Kurdish Studies
August 2022
Volume: 10, No: 01, pp 407-421
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.67145/ks.v10i1.4141

Submodular Selection of Orthographic Correspondences Bounds the Script Component of the Sorani-Kurmanji Lexical Divide

Luis Eduardo Muñoz Guerrero[1]

[1]PhD, Universidad Tecnológica de Pereira, Pereira, Colombia. Email: lemunozg@utp.edu.co. ORCID: https://orcid.org/0000-0002-9414-6187

Abstract

Background. Central Kurdish and Northern Kurdish are written here in different alphabets, so any comparison of their vocabularies depends on a conversion step that is a choice.

Methods. Correspondence rules are chosen after transliteration to maximise covered token mass under a rule budget: a monotone submodular problem, NP-hard in general, solved greedily and certified against the exact optimum.

Results. Transliteration alone bridges 0.586 of token mass onto forms attested at least five times; twenty rules raise it to 0.653, the ground set's best twenty. Transliteration returns 0.311 against a Zazaki outgroup and 0.953 against the other half of the source corpus. On article titles held out as pairs the sixty-rule selection bridges 167 tokens, against 28.0 for jumbled-target rules.

Conclusions. Kurdish lexical overlap is not a well-defined single figure but a curve and a position on it; whether the varieties are one language is not decided here.

Keywords: Central Kurdish; Northern Kurdish; maximum coverage; approximation algorithms; transliteration.

1. Introduction

Whether Central Kurdish (Sorani) and Northern Kurdish (Kurmanji) are two languages or two written standards of one is among the oldest questions in Kurdish studies, and it bears on which variety a school system adopts (Hassanpour, 1992; Kreyenbroek, 1992; Sheyholislami, 2011); measuring it means choosing which orthographic correspondences to apply, a choice this paper poses as a constrained optimisation and bounds. The descriptive literature holds them closely related but distinct (Jügel, 2014; MacKenzie, 1961, 1962; McCarus, 2009), Northern Kurdish varies considerably within itself (Öpengin & Haig, 2014), and the standardisation debate is active (Hassanpour et al., 2012) and inseparable from language policy and shift (Zeydanlıoğlu, 2012; Zeyneloğlu et al., 2016). A critical overview traces much of the confusion to a failure to separate linguistic evidence from socially negotiated identity (Haig & Öpengin, 2014).

The question resists measurement for a mechanical reason, not a linguistic one. Central Kurdish is written in an Arabic-based alphabet and Northern Kurdish, in the material used here, in a Latin-based one, though it too is written in Arabic script in Iraq. A direct comparison of written forms therefore returns near-total difference however close the varieties are, so any quantitative answer depends on a conversion step, and that step is a choice.

The field is often summarised as though a figure existed. Sheykh Esmaili and Salavati (2013) built the first substantial comparable corpora of the two varieties and documented their phonological, morphological and distributional differences, but reported no measure of shared vocabulary, closing by proposing to "examine the lexical differences" in future work (Sheykh Esmaili & Salavati, 2013: 304). That corpus and the retrieval collection built on it (Sheykh Esmaili et al., 2013, 2014) remain the standard resources. The one published number comes from Hassani and Medjedovic (2016), who report the same 208 shared entries three ways: three per cent of a combined list, seven of one dialect's slice, five of the other's. None is an overlap between corpora or lexicons: all are shares of a 6,792-entry classifier-weighting list, computed after an unpublished transliteration and with no stemming; the authors note the rate "is far less than the one that is observable in the dialect identification percentage" (Hassani & Medjedovic, 2016: 73). The missing statistic is a corpus-level one: this paper's gap.

This paper treats that conversion step as the object of study. Rather than fix a conversion by judgement, we fix a minimal, uncontroversial conversion, enumerate the further correspondences the data suggest, and ask a question with an explicit objective: given at most k correspondence rules, which maximise the share of Central Kurdish text landing on an attested Northern Kurdish form? The answer is a curve rather than a number, and an overlap figure computed after any one fixed conversion reports a single point on it.

Posed this way, the question is maximum coverage under a cardinality constraint: cover as much weighted material as possible within a rule budget. The problem is NP-hard (Karp, 1972) and inapproximable beyond a fixed factor (Feige, 1998), but its objective is monotone and submodular, so the greedy construction carries the classical guarantee (Nemhauser et al., 1978). It began in location and allocation (Church & ReVelle, 1974; Cornuéjols et al., 1977), extends to weighted budgets (Khuller et al., 1999), is surveyed by Krause and Golovin (2014) and accelerated by Badanidiyuru and Vondrák (2014). In

language technology it serves summarisation (Lin & Bilmes, 2011), training-data selection (Kirchhoff & Bilmes, 2014), translation-rule pruning (Nishino et al., 2016) and recording-script selection (Barbot et al., 2015; François & Boëffard, 2001). The obvious alternative, ranking the rules once by standalone value and taking the best k, is provably not the same construction, because it cannot see redundancy; it is measured below as a baseline.

The contribution is fourfold: correspondence-rule choice stated as cardinality-constrained maximum coverage; an algorithm with three bounds on its own quality, the tightest certified by an exact integer solve; a measurement on corpora of both varieties with an outgroup, size, chance and positive controls, a sensitivity analysis and two held-out sets; and a pipeline whose every reported number comes from a stored artefact checked by a gate that fails the build.

Computational work on Kurdish is thin and mostly about building resources: lexicon induction from small seeds (Walther & Sagot, 2010), a survey of the obstacles (Sheykh Esmaili, 2012), corpora and a retrieval collection (Sheykh Esmaili et al., 2013, 2014), dialect identification within Central Kurdish (Malmasi, 2016) and inference across the Western Iranian group (Littell et al., 2016). The nearest analogue is Arabic, where dialect difference in written text has been measured at scale, though within one script (Zaidan & Callison-Burch, 2014).

Inducing correspondences from near misses is standard in cognate detection and transliteration mining: correspondences have been induced from bilingual wordlists (Kondrak, 2002) and read off orthographic alignments (Ciobanu & Dinu, 2014), transliteration has been modelled as a transduction (Knight & Graehl, 1998), and character-level models are standard between closely related varieties (Nakov & Tiedemann, 2012). What is new is not the induction but the budget: the rules are not merely found but priced against each other under a constraint.

One claim is deliberately not made. What is measured is the written form of words in edited prose. It is not mutual intelligibility, which depends on phonetic and lexical distance and on exposure (Gooskens, 2007), nor a statement about what speakers of either variety can understand. A reader who takes the numbers below as a verdict on that question will be misreading them.

2. Data

The primary material is the encyclopaedic text of three Wikipedias in closely related varieties, from the database dumps of 20 December 2016: Central Kurdish, Northern Kurdish and Zazaki. Zazaki is the outgroup: written in the same Latin orthography as Northern Kurdish and spoken in overlapping territory, but usually classified as a separate Northwestern Iranian language, not a Kurdish variety (Paul, 1998), a classification itself disputed; on that view, any procedure recovering a Central-to-Northern correspondence should recover markedly less against Zazaki. Two further sources are held out for evaluation: the interlanguage links of the same dumps, pairing articles on the same subject across the two Kurdish Wikipedias, and a small set of aligned translation strings. In total, six files are used, 63,613,318 bytes in all, each recorded with its origin, licence and SHA-256 digest before it was read.

Markup was removed with an explicit, ordered set of rules, redirects discarded, and only main-namespace articles kept. The Central Kurdish dump contains 56,886 main-namespace entries, of which 38,578 are redirects, leaving 18,308 articles; eight are empty once markup is removed, so the corpus is 18,300 articles, not the larger figure. Table 1 reports each corpus after extraction.

Table 1. Size and lexical shape of the three corpora after markup removal, redirect exclusion and tokenisation. Counts are of running words and distinct word types; hapax types occur exactly once, and the last row is the type-token ratio, higher for Zazaki because its corpus is the smallest.

