Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Computational Morphology for Zazaki

Veysel Yildizhan

vyildizha@outlook.com

Independent Researcher

Sina Ahmadi

University of Zurich

Research Article

Keywords: Zazaki, low-resource languages, computational morphology, spell checking, Hunspell

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10490705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Computational Morphology for Zazaki

Veysel Yildizhan[1†] and Sina Ahmadi[2]

[1]Independent Researcher, Diyarbakır, Turkey.
[2]University of Zurich, Zurich, Switzerland.

Contributing authors: vyildizha@outlook.com; sina.ahmadi@uzh.ch;
[†]This article is derived from the first author's doctoral dissertation: *Zazaki için Hunspell: Vate Dergisi Derlemi Temelli Morfolojik Analizör ve Yazım Denetleyicisi* [Hunspell for Zazaki: A Morphological Analyzer and Spell Checker Based on the Vate Journal Corpus] (Thesis No. 993524), Dicle University, 2025. Available from the Turkish Council of Higher Education (YÖK) National Thesis Center: https://tez.yok.gov.tr. ORCID: V.Y. 0000-0002-0419-2789; S.A. 0000-0001-7904-6551.

Abstract

Zazaki, also known to its speakers as Kirdkî, Dimilî, and Kirmanckî, is a Kurdic language classified among the Northwestern Iranian languages and spoken by an estimated three to five million people in eastern Turkey. The 2010 UNESCO Atlas of the World's Languages in Danger classifies it as vulnerable; scholarly work on the language is limited, and it has had almost no technological support. This study describes the inflectional morphology and morphophonology of Zazaki from a computational perspective, with the aim of building a morphological analyzer and spell checker. Drawing on Korpusê Kovara Vateyî (KKV), a new corpus of about 1.7 million tokens compiled from 75 issues of the journal *Vate*, and on the dictionaries of the Vate Working Group, we present Zazaki Hunspell, a rule-based morphological analyzer and spell checker built on a lexicon of 43,322 flagged entries and 1,023 affix rules. On a manually annotated test set, the system reaches 92.85% lexical coverage, a lemmatization $\boldsymbol{F_1}$ of 87.30%, and a spelling error detection $\boldsymbol{F_1}$ of 90.81%. The corpus, lexicon, rule set, and evaluation framework are released under an open-source license.

Keywords: Zazaki, low-resource languages, computational morphology, spell checking, Hunspell

1 Introduction

Computational morphology develops formalisms and algorithms for analyzing and generating word forms (Aronoff & Fudeman, 2012); (Oflazer, 2022, p. 29). Finite-state methods, beginning with Koskenniemi (1983)'s two-level morphology, have been applied successfully to morphologically rich, low-resource languages such as Pashto (Zuhra & Khan, 2009) and Irish (Uí Dhonnchadha, Ní Pháidín, & van Genabith, 2003); for high-resource languages, the state of the art is set by neural models that require large amounts of annotated training material (Cotterell et al., 2016; Vylomova et al., 2020); (Baxi & Bhatt, 2024, p. 12). For low-resource languages, defined by a chronic scarcity of large, well-annotated datasets, standardized tools, and basic linguistic resources (Pakray, Gelbukh, & Bandyopadhyay, 2025, p. 185), rule-based systems grounded in explicit linguistic formalization remain a viable alternative (Hurskainen, 2023, p. 236).

The language at the center of this paper is Zazaki, spoken by an estimated three to five million people in the eastern regions of Turkey (Çağlayan, 2024; Zilan, 2020) and classified as vulnerable in the 2010 UNESCO Atlas of the World's Languages in Danger (UNESCO, 2010). Its speakers have historically referred to their language and to themselves with terms that directly mean *Kurdish/Kurd*, namely Kirdkî and Kirmanckî (Dogan, 2024). Zazaki is generally classified in the Northwestern Iranian group together with Kurmanjî and Soranî (Paul, 2009); its exact genealogical relationship to the Kurdish varieties has been debated since Mann and Hadank (1932). Rather than entering that debate, this study treats Zazaki as a low-resource Kurdic language (Anonby, Taheri-Ardali, et al., 2015–2018) in need of computational support. Typologically, Zazaki is an inflectional, non-agglutinative language with a two-stem verb system, grammatical gender, split ergative alignment in past-tense transitive clauses, and an ezafe construction inflected for gender, number, and case. These properties make its morphology difficult to model computationally. Although several corpora for Zazaki have appeared in recent years (Section 2), the language has so far had neither a morphological analyzer nor a spell checker, and the ongoing standardization efforts have concentrated on compilation, standardization, and reaching their readership; the group's digitization work remains incomplete.

Contributions

This paper summarizes the distinctive features of Zazaki morphology and reviews the available language resources and the constraints they place on tool development. It presents Korpusê Kovara Vateyî (KKV), a raw corpus of about 1.7 million tokens compiled from 75 issues of the journal *Vate*, and Zazaki Hunspell, the first rule-based, open-source morphological analyzer and spell checker for the language, and reports experiments on a manually annotated test set (Section 5).

2 Related Work

The first corpus work on Zazaki was carried out by Ahmadi (2020), who built a corpus for the Zaza–Gorani language family. Another resource that became accessible during the preparation of this study is the Zaza Text Corpus, the first comprehensive digital

text corpus for Zazaki, developed by Arslan (2025). UD Zazaki ZSD, developed by Dogan, Talamo, Vaz, and Verkerk (2025), is an open-access treebank based on field recordings and oral interviews conducted in the Dersim region.

None of the existing resources contains systematically annotated morphological data on this scale, and none of the descriptive tradition reviewed in Section 3 had been computationally formalized before this study. Zazaki speakers therefore lack spell checking in everyday writing environments, and downstream natural language processing (NLP) tasks lack the morphological layer they would normally build on. The neural route is closed to Zazaki for now, for practical reasons: the morphologically annotated data needed to train a competitive sequence-to-sequence or transformer-based analyzer does not yet exist.

Finite-state transducer (FST) systems in Koskenniemi (1983)'s two-level morphology framework offer certain advantages over dictionary-lookup architectures such as Hunspell: they can represent cyclic morphotactic dependencies, handle long suffix chains through compositional transduction, and both generate and analyze surface forms. Their main limitation in this context is the considerable linguistic engineering effort they require, which presupposes a level of formal morphological documentation that is only now being created for Zazaki.

Taken together, these constraints lead to the conclusion that a rule-based lexicon-and-affix architecture is currently the only viable starting point for Zazaki. This choice is not a rejection of neural or finite-state alternatives but an acknowledgment of the current resource ceiling: part of the linguistic infrastructure an FST implementation would require is being laid by this study.

3 Zazaki and Its Morphology

Zazaki is an inflectional language: distinct grammatical categories within a word (person, number, tense, mood) frequently fuse into a single morpheme (Comrie, 2017, p. 73). This section describes the language and its morphology, focusing on the nominal and verbal systems and on the morphophonological processes that a computational model must handle.

3.1 Classification

In dialectal terms, Zazaki divides into four main subgroups, Northern, Central, Southern, and Eastern, which differ at the phonological, morphological, and lexical levels and correspond closely to the names Kirmanckî, Kirdkî, Dimilî, and Zazakî used by speakers (Arslan, 2022; Dogan, 2024; Malmîsanij, 2021; Paul, 2009). The leading standardization initiative, the Vate Working Group (GXV), prefers the name Kirmanckî, giving Zazakî in parentheses (Grûba Xebatê ya Vateyî, 2023). Figure 1 shows the geographical distribution of the principal dialect zones and the local glossonyms (Kirmanckî, Kirdkî, Zazakî, Dimilî) associated with each area.

