota(w)}\mathrm{ar}_i(w)\cdot\mathrm{re}(w).$$

We will use this fact often without explicitly referencing it.

Secondly, since each arch contains each letter at least once, each arch has all subwords of length 1, that is, each arch is 1-universal. Therefore, each word w has all subwords of length $\iota(w)$ by choosing one letter from each arch. Furthermore, $\iota(w)$ is the maximum number with this property because if $\mathrm{m}(w)\cdot\mathtt{x} \preceq w$ were to hold for all $\mathtt{x} \in \Sigma$, then the first property would imply $\mathtt{x} \preceq \mathrm{re}(w)$ for all $\mathtt{x} \in \Sigma$, which is a contradiction against the definition of $\mathrm{re}(w)$.

2.3 Shuffle

Definition 2.6. Let $u, v \in \Sigma^*$. The shuffle of u, v is the set $u \sqcup\!\sqcup v \subseteq \Sigma^{|u|+|v|}$ given by any of the following equivalent definitions [22].

(1) $w \in u \sqcup\!\sqcup v$, if and only if, there exists an $\ell \in \mathbb{N}$ such that $u = \prod_{i=1}^{\ell} u_i$, $v = \prod_{i=1}^{\ell} v_i$ for $u_i, v_i \in \Sigma^*$ for $i \in [\ell]$ and $w = \prod_{i=1}^{\ell} u_i v_i$.

(2) By induction on $|u| + |v|$, we define

$$u \sqcup\!\sqcup v \coloneqq \begin{cases} \{u\}, & v = \varepsilon \\ \{v\}, & u = \varepsilon \\ u[1](u[2\ldots|u|] \sqcup\!\sqcup v) \cup v[1](u \sqcup\!\sqcup v[2\ldots|v|]), & \text{otherwise.} \end{cases}$$

(3) $w \in u \sqcup\!\sqcup v$, if and only if, there exist occurrences O_1, O_2 of u, v in w respectively, such that $O_1 \sqcup O_2 = [|w|]$.

Furthermore, for $S, T \subseteq \Sigma^*$ the shuffle of languages is given by

$$S \sqcup\!\sqcup T \coloneqq \bigcup_{\substack{u \in S \\ v \in T}} u \sqcup\!\sqcup v.$$

Remark 2.6.1. The shuffle of languages is an associative and commutative operation with $\{\varepsilon\}$ as neutral and $\emptyset$ as absorbing element [22, Proposition 6.3.12]. Furthermore, the shuffle distributes over the union of languages and thus $(2^{\Sigma^*}, \cup, \sqcup\!\sqcup, \emptyset, \{\varepsilon\})$ forms a commutative semiring with 1. Let $\Omega \subseteq \Sigma$, then $\pi_\Omega(u \sqcup\!\sqcup v) = \pi_\Omega(u) \sqcup\!\sqcup \pi_\Omega(v)$ holds. Note that, in general this is not the case for prolonging morphisms, that is, for morphisms mapping single letters to words v with $|v| \geq 2$.

Lastly, note that the inductive definition is left-right symmetric, that is, we could have defined the set analogously by removing letters from the ends of the shuffled words. Because these two are equivalent we will sometimes apply the inductive definition from the right.

Definition 2.7. Let $w \in \Sigma^*$. If there exists $v \in \Sigma^*$ such that $w \in v \sqcup\!\sqcup v$ then w is called a *shuffle square* and v is called a *shuffle root* of w. If there exists $v \in \Sigma^*$ such that $w \in v \sqcup\!\sqcup v^R$ then w is called a *reverse shuffle square* and v is called a *reverse shuffle root* of w.

Remark 2.7.1. The shuffle root of a word is not necessarily unique. For example, $\texttt{aabaabaa} \in (\texttt{aaba} \sqcup\!\sqcup \texttt{aaba}) \cap (\texttt{abaa} \sqcup\!\sqcup \texttt{abaa})$ because we can find occurrences $\underline{\texttt{aab}}\overline{\texttt{aa}}\underline{\overline{\texttt{b}}}\underline{\texttt{a}}\overline{\texttt{a}}$ and $\underline{\texttt{a}}\overline{\texttt{a}}\underline{\texttt{baa}}\overline{\texttt{b}}\overline{\texttt{aa}}$ respectively.

The same holds for reverse shuffle squares. In fact, for every $w \in v \sqcup\!\sqcup v^R$, v^R is also a reverse shuffle root since $w \in v \sqcup\!\sqcup v^R = v^R \sqcup\!\sqcup v = v^R \sqcup\!\sqcup (v^R)^R$ because the shuffle is commutative and $\cdot^R$ is an involution.

2.4 Classes of Simon's Congruence and Shuffles

In this section, we cover a well-known connection between shuffles and equivalence classes of Simon's congruence and introduce a result with respect to the structure of the congruence class. Furthermore, as an introductory example, we apply it to find a number of singleton class of words with length greater than k.

First, we connect shuffles and Simon's congruence. The following is well-known and for example presented similarly in *Combinatorics on Words* [22, Proposition 6.2.5]. Define for $v \in \Sigma^k$ the set $L_v \coloneqq v \sqcup\!\sqcup \Sigma^*$ which is known as a *shuffle ideal*. Because the equivalence classes of Simon's congruence partition words based on their subwords, we can equivalently describe the shuffle ideal of $v \in \Sigma^{\leq k}$ as

$$L_v = \bigcup_{v \preceq w} [w]_{\sim_k}.$$

Note that the union is finite because $\sim_k$ has finite index. Since intersections and complements of unions of congruence classes are again unions of congruence classes, we can express boolean combinations of shuffle ideals as unions of elements of $\Sigma^*/\sim_k$. Furthermore, we can express congruence class of $w \in \Sigma^*$ as

$$[w]_{\sim_k} = \bigcap_{\substack{v \preceq w \\ v \in \Sigma^{\leq k}}} L_v \cap \bigcap_{\substack{v \not\preceq w \\ v \in \Sigma^{\leq k}}} L_v^{\complement}.$$

Using this connection, it is possible to decide whether $S \subseteq \Sigma^{\leq k}$ defines a congruence class by calculating the number of words of length n in $L_S \coloneqq \bigcap_{w \in S} L_w$ and $L_{S'}$ for all $S' \supset S$ for a sufficiently large n. Thus, one could also study the classes and index of Simon's congruence by studying shuffle ideals but we will focus in Chapter 4 mostly on the classes.

We now introduce a well-known structure result regarding the classes of Simon's congruence. The following theorem was proven by Simon [28, 29], and appeared with an improved proof in the monograph *Combinatorics on Words* [22].

Theorem 2.8 ([22, Theorem 6.2.6])**.** *Let $k \in \mathbb{N}$ and let $u, v \in \Sigma^*$ such that $u \sim_k v$. Then there exists some $w \in \Sigma^*$ such that $u, v \preceq w$ and $u \sim_k w \sim_k v$.*

This theorem can even be formulated as a characterization by only allowing certain common upper bounds [29, Lemma 6]. Furthermore, it has an important corollary informing the structure of the classes.

Corollary 2.8.1 ([22, Corollary 6.2.8])**.** *For each $k \in \mathbb{N}$ the equivalence classes of $\sim_k$ are either infinite or singletons.*

Obviously, all elements of $\Sigma^{<k}$ have their own singleton equivalence class. Furthermore, each element of Σ^k has its own class (although not necessarily finite class), because for $u, v \in \Sigma^k$, $[u] = [v]$ implies $u \preceq v$ and $v \preceq u$ and thus $u = v$. We can now apply Corollary 2.8.1 to give a large number of singleton classes.

Lemma 2.9. *Let $w \in \Sigma^*$. If $\sigma^k \notin \mathrm{SubWords}_{\leq k}(w)$ for all $\sigma \in \Sigma$, then $[w]_{\sim_k} = \{w\}$.*

Proof. It suffices to prove that the class is finite, then it is a singleton by Corollary 2.8.1. By contraposition, assume that $[w]_{\sim_k}$ is infinite. Then it contains some word v with $|v| \geq k|\Sigma|$. By the pigeonhole-principle, v contains one letter at least k times. □

We can use this fact to conclude that each element of

$$F \coloneqq \bigsqcup_{\substack{\ell_{\sigma_i} < k \\ 1 \leq i \leq |\Sigma|}} \sigma_1^{\ell_{\sigma_1}} \sqcup \sigma_2^{\ell_{\sigma_2}} \sqcup \cdots \sqcup \sigma_{|\Sigma|}^{\ell_{\sigma_{|\Sigma|}}} = \{\sigma_1^\ell \mid \ell < k\} \sqcup \cdots \sqcup \{\sigma_{|\Sigma|}^\ell \mid \ell < k\}$$

has its own singleton class. Furthermore, for the elements of $\Sigma^{\leq k}$, we can conclude that only classes $[\sigma^k] = \sigma^{\geq k}$ for $\sigma \in \Sigma$ are infinite classes. The set F contains more representatives of singleton classes than $\Sigma^{\leq k} \setminus \{\sigma^k \mid \sigma \in \Sigma\}$. The converse of Lemma 2.9 does not hold, that is, F does not classify all singleton classes.

Example 2.10. Consider the word `bbabb` with respect to $\sim_4$. Its subwords of length four are `bbab`, `babb` and `bbbb`. Therefore, each word in its class contains exactly one `a` (since `aa` is not a subword but `a` is), at least two `b` succeeding and preceding the `a` (`bba` and `abb` are subwords) but not more than two `b` (`bbba` and `abbb` are not subwords). Therefore, `bbabb` is the only word in this class, but it contains $\mathtt{b}^4$.

Chapter 3

Shuffles and Shuffle-Squares

In this chapter, we consider shuffles of words and sets. We focus mostly on (reverse) shuffle squares and specifically how sets of (reverse) shuffle squares relate to (reverse) squares of the set itself.

First, we show preliminary results about shuffle squares and a useful lemma involving shuffles for Chapter 4 in Section 3.1. In Section 3.2, we give a number of results relating squares of sets to sets of shuffle squares. Afterwards, we show some language theoretic motivated results about sets of shuffle squares in Sections 3.3 and 3.4 where we study the mapping $L \mapsto \bigcup_{v \in L} v \sqcup\!\sqcup v$ and specifically the set $\bigcup_{v \in \Sigma^*} v \sqcup\!\sqcup v$.

3.1 Preliminary Results

First, we show a result regarding the two occurrences of a root in a shuffle square. The lemma is folklore. It states that for a shuffle square, there always exist two occurrence such that the i^{th} letter of the first occurs before the i^{th} letter of the second. The lemma is used explicitly as well as implicitly in multiple publications regarding shuffle squares. For completeness, we also prove it here.

The idea for the proof is the following. The two occurrences of a root v in a word w can be seen as markings on the letters of w. As soon as one occurrence overtakes the other, we know that the suffix starting at this index is a shuffle square whose root is a suffix of v. We can therefore swap the markings in this suffix and continue by induction.

Lemma 3.1. *Let $v \in \Sigma^n$ and $w \in v \sqcup\!\sqcup v$, then there exists occurrences $O_1 = \{i_k\}_{k \in [n]}, O_2 = \{j_k\}_{k \in [n]}$ of v in w such that $O_1 \sqcup O_2 = [2n]$ and $i_k \leq j_k$ for all $k \in [n]$.*

Proof. Let $w \in v \sqcup\!\sqcup v$, there exists a partition $[|w|] = O_1 \sqcup O_2$ with $O_1 = \{i_k\}_{k \in [n]}, O_2 = \{j_k\}_{k \in [n]}$. Without loss of generality, assume $i_1 < j_1$. If the claim already holds, we are done. Thus, assume the contrary and let $\ell \in [2..n]$ be minimal such that $j_\ell < i_\ell$. By minimality, we have $i_\ell > j_\ell > j_{\ell-1} > i_{\ell-1}$ and thus

$$[2\ell - 2] = \{i_k \mid k < \ell\} \sqcup \{j_k \mid k < \ell\} \quad \text{and} \quad [2\ell - 1..2n] = \{i_k \mid k \geq \ell\} \sqcup \{j_k \mid k \geq \ell\}.$$

Define $\tilde{O}_1 = \{\tilde{i}_k\}_{k \in [n]} \coloneqq \{i_k \mid k < \ell\} \sqcup \{j_k \mid k \geq \ell\}$ and $\tilde{O}_2 = \{\tilde{j}_k\}_{k \in [n]} \coloneqq \{j_k \mid k < \ell\} \sqcup \{i_k \mid k \geq \ell\}$. By construction, $w[\tilde{O}_1] = v = w[\tilde{O}_2]$. Furthermore, there exists no $\tilde{\ell} \in [2..\ell]$ with $j_{\tilde{\ell}} < i_{\tilde{\ell}}$ and thus the claim follows by induction. □

Remark 3.1.1. Note that the above proof yields an algorithm that, given two disjoint occurences of the shuffle root, constructs two occurences with the desired property in linear time.

We finish this section with a short result in combinatorial enumeration involving shuffles which we will use later in Chapter 4. We are interested in the number of words of a fixed length in the shuffle ideal $L_w = w \sqcup\!\sqcup \Sigma^*$, that is, the number of words of some fixed length that contain w as a subword. To find a formula we use a known correspondence between regular languages and rational generating functions.

Lemma 3.2. *Let $w \in \Sigma^*$, then*

$$|\Sigma^n \cap L_w| = \sum_{i=|w|}^{n} \binom{i-1}{|w|-1} \cdot (|\Sigma|-1)^{i-|w|} \cdot |\Sigma|^{n-i}.$$

Proof. Let $w \in \Sigma^*$ and define $a_n \coloneqq |\Sigma^n \cap L_w|$. An unambiguous regular expression for L_w is given by

$$\left(\prod_{i=1}^{|w|} (\Sigma \setminus w[i])^* \cdot w[i]\right) \cdot \Sigma^*.$$

Denote the ordinary generating function of $(a_n)_{n\in\mathbb{N}}$ by $f_w \coloneqq \sum_{i=0}^{\infty} a_n X^n$. This generating function can be derived from an unambiguous regular expression [7]. Then, algebraic manipulation yields

$$\begin{aligned}
f_w &= \left(\prod_{i=1}^{|w|} \frac{1}{1-(|\Sigma|-1)\cdot X} \cdot X\right) \cdot \frac{1}{1-|\Sigma|\cdot X} && \text{([7, Proposition I.2, A.7])} \\
&= \left(\prod_{i=1}^{|w|} \sum_{n=1}^{\infty} (|\Sigma|-1)^{n-1} \cdot X^n\right) \cdot \left(\sum_{n=0}^{\infty} |\Sigma|^n \cdot X^n\right) \\
&= \left(\sum_{n=|w|}^{\infty} \binom{n-1}{|w|-1} \cdot (|\Sigma|-1)^{n-|w|} \cdot X^n\right) \cdot \left(\sum_{n=0}^{\infty} |\Sigma|^n \cdot X^n\right) && (|w|\text{-compositions of } n) \\
&= \sum_{n=|w|}^{\infty} \sum_{i=|w|}^{n} \binom{i-1}{|w|-1} \cdot (|\Sigma|-1)^{i-|w|} \cdot |\Sigma|^{n-i} \cdot X^n.
\end{aligned}$$

Therefore, we have

$$|\Sigma^n \cap L_w| = [X^n] f_w = \sum_{i=|w|}^{n} \binom{i-1}{|w|-1} \cdot (|\Sigma|-1)^{i-|w|} \cdot |\Sigma|^{n-i}. \qquad \square$$

The above proof was used to derive the formula. With it in hand, we can also give a second, purely combinatorial proof.

Proof. There exists no word of length shorter than $|w|$ in L_w because $v \preceq u$ implies $|v| \leq |u|$ for all $u \in \Sigma^*$. Consider a word $v \in L_w$ with $|w| \leq |v|$. There exists a left most occurrence of w in v, that is, a minimal index i such that the occurrence of w in v ends at i. There are exactly $\binom{i-1}{|w|-1}$ ways to choose the $|w|$ indices of the occurrence of w in $v[1 \dots i]$ because they correspond exactly to $|w|$-compositions of i. For the remaining $i - |w|$ letters in $v[1 \dots i]$ choose an arbitrary letter that is not the one of w following it (because otherwise there would exist another occurrence of w further left than the one using the chosen positions). For the remaining $|v| - i$ indices choose arbitrary letters from Σ. □

3.2 Shuffles, Shuffle Squares, and Reverse Shuffle Squares of Sets

Given a set $L \subseteq \Sigma^*$ we can consider its set of shuffle squares, $\bigcup_{v \in L} v \sqcup\!\sqcup v$. This set is by definition a subset of $L \sqcup\!\sqcup L$. As a motivation for investigating sets of shuffle squares, we show that the two only coincide if $|L| \leq 1$, that is, if the unionend sets already coincide by definition.[1]

Proposition 3.3. *Let $L \subseteq \Sigma^*$, then we have*

$$|L| \leq 1 \iff L \sqcup\!\sqcup L = \bigcup_{v \in L} v \sqcup\!\sqcup v.$$

Proof. Let $L \subseteq \Sigma^*$ and assume $|L| \leq 1$. Then the claim follows by definition. Therefore, assume that

$$L \sqcup\!\sqcup L = \bigcup_{v \in L} v \sqcup\!\sqcup v. \tag{3.1}$$

We show the proposition in a series of steps.

Claim 1. *For all $u, v \in L$, $\mathtt{x} \in \Sigma$, we have $|u|_{\mathtt{x}} = |v|_{\mathtt{x}}$, that is, all elements of L are permutations of each other. Thus, the set L is finite.*

Suppose there exists $\mathtt{x} \in \Sigma$ such that not all elements of L have the same number of $\mathtt{x}$. Choose $u_1, u_2 \in L$ such that $|u_1|_{\mathtt{x}} < |u_2|_{\mathtt{x}}$ and $|u_2|_{\mathtt{x}} - |u_1|_{\mathtt{x}}$ is minimal. Fix some $w \in u_1 \sqcup\!\sqcup u_2 \subseteq L \sqcup\!\sqcup L$. By Equation 3.1, there exists $v \in L$, such that $w \in v \sqcup\!\sqcup v$. This implies

$$|v|_{\mathtt{x}} = \frac{|w|_{\mathtt{x}}}{2} = \frac{|u_1|_{\mathtt{x}} + |u_2|_{\mathtt{x}}}{2}.$$

Therefore, we have

$$|u_1|_{\mathtt{x}} = \frac{|u_1|_{\mathtt{x}} + |u_1|_{\mathtt{x}}}{2} < |v|_{\mathtt{x}} < \frac{|u_2|_{\mathtt{x}} + |u_2|_{\mathtt{x}}}{2} = |u_2|_{\mathtt{x}},$$

a contradiction to the minimality of $|u_2|_{\mathtt{x}} - |u_1|_{\mathtt{x}}$.

[1] We would like to thank Volker Diekert for pointing out the missing inductive case in the proof of Claim 3 in the proof of Proposition 3.3.

Claim 2. *It suffices to show the proposition for $L \subseteq \Sigma_2^*$.*

Assume the proposition already holds for $|\Sigma| \leq 2$. Suppose $|L| \geq 2$, then there exist $u, v \in L$ with $u \neq v$. Thus, there exists $\Omega \in \binom{\Sigma}{2}$ such that $\pi_\Omega(u) \neq \pi_\Omega(v)$ and in particular $|\pi_\Omega(L)| \geq 2$. Because Equation 3.1 implies

$$\pi_\Omega(L) \sqcup\!\sqcup \pi_\Omega(L) = \pi_\Omega(L \sqcup\!\sqcup L) = \pi_\Omega\Big(\bigcup_{v \in L} v \sqcup\!\sqcup v\Big) = \bigcup_{v \in L} \pi_\Omega(v) \sqcup\!\sqcup \pi_\Omega(v),$$

the assumption holds for $\pi_\Omega(L)$. A contradiction to the binary case of the proposition because $|\Omega| = 2$ and $|\pi_\Omega(L)| \geq 2$.

Claim 3. *The proposition holds for all $L \subseteq \Sigma_2^*$.*

We prove the claim by induction over the (by Claim 1) common length of elements of L. We have $|L| \leq 1$ for $L \subseteq \{\varepsilon\} = \Sigma_2^0$. If any, and thus by Claim 1 all, elements of L are unary, then Claim 1 implies that $|L| = 1$. Thus, assume the elements of L are non-empty, binary words and suppose $|L| > 1$. We distinguish two cases.

Case 1 ($L \subseteq \mathtt{x}^m\overline{\mathtt{x}}\Sigma^*$ for some $m \in \mathbb{N}$ and $\mathtt{x} \in \Sigma$). Define $\tilde{L} \coloneqq \mathtt{x}^{-m}L$ and note that $|L| = |\tilde{L}|$. Let $w \in \tilde{L} \sqcup\!\sqcup \tilde{L}$. Because $w \in \mathtt{x}^{-m}u_1 \sqcup\!\sqcup \mathtt{x}^{-m}u_2$ for $u_1, u_2 \in L$, we have $\mathtt{x}^{2m}w \in L \sqcup\!\sqcup L$. Therefore, there exists some $v \in L$ such that $\mathtt{x}^{2m}w \in v \sqcup\!\sqcup v$. By the assumption of Case 1, we have $\mathtt{x}^m\overline{\mathtt{x}} \in \mathrm{Pref}(v)$ and thus $w \in \mathtt{x}^{-m}v \sqcup\!\sqcup \mathtt{x}^{-m}v$ and thus

$$\tilde{L} \sqcup\!\sqcup \tilde{L} = \bigcup_{u \in \tilde{L}} u \sqcup\!\sqcup u.$$

Therefore, $\tilde{L}$ satisfies the assumption and the contradiction follows by induction.

Case 2 (otherwise). There exists some $\mathtt{x} \in \Sigma_2$ such that there exist a pair of words $u_1, u_2 \in L$ with $\mathrm{Pref}(u_1) \cap \mathtt{x}^* \neq \mathrm{Pref}(u_2) \cap \mathtt{x}^*$. Among those, we choose a pair of words, such that their longest common prefix is of maximal length. To be more precise, choose $m \in \mathbb{N}_0$ maximal, such that for $\ell, r, s \in \mathbb{N}$ there exist

$$u_1 \coloneqq \mathtt{x}^m\mathtt{x}^\ell\overline{\mathtt{x}}^r\tilde{u}_1 \in L \qquad \text{and} \qquad u_2 \coloneqq \mathtt{x}^m\overline{\mathtt{x}}^s\tilde{u}_2 \in L.$$

Among all candidates for u_1, we choose a word such that r is also maximal. Note that these maximal choices are well-defined because L is finite by Claim 1.

Consider $w \coloneqq \mathtt{x}^{2m+\ell}\overline{\mathtt{x}}^{r+s}\tilde{u}_1\tilde{u}_2 \in u_1 \sqcup\!\sqcup u_2$. By Equation 3.1, there exists some $v \in L$ with $w \in v \sqcup\!\sqcup v$. Therefore, there exist $n, t \in \mathbb{N}$ and $\tilde{v} \in \Sigma^*$ such that $v = \mathtt{x}^n\overline{\mathtt{x}}^t\tilde{v}$ and $\lceil \frac{2m+\ell}{2} \rceil \leq n \leq 2m + \ell$. Note that $m < m + \lceil \frac{\ell}{2} \rceil \leq n$. Therefore, if $n \neq m + \ell$, we have that v, u_1 have unary prefixes of different lengths and their longest common prefix is of length $m + \lceil \frac{\ell}{2} \rceil > m$, a contradiction to the maximality of m. If $n = m + \ell$ we have

$$w = \mathtt{x}^{2m+\ell}\overline{\mathtt{x}}^{r+s}\tilde{u}_1\tilde{u}_2 \in \mathtt{x}^{m+\ell}\overline{\mathtt{x}}^t\tilde{v} \sqcup\!\sqcup \mathtt{x}^{m+\ell}\overline{\mathtt{x}}^t\tilde{v}.$$

This implies $t \geq r + s > r$, a contradiction to the maximality of r. □

We can prove a similar result for reverse shuffle squares. We use the following characterization of binary reverse shuffle squares as abelian squares by Henshall, Rampersad, and Shallit [16].

