THEORIE-TAG 2023

A. Lin, G. Zetzsche (Eds.): Theorietag 2023, Kaiserslautern, 4. – 6.10.2023
RPTU & MPI-SWS Kaiserslautern, Technical Report, pp. 71–74.

RPTU

α-β-Factorisation and the Binary Case of Simon's Congruence

Pamela Fleischmann[(A)] Jonas Höfer[(B)] Annika Huch[(A)]
Dirk Nowotka[(A)]

[(A)]Kiel University, Kiel, Germany
fpa@informatik.uni-kiel.de.de,stu216885@mail.uni-kiel.de,dn@informatik.uni-kiel.de

[(B)]University of Gothenburg, Sweden
jonas.hofer@gu.se

Abstract

Based on the arch factorisation (Hébrard 1991), first the notion of k-richness and later the one of k-universality - both measure words by their scattered factors - were introduced. In 2022 Fleischmann et al. presented a generalisation by intersecting the arch factorisations of a word and its reverse. Here we use this α-β-factorisation in order to characterise the Simon congruence of k-universal words in terms of 1-universal words and apply these results to binary words obtaining a full characterisation of the index of the congruence.

1. Introduction

A *scattered factor*, *subsequence*, or *scattered subword* of a word w is a word that is obtained by deleting letters from w while preserving the order of the remaining ones, e.g., `tea` and `thora` are both scattered factors of `theorietag`. In contrast to a factor, like `eta`, a scattered factor is not necessarily contiguous. Here, we focus on Simon's congruence [8] $\sim_k$ for $k \in \mathbb{N}_0$: $u \sim_k v$ iff u, v share all scattered factors up to length k. A long outstanding question, posed by Sakarovitch and Simon [7], is the exact structure of the congruence classes of $\sim_k$ and the index of the relation. Currently, no exact formula is known. One approach for studying scattered factors in words is based on the notion of *scattered factor universality* [1, 2, 3]. A word w is called *ℓ-universal* if it contains all words of length ℓ as scattered factors. For instance, the word `alfalfa`[1] is 2-universal since it contains all words of length two over the alphabet $\{\mathtt{a}, \mathtt{l}, \mathtt{f}\}$ as scattered factors. A main tool in this line of research is the *α-β-factorization* [3]. Kosche et al. [6] implicitly used this factorisation to determine shortest absent scattered factors in words.

Our Contribution. We investigate the α-β-factorization and give necessary and sufficient conditions for the congruence of words in terms of their factors. We characterise $\sim_k$ in terms of 1-universal words through their $\alpha\beta\alpha$-factors. We use these results to characterize the classes of binary words and their cardinality, as well as, to calculate the index in this special case. Lastly, we start to transfer the previous results to the ternary alphabet.

[1] Alfalfa (Medicago sativa) is plant whose name means *horse food* in Old Persian

