

Sawtone v4

Technical Report

Omar Kamali

Omneity Labs

OFFICIAL LINKS

Model overview: <https://www.omneitylabs.com/models/sawtone>

Report PDF: <https://www.omneitylabs.com/sawtone-v4/report.pdf>

August 2026

Contents

1. The Word-Form Problem	2
1.1 Why the problem appears in real systems	2
2. Sawtone in One Page	3
2.1 Three system levels	3
2.2 Model family and scale	3
3. Relation to Other Byte and Token-Free Methods	4
4. Evidence Design	5
4.1 How to read the metrics	5
5. Word-Form Geometry	6
5.1 What a useful neighborhood looks like	6
5.2 The same geometry exposes failure	6
6. Context Changes the Job	7
6.1 Sawtone and DistilUSE show complementary strengths	7
7. Transliteration as Candidate Selection	8
7.1 One candidate and Sawtone-selected	8
7.2 Why exact match and CER can disagree	9
8. Moroccan Arabic Results	9
9. Cross-Script Transfer	10
9.1 The Pinyin-Hanzi boundary	11
10. Where Sawtone Enters a System	11
11. Strengths and Limits	12
11.1 Supported by the evaluation	12
11.2 Not established by the evaluation	12
12. Conclusion	12
References	13
Appendix A. Cross-Script Result Table	13

1. The Word-Form Problem

A reader can recognize related word forms even when the writing changes. Many NLP systems do not expose that relation directly.

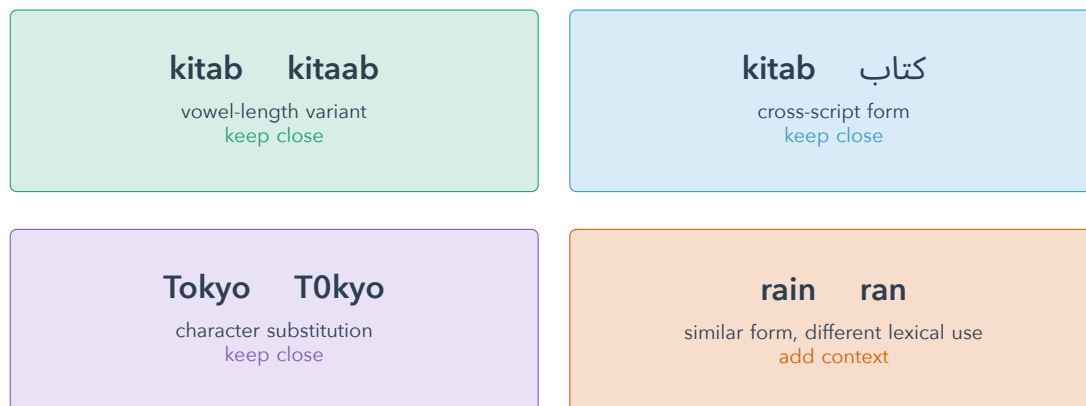


Figure 1: Four examples define the desired behavior. Sawtone should preserve form relations without making the final semantic decision.

These examples define the task:

- **kitab** and **kitaab** should be close because vowel length changes the spelling, not the likely lexical identity.
- **kitab** and **كتاب** should be close for a cross-script search or transliteration task.
- **Tokyo** and **T0kyo** should remain close after a common character substitution.
- **rain** and **ran** can be close in form but different in meaning. Form evidence alone must not make the final decision.

The central question is simple: **Can a system preserve form evidence until the task has enough context to use it correctly?**

1.1 Why the problem appears in real systems

Use case	User input	Existing record or target	Failure without form awareness
Cross-script search	kitab	كتاب	Exact search returns no match
Name matching	T0kyo	Tokyo	Character substitution splits the identity
Dialect normalization	kanmot	kanmout	One convention is treated as an error
Transliteration	one source form	several valid spellings	The first generated form is accepted without comparison
Semantic retrieval	rain	ran	Similar form creates a false positive if context is absent

Sawtone targets the first four cases. The fifth case defines an important limit.

2. Sawtone in One Page

Sawtone reads a complete encoded word form. It creates a state that keeps several kinds of evidence available:

- likely sounded form;
- written form;
- language and script context; and
- lexical or sequence context when the task supplies it.