Theorem 3.4 ([16, Theorem 7])**.**

(1) If there exists $v \in \Sigma^*$ such that $w \in v \sqcup\!\sqcup v^R$, then w is an abelian square.

(2) If $w \in \Sigma_2^*$ is an abelian square, then there exists $v \in \Sigma_2^*$ such that $w \in v \sqcup\!\sqcup v^R$.

Using some ideas from the proof of Proposition 3.3 and Theorem 3.4, we can give a proof of the following analogous proposition. In fact, we only need the the proposition (1) from Theorem 3.4.

Proposition 3.5. *Let $L \subseteq \Sigma^*$, then we have*

$$|L| \leq 1 \iff L \sqcup\!\sqcup L^R = \bigcup_{v \in L} v \sqcup\!\sqcup v^R.$$

Proof. Let $L \subseteq \Sigma^*$ and assume $|L| \leq 1$. Then the claim follows by definition. Therefore, assume that

$$L \sqcup\!\sqcup L^R = \bigcup_{v \in L} v \sqcup\!\sqcup v^R \tag{3.2}$$

We show the proposition in a series of steps.

Claim 1. *For all $u, v \in L$, $\mathtt{x} \in \Sigma$, we have $|u|_{\mathtt{x}} = |v|_{\mathtt{x}}$, that is, all elements of L are permutations of each other. Thus, the set L is finite.*

Suppose there exists $\mathtt{x} \in \Sigma$ such that not all elements of L have the same number of $\mathtt{x}$. Choose $u_1, u_2 \in L$ such that $|u_1|_{\mathtt{x}} < |u_2|_{\mathtt{x}}$ and $|u_2|_{\mathtt{x}} - |u_1|_{\mathtt{x}}$ is minimal. Fix some $w \in u_1 \sqcup\!\sqcup u_2^R \subseteq L \sqcup\!\sqcup L^R$. By Equation 3.2, there exist $v \in L$, such that $w \in v \sqcup\!\sqcup v$. This implies

$$|v|_{\mathtt{x}} = \frac{|w|_{\mathtt{x}}}{2} = \frac{|u_1|_{\mathtt{x}} + |u_2|_{\mathtt{x}}}{2}.$$

Therefore, we have

$$|u_1|_{\mathtt{x}} = \frac{|u_1|_{\mathtt{x}} + |u_1|_{\mathtt{x}}}{2} < |v|_{\mathtt{x}} < \frac{|u_2|_{\mathtt{x}} + |u_2|_{\mathtt{x}}}{2} = |u_2|_{\mathtt{x}},$$

a contradiction to the minimality of $|u_2|_{\mathtt{x}} - |u_1|_{\mathtt{x}}$.

Claim 2. *The proposition holds for all $L \subseteq \Sigma^*$.*

Let $u, v \in L$. Assume without loss of generality that $n \coloneqq |u| = |v|$ by Claim 1. By Equation 3.2, each element of $u \sqcup\!\sqcup v^R$ is a reverse shuffle square and therefore by Theorem 3.4 an abelian square. Therefore, we have for all $w \in u \sqcup\!\sqcup v^R$ that

$$\mathbf{p}(\mathrm{pref}_n(w)) = \mathbf{p}(\mathrm{suff}_n(w)).$$

Because $\{(\mathrm{pref}_\ell(u) \sqcup \mathrm{pref}_k(v^R))(\mathrm{suff}_k(u) \sqcup \mathrm{suff}_\ell(v^R)) \mid \ell + k = n\} \subseteq u \sqcup v^R$ and shuffles of words are never the empty set, we obtain

$$\forall \ell, k \in \mathbb{N}_0 . \ell + k = n \implies \mathbf{p}(\mathrm{pref}_\ell(u)) + \mathbf{p}(\mathrm{pref}_k(v^R)) = \mathbf{p}(\mathrm{suff}_k(u)) + \mathbf{p}(\mathrm{suff}_\ell(v^R)). \quad (3.3)$$

By Claim 3 it follows that $u = v$ and thus that $|L| \leq 1$.

Claim 3. *If Equation 3.3 holds for $u, v \in \Sigma^m$ with $n = m$, then $u = v$.*

If $m = 0$, then the claim follows trivially. Therefore, let $m > 0$ and $u, v \in \Sigma^m$. For $\ell = m$ and $k = 0$, we obtain $\mathbf{p}(u) = \mathbf{p}(v)$. For $\ell = m - 1$ and $j = 1$, we obtain $\mathbf{p}(u[1..|u| - 1]) + \mathbf{p}(v[|v|]) = \mathbf{p}(u[|u|]) + \mathbf{p}(v[1..|v| - 1])$, and thus $u[|u|] = v[|v|]$ because

$$\begin{aligned} 2\,\mathbf{p}(v[|v|]) &= 2\,\mathbf{p}(v[|v|]) + \mathbf{p}(u[1..|u| - 1]) + \mathbf{p}([|u|]) - \mathbf{p}(u) \\ &= \mathbf{p}(v[|v|]) + \mathbf{p}(u[|u|]) + \mathbf{p}(v[1..|v| - 1]) + \mathbf{p}([|u|]) - \mathbf{p}(u) \\ &= 2\,\mathbf{p}(u[|u|]) + \mathbf{p}(v) - \mathbf{p}(u) \\ &= 2\,\mathbf{p}(u[|u|]). \end{aligned}$$

Define $\tilde{u} \coloneqq u[1..|u| - 1]$ and $\tilde{v} \coloneqq v[1..|v| - 1]$. Now let $\tilde{\ell}, \tilde{k} \in \mathbb{N}_0$ such that $\tilde{\ell} + \tilde{k} = m - 1$. Then we have by assumption

$$\begin{aligned} \mathbf{p}(\mathrm{pref}_{\tilde{\ell}}(\tilde{u})) + \mathbf{p}(\mathrm{pref}_{\tilde{k}}(\tilde{v}^R)) + \mathbf{p}(v[|v|]) &= \mathbf{p}(\mathrm{pref}_{\tilde{\ell}}(u)) + \mathbf{p}(\mathrm{pref}_{\tilde{k}+1}(v^R)) \\ &= \mathbf{p}(\mathrm{suff}_{\tilde{k}+1}(u)) + \mathbf{p}(\mathrm{suff}_{\tilde{\ell}}(v^R)) \\ &= \mathbf{p}(\mathrm{suff}_{\tilde{k}}(\tilde{u})) + \mathbf{p}(\mathrm{suff}_{\tilde{\ell}}(\tilde{v}^R)) + \mathbf{p}(u[|u|]). \end{aligned}$$

Since $\mathbf{p}(u[|u|]) = \mathbf{p}(v[|v|])$, Equation 3.3 holds with $n = m - 1$ for $\tilde{u}$ and $\tilde{v}$ and the claim follows by induction. □

In the above prove of Proposition 3.5, we do not use the fact that the root of the reverse shuffle square is an element of L. Thus, we can conclude similarly that if $L \sqcup L^R$ for $L \subseteq \Sigma^n$ is a set of reverse shuffle squares, then $|L| \leq 1$. In particular, we can conclude the following proposition.

Corollary 3.5.1. *Let $u, v \in \Sigma^n$. If all elements of $u \sqcup v^R$ are reverse shuffle squares, then $u = v$.*

Proof. Let $w \in u \sqcup v$. By assumption, all elements of $u \sqcup v^R$ are reverse shuffle squares and thus by Theorem 3.4 an abelian squares. Therefore, we have for all $w \in u \sqcup v^R$ that

$$\mathbf{p}(\mathrm{pref}_n(w)) = \mathbf{p}(\mathrm{suff}_n(w)).$$

Because $\{(\mathrm{pref}_\ell(u) \sqcup \mathrm{pref}_k(v^R))(\mathrm{suff}_k(u) \sqcup \mathrm{suff}_\ell(v^R)) \mid \ell + k = n\} \subseteq u \sqcup v^R$ and shuffles of words are never the empty set, we obtain

$$\forall \ell, k \in \mathbb{N}_0 . \ell + k = n \implies \mathbf{p}(\mathrm{pref}_\ell(u)) + \mathbf{p}(\mathrm{pref}_k(v^R)) = \mathbf{p}(\mathrm{suff}_k(u)) + \mathbf{p}(\mathrm{suff}_\ell(v^R)).$$

By Claim 3 from the proof of Proposition 3.5 it now follows that $u = v$. □

A linked question is whether a similar fact holds for shuffle squares, that is, if $u \sqcup\!\sqcup v$ contains only shuffle squares, can we conclude that $u = v$? This proposition does not hold if $|u| \neq |v|$, consider for example $u = \mathtt{a}$ and $v = \mathtt{aaa}$. Since we were not able to prove the case that $|u| = |v|$, we claim the following conjecture for whose investigation we use other methods.

Conjecture 3.6. *Let $u, v \in \Sigma^n$, if every $w \in u \sqcup\!\sqcup v$ is a shuffle square, then $u = v$.*

3.3 Languages of Shuffle Squares

A question studied in the literature is the complexity of the language of shuffle squares. The problem of determining whether a word w is a shuffle square of some word v is known to be NP-complete [26, 6], even in the case of binary words [5]. Henshall, Rampersad, and Shallit [16] show that the languages of (binary) shuffle squares and shuffle cubes are both not context-free. In fact, using their techniques it is possible to generalize the result to k-fold shuffle powers.

The structure of the $k = 2$ case is similar to their proof, but the argument is slightly shorter by considering the overlap. We present it for completeness. The $k \geq 3$ case uses a generalization of their argument for $k = 3$. In both cases we heavily use the fact that a k-fold shuffle contains k pairwise disjoint occurrences of the root. Using this fact, we derive the contradictions by simply counting letters in some occurrence of the root.

Theorem 3.7. *Let $k \geq 2$, then the language $L_k \coloneqq \bigcup_{v \in \Sigma_2^*} \left(\bigsqcup\!\bigsqcup_{i=1}^{k} v \right)$ is not context-free.*

Proof. Let $k \in \mathbb{N}_{\geq 2}$ and suppose the language is context-free. We distinguish two cases.

Case 1 ($k = 2$). As done by Henshall, Rampersad, and Shallit [16], we show that

$$L_2 \cap \mathtt{a}(\mathtt{b}^+\mathtt{a}^+)\mathtt{ab}(\mathtt{b}^+\mathtt{a}^+)\mathtt{b} = \{\mathtt{a}(\mathtt{b}^i\mathtt{a}^j)\mathtt{ab}(\mathtt{b}^i\mathtt{a}^j)\mathtt{b} \mid i, j \in \mathbb{N}_{\geq 2}\}$$

is not context-free. Let $w = \mathtt{a}(\mathtt{b}^{i_1}\mathtt{a}^{j_1})(\mathtt{ab})(\mathtt{b}^{i_2}\mathtt{a}^{j_2})\mathtt{b}$ an element of the above intersection and v its shuffle root. By the inductive definition of the shuffle, we have $\mathtt{ab}^{i_1}\mathtt{a}^{j_1} \in \mathrm{Pref}(v)$ and $\mathtt{b}^{i_2}\mathtt{a}^{j_2}\mathtt{b} \in \mathrm{Suff}(v)$. If this prefix and suffix overlap, then we have $i_1 \geq i_2$ and $j_1 \leq j_2$. Furthermore, they have to be equal because we have $j_1 \geq j_2$ by

$$j_1 + j_2 + 2 = |w|_{\mathtt{a}} = 2 \cdot |v|_{\mathtt{a}} \geq 2 \cdot (|\mathtt{a}|_{\mathtt{a}} + |\mathtt{b}^{i_2}\mathtt{a}^{j_2}\mathtt{b}|_{\mathtt{a}}) = 2 + 2j_2,$$

and $i_1 \leq i_2$ by a symmetric argument. If the prefix and suffix do not overlap, we have $2 + j_1 + j_2 = |w|_{\mathtt{a}} = 2|v|_{\mathtt{a}} \geq 2(1 + j_1 + j_2) = 2 + 2j_1 + 2j_2$, a contradiction.

Case 2 ($k \geq 3$). By Ogden's lemma, there exists a constant p. Now consider

$$w \coloneqq (\mathtt{a}^p\mathtt{b})(\dot{\mathtt{a}}^p\mathtt{b})(\mathtt{a}^p\mathtt{b})^{k-2} \in L_k \cap (\mathtt{a}^*\mathtt{b})^k$$

where letters are marked with a dot. By Ogden's lemma there exists a decomposition $w = uvxyz$ such that $|vy|$ contains at least one and $|vxy|$ at most p markings. Because

all words of $L_k \cap (\mathtt{a}^*\mathtt{b})^k$ contain $\mathtt{b}$ exactly k times, we have a contradiction by pumping if $|vy|_\mathtt{b} \neq 0$.

The following cases are now similar to the ones by Henshall, Rampersad, and Shallit [16] for $k = 3$. For $\ell \in \mathbb{N}_0$ denote the k-fold shuffle root of $uv^\ell xy^\ell z$ by r_ℓ. Both v and y are contained in blocks of $\mathtt{a}$. Thus, $\mathtt{b}$ is a suffix of r_ℓ and therefore $r_\ell \in \mathtt{a}^*\mathtt{b}$ since $|w|_\mathtt{b} = k$.

Case 2.a (v contains $\dot{\mathtt{a}}$). We pump up once to obtain $\tilde{w} = uv^2xy^2z$. Since the first block is untouched, we obtain $|r_2| \leq p + 1$ since the first $\mathtt{b}$ has to belong to an occurrence of r_2 which can contain at most p other indices. A contradiction because $k(p + 1) \geq k \cdot |r_2| = |\tilde{w}| = |vy| + |w| = |vy| + k(p + 1)$.

Case 2.b (v does not contain $\dot{\mathtt{a}}$). Therefore, v is contained in the first and y in the second block because $|vy|$ contains a marking. We pump down and obtain $\tilde{w} = (\mathtt{a}^{p-q}\mathtt{b})(\mathtt{a}^{p-s}\mathtt{b})(\mathtt{a}^p\mathtt{b})^{k-2}$ for $q \in [p]_0$ and $s \in [p - 1]$. Because, we have at least three blocks, $\mathtt{a}^p\mathtt{b}$ is a suffix of $\tilde{w}$, and therefore a suffix of r_0. A contradiction because $k(p + 1) - |vy| = |w| - |vy| = |\tilde{w}| = k \cdot |r_0| \geq k(p + 1)$. □

One can give a nondeterministic linear space algorithm checking whether a word is a shuffle square, that is, show that the language of all shuffle squares is context-sensitive. This is done by guessing the positions of the first occurrence of the root and checking whether the remaining positions are an occurrence of the same word.

In between the two classes properly lies the class of indexed languages which was originally introduced by Aho [1]. We start investigating whether the language of shuffle squares is indexed. Index languages are generated by indexed grammars. They are similar to context-free grammars, but every non-terminal is equipped with a stack. In productions rules, the stack can be copied to any number of other non-terminals. Furthermore, production rules can depend on the top of the stack, and push to or pop from it when copying it. We do not give a precise definition of indexed languages because we claim the following conjecture which we approach indirectly.

Conjecture 3.8. *The language of (binary) shuffle squares is not indexed.*

It is possible to generate some elements of the language of shuffle squares using an indexed grammar, for example shuffle squares of the form $xyxzyz$. This conjecture is motivated by the observation that the shuffle language contains heavily nested words like $\overline{a}\overline{x}\underline{u}\underline{a}\overline{w}\underline{x}\overline{v}\underline{u}\overline{z}\underline{w}\overline{c}\underline{v}\underline{z}\underline{c} \in axuwvzc \sqcup\!\sqcup axuwvzc$ which do not have an obvious tree structure. Furthermore, the language of shuffle squares is conceptually similar to the well-known language

$$\mathrm{MIX} \coloneqq \{w \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}^* \mid |w|_\mathtt{a} = |w|_\mathtt{b} = |w|_\mathtt{c}\} = \bigcup_{n \in \mathbb{N}_0} \mathtt{a}^n \sqcup\!\sqcup \mathtt{b}^n \sqcup\!\sqcup \mathtt{c}^n.$$

This language has received attention in computational linguistics and the rational equivalent $\{w \in \{\mathtt{a}, \overline{\mathtt{a}}, \mathtt{b}, \overline{\mathtt{b}}\} \mid |w|_\mathtt{a} = |w|_{\overline{\mathtt{a}}}, |w|_\mathtt{b} = |w|_{\overline{\mathtt{b}}}\}$ in computational group theory. Whether these two languages are indexed is an open problem and the negative was first conjectured by Marsh [23]. A detailed list of references, a proof of equivalence, and a proof that MIX

is not an element of a proper subclass of the indexed languages is given by Kanazawa and Salvati [19] and Salvati [27].

There exist a number of necessary (although not sufficient) conditions for a language to be indexed, similar to the pumping lemmata for regular and context-free languages. We are aware of three necessary conditions by Hayashi [14], Gilman [13], and Smith [31]. Furthermore, there exists a Chomsky-Schützenberger type representation theorem by Fratani and Voundy [11, Theorem 18]. We investigate two of the necessary conditions. Unexpectedly both yield negative results but one of them yields a non-trivial combinatorial property of the language of shuffle squares. Furthermore, both of our proofs should be transferrable to MIX.

3.3.1 A Pumping Property

In this short section, we consider the first necessary condition, the pumping lemma by Smith [31]. First we state their result.

Theorem 3.9 ([31, Theorem 1])**.** *Let L be an indexed language. Then there is an $\ell \geq 0$ (which we will call a threshold for L) such that for any $w \in L$ with at least ℓ marked positions,*

(1) w can be written as $w = u_1 u_2 \cdots u_n$ and each u_i can be written $u_i = v_{i,1} v_{i,2} \cdots v_{i,n_i}$ (we will denote the set of subscripts of v, that is, $\{(i,j) \mid 1 \leq i \leq n \text{ and } 1 \leq j \leq n_i\}$, by I);

(2) there is a map $\varphi : I \to [n]$ such that if each $v_{i,j}$ is replaced with $u_{\varphi(i,j)}$, then the resulting word is still in L, and this process can be applied iteratively to always yield a word in L;

(3) for each $(i,j) \in I$, if $v_{i,j}$ contains a marked position then so does $u_{\varphi(i,j)}$;

(4) there is an $(i,j) \in I$ such that $\varphi(i,j) = i$, and there is at least one marked position in u_i but outside of $v_{i,j}$.

The next lemma shows that the language of shuffle squares satisfies this property. This follows mostly from the fact that the language is closed under concatenation.

Lemma 3.10. *The language $\bigcup_{v \in \Sigma} v \sqcup\!\sqcup v$ satisfies the pumping property by Smith [31].*

Proof. Define the threshold $\ell := 2$ and let $w \in v \sqcup\!\sqcup v$ with at least two marked positions. Write $w = u_1$ and $u_1 = v_1 v_2 \cdots v_{|w|}$ with $v_i \in \Sigma$ and let I as above. Now define $\varphi : I \to [1], (i,j) \mapsto 1$. Applying φ^k to w for some $k \in \mathbb{N}$ yields a power of w. Because the language of shuffle squares is closed under concatenation, the resulting word is always in the language. For all $(i,j) \in I$, we have $\varphi(i,j) = i = 1$. Furthermore, because w has at least 2 marked positions, $u_i = u_1$ contains a marked position which was not in $v_{i,j}$. □

3.3.2 A Shrinking Property

In this section, we prove that the shuffle language has the shrinking property by Gilman [13]. Before, we have to prove a technical lemma. The idea of the lemma is the following. If we have a word $\tilde{w} \preceq w$ for a shuffle square $w \in v \sqcup\!\sqcup v$, then we can extend $\tilde{w}$ to a shuffle square $\hat{w} \preceq w$. This is done by taking the parts of the two occurrences of v in $\tilde{w}$ and adding the partnering letter from the second occurrence. It is important, that this process can at most double the number of letters. Note that for occurrences $O_1 \subseteq O_2 \subseteq [|w|]$ we have that $w[O_1] \preceq w[O_2]$.

Lemma 3.11. *Let $w \in v \sqcup\!\sqcup v$ and $\tilde{w} \preceq w$. Then there exists $\hat{w}$ such that $\tilde{w} \preceq \hat{w} \preceq w$ and $|\hat{w}| \leq 2|\tilde{w}|$ and $\hat{v} \preceq v$ such that $\hat{w} \in \hat{v} \sqcup\!\sqcup \hat{v}$. Furthermore, formulated for occurrences, the following holds.*

Let $w \in \Sigma^*$ with $U_1 \sqcup U_2 = [|w|]$ such that $w[U_1] = w[U_2]$ and $\tilde{O} \subseteq [|w|]$. Then there exists $\hat{O}$ such that $\tilde{O} \subseteq \hat{O} \subseteq [|w|]$ and $|\hat{O}| \leq 2|\tilde{O}|$ and there exists $\hat{U}_1, \hat{U}_2$ such that $\hat{U}_i \subseteq U_i$ for all $i \in [2]$, $\hat{U}_1 \sqcup \hat{U}_2 = \hat{O}$ and $w[\hat{U}_1] = w[\hat{U}_2]$.

