Some distances and an unsolved problem

Gerhard Lischke *

Abstract

We investigate a Hamming distance-based measure between words and the sets of all generalized primitive or generalized periodic words, respectively. After repeating some known results we concentrate to the distance to the most general set of generalized periodic words. In the case of a two-letter alphabet the final answer is still unknown and may be a great challenge for research.

1 Introduction and preliminaries

A Hamming distance-based measure from coding theory [2] was used in several papers to study the distance between words and languages or bitween different languages, see, for instance, [5] and the references there. In [4], some special kinds of periodicity and primitivity for words have been introduced and investigated, see also [3,5]. There was the question for the distance between arbitrary words and the languages of generalized periodic or primitive words, respectively. After repeating the most important facts of these distances we concentrate to the distance between arbitrary words and the set of quasi-quasi-periodic words where the main question remains unsolved.

Let X be a fixed finite, nontrivial alphabet. This means, X is a finite set having at least two symbols denoted by a and b. X^* is the free monoid generated by X or the set of all words over X. The empty word is denoted by e, and $X^+ =_{Df} X^* \setminus \{e\}$.

For a word $p \in X^*$, $|p|$ denotes the length of p. For a natural number n, p^n denotes the concatenation of n copies of the word p. For $1 \leq i \leq |p|$, $p[i]$ is the letter at the i-th position of p. For words $p, q \in X^*$, p is a *prefix of* q, in symbols $p \sqsubseteq q$, if there exists $r \in X^*$ such that $q = pr$. p is a *strict prefix of* q, in symbols $p \sqsubset q$, if $p \sqsubseteq q$ and $p \neq q$. $Pr(q) =_{Df} \{p : p \sqsubset q\}$ is the *set of all strict prefixes*

*retired from Fakultät für Mathematik und Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 1-4, 07743 Jena, Germany; e-mail: gerhard.lischke@uni-jena.de

of q (including e if $q \neq e$). p is a *subword of* q, in symbols $p \top q$, if there exist $r, s \in X^*$ such that $q = rps$. $p \not\top q$ means that p does not occur as a subword of q.

For sets A, B, $A \subseteq B$ denotes their inclusion and $A \subset B$ denotes their strict inclusion.

In [4] a folding operation $\otimes$ was introduced which was a little more general than the following. For $p, q \in X^*$,

$$p \otimes q =_{Df} \begin{cases} \{p\} & \text{if } q = e \\ \{w_1 w_2 w_3 : \ w_3 \neq e \ \wedge \ w_1 w_2 = p \ \wedge \ w_2 w_3 = q\} & \text{otherwise,} \end{cases}$$

$p^{\otimes 0} =_{Df} \{e\}$, $\ p^{\otimes n} =_{Df} \bigcup \{w \otimes p : w \in p^{\otimes n-1}\}$ for $n \geq 1$.
For sets $A, B \subseteq X^*$, $\ A \otimes B =_{Df} \bigcup \{p \otimes q : p \in A \ \wedge \ q \in B\}$.

The reason for using this little restricted definition is discussed in [6], but all results and proofs in [3,4,5] remain unchanged under the new definition.

Next we cite the following definitions from [3,4,5].

$Per \ =_{Df} \ \{u : \exists v \exists n (v \sqsubset u \wedge n \geq 2 \wedge u = v^n)\}$
is the set of *periodic* words.

$Q \ =_{Df} \ X^+ \setminus Per$ is the set of *primitive* words.

$SPer \ =_{Df} \ \{u : \exists v \exists n (v \sqsubset u \wedge n \geq 2 \wedge u \in v^n \cdot Pr(v))\}$
is the set of *semi-periodic* words.

$SQ \ =_{Df} \ X^+ \setminus SPer$ is the set of *strongly primitive* words.

$QQPer \ =_{Df} \ \{u : \exists v \exists n (v \sqsubset u \wedge n \geq 2 \wedge u \in v^{\otimes n} \otimes Pr(v))\}$
is the set of *quasi-quasi-periodic* words.

$HHQ \ =_{Df} \ X^+ \setminus QQPer$ is the set of *hyperhyperprimitive* words.