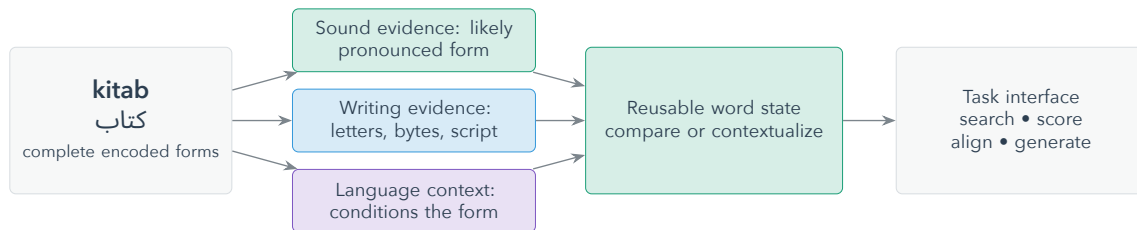


Figure 2: A concrete path through Sawtone. The model keeps complementary evidence available in one reusable word-form state.

The representation does not decide the application result. It supplies evidence to a search system, scorer, classifier, aligner, or generator.

2.1 Three system levels

Level	Input	Output	Example job
Word-form state	One complete form such as <i>kitaab</i>	Comparable form representation	Find <i>kitab</i> , کتاب, and spelling variants
Contextual state	A sequence of word states	Context-sensitive representation	Separate similar forms used with different meanings
Task interface	Form state plus task evidence	Score, label, rank, alignment, or output	Rank transliteration candidates or retrieve a document

Key boundary. Sawtone can say that two forms preserve related evidence. The downstream task decides whether that relation is relevant.

2.2 Model family and scale

Sawtone is a model family rather than one frozen checkpoint. This report uses the v4.4 word encoder for form geometry and the v4.5 contextual stack for sequence tests.

System	Approximate scale	Role in this report
Sawtone v4.4 word encoder	29.1M parameters	Whole-word geometry
Qwen3.5-0.8B	0.8B parameters	Input token-table diagnostic only
Sawtone v4.5 contextual stack	About 80M parameters	Form-aware contextual representation; aggregate estimate
DistilUSE multilingual	About 135M parameters; about 539 MB FP32	Mature multilingual sentence encoder

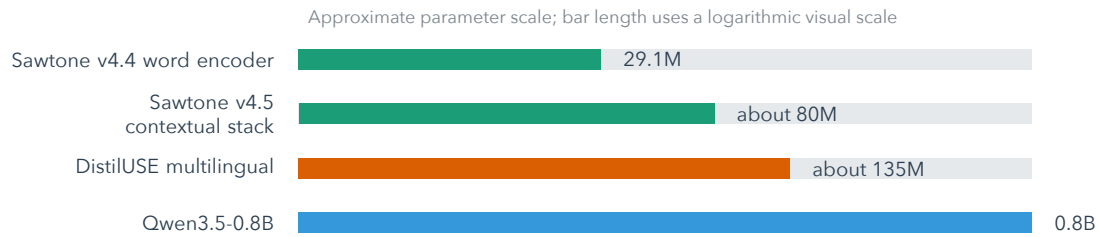


Figure 3: Scale context for the evaluated systems. The Qwen result uses only input embedding rows; the models do not perform the same job.

The systems perform different work. Parameter count gives scale context. It does not give a matched latency or memory result.

3. Relation to Other Byte and Token-Free Methods

The same raw input can enter several different computational paths.

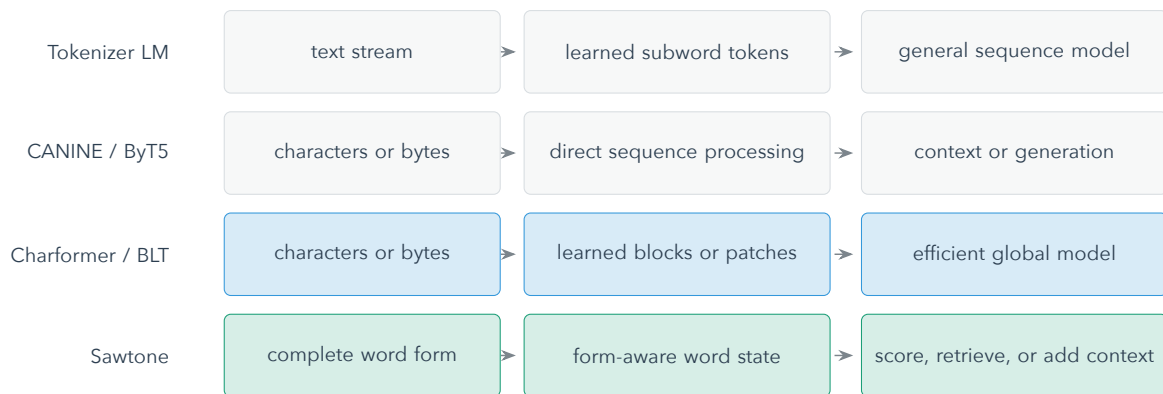


Figure 4: High-level processing paths. Sawtone makes the complete word form reusable before task-specific reasoning.