Proof. Let $\varphi : U_1 \to U_2$ be the unique monotonic bijection between the two sets, that is, the unique isomorphism of posets for $(U_1, \leq)$ and $(U_2, \leq)$. Since $w[U_1] = w[U_2]$, we have for $i \in U_1$ that $w[i] = w[\varphi(i)]$. This follows by induction on $|U_1| = |U_2|$ since

$$\begin{aligned} w[\min U_1] \cdot w[U_1 \setminus \{\min U_1\}] &= w[U_1] \\ &= w[U_2] \\ &= w[\min U_2] \cdot w[U_2 \setminus \{\min U_2\}] \\ &= w[\varphi(\min U_1)] \cdot w[U_2 \setminus \{\min U_2\}]. && (U_1 \cong_\varphi U_2) \end{aligned}$$

Let $\tilde{U}_i \coloneqq \tilde{O} \cap U_i \subseteq U_i$ for $i \in [2]$ and define $\hat{U}_1 \coloneqq \tilde{U}_1 \cup \varphi^{-1}(\tilde{U}_2)$ and $\hat{U}_2 \coloneqq \tilde{U}_2 \cup \varphi(\tilde{U}_1)$. We have

$$\begin{aligned} \hat{U}_1 \cap \hat{U}_2 &= (\tilde{U}_1 \cup \varphi^{-1}(\tilde{U}_2)) \cap (\tilde{U}_2 \cup \varphi(\tilde{U}_1)) \\ &\subseteq U_1 \cap U_2 && (\varphi^{-1}(\tilde{U}_2) \subseteq U_1 \text{ and } \tilde{U}_2 \cup \varphi(\tilde{U}_1) \subseteq U_2) \\ &= \emptyset. \end{aligned}$$

By construction, $\varphi(\hat{U}_1) = \hat{U}_2$ and thus $\hat{U}_1 \cong_\varphi \hat{U}_2$ as posets and in particular $|\hat{U}_1| = |\hat{U}_2|$. Define $\hat{O} \coloneqq \hat{U}_1 \sqcup \hat{U}_2$ and note that $\tilde{O} = (\tilde{O} \cap U_1) \sqcup (\tilde{O} \cap U_2) = \tilde{U}_1 \sqcup \tilde{U}_2 \subseteq \hat{O}$. Furthermore, we have

$$\begin{aligned} |\hat{O}| &= |\hat{U}_1| + |\hat{U}_2| \\ &\leq |\tilde{U}_1| + |\varphi^{-1}(\tilde{U}_2)| + |\tilde{U}_2| + |\varphi(\tilde{U}_1)| \\ &= |\tilde{U}_1| + |\tilde{U}_2| + |\tilde{U}_2| + |\tilde{U}_1| && (\varphi \text{ is bijection}) \\ &= 2|\tilde{O}|. \end{aligned}$$

It is left to show that $\hat{U}_1$ and $\hat{U}_2$ determine the same word. If $|\hat{U}_1| = |\hat{U}_2| = 0$, then $w[\hat{U}_1] = \varepsilon = w[\hat{U}_2]$. Therefore, assume $|\hat{U}_1| = |\hat{U}_2| \geq 1$, then we have

$$\begin{aligned}
w[\hat{U}_2] &= w[\min \hat{U}_2] \cdot w[\hat{U}_2 \setminus \{\min \hat{U}_2\}] \\
&= w[\varphi(\min \hat{U}_1)] \cdot w[\hat{U}_2 \setminus \{\min \hat{U}_2\}] && (\hat{U}_1 \cong_\varphi \hat{U}_2) \\
&= w[\varphi(\min \hat{U}_1)] \cdot w[\hat{U}_1 \setminus \{\min \hat{U}_1\}] && (\text{Induction}) \\
&= w[\min \hat{U}_1] \cdot w[\hat{U}_1 \setminus \{\min \hat{U}_1\}] && (w[j] = w[\varphi(j)] \text{ for all } j \in U_1) \\
&= w[\hat{U}_1]. && \square
\end{aligned}$$

With Lemma 3.11 in hand, we can show that the language of shuffle squares has a combinatorial property satisfied by all indexed languages. In the following we have to consider factorizations. For a word $w \in \Sigma^*$, we consider a factorization as an element of $(\mathrm{Fact}(w))^* \subseteq (\Sigma^*)^*$. We write factorizations in tuple notation and denote the concatenation of all factors in a factorization f by $\mu(f) \coloneqq \prod_{i=1}^{|f|} f[i]$.[2]

Lemma 3.12. *The language of shuffle squares has the shrinking property by Gilman [13]. This means that for each $m \in \mathbb{N}$, there exists $k \in \mathbb{N}$ and for each $w \in \bigcup_{v \in \Sigma^*} v \sqcup\!\sqcup v$ with $|w| \geq k$, there exists $r \in \mathbb{N}$ such that*

(1) $m < r \leq k$,

(2) w can be written as a product $\prod_{i=1}^{r} w_r$ such that $w_i \in \Sigma^+$ for all $i \in [r]$,

(3) for each $s \prec (w_1, w_2, \ldots, w_r)$ such that $|s| = m$, there exists $s \preceq \hat{s} \prec (w_1, w_2, \ldots, w_r)$ such that $\mu(\hat{s}) \in \bigcup_{v \in \Sigma^*} v \sqcup\!\sqcup v$.

Proof. Let $m \in \mathbb{N}$ and define $k \coloneqq 6m + 1$. Let $w \in \Sigma^{\geq k}$ such that $w \in v \sqcup\!\sqcup v$ for some $v \in \Sigma^*$. Factorize w as $w = x_1 x_2 \cdots x_{2m} \tilde{w}$ with $x_i \in \Sigma$ for all $i \in [m]$ and $\tilde{w} \in \Sigma^{\geq 4m+1}$. Because $w \in v \sqcup\!\sqcup v$, there exists a partition $[|w|] = O_1 \sqcup O_2$ with $O_1 = \{i_k\}_{k \in |O_1|}, O_2 = \{j_k\}_{k \in |O_2|}$ indexed ascending, such that $w[O_1] = v = w[O_2]$ and $i_k \leq j_k$ for all $k \in [|v|]$ by Lemma 3.1.

Now there exist the sets of indices $X_j \coloneqq [2m] \cap O_j$, of those x_i which are part of the j^{th} occurrence of v for $j \in [2]$. We have $|X_1| \geq |X_2|$ and $X_1 \sqcup X_2 = [2m]$ by construction. Thus, the number of the x_i that have their second occurrence in $\tilde{w}$ is given by

$$n \coloneqq |X_1| - |X_2| = |X_1| - (2m - |X_1|) = 2|X_1| - 2m.$$

Factorize $\tilde{w}$ and thus w at these second occurrences, that is, write

$$w = x_1 x_2 \cdots x_{2m} \cdot u_0 y_1 u_1 y_2 \cdots y_n u_n$$

for $u_i \in \Sigma^*$ for all $i \in [n]_0$ and $y_j \in \Sigma$ for all $j \in [n]$. Note that n is at most $2m$ and therefore the number of factors in the factorization $f \coloneqq (x_1, x_2, \ldots, x_{2m}, u_0, y_1, u_1, \ldots, u_n y_n)$

[2] Note that $\mu : (\Sigma^*)^* \to \Sigma^*$ exactly the component at Σ of the monad for monoids multiplication which is usually denoted μ_Σ. Thus we denote it similarly here.

is at most k. Let $\tilde{f} \preceq f$ such that $\mu(\tilde{f}) = f$, that is, let $\tilde{f}$ be the unique subsequence of f that omits empty factors. Since only u_i can be empty, we have that $\tilde{f}$ contains at least $2m > m$ factors. Let $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{\tilde{n}}$ denote the subsequence of nonempty u_i. It is left to show that $\tilde{f}$ has the shrinking property. By construction, we have

$$x_1 x_2 \cdots x_{2m} y_1 y_2 \cdots y_n \in v_p \sqcup v_p \qquad \text{and} \qquad \tilde{u}_0 \tilde{u}_1 \cdots \tilde{u}_{\tilde{n}} \in v_s \sqcup v_s$$

for $v_p \coloneqq w[O_1 \cap [2m]] \in \operatorname{Pref}(v)$ and $v_s \coloneqq v_p^{-1} v$. Let $s \prec \tilde{f}$ with $|s| = m$. If $s \prec (x_1, x_2, \ldots, x_{2m}, y_1, y_2, \ldots, y_n)$ then the claim follows directly from Lemma 3.11. Thus, assume $\tilde{u}_i \prec \tilde{f}$. Let $s_x \prec s$ be the subword of s containing all non u_i factors. By Lemma 3.11, there exists $s_x \preceq \hat{s}_x$ such that its concatenation is a shuffle square. Define $s \preceq \hat{s}$ as the unique subword containing all factors form $\hat{s}_x$ and all $\tilde{u}_i$. Now, $\mu(\hat{s})$ also describes a shuffle square because $\mu(\hat{s}_x)$ is a shuffle square, $\tilde{u}_1 \tilde{u}_2 \cdots \tilde{u}_{\tilde{n}}$ is a shuffle square by construction and the first occurrence of the shuffle root of $\mu(s_x)$ ends before $\tilde{u}_1$. Furthermore, $\hat{s}$ is short enough because

$$|\hat{s}| \leq 2 \cdot (m-1) + \tilde{n} < 2m + n + \tilde{n} = |\tilde{f}|. \qquad \square$$

3.4 The Set of all Squares Determines its Root

In this section, we show that the mapping $w \mapsto w \sqcup w$ is injective, or equivalently, a word is uniquely determined by its set of shuffle squares.

When shuffling a word $w\mathtt{a}$ with itself, the first application of the inductive definition of the shuffle simplifies because the shuffle is commutative, that is, $w\mathtt{a} \sqcup w\mathtt{a} = (w\mathtt{a} \sqcup w)\mathtt{a} \cup (w \sqcup w\mathtt{a})\mathtt{a} = (w\mathtt{a} \sqcup w)\mathtt{a}$. The following lemma is a generalization of this fact.

Lemma 3.13. *Let $u \in \Sigma^*$, $\mathtt{a}, \mathtt{b} \in \Sigma$ and $n \in \mathbb{N}$, then*

$$\bigcup_{i=1}^{n} (u\mathtt{b}\mathtt{a}^i \sqcup u\mathtt{b})\mathtt{a}^{n-i} = \bigcup_{i=0}^{n} (u\mathtt{b}\mathtt{a}^i \sqcup u)\mathtt{b}\mathtt{a}^{n-i}.$$

Proof. By induction on n. For $n = 1$ we have

$$\begin{aligned}\bigcup_{i=1}^{1} (u\mathtt{b}\mathtt{a}^i \sqcup u\mathtt{b})\mathtt{a}^{1-i} &= (u\mathtt{b}\mathtt{a} \sqcup u\mathtt{b}) = (u\mathtt{b} \sqcup u\mathtt{b})\mathtt{a} \cup (u\mathtt{b}\mathtt{a} \sqcup u)\mathtt{b} \\ &= (u\mathtt{b} \sqcup u)\mathtt{b}\mathtt{a} \cup (u\mathtt{b}\mathtt{a} \sqcup u)\mathtt{b} = \bigcup_{i=0}^{1} (u\mathtt{b}\mathtt{a}^i \sqcup u)\mathtt{b}\mathtt{a}^{1-i}.\end{aligned}$$

Let $n + 1 \geq 2$, then we have

$$\begin{aligned}\bigcup_{i=1}^{n+1} (u\mathtt{b}\mathtt{a}^i \sqcup u\mathtt{b})\mathtt{a}^{n+1-i} &= \bigcup_{i=1}^{n+1} \left((u\mathtt{b}\mathtt{a}^{i-1} \sqcup u\mathtt{b})\mathtt{a}^{(n+1)-i+1} \cup (u\mathtt{b}\mathtt{a}^i \sqcup u)\mathtt{b}\mathtt{a}^{n+1-i}\right) \\ &= \bigcup_{i=1}^{n+1} (u\mathtt{b}\mathtt{a}^{i-1} \sqcup u\mathtt{b})\mathtt{a}^{(n+1)-i+1} \cup \bigcup_{i=1}^{n+1} (u\mathtt{b}\mathtt{a}^i \sqcup u)\mathtt{b}\mathtt{a}^{(n+1)-i}.\end{aligned} \tag{3.4}$$

Note that $\bigcup_{i=1}^{n+1}(u\mathtt{ba}^i \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i} \subseteq \bigcup_{i=0}^{n+1}(u\mathtt{ba}^i \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i}$. Therefore, we consider the first set. By induction, we have

$$\begin{aligned}
\bigcup_{i=1}^{n+1}(u\mathtt{ba}^{i-1} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{(n+1)-i+1} &= \left(\bigcup_{i=1}^{n+1}(u\mathtt{ba}^{i-1} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{(n+1)-i}\right)\mathtt{a} \\
&= \left(\bigcup_{i=0}^{n}(u\mathtt{ba}^{i} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{n-i}\right)\mathtt{a} \\
&= \left((u\mathtt{b} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{n} \cup \bigcup_{i=1}^{n}(u\mathtt{ba}^{i} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{n-i}\right)\mathtt{a} \\
&= \left((u\mathtt{b} \sqcup\!\sqcup u)\mathtt{ba}^{n} \cup \bigcup_{i=0}^{n}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{n-i}\right)\mathtt{a} \\
&= \left(\bigcup_{i=0}^{n}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{n-i}\right)\mathtt{a} \\
&= \bigcup_{i=0}^{n}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i}.
\end{aligned} \tag{3.5}$$

Thus, we conclude by Equations 3.4 and 3.5 that

$$\begin{aligned}
\bigcup_{i=1}^{n+1}(u\mathtt{ba}^{i} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{n+1-i} &= \bigcup_{i=1}^{n+1}(u\mathtt{ba}^{i-1} \sqcup\!\sqcup u\mathtt{b})\mathtt{a}^{(n+1)-i+1} \cup \bigcup_{i=1}^{n+1}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i} \\
&= \bigcup_{i=0}^{n}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i} \cup \bigcup_{i=1}^{n+1}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i} \\
&= \bigcup_{i=0}^{n+1}(u\mathtt{ba}^{i} \sqcup\!\sqcup u)\mathtt{ba}^{(n+1)-i}. \qquad \square
\end{aligned}$$

Instead of proving that the mapping $w \mapsto w \sqcup\!\sqcup w$ is injective directly, we prove the following stronger, auxiliary lemma. Using Lemma 3.14 we can prove Proposition 3.15 almost immediately.

Lemma 3.14. *Let $n \in \mathbb{N}$, $u, v \in \Sigma^*$ and $\mathtt{a} \in \Sigma$. If*

$$\bigcup_{i=1}^{n}(u\mathtt{a}^{i} \sqcup\!\sqcup u)\mathtt{a}^{n-i} = \bigcup_{i=1}^{n}(v\mathtt{a}^{i} \sqcup\!\sqcup v)\mathtt{a}^{n-i}$$

then there exists some $\ell \in [|u|]_0$ such that $u = \operatorname{pref}_\ell(u) \cdot \mathtt{a}^{|u|-\ell}$, $v = \operatorname{pref}_\ell(v) \cdot \mathtt{a}^{|v|-\ell}$ and $\operatorname{pref}_\ell(u) \sqcup\!\sqcup \operatorname{pref}_\ell(u) = \operatorname{pref}_\ell(v) \sqcup\!\sqcup \operatorname{pref}_\ell(v)$.

Proof. Because all elements of a shuffle have the same length, we always have $|u| = |v|$. We proceed by induction on $|u| = |v|$. If $|u| = |v| = 0$, then $u = v = \varepsilon$ and the claim holds for $\ell = 0$.

Let $u\mathtt{b}, v\mathtt{c} \in \Sigma^+$ for $\mathtt{b}, \mathtt{c} \in \Sigma$ such that the assumption holds. By the assumption and two applications of Lemma 3.13, we have

$$\bigcup_{i=0}^{n}(u\mathtt{b}\mathtt{a}^i \sqcup u)\mathtt{b}\mathtt{a}^{n-i} = \bigcup_{i=1}^{n}(u\mathtt{b}\mathtt{a}^i \sqcup u\mathtt{b})\mathtt{a}^{n-i} = \bigcup_{i=1}^{n}(v\mathtt{c}\mathtt{a}^i \sqcup v\mathtt{c})\mathtt{a}^{n-i} = \bigcup_{i=0}^{n}(v\mathtt{c}\mathtt{a}^i \sqcup v)\mathtt{c}\mathtt{a}^{n-i}.$$

By definition, $x \sqcup y \neq \emptyset$ for all $x, y \in \Sigma^*$. Thus, the above implies $\mathtt{b}\mathtt{a}^{n-i} = \mathtt{c}\mathtt{a}^{n-i}$ and therefore $\mathtt{b} = \mathtt{c}$.

Case 1 ($\mathtt{a} \neq \mathtt{b} = \mathtt{c}$). Therefore, the suffixes of all words in the $n+1$ unioned sets are unique, and therefore we obtain the following disjoint union

$$\bigsqcup_{i=0}^{n}(u\mathtt{b}\mathtt{a}^i \sqcup u)\mathtt{b}\mathtt{a}^{n-i} = \bigsqcup_{i=0}^{n}(v\mathtt{c}\mathtt{a}^i \sqcup v)\mathtt{c}\mathtt{a}^{n-i}.$$

Furthermore, note that the sets unioned above are pairwise equal for each choice of $i \in [n]_0$. For $i = 0$, the claim follows for $\ell := |u| + 1$ because

$$u\mathtt{b} \sqcup u\mathtt{b} = (u\mathtt{b} \sqcup u)\mathtt{b} = (v\mathtt{c} \sqcup v)\mathtt{c} = v\mathtt{c} \sqcup v\mathtt{c}.$$

Case 2 ($\mathtt{a} = \mathtt{b} = \mathtt{c}$). We obtain

$$\bigcup_{i=0}^{n}(v\mathtt{a}\mathtt{a}^i \sqcup v)\mathtt{a}\mathtt{a}^{n-i} = \bigcup_{i=0}^{n}(v\mathtt{a}^{i+1} \sqcup v)\mathtt{a}^{(n+1)-i} = \bigcup_{i=1}^{n+1}(v\mathtt{a}^i \sqcup v)\mathtt{a}^{(n+1)-i+1},$$

and by an analogous claim for u, that

$$\left(\bigcup_{i=1}^{n+1}(u\mathtt{a}^i \sqcup u)\mathtt{a}^{(n+1)-i}\right)\mathtt{a} = \left(\bigcup_{i=1}^{n+1}(v\mathtt{a}^i \sqcup v)\mathtt{a}^{(n+1)-i}\right)\mathtt{a}.$$

The claim follows now by the injectivity of concatenation and induction. □

Proposition 3.15. *The mapping $w \mapsto w \sqcup w$ is injective.*

Proof. Let $u, v \in \Sigma^*$ such that $u \sqcup u = v \sqcup v$. Because all elements of a shuffle have the same length, we always have $|u| = |v|$. We proceed by induction on $|u| = |v|$. If $|u| = |v| = 0$, then $u = v = \varepsilon$ and the claim holds for $\ell = 0$.

Let $u\mathtt{b}, v\mathtt{c} \in \Sigma^+$ for $\mathtt{b}, \mathtt{c} \in \Sigma$ such that the assumption holds. We obtain

$$(u\mathtt{b} \sqcup u)\mathtt{b} = u\mathtt{b} \sqcup u\mathtt{b} = v\mathtt{c} \sqcup v\mathtt{c} = (v\mathtt{c} \sqcup v)\mathtt{c}.$$

Therefore, $\mathtt{b} = \mathtt{c}$ and $u\mathtt{b} \sqcup u = v\mathtt{c} \sqcup v$. Thus, by Lemma 3.14 there exists $\ell \in [|u|]_0$ such that $u = \operatorname{pref}_\ell(u)\mathtt{b}^{|u|-\ell}$, $v = \operatorname{pref}_\ell(v)\mathtt{b}^{|v|-\ell}$ and $\operatorname{pref}_\ell(u) \sqcup \operatorname{pref}_\ell(u) = \operatorname{pref}_\ell(v) \sqcup \operatorname{pref}_\ell(v)$. By induction, we obtain $\operatorname{pref}_\ell(u) = \operatorname{pref}_\ell(v)$. □

Remark 3.15.1. We noticed later, that the above result follows from a stronger theorem by Berstel and Boasson [3]. They prove that if a set L can be described as a shuffle $\bigsqcup_{i=1}^{n} w_i$ of n words, then the multiset of these w_i is unique if a convention for unary words is introduced. For these, it obviously holds that $\mathtt{a}^3 = \mathtt{a} \sqcup\!\sqcup \mathtt{a}^2 = \mathtt{a} \sqcup\!\sqcup \mathtt{a} \sqcup\!\sqcup \mathtt{a}$. Choosing a unique representation for these (in their case forbidding powers of single letters and instead shuffling multiple copies of them) makes the decomposition unique up to permutation.

In our case, both ambiguities cannot occur because we always shuffle two copies of the same word. We still kept our result because we give less technical proof which suffices for the claim about mapping we implicitly considered in this whole chapter.

3.5 Conclusion and Future Work

In this chapter, we investigated sets arising as shuffles of some set in connection (reverse) shuffle squares. We showed that a shuffle of a set with itself can only coincide with the set of shuffle squares of its elements, if the set is a singleton or empty. Furthermore, we showed that a similar result as well as a generalization holds for reverse shuffle squares. In Section 3.3 we studied the complexity of shuffle squares from a language theoretic point of view. We proved a generalization of a result by Henshall, Rampersad, and Shallit [16] and started an investigation of whether the language of shuffle squares is indexed. Lastly, in Section 3.4, we showed that a set of shuffle squares of a word w determines w uniquely, that is, that the map $w \mapsto w \sqcup\!\sqcup w$ is unique.

Potentially future works includes the claimed Conjectures 3.6 and 3.8. Conjecture 3.6 continues the line of propositions differentiating between (reverse) shuffles of set and there subsets of shuffle squares. Conjecture 3.8 concerns the complexity of the language of shuffle squares. An open question is whether it satisfies the last of the three necessary conditions by Hayashi [14]. Furthermore, Fratani and Voundy [11, Theorem 18] give a characterization of indexed languages which could also be used to approach this problem.

In Section 3.4, we studied the mapping $v \mapsto v \sqcup\!\sqcup v$. This mapping extends naturally to languages by $L \mapsto \bigcup_{v \in L} v \sqcup\!\sqcup v$. Because these sets arise as unions of shuffles, our Proposition 3.15 and the result by Berstel and Boasson [3] do not apply. An open question is whether this mapping is also injective. Furthermore, a generalization of this open questions is which sets arising as unions of shuffle sets, we can still decompose into a unique set of multisets.

Chapter 4

$\alpha\beta$-factorization and the Binary Case of Simon's Congruence

For some word w a word v of minimal length such that $v \not\preceq w$ is called a *shortest absent subword* (or *shortest absent subsequence*). These where investigated algorithmically by Kosche et al. [21] and combinatorial by Fleischmann et al. [10]. In this chapter we introduce the $\alpha\beta$-factorization, which was used by Fleischmann et al. [10] when investigating words with absent subwords of minimal length and characterizing those with a unique shortest absent subword.

We first define the factorization in Section 4.1 in its original context, investigate classes of words with a fixed number of shortest absent subwords and characterize two classes of words with a fixed number of absent subwords. In Section 4.2 we prove some preliminary results about the factorization in the general case. In Section 4.3, we apply these results in the special case of a binary alphabet, to completely characterize the equivalence classes. Using the characterization for the binary case, we can calculate the index of Simon's congruence for this case in Section 4.3.1. Lastly, in Section 4.4 we prove a characterization of k-equivalence of m-universal words in terms of their $\alpha\beta\alpha$-factors.

4.1 $\alpha\beta$-Factorization and Fixed Numbers of Subwords

We can refine the arch factorization and regain symmetry by factorizing in both direction.

Definition 4.1 ($\alpha\beta$-Factorization [10])**.** Let $w \in \Sigma^*$ and denote the arches modi, and rest of its backwards arch factorization by $\overleftarrow{\mathrm{ar}}_i(w) \coloneqq \mathrm{ar}^R_{\iota(w)-i+1}(w^R)$, $\overleftarrow{\mathrm{m}}_i(w) \coloneqq \mathrm{m}_{\iota(w)-i+1}(w^R)$ and $\overleftarrow{\mathrm{re}}(w) \coloneqq \mathrm{re}^R(w^R)$ and call them *backwards arches*, *backwards modi* and *backwards rest* respectively. Note that these correspond, as expected, to the i^{th} backwards arch or backwards modus when counting from left to right.

Choose $\alpha_0, \alpha_1, \ldots, \alpha_{\iota(w)}, \beta_1, \ldots, \beta_{\iota(w)} \in \Sigma^*$ such that

- $\alpha_{i-1}\beta_i = \mathrm{ar}_i(w)$ for all $i \in [\iota(w)]$,
- $\beta_i\alpha_i = \overleftarrow{\mathrm{ar}}_i(w)$ for all $i \in [\iota(w)]$,
- and $\alpha_{\iota(w)} = \mathrm{re}(w)$ and $\alpha_0 = \overleftarrow{\mathrm{re}}(w)$.

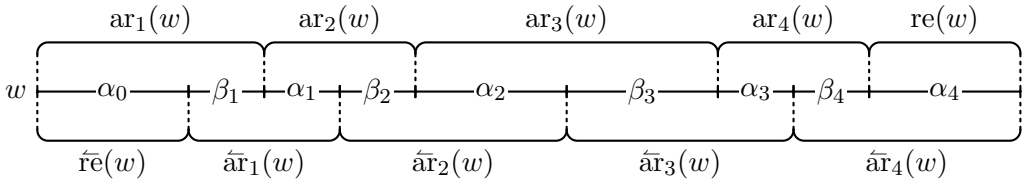


Figure 4.1: $\alpha\beta$-Factorization of a word w where $\alpha_i \coloneqq \mathrm{A}_i(w)$ for all $i \in [\iota(w)]_0$ and $\beta_i \coloneqq \mathrm{B}_i(w)$ for all $i \in [\iota(w)]$.