Three further kinds of periodicity and primitivity of words are also defined and investigated in [3,4,5] but they will not be considered here.

We have the following strict inclusions:
$Per \subset SPer \subset QQPer \subset X^+$, $\quad HHQ \subset SQ \subset Q \subset X^+$.

For $u \in X^+$, the shortest word v such that there exists a natural number n with $u \in v^{\otimes n} \otimes Pr(v)$ is called the *hyperhyperroot* of u, denoted by $hhroot(u)$.

For two words u and v of the same length, the *Hamming distance* is
$h(u, v) =_{Df} |\{i : 1 \leq i \leq |u| \ \wedge \ u[i] \neq v[i]\}|$, where $|M|$ for a set M denotes its cardinality.

For $p \in X^*$ and $L \subseteq X^*$, assuming L contains words of length $|p|$,
$d(p, L) =_{Df} \min\{h(p, q) : |q| = |p| \ \wedge \ q \in L\}$ is the *distance* between the word p and the language L.

Let $k = |X|$, $k \geq 2$. For a natural number n and a language $L \subseteq X^*$ which has words of length n, $\ md_k(n, L) =_{Df} \max\{d(p, L) : p \in X^* \ \wedge \ |p| = n\}$ is the *maximal distance between words of length* n *and the language* L.

As usual, for a real number r, $\lfloor r \rfloor$ denotes the greatest integer which is smaller or equal to r, and $\lceil r \rceil$ denotes the smallest integer which is greater or equal to r.

2 Known results

We are interested in the distances $md_k(n, L)$ where L is one of the sets Per, $SPer$, $QQPer$, Q, SQ, HHQ, and $n \geq 2$ (For $n < 2$ no periodic words of length n exist).

Theorem 1 [5]. If $p \in QQPer$ then there exists $q \in HHQ$ with $h(p, q) = 1$, and therefore $md_k(n, HHQ) = md_k(n, SQ) = md_k(n, Q) = 1$ for all $n, k \geq 2$.

Proof. Assume $p \in QQPer$. Let a be the first letter of p.
Case 1). There is no other letter in p, i.e. $p = a^i$, $i \geq 2$. Then $q = p^{i-1}b \in HHQ$.
Case 2). There is still another letter in p, let's say b, and i should be the greatest length of a subword of p consisting of letters b only. This means, $p = p_1 ab^i p_2$ where $p_1 = e$ or $a \sqsubseteq p_1$, $b^i \not\top p_1$ and $b^{i+1} \not\top b^i p_2$. Then let $q = p_1 bb^i p_2$. Assume $q \in QQPer$. Then $q \in v^{\otimes m} \otimes Pr(v)$ for some $v \sqsubset q$ and $m \geq 2$. Then $p_1 b^{i+1} \sqsubseteq v$ must follow and therefore $b^{i+1} \top p_2$ which is a contradiction.
In both cases we found $q \in HHQ$ with $h(p, q) = 1$ and therefore $md_k(n, HHQ) = 1$ for all $n, k \geq 2$. $md_k(n, SQ)$ and $md_k(n, Q)$ cannot be greater because of $HHQ \subset SQ \subset Q \subset X^+$. $\square$

Thus it is enough to change one letter (or one bit in the case of a two-letter alphabet) in a quasi-quasi-periodic or periodic word to transfer it into a hyperhyperprimitive or primitive word. But in the opposite direction, from some primitive word to a nonprimitive one the distance may be greater and is given by more complicated formulas or it is still unknown.

Theorem 2 [5]. For natural numbers $n, k \geq 2$ it holds that

$$md_k(n, Per) = \begin{cases} n - \frac{n}{s}(\lfloor \frac{s}{k} \rfloor + 1) & \text{if there is a divisor of } n \text{ which is} \\ & \text{not 1 and not dividable by } k, \text{ and} \\ & s \text{ is the smallest such divisor} \\ n - \frac{n}{k} & \text{if } k \text{ is prime and } n \text{ is a power of } k. \end{cases}$$

The complete proof is given in [5]. Remark, that there is no divisor of n which is not 1 and not dividable by k if and only if k is prime and n is a power of k.

