CDMTCS
Research
Report
Series

Prefix-Free Łukasiewicz Languages

Ludwig Staiger
Institut für Informatik,
Martin-Luther-Universität
Halle-Wittenberg,
Germany

CDMTCS-298
January 2007

Centre for Discrete Mathematics and
Theoretical Computer Science

Prefix-Free Łukasiewicz Languages

Ludwig Staiger
Institut für Informatik
Martin-Luther-Universität Halle-Wittenberg
D - 06099 Halle, Germany
email: ludwig.staiger@informatik.uni-halle.de

Abstract

Generalised Łukasiewicz languages are simply described languages having good information-theoretic properties. An especially desirable property is the one of being a prefix code. This paper addresses the question under which conditions a generalised Łukasiewicz language is a prefix code. Moreover, an upper bound on the delay of decipherability of a generalised Łukasiewicz language is derived.

Keywords: Prefix codes; Delay of decipherability; Deterministic context-free languages.
1991 Mathematics Subject Classification: 68Q45, 94A17

1 Introduction

The Łukasiewicz language over a two letter alphabet (see [1]) is the context-free language defined by the grammar $S \to aSS \mid b$. It is a simple deterministic language (cf. [5]), thus prefix-free. In the papers [6, 9] this language was generalised in two ways: First languages generated by grammars $S \to aS^n \mid b$ with $n \geq 2$ were admitted. The languages thus specified are also simple deterministic and prefix-free. In what follows we will not necessarily require $n \geq 2$. Unless specified otherwise, our results will hold for arbitrary $n \in \mathbb{N}$.

Secondly, we substitute the letters of a and b by languages C and B. This results in a language defined by the equation

$$Ł = C \cup B \cdot Ł^n \tag{1}$$

These languages Ł are—depending on the languages C and B—not necessarily context-free and will be called, in the sequel, (C, B)-n-Łukasiewicz-languages or, simply, generalised Łukasiewicz languages.

In [9] we confined to the case where $C \cup B$ is a code and $C \cap B = \emptyset$. In this case, the resulting (C, B)-n-Łukasiewicz-language Ł is also a code, but not necessarily prefix-free. The aim of this paper is to give some conditions under which the resulting code Ł is prefix-free. Moreover, a formula for the delay of decipherability of Ł depending on those of C and B is derived.

Next we introduce the notation used throughout the paper. By $\mathbb{N} = \{0, 1, 2, \ldots\}$ we denote the set of natural numbers. Let X be an alphabet of cardinality $|X| = r \geq 2$. By X^* we denote the set (monoid) of words on X, including the *empty word* e. For $w, v \in X^*$ let $w \cdot v$ be their *concatenation*. This concatenation product extends in an obvious way to subsets $W, V \subseteq X^*$. For a language W let $W^* := \bigcup_{i \in \mathbb{N}} W^i$ be the *submonoid* of X^* generated by W. Furthermore $|w|$ is the *length* of the word $w \in X^*$ and $\mathbf{A}(W)$ is the set of all finite prefixes of strings in $W \subseteq X^*$. We shall abbreviate $w \in \mathbf{A}(v)$ by $w \sqsubseteq v$.

As usual language $W \subseteq X^*$ is called a *code* provided $w_1 \cdots w_l = v_1 \cdots v_k$ for $w_1, \ldots, w_l, v_1, \ldots, v_k \in W$ implies $l = k$ and $w_i = v_i$.

A code W is called a *prefix code* provided $v \sqsubseteq w$ implies $v = w$ for $v, w \in W$.

2 Generalised Łukasiewicz languages

In this section we first present some general properties of the languages defined by Eq. (1) which can be found in Section 2 of [9]. Then we give a simple sufficient condition for the prefix-freeness of Łukasiewicz languages.

Lemma 1 *Let* $C, B \subseteq X^* \setminus \{e\}$ *and* Ł *be defined by Eq. (1).*

1. $Ł \subseteq C \cup B \cdot (C \cup B)^* \cdot C^n \subseteq (C \cup B)^*$
2. *Let* $C \cap B = \emptyset$. *If* $C \cup B$ *is a code then* Ł *is a code, and if* $C \cup B$ *is a prefix code then* Ł *is also a prefix code.*
3. *If* $w \in (B \cup C)^i$ *and* $v_1, \ldots, v_{i \cdot n} \in Ł$ *then* $w \cdot v_1 \cdots v_{i \cdot n} \in Ł^*$.
4. $\mathbf{A}(Ł^*) = \mathbf{A}((C \cup B)^*)$

We prove only 1. and 3, the other properties are proved in Proposition 2.1 of [9].
Proof. We have $Ł = \bigcup_{i=0}^{\infty} Ł_i$ where $Ł_0 := C$ and $Ł_{i+1} := Ł_i \cup B \cdot Ł_i^n$. Then one easily verifies by induction on i that $Ł_i \subseteq C \cup B \cdot (C \cup B)^* \cdot C^n$.

