hexagon:2610.00709v1

Upper bounds for the Chvátal–Sankoff constants γ_k, $k \geq 3$

Lueker's canonicity machine beyond the binary alphabet

David Niedbala Giraudin
ORCID 0009-0009-1526-1178 — Independent researcher, Meaux, France

20 September 2026
revised 21 September 2026; note added 31 October 2026

Abstract

Lueker (J. ACM 56(3), 2009) constructed a family of finite-state machines $\mathcal{M}_h$ whose associated transfer matrices yield upper bounds on the Chvátal–Sankoff constants γ_k, and proved (his Theorem 3.26) that these bounds converge to γ_k for *every* alphabet size k. He published values only for $k = 2$. For $k \geq 3$ the best upper bounds we have been able to find in the literature are still those of Dančík's 1994 thesis, thirty-two years on: subsequent work generalising Lueker has generalised his Section 2, the lower bound. We implement $\mathcal{M}_h$ for arbitrary k. One ingredient has no binary analogue — the generating function of a truncation class when *both* sides of a minimal match pair are truncated — and we derive it (Lemma 2). We then give a reduction by *relative forms* under which the transition multiplicities of the transfer matrix are falling factorials in k (Lemma 3); one alphabet-free construction consequently determines the exact matrix for every k. We obtain improved upper bounds for $k = 3, \ldots, 15$, each supported by a Collatz–Wielandt certificate verified in exact rational arithmetic.

1 Results

Let $\Sigma = \{0, \ldots, k-1\}$ and let $L_{n,n}$ be the length of a longest common subsequence of two independent uniform random words of length n over Σ. By subadditivity $\mathbb{E}[L_{n,n}]/n \to \gamma_k$ [1]. The constants are not known.

k	h	Dančík 1994	this note	gain	gain
3	4	0.765803	**0.747886276**	0.017917	2.34%
4	4	0.708236	**0.689063947**	0.019172	2.71%
5	4	0.664428	**0.645032787**	0.019395	2.92%
6	3	0.629316	**0.618256822**	0.011059	1.76%
7	3	0.600184	**0.589456726**	0.010727	1.79%
8	3	0.575407	**0.565041150**	0.010366	1.80%
9	3	0.553937	**0.543937275**	0.010000	1.81%
10	3	0.534855	**0.525416942**	0.009438	1.76%
11	3	0.517842	**0.508964025**	0.008878	1.71%
12	3	0.502591	**0.494199896**	0.008391	1.67%
13	3	0.488800	**0.480838934**	0.007961	1.63%
14	3	0.476198	**0.468660628**	0.007537	1.58%
15	3	0.464619	**0.457491357**	0.007128	1.53%

Every bound in the fourth column is certified in the sense of Section 6: it is the value of $2\log(kz_0)/\log\lambda$ for an explicit rational z_0 and an explicit rational λ for which the inequality $A_{z_0}v \le \lambda v$ has been checked coordinatewise in exact rational arithmetic.

2 Background, and why the gap existed

The framework is that of Chvátal and Sankoff [1]: writing $G(n,\ell)$ for the number of pairs of words of total length n whose longest common subsequence has length at least ℓ, if $G(2n, yn) = o(k^{2n})$ then $\gamma_k \le y$. All published upper bounds bound G by counting lists of *minimal match pairs*, the difficulty being to control the overcounting.

Dančík [2] and Dančík–Paterson [3] control it by *domination*; for $k = 2$ a hand-built 52-state automaton gives $\gamma_2 \le 0.837623$, and a much coarser three-class system gives the values for $3 \le k \le 15$ that are still quoted today (see for instance [7]).

Lueker [6] replaces domination by *canonicity*: the canonical longest common subsequence obtained by backtracking through the dynamic programming table with the fixed priority (up, left, diagonal). Every pair of words has exactly one canonical decomposition (his Lemma 3.4), which domination does not provide — indeed his Conjecture 3.19, that a pair is generated by only one undominated list, is false for every $k \ge 3$, and he gives counterexamples. His machine $\mathcal{M}_h$ remembers the last h characters of each word and the last h differences in the final row and column of the table. His Theorem 3.26 states that the bounds converge to γ_k for arbitrary k; his Table II is binary.

The subsequent literature generalised the *other* half of his paper. Kiwi and Soto [5] extended his Section 2 — the lower-bound recurrence — to arbitrary alphabet size and number of strings, and that extension has become the standard tool for lower bounds. Section 3 was apparently never revisited for $k \ge 3$. Professor Kiwi has confirmed (personal communication, 19 September 2026) that they neither implemented nor considered it, and that they know of no obstruction.