	Central	Northern	Zazaki
articles	18,300	22,863	7,243
tokens	2,262,145	2,787,883	672,760
word types	208,644	244,702	102,135
hapax types	125,957	143,864	61,125
types seen 5+ times	33,563	39,911	14,896
type-token ratio	0.092	0.088	0.152

Two properties shape everything that follows. The vocabularies are large and heavy-tailed: 125,957 of the 208,644 Central Kurdish types occur once. Comparing every Central Kurdish type against every Northern Kurdish type means 51,055,604,088 pairs. The frequency floors below, and dropping forms the script already matches, leave 28,059 residual source forms against 39,911 target forms, a product of 1,119,862,749, which is why the candidate search is indexed rather than exhaustive.

3. Problem formulation

A fixed function t maps each Central Kurdish type to a Latin string. Described in the next section, it is deliberately many-to-one, so the unit counted downstream is a transliterated Latin form, not a Sorani spelling.

Counts are merged before they are filtered, not after, and both vocabularies then carry a frequency floor and a length cap, Equation 1: a form seen once in an encyclopaedia is as likely a misspelling as a word. Write V for the source forms that survive, U for the target types, c(w) for the tokens reaching form w, and M for the sum of c over V. Table 2 collects the notation.

$$c(w) = \sum_{s\,:\,t(s)=w} c_0(s), \qquad V = \{\, w : c(w) \geq \tau_V,\ |w| \leq L \,\} \qquad (1)$$

Applying t partitions V. Let W_0 be the types whose transliteration is already an attested target form, and m_0 the mass they carry; these cost no rules and are the script-only baseline. The remainder, W, is what the rules work on.

Table 2. Notation used throughout. The distinction between W_0 and W separates what the script alone accounts for from what the selected rules add.

symbol	meaning
V, U	transliterated source forms; target types
c(w), M	token count of w, summed over a set; total source mass
t	the fixed transliteration
W_0, m_0	types already matched by t; their mass
W	the residual types, V without W_0
R	the ground set of rule bundles
B(w)	bundles that bridge w, a subset of R
C(r)	residual types that bundle r bridges
f(S)	mass bridged by rule set S
k	the rule budget
S_k, S*_k	the set greedy holds at k; an optimal set at k
ϱ(k)	share of M bridged with k rules
y_r, z_w	indicators in the relaxation
1[·]	one if the condition holds, zero otherwise
g, A(S)	the conjunctive objective; a bundle set's rewrites
B̃(k), γ(k)	the certificate at k; the ratio it certifies

A candidate correspondence is derived from a near miss. For each residual type w and each target form within an edit distance d (Levenshtein, 1966) of t(w), here two, Equation 2, the minimal edit script is computed and each operation generalised into a contextual rewrite rule: a substitution, insertion or deletion with one character of context on either side, word boundaries marked. Costs are uniform rather than learned (Ristad & Yianilos, 1998), since a learned cost model would absorb the conversion decision this paper prices. The set of rules implied by one edit script is a bundle, and a bundle bridges w when applying it to t(w) yields an attested target form, which is what B(w) collects. Budgets count bundles. All observed bundles form the ground set R, restricted to those supported by at least a fixed number of distinct types, a floor swept with the others below. The objective, in Equation 3, is the mass of residual types some chosen bundle bridges.

$$\mathcal{N}(w) = \{\, u \in U : \mathrm{lev}(t(w), u) \leq d \,\}, \qquad \mathcal{B}(w) = \{\, r \in \mathcal{R} : r(t(w)) \in U \,\} \qquad (2)$$

$$f(S) = \sum_{w \in W} c(w)\, \mathbb{1}[\, \mathcal{B}(w) \cap S \neq \emptyset \,] \qquad (3)$$

The problem is to choose a rule set of bounded size maximising that mass, Equation 4, where S ranges over subsets of R. The quantity reported throughout is the share of the whole source mass accounted for, Equation 5, beginning at the script-only baseline and rising with the budget.

$$S_k^\star \in \arg\max \{\, f(S) : S \subseteq \mathcal{R},\ |S| \leq k \,\} \qquad (4)$$

$$\rho(k) = \frac{m_0 + f(S_k)}{M}, \qquad M = \sum_{w \in V} c(w) \qquad (5)$$

The objective's structure makes the problem tractable in practice and hard in theory. Writing C(r) for the residual types bundle r bridges, f(S) is the weighted cardinality of the union of C(r) over the chosen r, a weighted coverage function: monotone, since adding a bundle can only enlarge the union, and submodular, since a type already covered contributes

nothing when covered a second time, the diminishing-returns inequality of Equation 6. Submodularity is exactly the property a single ranking cannot respect: two rules that are each valuable may be largely redundant.

$$f(S \cup \{r\}) - f(S) \geq f(T \cup \{r\}) - f(T), \quad S \subseteq T \subseteq \mathcal{R},\ r \notin T \tag{6}$$

The problem is NP-hard as a class. Set cover reduces to it: one residual type per element, with unit weight, and one bundle per subset; a rule set of size k bridges every type exactly when k subsets cover the universe. Set cover is NP-complete (Karp, 1972), and the decision form here lies in NP because f is computable in polynomial time. Approximability is separate, and its two thresholds are easy to conflate: set cover resists approximation below the natural logarithm of its size unless every problem in NP admits a slightly superpolynomial algorithm, and maximum coverage below 1 - 1/e unless P equals NP (Feige, 1998). Only the second carries over, and greedy attains it (Hochbaum & Pathria, 1998) under a cardinality constraint, so unless P equals NP it is, up to an arbitrarily small margin, the best a polynomial procedure can promise.

One modelling decision must be stated. A bundle is a single element of the ground set, so a residual type counts as bridged when some chosen bundle bridges it. Under the alternative reading, Equation 7, a type is bridged only if every rewrite of one of its bundles is chosen. That function is not submodular, since adding a rule can raise another's marginal value, and because a bundle holds at most two rules its maximisation contains the densest k-subgraph problem, for which no constant-factor approximation is known (Feige et al., 2001). The readings are related in a usable direction. Taking the union of the rewrites in a chosen set of bundles gives a conjunctive set at most twice as large that bridges everything the disjunctive one does, Equation 8, so the value reported at budget k is a lower bound on the conjunctive optimum at budget 2k, the conservative direction here.

$$g(S) = \sum_{w \in W} c(w)\, \mathbb{1}[\exists\, r \in \mathcal{B}(w) : r \subseteq S] \tag{7}$$

$$A(S) = \bigcup_{r \in S} r, \quad |A(S)| \leq 2|S|, \quad f(S) \leq g(A(S)) \leq g^{*}(2k) \text{ for } |S| \leq k \tag{8}$$

4. Algorithm and guarantees

The selection uses the accelerated, or lazy, greedy of Minoux (1978), shown as Algorithm 1. Where the unaccelerated form recomputes the marginal gain of every remaining bundle in every round, the accelerated form keeps them in a priority queue whose keys are gains from earlier rounds: for a submodular function an earlier gain bounds the gain now, so a bundle whose stale key falls below a freshly computed competitor cannot win the round. The coverage attained is identical to the unaccelerated form's at every budget at which both ran, which the implementation asserts rather than assumes. That is not a theorem: the two can break ties differently, and once their selections differ their later values need not agree. On this instance they do agree.