2. Preliminaries

Let $\mathbb{N} = \{1,2,\ldots\}$ and set $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$, $[m] = \{1,\ldots,m\}$, and $[m]_0 = \{0\} \cup [m]$. For the standard definitions of combinatorics on words, we refer to [7]. We abbreviate an alphabet of cardinality $i \in \mathbb{N}$ by Σ_i. If $w = xy$ we write $x^{-1}w$ for y and wy^{-1} for x. A word $u \in \Sigma^*$ of length $n \in \mathbb{N}_0$ is called a *scattered factor* of $w \in \Sigma^*$ if there exist $v_0,\ldots,v_n \in \Sigma^*$ with $w = v_0u[1]v_1\cdots v_{n-1}u[n]v_n$. Let $\mathrm{ScatFact}(w), \mathrm{ScatFact}_k$, and $\mathrm{ScatFact}_{\leq k}$ denote the sets of all, exactly of length k, up to length k resp. scattered factors of w. For comparing words w.r.t. their scattered factors, Simon introduced a congruence relation nowadays known as *Simon's congruence* [8]: two words $u, v \in \Sigma^*$ are called *Simon k-congruent ($u \sim_k v$)* iff $\mathrm{ScatFact}_{\leq k}(u) = \mathrm{ScatFact}_{\leq k}(v)$ for some $k \in \mathbb{N}$. A word $w \in \Sigma^*$ is called *k-universal* w.r.t. Σ if $\mathrm{ScatFact}_k(w) = \Sigma^k$. The maximal k such that w is k-universal is denoted by $\iota(w)$ and called w's *universality index*. For a word $w \in \Sigma^*$ the *arch factorisation* is given by $w = \mathrm{ar}_1(w)\cdots\mathrm{ar}_k(w)\,\mathrm{re}(w)$ for $k \in \mathbb{N}_0$ with $\mathrm{alph}(\mathrm{ar}_i(w)) = \Sigma$ for all $i \in [k]$, the last letter of $\mathrm{ar}_i(w)$ occurs exactly once in $\mathrm{ar}_i(w)$ for all $i \in [k]$, and $\mathrm{alph}(\mathrm{re}(w)) \subset \Sigma$. The words $\mathrm{ar}_i(w)$ are called *arches* and $\mathrm{re}(w)$ is the *rest* of w. Define the *modus of* w as $\mathrm{m}(w) = \mathrm{ar}_1(w)[|\mathrm{ar}_1(w)|]\cdots\mathrm{ar}_k(w)[|\mathrm{ar}_k(w)|] \in \Sigma^k$. Set $\mathrm{ar}_{i..j}(w) = \mathrm{ar}_i(w)\cdots\mathrm{ar}_j(w)$. The α-β-factorisation was introduced in [3] inspired by [6]. Define for the arch factorisation of w^R (read left to right) the i^{th} *reverse arch* $\overleftarrow{\mathrm{ar}}_i(w) = (\mathrm{ar}_{\iota(w)-i+1}(w^R))^R$, the *reverse rest* $\overleftarrow{\mathrm{re}}(w) = (\mathrm{re}(w^R))^R$, and set $\overleftarrow{\mathrm{m}}(w)$ as $\mathrm{m}(w^R)^R$ for the *reverse modus*.

Definition 2.1 *For $w \in \Sigma^*$ define w's* α-β-factorisation *by $w = \alpha_0\beta_1\alpha_1\cdots\alpha_{\iota(w)-1}\beta_{\iota(w)}\alpha_{\iota(w)}$ with $\mathrm{ar}_i(w) = \alpha_{i-1}\beta_i$ and $\overleftarrow{\mathrm{ar}}_i(w) = \beta_i\alpha_i$ for all $i \in [\iota(w)]$, $\overleftarrow{\mathrm{re}}(w) = \alpha_0$, as well as $\mathrm{re}(w) = \alpha_{\iota(w)}$. Define $\mathrm{core}_i = \beta_i[2..|\beta_i|-1]$ or ε if $|\beta_i| \leq 2$.*

Note that the α-β-factorisation is left-right-symmetric and that the i^{th} reverse arch always starts inside the i^{th} arch. We finish this section with three results from [4, 5, 8]

Lemma 2.2 *Let $u, v \in \Sigma^*$, $u', v' \in \Sigma^+$, and $\mathrm{x} \in \Sigma$.*
(1) If $u \sim_k v$ then $w_1uw_2 \sim_{\iota(w_1)+k+\iota(w_2)} w_1vw_2$.
(2) $u'v' \sim_k u'$ iff $u' = u'_1\cdots u'_k$ such that $\mathrm{alph}(u'_1) \supseteq \ldots \supseteq \mathrm{alph}(u'_k) \supseteq \mathrm{alph}(v')$.
(3) $uv \sim_k u\mathrm{x}v$ iff there exist $p, p' \in \mathbb{N}_0$ with $p + p' \geq k$ and $u\mathrm{x} \sim_p u$ and $\mathrm{x}v \sim_{p'} v$.

3. α-β-Factorisation

In this section, we investigate the α-β-factorisation based on results of [4]. The main result states that it suffices to look at 1-universal words in order to gain the information about the $\sim_k$ congruence classes. First, we show that *cutting of ℓ* arches from two k-congruent words each, leads to $(k-\ell)$-congruence. Then we connect the congruence of words to the congruence of their α factors, leading to a characterisation by the congruence of $\alpha\beta\alpha$-factors.