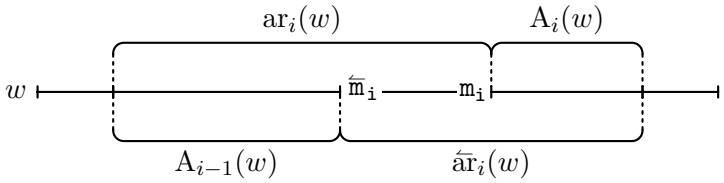


Figure 4.2: Structure of the i^{th} $\alpha\beta\alpha$-factor of a word w where $\mathtt{m}_\mathtt{i} \coloneqq \mathrm{m}_i(w)$ and $\overleftarrow{\mathtt{m}}_\mathtt{i} \coloneqq \overleftarrow{\mathrm{m}}_i(w)$.

Then $w = \alpha_0\beta_1\alpha_1 \cdots \beta_{\iota(w)}\alpha_{\iota(w)}$ is called the *$\alpha\beta$-factorization* of w. Denote by $\mathrm{A}(w)$ the $(\iota(w)+1)$-tuple $(\alpha_0, \alpha_1, \ldots, \alpha_{\iota(w)})$ and by $\mathrm{B}(w)$ the $\iota(w)$-tuple $(\beta_1, \beta_2, \ldots, \beta_{\iota(w)})$. Furthermore, denote by $\mathrm{A}_i(w)$ and $\mathrm{B}_i(w)$ the $(i+1)^{\text{th}}$ and the i^{th} element of the above tuples, respectively.

Remark 4.1.1. Figures 4.1 and 4.2 show a word in $\alpha\beta$-factorization and the structure of the single $\alpha\beta\alpha$-factors. The $\alpha\beta$-factorization is unique and always exists. The B_i are always non-empty since they contain the modi of $\mathrm{ar}_i(w)$ and $\overleftarrow{\mathrm{ar}}_i(w)$. Furthermore, the first and last letter, as modus of the corresponding arches, are unique in B_i. Because the last letter of each arch is unique, arches are factorized in α, β pairs, and B_i contains the unique letter of the arch, we have $\mathrm{alph}(\mathrm{A}_i(w)) \subset \Sigma$.

Example 4.2. The $\alpha\beta$-factorization of `abcabaabbbcbcccb` is given by `ab`$\cdot$`c`$\cdot$`aba`$\cdot$`abbbc`$\cdot$`bccb`. The $\alpha\beta$-factorization of `abaabcbaca` is given by `aba` $\cdot$ `abc` $\cdot\, \varepsilon \cdot$ `bac` $\cdot$ `a`. Note that the third factor is empty.

4.1.1 Words with Many Subwords

Fleischmann et al. [10] show the following characterization of words with exactly one absent subword of minimal length using the factorization. We restate the theorem in our notation, as a motivation for using $\alpha\beta$-factorization as a tool for studying subwords.

Theorem 4.3 ([10, Theorem 16])**.** *Let* $|\Sigma| \geq 2$*,* $k \in \mathbb{N}$ *and* $w \in \Sigma^*$*, then* $|\,\mathrm{SubWords}_k(w)| = |\Sigma|^k - 1$*, if and only if,* $|\,\mathrm{alph}(\mathrm{A}_i(w))| = |\Sigma| - 1$ *for all* $i \in [\iota(w)]_0$ *and* $\iota(w) = k - 1$*.*

Proof. Assume $|\,\mathrm{SubWords}_k(w)| = |\Sigma|^k - 1$. Suppose we have $\iota(w) < k - 1$. Then, for $\mathtt{x} \not\preceq \mathrm{re}(w)$, every $\mathrm{m}(w) \cdot v \in \Sigma^k$ with $\mathtt{x} \preceq v$ would not be a subword of w. Since

$|v| \geq 2$ and $|\Sigma| \geq 2$, there exist at least two absent subwords of length k, a contradiction. Furthermore, we have $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 1$ because otherwise $\prod_{j=1}^{i} \mathrm{m}_i(w) \cdot (\Sigma \setminus \operatorname{alph}(\mathrm{A}_i(w))) \cdot \prod_{j=i+1}^{k-1} \overleftarrow{\mathrm{m}}_i(w)$ would all not occur in w.

Assume $\iota(w) = k - 1$ and that all $\mathrm{A}_i(w)$ are missing exactly one letter. Let $v \in \Sigma^k$ with $v \not\preceq w$ by Remark 2.5.1. We show that each letter of $v[\ell + 1]$ is already uniquely determined for all $\ell \in [k-1]_0$. We have $v[1..\ell] \preceq \prod_{i=1}^{\ell} \mathrm{ar}_i(w)$ by ℓ-universality and $v[\ell + 2 \dots k] \preceq \prod_{i=\ell+1}^{k-1} \overleftarrow{\mathrm{ar}}_i(w)$ by $(k-1-\ell)$-universality. Therefore, $v[\ell+1] \not\preceq \mathrm{A}_\ell(w)$ because otherwise we would have $v \in \mathrm{SubWords}_{\leq k}(w)$. Thus, $v[\ell+1] \notin \operatorname{alph}(\mathrm{A}_\ell(w))$ and therefore $v[\ell+1]$ is the unique missing letter of $\mathrm{A}_\ell(w)$. □

The proof uses the following nice fact about the shortest absent subwords. If $v \in \Sigma^{\iota(w)+1} \setminus \mathrm{SubWords}_{\iota(w)+1}(w)$, then $v[i+1] \notin \operatorname{alph}(\mathrm{A}_i)$ for all $i \in [\iota(w)]_0$. Note that this is just a necessary condition, but not sufficient as $\mathtt{bc} \preceq \mathtt{a} \cdot \mathtt{bc} \cdot \mathtt{a}$.

In the case that exactly one subword is absent, the one missing letter of $\mathrm{A}_i(w)$ coincides exactly with the modus $\mathrm{m}_{i+1}(w)$. Because the alphabet of $\mathrm{A}_i(w)$ does not change when reversing the word, it also coincides with the modus of the corresponding reversed arch $\overleftarrow{\mathrm{m}}_i(w)$. We can generalize this idea to obtain rough upper and lower bounds on the number of absent subwords of length $\iota(w)+1$. Figures 4.3 and 4.4 show the constructions in the proof of Lemma 4.4.

Lemma 4.4. *Let $w \in \Sigma^*$ such that $k \coloneqq \iota(w) + 1$ and $\Sigma_i \coloneqq \operatorname{alph}(\mathrm{A}_i(w)) \subset \Sigma$, then*

$$\sum_{i=0}^{\iota(w)} \left|\Sigma_i^{\complement}\right| - \iota(w) = 1 + \sum_{i=0}^{\iota(w)} \left(\left|\Sigma_i^{\complement}\right| - 1\right) \leq |\Sigma|^k - |\operatorname{SubWords}_k(w)| \leq \prod_{i=0}^{\iota(w)} \left|\Sigma_i^{\complement}\right|.$$

Proof. Define

$$M_i(w) \coloneqq \prod_{j=1}^{i} \mathrm{m}_j(w) \cdot \Sigma_i^{\complement} \cdot \prod_{j=i+1}^{\iota(w)} \overleftarrow{\mathrm{m}}_j(w) \subseteq \Sigma^k.$$

We have $M_i(w) \subseteq \Sigma^k \setminus \mathrm{SubWords}_k(w)$ because for $v \in M_i$ we have $v \not\preceq w$, if and only if, $v[i+1] \not\preceq \mathrm{A}_i(w)$, which holds by definition of v. We show by induction that

$$\left|\bigcup_{i=0}^{\iota(w)} M_i(w)\right| = 1 + \sum_{i=0}^{\iota(w)} \left(\left|\Sigma_i^{\complement}\right| - 1\right).$$

For $\iota(w) = 0$, the claim holds trivially, since $|\Sigma_0^{\complement}| = 1 + |\Sigma_0^{\complement}| - 1$. Therefore, assume that $\iota(w) \geq 1$. By inclusion-exclusion, we immediately obtain

$$\begin{aligned} \left|\bigcup_{i=0}^{\iota(w)} M_i(w)\right| &= \left|M_0(w) \cup \bigcup_{i=1}^{\iota(w)} M_i(w)\right| \\ &= \left|M_0(w)\right| + \left|\bigcup_{i=1}^{\iota(w)} M_i(w)\right| - \left|M_0(w) \cap \bigcup_{i=1}^{\iota(w)} M_i(w)\right|. \end{aligned} \tag{4.1}$$

We handle the three sets separately. Note that for $\tilde{w} \coloneqq \mathrm{ar}_1^{-1}(w) \cdot w$, the function

$$\bigcup_{i=0}^{\iota(\tilde{w})} M_i(\tilde{w}) \to \bigcup_{i=1}^{\iota(w)} M_i(w), v \mapsto \mathrm{m}_1(w) \cdot v \tag{4.2}$$

is a bijection. Therefore, we can apply the induction hypothesis for the second set. For the third set we show an equivalence to calculate its cardinality. Note that we have

$$M_0(w) \cap \bigcup_{i=1}^{\iota(w)} M_i(w) \subseteq \left\{ \mathrm{m}_1(w) \cdot \prod_{j=1}^{\iota(w)} \overleftarrow{\mathrm{m}}_j(w) \right\} \tag{4.3}$$

because all elements of $\bigcup_{i=1}^{\iota(w)} M_i(w)$ share the common prefix $\mathrm{m}_1(w) \not\preceq \mathrm{A}_0(w)$. Furthermore, all elements of $M_0(w)$ have the common suffix $\prod_{j=1}^{\iota(w)} \overleftarrow{\mathrm{m}}_j(w)$. In addition to that, $\overleftarrow{\mathrm{m}}_1(w) \not\preceq \mathrm{A}_1(w)$, and thus

$$M_0(w) \cap \bigcup_{i=1}^{\iota(w)} M_i(w) \supseteq M_0 \cap M_1 \supseteq \left\{ \mathrm{m}_1(w) \cdot \prod_{j=1}^{\iota(w)} \overleftarrow{\mathrm{m}}_j(w) \right\}. \tag{4.4}$$

Now we have

$$\begin{aligned}
\left| \bigcup_{i=0}^{\iota(w)} M_i(w) \right| &= \left| M_0(w) \right| + \left| \bigcup_{i=1}^{\iota(w)} M_i(w) \right| - \left| M_0(w) \cap \bigcup_{i=1}^{\iota(w)} M_i(w) \right| && \text{(Equation 4.1)} \\
&= \left| \Sigma_0^{\complement} \right| + \left(1 + \sum_{i=1}^{\iota(w)} \left(\left| \Sigma_i^{\complement} \right| - 1 \right) \right) - \left| M_0(w) \cap \bigcup_{i=1}^{\iota(w)} M_i(w) \right| && \text{(Eq. 4.2 and IH)} \\
&= \left| \Sigma_0^{\complement} \right| + \left(1 + \sum_{i=1}^{\iota(w)} \left(\left| \Sigma_i^{\complement} \right| - 1 \right) \right) - 1 && \text{(Equations 4.3 and 4.4)} \\
&= 1 + \sum_{i=0}^{\iota(w)} \left(\left| \Sigma_i^{\complement} \right| - 1 \right).
\end{aligned}$$

It is left to show the upper bound. Let $v \in \Sigma^k \setminus \mathrm{SubWords}_k(w)$ and $j \in [k]$. Since

$$v[1..j-1] \preceq \prod_{i=1}^{j-1} \mathrm{ar}_i(w) \qquad \text{and} \qquad v[j+1..k] \preceq \prod_{i=j}^{\iota(w)} \overleftarrow{\mathrm{ar}}_i(w),$$

we have $v[j] \not\preceq \mathrm{A}_{j-1}(w)$ because otherwise we would have $v \preceq w$. Thus, $v[j] \notin \mathrm{alph}(\mathrm{A}_{j-1}(w))$ which is equivalent to $v[j] \in \Sigma \setminus \mathrm{alph}(\mathrm{A}_{j-1}(w)) = \Sigma_{j-1}^{\complement}$ and therefore

$$\Sigma^k \setminus \mathrm{SubWords}_k(w) \subseteq \prod_{i=0}^{\iota(w)} \Sigma_i^{\complement}. \qquad \square$$

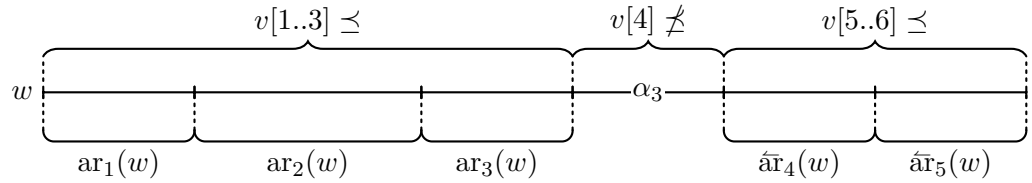


Figure 4.3: Occurrences of a prefix and a suffix of an absent subword $v \not\preceq w$ in the proof of Lemma 4.4.

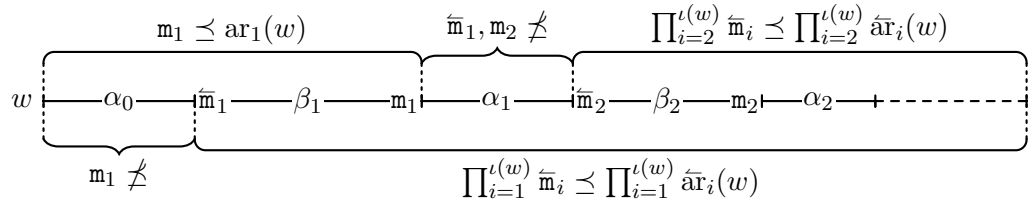


Figure 4.4: The absent subword in $M_0 \cap M_1$ for some w in the proof of Lemma 4.4 where $\mathtt{m}_i \coloneqq \mathrm{m}_i(w)$ and $\acute{\mathtt{m}}_i \coloneqq \acute{\mathrm{m}}_i(w)$ for all $i \in [\iota(w)]$.

Remark 4.4.1. In the above lemma, we only considered subwords of length $\iota(w)+1$. The used necessary and sufficient conditions can be generalized. For each $v_i \not\preceq \mathrm{A}_i(w)$, we have $\prod_{j=1}^{i} \mathrm{m}_j(w) \cdot v_i \cdot \prod_{j=i+1}^{\iota(w)} \acute{\mathrm{m}}_j(w) \not\preceq w$ by Remark 2.5.1. Furthermore, for $v \not\preceq w$ we have $v[i..i+|v|-\iota(w)-1] \not\preceq A_i(w)$. Therefore, we can construct analogous sets as in the above proof, but finding sufficiently general formulas for their cardinality becomes harder.

Corollary 4.4.2. *Let $w \in \Sigma^*$ with $k \coloneqq \iota(w)+1$ such that there exists $i \in [\iota(w)]_0$ such that $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - \ell$ and $|\operatorname{alph}(\mathrm{A}_j(w))| = |\Sigma| - 1$ for all $j \in [\iota(w)]_0 \setminus \{i\}$, then $|\operatorname{SubWords}_k(w)| = |\Sigma|^k - \ell$.*

Proof. By assumption, we have equality in the statement of Lemma 4.4. □

Fleischmann et al. [10, Theorem 34] give a number of necessary conditions for the structure of words with exactly two absent subwords of length k. We can give a straightforward characterization in terms of the A_i, similar to Theorem 4.3, because we can show that $\iota(w) = k-1$ and the missing letters of the A_i determine letters of the absent subwords. The claim for $\iota(w)$ will be the main problem going forward. If one A_ℓ is missing two letters, we can construct two different absent subwords by Lemma 4.4.

Proposition 4.5. *Let $w \in \Sigma^*$, then $|\operatorname{SubWords}_k(w)| = |\Sigma|^k - 2$, if and only if, $\iota(w) = k-1$ and there exists $\ell \in [k-1]_0$ such that $|\operatorname{alph}(\mathrm{A}_\ell(w))| = |\Sigma| - 2$ and for all $i \in [k]_0 \setminus \{\ell\}$ we have $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 1$.*

Proof. The backwards direction follows directly from Corollary 4.4.2. If $|\Sigma| \leq 1$, then the proposition is vacuously true. Therefore, assume that $|\Sigma| \geq 2$ and that w has exactly two absent subwords of length k.

Suppose $m \coloneqq \iota(w) < k-1$. Choose $\mathtt{x} \in \Sigma$ with $\mathtt{x} \not\preceq \mathrm{re}(w)$. Then for all $v \in \Sigma^{k-m}$ with $\mathtt{x} \preceq v$ we have $\mathrm{m}(w) \cdot v \not\preceq w$. Since there exist at least

$$\sum_{i=1}^{k-m} \binom{i-1}{1-1} \cdot (|\Sigma|-1)^{i-1} \cdot |\Sigma|^{k-m-i} \geq \sum_{i=1}^{2} (2-1)^{i-1} \cdot 2^{2-i} = 3$$

such words by Lemma 3.2, a contradiction.

Therefore, we have $\iota(w) = k-1$. By definition, we have $|\operatorname{alph}(\mathrm{A}_i(w))| \leq |\Sigma| - 1$ for all $i \in [m]_0$. Furthermore, we can assume $|\Sigma| - 2 \leq |\operatorname{alph}(\mathrm{A}_i(w))|$ for all $i \in [m]_0$ because otherwise, we could construct $|\Sigma| - |\operatorname{alph}(\mathrm{A}_i(w))| > 2$ absent subwords of length k by Lemma 4.4. We cannot have $|\Sigma| - 1 = \operatorname{alph}(\mathrm{A}_i(w))$ for all $i \in [m]_0$ by Theorem 4.3. Furthermore, we cannot have more than one $\mathrm{A}_i(w)$ missing more than one letter by the lower bound from Lemma 4.4. □

Corollary 4.5.1. *Let $|\Sigma| \geq 2$, then*

$$\left|\left\{w \in \Sigma^* \,\middle|\, |\operatorname{SubWords}_k(w)| = |\Sigma|^k - 2\right\} / \sim_k\right| = k \cdot \binom{|\Sigma|}{2} \cdot |\Sigma|^{k-1}.$$

Proof. Let $w \in \Sigma^*$ such that $|\operatorname{SubWords}_k(w)| = |\Sigma|^k - 2$. Then, the two absent subwords of length k are by Proposition 4.5 of the form

$$\prod_{i=1}^{\ell} \mathrm{m}_i(w) \cdot (\Sigma \setminus \operatorname{alph}(\mathrm{A}_\ell(w))) \cdot \prod_{i=\ell+1}^{m} \overleftarrow{\mathrm{m}}_i(w)$$

for some $\ell \in [\iota(w)]_0$. Furthermore, let $u, v \in \Sigma^*$ with $|uv| = k-1$ and $\Omega \in \binom{\Sigma}{2}$, then we can construct a word w with $\Sigma^k \setminus \operatorname{SubWords}_k(w) = u\Omega v$. Then, there exist exactly one class for each element of

$$\bigsqcup_{i \in [k]} \Sigma^{i-1} \times \binom{\Sigma}{2} \times \Sigma^{k-i} \cong [k] \times \binom{\Sigma}{2} \times \Sigma^{k-1},$$

such that for (u, Ω, v) the two absent subword are exactly $u\Omega v$. We give an inductive construction. For some $\Xi \subseteq \Sigma$, let $\pi(\Xi)$ denote the set of permutations of Ξ. If $|uv| = 0$, then let $w \in \pi(\Sigma \setminus \Omega)$. Thus, assume $|uv| = k+1$. We show how to extend u to the left, then the claim follows by symmetry. Let $u = \mathtt{a}u'$ and w by induction such that $\Sigma^k \setminus \operatorname{SubWords}_k(w) = u'\Omega v$. Write $w = \alpha_0 \cdot \tilde{w}$ where $\alpha_0 \coloneqq \mathrm{A}_0(w)$. Let

$$w' \in \pi(\Sigma \setminus \{\mathtt{a}\}) \cdot \pi(\Sigma \setminus (\operatorname{alph}(\alpha_0) \cup \{\mathtt{a}\})) \cdot \mathtt{a} \cdot \alpha_0 \cdot \tilde{w}.$$

By construction, we have $\mathrm{ar}_1(w') \in \pi(\Sigma \setminus \{\mathtt{a}\}) \cdot \pi(\Sigma \setminus \operatorname{alph}(\alpha_0) \setminus \{\mathtt{a}\}) \cdot \mathtt{a}$ and $\mathrm{A}_0(w') \in \pi(\Sigma \setminus \{\mathtt{a}\})$. Furthermore, the factorization of w does not change, implying by Proposition 4.5, that $u\Omega v$ are the two absent subwords of length k. □

Remark 4.5.2. Consider the words $w_1 = \mathtt{a}\cdot\mathtt{bcd}\cdot\mathtt{a}$ and $w_2 = \mathtt{a}\cdot\mathtt{bccd}\cdot\mathtt{a}$ in $\alpha\beta$-factorization. We have $w_1 \prec w_2$ and thus $\mathrm{SubWords}_2(w_1) \subseteq \mathrm{SubWords}_2(w_2)$. In particular, we have $\mathrm{SubWords}_2(w_2)\backslash\mathrm{SubWords}_2(w_1) = \{\mathtt{cc}\}$ and w_1 and w_2 have six and five absent subwords of length two respectively. Note that both w_1 and w_2 have the same A_i. Furthermore, $\mathtt{cc}$ does not contain any modi. Therefore, we cannot characterize words with ≥ 3 absent subwords, by the structure of their A alone.

This effect can also be seen in the last characterization in this section. We give an equivalent description of words with exactly three absent subwords. Note that for these words we do not always have $\iota(w) = k-1$ and that the structure of their B_i is relevant.

Proposition 4.6. *Let $w \in \Sigma^*$, then $|\operatorname{SubWords}_k(w)| = |\Sigma|^k - 3$, if and only if, either $w \in \mathtt{x}\mathtt{x}^+ \subseteq \Sigma_2^*$ for $\mathtt{x} \in \Sigma$ and $k = 2$ or $\iota(w) = k-1$, one of the following holds*

(1) there exists $i \in [k-1]_0$ with $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 3$

(2) there exists $i \in [k-1]$ with $|\operatorname{alph}(\mathrm{A}_j(w))| = |\Sigma| - 2$ for all $j \in \{i-1, i\}$ and $\mathtt{ab} \preceq \mathrm{B}_i(w)$ where $\mathrm{m}_i(w) \neq \mathtt{a} \not\preceq \mathrm{A}_{i-1}(w)$ and $\overleftarrow{\mathrm{m}}_i(w) \neq \mathtt{b} \not\preceq \mathrm{A}_i(w)$ for $\mathtt{a}, \mathtt{b} \in \Sigma$,

and all other $\mathrm{A}_\ell(w)$ are missing exactly one letter.

Proof. If $|\Sigma| \leq 1$, then the proposition is vacuously true, thus assume $|\Sigma| \geq 2$. We first consider the backwards direction. If $k = 2$ and $w = \mathtt{x}\mathtt{x}^+ \subseteq \Sigma_2$, then $\mathrm{SubWords}_2(w) = \{\mathtt{x}^2\}$ and the claim follows immediatly. If exactly one $\mathrm{A}_i(w)$ is missing three letters, then the claim follows directly from Lemma 4.4. Therefore, assume $\mathrm{A}_\ell(w), \mathrm{A}_{\ell+1}(w)$ are missing two letters each, that is, let $\mathtt{a}, \mathtt{b} \in \Sigma$ such that $\mathrm{m}_{\ell+1}(w) \neq \mathtt{a} \not\preceq \mathrm{A}_\ell(w)$ and $\overleftarrow{\mathrm{m}}_{\ell+1}(w) \neq \mathtt{b} \not\preceq \mathrm{A}_{\ell+1}(w)$. By Lemma 4.4, there exist four potentially subwords. By assumption, we have $\mathtt{ab} \preceq \mathrm{B}_\ell(w)$ and thus

$$\prod_{i=1}^{\ell} \mathrm{m}_i(w) \cdot \mathtt{ab} \cdot \prod_{i=\ell+2}^{\iota(w)} \overleftarrow{\mathrm{m}}_i(w) \preceq w.$$

Therefore, the upper bound from Lemma 4.4 is not tight and w has exactly 3 absent subwords by the lower bound.