Algorithm 1 Lazy greedy rule selection

```
Input   R, ground set of rule bundles
        C(r), residual types r bridges
        c, token weights;  k, the budget
Output  S, a subset of R with |S| <= k
Invariant  every key in Q is at least
        the true marginal gain against S

 1  Q <- empty max-priority queue
 2  S <- {};  X <- {}
 3  for r in R do push (c(C(r)), r, 0)
 4  for i <- 1 to k do
 5     loop
 6        if Q is empty then return S
 7        (g, r, s) <- pop(Q)
 8        if s = i then   // key fresh
 9           S <- S+{r};  X <- X+C(r)
10           break
11        g' <- c(C(r) \ X) // recompute
12        if Q is empty or g' >= key of
            top(Q) then
13           S <- S+{r};  X <- X+C(r)
14           break
15        else push (g', r, i) onto Q
16  return S
```

The invariant stated in the algorithm, that no key ever understates its bundle's true marginal gain, is what makes the early exit sound. It holds at initialisation because the keys are exact, and is preserved because a key is only ever replaced by the exact current gain and true gains never increase as the selection grows. Since a bundle is accepted only when its key is known to be exact, each round chooses a true maximiser, the construction attains the classical ratio of Nemhauser et al. (1978) given in Equation 9, and no procedure making polynomially many evaluations of the objective can promise better (Nemhauser & Wolsey, 1978).

$$f(S_k) \;\geq\; \left[1-\left(1-\frac{1}{k}\right)^{k}\right] f(S_k^{\star}) \;\geq\; \left(1-\frac{1}{e}\right) f(S_k^{\star}) \tag{9}$$

The worst-case ratio concerns the hardest instance, not this one, so two computable bounds are evaluated too.
The first follows from submodularity applied to any set in hand: the optimum at budget k cannot exceed the value attained plus the largest gains still available over the k bundles outside S, written T(S) in Equation 10. This online bound, used in sensor placement (Leskovec et al., 2007), is read off the priority queue, where stale keys are upper bounds and keep it valid, and minimised over every prefix, Equation 11, which is what the series below reports. The second is the optimum of the linear-programming (LP) relaxation of the natural integer program, Equation 12, relaxing the rule and coverage indicators to the unit interval.

$$f(S_k^{\star}) \;\leq\; f(S) + \sum_{r \in \mathcal{T}(S)} \delta_r(S), \qquad \delta_r(S) = f(S \cup \{r\}) - f(S) \tag{10}$$

$$\hat{B}_{\mathrm{on}}(k) \;=\; \min_{0 \leq j \leq k} \left[f(S_j) + \sum_{r \in \mathcal{T}_k(S_j)} \delta_r(S_j) \right] \tag{11}$$

$$\max \sum_{w \in W} c(w)\, z_w \quad \text{s.t.} \quad z_w \leq \sum_{r \in \mathcal{B}(w)} y_r, \quad \sum_{r \in \mathcal{R}} y_r \leq k, \quad 0 \leq y, z \leq 1 \tag{12}$$

The two bounds are not redundant: the online bound costs nothing and is loose, the relaxation costs a solve and can be tight. The smaller certifies that the selection attains at least a stated fraction of the optimum, Equation 13, much stronger than the worst-case ratio when the instance is benign. Its second form carries the baseline on both sides and flatters it. At this size the optimum can be computed outright, so the integer program behind the relaxation is solved by branch and bound at every budget, and the implementation refuses to proceed unless greedy value, integer optimum and relaxation are correctly ordered.

$$\gamma(k) = \frac{f(S_k)}{\hat{B}(k)}, \quad \tilde{\gamma}(k) = \frac{m_0 + f(S_k)}{m_0 + \hat{B}(k)}, \quad \hat{B}(k) = \min\{\hat{B}_{\mathrm{on}}(k), \hat{B}_{\mathrm{LP}}(k)\} \tag{13}$$

The unaccelerated greedy makes at most k passes over the ground set; the accelerated form shares that bound and is much faster in practice, as measured below. Building the ground set is dominated by the candidate search. Comparing every residual source form with every target form is quadratic in vocabulary size; a deletion-neighbourhood index reduces each query to hash lookups whose number depends only on word length and the distance bound. Two strings within edit distance d share a string obtainable by deleting at most d characters from each (Boytsov, 2011), so indexing every target form's deletion neighbourhood (Mor & Fraenkel, 1982; Muth & Manber, 1996) is exact rather than heuristic, and Equation 14 bounds a query's cost.
Bounded edit distance uses the banded form that stops once the bound is exceeded (Ukkonen, 1985); Navarro (2001) surveys the field.

$$\mathrm{lev}(x,y) \leq d \;\Rightarrow\; D_d(x) \cap D_d(y) \neq \emptyset, |D_d(x)| \;\leq\; \sum_{i=0}^{d} \binom{|x|}{i} \tag{14}$$

5. Pipeline and reproducibility
The transliteration t is deterministic and applied before anything else. It maps each letter of the Arabic-based alphabet to its Latin counterpart, following the standard reference grammar and published romanisation table (American Library Association & Library of Congress, 2012; Thackston, 2006b), resolves by position the two letters orthography leaves genuinely ambiguous, and changes nothing else. The digraph for the long back vowel is read before the single letter

composing it; word-final forms of the letter writing both a consonant and a short vowel are normalised. The variant spellings Sorani web text mixes freely are unified first: the Arabic and Persian forms of yeh and kaf, the several spellings of final e, and the zero-width non-joiner. Because the map is many-to-one, distinct Sorani spellings can arrive at one Latin string; their counts are added, and every figure below follows that merging.

One systematic difference is deliberately not handled, because handling it would decide the result in advance. Sorani writes every vowel but one: the short i is not represented inside a word (Thackston, 2006b: 5, 75), whereas Kurmanji's Latin orthography writes it as a full letter (American Library Association & Library of Congress, 2012; Bedir Khan & Lescot, 1970; Thackston, 2006a). Supplying it is the kind of adjustment a fixed conversion makes silently, and it moves the answer, so the vowel is left for the rule set to find and the budget curve to price, as are the inflectional endings on which the varieties differ and the lexical divergences in the reference lexicography (Chyet, 2003).

Figure 1 shows the stages. Each writes a file the next reads, no value passes between stages in memory, and the costly stages are timed. The program that writes this document reads every reported measurement from those files, and gates fail the build if the document carries a numeral matching no stored value, mentions a figure or table only after it appears, exceeds the abstract limit, uses an abbreviation before defining it, or numbers its equations out of order. The gates are themselves under test, and those tests run before the build. Random choices are seeded, and the mining, selection and held-out stages give identical result files once timings are set aside.

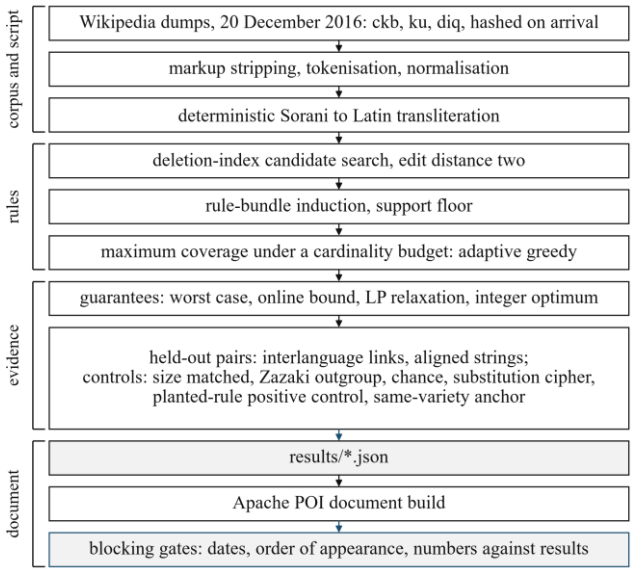

Figure 1. Stages of the pipeline. Each box writes an artefact the next reads, so any stage can be rerun alone. The controls rerun the same stages against another target. The gates come last because they consume the finished document rather than feed it.

6. Results

Transliteration alone accounts for a large share of the text. The denominator is easy to state loosely: transliterating collapses distinct Sorani spellings onto one Latin string, and the frequency floor applies to the merged forms, leaving 32,769 forms carrying 2,023,736 tokens. Of that mass, 0.586 lands on a Northern Kurdish form itself attested at least five times, with no rules at all. Against every token in the corpus instead, the same count gives 0.524. Against a chance target of the same size, built from the same letters, it is 0.199, the number the full-size figures below should be read against.