Lemma 3.1 *Let $w, \tilde{w} \in \Sigma^*$ with $w \sim_k \tilde{w}$ and $\iota(w) = \iota(\tilde{w}) < k$, then $\mathrm{ar}_1^{-1}(w)\cdot w \sim_{k-1} \mathrm{ar}_1^{-1}(\tilde{w})\cdot\tilde{w}$ and $\alpha_i\beta_{i+1}\alpha_{i+1}\cdots\alpha_j \sim_{k-\iota(w)+j-i} \tilde{\alpha}_i\tilde{\beta}_{i+1}\tilde{\alpha}_{i+1}\cdots\tilde{\alpha}_j$ for all $0 \leq i \leq j \leq \iota(w)$.*

Proposition 3.2 *For all $w, \tilde{w} \in \Sigma^*$ with $m = \iota(w) = \iota(\tilde{w}) < k$ such that $\beta_i = \tilde{\beta}_i$ for all $i \in [m]$, we have $w \sim_k \tilde{w}$ iff $\alpha_i \sim_{k-m} \tilde{\alpha}_i$ for all $i \in [m]_0$. Thus, $w \sim_k \tilde{w}$ iff $\alpha_i \sim_{k-m} \tilde{\alpha}_i$ for all $i \in [m]_0$ and for $w' = \alpha_0\tilde{\beta}_1\alpha_1\cdots\tilde{\beta}_m\alpha_m$ we have $w \sim_k w'$.*

Theorem 3.3 *Let $w, \tilde{w} \in \Sigma^*$ with $m = \iota(w) = \iota(\tilde{w}) < k$. Then, $w \sim_k \tilde{w}$ iff $\alpha_{i-1}\beta_i\alpha_i \sim_{k-m+1} \tilde{\alpha}_{i-1}\tilde{\beta}_i\tilde{\alpha}_i$ for all $i \in [m]$.*

In the light of Theorem 3.3, in the following, we consider some special cases of these triples w.r.t. the alphabet of the both involved α. Hence, let $w, \tilde{w} \in \Sigma^*$ with $1 = \iota(w) = \iota(\tilde{w})$.

Proposition 3.4 *(1) Let $\alpha_0 = \alpha_1 = \tilde{\alpha}_0 = \tilde{\alpha}_1 = \varepsilon$. Then $w \sim_k \tilde{w}$ iff $k = 1$ or $k \geq 2$, $\mathrm{m}(w) = \mathrm{m}(\tilde{w})$, $\overleftarrow{\mathrm{m}}(w) = \overleftarrow{\mathrm{m}}(\tilde{w})$, and $\mathrm{core}_1 \sim_k \widetilde{\mathrm{core}}_1$.*
(2) Let $\mathrm{alph}(\alpha_i) = \mathrm{alph}(\tilde{\alpha}_i) \in \binom{\Sigma}{|\Sigma|-1}$, then $w \sim_k \tilde{w}$ iff $\alpha_i \sim_{k-1} \tilde{\alpha}_i$ for all $i \in [1]_0$.

The last proposition does not hold if not both $\mathrm{m}(w)$ and $\overleftarrow{\mathrm{m}}(w)$ are identical: consider $w = \mathtt{ababeabab} \cdot \mathtt{abecd} \cdot \mathtt{cdcdcd} \sim_4 \mathtt{ababeabab} \cdot \mathtt{baedc} \cdot \mathtt{cdcdcd} = \tilde{w}$ with $\mathrm{m}(w) = \mathtt{d} \neq \mathtt{c} = \mathrm{m}(\tilde{w})$ and $\overleftarrow{\mathrm{m}}(w) = \mathtt{a} \neq \mathtt{b} = \overleftarrow{\mathrm{m}}(\tilde{w})$. In the next proposition, we give a necessary condition for the α-factors.