Theorem 3 [5]. For natural numbers $n, k \geq 2$ it holds that

$$md_k(n, SPer) = \begin{cases} \lceil \frac{n}{3} \rceil & \text{if } k = 2 \\ \lceil \frac{n}{2} \rceil & \text{if } k > 2. \end{cases}$$

Theorem 4 [5]. For natural numbers $n \geq 2$ and $k \geq 3$ it holds that $md_k(n, QQPer) = md_3(n, SPer) = \lceil \frac{n}{2} \rceil$.

It is clear that $md_k(n, QQPer) \leq md_k(n, SPer)$ because of $SPer \subset QQPer$. To show the equality we have to find for each $n \geq 2$ a word p of length n over a k-letter alphabet such that there exists $q' \in QQPer$ with $|q'| = n$ and $h(p, q') = \lceil \frac{n}{2} \rceil$, and there is no $q \in QQPer$ with $|q| = n$ and $h(p, q) < \lceil \frac{n}{2} \rceil$. Such a word p is called a *witness word*.

Let a, b, c be three pairwise different letters from the alphabet X. Then it is not hard to see that $p = a^{\lceil \frac{n}{3} \rceil} b^{\lfloor \frac{n+1}{3} \rfloor} c^{\lfloor \frac{n}{3} \rfloor}$ may act as a witness word, which is shown in [5].

3 The open case

It remains to determine the distance $md_2(n, QQPer)$, and it is clear that $md_2(n, QQPer) \leq md_2(n, SPer) = \lceil \frac{n}{3} \rceil$. But there are great problems because of the overlaps in the quasi-quasi-periodic words. The goal must be to find for each $n \geq 2$ a formula m_n in n and a word p of length n over $\{a, b\}$ such that

(1) there exists $q' \in QQPer$ with $|q'| = n$ and $h(p, q') = m_n$, and

(2) to show that there is no $q \in QQPer$ with $|q| = n$ and $h(p, q) < m_n$.

Then $md_2(n, QQPer) = m_n$, and such a word p is called a *witness word* again.

In looking for witness words we may restrict to words beginning with a since all facts regarding generalized periodicity or primitivity of words over $\{a, b\}$ are also true for the dual words (it means, by exchanging a and b). Therefore we have to inspect 2^{n-1} words of length n beginning with a. The first attempt for witness words was $p = a^{\lceil \frac{n}{3} \rceil} b^{\lfloor \frac{2n}{3} \rfloor}$ and $m_n = \lceil \frac{n}{3} \rceil$. But then (2) is not fulfilled for each n. The smallest counterexample, found by Georg Lohmann [7] is for $n = 13$: $p = a^5 b^8$. It is $h(p, q) = 4 < \lceil \frac{n}{3} \rceil$ for $q = abbaabbabbabb \in (abba)^{\otimes 3} \otimes (abb)$ and thus $q \in QQPer$ with $hhroot(q) = abba$. Even more, we could show

Lemma 5. For $p = a^{\lceil \frac{n}{3} \rceil} b^{\lfloor \frac{2n}{3} \rfloor}$ holds that $d(p, QQPer) \leq \lceil \frac{n}{3} \rceil - \ell$ if $n = |p| \geq 21\ell$.

Proof. Case 1) $n = 3j$. Then $\lceil \frac{n}{3} \rceil = j$, $p = a^j b^{2j}$ and $q = a^\ell b^{j-3\ell} a^{2\ell} b^{j-3\ell} a^\ell b^{j-3\ell} a^\ell b^{4\ell} \in QQPer$ with $hhroot(q) = a^\ell b^{j-3\ell} a^\ell$ and $h(p, q) = j - \ell$ if $j - 3\ell \geq 4\ell$ and therefore $n \geq 21\ell$.

Case 2) $n = 3j + 1$. Then $\lceil \frac{n}{3} \rceil = j + 1$, $p = a^{j+1}b^{2j}$ and $q = a^\ell b^{j+1-3\ell} a^{2\ell} b^{j+1-3\ell} a^\ell b^{j+1-3\ell} a^\ell b^{4\ell-2} \in QQPer$ with $hhroot(q) = a^\ell b^{j+1-3\ell} a^\ell$ and $h(p,q) = j + 1 - \ell$ if $j + 1 - 3\ell \geq 4\ell - 2$ and therefore $n \geq 21\ell - 8$.