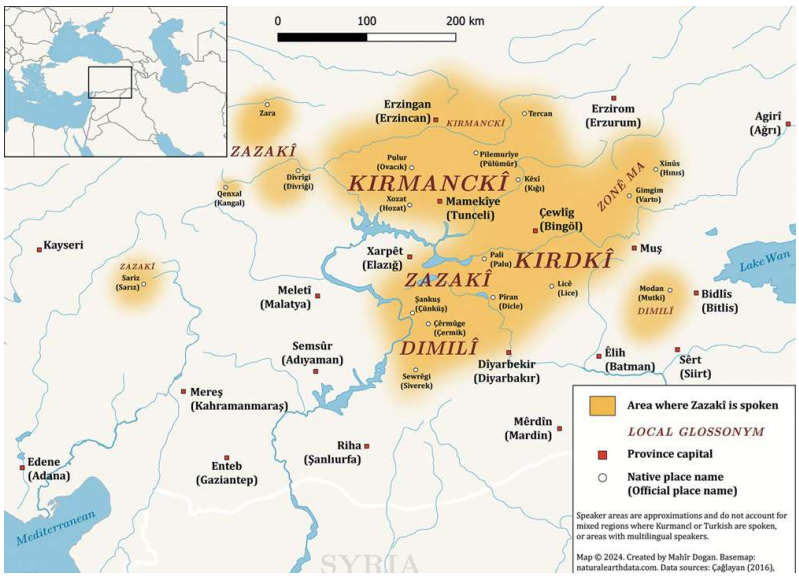

Fig. 1 Geographical distribution of the principal Zazaki dialect zones. Map reproduced from Dogan (2024, p. 253), published open access by De Gruyter Mouton.

3.2 Writing and Standardization

Setting aside the Arabic-based script used in the classical works and rarely consulted since, Zazaki is written mainly in the Jacobson (Keskin, 2015) and Bedirxan alphabets; some 14 different alphabets, almost all Latin-based, have been proposed or used for the language (Malmîsanij, 2015, pp. 23–24). This study uses the Bedirxan alphabet, the one adopted by the journal *Vate*, our source material. Written Zazaki has moved toward a degree of standardization through the work of Gruba Xebate ya Vateyî (GXV) and its organ Kovara Vate (KV) (Gruba Xebate ya Vateyî, 2023); the lexicon and rule set presented here follow this written norm.

3.3 Grammatical Overview

The grammatical description of Zazaki spans more than a century and a half, from the first sketches of Müller (1865) and Mann and Hadank (1932), through the reference grammars of Todd (1985, 2008), Paul (1998), and Selcan (1998), to the recent work produced around the Vate standardization effort (Çem, 2003; Kırkan, 2021; Malmîsanij, 2015). This study relies mainly on the last group and follows their orthographic and analytical conventions.

3.4 Nouns

The Zazaki nominal system is organized along the axes of gender, number, case, definiteness, and animacy. The interaction of these categories produces complex morphological patterns, particularly in the ezafe construction. Table 1 summarizes the main nominal affixes with examples.

Gender

Nouns fall into two grammatical genders, masculine and feminine (Malmîsanij, 2015, p. 67); (Kırkan, 2017, p. 103). The distinction is morphologically visible in singular nouns and neutralized in the plural. Masculine nouns typically end in a consonant (e.g., *merdim* 'man', *kitab* 'book'), while feminine nouns frequently end in {-e} (e.g., *dewe* 'village'), {-a} (e.g., *kêna* 'girl'), or {-î} (e.g., *kardî* 'knife'); these are tendencies, not exceptionless rules.

Grammatical gender surfaces through agreement in adjectival and ezafe constructions: *gayo sur* (red ox) with a masculine noun, *dewa girde* (big village) with a feminine one. As Corbett (1991) notes, gender assignment draws on formal, semantic, and phonological information. In Zazaki, for instance, abstract nouns are largely masculine and fruit names largely feminine (Çem, 2003).

Number

Nouns inflect for singular and plural. The nominative plural is formed with the suffix {-î} and the oblique plural with the suffix {-an}: *dest* (hand) → *destî* (hands, nominative), *destî* (hand, singular, oblique), *destan* (hands, plural, oblique). Plural formation varies with the phonological shape of the stem.

Case

Nouns inflect for two basic cases: the nominative (direct/absolutive) and the oblique (accusative); some studies treat the vocative as an independent case (Malmîsanij, 2015, p. 81), while others also recognize a locative and a dative (Todd, 2008, p. 37). The nominative marks the subject in present-tense clauses, the object in past-tense transitive clauses, and the subject of intransitive clauses. The oblique marks the definite object in the present tense, the subject of a transitive verb in the past tense, and the possessor in possessive noun phrases. In masculine singular nouns the nominative–oblique contrast is realized by the suffix {-î}, whereas in feminine singular nouns the two cases are mostly identical in form.

Definiteness

Definiteness is marked morphologically rather than by separate articles. Indefiniteness is expressed either by placing *yew* 'one' before the noun or by adding allomorphs such as {-êk}/{-ê}/{-ên}/{-êr}: *cinî* (woman) → *cinîyêk/cinîyê* 'a woman'. In the plural, *tayê* 'some' serves as the indefiniteness marker. Definiteness is expressed through the oblique case form; this pattern coincides with Zazaki's system of Differential Object Marking (DOM): *Mi nan werd.* (I ate bread.) ∼ *Mi nanî werd.* (I ate the bread.).

Ezafe

The ezafe, a characteristic feature of the Iranian languages, is the most complex part of the Zazaki morphological system. Historically traceable to the Old Persian relative pronoun *hya*, the ezafe links an adjective or a possessor noun to the head noun (Arslan, 2022; Karimi, 2007). The form of the ezafe morpheme depends on the gender and number of the head noun and on the case of the whole noun phrase. In

Table 1 Main nominal affixes of Zazaki with examples attested in the KKV corpus

Function	Affix	Example	Gloss
Nominative plural	{-î}	*dest* → *destî*	‘hand’ → ‘hands’
Oblique plural	{-an}	*dest* → *destan*	‘hands (oblique)’
Oblique singular (masc.)	{-î}	*nan* → *nanî*	‘the bread (oblique)’
Indefiniteness	{-êk}/{-ê}/{-ên}/{-êr}	*cinî* → *cinîyêk*	‘a woman’
Typical feminine endings	{-e}, {-a}, {-î}	*dewe*, *kêna*, *kardî*	‘village’, ‘girl’, ‘knife’
Ezafe (masc., nominative)	{-o}/{-ê}	*çim-o vêşan*	‘hungry eye’
Ezafe (masc., oblique)	{-ê}	*veng-ê lejekî*	‘the child’s voice’
Ezafe (feminine)	{-a}	*dew-a girde*	‘big village’
Ezafe (oblique, plural)	{-anê}	*rîpel-anê kitabî*	‘the pages of the book (oblique)’
Differential object marking	oblique {-î}	*Mi nan werd.* ∼ *Mi nanî werd.*	‘I ate bread. / I ate the bread.’

adjectival phrases the attributive adjective follows the head noun, an ordering that distinguishes Zazaki from some other Northwestern Iranian languages (Samvelian, 2018). In possessive noun phrases the possessed element stands on the left and the possessor on the right, and the possessor is always in the oblique case: *vengê lejekî* (the child’s voice). Ezafe agreement is most clearly visible in feminine and plural noun phrases and in all oblique constructions. Figure 2 illustrates the allomorph selection conditions with lexical examples; the full nominal inflection inventory is given in Table A1 in Appendix A.

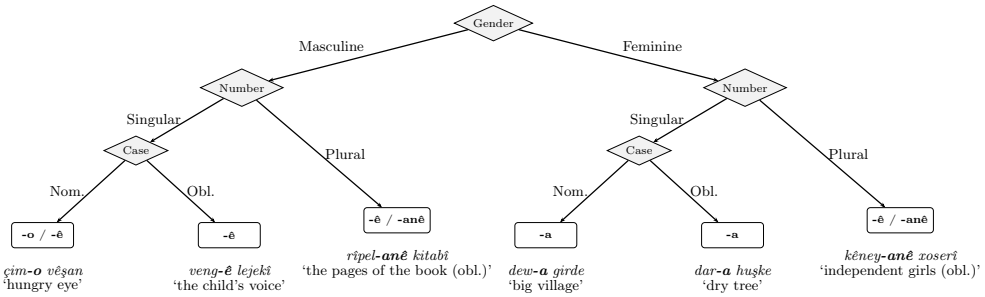

Fig. 2 Ezafe allomorph selection tree in Zazaki: the effect of the head noun’s gender, number, and case on allomorph choice. Allomorph forms follow the GXV orthographic guides; the masculine nominative singular takes {-o} with adjectives and {-ê} with nominal complements.

