A Quantitative Approach to the Primitive Words Conjecture

Ryoma Sin'ya
(Akita University)

Computational Logic and Applications
12 Oct 2020

Outline

1. The primitive words conjecture
2. Quantitative properties of Q
3. Conclusion and open problems

Outline

1. The primitive words conjecture
2. Quantitative properties of Q
3. Conclusion and open problems

The Primitive Words Conjecture

[Dömösi-Horvath-Ito 1991]

- A non-empty word w is said to be **primitive** if it can not be represented as a power of shorter words, i.e., $w = u^n \Rightarrow u = w$ (and $n = 1$).
  Q_A denotes the set of all primitive words over A.

- Here after we only consider the case $A = \{a, b\}$ for Q_A, and simply write Q.

Example： $ababa \in Q$ $ababab = (ab)^3 \notin Q$

Conjecture: Q is not context-free.

Why is "primitivity" important?

- Primitive words are like prime numbers.
  Fact: For every non-empty word w, there exists a unique primitive word v such that $w = v^k$ for some $k \geq 1$.

$$A^* = \{\varepsilon\} \uplus \mathrm{Q} \uplus \mathrm{Q}^{(2)} \uplus \mathrm{Q}^{(3)} \uplus \cdots$$

$$\text{where } \mathrm{Q}^{(n)} = \{w^n \mid w \in \mathrm{Q}\}$$

Why is "primitivity" important?

- Primitive words are like prime numbers.
  Fact: For every non-empty word w, there exists a unique primitive word v such that $w = v^k$ for some $k \geq 1$.

- For a word $w = uv$, we denote its *conjugate* (by u) vu by $u^{-1}wu = vu$.
  If u and v are non-empty, $u^{-1}wu$ is called a *proper* conjugate.
  Fact: w is primitive $\Leftrightarrow$ $w \neq u^{-1}wu$ for every proper conjugate.

  Note: if we regard a conjugation as a (partial) morphism on words, "w is primitive" means "w has no non-trivial automorphism" (cf. rigid graphs, rigid models in model theory) .

Why is “primitivity” important?

- Primitive words are like prime numbers.
  Fact: For every non-empty word w, there exists a unique primitive word v such that $w = v^k$ for some $k \geq 1$.

- For a word $w = uv$, we denote its *conjugate* (by u) vu by $u^{-1}wu = vu$.
  If u and v are non-empty, $u^{-1}wu$ is called a *proper* conjugate.
  Fact: w is primitive ⇔ $w \neq u^{-1}wu$ for every proper conjugate.

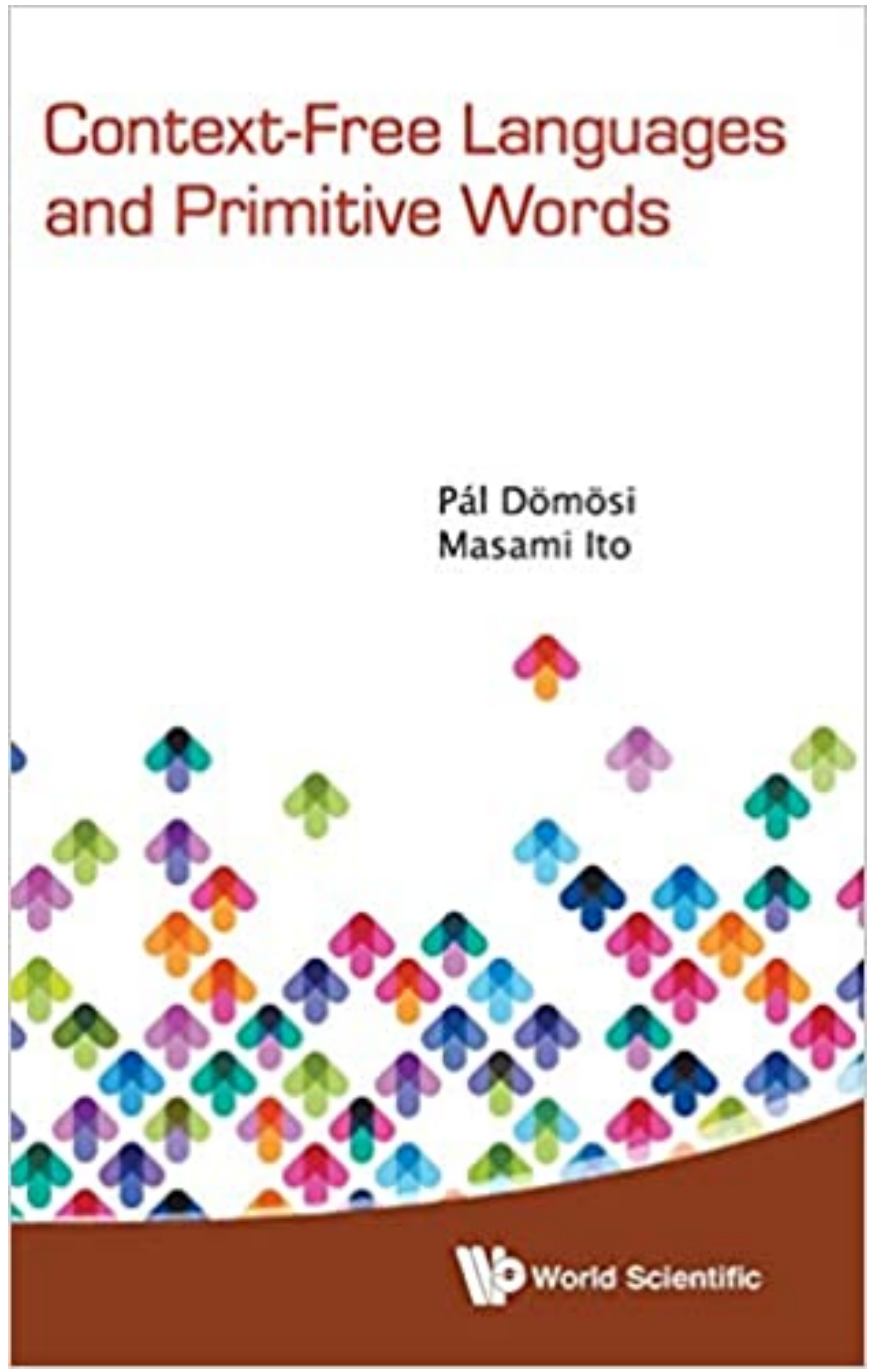
- Primitive words and its special class called *Lyndon words* play a central role in algebraic coding theory and combinatorics on words, also in text compression (cf. Lyndon factorisation, Burrows–Wheeler transformation).