For the proof of 3. we show by induction on i that the assertion holds for every $w \in (B \cup C)^i$.

If $w \in (B \cup C)^0 = \{e\}$ then $w \in Ł^*$. Assume $w \in (B \cup C)^{i+1}$. Then $w = v \cdot u$ for $v \in (B \cup C)$ and $u \in (B \cup C)^i$. By the induction hypothesis, $u \cdot v_1 \cdots v_{i \cdot n} \in Ł^m$ for suitable $m \in \mathbb{N}$. Consequently, $u \cdot v_1 \cdots v_{(i+1) \cdot n} \in Ł^{m+n}$ has a decomposition $u \cdot v_1 \cdots v_{(i+1) \cdot n} = u_1 \cdots u_n \cdot u'$ where $u_j \in Ł$ and $u' \in Ł^m$.

If $v \in C$ then $v \cdot u \cdot v_1 \cdots v_{(i+1)\cdot n} \in Ł^{m+n+1}$ and the assertion is true. If $v \in B$ then $v \cdot u_1 \cdots u_n \in Ł$ whence $v \cdot u_1 \cdots u_n \cdot u' \in Ł^*$ and the assertion is also true. ❑

Next we are going to investigate, under which conditions the Łukasiewicz language Ł is a prefix code. It is known that Ł might be a prefix code if $C \cup B$ is not a prefix code.

It is obvious that Ł cannot be a prefix code unless C is a prefix code and $\mathbf{A}(B \cdot Ł^n) \cap C = \emptyset = \mathbf{A}(C) \cap B \cdot Ł^n$. The following theorem shows that in addition to these necessary conditions it suffices to require that B be a prefix code.

Lemma 2 *Let* C, B *be prefix codes and* Ł *be defined by the equation* $Ł = C \cup B \cdot Ł^n$. *If* $\mathbf{A}(B \cdot Ł^n) \cap C = \emptyset = \mathbf{A}(C) \cap B \cdot Ł^n$ *then* Ł *is a prefix code.*

Proof. Assume $w, w' \in Ł$ are words with $w \sqsubset w'$ and $|w| + |w'|$ be minimal. From the hypothesis it follows immediately that $w, w' \notin C$.

Then $w = v \cdot w_1 \cdots w_n$ and $w' = v' \cdot w'_1 \cdots w'_n$ where $v, v' \in B$ and $w_j, w'_j \in Ł$. Since B is a prefix code, $v = v'$. Thus $w_j \sqsubset w'_j$ or $w'_j \sqsubset w_j$ for some $j = 1, \ldots, n$ contradicting the minimality of $|w| + |w'|$. ❑

Example 2.2 of [9] shows that the converse of Lemma 2 is not true, that is, Ł might be a prefix code without requiring that B be a prefix code.

Moreover, the following examples show that the conditions $\mathbf{A}(B \cdot Ł^n) \cap C = \emptyset$ and $\mathbf{A}(C) \cap B \cdot Ł^n = \emptyset$ are likewise independent under the assumptions that $C \cup B$ is a code, $C \cap B = \emptyset$ and C, B are prefix codes.

Example 3 Consider $B := \{aa\}$ and $C := \{aab, bba\}$ and $Ł = C \cup B \cdot Ł^2$. Then $C \cup B$ is a code, C and B are prefix codes, $C \cap B = \emptyset$ and $\mathbf{A}(C) \cap B \cdot Ł^n = \emptyset$ but Ł is not a prefix code because $C \subseteq \mathbf{A}(B \cdot Ł^n)$. ❑

Example 4 Let $C := \{aab, aaabaabb\}$, $B := \{a\}$ and $Ł = C \cup B \cdot Ł^2$. Then $C \cup B$ is a code, C and B are prefix codes, $C \cap B = \emptyset$ and $\mathbf{A}(B \cdot Ł^n) \cap C = \emptyset$ but $a \cdot aab \cdot aab \sqsubseteq aaabaabb$, and Ł is not a prefix code. ❑

3 Factorisation and the Delay of Decipherability

In this section we investigate the delay of decipherability of $Ł \subseteq X^*$. To this end we consider factorisations of words $w \in W^*$. As usual, a tuple $(w_1, \ldots, w_l)$ is called a *W-factorisation* of w provided $w = w_1 \cdots w_l$. For the sake of brevity we shall say that w has the W-factorisation $w = w_1 \cdots w_l$. Thus, a language W is a code if and only if every word $w \in X^*$ has at most one W-factorisation.