Adjectival Inflection

Adjectives inflect for the gender, number, and case of the noun they modify, follow the noun, and are linked to it by the ezafe (Malmîsanij, 2015). Gradation is expressed with the suffixes {-êr}/{-tir} and with adverbs such as *hende*, *hîna*, and *tewr*; the adjectival inflection suffixes appear in the adjective columns of Table A1.

3.5 Verbs

The Zazaki verbal system encodes tense-aspect, mood, person, number, and voice, and is shaped throughout by the language's split ergative alignment. Zazaki verbs are built on two distinct stems (Kırkan, 2021, p. 91):

- **Present stem:** the basis of present- and future-tense clauses (Kırkan, 2025; Toosarvandani & Van Urk, 2014), e.g. *wan-* 'read' (cf. subjunctive *biwano*);
- **Past stem:** the basis of past-tense clauses, e.g. *kerd-* 'did' (cf. *mi kerd* 'I did (it)');
- **Infinitive:** formed with the suffix {-ene} or with {-iş}/{-ayîş} (*kerdene* versus *kerdiş*); in the GXV orthographic rules {-ene} is distinguished as the infinitive marker while {-iş} functions as a nominal suffix (Kırkan, 2021, p. 83).

Tables A3 and A4 in Appendix A give the full inflection patterns for a transitive and an intransitive verb, and Table A2 presents the verbal affix inventory.

Tense and Aspect

Zazaki verbs inflect for present, past, and future tense, each of which also encodes perfective and imperfective aspect (Kırkan, 2021, p. 87). The imperfective is formed with the suffixes {-en}/{-n}; this suffixal strategy differs from the prefixal {de-}/{e-} marking of Soranî and Kurmanjî (Kurdo, 2016, p. 178). In past-tense inflection, the gender and number of the object directly determine the inflected form in transitive clauses (Bingöl, 2026).

Mood

Zazaki distinguishes the indicative from the subjunctive-conditional (Kırkan, 2021, p. 173). The latter covers imperative, optative, necessitative, and desiderative functions, and the subjunctive and imperative are formed identically. Negation is realized with {nê-} in the indicative and with {me-} in the prohibitive imperative; the prefix {bi-} forms the subjunctive/optative and is dropped in certain contexts.

Agreement

Zazaki verbs agree with the subject in person and number; the second person singular (in the copula only) and the third person singular (in both pronouns and the copula) additionally distinguish gender. How this agreement is distributed across tense and transitivity is described below.

Split Ergativity

The defining typological feature of Zazaki is its split ergative alignment, in which past and present tenses show different case-marking patterns (Kırkan, 2021, p. 222). In the

present/future, subjects are in the nominative, definite objects are in the oblique, and the verb agrees with the subject (nominative–accusative alignment). In the past tense with transitive verbs, the subject is in the oblique, the object is in the nominative, and the verb agrees with the object in gender and number (ergative–absolutive alignment); the person of the subject plays no role. Past-tense intransitive verbs retain the nominative subject and subject agreement. Figure 3 diagrams these patterns; encoding them as flags in a rule-based analyzer requires knowing the verb's transitivity, its stem-final phoneme, and the agreement patterns of the target dialect.

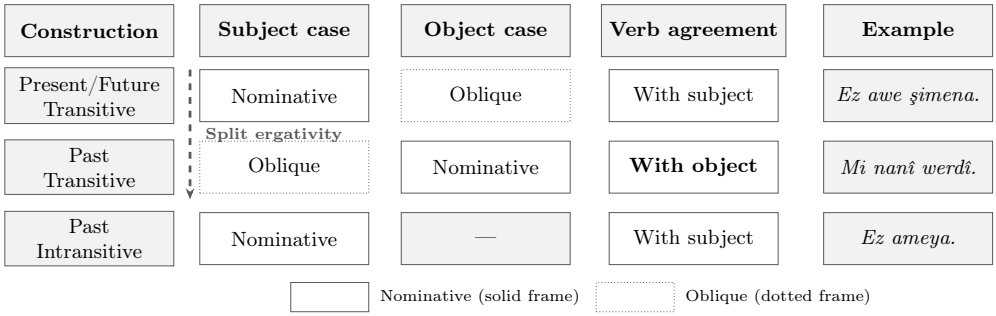

Fig. 3 Split ergative alignment in Zazaki: case-marking and verb agreement patterns across the axes of tense and transitivity. The dashed arrow indicates the structural shift in alignment characterizing the split ergativity between present/future and past transitive constructions.

Voice

Zazaki has active, passive, reflexive, causative, and reciprocal constructions. The passive is formed either morphologically (*werdene* → *werîya* 'was eaten') or lexically with the construction *ame werdiş/werdene*; reflexivity is marked with the pronoun *xo*; the causative is formed with the suffix {-na} or with the auxiliary *dayene*; and reciprocity is expressed with pronouns such as *yewbîn* and *pê* (Kırkan, 2021, pp. 68–77).

3.6 Adverbs and Adpositions

Adverbs express place, time, manner, quantity, or interrogation; many attributive adjectives can be used adverbially without morphological change, and adverbs are derived with suffixes such as {-ane}, {-kanî}, and {-kî} or through reduplication (Kırkan, 2024; Laçin, 2018; Malmîsanij, 2015). Adpositions fall into three categories according to how they attach to the noun: prepositions (e.g. *bi* 'with'), postpositions (e.g. *ra* 'from'), and circumpositions (e.g. *di* … *de* 'in') (Kaçan, 2018; Kırkan, 2021; Malmîsanij, 2015); both classes are largely invariable with respect to inflection.

3.7 Morphophonological Processes

Morphophonological (morphophonemic) processes are regular sound changes that arise at morpheme boundaries (Booij, 2005, p. 317); in Zazaki they occur mainly at stem–suffix and suffix–suffix boundaries, and each requires separate suffix allomorphs

or transformation rules in analyzer design. Four operations are central. Glide epenthesis inserts {-y-} between adjacent vowels (*ga* + {-î} → *ga-y-î* 'oxen'). Syncope deletes stem-final {-e} before plural or ezafe suffixes (*tîre* → *tîr-î*). Epenthesis inserts a vowel to resolve disallowed consonant clusters, particularly in Arabic loanwords (Ar. *hebs* → Zaz. *hepis*) (Aronoff & Fudeman, 2012, p. 76). Assimilation, found mainly at suffix–suffix boundaries and in some dialects, adapts a sound to its neighbor in voicing or place of articulation (Korkmaz, 2007, p. 84). Figure 4 summarizes the four alternation operations.

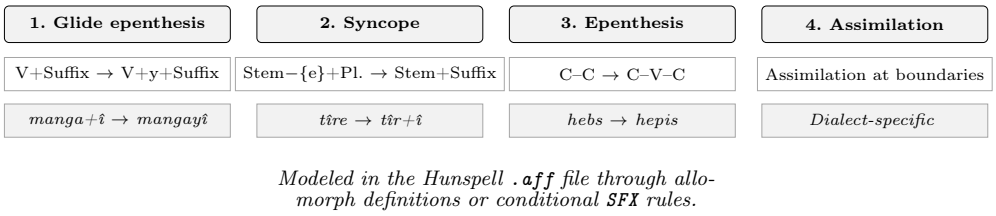

Fig. 4 Morphophonological processes in Zazaki: the four main alternation operations with rules, descriptions, and lexical examples.