Method	Unit and objective	Best high-level comparison with Sawtone
Tokenizer-backed Transformer	Learned subword tokens for general sequence modeling	Strong context model, but rare or noisy forms can fragment before word-level comparison
CANINE	Characters with downsampling for tokenization-free encoding	General contextual encoder, not a word-form comparison interface
ByT5	Raw bytes for end-to-end text-to-text modeling	Stronger reference for token-free generation at large scale
Charformer	Learned latent blocks for efficient token-free processing	Learns computational units for the sequence model
Meta BLT	Dynamic byte patches for scalable language modeling	Allocates global computation over learned patches; targets LM efficiency
Sawtone	Complete encoded word forms with optional sequence context	Exposes a reusable form-aware state before task-specific reasoning

The distinction is the first object that the system makes reusable. BLT, ByT5, CANINE, and Charformer redesign general sequence processing. Sawtone first makes the complete word form available for comparison. A later model can then add sentence context or generation.

Use the other methods when: the primary job is end-to-end language modeling, broad contextual encoding, or generation.

Use Sawtone when: the primary job depends on spelling, sound, script variation, transliteration, or form-sensitive retrieval.

4. Evidence Design

The evaluation uses four questions. Each question tests a different system level.

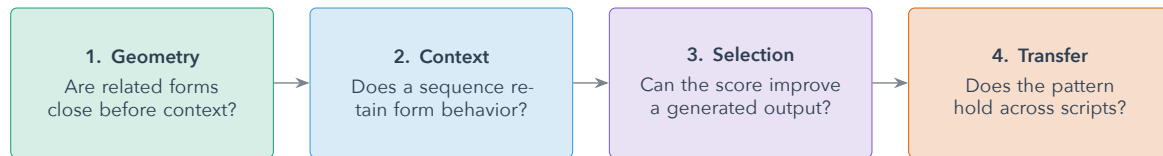


Figure 5: The evidence follows the representation from word geometry to contextual use and cross-script transfer.

Question	Comparison	Main measure
Does the base representation preserve form relations?	Sawtone v4.4 vs sampled Qwen input-token rows	Pairwise cosine similarity
Does context retain form-sensitive behavior?	Sawtone v4.5 vs DistilUSE	Spearman, P@1, MAP, separation
Can the representation improve a generated output?	One candidate vs Sawtone selection from up to eight	Exact match, CER, script validity
Does the behavior transfer across scripts?	Seven additional language-script sets	Exact match, CER, script validity

No overall score combines these tracks. Each track answers a different question.

4.1 How to read the metrics

Metric	Reader question	Important limitation
Cosine similarity	Are two representations close?	Close form does not prove identical meaning
Normalized exact match	Does the output equal the reference convention?	A plausible spelling variant still counts as wrong
Character error rate (CER)	How much character editing remains?	Low error does not prove lexical correctness
Target-script validity	Is the output in the expected script?	Correct script does not prove the correct word
P@1 or MAP	Did retrieval rank the right item highly?	The result depends on the task set and relevance labels

5. Word-Form Geometry

5.1 What a useful neighborhood looks like

The base diagnostic compares complete Sawtone word vectors with pooled input-token rows from Qwen3.5-0.8B. It tests the representation before sentence reasoning.

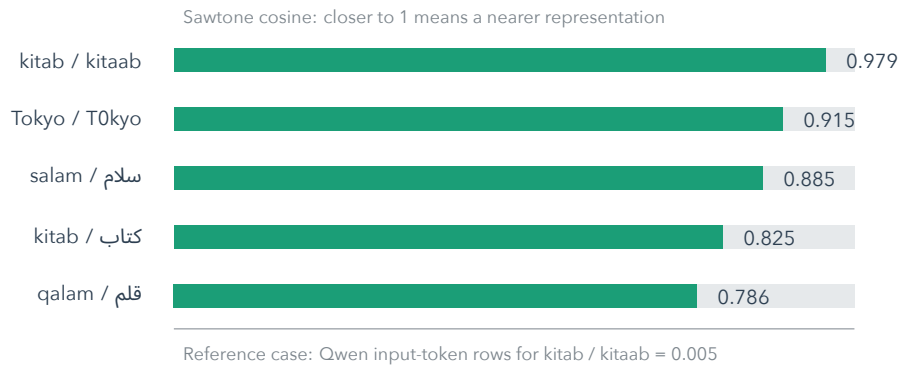


Figure 6: Concrete word neighborhoods. Sawtone keeps several spelling and cross-script variants close.