The deciphering (or decoding) of a message $w \in W^*$ consists in finding the (unique) W-factorisation $w = w_1 \cdots w_l$ of w. If we consider the deciphering as a parsing process proceeding from left to right we are confronted with the task, given a prefix $w' \sqsubseteq w$, to estimate as soon as possible valid W-factorisations of this prefix w'. This can be based only on the possible W-factorisations of prefixes $w'' \sqsubseteq w'$.

According to [2, 3, 4, 8], a code W has a *delay of decipherability (or deciphering delay)* $d \geq 0$ if and only if for all $v, v' \in W$ the relation $v \cdot v_1 \cdots v_d \sqsubseteq v' \cdot u$ where $v_1, \ldots, v_d \in W$ and $u \in W^*$ implies $v = v'$. For the above mentioned parsing process this means that once we have a W-factorisation of a prefix $w'' \sqsubseteq w$ into $d+1$ factors $w'' = v \cdot v_1 \cdots v_d$ then the first factor v is definitively valid, that is, every W-factorisation of w starts with the factor $v \in W$.

Observe that a code $W \subseteq X^*$ of delay of decipherability d satisfies

$$W^{d+1} \cap \mathbf{A}(W) = \emptyset\,. \tag{2}$$

We say that a code W has a *bounded delay of decipherability* provided W has a delay of decipherability d for some $d \in \mathbb{N}$. In particular, W is a prefix code iff W has a delay of decipherability 0.

A further generalisation of the delay of decipherability is the *delay function* $d_W : W \to \mathbb{N} \cup \{\infty\}$ (see [3, 4, 8]) defined as follows. First, we say that a word $v \in W$ has a delay of decipherability of d with respect to $W \subseteq X^*$ provided for all $v' \in W$ the relation $v \cdot v_1 \cdots v_d \sqsubseteq v' \cdot u$ where $v_1, \ldots, v_d \in W$ and $u \in W^*$ implies $v = v'$. Cast into words of our parsing process this means that once we have a W-factorisation of a prefix $w'' \sqsubseteq w' \sqsubseteq w$ into $d+1$ factors starting with the factor $v \in W$ this factor v is definitively valid. Now define $d_W : W \to \mathbb{N} \cup \{\infty\}$ as follows.

$$d_W(w) := \inf\{d : w \text{ has a delay of decipherability of } d \text{ w.r.t. } W\} \tag{3}$$

Thus W has a delay of decipherability of d iff $d_W(w) \leq d$ for all $w \in W$, and W has a bounded delay of decipherability iff the function d_W is bounded on W. We say that a code W has a *finite delay of decipherability*[1], provided the function d_W maps W to $\mathbb{N}$, that is, has never the value ∞.

We conclude this section with a theorem on Łukasiewicz languages derived from codes having a bounded delay of decipherability.

Theorem 5 *Let C be a prefix code, $\mathbf{A}(B \cdot Ł^n) \cap C = \emptyset$ and let $C \cup B$ be a code having a delay of decipherability of at most n. Then the generalised Łukasiewicz language Ł defined by $Ł = C \cup B \cdot Ł^n$ is a prefix code.*

Proof. Assume the contrary, that is, there are $v, v' \in Ł$ such that $v \sqsubset v'$ and assume $|v| + |v'|$ to be minimal. It is obvious that $v \notin C$. Then $v = u \cdot w_1 \cdots w_n$ with $u \in B$ and $w_i \in Ł \subseteq (C \cup B)^* \setminus \{e\}$. Now, Eq. (2) shows that $v' \notin C$.

Consequently, $v' = u' \cdot w_1' \cdots w_n'$ where $u' \in B$ and $w_i' \in Ł \subseteq (C \cup B)^* \setminus \{e\}$. Since $C \cup B$ has a delay of decipherability of n this implies $u = u'$ which in turn implies, that there is a j, $1 \leq j \leq n$, such that $w_j \sqsubset w_j'$ or $w_j' \sqsubset w_j$, contradicting the minimality of $|v| + |v'|$. ❏

The subsequent examples show, on the one hand, that the bound n in Theorem 5 cannot be improved but, on the other hand, that there are cases where it is not tight.