The Primitive Words Conjecture

[Dömösi-Horvath-Ito 1991]

Masami Ito

Pál Dömösi



[Dömösi-Ito 2014]

The Primitive Words Conjecture

[Dömösi-Horvath-Ito 1991]

Masami Ito

Pál Dömösi

Szilárd Fazekas

Known approaches

- Generating function method: it is known that Q is not an *unambiguous* context-free language.
  However, no "good theory" of generating functions of **general** context-free languages is known.

Note: The generating function of every unambiguous context-free language is algebraic (Chomsky-Schützenberger), while the generating function of Q:

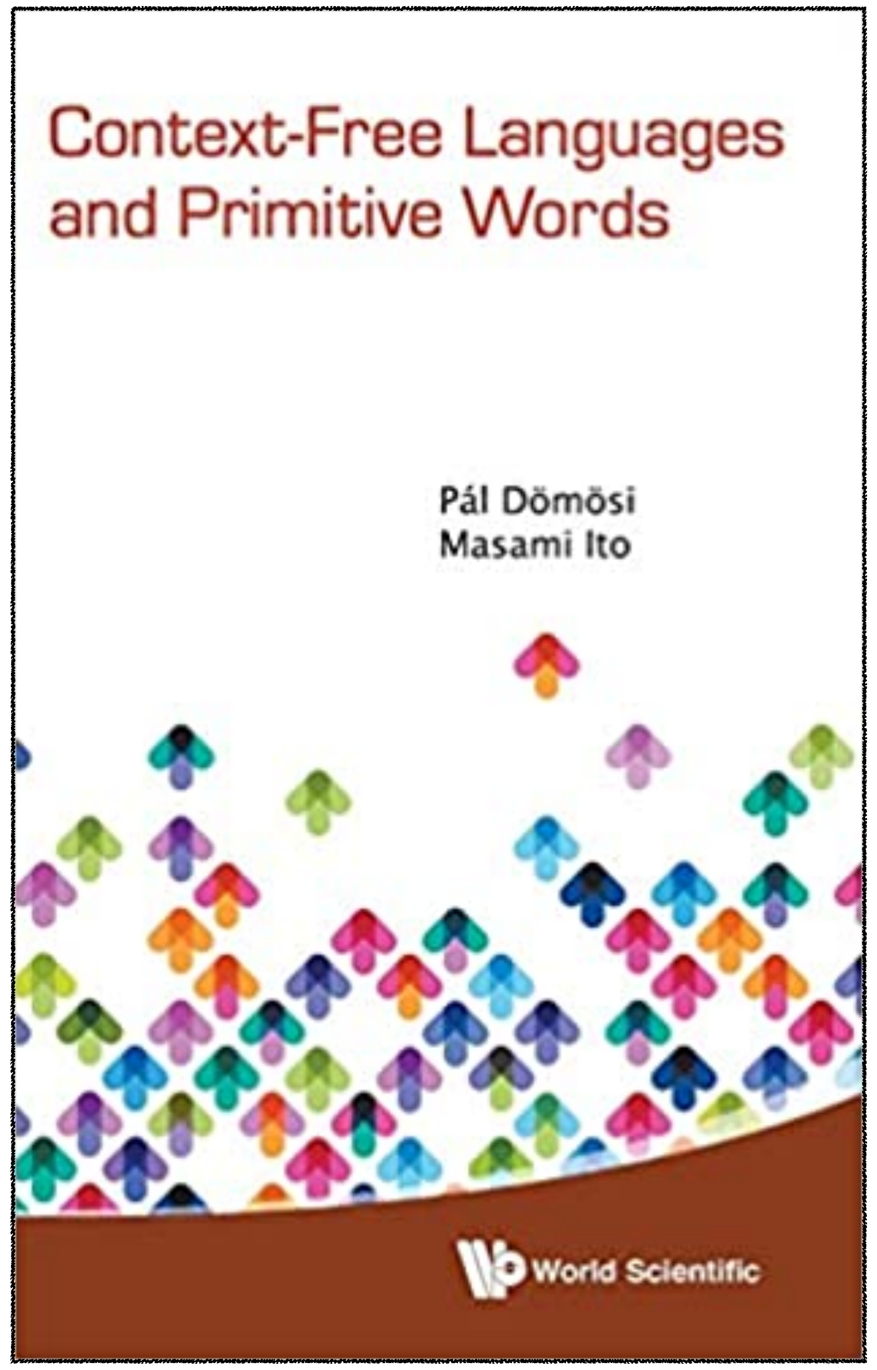
$$\sum_{n=0}^{\infty} \#(Q \cap A^n)\, z^n = \sum_{n=0}^{\infty} \left(\sum_{d|n} \mu(d) 2^{n/d} \right) z^n$$ is not algebraic (cf. [Petersen 1994]).