For the other direction, let $w \in \Sigma^*$ with $|\operatorname{SubWords}_k(w)| = |\Sigma|^k - 3$. Suppose $\iota(w) < k-1$. We distinguish two cases.

Case 1 ($|\Sigma| \geq 3$). Choose $\mathtt{x} \in \Sigma$ with $\mathtt{x} \not\preceq \mathrm{re}(w)$. Then for all $v \in \Sigma^{k-\iota(w)}$ with $\mathtt{x} \preceq v$ we have $\mathrm{m}(w) \cdot v \not\preceq w$. Since there exist at least

$$\sum_{i=1}^{k-\iota(w)} \binom{i-1}{1-1} \cdot (|\Sigma|-1)^{i-1} \cdot |\Sigma|^{k-\iota(w)-i} \geq \sum_{i=1}^{2} (3-1)^{i-1} \cdot 3^{2-i} = 1 \cdot 3 + 2 \cdot 1 = 5$$

such words by Lemma 3.2, a contradiction against $|\Sigma|^k - |\operatorname{SubWords}_k(w)| = 3$.

Case 2 ($|\Sigma| = 2$). If $\iota(w) = 0$ we have $w = \mathtt{x}^n$ for $\mathtt{x} \in \Sigma_2, n \in \mathbb{N}_0$. If $k \geq 3$ then we have a contradiction by Lemma 3.2. If $k \leq 2$, then we have $k = 2$ by the supposition. If $n \leq 1$, then $|w| \leq 1$ and thus $\mathrm{SubWords}_k(w) = \emptyset$. Otherwise, if $n \geq 2$, $\mathtt{x}\overline{\mathtt{x}}, \overline{\mathtt{x}}\mathtt{x}, \overline{\mathtt{x}}^2$ are exactly the absent subwords.

Now assume $\iota(w) \geq 1$. We consider absent subwords of the form $v \cdot \hat{\mathrm{m}}(w)$ and $\mathrm{m}(w) \cdot u$ for $v, u \in \Sigma^{k-\iota(w)}$. By Lemma 3.2, we have at least three different choices for v and u each. Since $\iota(w) \geq 1$, all absent subwords of the form $\mathrm{m}(w) \cdot u$ start with the same letter. Because there exist choices for v with different first letters, we have at least four absent subwords.

Thus, we either have $\iota(w) = k - 1$ or $w = \mathtt{x}^2\mathtt{x}^* \subseteq \Sigma_2^*$ and $k = 2$. We only have to consider the first case. We can assume $|\Sigma| - 3 \leq |\operatorname{alph}(\mathrm{A}_i(w))| \leq |\Sigma| - 1$ for all $i \in [m]_0$ by the lower bound from Lemma 4.4 and definition. The following cases yield the valid combinations of the A_i's alphabets.

Case 1 (there exists i such that $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 3$). There cannot exist $j \neq i$ with $|\Sigma| - |\operatorname{alph}(\Sigma_j)| \geq 2$, by the lower bound from Lemma 4.4.

Case 2 ($|\operatorname{alph}(\mathrm{A}_i(w))| \geq |\Sigma| - 2$ for all $i \in [k-1]_0$). We have to consider the two subcases that either three different A are missing two letters or the two A are not next to each other.

Case 2.a (There exist $x < y < z$ s.t. $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 2$ for $i \in \{x, y, z\}$). There cannot exist x, y, z pairwise different with $|\Sigma| - |\operatorname{alph}(\Sigma_i)| = 2$ for all $i \in \{x, y, z\}$, by the lower bound from Lemma 4.4.

Case 2.b (There exist $x + 1 = y$ s.t. $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 2$ for $i \in \{x, y\}$). By Subcase 2.a, we can assume that all other $\mathrm{A}_i(w)$ are missing exactly one letter. Let $\Sigma \setminus \operatorname{alph}(\mathrm{A}_x(w)) = \{\mathrm{m}_{x+1}(w), \mathtt{a}\}$ and $\Sigma \setminus \operatorname{alph}(\mathrm{A}_y(w)) = \{\hat{\mathrm{m}}_y(w), \mathtt{b}\}$ by Remark 4.1.1. There exist four potential absent subwords, three of which are always absent by construction (the ones from the construction of the lower bound in Lemma 4.4). The fourth word $\prod_{i=1}^{x} \mathrm{m}_i(w) \cdot \mathtt{ab} \cdot \prod_{i=y+1}^{\iota(w)} \hat{\mathrm{m}}_i(w)$ occurs, if and only if, $\mathtt{ab} \preceq \mathrm{B}_{x+1}(w) = \mathrm{B}_y(w)$.

Case 2.c (There exist $x + 1 < y$ s.t. $|\operatorname{alph}(\mathrm{A}_i(w))| = |\Sigma| - 2$ for $i \in \{x, y\}$). By Subcase 2.a, we can assume that all other $\mathrm{A}_i(w)$ are missing exactly one letter. Let $\Sigma \setminus \operatorname{alph}(\mathrm{A}_x(w)) = \{\mathrm{m}_{x+1}(w), \mathtt{a}\}$ and $\Sigma \setminus \operatorname{alph}(\mathrm{A}_y(w)) = \{\hat{\mathrm{m}}_y(w), \mathtt{b}\}$ by Remark 4.1.1. There exist four potential absent subwords, three of which are always absent by construction (the ones from the construction of the lower bound in Lemma 4.4).

Define $\Sigma_i^{\complement} \coloneqq \Sigma \setminus \operatorname{alph}(\mathrm{A}_i(w))$ and note that $\Sigma_j^{\complement} = \{\hat{\mathrm{m}}_j(w) = \mathrm{m}_{j+1}(w)\}$ for $j \in [k-1]_0 \setminus \{x, y\}$. We consider the fourth potentially absent subword[1]

$$v \coloneqq \prod_{i=0}^{x-1} \Sigma_i^{\complement} \cdot \mathtt{a} \cdot \prod_{i=x+1}^{y-1} \Sigma_i^{\complement} \cdot \mathtt{b} \cdot \prod_{i=y+1}^{\iota(w)} \Sigma_i^{\complement} = \prod_{i=1}^{x} \mathrm{m}_i(w) \cdot \mathtt{a} \cdot \prod_{i=x+1}^{y-1} \Sigma_i^{\complement} \cdot \mathtt{b} \cdot \prod_{i=y+1}^{\iota(w)} \hat{\mathrm{m}}_i(w).$$

[1] In an abuse of notation, we equate singleton sets with their single element to simplify some notation.

By Theorem 4.3 and the uniqueness of $\mathrm{m}_\ell(w)$ and $\overleftarrow{\mathrm{m}}_\ell(w)$ in $\mathrm{B}_\ell(w)$, we have that

$$\prod_{i=x+1}^{y-1} \Sigma_i^{\complement} \not\preceq \overbrace{\overleftarrow{\mathrm{m}}_{x+1}^{-1}(w) \cdot \mathrm{B}_{x+1}(w)}^{\Sigma_{x+1}^{\complement}=\{\overleftarrow{\mathrm{m}}_{x+1}(w)\}\not\preceq} \cdot \overbrace{\mathrm{A}_{x+1}(w) \cdots \mathrm{A}_{y-1}(w)}^{\prod_{x+1}^{y-1}\Sigma_i^{\complement}\not\preceq} \cdot \overbrace{\mathrm{B}_y(w) \cdot \mathrm{m}_y^{-1}(w)}^{\Sigma_{y-1}^{\complement}=\{\mathrm{m}_y(w)\}\not\preceq}.$$

Furthermore, it holds that $\mathtt{a} \not\preceq \mathrm{A}_x(w)$ and $\mathtt{b} \not\preceq \mathrm{A}_y(w)$. Using this fact and the above equation, we conclude that

$$\begin{aligned}\mathtt{a} \cdot \prod_{i=x+1}^{y-1} \Sigma_i^{\complement} \cdot \mathtt{b} &\not\preceq \mathrm{A}_x(w) \cdot \mathrm{B}_{x+1}(w) \cdot \mathrm{A}_{x+1}(w) \cdots \mathrm{A}_{y-1}(w) \cdot \mathrm{B}_y(w) \cdot \mathrm{A}_y(w) \\ &= \Bigg(\prod_{i=1}^{x} \mathrm{ar}_i(w)\Bigg)^{-1} \cdot w \cdot \Bigg(\prod_{i=y+1}^{\iota(w)} \mathrm{ar}_i(w)\Bigg)^{-1},\end{aligned}$$

which is equivalent to $v \not\preceq w$ by Remark 2.5.1 and the definition of v. □

4.1.2 Words with Few Subwords

We now consider words with very small numbers of subwords of length k. Fleischmann et al. [10, Proposition 32] show a condition on the alphabet of subwords, for words with exactly two subwords of a fixed length k. We generalize this proposition. For words with less than two subwords of length k, the alphabet can obviously be as large as k. Note that, every word with at least two different subwords of length k, is at least $k+1$ letters long. We therefore impose a length condition on the words instead of a lower bound on the number of different subwords.

Proposition 4.7. *Let $w \in \Sigma^{>k}$, then $|\operatorname{alph}(w)| \leq |\operatorname{SubWords}_k(w)|$.*

Proof. Let $w \in \Sigma^{>k}$ and $\alpha \coloneqq |\operatorname{alph}(w)|$. We show that w has at least α subwords of length k.

Case 1 ($\alpha > k$). By choosing the first occurrence of each letter, we obtain a subword of w containing exactly α different letters. There exist at least $\binom{\alpha}{k} \geq \alpha$ different subwords of length k of this word because $0 < k < \alpha$.

Case 2 ($\alpha \leq k$). Since $|w| > k$, w has at least one subword of length $k+1$. Choose $\ell \in \mathbb{N}$ maximal such that there exist $\mathtt{a}_i \in \Sigma, \ell_i \in \mathbb{N}$ for all $i \in [\ell]_0$, such that $\mathtt{a}_{i-1} \neq \mathtt{a}_i$ for all $i \in [\ell]$ and

$$v \coloneqq \prod_{i=0}^{\ell} \mathtt{a}_i^{\ell_i} \in \operatorname{SubWords}_{k+1}(w).$$

Note that $\alpha \leq \ell + 1$ because v is long enough to contain all different letters of w. Now define for all $j \in [\ell]_0$

$$v_j \coloneqq \prod_{i=0}^{j-1} \mathtt{a}_i^{\ell_i} \cdot \mathtt{a}_j^{\ell_j - 1} \cdot \prod_{i=j+1}^{\ell} \mathtt{a}_i^{\ell_i} \preceq v \preceq w.$$

Suppose $v_x = v_y$ for some $x, y \in [\ell]_0$ with $x \neq y$ and assume without loss of generality that $x < y$. Then we have

$$\mathtt{a}_x^{\ell_x - 1} \cdot \prod_{i=x+1}^{y-1} \mathtt{a}_i^{\ell_i} \cdot \mathtt{a}_y^{\ell_y} = \mathtt{a}_x^{\ell_x} \cdot \prod_{i=x+1}^{y-1} \mathtt{a}_i^{\ell_i} \cdot \mathtt{a}_y^{\ell_y - 1}.$$

By the *xy-yz* lemma [22, Proposition 1.3.4], we obtain $\mathtt{a}_x = \mathtt{a}_{x+1} = \cdots = \mathtt{a}_y$, a contradiction. Thus, at least $|[\ell]_0| \geq |[\alpha - 1]_0| = \alpha$ subwords are distinct. □

4.2 General Results on $\alpha\beta$-Factorization

The following well-known result, shows that k-equivalent words are either both trivial or share the same number of arches [2]. Because the first case yields just a single class, we can focus on the second case. Therefore, when characterizing k-equivalence, we can always assume that two words share the same number of arches. In the following we therefore only consider words $w \in \Sigma^*$ with $\iota(w) < k$ with respect to $\sim_k$.

Lemma 4.8. *Let $w, w' \in \Sigma'$ such that $w \sim_k w'$ for some $k \in \mathbb{N}_0$. Then either both w, w' have k or more arches or they both have less than k and the same number of arches.*

- *If* $\mathrm{SubWords}_{\leq k}(w) = \mathrm{SubWords}_{\leq k}(w') = \Sigma^k$ *then* $\iota(w), \iota(w') \geq k$.
- *If* $\mathrm{SubWords}_{\leq k}(w) = \mathrm{SubWords}_{\leq k}(w') \subset \Sigma^k$ *then* $\iota(w) = \iota(w') \in [k-1]_0$.

Proof. For some word v, by Remark 2.5.1, $\iota(v)$ is the maximum number such that v has all subwords of this length. Therefore, if $\mathrm{SubWords}_{\leq k}(w) = \Sigma^k$ we have $\iota(w) \geq k$ and otherwise if $\mathrm{SubWords}_{\leq k}(w) \subset \Sigma^k$ we have $\iota(w) < k$. Since $w \sim_k w'$ is by definition equivalent to $\mathrm{SubWords}_{\leq k}(w) = \mathrm{SubWords}_{\leq k}(w')$, the claim follows. □

The following lemma uses a well-known technique and was shown in this form by Karandikar, Kufleitner, and Schnoebelen [20] using different notation. The main idea is to modify an occurrence of a subword, by distributing letters into arches, as each of them is guaranteed to contain each element of Σ at least once. We will use this lemma to break up more complicated constructions into smaller ones.

Lemma 4.9 ([20, Lemma 4.1])**.** *Let $w, w' \in \Sigma^*$ such that $w \sim_k w'$, then for all $u, v \in \Sigma^*$ we have $uwv \sim_{\iota(u)+k+\iota(v)} uw'v$.*

Proof. It suffices to prove that $uw \sim_{\iota(u)+k} uw'$, then an analogous claim for the right side follows by symmetry, and therefore, the claim by applying both.

Let $v \in \mathrm{SubWords}_{\leq \iota(u)+k}(uw)$. Now there exists a prefix v_u of maximal length of v such that $v_u \preceq u$. This prefix is at least of length $\iota(u)$ because $\Sigma^{\iota(u)} \subseteq \mathrm{SubWords}(u)$. Therefore, by the maximality of $|v_u|$, $v_w \coloneqq v_u^{-1}v$ is a subword of w and is at most of the length k. Thus, we have $v_w \preceq w \sim_k w'$, and therefore, $v = v_u v_w \preceq uw'$. □

Factors consisting of whole arches can be used in a twofold, quasi-inverse way. The first is given by the above lemma. If we have two k-equivalent words w, w', we can append the same word v to both. The resulting words $wv, w'v$ are not only k-equivalent (as given by $\sim_k$ being a congruence relation) but $k + \iota(v)$ equivalent. Note that we just used the universality of v, not the structure of its arches.

For the second way, consider two k-equivalent words w, w'. As they share all subwords of up to length k, they especially share the sets $\mathrm{m}(w') \cdot \mathrm{SubWords}_{\leq k-\iota(w')}(\mathrm{re}(w))$ and $\mathrm{m}(w) \cdot \mathrm{SubWords}_{\leq k-\iota(w')}(\mathrm{re}(w'))$. Therefore, we can conclude $\mathrm{re}(w) \sim_{k-\iota(w)} \mathrm{re}(w')$.

Using the $\alpha\beta$-factorization, we can extend the above idea from $\mathrm{re}(w) = \mathrm{A}_{\iota(w)}(w)$ to other $\mathrm{A}_i(w)$ as done in Corollary 4.10.1 and even further, two whole factors starting and ending with $\mathrm{A}_i(w)$ and $\mathrm{A}_j(w)$ for $i \leq j$ as done in Lemma 4.10 itself.

Lemma 4.10. *Let $w, w' \in \Sigma^*$ with $w \sim_k w'$ and $m = \iota(w) = \iota(w') < k$ and denote their $\alpha\beta$-factorizations by $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and $w' = \alpha'_0\beta'_1\alpha'_1 \cdots \beta'_m\alpha'_m$. Then, for all $i, j \in [m]_0$ with $i \leq j$ we have*

$$\alpha_i\beta_{i+1}\alpha_{i+1} \cdots \beta_j\alpha_j \sim_{k-m+(j-i)} \alpha'_i\beta'_{i+1}\alpha'_{i+1} \cdots \beta'_j\alpha'_j.$$

Proof. Let $w \sim_k w'$ and write $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and $w' = \alpha'_0\beta'_1\alpha'_1 \cdots \beta'_m\alpha'_m$. For $i \leq j$, define $v \coloneqq \alpha_i\beta_{i+1}\alpha_{i+1} \cdots \beta_j\alpha_j$ and $v' \coloneqq \alpha'_i\beta'_{i+1}\alpha'_{i+1} \cdots \beta'_j\alpha'_j$. We have

$$\begin{aligned}
\prod_{k=1}^{i} \mathrm{m}_k(w') \cdot \mathrm{SubWords}_{\leq k-m+(j-i)}(v) \cdot \prod_{k=j+1}^{m} \overleftarrow{\mathrm{m}}_k(w') &\subseteq \Sigma^i \cdot \mathrm{SubWords}_{\leq k-m+(j-i)}(v) \cdot \Sigma^{m-j} \\
&\subseteq \mathrm{SubWords}_{\leq i+(k-m+(j-i))+(m-j)}(w) \\
&= \mathrm{SubWords}_{\leq k}(w) \\
&= \mathrm{SubWords}_{\leq k}(w')
\end{aligned}$$

because $\alpha_0\beta_1 \cdots \alpha_{i-1}\beta_i$ is i-universal and $\beta_{j+1}\alpha_{j+1} \cdots \beta_m\alpha_m$ is $(m-j)$-universal. This implies $\mathrm{SubWords}_{\leq k-m+(j-i)}(v) \subseteq \mathrm{SubWords}_{\leq k-m+(j-i)}(v')$ by Remark 2.5.1. By symmetry, we obtain the other inclusion and thus $v \sim_{k-m+(j-i)} v'$. □

Corollary 4.10.1. *Let $w, w' \in \Sigma^*$ with $w \sim_k w'$ and $m = \iota(w) = \iota(w') < k$, then $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ for all $i \in [m]_0$.*

Proof. Follows directly from Lemma 4.10 for $i = j$. □

Corollary 4.10.2. *Let $w, w' \in \Sigma^*$ with $w \sim_k w'$ and $m \coloneqq \iota(w) = \iota(w') < k$, then $\mathrm{alph}(\mathrm{A}_i(w)) = \mathrm{alph}(\mathrm{A}_i(w'))$ for all $i \in [m]_0$. Furthermore, for all $i \in [m]$ we have $\mathrm{alph}(\mathrm{B}_i(w)) \supseteq \Sigma \setminus (\mathrm{alph}(\mathrm{A}_{i-1}(w)) \cap \mathrm{alph}(\mathrm{A}_i(w)))$.*

Proof. $\mathrm{A}_i(w)$ and $\mathrm{A}_i(w')$ are at least 1-equivalent for all $i \in [m]_0$. 1-equivalent words have exactly the same alphabet.

Write $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and let $i \in [m]$. We have $\mathrm{ar}_i(w) = \alpha_{i-1}\beta_i$ and $\overleftarrow{\mathrm{ar}}_i(w) = \beta_i\alpha_i$. Since $\Sigma = \mathrm{alph}(\mathrm{ar}_i(w)) = \mathrm{alph}(\overleftarrow{\mathrm{ar}}_i(w))$ the claim follows. □

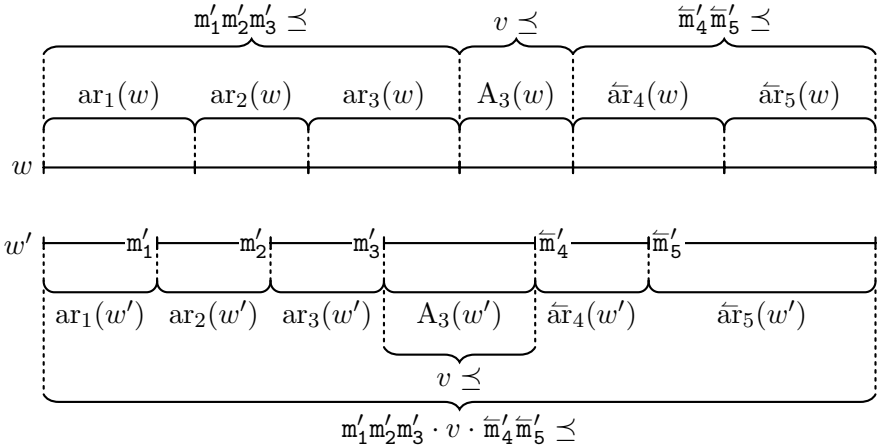


Figure 4.5: $(k-5)$-equivalence of A_3 in k-equivalent words where $\mathtt{m}'_\mathtt{i} \coloneqq \mathrm{m}_i(w')$, $\overleftarrow{\mathtt{m}}'_\mathtt{i} \coloneqq \overleftarrow{\mathrm{m}}_i(w')$ and $v \in \mathrm{SubWords}_{k-5}(\mathrm{A}_3(w))$.

Let $w, w' \in \Sigma^*$. If we have $\mathrm{B}(w) = \mathrm{B}(w')$, then we already obtain the converse of the above proposition. The main idea for the proof is the following. We consider a subword v of w. Letters inside $\mathrm{B}_i(w) = \mathrm{B}_i(w')$ can directly be moved to the other word. If the subword v of w uses more than $k-m$ letters of some $\mathrm{A}_i(w)$, then either to the left or right, it uses fewer letters than w has arches which precede or succeed $\mathrm{A}_i(w)$. Therefore, we can shift some letters outside $\mathrm{A}_i(w)$.

Proposition 4.11. *Let $w, w' \in \Sigma^*$ with $m \coloneqq \iota(w) = \iota(w') < k$ such that $\mathrm{B}(w) = \mathrm{B}(w')$, then $w \sim_k w'$, if and only if, $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ for all $i \in [m]_0$.*

Proof. By Corollary 4.10.1, we directly obtain one direction. Therefore, let $w, w' \in \Sigma^*$ such that $\mathrm{B}(w) = \mathrm{B}(w')$ and $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ for all $i \in [m]_0$. Write $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and $w' = \alpha'_0\beta_1\alpha'_1 \cdots \beta_m\alpha'_m$.

We show that $\mathrm{SubWords}_{\leq k}(w) \subseteq \mathrm{SubWords}_{\leq k}(w')$, then the claim follows by symmetry. Let $v \in \mathrm{SubWords}_{\leq k}(w)$. Then there exists a factorization $v = v_{\alpha_0}v_{\beta_1}v_{\alpha_1} \cdots v_{\beta_m}v_{\alpha_m}$ such that $v_{\alpha_i} \preceq \alpha_i$ and $v_{\beta_j} \preceq \beta_j$ for all $i \in [m]_0$ and $j \in [m]$. For simplicity, we proceed by induction on the value t assigned to the factorization of v satisfying the above properties, where

$$t \coloneqq \sum_{\substack{i=0 \\ |v_{\alpha_i}|>k-m}}^{m} |v_{\alpha_i}|.$$

If $t = 0$, then we have $v_{\alpha_i} \preceq \alpha_i \sim_{k-m} \alpha'_i$ for all $i \in [m]_0$ and thus $v = v_{\alpha_0}v_{\beta_1}v_{\alpha_1} \cdots v_{\beta_m}v_{\alpha_m} \preceq \alpha'_0\beta_1 \cdots \beta_m\alpha'_m = w'$. Therefore, assume $t > 0$. Thus, there exists some v_{α_i} with $|v_{\alpha_i}| > k-m$. Let

$$x \coloneqq \prod_{j=1}^{i} v_{\alpha_{j-1}}v_{\beta_j} \quad \text{and} \quad y \coloneqq \prod_{j=i+1}^{m} v_{\beta_j}v_{\alpha_j}.$$

We have $|xy| = |v| - |v_{\alpha_i}| < |v| - k + m \leq m$. Therefore, we have $|x| < i$ or $|y| < m - i$. By left-right symmetry, assume without loss of generality $|x| < i$. Factorize $v_{\alpha_i} = \operatorname{pref}_{i-|x|}(v_{\alpha_i}) \cdot \tilde{v}_{\alpha_i}$ for $\tilde{v}_{\alpha_i} \in \Sigma^{<|v_{\alpha_i}|}$. Now we have

$$x \cdot \operatorname{pref}_{i-|x|}(v_{\alpha_i}) \preceq \prod_{j=1}^{i} \alpha_{j-1}\beta_j \quad \text{and} \quad \tilde{v}_{\alpha_i} \preceq v_{\alpha_i} \preceq \alpha_i \quad \text{and} \quad y \preceq \prod_{j=i+1}^{m} \beta_j\alpha_j$$

by i-universality, transitivity and assumption respectively.