Proposition 3.5 *Let $w \in \Sigma^*$ with $\iota(w) = 1$, $k \in \mathbb{N}$, and $\overleftarrow{\mathrm{M}} = \{\overleftarrow{\mathrm{m}}(\tilde{w})[1] \mid \tilde{w} \in [w]_{\sim_k}\}$. If $|\overleftarrow{\mathrm{M}}| \geq 2$ then there exists a factorisation $\alpha_0 = u_1 \cdots u_{k-1}$ with $\mathrm{alph}(u_1) \supseteq \ldots \supseteq \mathrm{alph}(u_{k-1}) \supseteq \overleftarrow{\mathrm{M}}$.*

4. The Binary and Ternary Case of Simon's Congruence

First, we apply our results to the binary alphabet. Note that for a given w with $\iota(w) \leq k$, we have $|\{\overleftarrow{\mathrm{m}}(\tilde{w}) \mid \tilde{w} \in [w]_{\sim_k}\}| = 1$.

Proposition 4.1 *For all $w \in \Sigma_2^*$, we have for all $i \in [\iota(w)]$, $\beta_i \in \{\mathtt{a}, \mathtt{b}, \mathtt{ab}, \mathtt{ba}\}$. If $\beta_i = \mathtt{x}$, then $\alpha_{i-1}, \alpha_i \in \overline{\mathtt{x}}^+$ with $\mathtt{x} \in \Sigma_2$ and if $\beta_i = \mathtt{x}\overline{\mathtt{x}}$, then $\alpha_{i-1} \in \mathtt{x}^*$ and $\alpha_i \in \overline{\mathtt{x}}^*$ with $\mathtt{x} \in \Sigma_2$.*

In the binary case the k-congruence of two words with identical $\iota < k$ leads to the same modi and same β giving a characterisation of $\sim_k$ for binary words in terms of unary words.

Lemma 4.2 *Let $w, w' \in \Sigma_2^*$ with $w \sim_k w'$ and $m = \iota(w) = \iota(w') < k$, then $\mathrm{m}(w) = \mathrm{m}(w')$ and thus, $\beta_i = \beta_i'$ for all $i \in [m]$.*

Theorem 4.3 *Let $w, w' \in \Sigma_2^*$ such that $m = \iota(w) = \iota(w') < k$, then $w \sim_k w'$ iff $\beta_i = \beta_i'$ for all $i \in [m]$ and $\alpha_i \sim_{k-m} \alpha_i'$ for all $i \in [m]_0$.*

Theorem 4.3 implies if $|[w]_{\sim_k}| = \infty$, then $\mathtt{x}^k \in \mathrm{ScatFact}_k(w)$ for some $\mathtt{x} \in \Sigma$ (the contrary is generally not true: $v = \mathtt{bbabb}$ w.r.t. $\sim_4$). The following theorem characterises the binary case.

Theorem 4.4 *Let $w \in \Sigma_2^*$, then $|[w]_{\sim_k}| < \infty$. We have $|[w]_{\sim_k}| = 1$ iff $\iota(w) < k$ and $|\alpha_i| < k - \iota(w)$ for all $i \in [\iota(w)]_0$.*

We present a formula for the precise value of $|\Sigma_2^*/\sim_k|$ counting classes based on the valid combinations of β-factors and number of classes for each α-factors and giving the index.

Theorem 4.5 *The number of congruence classes of $\Sigma_2^*/\sim_k$ of words with $m < k$ arches is given by $\| \binom{k-m\ \ k-m\ \ k-m}{1\ \ \ 2\ \ \ 1\ \ \ } \binom{k-m}{1}^m \cdot \binom{k-m}{k-m} \|_1 = c_k^m$ where $c_k^{-1} = 1$, $c_k^0 = 2k+1$, and $c_k^m = 2 \cdot (k-m+1) \cdot c_{k-1}^{m-1} - 2 \cdot (k-m) \cdot c_{k-2}^{m-2}$ where $\|\cdot\|_1$ denotes the 1-norm.*

Corollary 4.6 *For $k \in \mathbb{N}_0$ we have $|\Sigma_2^*/\sim_k| = 1 + \sum_{m=0}^{k-1} c_k^m$.*

Now, we consider Σ_3. Note that if $\mathrm{m}_1(w) = \overleftarrow{\mathrm{m}}_1(w)$ then $\mathrm{core}_1 = \varepsilon$. If $\mathrm{m}_1(w) \neq \overleftarrow{\mathrm{m}}_1(w)$, then $\mathrm{core}_1 \in (\Sigma \setminus \{\mathrm{m}_1(w), \overleftarrow{\mathrm{m}}_1(w)\})^*$, i.e., the cores are unary and denoted by $\mathrm{y} \in \Sigma_3$. Define for a boolean predicate P, $\delta_{P(x)} = 1$ if $P(x)$ is true and 0 otherwise. We assume $k \geq 2$ (we characterise 1-universal words) and $w, \tilde{w} \in \Sigma_3^*$ with $1 = \iota(w) = \iota(\tilde{w})$.