3 The machine

We recall $\mathcal{M}_h$ only as far as needed. A *minimal match pair* is $p = (ua, va)$ with $a \in \Sigma$, $a \notin \mathrm{alph}(u) \cup \mathrm{alph}(v)$ and $\mathrm{alph}(u) \cap \mathrm{alph}(v) = \emptyset$. The map μ truncates each side of p from the left to at most h characters; M_0 denotes the (finite) set of minimal match pairs with both sides of length at most h. A state is $q = (x, y, dx, dy)$ with $x, y \in \Sigma^h$ and $dx, dy \in \{0,1\}^h$, and δ is computed as in [6, Fig. 7]: the limited-history table is extended, missing entries being replaced by the lower bounds available, and the input is accepted exactly when the canonical backtracking from the new corner returns to the stored corner.

Remark 1 (implementation note)**.** The canonicity test of [6, Fig. 7] is printed as `while` $i \ge h$ `and` $j \ge h$. Read literally the loop continues past (h,h) into the region where the limited-history table holds no values, and every input is rejected. The intended semantics — backtrack while strictly outside the stored corner, then test arrival at (h,h) — is `while` $i > h$ `or` $j > h$. With this reading the machine reproduces Table II of [6] exactly; see Section 7.

By [6, Thm. 3.3], if $z \in (0, 1/k)$ and $\lambda(z) \in (0,1)$ bounds the growth of the generating function of accepted lists, then

$$\gamma_k \;\le\; \frac{2\log(kz)}{\log\lambda(z)}. \tag{1}$$

Both logarithms are negative, so the right-hand side is *increasing* in λ: overestimating λ overestimates the bound, which is the safe direction. This is what makes the certificates of Section 6 sufficient.

4 The truncation weights for $k \geq 3$

The transfer matrix is $A_z = (a_{q,q'})$ with $a_{q,q'} = \sum_{p \in M_0} w_{\mu,z}(p)\,[\delta(q,p) = q']$, where $w_{\mu,z}(p) = \sum_{\mu(p')=p} z^{s(p')}$ sums over all minimal match pairs truncating to p [6, eq. (36)]. For $k = 2$ a minimal match pair always has one side of length 1, so at most one side is ever truncated and $w_{\mu,z}$ is immediate. For $k \geq 3$ both sides can be truncated at once, and the two extensions are not independent: they may not share a letter. This is the only step of [6, Sec. 3] that does not carry over verbatim.

Lemma 2. *Let $p = (ua, va) \in M_0$, $\ell_u = |ua|$, $\ell_v = |va|$, $A = \mathrm{alph}(u)$, $B = \mathrm{alph}(v)$ and $f = k - 1 - |A| - |B|$. Then*

$$w_{\mu,z}(p) = \begin{cases} z^{\ell_u+\ell_v}, & \ell_u < h,\ \ell_v < h, \\ \dfrac{z^{\ell_u+\ell_v}}{1-(|A|+f)z}, & \ell_u = h,\ \ell_v < h, \\ \dfrac{z^{\ell_u+\ell_v}}{1-(|B|+f)z}, & \ell_v = h,\ \ell_u < h, \\ z^{2h} \displaystyle\sum_{r+s \leq f} \binom{f}{r}\binom{f-r}{s}\, e_{|A|}(r)\, e_{|B|}(s), & \ell_u = \ell_v = h, \end{cases}$$

where $e_a(r) = \sum_{j=0}^{r} \binom{r}{j}(-1)^{r-j}\big(1-(a+j)z\big)^{-1}$.

Proof. The preimages of p under μ are the minimal match pairs $(w\,ua,\ w'\,va)$ with $w, w' \in \Sigma^*$, subject to $a \notin \mathrm{alph}(w\,u) \cup \mathrm{alph}(w'v)$ and $\mathrm{alph}(w\,u) \cap \mathrm{alph}(w'v) = \emptyset$. Let F be the set of f letters outside $\{a\} \cup A \cup B$. The constraints are exactly $\mathrm{alph}(w) \subseteq A \cup F$, $\mathrm{alph}(w') \subseteq B \cup F$ and $\mathrm{alph}(w) \cap \mathrm{alph}(w') \cap F = \emptyset$. If $\ell_u < h$ no truncation occurred on the top, so w is empty; likewise for the bottom. This gives the first three cases, the geometric series counting words over $A \cup F$ (resp. $B \cup F$).