Case 3) $n = 3j + 2$. Then $\lceil \frac{n}{3} \rceil = j + 1$, $p = a^{j+1}b^{2j+1}$ and $q = a^\ell b^{j+1-3\ell} a^{2\ell} b^{j+1-3\ell} a^\ell b^{j+1-3\ell} a^\ell b^{4\ell-1} \in QQPer$ with $hhroot(q) = a^\ell b^{j+1-3\ell} a^\ell$ and $h(p,q) = j + 1 - \ell$ if $j + 1 - 3\ell \geq 4\ell - 1$ and therefore $n \geq 21\ell - 4$. $\square$

Corollary. If $m_n = \lceil \frac{n}{3} \rceil$ then the words $a^{\lceil \frac{n}{3} \rceil} b^{\lfloor \frac{2n}{3} \rfloor}$ are not suitable as witness words for $n \in \{13, 16, 17\}$ or $n \geq 19$.

Péter Burcsi [1] in Budapest developed computer programs to list for each n and each $d \leq \lceil \frac{n}{3} \rceil$ all words p of length n together with all $q \in QQPer$ where $h(p,q) = d$. These lists have been bounded by time capacity first to $n < 30$, later on to $n \leq 32$. We found that for $2 \leq n < 13$, $m_n = \lceil \frac{n}{3} \rceil$ is true with witness words $a^{\lceil \frac{n}{3} \rceil} b^{\lfloor \frac{2n}{3} \rfloor}$. We guessed that for $n \geq 13$, $m_n = \lfloor \frac{n+1}{3} \rfloor$ which is by 1 smaller than $\lceil \frac{n}{3} \rceil$ if $n \equiv 1 \pmod 3$. The appropriate witness words seemed to be $a^{\lfloor \frac{n}{3} \rfloor + 2} b^{\lceil \frac{2n}{3} \rceil - 2}$. For instance, if $n = 13$ then $p = a^6 b^7$ and there exists a single $q \in QQPer$ with $h(p,q) = 4$, namely $q = ab^3a^2b^3ab^3$. This guess was confirmed by Péter Burcsi's lists for $n \leq 32$ without $n = 31$. He found that for $n = 31$ there is a single word p of length n beginning with a such that again $d(p, QQPer) = \lfloor \frac{n+1}{3} \rfloor + 1 = \lceil \frac{n}{3} \rceil$. It is $p = a^{11}ba^4b^{15}$ with $h(p,q) = 11$ for $q = a^3bab^3a^3ba^4bab^3a^3bab^3a^2 \in QQPer$ and $hhroot(q) = a^3bab^3a^3ba$.

Because of the rather complicated structure of words with the presumable greatest distance between HHQ and $QQPer$ and their hyperhyperroots we could not yet found for each $n \geq 2$ the exact values m_n and appropriate witness words with the proof of (2). To solve this problem may be a great challenge for researchers in combinatorics on words and number theory.

Acknowledgement. The author is very grateful to Péter Burcsi in Budapest and to Georg Lohmann in Apolda for their interest and the assistance with their computer programs.

References

[1] P. Burcsi, Personal communication (May 2012, November 2013).

[2] R. W. Hamming, Error detecting and error correcting codes, *Bell System Techn. Journ.* 29 (1950), 147–160.

[3] P. Dömösi, M. Ito, *Context-Free Languages and Primitive Words*, World Scientific, Singapore, 2014.

[4] M. Ito, G. Lischke, Generalized periodicity and primitivity for words, *Math. Log. Quart.* 53 (2007), 91–106.

[5] G. Lischke, The primitivity distance of words, in: M. Ito, Y. Kobayashi, K. Shoji (Eds.), *Automata, Formal Languages and Algebraic Systems, Proceedings of AFLAS 2008*, World Scientific, 2010, 125–137.

[6] G. Lischke, Root clustering of words, *RAIRO-Theor. Inf. Appl.* 48 (2014), 267–280.

[7] G. Lohmann, Personal communication (August 2011).