Lemma 4.7 *Let $\mathrm{m}(w) = \mathrm{m}(\tilde{w})$ and $\overleftarrow{\mathrm{m}}(w) = \overleftarrow{\mathrm{m}}(\tilde{w})$, we have $w \sim_k \tilde{w}$ iff $\alpha_i \sim_{k-1} \tilde{\alpha}_i$ for all $i \in [1]_0$ and $\mathrm{core}_1 \sim_{k-c} \widetilde{\mathrm{core}}_1 \in \mathrm{y}^*$ where $c \coloneqq \iota(\alpha_0) + \delta_{\mathrm{y} \preceq \mathrm{re}(\alpha_0)} + \iota(\alpha_1) + \delta_{\mathrm{y} \preceq \overleftarrow{\mathrm{re}}(\alpha_1)}$.*

We finish with a characterisation in the ternary case.

Theorem 4.8 *For $w, \tilde{w} \in \Sigma_3^*$ we have $w \sim_k \tilde{w}$ iff $\alpha_i \sim_{k-1} \tilde{\alpha}_i$ for all $i \in [1]_0$, and*
(1) $|\mathrm{alph}(\alpha_i)| = 2$, $\mathrm{alph}(\alpha_{1-i}) \cap \mathrm{alph}(\alpha_i) = \emptyset$, and $\iota(\alpha_i) \geq k-1$ for some $i \in [1]_0$, or
(2) $\mathrm{m}(w) = \mathrm{m}(\tilde{w})$, $\overleftarrow{\mathrm{m}}(w) = \overleftarrow{\mathrm{m}}(\tilde{w})$, and $\mathrm{core} \sim_{k-c} \widetilde{\mathrm{core}}$ where $c \coloneqq \iota(\alpha_0) + \delta_{\mathrm{y} \preceq \alpha_0} + \iota(\alpha_1) + \delta_{\mathrm{y} \in \alpha_1}$.

5. Conclusion

In this paper, we investigated the α-β-factorisation (cf. [6, 3]) as an object of intrinsic interest. This leads to a result characterising k-congruence of m-universal words in terms of their 1-universal $\alpha\beta\alpha$-factors. In the case of the binary and ternary alphabet, we fully characterised the congruence of words in terms of their single factors. Extending this idea of the α-β-factorisation to lower layers (arches w.r.t. some $\Omega \subset \Sigma$), is left as future work.

References

[1] L. Barker, P. Fleischmann, K. Harwardt, F. Manea, D. Nowotka, Scattered factor-universality of words. In: *DLT*. Springer, 2020, 14–28.

[2] P. Fleischmann, S. Germann, D. Nowotka, Scattered Factor Universality–The Power of the Remainder. *preprint arXiv:2104.09063 (published at RuFiDim)* (2021).

[3] P. Fleischmann, L. Haschke, A. Huch, A. Mayrock, D. Nowotka, Nearly k-universal words-investigating a part of simon's congruence. In: *DCFS*. 2022, 57–71.

[4] P. Karandikar, M. Kufleitner, P. Schnoebelen, On the index of Simon's congruence for piecewise testability. *Inf. Process. Lett.* **115** (2015) 4, 515–519.

[5] P. Karandikar, P. Schnoebelen, The height of piecewise-testable languages and the complexity of the logic of subwords. *LICS* **15** (2019) 2.

[6] M. Kosche, T. Koss, F. Manea, S. Siemer, Absent subsequences in words. In: *RP*. Springer, 2021, 115–131.

[7] M. Lothaire, *Combinatorics on Words*. Cambridge Mathematical Library, Cambridge University Press, 1997.

[8] I. Simon, Piecewise testable events. In: *Autom. Theor. Form. Lang., 2nd GI Conf.*. LNCS 33, Springer, 1975, 214–222.