3.8 Derivation

Derivation creates new lexical items by adding prefixes, suffixes, and in some cases infixes to roots. The sources consulted treat derivation under three basic categories: nominal, adjectival, and verbal (Kırkan, 2021; Malmîsanij, 1996, 1997; Malmîsanij & Mosa, 2016, 2017; Paul, 2009; Werner, 2012). Table 2 summarizes the main strategies. A more detailed description of Zazaki grammar, including pronouns and other free morphemes, is given in Paul (1998), Todd (2008), and Kırkan (2021).

Table 2 Morphological derivation strategies in Zazaki

Nominal derivation	**Adjectival derivation**	**Verbal derivation**	**Adverbial derivation**
{-i}, {-iye}, {-ey}, {-e} (e.g., *azadi*, *rasti*) {-ij} (origin), {-ey} (abstractness), {-xane}, {-ek}/{-ik} (diminutive)	{bê-} (privative: *bêaqil*, *bêkar*) {-î}, {-în}/{-ên}, {-dar}, {-ên}/{-în} (derived from nouns/adverbs)	**Prefixes:** {a-}, {ci-}, {cêra-}, {pira-}, {we-}, {te-}, {ro-}, {ra-}, {tê-} **Valency:** Intransitive → Transitive (e.g., *remnayene*)	Adjective to adverb {-ane}/{-one}/{-ûne}, {-kanî}, {-kî} Reduplication and converbs

4 Methodology

Given the resource constraints described in Section 2, we adopted a fully rule-based approach built on the Hunspell architecture. The system consists of two core files: a dictionary file (`.dic`) containing word roots and the rule flags applicable to each root, and an affix file (`.aff`) defining what those flags mean, that is, the affixation rules (prefixes: `PFX`; suffixes: `SFX`), the compounding rules, and the morphophonemic alternations (Ooms, 2024). Setting statistical approaches aside, the method concentrates on the following rule-based steps:

- **Lexicon construction:** compiling a dictionary (`.dic` file) that lists Zazaki word roots—nouns, verbs, adjectives, and other categories—together with their rule flags.
- **Morphological rule set development:** encoding a rule set in Hunspell format (`.aff` file) that defines the affixal structure of Zazaki—especially its inflectional suffixes—based on the morphological description in Section 3.
- **Testing and validation:** running the resulting analyzer and spell checker on a prepared test dataset to assess the accuracy and coverage of the rules.

Figure 5 shows the processing pipeline from the data sources to the system's two outputs.

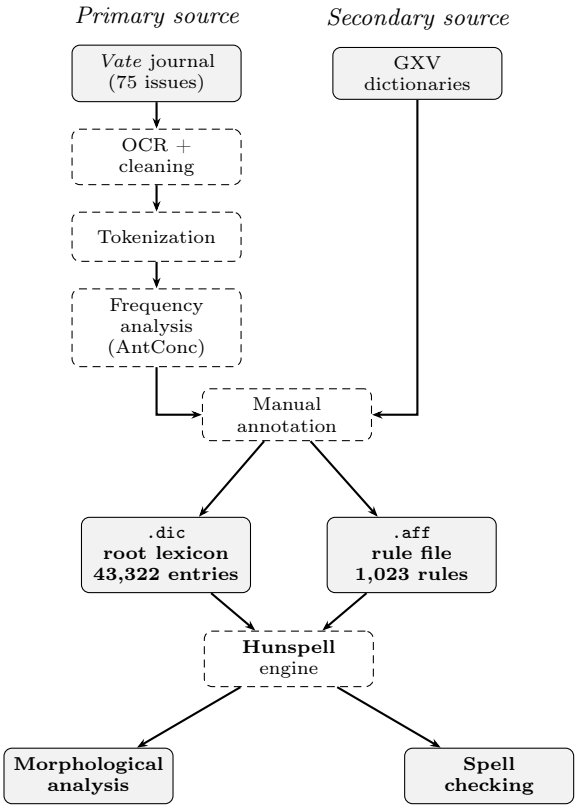

Fig. 5 Architecture of the Zazaki Hunspell system: the processing pipeline from the data sources to the morphological analysis and spell checking outputs.

4.1 Data

As the primary data source we selected Kovara Vate (KV), the journal published continuously since 1997 and central to written Zazaki literature. As the secondary linguistic resource we used the dictionaries of the Vate Working Group (Grûba Xebate ya Vateyî; GXV).

Entries organized as glossaries or word lists were excluded from the corpus; issues available only in print were digitized with OCR, and OCR-induced errors were corrected manually. The texts were split into *tokens* using a standard whitespace-and-punctuation method.

The root lexicon was built in two stages, combining corpus-driven extraction with dictionary transfer. In the first stage, frequency and distribution analyses were run on the KKV corpus with AntConc (Anthony, 2022). The 10,000 most frequent tokens were extracted and annotated manually: each token was examined in context, assigned a part-of-speech (PoS) flag, and stripped of inflectional prefixes and suffixes to reach its citation form. This process produced 3,670 distinct root entries, each verified manually against the published grammatical descriptions of the Vate Working Group.

In the second stage, the full headword list of the Kirmanckî–Tirkî dictionary (Grûba Xebatê ya Vateyî, 2023) was incorporated into the `.dic` file. As the product of roughly three decades of collaborative lexicographic work, these entries were taken to be reliably normalized and were not subjected to further token-level review. The final merged lexicon contains 43,322 flagged entries: 3,670 of them verified at the token level, with the remaining ∼39,600 inheriting the editorial quality controls of the GXV dictionary. This distinction matters for interpreting out-of-vocabulary (OOV) rates in evaluations that target specialized domains not covered by the GXV headword list.

The dataset consists of 75 journal issues spanning 29 genres, with 1,397 texts and 1,658,744 tokens in total (a mean of 1,187.4 tokens per text; median 643). Texts by male authors make up 76.59% of the corpus; summary statistics are given in Table 3, and the full genre and author-gender distributions are shown in Figure 6.

Table 3 Descriptive statistics of the KKV corpus: distribution of text genre, author gender, and token-level measures

Category / Dimension	*n*	%	**Mean**	**Median**	**Min**	**Max**	**Total tokens**
Corpus-level summary							
Total journal issues	75	—	—	—	—	—	—
Total texts	1,397	—	1,187.4	643	21	6,451	1,658,744
Genre distribution (top 5 genres by frequency; the remaining 24 genres are merged)							
Poetry	404	28.92	—	—	—	—	—
Anthology/compilation	224	16.03	2,430	1,977	98	6,451	521,887
Short story	161	11.52	1,634	1,425	112	5,876	263,031
Article	129	9.23	2,069	1,835	95	6,102	266,922
Interview	82	5.87	2,053	1,796	126	5,987	168,363
Other (24 genres)	397	28.43	—	—	—	—	—
Author gender distribution							
Male	1,070	76.59	1,112	—	—	—	—
Collective	132	9.45	1,654	—	—	—	—
Female	117	8.38	1,087	—	—	—	—
Interview (M–M)	54	3.87	2,076	—	—	—	—
Interview (M–F)	15	1.07	1,987	—	—	—	—
Other[a]	9	0.64	—	—	—	—	—

Note. n = number of texts; % = share of the corpus total; token counts are based on whitespace-and-punctuation tokenization. Cells marked '—' indicate categories for which token-level statistics were not computed. [a]Includes female–female pairs and interviews with more than two participants.