A same-variety anchor matters as much as a lower reference. Run against the other half of its own corpus, with the same transliteration and floors, the procedure reaches 0.953 before any rule is selected, so 4.7 points are lost to corpus sparsity and the floor alone, and the figure for the two varieties is 61 per cent of that, conservatively: the anchor's target is half the size of the Northern Kurdish one, and is a mean over three splits.

From the residual types the candidate search produced a ground set of 903,801 distinct bundles, the overwhelming majority bridging exactly one word type. A bundle supported by a single type is an orthographic accident, not a correspondence; admitting such bundles would let the procedure claim credit for coincidences. Requiring at least five distinct types leaves 4,570 bundles, 1,599 a single rewrite and 2,971 a pair, reaching 10,855 residual types between them. Selecting all of them would raise the bridged share to 0.808, the ceiling against which any budget should be read.

Figure 2 gives coverage against the rule budget. The adaptive greedy construction reaches 0.633 at ten rules and 0.653 at twenty, against a script-only baseline of 0.586. A ranking of the same bundles by standalone yield, taken once and never revised, reaches only 0.619 at twenty, and uniformly random rule sets reach 0.590. The gap between the adaptive and the static construction is the price of ignoring redundancy: 5.2 percentage points at sixty rules, about as much as the last forty of the sixty rules add at all, 4.0 points.

The certificate is the more interesting half of the figure. The LP relaxation of this instance is tight at fifty-nine of the sixty budgets solved: at those the integer program solved to optimality returns the relaxation's value, and at the one exception falls below it by 15.5 tokens. Up to twenty-five rules greedy attains it, so over that range the selection is not merely within a guaranteed factor of the optimum but optimal for this ground set.

Beyond it greedy falls short by 1,008 tokens at thirty-nine rules and 1,209 at sixty. On the objective the guarantee concerns, the residual mass without the script-only baseline, the ratio at sixty rules is 0.99448, where the worst case would permit a shortfall of nearly thirty-seven per cent; counting the baseline on both sides flatters it to 0.99914.

The two computable bounds are not equally useful here. At twenty rules the online bound alone certifies 0.746 of the optimum objective while the relaxation certifies optimality outright, and Figure 3 sets the two against each other over every budget: the online bound costs nothing, is loosest in the middle of the range, and is the only one available where the relaxation does not solve.

Two further checks share no code with the solver: an exhaustive search over pairs, pruning only pairs that cannot win, returns the greedy optima at the two smallest budgets, and the accelerated and unaccelerated greedy agree at every budget up to twenty, the range over which both ran.

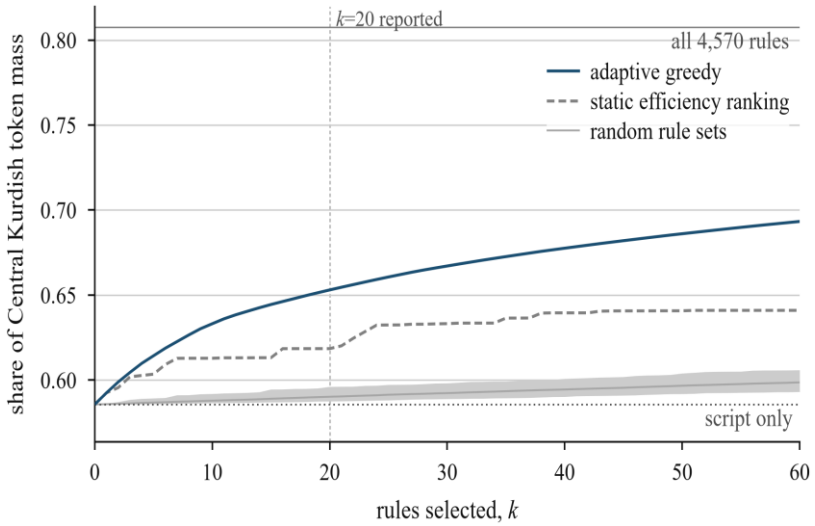

Figure 2. Share of Central Kurdish token mass bridged against the number of rules selected. The grey band spans the middle ninety-five per cent of 200 random rule sets; the horizontal lines mark the script-only baseline and the share the whole ground set would reach; the dashed line is that set taken once in order of standalone yield.

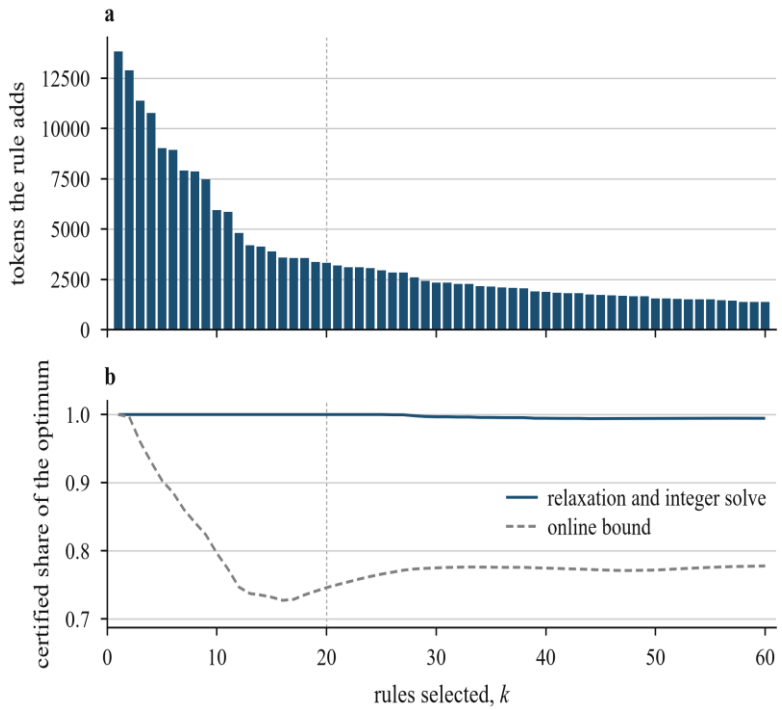

Figure 3. Upper panel: the tokens each further rule adds, in the order chosen. Lower panel: the share of the optimum each certificate establishes, on the residual mass the guarantee concerns.

Table 3 lists the highest-yielding rules with the mass each adds when it is chosen.

Table 3. The eight rules the adaptive construction selects first, in the order chosen. A rule is up to two rewrites in context, separated by a semicolon: the underscore marks the rewritten material, the dollar sign the right word boundary, and a zero the empty string. The last column is the mass the rule adds when selected; the types column is its standalone support, so the two are not on the same basis.

rank	rule	types	added
1	y -> 0 / e _ $	241	13,831
2	0 -> y / û _ e	10	12,900
3	î -> a / n _ $	84	11,391
4	0 -> i / k _ r	31	10,768
5	y -> 0 / a _ $	111	9,021
6	0 -> i / t _ r	33	8,948
7	î -> 0 / r _ $	146	7,909
8	î -> 0 / k _ $; k -> 0 / ê _ î	226	7,865

The rules are not arbitrary, worth saying because a coverage objective is under no obligation to produce anything interpretable. Five of the first eight concern the ends of words: ranks three, seven and eight rewrite a final î after a consonant, among them î -> a / n _ $ and î -> 0 / r _ $, and ranks one and five delete a final y after a vowel. That is where the varieties differ in their nominal endings, and the loss of nominal case in Central Kurdish is among the divergences catalogued by Sheykh Esmaili and Salavati (2013). Rank two, 0 -> y / û _ e, inserts a glide between two vowels; it bridges only 10 word types, but frequent ones, exactly the case a type-level criterion discards and a token-weighted one keeps. Rank six is worth pausing on. It is 0 -> i / t _ r, the insertion of a short i between two consonants: the vowel Sorani orthography does not write and Kurmanji does, recovered from the data without having been told about it.