The positive cases cover four different changes:

Pair	Sawtone cosine	Change represented
kitab / kitaab	0.979	Vowel length
kitab / كتاب	0.825	Latin-Arabic script change
qalam / قلم	0.786	Cross-script form with omitted short vowels
salam / سلام	0.885	Cross-script greeting form
Tokyo / T0kyo	0.915	Character substitution

For kitab / kitaab, the pooled Qwen token-table cosine is 0.005. The comparison concerns the input table, not the full Qwen transformer.

5.2 The same geometry exposes failure

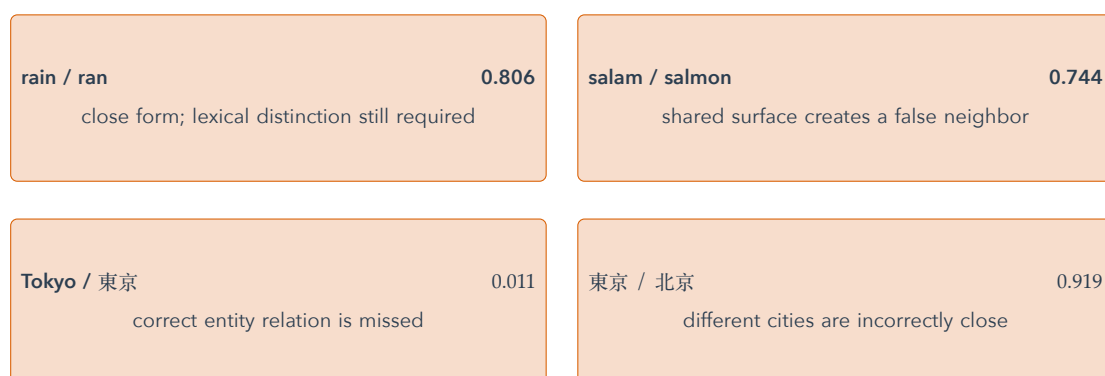


Figure 7: The failure examples are as important as the positive examples. Form geometry needs semantic or entity evidence at decision time.

Pair	Sawtone cosine	Why the result is a problem
rain / ran	0.806	Related form can hide a lexical difference
salam / salmon	0.744	Shared surface material can create a false neighbor
Tokyo / 東京	0.011	The correct cross-script entity relation is missed
東京 / 北京	0.919	Two different Han-script cities are placed too close

This is the operating boundary. The representation is useful for candidate discovery. Semantic context or domain rules must make the final identity decision.

6. Context Changes the Job

Form similarity is useful until two similar forms carry different meanings.

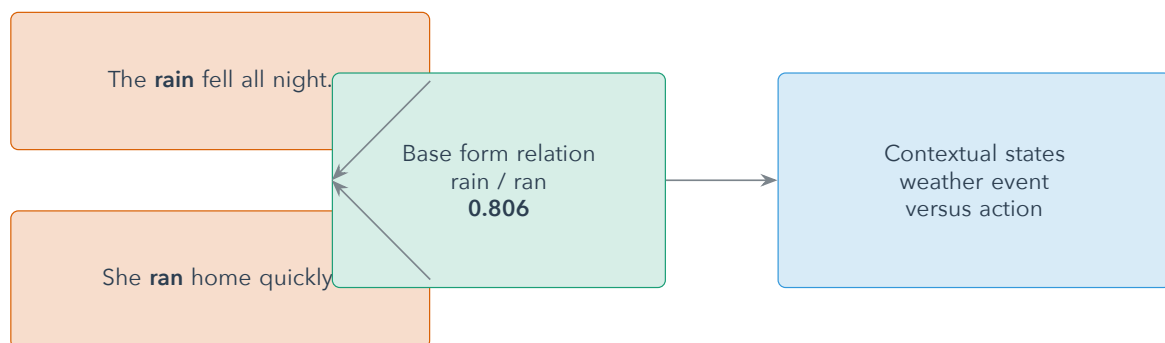


Figure 8: Illustrative context path. A useful form relation becomes a semantic error if the application stops before context.

The contextual stage operates above Sawtone word states. It lets neighboring words change the representation while the base form evidence remains available.

6.1 Sawtone and DistilUSE show complementary strengths

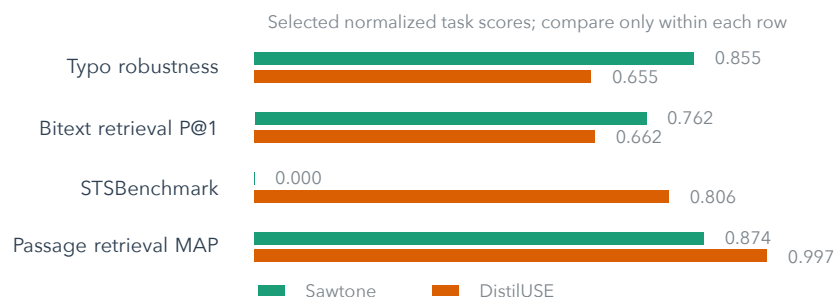


Figure 9: The task profile matters more than one aggregate winner. Sawtone leads on form-sensitive tests; DistilUSE leads on broad semantic tests.