Here $d \mid n$ means "d divides n" and μ is the classical Möbius function

Known approaches

- Generating function method: it is known that Q is not an *unambiguous* context-free language.
  However, no “good theory” of generating functions of **general** context-free languages is known.
- Constructing a regular language R such that $Q \cap R$ is not context-free:

  By some results of L. Kászonyi and M. Katsura, this approach also seems to be hopeless (cf. Kászonyi-Katsura theory).

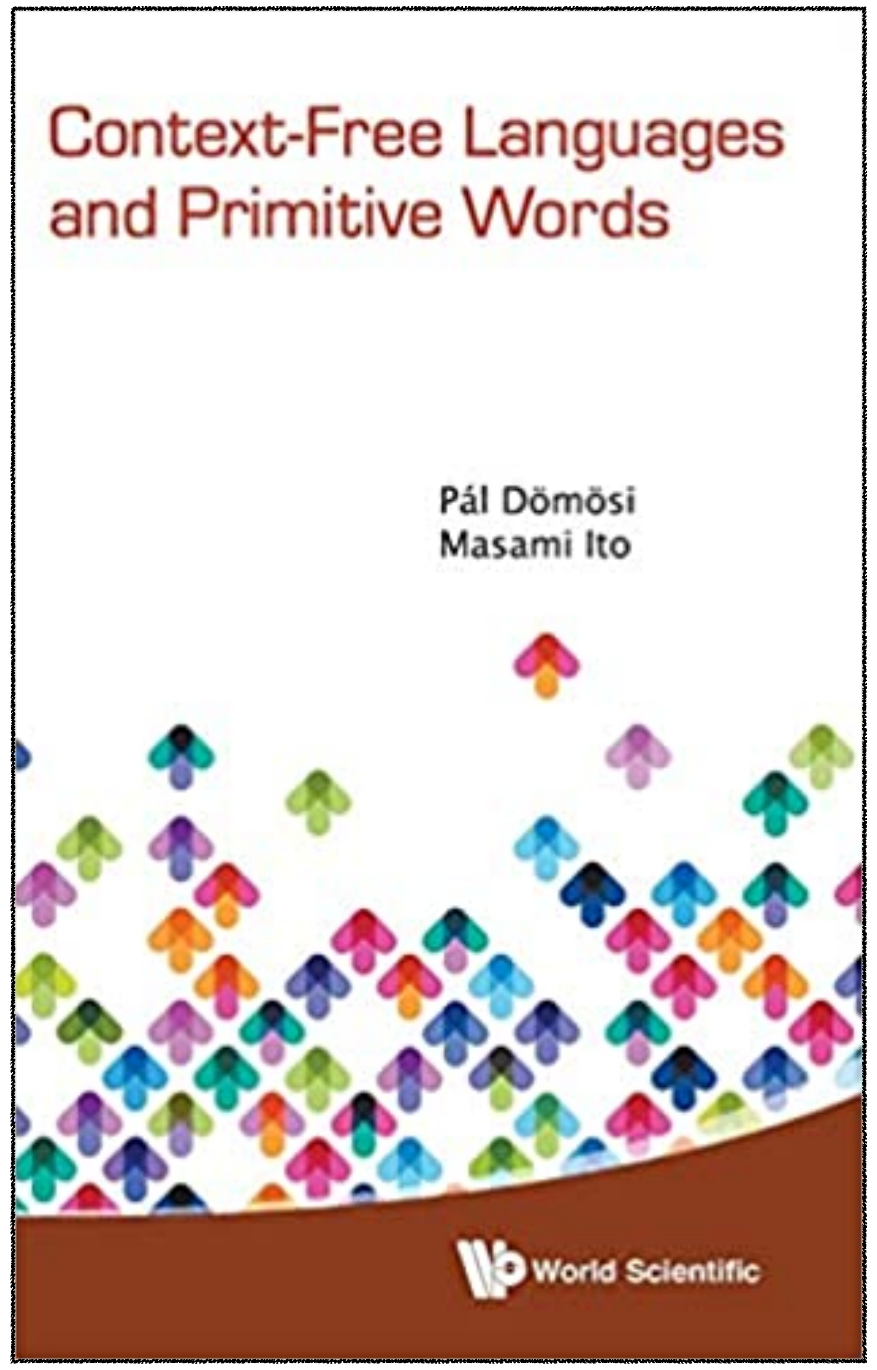


Note: if L is context-free and R is regular, then $L \cap R$ is always context-free.

Known approaches

- Generating function method: it is known that Q is not an *unambiguous* context-free language.
  However, no “good theory” of generating functions of **general** context-free languages is known.
- Constructing a regular language R such that $\mathrm{Q} \cap R$ is not context-free:

  By some results of L. Kászonyi and M. Katsura, this approach also seems to be hopeless (cf. Kászonyi-Katsura theory).



- Pumping-lemma-like tests:

  Q resists almost all well-known tests of context-freeness.

Outline

1. The primitive words conjecture
2. Quantitative properties of Q
3. Conclusion and open problems

Density of formal languages

- The (*asymptotic*) *density* $\delta_A(L)$ of a language L over A is defined as

$$\delta_A(L) = \lim_{n\to\infty} \frac{\#(L \cap A^n)}{\#(A^n)}$$

Not null: measure theoretic "largeness"
Dense: topological "largeness"