Supplied in the transliteration, as it easily could have been, it would have gone into the script-only baseline and made the rule budget look cheaper than it is. Leaving it to be found is what makes the division between the two components mean anything.

Each of the following reruns the mining and selection against a different target. The Zazaki outgroup is processed identically. A size-matched control subsamples the Northern Kurdish corpus by whole articles to the Zazaki token total, three times, since a larger target vocabulary offers more chances to match by accident. A chance level permutes the letters inside each target word, keeping word lengths, letter frequencies and token weights, and rejects an image identical to its source; where a word has no other anagram, its letters are drawn from the target's letter distribution instead. An image coinciding with some other real word is kept, since an unrelated vocabulary would produce such coincidences too. The chance level runs at full size and at the outgroup's size. A substitution cipher relabels the target alphabet. Figure 4 shows all of them; a positive control with planted rules is described below, and the same-variety anchor above is a rerun of the same kind, not plotted here.

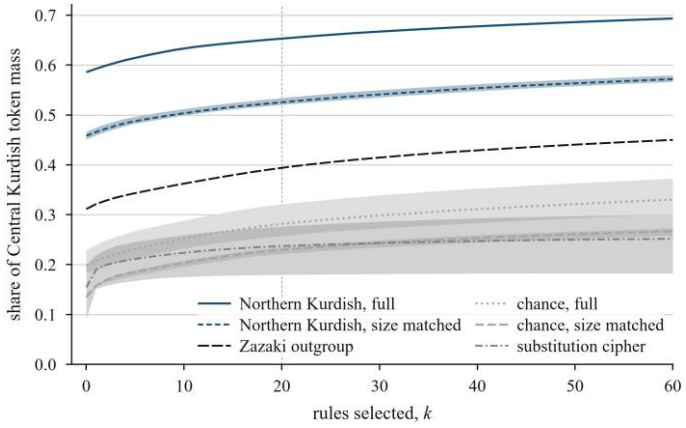

Figure 4. Coverage against rule budget for the six conditions Table 4 reads at twenty rules, banded from the lowest seed to the highest. The comparison controlling for vocabulary size is the size-matched curve against the outgroup and the chance level of the same size.

With no rules at all Northern Kurdish stands to the outgroup in a ratio of 1.88, but much of that is an artefact of corpus size: cut to the Zazaki token total it falls to 1.47. The second comparison is the honest one: the ratio is smaller, and it survives; reporting only the first would have overstated how far Northern Kurdish stands from Zazaki by a wide margin.

The chance level needs reading with care. With the letters of each target word jumbled, the script-only baseline is still 0.199, but 98.3 per cent of that is source forms landing on strings that are themselves attested Northern Kurdish words: a permuted short word lands in a space the real vocabulary fills. What rests on strings the target does not contain is 0.003. The jumble bounds chance from above rather than estimating it. Comparisons must be made within a corpus size. The ordering holds within each size, but no one of these levels measures anything on its own.

The positive control is a target with known correspondences. The ground set contains 33 single rewrites bridging at least 50 word types; ten are planted at random in a copy of the transliterated source, never two on the same site and never one

whose output the source already has, so the copy differs only by those rewrites. In each of three draws the first ten selections contain at least nine of the ten planted rules and bridge at least 99.6 per cent of the planted mass, and twenty rules bridge 100.0 per cent of it, where the same ground set ranked once by standalone yield recovers at most three. When a planted rule is missed, its forms are bridged by another landing them on a different attested word, which the objective cannot distinguish from the right one: the limit the held-out sets expose below, here with the answer known. The cipher was meant to be this control and is not: undoing it needs a rewrite at almost every letter, a bundle allows two, and the rules chosen against it are the accidental deletions the jumbled target yields. It is a second chance level, read as one.

That leaves a question the coverage curves cannot answer, and the answer runs against the paper's headline. Table 4 sets the six conditions side by side. Broadly, the lower the baseline the more the rules add, and the two real-target conditions, which start highest, gain least of all; as a share of the residual available they are of the same order, the real pair the highest. In percentage points the increment is not evidence about these two varieties; it measures the near-miss density of a vocabulary this size. Whether the rules chosen are real is a question for held-out data.

Table 4. The six conditions of Figure 4, each the mining and selection rerun against a different target. Gain is in percentage points of source mass; the last is that gain against the mass the script left unbridged, which is what the rules compete for. Seeded conditions are means over three draws.

	no rules	at twenty	gain	residual
Northern Kurdish, full	0.586	0.653	6.7	16.3
Northern Kurdish, size matched	0.458	0.525	6.7	12.4
Zazaki outgroup	0.311	0.394	8.3	12.0
substitution cipher	0.154	0.237	8.3	9.8
chance, full	0.199	0.281	8.3	10.3
chance, size matched	0.135	0.229	9.5	10.9

The coverage objective counts a type as bridged when the rewritten form is attested anywhere in the target vocabulary, which is weaker than it sounds: it can land on an unrelated word. Two held-out sets test the stronger claim; Table 5 reports them.

The interlanguage links give 4,841 title pairs denoting the same subject. Across them 0.244 of source tokens land on a word in the paired title, against 0.0002 under random pairing, and most of that is not the rules: transliteration alone accounts for 0.223 and the selected rules add 0.021.

Titles whose words are usually capitalised in Northern Kurdish prose bridge at 0.282 and the others at 0.160, a ratio of 1.77. That contrast does not survive its own contamination. A third of the word-like tokens, 856 of 2,601, come from titles beginning with a numeral, dates almost all of them, none of which bridges. Set them aside and the word-like rate rises to 0.238, the ratio falls to 1.19, and what looked like a difference between names and words is mostly one between words and dates. The classification matters more than the numbers: reading the case of the title instead, which the wiki software capitalises without exception, leaves only 1,144 tokens in the other bucket, 847 numeral-initial, and gives a ratio of 5.36, overstating the prose-classified contrast 3.0-fold.

Table 5. Held-out evaluation with the sixty-rule selection. Script is the share of source tokens reaching a word of the paired expression by transliteration alone, rules the further share the selected rules add, total both together, and chance the same against random pairings. Titles are sorted by whether their words are usually capitalised in Northern Kurdish prose; those too rare to judge are kept apart. The aligned strings are a second, unrelated held-out set of user-interface translation pairs. Chance is a mean over 200 reshufflings, and the near-zero rates carry one more decimal.

	tokens	script	rules	total	chance
linked titles, all	7,918	0.223	0.021	0.244	0.0002
name-like titles	3,897	0.262	0.020	0.282	0.0003
word-like titles	2,601	0.145	0.015	0.160	0.0002
too rare to judge	1,420	0.259	0.036	0.295	0.0002
aligned strings	638	0.0298	0.0016	0.0313	0.0020

The aligned strings are user-interface translations from an open parallel collection (Tiedemann, 2012). Of 1,614 pairs in the file only 217 carry Central Kurdish script on the source side, the rest untranslated, leaving 638 tokens to test. Of those, 0.031 reach a word in the actual translation, all but one by transliteration alone, against 0.0020 under random pairing, so the set bears on precision alone. Reading the pairs explains the low figure: the two localisations were made independently, often choose different words for the same element, and sometimes render different strings. Of the tokens the objective would have declared bridged, only 6.2 per cent land on a word in the paired translation. That is a lower bound, and the clearest statement here of what the corpus-level number does and does not mean.

The held-out links bear on the question the controls left open. The same evaluation was run with rules selected against the real target and against each jumbled one.

Exact matches involve no rules and come out identical, at 1,766 tokens, confirming that nothing else differed. The rules do not, Table 6: chance-selected correspondences transfer at a fraction of the rate.