Task	Sawtone v4.5	DistilUSE	What the comparison shows
SimLex-999 Spearman	0.5709	0.5037	Sawtone leads on this word-similarity set
Bitext retrieval P@1	0.7623	0.6615	Sawtone leads on the evaluated cross-lingual retrieval set
Darija Arabic-Arabizi P@1	0.7000	0.5000	Form and script variation matter directly
Typo robustness	0.8546	0.6550	Sawtone retains more form stability
Multilingual cluster separation	0.4842	0.2110	Sawtone separates the evaluated clusters more strongly
RareWord Spearman	0.3744	0.5174	DistilUSE has stronger rare-word semantics
STSBenchmark Spearman	-0.0143	0.8062	Sawtone is not yet a competitive sentence-similarity encoder
Passage retrieval MAP	0.8739	0.9972	DistilUSE remains stronger for broad semantic retrieval

The result is not one winner. Sawtone leads where form and script structure dominate. DistilUSE leads where sentence semantics and passage relevance dominate.

7. Transliteration as Candidate Selection

Transliteration is not a character substitution table. A source can omit vowels. Users can follow different romanization conventions. A reference usually contains only one accepted spelling.

7.1 One candidate and Sawtone-selected

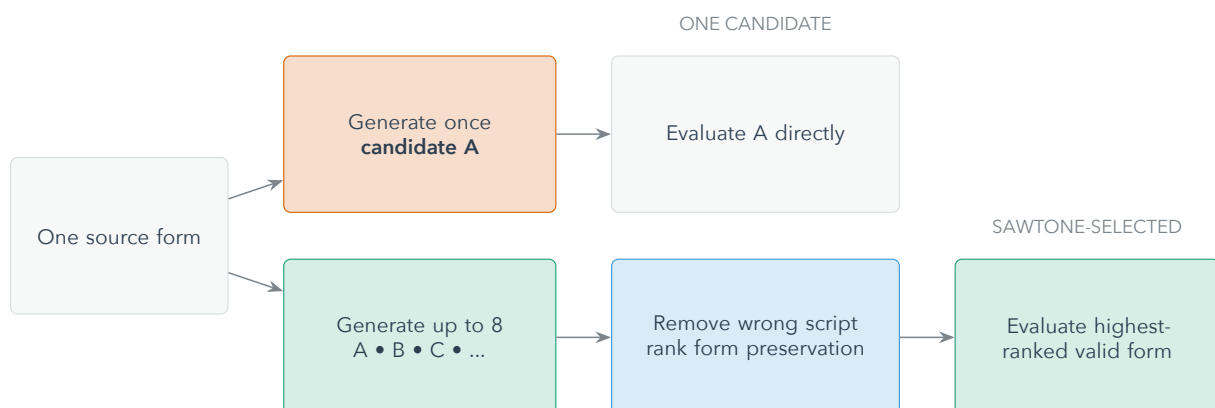


Figure 10: The generator is fixed. The second condition adds candidate coverage, a script check, and Sawtone ranking.

The two conditions use the same generator:

- **One candidate:** generate one output and evaluate it directly.

- **Sawtone-selected:** generate up to eight outputs, remove outputs in the wrong script, rank the rest by form preservation, and evaluate the highest-ranked output.

The gain comes from candidate coverage plus selection. It does not come from a larger generator.

7.2 Why exact match and CER can disagree

The evaluated outputs contain convention-level differences such as these:

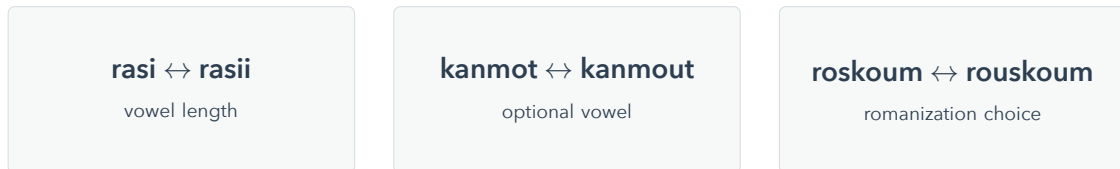


Figure 11: Observed error types in the Moroccan evaluation. Exact match treats each pair as different even when the form change is small.