Fact1 (cf. [Berstel 1972]):
If a regular language L has a density, then it is always rational.

Fact2 (cf. [S2]): A regular language L is *not null* (i.e., $\delta_A(L) \neq 0$) if and only if L is *dense* (i.e.,$L \cap A^*wA^* \neq \emptyset$ for any $w \in A^*$).

Note: "L is not null $\Rightarrow$ L is dense" is true for any language L, but
"L is dense $\Rightarrow$ L is not null" is false for general non-regular languages.

Density of formal languages

Note: "L is not null $\Rightarrow$ L is dense" is true for any language L, but
"L is dense $\Rightarrow$ L is not null" is false for general non-regular languages.

Infinite Monkey Theorem (cf. [Borel 1913]): $\delta_A(A^*wA^*) = 1$ for any $w \in A^*$.

L is not dense means that there exists w such that $L \cap A^*wA^* = \emptyset$
(such word is called a *forbidden word* of L),
thus $\delta_A(L) \leq 1 - \delta_A(A^*wA^*) = 0$ by the infinite monkey theorem.

The *semi-Dyck* language $\mathrm{D} = \{\varepsilon, (), (()), ()(), ((())), \ldots\}$ over $A = \{(,)\}$
is dense, but actually null.

$($ $)(()($ $))$

Q is "very large"

Theorem (cf. [S1]): Q is co-null, i,e,. $\delta_A(\mathrm{Q}) = 1.$

Proof: we show that the complement $\overline{\mathrm{Q}}$ (set of non-primitive words) is null.

Because $n \in \mathbb{N}$ has at most $2\sqrt{n}$ divisors and $w = v^m$ $(|w| = n, m \geq 2)$ implies $|v| \leq n/2$, we have $\#(\overline{\mathrm{Q}} \cap A^n) \leq 2\sqrt{n} \cdot \#(A)^{n/2+1}$.