[1]The reader is warned that the book [2] uses the term finite delay of decipherability to denote codes of bounded delay of decipherability.

The first example is a code $C \cup B$ of delay of decipherability $n + 1$ for which $Ł = C \cup B \cdot Ł^n$ is not a prefix code. Its construction is similar to the one of Theorem 5.1 in [4].

Example 6 Let $B := \{b, bab\}$, $C := \{a^j b a^{j+1} b : 1 \le j \le 2n\}$ and $Ł = C \cup B \cdot Ł^n$. Then C is a prefix code and $C \cap \mathbf{A}(B \cdot Ł^n) = \emptyset$.

Moreover, $C \cup B$ has a delay of decipherability $n + 1$, for the words $w_1 := b \cdot aba^2b \cdots a^{2n-1}ba^{2n}b$ and $w_2 := bab \cdot a^2ba^3b \cdots a^{2n}ba^{2n+1}b$ are the words having the longest $(C \cup B)$-factorisations in which the first factors differ. Since $w_1, w_2 \in B \cdot C^n$, this shows also that Ł is not a prefix code. ❑

The second example is a code $C \cup B$ having a delay of decipherability of d for which Ł, independently of $n \ge 1$, is a prefix code.

Example 7 Let $X = \{a, b\}$, $C := \{a^d b\}$ and $B := \{a\}$. Then $C \cup B$ has a delay of decipherability of d and $Ł = C \cup B \cdot Ł^n$ $(n \ge 1)$ is a prefix code. ❑

4 The Delay of Decipherability of Łukasiewicz Languages

In the preceding section we observed that the construction of Eq. (1) might turn non-prefix codes into prefix codes. By a more subtle consideration of the delay function $d_{C \cup B}$ for specific words we are able to obtain a result more general than Theorem 5. First we mention the following result, which is a simple consequence of the fact that the $(C \cup B)$-factorisation of a word $w \in Ł^*$ is longer than its Ł-factorisation.

Corollary 8 *Let $C \cup B$ a code, $C \cap B = \emptyset$ and $Ł = C \cup B \cdot Ł^n$. Then $d_Ł(w) \le d_{C \cup B}(w)$, for every $w \in C$.*

The aim of this last section is to investigate in more detail the delay function of decipherability of generalized Łukasiewicz languages. Before proceeding to the main result of this section we need an auxiliary property of Łukasiewicz languages.

Proposition 9 *Let $C \cup B$ be a code, $C \cap B = \emptyset$, $Ł = C \cup B \cdot Ł^n$ and let $w, v \in Ł$ having $(C \cup B)$-factorisations $w = w_1 \cdots w_l$ and $v = w_1 \cdots w_k$ with $l \le k$, respectively. Then $k = l$.*

Proof. If $l < k$ by Lemma 1.3 we have $w_{l+1} \cdots w_k \cdot u^{(k-l) \cdot n} \in Ł^*$, for $u \in C \subseteq Ł$. Consequently, the word $w' = w \cdot (w_{l+1} \cdots w_k \cdot u^{(k-l) \cdot n}) = v \cdot (u^{(k-l) \cdot n})$ has two different Ł-factorisations, which contradicts Lemma 1.2. ❑

After this preparation we derive the main theorem of this section.

Theorem 10 *Let $C \cup B$ be a code, and let $d_{C \cup B} : C \cup B \to \mathbb{N} \cup \{\infty\}$ its delay function. If $Ł = C \cup B \cdot Ł^n$ and $w \in Ł$ has a $(C \cup B)$-factorisation $w = w_0 \cdot w_1 \cdots w_l$ then*

$$d_Ł(w) \le \max\{d_{C \cup B}(w_i) - l + i : 0 \le i \le l\} \cup \{0\}.$$

Proof. For $w \in C$ the assertion is Corollary 8.

Let $w \in B \cdot Ł^n$ and consider its $(C \cup B)$-factorisation $w = w_0 \cdot w_1 \cdots w_l$. We proceed by induction on i. Let $d := \max\{d_{C\cup B}(w_i) - l + i : 0 \le i \le l\} \cup \{0\}$ and consider the relation $w \cdot (u_1 \cdots u_d) \sqsubseteq v \cdot u$ where $u_i, v \in Ł$ and $u \in Ł^*$.

Since $Ł \subseteq (C \cup B)^*$, we may consider the $(C \cup B)$-factorisations $w_0 \cdot w_1 \cdots w_l \cdot (u_1 \cdots u_d) \sqsubseteq v_0 \cdot v_1 \cdots v_k \cdot u$ where $v = v_0 \cdot v_1 \cdots v_k$ is the $(C \cup B)$-factorisation of $v \in Ł$. From $d_{C\cup B}(w_0) \le d + l$ we obtain $w_0 = v_0$.