An exact-match metric treats each pair as different. CER shows that only a small edit remains. Sawtone asks a third question: does the alternative preserve the source form?

8. Moroccan Arabic Results

The Moroccan evaluation uses 500 directional samples from a controlled DODa-derived subset.

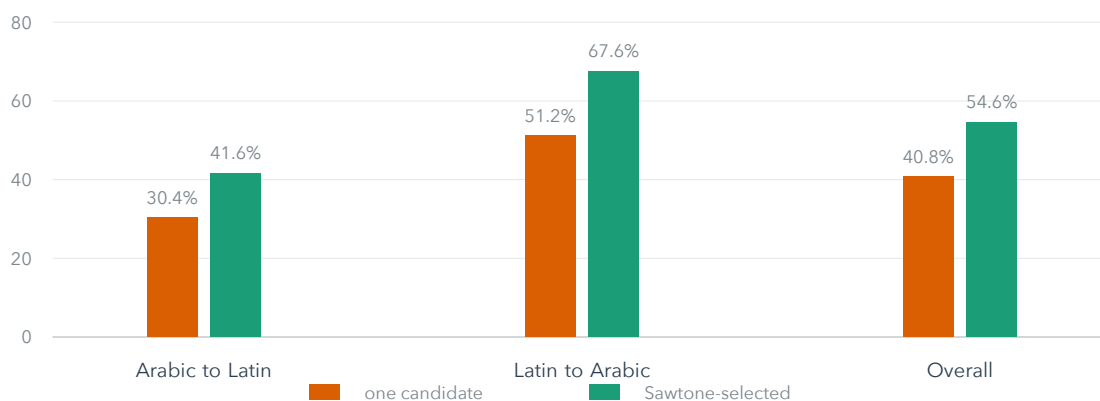


Figure 12: Normalized exact match on 500 Moroccan directional samples. Sawtone selection improves both directions.

Direction	Exact: one candidate	Exact: Sawtone-selected	CER: one candidate	CER: Sawtone-selected
Arabic to Latin	30.4%	41.6%	0.174	0.124
Latin to Arabic	51.2%	67.6%	0.143	0.078
Overall	40.8%	54.6%	0.158	0.101

The strongest gain appears in Latin-to-Arabic exact match: +16.4 percentage points. Overall target-script validity reaches 100% under Sawtone selection.

What improved: the candidate set often contains a form that better preserves the source.

What remains: spelling convention, optional vowels, doubled letters, and generator coverage.

9. Cross-Script Transfer

The broader evaluation contains 3,000 predictions across Bengali, Hindi, Punjabi, Tamil, Urdu, Malay Jawi, and Pinyin-Hanzi sets.

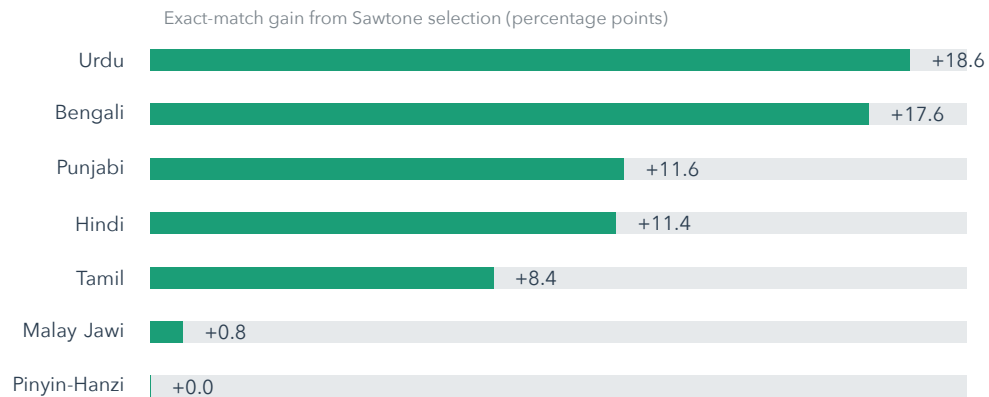


Figure 13: Exact-match gain by set. CER also improves in every set, including Pinyin-Hanzi where exact match remains zero.

Set	Exact gain	CER reduction	Main reading
Bengali	+17.6 pp	0.058	Strong improvement under both measures
Hindi	+11.4 pp	0.041	Candidate selection helps both directions in aggregate
Punjabi	+11.6 pp	0.072	Largest CER reduction among the Indic sets
Tamil	+8.4 pp	0.050	Graded improvement remains visible in CER
Urdu	+18.6 pp	0.074	Largest exact-match gain and CER reduction
Malay Jawi	+0.8 pp	0.010	Candidate coverage or convention remains the main limit
Pinyin-Hanzi	0.0 pp	0.045	Form evidence cannot recover lexical identity

Aggregate exact match rises from 40.7% to 52.1%. Aggregate CER falls from 0.259 to 0.205.