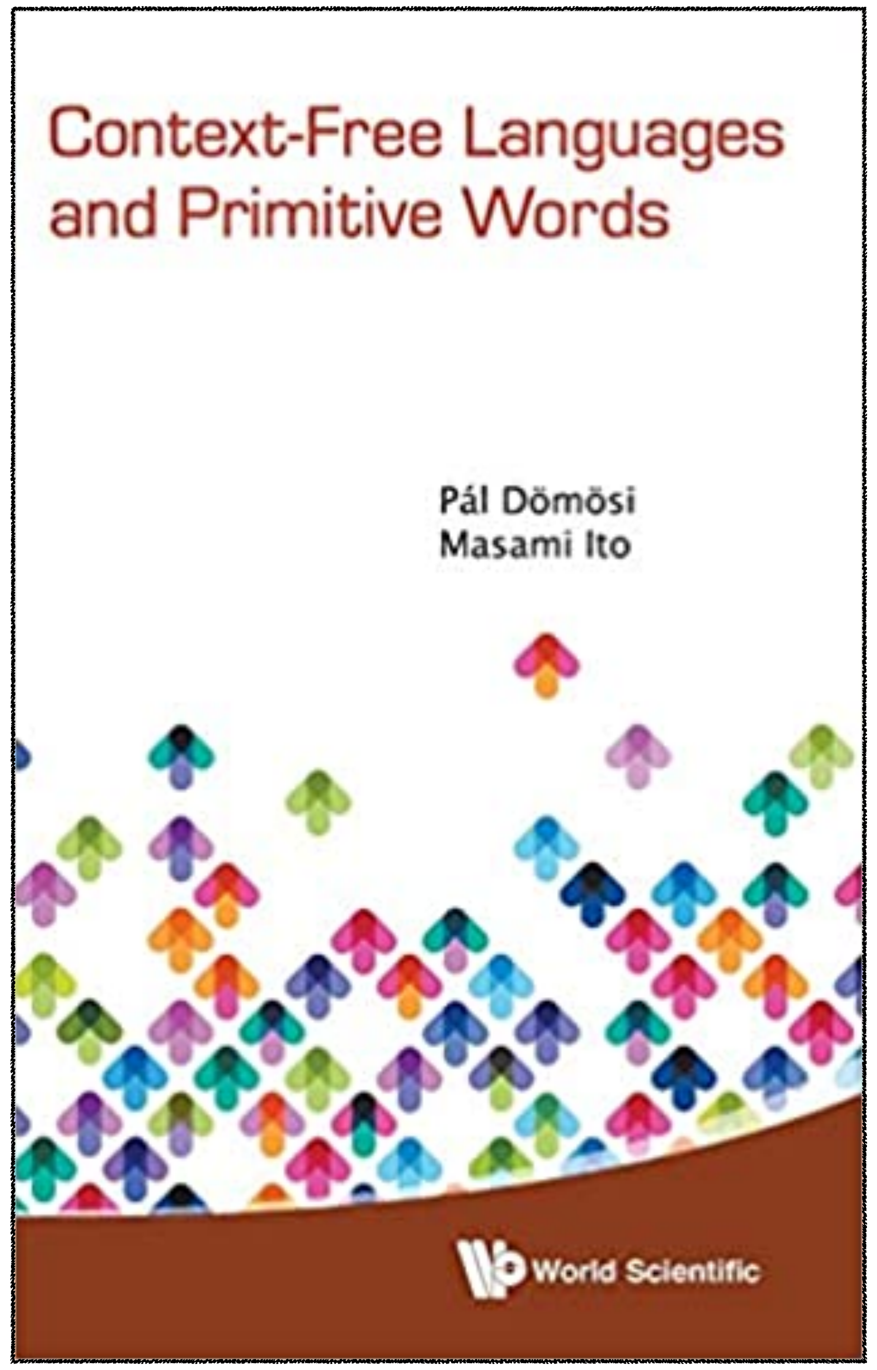
$$\frac{\#(\overline{\mathrm{Q}} \cap A^n)}{\#(A^n)} \leq \frac{2\sqrt{n} \cdot \#(A)^{n/2+1}}{\#(A)^n} \leq \frac{2\sqrt{n}}{2^{n/2-1}} \quad (\rightarrow 0 \text{ if } n \rightarrow \infty).$$

Q is “very large”

Theorem (cf. [S1]): Q is co-null, i,e,. $\delta_A(\mathrm{Q}) = 1.$

- This fact is a rough (but good) intuition that Q fulfills various extensions of pumping-lemma-like test of context-freeness. Because any pumping sequence *can not escape from* Q!!!