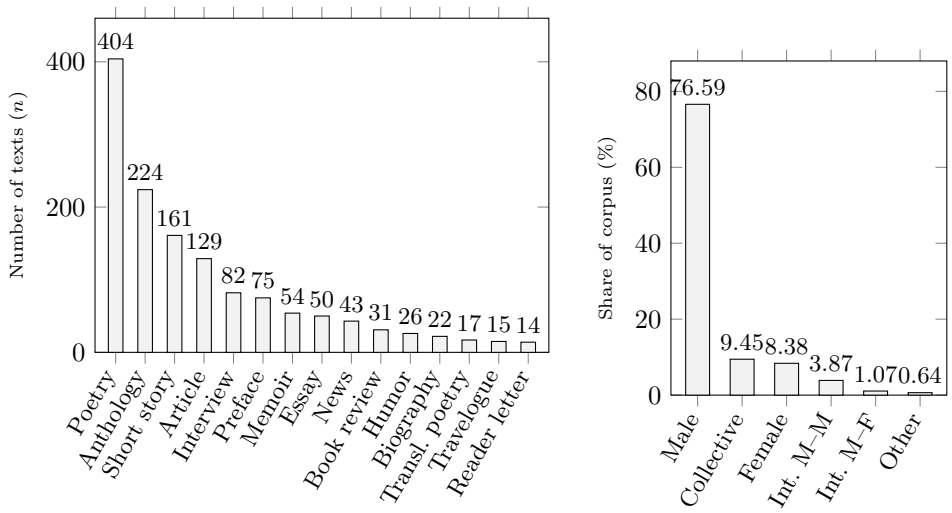

Fig. 6 Composition of the KKV corpus ($N_{\text{total}} = 1{,}397$). Left: distribution of text genres; poetry is the largest category, followed by compilations and short stories. Right: distribution of texts by author gender; texts by male authors make up 76.59% of the corpus.

4.2 Modeling Zazaki in Hunspell

Tag Set

The core tag set consists of thirteen categories: Noun (M for masculine, F for feminine, K for plural-only nouns, a small class lacking a singular), Proper Noun (Z), Verb (V), Adjective (A), Adverb (R), Pronoun (B), Adposition (P), Conjunction (C), Particle (G), Numeral (E), and Punctuation (H). Since transitivity determines which argument controls verb agreement, transitive verbs are flagged T and intransitive verbs I throughout this study; the present stems of all verb types carry the V flag. These part-of-speech categories are attached as flags to each word root in the Hunspell dictionary file (`.dic`), indicating which affix class defined in the `.aff` file the root belongs to and which morphological rules govern it.

Rule Format and Conditions

The general format of these rules in the Hunspell `.aff` file is as follows (Mankier, 2017): the first line defines the rule class:

```
PFX/SFX <flag> <cross_product> <rule_count>
```

followed by one or more lines, each corresponding to a rule variant:

```
PFX/SFX <flag> <strip> <affix> <condition> [morphological_info]
```

The condition mechanism of Hunspell's *PFX* and *SFX* rules models complex linguistic dependencies. For example, a plural suffix that takes a different surface form depending on whether the host noun ends in a vowel or a consonant can be captured in the condition field with regular-expression-like patterns such as *[aeêiîouû]* (any vowel) or *[^aeêiîouû]* (any consonant). Suffix stacking, the successive attachment of several suffixes to a root in a fixed order, is common in Zazaki. Hunspell manages such combinations—among other mechanisms—through the `<cross_product>` parameter of a *PFX*/*SFX* rule: when this parameter is set to *Y* (Yes) (Spylls Project, 2024), it indicates that a form may take further affixes after the first affix has applied. Table 4 parses actual rule examples from the `.aff` file.

Table 4 Actual `SFX/PFX` rule examples from the Hunspell `.aff` file: the function of each field and its Zazaki counterpart.

Rule / Flag	Strip	Affix	Condition	Description and example
`SFX M`	`0`	`î`	`[^aeêiîouû]`	Masculine nominative plural: consonant-final stems. *nan* + {-î} → *nanî*
`SFX M`	`0`	`yî`	`[aeêiîouû]`	Glide epenthesis: vowel-final stems. *ga* + {-yî} → *gayî*
`SFX F`	`e`	`î`	`e`	Feminine plural: {-e} is stripped and {-î} added. *tîre* + {-î} → *tîrî*
`SFX T`	`0`	`o`	`[^aeêiîouû]`	Ergative past, masculine object: consonant-final stem. *kerd* + {-o} → *kerdo*
`SFX T`	`0`	`a`	`[^aeêiîouû]`	Ergative past, feminine object: consonant-final stem. *kerd* + {-a} → *kerda*
`PFX V`	`0`	`bi`	`.`	Subjunctive prefix: all verb stems. {bi-} + *wano* → *biwano*

Note. `Strip`: the portion removed from the stem end; `Affix`: the unit added; `Condition`: the application condition as a regular expression. `[^aeêiîouû]` = any non-vowel character. Rules with the cross-product value set to `Y` allow an affix to attach to an already inflected base.

Modeling Ezafe Agreement

The agreement mechanisms of the ezafe construction are among the most challenging aspects of the Hunspell modeling. The different allomorphs of the ezafe morpheme ({-o}, {-a}, {-ê}, {-anê}, etc.) must be defined as separate affixes in the `.aff` file, and the conditions governing their application must be managed through the gender flags carried by the head noun. Verb-agreement asymmetries likewise require that the flags assigned to verb roots and tense/person suffixes reflect unambiguously which

argument (subject or object) the verb agrees with and under which conditions (tense, transitivity).

Modeling Morphophonemic Alternations

Phonological processes at root–affix boundaries—syncope, epenthesis (for example the {-y-} glide as a linking consonant), and assimilation—are represented in the `.aff` file as allomorphic variants of the affected affixes. For example, the realization of the plural suffix {-î} as {-yî} after a vowel-final root is expressed either as two separate `SFX` definitions or as a single conditional rule encoding both allomorphs. Hunspell's `REP` (replacement) directive can also model certain regular alternations, but its side effects have to be weighed: defining the allomorphs separately and selecting the correct one through root flags or affix conditions is the safer and more transparent approach. Tables 5 and 6 summarize the resulting rule distribution and the full flag inventory.

Table 5 Zazaki morphological classes and rule distribution (.aff)

Class	Flags	Rules (total)	Functions	Example
Verb	T, I, V, W	491	Transitive past (T), intransitive past (I), present (V), irregular (W)	bikerdêne
Noun	M, F	296	Masculine (M), feminine (F)	nanî
Adjective	A	59	Adjectival inflection	şîrîn
Ordinal	E	19	Ordinal inflection	çarin

Note. The four inflectional classes listed here account for 865 of the 1,023 rules in the `.aff` file; the remaining rules cover pronominal inflection, the copula, and other affix classes, plus morphophonemic and suggestion directives. The subjunctive/negation prefix block (79 rules) shared by the verb subclasses V, I, and W is counted once.

Table 6 Alphabetical distribution of morphological flags with examples (N = 43,322) (.dic)

Flag	Description	Count	%	Example
A	Adjective	445	1.03%	şîrîn
B	Pronoun	87	0.20%	ez
C	Conjunction	66	0.15%	û
E	Numeral	62	0.14%	hîrê
F	Feminine gender	16,787	38.75%	maye
FM	Masculine and feminine	108	0.25%	embaz
G	Ezafe	3	0.01%	-ê
H	Punctuation	55	0.13%	!
I	Intransitive past stem	200	0.46%	ame-
K	Plural-only	581	1.34%	ardî
M	Masculine gender	22,470	51.87%	nan
MA	Masculine adjective	6	0.01%	pîl
P	Adposition	82	0.19%	de
R	Adverb	87	0.20%	nika
T	Transitive past stem	247	0.57%	kerd-
V	Verb root / present stem	449	1.04%	wan-
W	Special inflections (copula)	61	0.14%	-o
XF	Primary infinitive (feminine)	443	1.02%	kerdene
XM	Secondary infinitive (masculine)	447	1.03%	kerdiş
Z	Proper noun	635	1.47%	Ehmed

4.3 Test Set Construction and Error Generation

To evaluate the system, we built a manually annotated test dataset of 1,188 tokens from 100 sentences randomly sampled from the KKV corpus. Words were selected with attention to morphological complexity, aiming for a balanced representation of word classes, common and rare morphological constructions, various suffix combinations, and possible morphophonemic alternations. Two annotators with linguistic training, following a written annotation guide, independently segmented each word into its root and morphemes; disagreements were resolved through discussion. The resulting annotations constitute the reference data for evaluation. Table 7 summarizes the composition of the test set.