Now, canceling equal factors from the left, we proceed by induction and obtain $w_i = v_i$ for $i \le \min\{l, k\}$. Then $w = v$ follows from Proposition 9. ❑

With Lemma 1.1 we obtain the following corollary.

Corollary 11 *Let* $C \cup B$ *be a code of a delay of decipherability* d, $C \cap B = \emptyset$ *and* $Ł = C \cup B \cdot Ł^n$. *Moreover let* $d_{C\cup B}(v) \le d'$ *for all* $v \in C$. *Then* $Ł$ *has a delay of decipherability* $d_Ł \le \max\{d', d - n\}$.

Proof. It suffices to show that every $w \in B \cdot Ł^n$ has $d_Ł(w) \le d - n$. To this end consider a $C \cup B$-factorisation $w = w_0 \cdot w_1 \cdots w_l$. According to Lemma 1.1 the last n words $w_{l-n+1}, \ldots w_l$ are in C. Now the assertion follows from Theorem 10. ❑

Again Examples 6 and 7 show, on the one hand, that the bound in Theorem 10 and Corollary 11 cannot be improved but, on the other hand, that there are cases where it is not tight.

Example 6. (continued) We have $d_{C\cup B}(v) = 0$ for $v \in C$, $d_{C\cup B}(bab) = n$ and $d_{C\cup B}(b) = n + 1$. This shows $d_Ł = 1$ because Ł is not a prefix code. ❑

Example 7. (continued) Here $d_{C\cup B}(a^d b) = 0$ and $d_{C\cup B}(a) = d$, but $Ł = C \cup B \cdot Ł^n$ $(n \ge 1)$ is a prefix code. ❑

Concluding Remark

It was observed in [6, 9] that Łukasiewicz languages have, on the one hand, remarkable information-theoretic properties and are, on the other hand, simply to describe (cf. [1, 5]). A simpler class of languages is the class of languages definable by finite automata which does not possess those properties.

In the papers [10, 11] the information-theoretic properties of generalised Łukasiewicz languages were employed to demonstrate in a (computationally or language-theoretic) simple way the non-coincidence of Hausdorff dimension and Hausdorff measure for sets defined by infinite iterated function systems and their closures. Here mainly Łukasiewicz languages Ł derived from prefix codes $C \cup B$ were used in order to obtain iterated function systems with nice topological properties. The present paper shows that it is possible to obtain prefix-free Łukasiewicz languages also from non-prefix codes. This might hint to further examples in the sense of the papers [10, 11].

The preceding section (Corollaries 8 and 11 and Theorem 10) shows that at the cost of a slightly more complicated internal structure of a generalised Łukasiewicz language its deciphering behaviour (as a code) might become less complicated.

References

[1] J.-M. Autebert, J. Berstel and L. Boasson, Context-Free Languages and Pushdown Automata, in: [7], Vol. 1, pp. 111–174.

[2] J. Berstel and D. Perrin. *Theory of Codes*. Academic Press, 1985.

[3] J. Devolder, M. Latteux, I. Litovsky and L. Staiger. Codes and Infinite Words. *Acta Cybernetica*, 11:241–256, 1994.

[4] H. Fernau, K. Reinhardt and L. Staiger. Decidability of Code Properties, in: *Developments in Language Theory IV* (G. Rozenberg and W. Thomas eds.), World Scientific, Singapore, pp. 153 – 163, 2000.

[5] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata and Language Theory*, Addison-Wesley, Reading MA, 1979.

[6] W. Kuich. On the Entropy of Context-free Languages. *Inf. & Contr.*, 16:173–200, 1970.

[7] G. Rozenberg and A. Salomaa (eds.) *Handbook of Formal Languages*, Springer, 1997.

[8] L. Staiger. On Infinitary Finite Length Codes, *Theoret. Informatics & Applications*, 20:483–494, 1986.

[9] L. Staiger. The Entropy of Łukasiewicz-languages, *Theoret. Informatics & Applications*, 39:621–639, 2005.

[10] L. Staiger, Infinite Iterated Function Systems in Cantor Space and the Hausdorff Measure of ω-power Languages, *Intern. J. Found. Comput. Sci.* 16:787–802, 2005.

[11] L. Staiger. Hausdorff Measure and Łukasiewicz Languages, *J. of Universal Computer Science*, 11:2114–2124, 2006.