We can factorize $x \cdot \operatorname{pref}_{i-|x|}(v_{\alpha_i})$ such that $(x \cdot \operatorname{pref}_{i-|x|}(v_{\alpha_i}))[j] \preceq \alpha_{j-1}\beta_j$ for all $j \in [i]$. Define $\tilde{v}_{\alpha_{j-1}}, \tilde{v}_{\beta_j}$ such that $\tilde{v}_{\alpha_{j-1}}\tilde{v}_{\beta_j} = (x \cdot \operatorname{pref}_{i-|x|}(v_{\alpha_i}))[j]$ and $\tilde{v}_{\alpha_{j-1}} \preceq \alpha_{j-1}$ and $\tilde{v}_{\beta_j} \preceq \beta_j$ for all $j \in [i-1]$. Define $\tilde{v}_{\alpha_j} \coloneqq v_{\alpha_j}$ and $\tilde{v}_{\beta_j} \coloneqq v_{\beta_j}$ for $i < j \leq m$. Then $v = \tilde{v}_{\alpha_0}\tilde{v}_{\beta_1} \cdots \tilde{v}_{\beta_m}\tilde{v}_{\alpha_m}$ such that $\tilde{v}_{\alpha_j} \preceq \alpha_j$ and $\tilde{v}_{\beta_k} \preceq \beta_k$ for all $j \in [m]_0$ and $k \in [m]$. Furthermore, for this factorization holds

$$\begin{aligned}
\sum_{\substack{j=0 \\ |\tilde{v}_{\alpha_j}|>k-m}}^{m} |\tilde{v}_{\alpha_j}| &\leq \sum_{\substack{j=0 \\ |\tilde{v}_{\alpha_j}|>k-m}}^{i-1} |\tilde{v}_{\alpha_j}| + |\tilde{v}_{\alpha_i}| + \sum_{\substack{j=i+1 \\ |\tilde{v}_{\alpha_j}|>k-m}}^{m} |\tilde{v}_{\alpha_j}| && \text{(Inequality if } |\tilde{v}_{\alpha_i}| \leq k-m) \\
&= 0 + |\tilde{v}_{\alpha_i}| + \sum_{\substack{j=i+1 \\ |v_{\alpha_j}|>k-m}}^{m} |v_{\alpha_j}| && (|\tilde{v}_{\alpha_j}| \leq 1 \text{ for } j \in [i-1]_0) \\
&< \sum_{\substack{j=0 \\ |v_{\alpha_j}|>k-m}}^{i-1} |v_{\alpha_j}| + |v_{\alpha_i}| + \sum_{\substack{j=i+1 \\ |v_{\alpha_j}|>k-m}}^{m} |v_{\alpha_j}| = t. && (|\tilde{v}_{\alpha_i}| < |v_{\alpha_i}|)
\end{aligned}$$

Therefore, the claim follows by induction. □

Using a similar idea as Karandikar, Kufleitner, and Schnoebelen [20, Lemma 4.2], we can give a much shorter proof of the proposition. In contrast to Karandikar, Kufleitner, and Schnoebelen [20], we choose different factors, allowing us to give a characterization of the equivalence. In contrast to the above proof, we use the transitivity of $\sim_k$. Thus, the construction of the occurrence of a subword $v \preceq w$ in w' is not as explicit.

Note that exchanging A_i with different words with the same alphabet does not change the $\alpha\beta$-factorization because the unique last letter of the (reverse) arches does not change. This fact will be used in the next alternative proof of Proposition 4.11 and the proof of Corollary 4.11.1.

Proof. By Corollary 4.10.1, we directly obtain one direction. Therefore, let $w, w' \in \Sigma^*$ such that $\mathrm{B}(w) = \mathrm{B}(w')$ and $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ for all $i \in [m]_0$. Write $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and $w' = \alpha'_0\beta_1\alpha'_1 \cdots \beta_m\alpha'_m$. We obtain by Lemma 4.9 that

$$\begin{aligned}
\alpha_0\beta_1\alpha_1\beta_2\alpha_2 \cdots \beta_m\alpha_m &\sim_k \alpha'_0\beta_1\alpha_1\beta_2\alpha_2 \cdots \beta_m\alpha_m \\
&\sim_k \alpha'_0\beta_1\alpha'_1\beta_2\alpha_2 \cdots \beta_m\alpha_m \\
&\cdots \\
&\sim_k \alpha'_0\beta_1\alpha'_1\beta_2\alpha'_2 \cdots \beta_m\alpha'_m.
\end{aligned}$$

□

Corollary 4.11.1. *Let $w, w' \in \Sigma^*$ with $m \coloneqq \iota(w) = \iota(w') < k$, then $w \sim_k w'$, if and only if, $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ for all $i \in [m]_0$ and for $\tilde{w} \coloneqq \mathrm{A}_0(w)\,\mathrm{B}_1(w')\,\mathrm{A}_1(w) \cdots \mathrm{B}_m(w')\,\mathrm{A}_m(w)$ we have $w \sim_k \tilde{w}$.*

Proof. Note that, the $\alpha\beta$-factorization of $\tilde{w}$ is as in the definition because the exchanged A_i are equivalent. Assume $w \sim_k w'$. By Corollary 4.10.1 we have $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$. Define $\tilde{w}$ as above. We have $w' \sim_k \tilde{w}$ by Proposition 4.11 and therefore, by transitivity $w \sim_k \tilde{w}$.

Now assume the converse. By $\mathrm{A}_i(\tilde{w}) = \mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ and $\mathrm{B}(\tilde{w}) = \mathrm{B}(w')$ we obtain by Proposition 4.11 that $\tilde{w} \sim_k w'$. By the assumption and transitivity we obtain $w \sim_k w'$. □

Example 4.12. We have $\mathtt{bcbc} \cdot \mathtt{bca} \cdot \mathtt{a} \sim_3 \mathtt{bcbc} \cdot \mathtt{cba} \cdot \mathtt{a}$ but $\mathtt{bca} \neq \mathtt{cba}$. Note that in this example, we have equivalence because $\mathrm{A}_0(w)$ and $\mathrm{A}_0(w')$ have enough arches with respect to $\{\mathtt{b}, \mathtt{c}\}$.

The above example shows, that the extra assumption is too strong, that is, there exist equivalent words not satisfying it. We can use Corollary 4.10.1 and Proposition 4.11 already to obtain a partial normal form, by choosing unique representatives for the A_i. This is given by Corollary 4.11.1. Then, equal B_i are a sufficient but not necessary condition for equivalence.

4.3 A Characterization of Classes for Binary Alphabets

In this section, we apply the results to the special case of binary words. In the following, denote the elements of Σ_2 by $\mathtt{a}$ and $\mathtt{b}$. Furthermore, recall that for $\mathtt{x} \in \Sigma_2$, we denote the unique second letter in Σ_2 by $\overline{\mathtt{x}}$.

Proposition 4.13. *Let $w \in \Sigma_2$ and $m = \iota(w)$, then*

(1) for all $i \in [m]$, we have $\mathrm{B}_i(w) \in \{\mathtt{a}, \mathtt{b}, \mathtt{ab}, \mathtt{ba}\}$,

(2) for all $i \in [m]$ if $\mathrm{B}_i(w) = \mathtt{x}$, then $\mathrm{A}_{i-1}(w), \mathrm{A}_i(w) \in \overline{\mathtt{x}}^+$,

(3) for all $i \in [m]$ if $\mathrm{B}_i(w) = \mathtt{x}\overline{\mathtt{x}}$, then $\mathrm{A}_{i-1}(w) \in \mathtt{x}^*$ and $\mathrm{A}_i(w) \in \overline{\mathtt{x}}^*$.

Proof. Let $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$. By Remark 4.1.1, the first and last letter of β_i are unique. Therefore, if $\beta_i[1] = \beta_i[|\beta_i|]$ we have $|\beta_i| = 1$. Furthermore, if $\beta_i[1] \neq \beta_i[|\beta_i|]$ we have $|\beta_i| = 2$ because $|\Sigma_2| = 2$.

By Remark 4.1.1 the α_i are unary words for all $i \in [m]_0$. By symmetry, we only have to show the claim for α_{i-1}. The restrictions on the alphabet of α_{i-1} follow directly from $\mathrm{ar}_i(w) = \alpha_{i-1}\beta_i$ and the uniqueness of the last letter. Furthermore, if $\beta_i = \mathtt{x} \in \Sigma_2$ then $\alpha_{i-1} \neq \varepsilon$ because $\mathtt{x}, \overline{\mathtt{x}} \preceq \mathrm{ar}_i(w) = \alpha_{i-1}\beta_i$. □

Remark 4.13.1. Because $\mathrm{ar}_i(w) = \mathrm{A}_{i-1}(w)\,\mathrm{B}_i(w)$, we have $\mathrm{alph}(\mathrm{A}_{i-1}(w)\,\mathrm{B}_i(w)) = \Sigma_2$. This implies the different conditions on the equivalence class of $\mathrm{A}_{i-1}(w)$ based on the choice of $\mathrm{B}_i(w)$.

The following lemma holds only for binary words. It will be the main result, specific to binary words used for the characterization.

Lemma 4.14. *Let $w, w' \in \Sigma_2^*$ with $w \sim_k w'$ and $\iota(w) = \iota(w') < k$, then $\mathrm{m}(w) = \mathrm{m}(w')$.*

Proof. The claim obviously holds for $\iota(w) = \iota(w') = 0$. Assume $\iota(w) = \iota(w') = m + 1$ and $w \sim_{m+2} w'$. It suffices to show that the first modi coincide, then $(\alpha_0\beta_1)^{-1}w \sim_{m+1} (\alpha'_0\beta'_1)^{-1}w'$ and the claim follows by induction. Let

$$w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m\beta_{m+1}\alpha_{m+1} \qquad \text{and} \qquad w' = \alpha'_0\beta'_1\alpha'_1 \cdots \beta'_m\alpha'_m\beta'_{m+1}\alpha'_{m+1}.$$

Because $w \sim_k w'$, we have $\mathrm{alph}(\alpha_0) = \mathrm{alph}(\alpha'_0)$ by Corollary 4.10.2. If $\mathrm{alph}(\alpha_0) \neq \emptyset$, then the arches $\alpha_0\beta_1$ and $\alpha'_0\beta'_1$ start with the same letter and thus their modi coincide. Therefore, assume $\alpha_0 = \alpha'_0 = \varepsilon$, and suppose $\beta_1 = \mathtt{y}\overline{\mathtt{y}}$ and $\beta'_1 = \overline{\mathtt{y}}\mathtt{y}$ by Proposition 4.13. Again, by Corollary 4.10.2, and the construction of the reverse arches, we have

$$\overleftarrow{\mathrm{m}}_1(w) = \mathtt{y} \notin \mathrm{alph}(\alpha_1) = \mathrm{alph}(\alpha'_1) \not\ni \overline{\mathtt{y}} = \overleftarrow{\mathrm{m}}_1(w')$$

and thus $\alpha_1 = \alpha'_1 = \varepsilon$. Let $v \coloneqq \overline{\mathtt{y}}\mathtt{y} \cdot \overleftarrow{\mathrm{m}}_2(w) \cdots \overleftarrow{\mathrm{m}}_m(w)\,\overleftarrow{\mathrm{m}}_{m+1}(w)$. Then, we have, $v \not\preceq w$ because $\overline{\mathtt{y}}\mathtt{y} \not\preceq \alpha_0\beta_1\alpha_1$, but $v \preceq w'$ because $\overline{\mathtt{y}}\mathtt{y} \preceq \mathrm{ar}_1(w') = \alpha'_0\beta'_1$, a contradiction. □

Corollary 4.14.1. *Let $w, w' \in \Sigma_2^*$ with $w \sim_k w'$ and $\iota(w) = \iota(w') < k$, then $\mathrm{B}(w) = \mathrm{B}(w')$.*

Proof. By Lemma 4.14 we have $\mathrm{m}(w) = \mathrm{m}(w')$ and $\overleftarrow{\mathrm{m}}(w) = \mathrm{m}(w^R) = \mathrm{m}(w'^R) = \overleftarrow{\mathrm{m}}(w')$. By Proposition 4.13 these already determine $\mathrm{B}(w)$ and $\mathrm{B}(w')$ completely. □

Theorem 4.15 (Binary Characterization)**.** *Let $w, w' \in \Sigma_2^*$ such that $m \coloneqq \iota(w) = \iota(w') < k$, then $w \sim_k w'$, if and only if, $\mathrm{B}(w) = \mathrm{B}(w')$ and $\mathrm{A}_i(w) \sim_{k-m} \mathrm{A}_i(w')$ for all $i \in [m]_0$.*

Proof. Follows directly from Corollaries 4.14.1 and 4.10.1 and Proposition 4.11. □

The above characterization explains the earlier Example 2.10 which was a singleton class that was not captured by Lemma 2.9. We can immediately conclude that $[w]_{\sim_k}$ for $w \in \Sigma_2^*$ is a singleton, if and only if, $\iota(w) < k$ and $[\mathrm{A}_i(w)]_{\sim_{k-\iota(w)}}$ is a singleton for all $i \in [\iota(w)]_0$.

Using the characterization, we can also give a linear time algorithm for finding the largest k such that $u \sim_k v$ for $u, v \in \Sigma_2^*$. This special case was originally solved by Hébrard [15] with an algorithm considering just arches. Furthermore, a linear time algorithm for arbitrary alphabets was recently presented by Gawrychowski et al. [12]. Nonetheless, we give Algorithm 1, as it is a conceptually simple algorithm. Furthermore, it exploits that $\mathrm{A}_i(w)$ factors can be treated similar to $\mathrm{re}(w)$ in the arch factorization.

Algorithm 1: MaxSimK for binary words

Input: $u, v \in \Sigma_2^*$
Result: if $u = v$ then ∞ and otherwise the maximum k such that $u \sim_k v$

```
(α_0, β_1, ..., α_{ι(u)}) := αβ-Fact(u);                    // w.r.t. Σ_2
(α'_0, β'_1, ..., α'_{ι(v)}) := αβ-Fact(v);
if ι(u) ≠ ι(v) ∨ alph(u) ≠ alph(v) then    // 2nd condition for u = x^i, v = x̄^j
    return min(ι(u), ι(v));
else if β_1 = β'_1 ∧ ··· ∧ β_{ι(u)} = β'_{ι(v)} then
    for i ∈ [ι(u)]_0 do                          // solve MaxSimK for unary α pairs
        e_i := if |α_i| = |α'_i| then ∞ else min(|α_i|, |α'_i|);
    return ι(u) + min{e_i | i ∈ [ι(u)]_0};
else
    return ι(u);
```

4.3.1 Counting Classes in the Binary Case

In the following, we will use Theorem 4.15 to derive a formula for the concrete value of $|\Sigma_2^*/\sim_k|$. The first values for the index of Simon's congruence for different alphabets are given in Table 4.1. We will derive a formula for the first column. Note that in the unary case, the empty word has its own class. Then the formula for the first column is obvious.

By Lemma 4.8 we know that there exists exactly one class of words with k arches and that we can consider the other classes separated by the number arches. By Theorem 4.15, we can count classes based on the valid combinations of B and number of classes for each A. Because the A are unary, we already know the number of classes. The valid combinations are exactly given by Proposition 4.13. The first values for the number of classes separated by number of arches are given in Table 4.2.

Theorem 4.16. *The number of congruence classes of $\Sigma_2^*/\sim_k$ of words with $m < k$ arches (the entries of Table 4.2) is given by*

$$\left\| \begin{pmatrix} k-m & k-m & k-m \\ 1 & 2 & 1 \\ k-m & k-m & k-m \end{pmatrix}^m \cdot \begin{pmatrix} k-m \\ 1 \\ k-m \end{pmatrix} \right\|_1 = c_k^m$$

where $c_k^{-1} := 1$, $c_k^0 := 2k+1$ and $c_k^m := 2 \cdot (k-m+1) \cdot c_{k-1}^{m-1} - 2 \cdot (k-m) \cdot c_{k-2}^{m-2}$.

Proof. First, we show that the matrix representation produces the correct values, then we show the characterization as recurrence. Note that $k - m$ is fixed on the diagonals of Table 4.2. Therefore, increasing both, increases just the exponent of the matrix. We show that the first column is correct and then proceed by induction along the diagonals. Denote that above matrix by $D_{k,m}$.

Let $k \in \mathbb{N}_0$ and $w \in \Sigma^*$ with $m := \iota(w) < k$. For $i \in [m]_0$, all elements of $v \in [w]_{\sim_k}$ have $k - m$ equivalent $\mathrm{A}_i(v)$. By Remark 4.1.1, their alphabets are proper subsets of Σ_2. Therefore, they are either empty or non-empty unary words consisting of just `a` or

Length of Subwords	Number of Letters								
	1	2	3	4	5	6	7	8	σ
0	1	1	1	1	1	1	1	1	1
1	2	4	8	16	32	64	128	256	2^σ
2	3	16	152						
3	4	68	5 312						
4	5	312	334 202						
5	6	1 560	38 450 477						
6	7	8 528	$\geq 39 \cdot 10^7$						
7	8	50 864							
8	9	329 248							
9	10	2 298 592							
10	11	17 203 264							
11	12	137 289 920							
k	$k+1$	Corollary 4.16.3							

Table 4.1: Index of Simon's Congruence as computed by Karandikar, Kufleitner, and Schnoebelen [20] for different alphabet sizes and subword lengths.

b. We separate the choice of A_i into these three cases. Let $M_\varepsilon^\ell \coloneqq \{[w] \in \Sigma_2^*/{\sim_{(k-m)+\ell}} \mid \iota(w) = \ell, \mathrm{A}_0(w) \sim_{k-m} \varepsilon\}$ and $M_{\mathtt{x}}^\ell \coloneqq \{[w] \in \Sigma_2^*/{\sim_{(k-m)+\ell}} \mid \iota(w) = \ell, \mathrm{A}_0(w) \sim_{k-m} \mathtt{x}^+\}$ for $\mathtt{x} \in \Sigma_2$ be sets of $\ell + (k-m)$ equivalence classes of words with ℓ arches, separated by the alphabet of A_0. Denote by $e_{k,m} \coloneqq (|M_{\mathtt{a}}^0|, |M_\varepsilon^0|, |M_{\mathtt{b}}^0|)^\intercal = (k-m, 1, k-m)^\intercal$ the number of classes for zero arches. We show $\|D_{k,m}^\ell \cdot e_{k,m}\|_1 = (|M_{\mathtt{a}}^\ell|, |M_\varepsilon^\ell|, |M_{\mathtt{b}}^\ell|)^\intercal$. There are four choices for B_i which are given by Proposition 4.13. Each choice of B_{i+1} depends on the preceding A_i and limits the choices for the succeeding A_{i+1}. These are given by Proposition 4.13, and correspond to the entries of the matrix because for $\ell \geq 1$ we have

$$\begin{aligned}
M_\varepsilon^\ell &= \{[w]_{\sim_{(k-m)+\ell}} \in M_\varepsilon^\ell \mid \mathtt{x} \in \Sigma_2, \mathrm{B}_1(w) = \overline{\mathtt{x}}\mathtt{x}, \mathrm{A}_1(w) \sim_{k-m} \varepsilon\} \\
&\qquad \sqcup \{[w]_{\sim_{(k-m)+\ell}} \in M_\varepsilon^\ell \mid \mathtt{x} \in \Sigma_2, \mathrm{B}_1(w) = \overline{\mathtt{x}}\mathtt{x}, \mathrm{A}_1(w) \sim_{k-m} \mathtt{x}^+\} \\
&\cong \{\mathtt{ab}, \mathtt{ba}\} \times M_\varepsilon^{\ell-1} \sqcup \{\mathtt{ab}\} \times M_{\mathtt{b}}^{\ell-1} \sqcup \{\mathtt{ba}\} \times M_{\mathtt{a}}^{\ell-1} \\
M_{\mathtt{x}}^\ell &= \{[w]_{\sim_{(k-m)+\ell}} \in M_{\mathtt{x}}^\ell \mid \mathrm{B}_1(w) = \overline{\mathtt{x}}\} \\
&\qquad \sqcup \{[w]_{\sim_{(k-m)+\ell}} \in M_{\mathtt{x}}^\ell \mid \mathrm{B}_1(w) = \mathtt{x}\overline{\mathtt{x}}, \mathrm{A}_1(w) \sim_{k-m} \overline{\mathtt{x}}^+\} \\
&\qquad \sqcup \{[w]_{\sim_{(k-m)+\ell}} \in M_{\mathtt{x}}^\ell \mid \mathrm{B}_1(w) = \mathtt{x}\overline{\mathtt{x}}, \mathrm{A}_1(w) \sim_{k-m} \varepsilon\} \\
&\cong [k-m] \times \big(\{\overline{\mathtt{x}}\} \times M_{\mathtt{x}}^{\ell-1} \sqcup \{\mathtt{x}\overline{\mathtt{x}}\} \times M_{\overline{\mathtt{x}}}^{\ell-1} \sqcup \{\mathtt{x}\overline{\mathtt{x}}\} \times M_\varepsilon^{\ell-1}\big).
\end{aligned}$$

Therefore, each multiplication with the matrix increases the number of arches ℓ by one. Thus, for $m = \ell$ we have the desired value as M_ε^m and $M_{\mathtt{x}}^m$ are sets of k equivalence classes with m arches. Therefore, $\|D_{k,m}^m \cdot e_{k,m}\|_1$ corresponds to the number of classes with respect to $\sim_k$ of words with m arches.

		Number of Arches								
		0	1	2	3	4	5	6	7	m
Subword length	1	3	1							
	2	5	10	1						
	3	7	26	34	1					
	4	9	50	136	116	1				
	5	11	82	358	712	396	1			
	6	13	122	748	2 564	3 728	1 352	1		
	7	15	170	1 354	6 824	18 364	19 520	4 616	1	
	k	$2k+1$								

Table 4.2: Index of Simon's congruence restricted to binary words with a fixed number of arches

The equivalence of the two formulas is left to show. The characteristic polynomial of $D_{k,m}$ is given by

$$\chi_{D_{k,m}} = \det(D_{k,m} - \lambda I) = -\lambda^3 + 2\lambda^2 + 2(k-m)\lambda^2 - 2(k-m)\lambda.$$

By the Cayley-Hamilton theorem, $D_{k,m}$ is a root of its characteristic polynomial and thus satisfies the recurrence

$$\begin{aligned} D_{k,m}^{\ell+2} &= 2 \cdot D_{k,m}^{\ell+1} + 2 \cdot (k-m) \cdot D_{k,m}^{\ell+1} - 2 \cdot (k-m) \cdot D_{k,m}^{\ell} \\ &= 2 \cdot (k-m+1) \cdot D_{k,m}^{\ell+1} - 2 \cdot (k-m) \cdot D_{k,m}^{\ell} \end{aligned}$$

for $\ell \in \mathbb{N}$. Note that $e_{k,m} = e_{k+\ell,m+\ell}$ for all $\ell \in \mathbb{N}_0$. Now we conclude by induction that

$$\begin{aligned} \left\| D_{k+2,m+2}^{m+2} \cdot e_{k+2,m+2} \right\|_1 &= \left\| D_{k,m}^{m+2} \cdot e_{k,m} \right\|_1 \\ &= \left\| (2 \cdot (k-m+1) \cdot D_{k,m}^{m+1} - 2 \cdot (k-m) \cdot D_{k,m}^{m}) \cdot e_{k,m} \right\|_1 \\ &= 2 \cdot (k-m+1) \cdot \left\| D_{k,m}^{m+1} \cdot e_{k,m} \right\|_1 - 2 \cdot (k-m) \cdot \left\| D_{k,m}^{m} \cdot e_{k,m} \right\|_1 \\ &= 2 \cdot (k-m+1) \cdot c_{k+1}^{m+1} - 2 \cdot (k-m) \cdot c_k^m \\ &= c_{k+2}^{m+2}, \end{aligned}$$

because $\|u \pm v\|_1 = \|u\|_1 \pm \|v\|_1$ for all $u = (u_i), v = (v_i) \in \mathbb{R}^n$ for which $u_j v_j \geq 0$ for all $j \in [n]$. □

Remark 4.16.1. Note that by setting $\Delta \coloneqq k - m$, the family of recurrences depends only on one variable Δ, because $k - m = (k - \ell) - (m - \ell)$.