Misspelled test forms were generated in two stages. In the first stage, a keyboard-layout-based error generation algorithm was applied to the correctly annotated word forms. The algorithm maps each character, including Zazaki-specific characters such as î, ê, û, and ç, to its physically adjacent keys on the Kurdish Q keyboard layout and applies one of the following four error types to each token:

- **Keyboard substitution:** a character is replaced with an adjacent-key character;
- **Character doubling:** a character is repeated;
- **Character deletion:** a single character is omitted;
- **Transposition:** two adjacent characters are swapped.

Table 7 Properties of the standard test set

Property / PoS tag	**Frequency (n)**	**Percentage (%)**
Total tokens	1188	–
Morphological phenomena	Nominal and verbal inflections (roots)	
ADJ	85	7.2%
ADP	137	11.5%
ADV	85	7.2%
CONJ	92	7.7%
INTJ	2	0.2%
NOUN	374	31.5%
NUM	18	1.5%
PART	12	1.0%
PRON	170	14.3%
VERB	213	17.9%
Total	**1188**	**100.0%**

The error type was chosen at random for each token, and only tokens of four or more characters were included, to avoid producing trivially unrecognizable forms. In the second stage, all automatically generated misspellings were reviewed manually to verify that each erroneous form (a) did not coincidentally match a valid Zazaki word and (b) represented a plausible typing error rather than a random string. This keeps the evaluation close to realistic spelling errors. The final spell checking test set consists of 370 misspelled forms.

The size of the test set is consistent with comparable Hunspell evaluations for low-resource languages; Ahmadi (2021) used a set of similar size for Sorani Kurdish. The same *token* pool was used for both the morphological analysis and the spell checking evaluations, with the misspelled variants generated artificially from the correctly annotated forms, so that every misspelled test item has a verified correct counterpart. What this design means for the interpretation of the results is taken up in the Limitations section.

5 Results and Discussion

5.1 Morphological Analysis and Spell Checking

The system was evaluated along four dimensions: lexical coverage, lemmatization, part-of-speech (POS) tagging, and spell checking. Table 8 presents the full results; Figure 7 visualizes the core metrics with confidence intervals.

Three points in Table 8 require context. In lemmatization, the gap between precision (90.66%) and recall (84.18%) reflects out-of-vocabulary (OOV) items: when the analyzer returns no candidate, the *token* counts as a recall failure but does not hurt precision. POS tagging trails lemmatization by about five points (F_1: 81.89% versus 87.30%), consistent with the higher ambiguity of functional categories in Zazaki (Section 5.2). The joint lemma+POS F_1 of 76.39% is a lower bound on practical usefulness, since many downstream applications (e.g. index search, spell checking) require

Table 8 Comprehensive evaluation of Zazaki Hunspell: morphological analysis, spell checking, and candidate ranking

Task / Dimension	Metric	Value
Lexical coverage[a]	Coverage	92.85%
	Out-of-vocabulary (OOV) rate	7.15%
Lemmatization	Accuracy	90.66%
	Precision	90.66%
	Recall	84.18%
	F_1 score	87.30%
POS tagging	Accuracy	85.04%
	Precision	85.04%
	Recall	78.96%
	F_1 score	81.89%
Joint lemma + POS analysis	Accuracy	79.33%
	Precision	79.33%
	Recall	73.65%
	F_1 score	76.39%
Spelling error detection[b]	Precision / Recall / F_1	90.81%
Spelling correction[c]	Correct suggestion offered	99.40%
Candidate ranking quality[d]	Top-1 accuracy	68.15%
	Top-3 accuracy	86.90%
	Top-5 accuracy	93.75%
	Mean Reciprocal Rank (MRR)	0.794
	Avg. suggestion position (ASP)	1.9

Note. All percentage-based metrics are macro-averaged. [a]N_{total} = 1,188, N_{analyzed} = 1,103. [b]$N_{\text{misspelled}}$ = 370, correctly flagged: 336. [c]Correct suggestions: 334 / 336; the system returned no candidates for 2 flagged items. [d]Top-k counts over n = 336 flagged *tokens*: 229 (k=1), 292 ($k\leq$3), 315 ($k\leq$5).

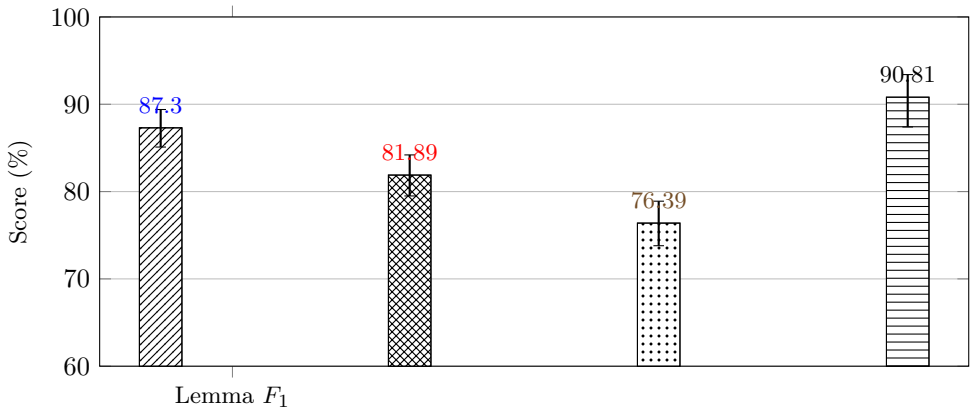

Fig. 7 Core performance metrics of Zazaki Hunspell with 95% Wilson confidence intervals (N_{morph} = 1,103; N_{spell} = 370). Error bars are computed with the Wilson score method.

only the lemma identity. Spelling error detection and suggestion accuracy (90.81% and 99.40%) belong to a separate subtask and are reported separately.

To assess the reliability of the reported figures, 95% Wilson score confidence intervals were computed for each core metric using the sample sizes of each evaluation subtask: the set of analyzed *tokens* for morphological analysis (N = 1,103) and the artificially generated misspelled test items for spelling error detection (N = 370). The resulting intervals are reported in Table 9; the top-k suggestion counts underlying Table 8 are given in note d.

Table 9 Core evaluation metrics with 95% confidence intervals (Wilson score method)

Metric	**Point estimate (%)**	**95% CI (%)**		***N***
		Lower	**Upper**	
Lemma F_1	87.30	85.1	89.4	1,103
POS F_1	81.89	79.5	84.2	1,103
Joint lemma + POS F_1	76.39	73.8	78.9	1,103
Spelling error detection F_1	90.81	87.4	93.4	370
Top-1 suggestion accuracy	68.15	63.0	72.9	336

Note. Confidence intervals are computed with the Wilson score method (Wilson, 1927). Point estimates correspond to the values reported in Table 8. N indicates the sample size for each metric.

5.2 Error Analysis

The system handled common inflectional affixes (case, possession, tense, person) successfully in both analysis and spell checking. Most of the difficulties revealed by the error analysis stem from mismatches between the linguistic properties of Zazaki and the rule-based architecture of Hunspell, which operates over a root lexicon and a finite set of affix rules. The underlying question is to what extent a finite rule set can adequately represent Zazaki's capacity to generate a potentially unbounded number of word forms.

The following phenomena proved particularly problematic:

- words with long, complex affix sequences arising from the agglutinative side of Zazaki;
- verbs or nouns with irregular stem alternations;
- derived forms that are rare or absent in the lexicon;
- determining the internal structure and boundaries of compounds;
- insufficiently modeled complex morphophonemic alternations (e.g. vowel harmony, vowel/consonant deletion, and assimilation);
- ambiguity cases in which a single word form admits more than one valid morphological analysis.

Ahmadi (2021, p. 10) likewise identifies "undefined morphological rules" and "ambiguity" as recurring problems for Sorani.