In the fourth case put $P = \mathrm{alph}(w) \cap F$ and $Q = \mathrm{alph}(w') \cap F$; the constraint is $P \cap Q = \emptyset$. For fixed P, the generating function of words w with $\mathrm{alph}(w) \subseteq A \cup P$ and $\mathrm{alph}(w) \cap F = P$ is, by inclusion–exclusion over the subsets of P, $\sum_{P' \subseteq P} (-1)^{|P|-|P'|}\big(1-(|A|+|P'|)z\big)^{-1} = e_{|A|}(|P|)$, which depends on P only through $|P|$. Summing over ordered pairs of disjoint subsets of F of cardinalities r and s, of which there are $\binom{f}{r}\binom{f-r}{s}$, gives the stated expression; the factor z^{2h} accounts for ua and va themselves. □

That the summand depends only on the *cardinalities* r, s is what makes the computation feasible: the double sum has $O(f^2)$ terms instead of 3^f.

5 Relative forms, and one construction for every k

Canonical states have letter set $\{0, \ldots, m-1\}$ for some m. For $p \in M_0$ let $\mathrm{rel}_m(p)$ be obtained by keeping the letters $< m$ and renumbering the letters $\geq m$ by order of first appearance, reading the top component then the bottom one; let $\varphi(p)$ be the number of letters so renumbered.

Lemma 3. *Let q be a canonical state with m letters.*

(i) $\delta(q,p)$, its canonical class, and the pair $(\ell_u, \ell_v, |A|, |B|)$ depend on p only through $\mathrm{rel}_m(p)$.

(ii) For each relative form r with $\varphi(r) = \varphi$, the number of $p \in M_0$ with $\mathrm{rel}_m(p) = r$ *is the falling factorial $(k-m)(k-m-1)\cdots(k-m-\varphi+1)$, which vanishes as soon as $\varphi > k-m$.*

Proof. Two symbols with the same relative form differ by a permutation of Σ fixing $\{0,\dots,m-1\}$ pointwise. The computation of δ uses only equality tests between characters, so it commutes with any permutation of Σ; one fixing the letters of q leaves q unchanged, whence (i). For (ii), such a symbol is obtained from r by choosing the images of its φ fresh letters among the $k-m$ letters absent from q, injectively and in the prescribed order of first appearance. □

Corollary 4. *Each entry of the aggregated transfer matrix is $\sum_{\varphi} N_{\varphi}\,(k-m)(k-m-1)\cdots(k-m-\varphi+1)$, where N_{φ} is the number of accepted relative forms with φ fresh letters. A single alphabet-free construction, enumerating relative forms rather than symbols, therefore determines the exact matrix for every k; the terms with $\varphi > k-m$ vanish of their own accord, and the states that become unreachable are removed by a graph traversal.*

Corollary 5. *The transition structure is independent of k once $k \geq 4h-2$. Indeed $m \leq 2h-1$, because the two words stored in a state end with the same character — the match character of the last pair read — and $\varphi \leq 2h-1$, because a minimal match pair with both sides of length at most h uses at most $(h-1)+(h-1)+1$ distinct letters.*

At $h=3$ both bounds are attained ($m_{\max} = \varphi_{\max} = 5$), so the threshold is exactly 10. We stress what does *not* stabilise: the multiplicities themselves are polynomials in k and keep changing. At $h=3$, 24 628 of the 28 467 entries of the aggregated matrix differ between two consecutive values of k.

6 Certificates

For a nonnegative matrix A and any vector $v > 0$ one has $\rho(A) \leq \max_i (Av)_i / v_i$ (Collatz–Wielandt), whatever the provenance of v; this is the criterion Lueker uses [6, eq. (41)]. We take z_0 rational, compute an approximate Perron vector in floating point, round it to a rational vector $v > 0$ with denominator 10^8, and then evaluate $\lambda = \max_i (A_{z_0} v)_i / v_i$ and verify $A_{z_0} v \leq \lambda v$ coordinatewise *in exact rational arithmetic*. By the monotonicity noted after (1), $2\log(kz_0)/\log\lambda$ is then a valid upper bound for γ_k. No floating-point error analysis is required.

k	h	states	entries	z_0	λ (exact, shown truncated)	bound
3	4	5116	366370	18207/100000	0.198447797322	0.747886276
4	4	9427	2914973	1753/12500	0.186757381843	0.689063947
5	4	11093	7053339	11493/100000	0.179474183475	0.645032787
6	3	316	28466	2281/25000	0.142411285045	0.618256822
7	3	316	28467	8027/100000	0.141441938103	0.589456726
8	3	316	28467	3593/50000	0.140931417339	0.565041150
9	3	316	28467	6519/100000	0.140772426086	0.543937275
10	3	316	28467	2987/50000	0.140719207043	0.525416942
11	3	316	28467	69/1250	0.140796270761	0.508964025
12	3	316	28467	1027/20000	0.140933803949	0.494199896
13	3	316	28467	1201/25000	0.141122452238	0.480838934
14	3	316	28467	1129/25000	0.141340858176	0.468660628
15	3	316	28467	4263/100000	0.141596014720	0.457491357