How held out this is deserves a plain answer, because the honest figure is the stronger one. The pairs were not used to induce or select rules, but their words were: 77.3 per cent of the source tokens are forms the objective summed over, and 61 of the 167 rule hits are pairs the candidate search had already mined, credited to the rule the objective was scored on. The last two rows of the table set those aside and then keep only source forms the selection never saw. Removing the mined pairs alone lowers the ratio; restricting to forms the selection never saw raises it well above where it began, which is the direction that matters and the strictest test. Over all held-out tokens an exact McNemar test on the discordant pairs (160 against 41 at the weakest of the three seeds) rejects equality against every chance selection at p below 0.001. The binomial is exact rather than continuity-corrected because at these counts the approximation is the wrong test, not a second opinion. The strict figure rests on 54 tokens, which is not many.

Table 6. Tokens the selected rules bridge on the held-out title pairs, against the mean of the three chance selections. Each of the last two rows is a harder test than the one above: pairs the candidate search had already mined are set aside, and then only source forms the objective never summed over are kept. Exact matches use no rules and are identical in both arms.

	tokens	real	chance	ratio
all held-out tokens	7,918	167	28.0	6.0
name-like titles	3,897	78	16.0	4.9
word-like titles	2,601	38	6.3	6.0
too rare to judge	1,420	51	5.7	9.0
mined pairs removed	7,857	106	26.7	4.0
unseen source forms	1,796	54	4.3	12.5

Table 7 gives the measured cost of each stage. What matters is the candidate search: comparing every residual source form with every target form at the working floor is 1,119,862,749 pairs, and on nested subsamples the exhaustive scan runs at 680,184 pairs per second, projecting to 1,646 seconds at full size against 10.2 seconds for the index. The two agree at every size at which both ran.

At twenty rules the accelerated greedy performs 4,662 marginal-gain evaluations against 91,210 for the unaccelerated form, a factor of 19.6, and is 13.5 times faster in wall-clock time.

Fitted on logarithmic axes across the nested subsamples, the scan's time grows with the 2.03 power of the sampled fraction of both vocabularies, the index's with the 1.19 power, or the 1.36 power once the full-size run is included. The largest index-over-scan advantage measured where both ran is a factor of 99.3 and the projected advantage at full size about 161, not the orders of magnitude the asymptotics suggest, because generating a query's deletion neighbourhood is itself expensive in an interpreted language.

Table 7. Wall-clock time and peak traced allocation by stage, on a single eight-core desktop processor with 16 GB of memory, single-threaded; memory is what the interpreter's allocation tracker reports, not resident set size. Both greedy rows use a budget of twenty rules. The indexed-search row is part of the rule-induction row, timed alone, and not to be added to it. The exhaustive search row is projected; the last two rows each solve sixty programs. Times are measured with the tracker off, because tracing costs the indexed paths several times their running time; memory comes from a separate pass. A dash marks a quantity not recorded.

stage	seconds	peak MiB
exhaustive search (projected)	1,646	—
indexed search	10.2	202
rule induction	26.6	—
greedy, unaccelerated	0.23	—
greedy, accelerated	0.017	—
LP relaxation, every budget	9.7	—
integer optimum, every budget	170.2	—

Five of the quantities set by judgement were swept, each varied with the others fixed, Table 8. One of them is the transliteration itself, whose two published readings of the trill and the velarised lateral are a choice like any other. Figure 5 plots the three that take more than two values.

One threshold dominates the bridged share; the other three matter much less. Moving the target frequency floor from one to twenty moves the bridged share at twenty rules by 27.8 percentage points, more than twice the 13.2-point gap between the size-matched target and the outgroup. The ceiling is a different matter and the dominance does not hold there: the support floor, which barely moves the budgeted share, moves the ceiling as far as the target floor does, and the other three move it at most 5.6 points. A ceiling quoted without its thresholds is therefore no better defined than a single overlap figure.

Table 8. Each threshold swept with the others at their working values. The second column is how far the bridged share at twenty rules moves across the settings tried, in percentage points; the last two are the lowest and highest ceiling the ground set reaches. Three thresholds move the ceiling and two barely touch it, and they are not the ones that move the budgeted share most.

threshold	at k	ceiling	
target frequency floor	27.8	0.676	0.925
source frequency floor	8.8	0.803	0.808
rule support floor	2.0	0.644	0.878
edit-distance bound	0.3	0.751	0.808
transliteration table	1.3	0.793	0.808

No absolute figure here is therefore a property of the two varieties alone; each is also a property of a threshold. What can be tested is whether the ordering survives the threshold that moves everything most, so the size-matched target, the outgroup and the size-matched chance level were rerun at the lowest and highest target floors. At those and the working floors, and for every seed, the size-matched target sits above the outgroup and the outgroup above chance, with no rules and at twenty. The first gap is not stable: at twenty rules it runs from 9.1 to 18.3 percentage points across the floors. The other three thresholds were not crossed with the controls, so the claim is for the target floor alone.

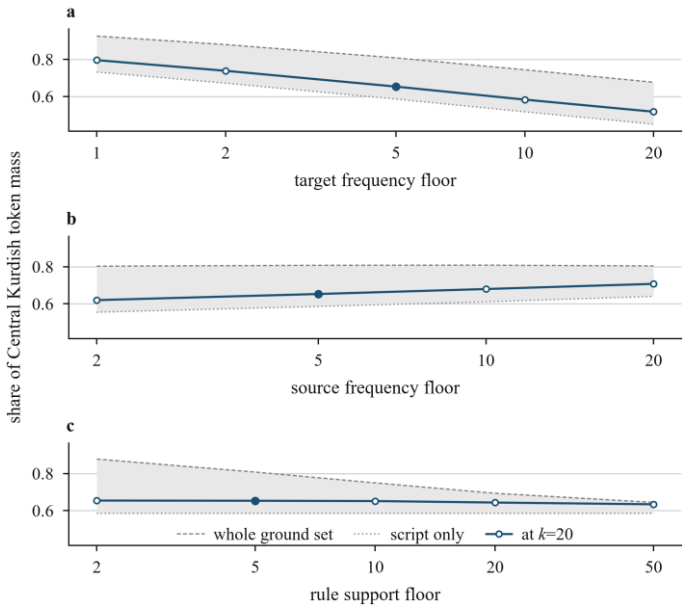

Figure 5. Share of Central Kurdish token mass bridged at twenty rules as each threshold is varied with the others held, shown by the filled marker. Horizontal axes are logarithmic; the panels share one vertical range and scale, and each axis label carries the span its panel covers. The two thresholds that take only two values are in the table, not here.

7. Discussion

The central finding is a shape rather than a number. There is no single quantity called the lexical overlap of Central and Northern Kurdish, only a curve from what the writing systems alone account for to what an unbounded set of correspondences could, and a study reporting one number has chosen a point on it.

The range reported here, on one ground set, spans a factor of 26: from 0.031 for interface strings measured against their translations, to 0.808 for the frequency-weighted share the whole ground set would reach.

Neither is wrong; they answer different questions. Figure 6 sets the measured quantities side by side; the ceiling is not among them, because no measurement reaches it.

The lowest lies close to the three per cent of Hassani and Medjedovic (2016), computed over different objects, so the resemblance is not agreement.

Name-like titles bridge somewhat more often than word-like ones, but the margin is small once date titles are removed, and the plausible reason, that names are transliterated where common words are translated, is not established here: the sets differ in other uncontrolled ways.

The design does show how easily such a contrast is manufactured: classifying by the case of the title, which the wiki software sets, overstates it 3.0-fold, and leaving the dates in overstates it again.

The implication for the standardisation debate is narrow. At sixty rules the selection has realised about half the mass the ground set makes reachable and is still gaining, so these data do not locate where the curve flattens, and the ground set's ceiling moves by more than twenty points with the thresholds.

Nor is it established what the residue consists of: this study measures how much of it bounded rewriting removes, not what remains.

An argument for a common written standard resting on the difference being mostly one of alphabet, a question the standardisation literature has long debated (Hassanpour, 1992; Hassanpour et al., 2012), is not settled by this evidence either way.
What does follow is how to read the figures in circulation: each says what one conversion, at one budget, on one pair of corpora could bridge, and two cannot be compared without those three quantities.