The test logs show two main error tendencies: false positives and false negatives. The erroneous flagging of verb inflections such as *biwanî* and *biceribnî*, proper nouns such as *Îdrîs* and *zilkîf*, and adjectives such as *qaymeqamî* (false positives) points to gaps in lexicon coverage and affix definitions. The acceptance of misspelled forms such as *pirrdê* (correctly *pirdê*) and *alunê* (false negatives) shows that the spell checker struggles with real-word errors, where an erroneous form collides with another valid entry.

Frequent tagging problems include adjective–adverb (ADJ–ADV) confusion in words such as *nê* and *xeylê*; the consistent tagging of *yew* as a numeral (NUM) even where it functions as an indefinite adjective; and the confusion of pronouns (PRON) with nouns or adjectives. The misanalysis of verb inflections such as *dano* and *şî* as nominal roots shows that resolving morphological ambiguity would require syntactic context. Table 10 categorizes the error types with examples and root causes.

Table 10 Zazaki Hunspell error analysis: false positive and false negative categories, examples, and root causes.

Type	Category	***Flagged form***	***Correct form***	**Root cause**
FP	Verb inflection	*biwanî*	*biwanî*	Missing subjunctive {bi-} definition
	Verb inflection	*biceribnî*	*biceribnî*	Missing rare inflection rule
	Proper noun	*Îdrîs*	*Îdrîs*	Missing Z-flagged entry
	Adjective	*qaymeqamî*	*qaymeqamî*	Insufficient lexicon coverage
FN	Misspelling	*pirrdê*	*pirdê*	Real-word error
	Misspelling	*alunê*	—	Collision with another valid entry
Amb.	ADJ / ADV	*nê*	—	Multiple valid analyses
	NUM / ADJ	*yew*	—	Contextual category ambiguity
	PRON / NOUN	*kê*	—	Syntactic context required

Note. False positive (FP): the system erroneously flags valid words. False negative (FN): an orthographically incorrect form is accepted as valid. Ambiguity (Amb.): more than one valid morphological analysis exists for a single surface form.

5.3 Comparison with Other Hunspell-Based Systems

Comparing Hunspell-based systems across Zazaki, Azerbaijani, and Sorani Kurdish is informative but calls for caution, because the three studies differ substantially in

evaluation protocol, test set composition, and the morphological profiles of the target languages. The figures below are contextual reference points, not evidence of relative superiority. The Zazaki figures in Table 11 were computed on running text with sentence sampling. For Sorani, Ahmadi (2021) reported 27.16% coverage over a list of unique word types rather than running *tokens*; this different denominator inflates the apparent coverage gap, because running-text samples overrepresent high-frequency, morphologically regular items. Ahmadi (2021) also reported 80.14% accuracy in morphological segmentation and 86.02% accuracy in POS tagging, with a verb stemming accuracy of 63.75%. For Azerbaijani, Gasim (2025) reported no specific morphological analysis metric, noting instead that the system struggled with "highly agglutinative forms".

In spell checking, the Zazaki implementation reached 90.81% error detection accuracy. The Sorani study reported the rate at which the correct suggestion appears in the suggestion list (Sugg>3) as 47.66% on test set T1 and 31.52% on test set T2 (Ahmadi, 2021). The Azerbaijani study reported overall correction performance, including, for example, a precision of 83.2% (Gasim, 2025). Table 11 summarizes the comparison with protocol caveats.

Table 11 Comparative evaluation of Hunspell-based morphological analyzer systems. Direct comparison should be avoided owing to methodological differences in test protocols.

System	**Language**	**Coverage / F_1**	**Spelling F_1**	**Test N**	**Protocol note**
This work	Zazaki	92.85% / 87.30%	90.81%	1,188 *tokens*	Running text; sentence sampling
Ahmadi (2021)	Sorani Kurdish	27.16%[†] / 46–98%[*]	46.00%	Type list[†]	Unique type list; different denominator
Gasim (2025)	Azerbaijani	Not reported	84.20%	Not specified	Dictionary-based; problems with highly agglutinative forms

[†]Ahmadi's coverage metric is computed over a list of unique word types and is therefore not directly comparable with running-text sampling. [*]Ahmadi reports F_1 scores ranging from 46% to 98% and a POS tagging accuracy of 86.02%.

6 Conclusion

This study has described the morphology of Zazaki and reviewed the current state of its language resources. Three concrete outputs were produced: the KKV corpus, covering a range of text genres, many authors, and several regional varieties; Zazaki Hunspell; and an evaluation framework for Hunspell-based systems in Zazaki, reported with confidence intervals. Together these provide a baseline for future natural language processing tools for Zazaki.

Formalizing Zazaki inflectional morphology and its morphophonological processes within Hunspell rules also yields a computable description of the language's structure that can serve standardization work and language teaching. All resources are publicly released under an open-source license (repository links are given in the Declarations), and the Zazaki Hunspell files have been integrated into the Kurdish Hunspell collection maintained by Ahmadi (2021).[1] We propose integrating the system into Hunspell-based text editors and word processors, enriching the lexicon with dialect-specific and domain-specific terms, and building on these resources a Zazaki POS tagger as a shared foundation for future research on Zazaki syntax (Hunspell Project, 2025).

Limitations

One limitation concerns the scale and composition of the evaluation. The test set (Section 4.3) is small and drawn from a single corpus slice, so low-frequency phenomena such as rare derivational affixes, dialect-specific allomorphs, and complex affix sequences are likely underrepresented; moreover, because the morphological and spell checking evaluations rest on the same *token* pool, the two tasks are not fully independent. The reported figures should therefore be read as indicators of performance on high-frequency, well-represented Zazaki vocabulary rather than as a coverage estimate over the language's entire morphological space. In addition, only inflectional morphology was addressed; derivational paradigms were left out of scope. The lexicon and rule set reflect the written norm of the Vate Working Group, which is closest to the Central and Northern varieties; this is a defensible anchor, since the GXV standard is currently the only maintained, internally consistent written norm for Zazaki, but performance on lexical items or morphophonemic patterns specific to the Southern or Eastern varieties may fall below the figures reported here (Kırkan, 2021; Paul, 2009). We plan to address these points with larger, multi-source, dialectally stratified test sets and by adding dialect-specific lexicon entries and allomorph rules to the `.dic` and `.aff` files.

Another limitation stems from the Hunspell architecture itself. The system lacks mechanisms to resolve ambiguity when it finds more than one valid analysis for a single word form; it has no effective algorithm for segmenting compounds, and its general disregard for word-internal spaces can cause problems in the orthographic representation of Zazaki compounds (Mankier, 2017). Complex morphophonemic alternations and irregular inflections can only be approximated through allomorph definitions and conditional rules. We expect these constraints to be overcome by moving to the finite-state transducer architecture discussed in Section 2.

[1] https://github.com/sinaahmadi/KurdishHunspell/tree/main/zza

Acknowledgements. The authors thank Dr. Pınar Yıldız and Dr. Haci İbrahim Aytekin, who annotated and cross-checked the 100-sentence test set; Assoc. Prof. Hacı Önen, Prof. Vehbi Türel, and Assoc. Prof. Ahmet Kırkan, who read the dissertation in its formatted form and provided valuable feedback; and the Vate Working Group (GXV), whose long-standing work on the standardization of Zazaki provided the linguistic foundation on which this research rests.

Declarations

Funding The authors declare that no funding was received for this research.

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials The corpus, lexicon, rule set, and evaluation framework are openly available under an open-source license at https://github.com/Wyldhan/Hunspell-Tool, https://github.com/Wyldhan/Zazaki_Corpus, and https://github.com/Wyldhan/Zazaki_Tag_set. The Vate Working Group and the authors of Kovara Vate granted permission for the use of the journal texts in this research.

Code availability See Availability of data and materials above.