7 Validation

Reproduction of Table II. Our implementation, run for $k = 2$, returns 7, 98 and 1020 accepting states at $h = 2, 3, 4$, matching [6, Table II], and certified values $\lambda = 0.131204212$, 0.177752613, 0.217017514 against the published 0.131204211, 0.177752610, 0.217017510. Our values *exceed* the published ones — the safe direction — by less than $4 \cdot 10^{-9}$, the difference being the rounding of the certificate vector. The resulting bounds 0.845855, 0.839250, 0.835072 agree with the published six decimals.

Thirteen further published values. The framework (1) was calibrated beforehand against the Chvátal–Sankoff bounds of 1975 and 1983 for $k = 2, \ldots, 6$ — computed 0.905118, 0.829982, 0.772908, 0.727666, 0.690556 and 0.866595, 0.786473, 0.729705, 0.686117, 0.650983 against published 0.90512, 0.82999, 0.77291, 0.72767, 0.69056 and 0.86660, 0.78648, 0.72971, 0.68612, 0.65099 — and against the values attributed to Deken in [2, Fig. 4.2] for $k = 3, 4, 5$, obtained by taking minimal match pairs alone: 0.776817, 0.718092, 0.673221 against 0.77682, 0.71810, 0.67323.

Remark 6. For $k = 2$ the same computation gives $\lambda(z) = 2z^2(1 + z)/(1 - z)$ and hence 0.860191, not the 0.85750 listed for Deken in [2, Fig. 4.2]. The reconstruction of [2, Sec. 4.2] therefore does not reproduce that particular entry, and it should not be used as a calibration point.

End-to-end test. For all pairs of words over Σ up to a given length we computed the canonical decomposition directly from the dynamic programming table and checked that $\mathcal{M}_h$ accepts it; and, conversely, that lists of minimal match pairs of total size at most h that are not canonical are rejected, as required by [6, Def. 3.12(2)]. Over (k, h) equal to $(2, 3)$, $(2, 4)$, $(2, 6)$, $(3, 3)$, $(3, 4)$ and $(4, 3)$ this covers 56 788 canonical decompositions: **none is rejected**, and of the 195 short lists tested (four of them not canonical) **none is wrongly accepted**. This is the test that would invalidate the bounds, since coverage of Σ^* is what [6, Thm. 3.3] requires.

The falling-factorial law. Corollary 4 was checked against direct constructions: 85 401 comparisons of individual matrix entries at $k = 7, 8, 9$, no discrepancy. The alphabet-free construction was then compared entry by entry with the direct one for $k = 3, 4, 5, 7, 10, 15$: identical state sets (including the smaller ones, 249 states at $k = 3$ and 309 at $k = 4$) and identical multiplicities throughout.

Internal consistency. The bounds are strictly decreasing in k. Kiwi, Loebl and Matoušek [4] proved $\sqrt{k}\,\gamma_k \to 2$; our $\sqrt{k}\times$ bound increases monotonically and stays below 2:

k	bound	$\sqrt{k}$·bound	$\sqrt{k}$·Dančík
3	0.747886276	1.29538	1.32641
4	0.689063947	1.37813	1.41647
5	0.645032787	1.44234	1.48571
6	0.618256822	1.51441	1.54150
7	0.589456726	1.55956	1.58794
8	0.565041150	1.59818	1.62750
9	0.543937275	1.63181	1.66181
10	0.525416942	1.66151	1.69136
11	0.508964025	1.68804	1.71749
12	0.494199896	1.71196	1.74103
13	0.480838934	1.73369	1.76239
14	0.468660628	1.75357	1.78177
15	0.457491357	1.77186	1.79946

8 Limitations

The bounds for $k \geq 6$ are computed at $h = 3$, the first rung of the ladder. At $k = 3$ the ladder reads 0.764971 ($h = 2$), 0.754807 ($h = 3$), 0.747886 ($h = 4$); at $k = 4$, 0.696850 ($h = 3$) and 0.689064 ($h = 4$); at $k = 5$, 0.653114 ($h = 3$) and 0.645033 ($h = 4$). Each further rung is worth roughly 0.007–0.008. We did not reach $h = 4$ for $k \geq 6$: the number of relative forms per state at $h = 4$ is 81, 414, 1895, 7740, 28 149, 91 226, 264 699, 693 480 for $m = 0, \ldots, 7$, and with some 10^4 states this puts $k = 6$ at about $4 \cdot 10^8$ evaluations of δ and $k = 7$ at about $2 \cdot 10^9$. These are not bounds of the method, only of the hardware used here; anyone with more compute will improve every line of the table.