Remark 4.16.2. The proof of Theorem 4.16 can be seen as a generalized counting problem for domino sequences of length m. Each $\mathrm{A}_{i-1}(w)\,\mathrm{B}_i(w)\,\mathrm{A}_i(w)$ corresponds to one domino piece and class of words with m arches corresponds, by Theorem 4.15, to a sequence of length m. This idea reappears later in a generalized form in Corollary 4.17.1.

		Number of Arches								
		0	1	2	3	4	5	6	7	8	m
Subword length	2	1	4	1							
	3	1	6	14	1						
	4	1	8	32	48	1					
	5	1	10	58	168	164	1				
	6	1	12	92	416	880	560	1			
	7	1	14	134	840	2 980	4 608	1 912	1		
	8	1	16	184	1 488	7 664	21 344	24 344	6 528	1	
	k	1	$2k$								

Table 4.3: Number of classes of perfect universal binary words restricted to words with a fixed number of arches

Corollary 4.16.3. *Let $k \in \mathbb{N}_0$. Over a binary alphabet, the number of congruence classes of $\sim_k$ is given by*

$$|\Sigma_2^*/\sim_k| = 1 + \sum_{m=0}^{k-1} c_k^m.$$

Proof. By Lemma 4.8, we can count the number of classes separated by the number of arches for words with less than k arches. □

Some sequences in Table 4.2 are known sequences. The first and second diagonal are A007052 and A018903 respectively in *The On-Line Encyclopedia of Integer Sequences* [24]. In fact, both sequences appear in the same context, as number of compositions of n when there exist three and five different ones respectively [18]. Furthermore, the sequences c_k^m seem to be equivalent to the family of sequences (s_n) where $s_0 = 1$ and s_1 is fixed and s_{n+2} is the smallest number such that $\frac{s_{n+2}}{s_{n+1}} > \frac{s_{n+1}}{s_n}$. These sequences where studied by Boyd [4].

We can use the idea of Theorem 4.16 to also count the number of perfect universal words with a certain universality for some k. Since a perfect universal word has no rest, we can count them by replacing the vector with the initial distribution of A_i values with $(0, 1, 0)^\intercal$. Thus, the formula counts words starting or ending with an empty A. Because the matrix does not change, we obtain the same recurrence with different initial values. The k^{th} diagonal is now given by the Lucas sequence of the first kind $U(2 \cdot k + 2, 2 \cdot k)$, where $U_n(P, Q)$ is given by

$$U_0(P, Q) = 0, \qquad U_1(P, Q) = 1, \qquad U_n(P, Q) = P \cdot U_{n-1}(P, Q) - Q \cdot U_{n-2}(P, Q).$$

The first calculated values are given in Table 4.3. The first three diagonals of the table are known integer sequences, namely A007070, A084326, and A190978 in *The On-Line Encyclopedia of Integer Sequences* [24].

4.4 Decomposition into $\alpha\beta\alpha$-Factors

In this section, we prove a general result about the structure of k-universal words for $k \geq 2$. Corollary 4.10.1 does only yield an equivalence of A_i. In Proposition 4.11 we added equality of B_i as an extra assumption to obtain a characterization for this special case. By applying Lemma 4.10 for $i+1=j$, we obtain an equivalence for the shortest factors containing the B_i. The next proposition shows that by requiring equivalence of these larger factors, we can give a general characterization of k-equivalence in terms of these longer factors. This is in contrast to Lemma 4.10.

The idea of the proof is as follows. If any of the equivalent factors is overfilled by an occurrence of a subword, then we can shift letters into the arches not contained in the factor, and use the equivalences to transport it over. Otherwise, we show that we can shift letters into the overlapping A_i by using the equivalences of the two overlapping factors and the equivalence of the A_i, reducing the problem to the first case.

Proposition 4.17. *Let $w, w' \in \Sigma^*$ with $m = \iota(w) = \iota(w') < k$ and denote their $\alpha\beta$-factorizations by $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and $w' = \alpha'_0\beta'_1\alpha'_1 \cdots \beta'_m\alpha'_m$. Then $w \sim_k w'$, if and only if, $\alpha_0\beta_1\alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1\alpha'_1$ and $\alpha_1\beta_2 \cdots \beta_m\alpha_m \sim_{k-1} \alpha'_1\beta'_2 \cdots \beta'_m\alpha'_m$.*

Proof. Assume $w \sim_k w'$. By two applications of Lemma 4.10 for $i = 0, j = 1$ and $i = 1, j = m$ we obtain the equivalences.

Assume $\alpha_0\beta_1\alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1\alpha'_1$ and $\alpha_1\beta_2\alpha_2 \cdots \beta_m\alpha_m \sim_{k-1} \alpha'_1\beta'_2\alpha'_2 \cdots \beta'_m\alpha'_m$. We show $\mathrm{SubWords}_{\leq k}(w) \subseteq \mathrm{SubWords}_{\leq k}(w')$, then the claim follows by symmetry. Let $v \in \mathrm{SubWords}_{\leq k}(w)$. Then there exists a factorization $v = v_{\alpha\beta}v_\alpha v_{\beta\alpha}$ such that $v_{\alpha\beta} \preceq \alpha_0\beta_1$, $v_\alpha \preceq \alpha_1$ and $v_{\beta\alpha} \preceq \beta_2\alpha_2 \cdots \beta_m\alpha_m$ such that $|v_\alpha|$ is *maximal*. By Corollary 4.10.1, we obtain that $\alpha_i \sim_{k-m} \alpha'_i$ for all $i \in [m]_0$.

Case 1 ($|v_\alpha v_{\beta\alpha}| \geq k-1$). Define

$$u_1 \coloneqq v_{\alpha\beta} \cdot \mathrm{pref}_{|v_\alpha v_{\beta\alpha}|-(k-1)}(v_\alpha v_{\beta\alpha}) \qquad \text{and} \qquad u_2 \coloneqq \mathrm{suff}_{k-1}(v_\alpha v_{\beta\alpha}).$$

We have $|u_1| = |v_{\alpha\beta}| + |v_\alpha v_{\beta\alpha}| - (k-1) \leq 1$. Therefore, by the 1-universality of $\alpha'_0\beta'_1$ we have $u_1 \preceq \alpha'_0\beta'_1$. Furthermore, by $u_2 \preceq v_\alpha v_{\beta\alpha} \preceq \alpha_1 \cdot \prod_{i=2}^m \beta_i\alpha_i$ we have

$$u_2 \preceq \alpha_1 \cdot \prod_{i=2}^m \beta_i\alpha_i \sim_{k-1} \alpha'_1 \cdot \prod_{i=2}^m \beta'_i\alpha'_i,$$

and thus $v = u_1 \cdot u_2 \preceq \alpha'_0\beta'_1 \cdot \alpha'_1 \prod_{i=2}^m \beta'_i\alpha'_i = w'$.

Case 2 ($|v_{\alpha\beta}v_\alpha| \geq k-(m-1)$). Define

$$u_1 \coloneqq \mathrm{pref}_{k-(m-1)}(v_{\alpha\beta}v_\alpha) \qquad \text{and} \qquad u_2 \coloneqq \mathrm{suff}_{|v_{\alpha\beta}v_\alpha|-(k-(m-1))}(v_{\alpha\beta}v_\alpha) \cdot v_{\beta\alpha}.$$

We have $|u_2| = |v_{\alpha\beta}v_\alpha| - (k-(m-1)) + |v_{\beta\alpha}| \leq m-1$. Therefore, by the $(m-1)$-universality of $\prod_{i=2}^m \beta'_i\alpha'_i$ we have $u_2 \preceq \prod_{i=2}^m \beta'_i\alpha'_i$. Furthermore, by $u_1 \preceq v_{\alpha\beta}v_\alpha \preceq \alpha_0\beta_1 \cdot \alpha_1$ we have

$$u_1 \preceq \alpha_0\beta_1\alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1\alpha'_1,$$

and thus $v = u_1 \cdot u_2 \preceq \alpha'_0\beta'_1\alpha'_1 \cdot \prod_{i=2}^m \beta'_i\alpha'_i = w'$.

Case 3 ($|v_\alpha| \geq k-m$). If $|v_\alpha| > k-m$ or $|v_{\alpha\beta}| > 0$, then $|v_{\alpha\beta}v_\alpha| \geq k-m+1 = k-(m-1)$ and we are in Case 2. Therefore, assume $|v_{\alpha\beta}| = 0$ and $|v_\alpha| = k-m$. If $|v_{\beta\alpha}| = m$, then $|v_\alpha v_{\beta\alpha}| = k$ and we are in Case 1. Therefore, assume $|v_{\beta\alpha}| \leq m-1$. In this case we have $v_{\beta\alpha} \preceq \prod_{i=2}^{m} \beta'_i\alpha'_i$ by the $(m-1)$-universality of $\prod_{i=2}^{m} \beta'_i\alpha'_i$ and

$$v_\alpha = v_{\alpha\beta} \cdot v_\alpha \preceq \alpha_0\beta_1 \cdot \alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1 \cdot \alpha'_1.$$

Therefore, $v = v_\alpha \cdot v_{\beta\alpha} \preceq \alpha'_0\beta'_1\alpha'_1 \cdot \prod_{i=2}^{m} \beta'_i\alpha'_i$.

Case 4 ($|v_{\alpha\beta}v_\alpha| < k-(m-1)$ and $|v_\alpha v_{\beta\alpha}| < k-1$ and $|v_\alpha| < k-m$). We have

$$v_\alpha \cdot v_{\beta\alpha} \preceq \alpha_1 \cdot \prod_{i=2}^{m} \beta_m\alpha_m \sim_{k-1} \alpha'_1 \cdot \prod_{i=2}^{m} \beta'_m\alpha'_m,$$

by definition of $v_\alpha, v_{\beta\alpha}$ and the length assumption of Case 4. If $v_{\alpha\beta} \preceq \alpha'_0\beta'_1$ the claim follows. Therefore, assume $v_{\alpha\beta} \not\preceq \alpha'_0\beta'_1$. Again, by definition of $v_{\alpha\beta}, v_\alpha$ and the length assumption of Case 4, we have

$$v_{\alpha\beta} \cdot v_\alpha \preceq \alpha_0\beta_1 \cdot \alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1\alpha'_1.$$

Therefore, there exists a factorization $v_{\alpha\beta} = u_1u_2$ with $u_1 \in \Sigma^*$ and $u_2 \in \Sigma^+$, such that, $u_1 \preceq \alpha'_0\beta'_1$ and $u_2v_\alpha \preceq \alpha'_1$. Therefore, define

$$\hat{v}_{\alpha\beta} \coloneqq v_{\alpha\beta}[1..|v_{\alpha\beta}|-1] \preceq v_{\alpha\beta} \preceq \alpha_0\beta_1, \qquad \hat{v}_{\beta\alpha} \coloneqq v_{\beta\alpha} \preceq \beta_2\alpha_2 \cdots \beta_m\alpha_m,$$

and $\hat{v}_\alpha \coloneqq v_{\alpha\beta}[|v_{\alpha\beta}|] \cdot v_\alpha \preceq u_2 \cdot v_\alpha \preceq \alpha'_1$. By the length assumption from Case 4, we have $\hat{v}_\alpha = v_{\alpha\beta}[|v_{\alpha\beta}|] \cdot v_\alpha \preceq \alpha'_1 \sim_{k-m} \alpha_1$. A contradiction against the maximality of $|v_\alpha|$ since $\hat{v}_{\alpha\beta} \preceq v_{\alpha\beta} \preceq \alpha_0\beta_1$, $\hat{v}_{\beta\alpha} = v_{\beta\alpha} \preceq \prod_{i=2}^{m} \beta_i\alpha_i$ and

$$\hat{v}_{\alpha\beta} \cdot \hat{v}_\alpha \cdot \hat{v}_{\beta\alpha} = v_{\alpha\beta}[1..|v_{\alpha\beta}|-1] \cdot v_{\alpha\beta}[|v_{\alpha\beta}|] \cdot v_\alpha \cdot v_{\beta\alpha} = v_{\alpha\beta} \cdot v_\alpha \cdot v_{\beta\alpha} = v. \qquad \square$$

Similar to Proposition 4.11, we can decompose the redistribution argument into smaller ones using perfect universal words and Lemma 4.9. Again, this yields a shorter although less informative proof.

Proof. Assume $w \sim_k w'$. By two applications of Lemma 4.10 for $i = 0, j = 1$ and $i = 1, j = m$ we obtain the equivalences.

Assume $\alpha_0\beta_1\alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1\alpha'_1$ and $\alpha_1\beta_2\alpha_2 \cdots \beta_m\alpha_m \sim_{k-1} \alpha'_1\beta'_2\alpha'_2 \cdots \beta'_m\alpha'_m$. By two applications of Lemma 4.10, we obtain that $\alpha_i \sim_{k-m} \alpha'_i$ for all $i \in [m]_0$. We have by repeated applications of Lemma 4.9 that

$$\begin{aligned}
\alpha_0\beta_1 \cdot \alpha_1 \cdot \prod_{i=2}^{m} \beta_i\alpha_i &\sim_k \alpha'_0\beta'_1 \cdot \alpha'_1 \cdot \prod_{i=2}^{m} \beta_i\alpha_i && (\alpha_0\beta_1\alpha_1 \sim_{k-(m-1)} \alpha'_0\beta'_1\alpha'_1)\\
&\sim_k \alpha'_0\beta'_1 \cdot \alpha_1 \cdot \prod_{i=2}^{m} \beta_i\alpha_i && (\alpha_1 \sim_{k-m} \alpha'_1)\\
&\sim_k \alpha'_0\beta'_1 \cdot \alpha'_1 \cdot \prod_{i=2}^{m} \beta'_i\alpha'_i, && (\alpha_1\beta_2\alpha_2 \cdots \beta_m\alpha_m \sim_{k-1} \alpha'_1\beta'_2\alpha'_2 \cdots \beta'_m\alpha'_m)
\end{aligned}$$

because $\iota(\alpha_0\beta_1) = 1$ and $\iota(\prod_{i=2}^{m} \beta_i\alpha_i) = m-1$. $\square$

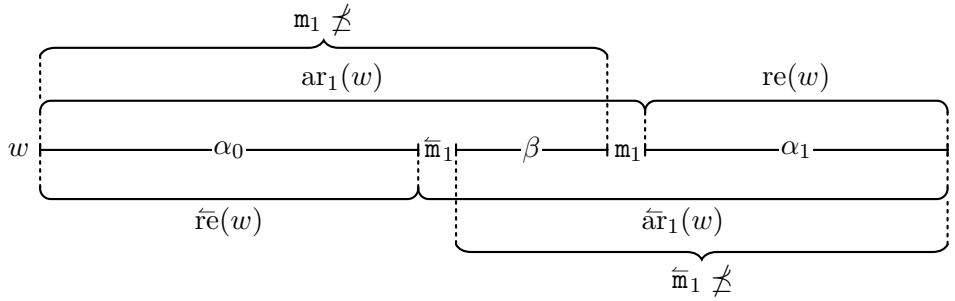


Figure 4.6: $\alpha\beta$-factorization of a 1-universal word w where $\overleftarrow{\mathtt{m}}_1\beta\mathtt{m}_1 \coloneqq \mathrm{B}_1(w)$ and $\alpha_i \coloneqq \mathrm{A}_i(w)$.

Corollary 4.17.1. *Let $w, w' \in \Sigma^*$ such that $m \coloneqq \iota(w) = \iota(w') < k$ and denote their $\alpha\beta$-factorizations by $w = \alpha_0\beta_1\alpha_1 \cdots \beta_m\alpha_m$ and $w' = \alpha'_0\beta'_1\alpha'_1 \cdots \beta'_m\alpha'_m$. Then, $w \sim_k w'$, if and only if, for all $i \in [m]$ we have $\alpha_{i-1}\beta_i\alpha_i \sim_{k-m+1} \alpha'_{i-1}\beta'_i\alpha'_i$.*

Proof. By induction over the number of arches using Proposition 4.17. □

Remark 4.17.2. Corollary 4.17.1 reduces the problem to finding the number of equivalence classes of 1-universal words with respect to all $k \geq 2$ and separating them by the classes of their A_0 and A_1.

4.4.1 Classes of 1-Universal Words

In the light of Corollary 4.17.1, we finish this section by considering words with exactly one arch. We give some preliminary results for them as well as some necessary and some sufficient conditions for their equivalence. Figure 4.6 shows the $\alpha\beta$-factorization of such a word. Because of the uniqueness of the last letter of an arch, we can factorize the word into three parts with a smaller alphabet each.

We start with results regarding the modi. In general, to decide whether two 1-universal words are equivalent is hard but in the case that $\mathrm{B}_1(w)$ starts and ends with the same letter we can give a simple characterization.

Lemma 4.18. *Let $w, w' \in \Sigma$ such that $\iota(w) = \iota(w') = 1$ and $\mathrm{B}_1(w), \mathrm{B}_1(w') \in \Sigma$, then $w \sim_k w'$, if and only if, $\mathrm{B}_1(w) = \mathrm{B}_1(w')$ and $\mathrm{A}_i(w) \sim_{k-1} \mathrm{A}_i(w')$ for $i \in [1]_0$.*

Proof. The backwards direction follows directly from Proposition 4.11. Thus, assume $w \sim_k w'$ and write $w = \alpha_0\mathtt{x}\alpha_1$ and $w = \alpha'_0\mathtt{x}'\alpha'_1$ for $\mathtt{x}, \mathtt{x}' \in \Sigma$. By Lemma 4.10 we obtain $\alpha_i \sim_{k-1} \alpha'_i$ for $i \in [1]_0$ and thus $\mathrm{alph}(\alpha_0) = \mathrm{alph}(\alpha'_0)$. By Remark 4.1.1, $\mathrm{alph}(\mathrm{ar}_0(w)) = \mathrm{alph}(\mathrm{A}_0(w)) \sqcup \{\mathtt{x}\}$ and $\mathrm{alph}(\mathrm{ar}_0(w')) = \mathrm{alph}(\mathrm{A}_0(w')) \sqcup \{\mathtt{x}'\}$ and thus $\mathtt{x} = \mathtt{x}'$. □

In the following we therefore assume $\mathrm{m}_1(w) \neq \overleftarrow{\mathrm{m}}_1(w)$, that is, $\overleftarrow{\mathtt{m}}_1 \neq \mathtt{m}_1$ in Figure 4.6. Furthermore, if the A_i have enough letters we can directly conclude that the modi already coincide. The assumption is equivalent to the fact that $\mathrm{A}_0(w)$ is 1-universal with respect to the alphabet $\Sigma \setminus \{\mathrm{m}_0(w)\}$.

Lemma 4.19. *Let $w, w' \in \Sigma^*$ such that $\iota(w) = \iota(w') = 1$ and $w \sim_k w'$ for $k \geq 2$.*

(1) If $|\operatorname{alph}(\mathrm{A}_0(w))| = |\Sigma| - 1$, then $\mathrm{m}_1(w) = \mathrm{m}_1(w')$.

(2) If $|\operatorname{alph}(\mathrm{A}_1(w))| = |\Sigma| - 1$, then $\overleftarrow{\mathrm{m}}_1(w) = \overleftarrow{\mathrm{m}}_1(w')$.

Proof. It suffices to show the first claim. By assumption, $|\operatorname{alph}(\mathrm{A}_0(w))| = |\Sigma| - 1$ and thus $\mathrm{m}_1(w)$ is the unique missing letter. By Lemma 4.10, we obtain $\mathrm{A}_0(w) \sim_{k-1} \mathrm{A}_0(w')$ and thus $\operatorname{alph}(\mathrm{A}_0(w)) = \operatorname{alph}(\mathrm{A}_0(w'))$. Therefore, $\mathrm{A}_0(w')$ is missing the same letter and thus $\mathrm{m}_1(w) = \mathrm{m}_1(w')$. □

In this last lemma, we show another decomposition, this time of 1-universal words with non-trivial A_i. In the lemma and proof we consider factorizations of A_i. We regard $\mathrm{A}_0(w)$ and $\mathrm{A}_1(w)$ by Remark 4.1.1 as words over the alphabets $\Sigma \setminus \{\mathrm{m}_1(w)\}$ and $\Sigma \setminus \{\overleftarrow{\mathrm{m}}_1(w)\}$ respectively. Furthermore, for simplicity, we assume that the A_i do not have more than $k-1$ arches, as we can always exchange A_i with more than $k-1$ arches by Proposition 4.11.

Lemma 4.20. *Let $u, v \in \Sigma^*$ such that $1 = \iota(u) = \iota(v)$ and $\iota(\mathrm{A}_i(u)), \iota(\mathrm{A}_i(v)) \leq k - 1$ for $i \in [1]_0$. Then $u \sim_k v$, if and only if,*

- *$\mathrm{A}_i(u) \sim_{k-1} \mathrm{A}_i(v)$ for all $i \in [1]_0$,*
- *$\operatorname{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \operatorname{re}(\mathrm{A}_1(u)) \sim_{\max(0,k-\ell_0-\ell_1)} \operatorname{re}(\mathrm{A}_0(v)) \cdot \mathrm{B}_1(v) \cdot \operatorname{re}(\mathrm{A}_1(v))$ where $\ell_i = \iota(\mathrm{A}_i(u)) = \iota(\mathrm{A}_i(v))$ for all $i \in [1]_0$.*

Proof. Let $u, v \in \Sigma^*$ such that $u \sim_k v$ and $1 = \iota(u) = \iota(v)$. By Corollary 4.10.1, we obtain $\mathrm{A}_i(u) \sim_{k-1} \mathrm{A}_i(v)$. If $\ell_0 + \ell_1 \geq k$, then the second condition follows because all words are 0-equivalent. Therefore, assume $\ell_0 + \ell_1 < k$ and thus $\ell \coloneqq k - \ell_0 - \ell_1 \geq 1$. Now we have for $\tilde{u} \coloneqq \operatorname{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \operatorname{re}(\mathrm{A}_1(u))$

$$\prod_{i=1}^{\ell_0} \mathrm{m}_i(\mathrm{A}_0(v)) \cdot \operatorname{SubWords}_{\leq \ell}(\tilde{u}) \cdot \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{m}}_i(\mathrm{A}_1(v)) \subseteq \operatorname{SubWords}_{\leq k}(u) = \operatorname{SubWords}_{\leq k}(v).$$

By Remark 2.5.1, we conclude

$$\operatorname{SubWords}_{\leq \ell}(\tilde{u}) \subseteq \operatorname{SubWords}_{\leq k}\left(\left(\prod_{i=1}^{\ell_0} \mathrm{ar}_i(\mathrm{A}_0(v))\right)^{-1} \cdot v \cdot \left(\prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v))\right)^{-1}\right) = \operatorname{SubWords}_{\leq \ell}(\operatorname{re}(\mathrm{A}_0(v)) \cdot \mathrm{B}_1(v) \cdot \operatorname{re}(\mathrm{A}_1(v))).$$

By a symmetric argument, we obtain $\operatorname{SubWords}_{\leq \ell}(\operatorname{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \operatorname{re}(\mathrm{A}_1(u))) = \operatorname{SubWords}_{\leq \ell}(\operatorname{re}(\mathrm{A}_0(v)) \cdot \mathrm{B}_1(v) \cdot \operatorname{re}(\mathrm{A}_1(v)))$.