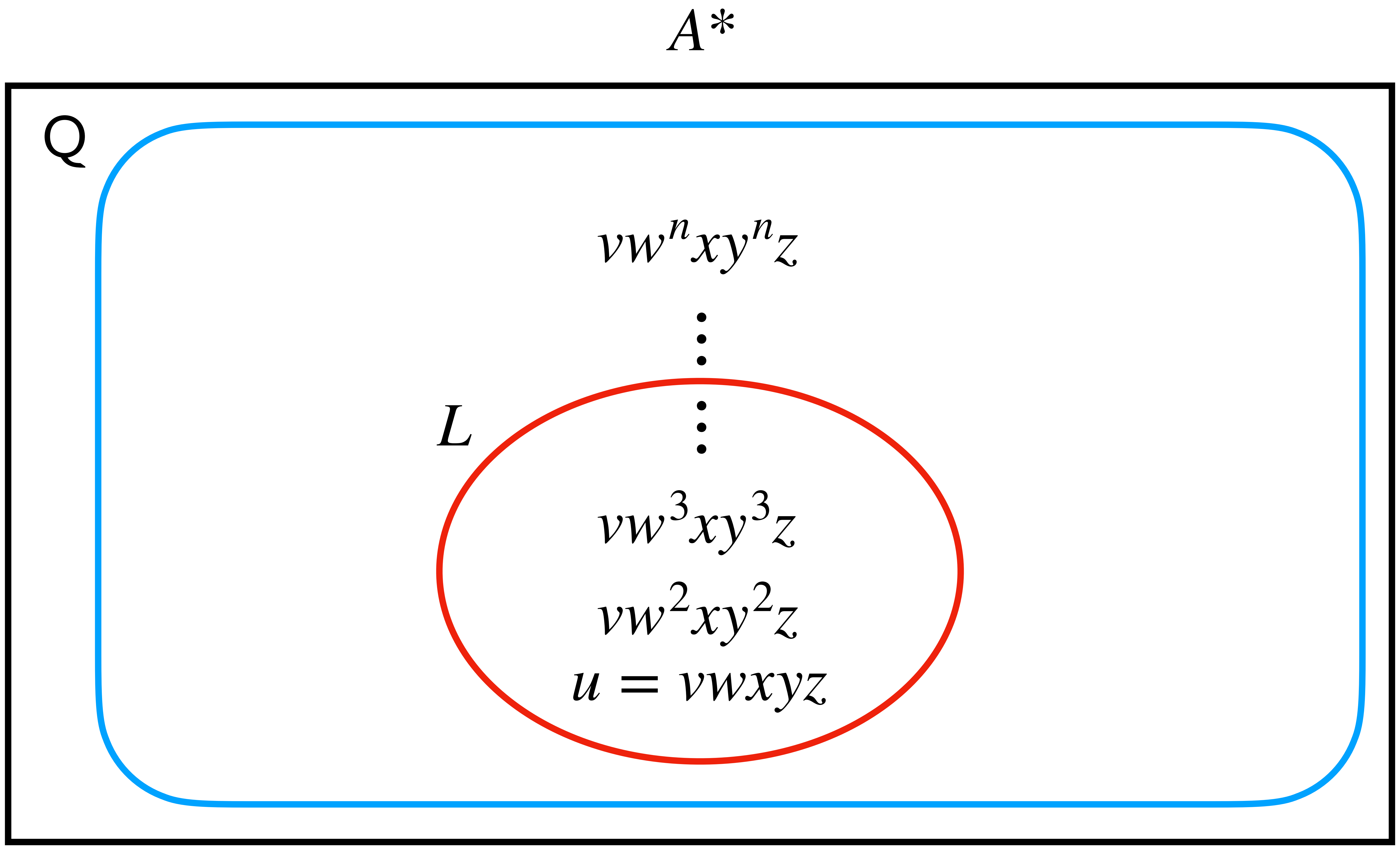
Q resists almost all well-known tests of context-freeness.



[Pumping lemma] for every context-free language L, there exists $p \geq 1$ such that: every word $u \in L$ longer than p can be factorised as $u = vwxyz$ satisfying

(1) $|wy| \geq 1$ (i.e., pumping part is non-empty), (2) $|wxy| \leq p$ and

(3) $vw^ixy^iz \in L$ for every $i \geq 0$ (i.e., every pumping sequence is in L).



L is not context-free!

…but any pumping sequence *can not escape from* Q, since it is very large!

Every regular subset of Q is null

Theorem [S1]: Every non-null regular language contains non-primitive words.

- While Q is very large (i.e., co-null), every regular subset of Q is null.