9.1 The Pinyin-Hanzi boundary

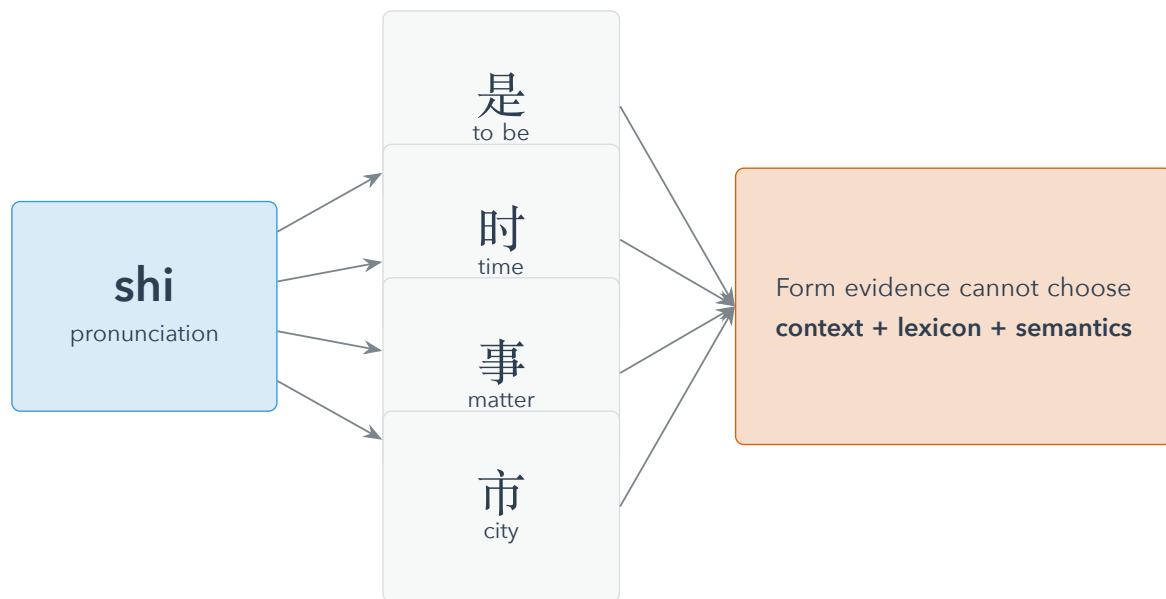


Figure 14: Illustrative Pinyin ambiguity. Correct pronunciation and script still do not identify the intended character.

A pronunciation such as shi can correspond to several characters and meanings. A form scorer can check pronunciation and script. It cannot infer the intended lexical item without context or a lexicon.

In the evaluated set, exact match remains 0.0% in both conditions. Script validity improves from 27.6% to 62.4%. This is useful diagnostics, but not successful lexical recovery.

10. Where Sawtone Enters a System

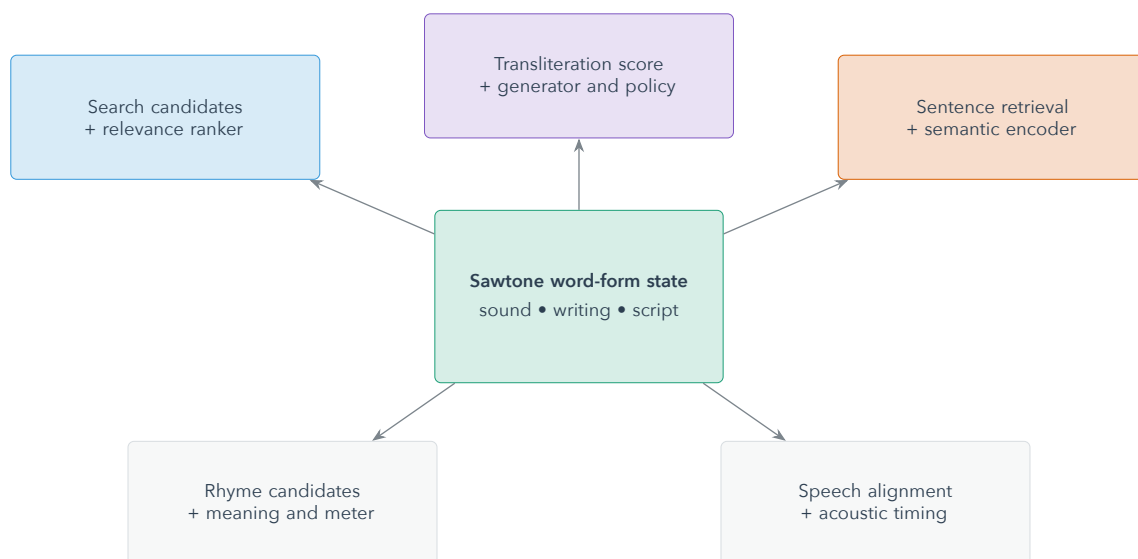


Figure 15: Sawtone improves the form-sensitive stage. Each application adds the evidence needed for its final decision.