Assume $\mathrm{A}_i(u) \sim_{k-1} \mathrm{A}_i(v)$ for all $i \in [1]_0$ and $\operatorname{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \operatorname{re}(\mathrm{A}_1(u)) \sim_{\max(0,k-\ell_0-\ell_1)} \operatorname{re}(\mathrm{A}_0(v)) \cdot \mathrm{B}_1(v) \cdot \operatorname{re}(\mathrm{A}_1(v))$ where $\ell_i \coloneqq \iota(\mathrm{A}_i(u)) = \iota(\mathrm{A}_i(v))$ for all $i \in [1]_0$. Because

$\mathrm{A}_0(u) \sim_{k-1} \mathrm{A}_0(v)$, we have $\mathrm{re}(\mathrm{A}_0(u)) \sim_{k-1-\ell_0} \mathrm{re}(\mathrm{A}_0(v))$ by Corollary 4.10.1 and therefore by Proposition 4.11 and $\mathrm{A}_0(v) \sim_{k-1} \mathrm{A}_0(u)$ that

$$\mathrm{A}_0(u) \sim_{k-1} \prod_{i=1}^{\ell_0} \mathrm{ar}_i(\mathrm{A}_0(v)) \cdot \mathrm{re}(\mathrm{A}_0(u)) \text{ and } \mathrm{A}_1(u) \sim_{k-1} \overleftarrow{\mathrm{re}}(\mathrm{A}_1(u)) \cdot \prod_{i=1}^{\ell_0} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v)) \quad (4.5)$$

by a symmetric argument. Next, we show

$$\tilde{u} \coloneqq \prod_{i=1}^{\ell_0} \mathrm{ar}(\mathrm{A}_0(v)) \cdot \mathrm{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(u)) \cdot \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}(\mathrm{A}_1(v)) \sim_k v. \quad (4.6)$$

Let $xyz \in \mathrm{SubWords}_{\leq k}(\tilde{u})$ such that x and z contain all occurrences of $\overleftarrow{\mathrm{m}}_1(u) = \overleftarrow{\mathrm{m}}_1(\tilde{u})$ and $\mathrm{m}_1(u) = \mathrm{m}_1(u)$ respectively, and are of minimal length. Therefore, $\mathrm{m}_1(u), \overleftarrow{\mathrm{m}}_1(u) \not\preceq y$.

Case 1 ($|x| \leq \ell_0$ and $|z| \leq \ell_1$). We define

$$p \coloneqq x \cdot \mathrm{pref}_{\ell_0-|x|}(y), \qquad r \coloneqq \mathrm{suff}_{\ell_1-|z|}(x \cdot p^{-1} \cdot y) \cdot z \qquad q \coloneqq xp^{-1}yr^{-1}z.$$

We have $p \preceq \prod_{i=1}^{\ell_0} \mathrm{ar}_i(\mathrm{A}_0(v))$ and $r \preceq \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v))$ by ℓ_0- and ℓ_1-universality respectively. Therefore, there exists an occurrence of xyz in $\tilde{u}$ such that at most $|q| \leq \max(0, k - \ell_0 - \ell_1)$ letters are chosen from $\mathrm{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(u))$.

Case 2 ($|x| > \ell_0$ and $|z| > \ell_1$). We define

$$p \coloneqq \mathrm{pref}_{\ell_0}(x), \qquad r \coloneqq \mathrm{suff}_{\ell_1}(z), \qquad q \coloneqq \mathrm{suff}_{|x|-\ell_0}(x) \cdot y \cdot \mathrm{pref}_{|z|-\ell_1}(z).$$

We have $p \preceq \prod_{i=1}^{\ell_0} \mathrm{ar}_i(\mathrm{A}_0(v))$ and $r \preceq \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v))$ by ℓ_0- and ℓ_1-universality respectively. Therefore, there exists an occurrence of xyz in $\tilde{u}$ such that at most $|q| \leq \max(0, k - \ell_0 - \ell_1)$ letters are chosen from $\mathrm{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(u))$.

Case 3 ($|x| > \ell_0$ and $|z| \leq \ell_1$ ($|x| \leq \ell_0$ and $|z| > \ell_1$ by symmetry)). We distinguish two cases whether the center factor y together with z fills all ℓ_1 arches.

Case 3.a ($|yz| \geq \ell_1$). We define

$$p \coloneqq \mathrm{pref}_{\ell_0}(x), \qquad r \coloneqq \mathrm{suff}_{\ell_1}(yz), \qquad q \coloneqq \mathrm{suff}_{|x|-\ell_0}(x) \cdot \mathrm{pref}_{|yz|-\ell_1}(yz).$$

We have $p \preceq \prod_{i=1}^{\ell_0} \mathrm{ar}_i(\mathrm{A}_0(v))$ and $r \preceq \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v))$ by ℓ_0- and ℓ_1-universality respectively. Therefore, there exists an occurrence of xyz in $\tilde{u}$ such that at most $|q| \leq \max(0, k - \ell_0 - \ell_1)$ letters are chosen from $\mathrm{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(u))$.

Case 3.b ($|yz| < \ell_1$). By definition, $x = \tilde{x} \cdot \overleftarrow{\mathrm{m}}_1(\tilde{u})$. Since $x \preceq xyz \preceq \tilde{u}$, we have

$$\tilde{x} \preceq \prod_{i=1}^{\ell_0} \mathrm{ar}(\mathrm{A}_0(v)) \cdot \mathrm{re}(\mathrm{A}_0(u)) \sim_{k-1} \mathrm{A}_0(v)$$

by Equation 4.5 and transitivity. Furthermore, $yz \preceq \prod_{i=2}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v))$ by $(\ell_1 - 1)$-universality. Lastly, $\mathrm{m}_1(\tilde{u}) \preceq \mathrm{B}_1(v) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(v)) \cdot \overleftarrow{\mathrm{ar}}_1(\mathrm{A}_1(v))$ by 1-universality. Thus,

$$xyz = \tilde{x} \cdot \overleftarrow{\mathrm{m}}_1(\tilde{u}) \cdot yz \preceq \mathrm{A}_0(v) \cdot \mathrm{B}_1(v) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(v)) \cdot \overleftarrow{\mathrm{ar}}_1(\mathrm{A}_1(v)) \cdot \prod_{i=2}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(\mathrm{A}_1(v)) = v.$$

The other inclusion follows by a symmetric argument. Using these three equivalences, we can conclude

$$
\begin{aligned}
u &\sim_k \prod_{i=1}^{\ell_0} \mathrm{ar}_i(v) \cdot \mathrm{re}(\mathrm{A}_0(u)) \cdot \mathrm{B}_1(u) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(u)) \cdot \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(v) && \text{(Proposition 4.11 and Eq 4.5)} \\
&\sim_k \prod_{i=1}^{\ell_0} \mathrm{ar}_i(v) \cdot \mathrm{re}(\mathrm{A}_0(v)) \cdot \mathrm{B}_1(v) \cdot \overleftarrow{\mathrm{re}}(\mathrm{A}_1(v)) \cdot \prod_{i=1}^{\ell_1} \overleftarrow{\mathrm{ar}}_i(v) && \text{(Eq. 4.6)} \\
&= v. && \square
\end{aligned}
$$

Remark 4.20.1. In the forward direction of the proof, we use a similar idea to Lemma 4.10 for arches of the A_i factors. These arches can be extended into the B_1 factor. Therefore, we could also obtain an equivalence of (in general proper) factors of B_1.

4.5 Conclusion and Future Work

In this chapter, we investigated the classes of Simon's congruence by using $\alpha\beta$-factorization. We studied the $\alpha\beta$-factorization in its original context and proved structure results for words w with exactly two and three absent subwords of length $\iota(w) + 1$ in Section 4.1. These results relied on Lemma 4.4 which is only applicable in the case of *shortest* absent subwords. For the characterization of words with three absent subwords of length k, we found cases in which these absent subwords where not of minimal length, and therefore Lemma 4.4 was no longer applicable. For these longer absent subwords, the used necessary and sufficient conditions used to prove Lemma 4.4 again hold but calculating the bound becomes harder due to overlaps. Trying to find an analog of Lemma 4.4 for longer absent subwords, could allow for characterization of classes of words with absent subwords of length less than $k - 1$. Furthermore, the structure results show that the absent subwords have to be similar, that is, differ in common positions. This suggests that absent (or occurring) subwords of some word, have to be in some sense *similar* or *close* to each other. This could be explored when trying to characterize which subsets of $\Sigma^{\leq k}$ can appear as sets of subwords.

Due to the results of the last section, we focused on another approach in Section 4.2 in which we separated the classes by the number of arches. We investigated the $\alpha\beta$-factorization in this context and showed a number of results showing the equivalence of α bordered parts of words and a characterization for words with already equal B. We applied these results in Section 4.3.1 to show a complete characterization of classes of binary words and for this special case where able to calculate the index of the congruence. The concrete numbers for the index of $\sim_k$ in the binary case suggest connections to number theory [4] and other areas of combinatorics [18]. Exploring them could reveal different approaches for analyzing subwords.

Lastly, in Section 4.4, we showed a characterization of equivalence classes in terms of equivalent classes of 1-universal words by decomposing words into $\alpha\beta\alpha$-factors. This result reduces the problem to the analysis of 1 universal words. Analyzing factors using repeated arch- or $\alpha\beta$-factorizations proved useful in some results in Section 4.4. In particular, we

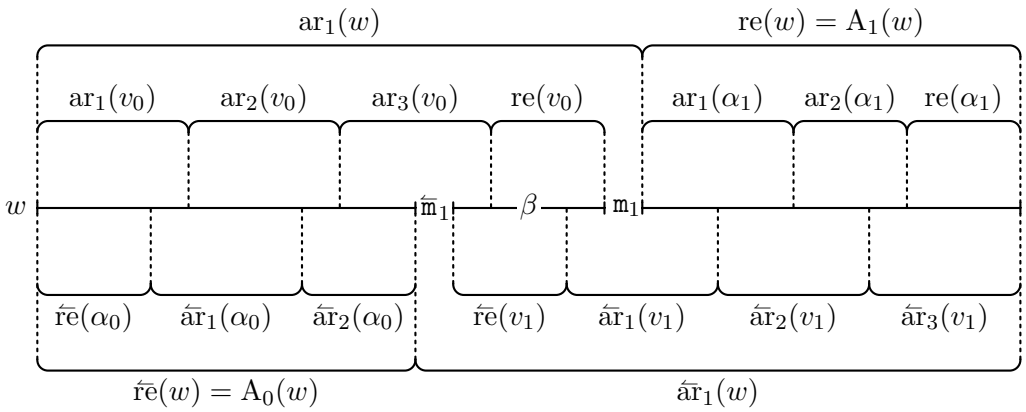


Figure 4.7: 1-universal word with a second layer of factorization. For the inner factorization let $v_0 \coloneqq \mathrm{ar}_1(w) \cdot \mathrm{m}_1^{-1}(w) \in (\Sigma \setminus \{\mathrm{m}_1(w)\})^*$ and v_1 analogous. For the factorization of the rests, we also consider $\mathrm{re}(w) \in (\Sigma \setminus \{\overleftarrow{\mathrm{m}}_1(w)\})^*$ and $\overleftarrow{\mathrm{re}}(w)$ analogous.

would like to point out the connection between repeated nested arch factorizations in both directions and the shortlex normalform by Fleischer and Kufleitner [8] or equivalently the characterization of minimal elements of $[w]_{\sim_k}$ by Simon [28] stated in Lothaire [22, Theorem 6.2.9].

For each arch, we can again factorize its longest proper prefix (with respect to its alphabet without the arch's modus). A similar idea was used by Simon [29, Lemma 3] and again in a similar fashion in *Combinatorics on Words* [22, Theorem 6.2.15] although not as a simultaneous factorization of the whole word. Figure 4.7 shows the second layer of the factorization for a 1-universal word. The number of layers is exactly $|\Sigma|$. This factorization therefore corresponds to one for $|\Sigma| = 3$ omitting the unary layer for readability, as all the unary arches are of length one and the rest is always empty. For some word w, each letter $w[i]$ is the modus of some (potentially inner) arch. Counting the (reverse) arches to this letter yields the optimal ranker chains and thus which letters can be omitted or permuted. Permutable factors occur whenever nested arches end exactly one after another in both directions. Exploring these connections could be insightful and allow for a more structural view on the minimal elements of the equivalence classes in general and not just for a specific word w.

Chapter 5

Conclusion

In this thesis, we investigated subword structures in words. We examined subwords from two different points of views, shuffles and Simon's congruence, that is, word decomposing into disjoint occurrences of subwords and words sharing all subwords up to a certain length.

In Chapter 3, we considered shuffles of words and explicitly shuffle squares and reverse shuffle squares. We showed that a (reverse) shuffle of a set with itself only coincides with the set of all (reverse) shuffle squares of its elements in the trivial case. Furthermore, we investigated the language theoretic complexity of languages of shuffles. We generalized a result by Henshall, Rampersad, and Shallit [16], showing that the languages of k-fold shuffle-powers are never context-free for $k \geq 2$. Furthermore, we started an investigation into whether the language of all shuffle squares is indexed where we considered necessary properties described by Gilman [13] and Smith [31]. Lastly, we showed that the mapping $v \mapsto v \sqcup\!\sqcup v$ is injective, that is, that a word is determined by the set of all its shuffle squares.

In Chapter 4, investigated the structure of some classes of Simon's congruence. First, we continued the work by Barker et al. [2] and Fleischmann et al. [10] and considered words with a fixed number of subwords. We characterized two further sets of classes by considering their shortest absent subwords. Furthermore, we showed a number of results about the $\alpha\beta$-factorization, among them, a theorem characterizing k-equivalence of ℓ-universal words in terms of their 1-universal $\alpha\beta\alpha$-factors. Lastly, we applied our results to the special case of binary words for which we proved a characterization for the equivalence of words and gave a formula for the number of classes of binary words for each k.

Bibliography

[1] Alfred Vaino Aho. "Indexed Grammars – An Extension of Context-Free Grammars". In: *Journal of the ACM* 15.4 (1968). doi: 10.1145/321479.321488.

[2] Laura Barker, Pamela Fleischmann, Katharina Harwardt, Florin Manea, and Dirk Nowotka. "Scattered Factor-Universality of Words". In: *Developments in Language Theory – 24th International Conference, DLT 2020, Proceedings* (Tampa, FL, USA, May 11–15, 2020). Vol. 12086. Lecture Notes in Computer Science. 2020. doi: 10.1007/978-3-030-48516-0_2.

[3] Jean Berstel and Luc Boasson. "Shuffle Factorization is Unique". In: *Theoretical Computer Science* 273.1-2 (2002). doi: 10.1016/S0304-3975(00)00433-3.

[4] David William Boyd. "Linear Recurrence Relations for some Generalized Pisot Sequences". In: *Thrid Conference of the Canadian Number Theory Association, Proceedings* (Queen's University, Kingston, Canada, Aug. 18–24, 1991). Advances in Number Theory. 1993. isbn: 9780198536680.

[5] Laurent Bulteau and Stéphane Vialette. "Recognizing Binary Shuffle Squares is NP-hard". In: *Theoretical Computer Science* 806 (2020). doi: 10.1016/j.tcs.2019.01.012.

[6] Sam Buss and Michael Soltys. "Unshuffling a Square is NP-hard". In: *Journal of Computer and System Sciences* 80.4 (2014). doi: 10.1016/j.jcss.2013.11.002.

[7] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. 2009. isbn: 978-0-521-89806-5. doi: 10.1017/CBO9780511801655.

[8] Lukas Fleischer and Manfred Kufleitner. "Testing Simon's Congruence". In: *43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018* (Liverpool, UK, Aug. 27–31, 2018). Vol. 117. Leibniz International Proceedings in Informatics. 2018. doi: 10.4230/LIPIcs.MFCS.2018.62.

[9] Pamela Fleischmann, Sebastian Bernhard Germann, and Dirk Nowotka. "Scattered Factor Universality – The Power of the Remainder". In: (2021). arXiv: 2104.09063 [cs.CL].

[10] Pamela Fleischmann, Lukas Haschke, Annika Huch, Annika Mayrock, and Dirk Nowotka. "Nearly k-Universal Words – Investigating a Part of Simon's Congruence". In: *Descriptional Complexity of Formal Systems – 24th IFIP WG 1.02 International Conference, DCFS 2022, Proceedings* (Debrecen, Hungary, Aug. 29–31,

2022). Vol. 13439. Lecture Notes in Computer Science. 2022. DOI: 10.1007/978-3-031-13257-5_5.

[11] Séverine Fratani and El Makki Voundy. "Homomorphic Characterizations of Indexed Languages". In: *Language and Automata Theory and Applications – 10th International Conference, LATA 2016, Proceedings* (Prague, Czech Republic, Mar. 14–18, 2016). Vol. 9618. Lecture Notes in Computer Science. 2016. DOI: 10.1007/978-3-319-30000-9_28.

[12] Pawel Gawrychowski, Maria Kosche, Tore Koss, Florin Manea, and Stefan Siemer. "Efficiently Testing Simon's Congruence". In: *38th International Symposium on Theoretical Aspects of Computer Science, STACS 2021* (Saarbrücken, Germany (Virtual Conference), Mar. 16–19, 2021). Vol. 187. Leibniz International Proceedings in Informatics. 2021. DOI: 10.4230/LIPIcs.STACS.2021.34.

[13] Robert Gilman. "A Shrinking Lemma for Indexed Languages". In: *Theoretical Computer Science* 163.1&2 (1996). DOI: 10.1016/0304-3975(96)00244-7.

[14] Takeshi Hayashi. "On Derivation Trees of Indexed Grammars, an Extension of the *uvwxy*-theorem". In: *Publications of the Research Institute for Mathematical Sciences* 9.1 (1973). DOI: 10.2977/PRIMS/1195192738.

[15] Jean-Jacques Hébrard. "An Algorithm for Distinguishing Efficiently Bit-Strings by their Subsequences". In: *Theoretical Computer Science* 82.1 (1991). DOI: 10.1016/0304-3975(91)90170-7.

[16] Dane Henshall, Narad Rampersad, and Jeffrey Outlaw Shallit. "Shuffling and Unshuffling". In: *Bulletin of the European Association for Theoretical Computer Science* 107 (2012). URL: http://eatcs.org/beatcs/index.php/beatcs/article/view/71.

[17] Graham Higman. "Ordering by Divisibility in Abstract Algebras". In: *Proceedings of the London Mathematical Society* s3-2.1 (1952). ISSN: 0024-6115. DOI: 10.1112/plms/s3-2.1.326.

[18] Milan Janjic. "Generalized Compositions with a Fixed Number of Parts". In: (2010). arXiv: 1012.3892 [math.CO].

[19] Makoto Kanazawa and Sylvain Salvati. "MIX Is Not a Tree-Adjoining Language". In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings* (Jeju Island, Korea, July 8–14, 2012). Vol. 1: Long Papers. 2012. URL: https://aclanthology.org/P12-1070/.

[20] Prateek Karandikar, Manfred Kufleitner, and Philippe Schnoebelen. "On the Index of Simon's Congruence for Piecewise Testability". In: *Information Processing Letters* 115.4 (2015). DOI: 10.1016/j.ipl.2014.11.008.

[21] Maria Kosche, Tore Koss, Florin Manea, and Stefan Siemer. "Absent Subsequences in Words". In: *Reachability Problems – 15th International Conference, RP 2021, Proceedings* (Liverpool, UK, Oct. 25–27, 2021). Vol. 13035. Lecture Notes in Computer Science. 2021. DOI: 10.1007/978-3-030-89716-1_8.

[22] M. Lothaire. *Combinatorics on Words*. 2nd ed. Cambridge Mathematical Library. 1997. ISBN: 978-0-521-59924-5. DOI: 10.1017/CBO9780511566097.

[23] William Marsh. "Some Conjectures on Indexed Languages". In: *Association for Symbolic Logic Meeting* (Stanford University, CA, USA, July 15–19, 1985). 1985.

[24] OEIS Foundation Inc. *The On-Line Encyclopedia of Integer Sequences*. Published electronically at http://oeis.org. 2022.

[25] Rohit Jivanlal Parikh. "On Context-Free Languages". In: *Journal of the ACM* 13.4 (1966). ISSN: 0004-5411. DOI: 10.1145/321356.321364.

[26] Romeo Rizzi and Stéphane Vialette. "On Recognizing Words That Are Squares for the Shuffle Product". In: *Computer Science – Theory and Applications – 8th International Computer Science Symposium in Russia, CSR 2013, Proceedings* (Ekaterinburg, Russia, June 25–29, 2013). Vol. 7913. Lecture Notes in Computer Science. 2013. DOI: 10.1007/978-3-642-38536-0_21.

[27] Sylvain Salvati. "MIX is a 2-MCFL and the Word Problem in Z^2 is captured by the IO and the OI hierarchies". In: *Journal of Computer and System Sciences* 81.7 (2015). DOI: 10.1016/j.jcss.2015.03.004.

[28] Imre Simon. "Hierarchies of Events with Dot-depth One". PhD thesis. University of Waterloo, Department of Applied Analysis and Computer Science, 1972.

[29] Imre Simon. "Piecewise Testable Events". In: *Automata Theory and Formal Languages, 2nd GI Conference* (Kaiserslautern, Germany, May 20–23, 1975). Vol. 33. Lecture Notes in Computer Science. 1975. DOI: 10.1007/3-540-07407-4_23.

[30] Imre Simon. "Words distinguished by their subwords (extended abstract)". In: *Proceedings of WORDS'03, 4th International Conference on Combinatorics on Words* (Turku, Finland, Sept. 10–13, 2003). TUCS General Publication 27. 2003. ISBN: 952-12-1211-X.

[31] Tim Smith. "A new Pumping Lemma for Indexed Languages, with an Application to Infinite Words". In: *Information and Computation* 252 (2017). DOI: 10.1016/j.ic.2016.11.002.

Acknowledgements

First and foremost, I want to thank my research supervisor Prof. Dr. Dirk Nowotka for introducing me into the field of combinatorics on words with his fascinating lectures, and providing me with the opportunity to write a master's thesis about a topic where I was able to conduct my own research. Second, I want to thank my second supervisor Dr. Pamela Fleischmann for our weakly insightful conversations, and her regular feedback on the thesis. Without her, this thesis would not be in the state it is now. I am looking forward to working with you both, as part of the dependable systems group.

Next, I want to thank my family: my parents, my sister Lilli, and of course our cat Henry, for their emotional support and their continued encouragement during my studies in general, and in particular this master's thesis.

Last but not least, I would like to thank my friends Alex, Anne, Björn, Malte, Max, Thore, and Yannik[1] for their feedback on the thesis, our conversations, about my topic, as well as, those explicitly not about it, and our game nights as an occasional break.

[1] Ordered by the lexicographical extension of the usual order on $\Sigma = \{\mathrm{A}, \ldots, \mathrm{Z}\}$.

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Weiterhin versichere ich, dass diese Arbeit noch nicht als Abschlussarbeit an anderer Stelle vorgelegen hat.

Kiel, 16. Mai 2026