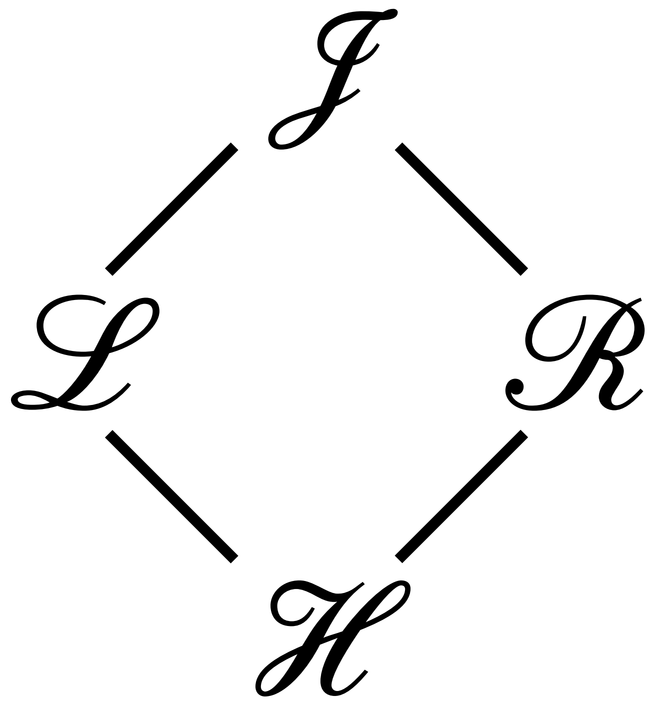
  Intuitively, this means that there is no "*good-lower-approximation of* Q by a regular language".

The proof uses basic semigroup theory: *Green's relations* and *Green's theorem*.

Quick introduction to Green's theorem

Let M be a monoid.

Green's four relations $\mathscr{J}, \mathscr{L}, \mathscr{R}$ and $\mathscr{H}$ are defined as follows:



$$a \mathscr{J} b \Leftrightarrow MaM = MbM$$

$$\Leftrightarrow \exists x, y, x', y' \in M\,[xay = b \wedge x'by' = a]$$

$\Leftrightarrow$ a and b belong to the same *strongly-connected component* in the Cayley graph of M.

$a \mathscr{L} b \Leftrightarrow Ma = Mb$ $\quad$ $a \mathscr{R} b \Leftrightarrow aM = bM$ $\quad$ $a \mathscr{H} b \Leftrightarrow a \mathscr{L} b \wedge a \mathscr{R} b$

Theorem [Green]: Let M be a monoid and a be its element.

$\mathscr{H}_a = \{b \in M \mid a \mathscr{H} b\}$ contains e such that $e = e^2$ $\Leftrightarrow$ $\mathscr{H}_a$ is a subgroup of M

(*idempotent* element) $\quad$ whose identity element is e.

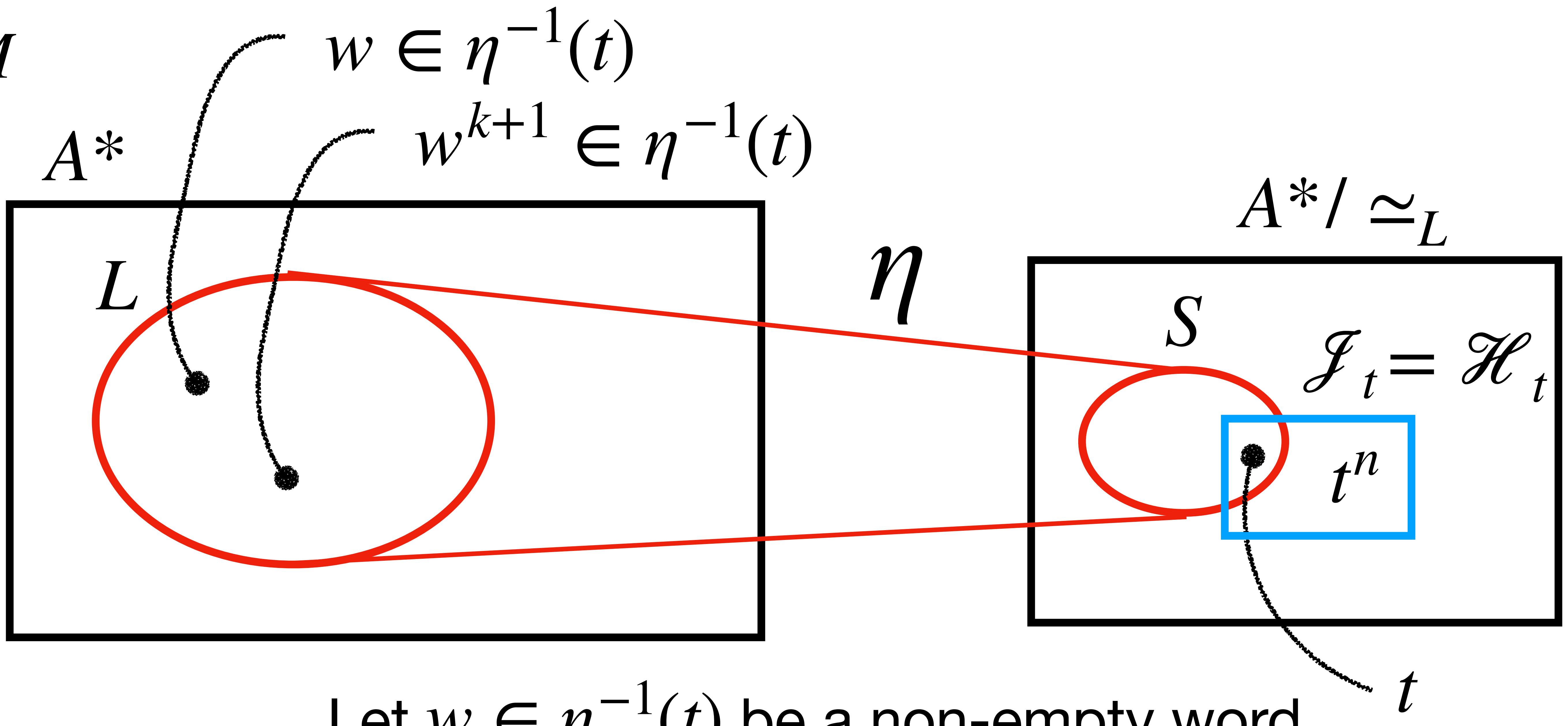
Theorem [S1]: Every non-null regular language contains non-primitive words.