Application	Sawtone supplies	The application must still supply
Cross-script search	Form-aware candidates	Relevance, entity, and domain ranking
Transliteration	Source-candidate form score	Candidate generator and orthographic policy
Dataset quality	Suspicious or aligned pair score	Audit labels and acceptance rules
Rhyme and writing tools	Sound-aware neighbors	Meaning, meter, style, and user choice
Speech-text alignment	Text-side form evidence	Acoustic timing and alignment confidence
Sentence retrieval	Form-aware features	A strong semantic sentence or passage encoder

Use Sawtone as the first-stage representation when form variation defines the recall problem. Add semantic or task-specific evidence before the final decision.

11. Strengths and Limits

11.1 Supported by the evaluation

1. Sawtone preserves several spelling, typo, and cross-script relations at the word-form level.
2. The contextual stack retains useful form-sensitive behavior.
3. Candidate selection improves exact match and CER across several writing systems.
4. One scoring interface transfers across multiple language-script pairs.

11.2 Not established by the evaluation

1. Sawtone is not a universal semantic or sentence encoder.
2. Form similarity does not prove lexical identity.
3. Candidate selection cannot recover a form that the generator did not produce.
4. Pinyin-to-Hanzi conversion needs lexical and semantic disambiguation.
5. The current comparisons do not provide matched hardware latency or memory measurements.
6. Search, rhyme, alignment, and other application interfaces need their own public task benchmarks.

12. Conclusion

Sawtone preserves evidence that multilingual systems often lose early: how a word is written, how it can sound, and how its form changes across scripts.

The useful mental model is a two-stage system. Sawtone finds and scores form relations. Context, semantics, and task policy decide what those relations mean.

The results support this role. Sawtone creates useful neighborhoods for spelling and script variants. Its contextual stack leads on several form-sensitive tests. Its candidate scorer improves transliteration across Moroccan Arabic and additional scripts. The same tests also show the boundary: form alone cannot resolve lexical meaning or logographic identity.

References

- O. Kamali, “[Sawtone: A Universal Framework for Phonetic Similarity and Alignment Across Languages and Scripts](#)”, *Lingua Posnaniensis*.
- Omneity Labs, [Sawtone model overview](#).
- A. Outchakoucht and H. Es-Samaali, “[Moroccan Dialect - Darija - Open Dataset](#)”.
- Y. Madhani et al., “[Aksharantar: Open Indic-language Transliteration Datasets and Models for the Next Billion Users](#)”.
- U. Hermjakob, J. May, and K. Knight, “[uroman: Out-of-the-box Universal Romanization Tool](#)”.
- L. Xue et al., “[ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models](#)”.
- J. H. Clark et al., “[CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation](#)”.
- Y. Tay et al., “[Charformer: Fast Character Transformers via Gradient-based Subword Tokenization](#)”.
- N. Reimers and I. Gurevych, “[Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#)”.
- A. Pagnoni et al., “[Byte Latent Transformer: Patches Scale Better Than Tokens](#)”.
- Qwen Team, [Qwen3.5-0.8B Base model card](#).
- Sentence-Transformers, [DistilUSE multilingual model card](#).

Appendix A. Cross-Script Result Table

Set	n	Exact: one candidate	Exact: Sawtone-selected	Change	CER: one candidate	CER: Sawtone-selected
Bengali	500	54.4%	72.0%	+17.6 pp	0.122	0.063
Hindi	500	55.6%	67.0%	+11.4 pp	0.123	0.082
Punjabi	500	40.6%	52.2%	+11.6 pp	0.210	0.138
Tamil	500	40.8%	49.2%	+8.4 pp	0.187	0.137
Urdu	500	41.6%	60.2%	+18.6 pp	0.170	0.096
Malay Jawi	250	22.8%	23.6%	+0.8 pp	0.208	0.198
Pinyin-Hanzi	250	0.0%	0.0%	0.0 pp	1.271	1.226
Overall	3,000	40.7%	52.1%	+11.3 pp	0.259	0.205