For $k = 2$ we reproduce [6] and add nothing: his $h = 8$ machine has 6.4 million accepting states and gives $\gamma_2 \leq 0.826280$.

9 Code

A single self-contained script, `VERIF_CS.py`, rebuilds the machine from scratch, self-tests against Table II of [6], and re-verifies $A_{z_0} v \leq \lambda v$ in exact rational arithmetic. It shares no code with the program that produced the bounds. As distributed it covers twelve of the thirteen lines of the table — the ten lines at $h = 3$, together with $k = 3$ and $k = 4$ at $h = 4$ — in a few minutes, and reproduces each of them to the last digit printed here. The thirteenth line, $k = 5$ at $h = 4$, has 11 093 states and some $7 \cdot 10^6$ matrix entries and does not fit in a single short session; it was re-verified with the script's own routines in a checkpointed run of about a quarter of an hour, which reconstructs the same automaton (11 093 states, 7 053 339 entries) and the same certificate. All thirteen lines have thus been reproduced, to the last digit printed here, along a second code path. We note the limit of this independence in any case: the two programs were written by the same author from the same reading of [6, Fig. 7], so they are independent as code paths, not as readings. The script is attached to this deposit as `anc/VERIF_CS.py` (Python 3 and numpy, no argument); the line $k = 5$, $h = 4$ is obtained by uncommenting one line.

Note added, 31 October 2026

This note was circulated in the version above on 20 and 21 September 2026. On 25 October 2026, L. Jack made public a repository [8] with upper bounds for $3 \le k \le 15$ obtained from the same machine of Lueker. For $k = 3, 4$ ($h \le 4$) and $k = 6, \dots, 15$ ($h = 3$) it reports the same numbers of states and the same bounds to six decimals as the present note, and for $k = 5$, $h = 3$ the bound 0.653115, which agrees with the value 0.653114 of Section 8 up to the rounding of the last digit: an independent confirmation of these computations. It also reports $\gamma_3 < 0.737495$ ($h = 7$) and $\gamma_4 < 0.679701$ ($h = 6$), which improve the values of the present note for $k = 3, 4$. For $k = 5$ it stops at $h = 3$; to our knowledge the bound $\gamma_5 \le 0.645033$ ($h = 4$) above is the best published one, and for $6 \le k \le 15$ the two sets of values coincide.

Acknowledgements

I thank Professor Marcos Kiwi for two prompt replies, and in particular for confirming that Section 3 of [6] had been neither implemented nor considered for $k \ge 3$, and that no obstruction was known to him. The computations and the drafting of this note were carried out with the assistance of an AI system (Claude, Anthropic); every statement attributed to the literature has been checked against the primary source cited, and every numerical claim is backed by the certificates of Section 6.

References

[1] V. Chvátal and D. Sankoff, Longest common subsequences of two random sequences, *J. Appl. Probab.* **12** (1975), 306–315.

[2] V. Dančík, *Expected length of longest common subsequences*, Ph.D. dissertation, Department of Computer Science, University of Warwick, 1994.

[3] V. Dančík and M. Paterson, Upper bounds for the expected length of a longest common subsequence of two binary sequences, *Random Structures Algorithms* **6** (1995), no. 4, 449–458.

[4] M. Kiwi, M. Loebl and J. Matoušek, Expected length of the longest common subsequence for large alphabets, *Adv. Math.* **197** (2005), no. 2, 480–498.

[5] M. Kiwi and J. Soto, On a speculated relation between Chvátal–Sankoff constants of several sequences, *Combin. Probab. Comput.* **18** (2009), no. 4, 517–532.

[6] G. S. Lueker, Improved bounds on the average length of longest common subsequences, *J. ACM* **56** (2009), no. 3, Article 17, 38 pp.

[7] M. Rosenfeld, Upper bounds on the average edit distance between two random strings, arXiv:2407.18113 (2024).

[8] L. Jack, Improved upper bounds on the Chvátal–Sankoff constants for alphabets of size 3 to 15, GitHub repository `leandrejack-afk/chvatal-sankoff-bound`, version `ipl-2026`, 25 October 2026.