Proof sketch:

Let L be a regular language over A with $\delta_A(L) > 0$.

Let $\eta : A^* \to A^*/\simeq_L$ be the syntactic morphism of L and $S = \eta(L) \subseteq A^*/\simeq_L$ be the image of L (where $\simeq_L$ is the syntactic congruence: $u \simeq_L v$ iff $\forall x, y \in A^*[xuv \in L \Leftrightarrow xvy \in L]$).

Notation: $a \leq_{\mathcal{J}} b \Leftrightarrow MaM \subseteq MbM$



Claim 1

"$\delta_A(L) > 0$" and "$A^*/\simeq_L$ is finite" implies "S contains a $\leq_{\mathcal{J}}$-minimal element t".

Claim 2

"t is $\leq_J$-minimal" implies "$t \mathcal{J} t^n$ for all $n \geq 1$".

Claim 3

"$A^*/\simeq_L$ is finite" and "$t \mathcal{J} t^n$" implies "$t \mathcal{H} t^n$".

Claim 4

"$A^*/\simeq_L$ is finite" implies "$t^k t^k = t^k$ for some k".

By Green's theorem, $\mathcal{H}_t$ is a group with the identity t^k.

Let $w \in \eta^{-1}(t)$ be a non-empty word

$$\eta(w^{k+1}) = \eta(w)^{k+1} = t^{k+1} = t^k t = t \in S$$

Thus $w^{k+1} \in L$

Outline

1. The primitive words conjecture
2. Quantitative properties of Q
3. Conclusion and open problems

Conclusion

- We gave an introduction to the primitive words conjecture, including a short survey of several known approaches and a brief intuition why this problem is hard to solve.

- We also describe a special quantitative property of Q:
  While Q is "very large" (co-null), any regular subset of Q is "very small" (null).

- For tackling this conjecture, I think a study of the theory of "large context-free languages" is important.

Open problems

1. Does every non-null context-free language contain non-primitive words?

   Note: for the regular case, the answer of this problem is “yes” [S1].

2. Does every co-null context-free language contain non-primitive words?

3. Can we give an alternative characterisation of the class of null (resp. co-null) context-free languages?

   Note: there are several different characterisation of the class of null (resp. co-null) regular languages [S2].

(Akita-Inu)

References

- [Berstel 1972] Sur la densité asymptotique de langages formels, *ICALP1972*.
- [Borel 1913] Mécanique Statistique et Irréversibilité, *J. Phys*.
- [Dömösi-Ito 2014] Context-Free Languages And Primitive Words.
- [Dömösi-Horvath-Ito 1991] On the Connection between Formal Languages and Primitive Words.
- [Petersen 1996] On the language of primitive words, *TCS*.
- [S1] Asymptotic Approximation by Regular Languages, *SOFSEM2021* (to appear).
- [S2] An Automata Theoretic Approach to the Zero-One Law for Regular Languages, *GandALF2015*.

The full versions of [S1] and [S2] are all available at
http://www.math.akita-u.ac.jp/~ryoma