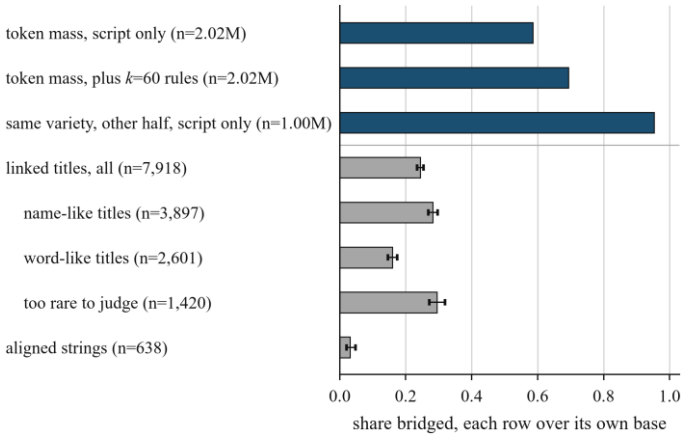

Figure 6. Eight ways of stating the same result. The three darker bars are frequency-weighted shares of running text, dominated by the most frequent words; the lighter bars are unweighted token shares over held-out pairs, with ninety-five per cent intervals. The second row adds the sixty selected rules to the first; the third is the same measurement against the other half of Central Kurdish's own corpus, over a base half the size. The three indented rows divide the linked-title row above. No single number can be quoted without saying which it is.

Two things transfer beyond Kurdish. First, making the preprocessing choice an explicit objective with a budget turns an argument about methods into a measurement. Dialectometry compares varieties with string distances (Heeringa et al., 2006; Kessler, 1995; Nerbonne & Heeringa, 1997), reaching back to lexicostatistics (Swadesh, 1952) and the classical surveys (Chambers & Trudgill, 1998), with careful accounts of what they capture (Nerbonne, 2009; Wichmann et al., 2010; Wieling & Nerbonne, 2015); a coverage formulation adds a budget, and so a curve. Second, the worst-case guarantee was far from binding: the relaxation certified greedy optimal up to twenty-five rules and the integer solve measured its shortfall beyond. The certificate tells a practitioner what the guarantee cannot, at the cost of one solve per budget.

8. Limitations

The seeded conditions are replicated three times, which prices the noise of subsampling one corpus, not variation across varieties of that size; the outgroup is a single corpus with no replicate, so the comparison that carries the linguistic claim is one against one.
Three draws do not resolve the third decimal these means carry. The size-matched control equalises token mass, not the type count, which it does not reach. The Northern Kurdish corpus also carries passages of Central Kurdish and Zazaki transcribed in Latin, which nothing here removes.
The corpora are encyclopaedic prose written by volunteers, and frequency-weighted statistics lean towards whichever articles are most numerous; nothing corrects for that. The held-out figures, counting each token once rather than by corpus frequency, are less exposed. Every absolute figure also depends on the target frequency floor, and only the ordering of the conditions was shown to survive it.
What is measured is vocabulary shared in writing, whatever its origin: Arabic, Persian and Turkish loans count exactly as inherited words do, so the figures are not a measure of common descent and would be lower on a basic-vocabulary list (Swadesh, 1952). And the claim that the selected rules are interpretable is the author's reading of Table 3, not a Kurdish linguist's.
The transliteration resolves two ambiguous letters by position. That convention is standard but not always right, and its errors fall into the residual, where the selected rules can repair some, so the split between script and rules is not perfectly sharp; the total is unaffected.
A bundle counts as bridging a type whenever it produces any attested target form, and the held-out evaluation shows many such landings are on words that are not the counterpart. The corpus-level curve is therefore an upper bound on correspondence and the held-out figures a lower one, and this work does not locate the truth between them. The optimality established above is over the ground set the candidate search induced, under the thresholds that shaped it.
The title pairs are held out from rule induction and selection as pairs, not as words: their Northern Kurdish words can also occur in the running text the target vocabulary was built from, so the evaluation is not independent of the target lexicon.

The aligned-translation set is small and drawn from a single technical register, and the selected rules bridge one of its 638 testable tokens, so it bears on precision alone. Finally, only two varieties are treated in depth. Southern Kurdish is absent, as are the Gorani varieties usually grouped with Zazaki rather than Kurdish, because the source offered no comparable text for them in the period covered.

9. Conclusion

Choosing which orthographic correspondences to apply before comparing two varieties is a combinatorial optimisation problem; treating it as one makes the figures comparable. On the December 2016 corpora, script conversion alone bridges 0.586 of source token mass, twenty selected rules bring it to 0.653, the best any twenty bundles in the ground set can do, and the whole ground set would reach 0.808. On held-out expression pairs the selection bridges 0.282 of the tokens in name-like titles and 0.238 in word-like ones once date titles are set aside, and rules learned from the real target transfer at 6.0 times the rate of those learned from a jumbled one. No single figure for the lexical overlap survives scrutiny: what can be reported is a curve, the ground set it was computed over, and where on it a study sits. Whether the two are one language is a question these measurements inform but do not settle.

Declarations

Funding. This research received no specific grant from any funding agency.
Conflicts of interest. The author declares none.
Data availability. All primary data are public. The dumps of 20 December 2016 and the interlanguage-link tables are archived at the Internet Archive under the Creative Commons Attribution-ShareAlike licence and are the work of their contributors (Wikimedia Foundation, 2016); the live mirror no longer carries them. The aligned strings come from the Ubuntu corpus distributed by OPUS (Tiedemann, 2012), a later repackaging of strings from a 2014 release, licensed per string.
Ethics. The study uses published text and involves no human participants.

References

1. CAmerican Library Association & Library of Congress. (2012). Kurdish romanization table (2012 version). In ALA-LC romanization tables. Washington, DC: Library of Congress.
2. Badanidiyuru, A., & Vondrák, J. (2014). Fast algorithms for maximizing submodular functions. In Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 1497–1514). Philadelphia: SIAM.
3. Barbot, N., Boëffard, O., Chevelu, J., & Delhay, A. (2015). Large linguistic corpus reduction with SCP algorithms. Computational Linguistics, 41(3), 355–383.
4. Bedir Khan, E. D., & Lescot, R. (1970). Grammaire kurde (dialecte kurmandji). Paris: Librairie d'Amérique et d'Orient Adrien Maisonneuve.
5. Boytsov, L. (2011). Indexing methods for approximate dictionary searching: Comparative analysis. ACM Journal of Experimental Algorithmics, 16, Article 1.1.
6. Chambers, J. K., & Trudgill, P. (1998). Dialectology (2nd ed.). Cambridge: Cambridge University Press.
7. Church, R., & ReVelle, C. (1974). The maximal covering location problem. Papers of the Regional Science Association, 32(1), 101–118.
8. Chyet, M. L. (2003). Kurdish-English dictionary = Ferhenga Kurmancî-Inglîzî (Yale Language Series; with selected etymologies by M. Schwartz). New Haven: Yale University Press.
9. Ciobanu, A. M., & Dinu, L. P. (2014). Automatic detection of cognates using orthographic alignment. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 99–105). Baltimore: Association for Computational Linguistics.
10. Cornuéjols, G., Fisher, M. L., & Nemhauser, G. L. (1977). Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. Management Science, 23(8), 789–810.
11. Feige, U. (1998). A threshold of ln n for approximating set cover. Journal of the ACM, 45(4), 634–652.
12. Feige, U., Peleg, D., & Kortsarz, G. (2001). The dense k-subgraph problem. Algorithmica, 29(3), 410–421.
13. François, H., & Boëffard, O. (2001). Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. In Proceedings of Eurospeech 2001 (pp. 829–832). Aalborg: ISCA.
14. Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. Journal of Multilingual and Multicultural Development, 28(6), 445–467.
15. Haig, G., & Öpengin, E. (2014). Introduction to special issue – Kurdish: A critical research overview. Kurdish Studies, 2(2), 99–122.
16. Hassani, H., & Medjedovic, D. (2016). Automatic Kurdish dialects identification. Computer Science & Information Technology, 6(3), 61–78.
17. Hassanpour, A. (1992). Nationalism and language in Kurdistan, 1918–1985. San Francisco: Mellen Research University Press.
18. Hassanpour, A., Sheyholislami, J., & Skutnabb-Kangas, T. (2012). Introduction. Kurdish: Linguicide, resistance and hope. International Journal of the Sociology of Language, 2012(217), 1–18.

19. Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In Proceedings of the Workshop on Linguistic Distances (pp. 51–62). Sydney: Association for Computational Linguistics.
20. Hochbaum, D. S., & Pathria, A. (1998). Analysis of the greedy approach in problems of maximum k-coverage. Naval Research Logistics, 45(6), 615–627.
21. Jügel, T. (2014). On the linguistic history of Kurdish. Kurdish Studies, 2(2), 123–142.
22. Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller & J. W. Thatcher (Eds.), Complexity of computer computations (pp. 85–103). New York: Plenum Press.
23. Kessler, B. (1995). Computational dialectology in Irish Gaelic. In Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (pp. 60–66). Dublin: Association for Computational Linguistics.
24. Khuller, S., Moss, A., & Naor, J. (1999). The budgeted maximum coverage problem. Information Processing Letters, 70(1), 39–45.
25. Kirchhoff, K., & Bilmes, J. (2014). Submodularity for data selection in machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 131–141). Doha: Association for Computational Linguistics.
26. Knight, K., & Graehl, J. (1998). Machine transliteration. Computational Linguistics, 24(4), 599–612.
27. Kondrak, G. (2002). Determining recurrent sound correspondences by inducing translation models. In Proceedings of the 19th International Conference on Computational Linguistics. Taipei: Association for Computational Linguistics.
28. Krause, A., & Golovin, D. (2014). Submodular function maximization. In L. Bordeaux, Y. Hamadi, & P. Kohli (Eds.), Tractability: Practical approaches to hard problems (pp. 71–104). Cambridge: Cambridge University Press.
29. Kreyenbroek, P. G. (1992). On the Kurdish language. In P. G. Kreyenbroek & S. Sperl (Eds.), The Kurds: A contemporary overview (pp. 68–83). London: Routledge.
30. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 420–429). New York: ACM.
31. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), 707–710.
32. Lin, H., & Bilmes, J. (2011). A class of submodular functions for document summarization. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 510–520). Portland: Association for Computational Linguistics.
33. Littell, P., Mortensen, D. R., Goyal, K., Dyer, C., & Levin, L. (2016). Bridge-language capitalization inference in Western Iranian: Sorani, Kurmanji, Zazaki, and Tajik. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (pp. 3318–3324). Portorož: European Language Resources Association.
34. MacKenzie, D. N. (1961). Kurdish dialect studies I (London Oriental Series, Vol. 9). London: Oxford University Press.
35. MacKenzie, D. N. (1962). Kurdish dialect studies II (London Oriental Series, Vol. 10). London: Oxford University Press.
36. Malmasi, S. (2016). Subdialectal differences in Sorani Kurdish. In Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (pp. 89–96). Osaka: The COLING 2016 Organizing Committee.
37. McCarus, E. N. (2009). Kurdish. In G. Windfuhr (Ed.), The Iranian languages (Routledge Language Family Series, pp. 587–633). London: Routledge.
38. Minoux, M. (1978). Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer (Ed.), Optimization techniques (Lecture Notes in Control and Information Sciences, Vol. 7, pp. 234–243). Berlin: Springer.
39. Mor, M., & Fraenkel, A. S. (1982). A hash code method for detecting and correcting spelling errors. Communications of the ACM, 25(12), 935–938.
40. Muth, R., & Manber, U. (1996). Approximate multiple string search. In Proceedings of the 7th Annual Symposium on Combinatorial Pattern Matching (Lecture Notes in Computer Science, Vol. 1075, pp. 75–86). Berlin: Springer.
41. Nakov, P., & Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 301–305). Jeju Island: Association for Computational Linguistics.
42. Navarro, G. (2001). A guided tour to approximate string matching. ACM Computing Surveys, 33(1), 31–88.
43. Nemhauser, G. L., & Wolsey, L. A. (1978). Best algorithms for approximating the maximum of a submodular set function. Mathematics of Operations Research, 3(3), 177–188.
44. Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 14(1), 265–294.
45. Nerbonne, J. (2009). Data-driven dialectology. Language and Linguistics Compass, 3(1), 175–198.

46. Nerbonne, J., & Heeringa, W. (1997). Measuring dialect distance phonetically. In J. Coleman (Ed.), Computational phonology: Third meeting of the ACL special interest group in computational phonology (pp. 11–18). Madrid: Association for Computational Linguistics.
47. Nishino, M., Suzuki, J., & Nagata, M. (2016). Phrase table pruning via submodular function maximization. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 406–411). Berlin: Association for Computational Linguistics.
48. Öpengin, E., & Haig, G. (2014). Regional variation in Kurmanji: A preliminary classification of dialects. Kurdish Studies, 2(2), 143–176.
49. Paul, L. (1998). The position of Zazaki among West Iranian languages. In N. Sims-Williams (Ed.), Proceedings of the Third European Conference of Iranian Studies, Part I: Old and Middle Iranian Studies (pp. 163–177). Wiesbaden: Reichert.
50. Ristad, E. S., & Yianilos, P. N. (1998). Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(5), 522–532.
51. Sheyholislami, J. (2011). Kurdish identity, discourse, and new media (The Palgrave Macmillan Series in International Political Communication). New York: Palgrave Macmillan.
52. Sheykh Esmaili, K. (2012). Challenges in Kurdish text processing. arXiv preprint arXiv:1212.0074.
53. Sheykh Esmaili, K., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S., & Hakimi, S. (2013). Building a test collection for Sorani Kurdish. In Proceedings of the 2013 ACS International Conference on Computer Systems and Applications (pp. 1–7). Ifrane: IEEE.
54. Sheykh Esmaili, K., & Salavati, S. (2013). Sorani Kurdish versus Kurmanji Kurdish: An empirical comparison. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 300–305). Sofia: Association for Computational Linguistics.
55. Sheykh Esmaili, K., Salavati, S., & Datta, A. (2014). Towards Kurdish information retrieval. ACM Transactions on Asian Language Information Processing, 13(2), Article 7.
56. Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. Proceedings of the American Philosophical Society, 96(4), 452–463.
57. Thackston, W. M. (2006a). Kurmanji Kurdish: A reference grammar with selected readings. Cambridge, MA: Harvard University. [Self-published teaching text, undated; commonly cited as 2006.]
58. Thackston, W. M. (2006b). Sorani Kurdish: A reference grammar with selected readings. Cambridge, MA: Harvard University, Iranian Studies. [Self-published teaching text; carries a 2006 copyright line.]
59. Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (pp. 2214–2218). Istanbul: European Language Resources Association.
60. Ukkonen, E. (1985). Algorithms for approximate string matching. Information and Control, 64(1–3), 100–118.
61. Walther, G., & Sagot, B. (2010). Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish. In Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages. Valletta: European Language Resources Association.
62. Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. Physica A: Statistical Mechanics and Its Applications, 389(17), 3632–3639.
63. Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. Annual Review of Linguistics, 1(1), 243–264.
64. Wikimedia Foundation. (2016). Central Kurdish, Northern Kurdish and Zazaki Wikipedia [Database dumps of 20 December 2016]. San Francisco: Wikimedia Foundation.
65. Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. Computational Linguistics, 40(1), 171–202.
66. Zeydanlıoğlu, W. (2012). Turkey's Kurdish language policy. International Journal of the Sociology of Language, 2012(217), 99–125.
67. Zeyneloğlu, S., Sirkeci, I., & Civelek, Y. (2016). Language shift among Kurds in Turkey: A spatial and demographic analysis. Kurdish Studies, 4(1), 25–50.