Authors' contributions V.Y. designed the study, compiled the text corpus, contributed to the development of the Hunspell system, designed and conducted the experiments, and drafted the article. S.A. compiled the initial dictionary data (particularly for verbs), created the prototype for the supplementary file (`.aff`), provided expert guidance on resource-constrained NLP methodology, and reviewed the article from a critical perspective. Both authors have read and approved the final version of the article.

References

Ahmadi, S. (2020, December 13). Building a corpus for the Zaza-Gorani language family. *Proceedings of the 7th VarDial workshop on NLP for similar languages, varieties and dialects* (pp. 70–78). Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL). Retrieved from https://aclanthology.org/2020.vardial-1.7/

Ahmadi, S. (2021). *Hunspell for Sorani Kurdish spell checking and morphological analysis.* Retrieved from https://arxiv.org/abs/2109.06374 (arXiv preprint arXiv:2109.06374)

Anonby, E., Taheri-Ardali, M., et al. (2015–2018). *Atlas of the languages of Iran (ALI).* Ottawa: Carleton University. Retrieved from http://iranatlas.net/ (Ongoing project and published research series)

Anthony, L. (2022). AntConc (version 4.2.0) [computer software] [Computer software manual]. Tokyo, Japan: Waseda University. Retrieved from

https://www.laurenceanthony.net/software/antconc/

Aronoff, M., & Fudeman, K.A. (2012). *What is morphology?* (2nd ed.). Pondicherry: Wiley-Blackwell.

Arslan, İ. (2022). Ezafe (genitive) in Zaza language. *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, *10*(1), 405–416, https://doi.org/10.18506/anemon.946031

Arslan, İ. (2025). *Zaza text corpus.* Dataset/Corpus Project. (Academic Research Resource)

Baxi, J., & Bhatt, B. (2024). Recent advancements in computational morphology: A comprehensive survey. *arXiv preprint arXiv:2406.05424*, , Retrieved from https://arxiv.org/abs/2406.05424 arXiv:2406.05424 [cs.CL]

Bingöl, İ. (2026). Verbal morphology and alignment in Zazaki: A comparative study of the Bingöl and Varto dialects. *Kurdiname*(14), 39–62, https://doi.org/10.55106/kurdiname.1885943 Retrieved from https://doi.org/10.55106/kurdiname.1885943

Booij, G. (2005). *The grammar of words.* New York: Oxford University Press.

Çağlayan, E. (2024). *Zazalar tarih, kültür ve kimlik* (4. Baskı ed.). İstanbul: İstanbul Bilgi Üniversitesi.

Çem, M. (2003). *Türkçe açıklamalı Kirmancca (Zazakî) gramer.* İstanbul: Deng Yayınları.

Comrie, B. (2017). *Dil evrensellikleri ve dilbilim tipolojisi* (2. Baskı ed.; İ. Ulutaş, Trans.). İstanbul: Hece.

Corbett, G.G. (1991). *Gender.* Cambridge: Cambridge University Press.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., Hulden, M. (2016). The SIGMORPHON 2016 shared task — morphological reinflection. *Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology* (pp. 10–22). Berlin, Germany: Association for Computational Linguistics.

Dogan, M. (2024). Problems in Zazakî nomenclature. S.O. Karim & S. Gholami (Eds.), *Gorani in its historical and linguistic context* (pp. 249–276). Berlin/Boston: De Gruyter Mouton. (Open Access)

Dogan, M., Talamo, L., Vaz, H., Verkerk, A. (2025, November). *Universal Dependencies treebank for Zazaki ZSD (version 2.17).* https://github.com/

UniversalDependencies/UD_Zazaki-ZSD. GitHub. (UD Zazaki-ZSD Data Set)

Gasim, I. (2025). Evaluating Hunspell, SymSpell, Norvig, and N-gram spellcheckers for Azerbaijani text. *International Journal of Innovative Science and Research Technology*, *10*(5), 2574–2577,

Gruba Xebate ya Vateyî (2023). *Rastnuştişê Kirmanckî (Zazakî)* (3. Baskı ed.). İstanbul: Weşanxaneyê Vateyî.

Grûba Xebatê ya Vateyî (2023). *Kirmanckî (Zazakî) – Tirkî sözlük / Kırmancca (Zazaca) – Türkçe sözlük* (6th ed.). Istanbul, Turkey: Vate Yayınları.

Hunspell Project (2025). *Hunspell spellchecker.* https://github.com/hunspell/hunspell. (GitHub Repository, Accessed: 2026-05-09)

Hurskainen, A. (2023). Rule-based language technology for African languages. *Rule-based language technology* (pp. 228–239). Tartu: Northern European Association for Language Technology (NEALT). Retrieved from https://dspace.ut.ee/handle/10062/89595

Kaçan, M.E. (2018). *Zazakî de kategorîzekerdişê edatan û metnanê Zazakî ra nimûneyî* (Yüksek Lisans Tezi). Bingöl Üniversitesi, Bingöl, Turkey.

Karimi, Y. (2007). Kurdish ezafe construction: Implications for DP structure. *Lingua*, *117*(12), 2159–2177,

Keskin, M. (2015). Zaza dili. *Bingöl Üniversitesi Yaşayan Diller Enstitüsü Dergisi*, *1*(1), 93–114, Retrieved from https://dergipark.org.tr/tr/pub/buyeder/issue/19131/199341

Kırkan, A. (2021). Zazakî de edatî. *Kurdiname*(4), 206–225, Retrieved from https://dergipark.org.tr/tr/pub/kurdiname/issue/61466/874133

Kırkan, A. (2024). Zazakî de zerfî. *Şarkiyat*, *16*(1), 208–233, https://doi.org/10.26791/sarkiat.1440088 Retrieved from https://doi.org/10.26791/sarkiat.1440088

Kırkan, A. (2025). In terms of nominal morphology comparing Farsi and Zazaki-name. *International Journal of Kurdish Studies*, *11*(1), 38–75, https://doi.org/10.21600/ijoks.1605054 Retrieved from https://doi.org/10.21600/ijoks.1605054

Korkmaz, Z. (2007). *Gramer terimleri sözlügü.* Ankara: Türk Dil Kurumu Yayınları.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production* (No. 11). Helsinki, Finland: University of Helsinki, Department of General Linguistics.

Kurdo, Q. (2016). *Lêgerîna Zaza.* Bonn: Vejîn.

Kırkan, A. (2017). Fonetik û morfolojîya Kirmanckî rê girizgah. R. Geyik & K. Subaşı (Eds.), *Gotarên zimannasiyê.* Fatih, İstanbul: Nûbihar.

Kırkan, A. (2021). *Zazakî de kar.* Wan: Peywend.

Laçin, Ş. (2018). *Zazakî de zerfî û goreyê raweyan şuxulnayîşê zerfan* (Master's thesis). Bingöl Üniversitesi, Bingöl, Turkey.

Malmîsanij, M. (1996). Kirmanckî (Zazakî) de suffîksî. *Çira*, *6*, 48–63,

Malmîsanij, M. (1997). Kirmanckî (Kirdkî) de suffîksî. *Vate*, *1*, 81–93,

Malmîsanij, M. (2021). The Kirmanjki (Zazaki) dialect of Kurdish language and the issues it faces. H. Bozarslan, C. Gunes, & V. Yadirgi (Eds.), *The Cambridge history of the Kurds* (1st ed., pp. 663–684). Cambridge: Cambridge University Press. (Chapter 25)

Malmîsanij, M., & Mosa, A. (2016). Li gor Çêbûna xwe bêjeyên Kirmanckî (Zazakî). *Humanities Journal of University of Zakho*, *4*(1), 1–11,

Malmîsanij, M., & Mosa, A.X. (2017). Berawirdkirina pêşgir, paşgir û navbergiran dinavbera Kirmanckî (Zazakîyê) û Kurmancî yê da. *Humanities Journal of University of Zakho*, *5*(2), 508–526,

Malmîsanij, M. (2015). *Kurmancca İle karşılaştırmalı Kırmancca (Zazakî) dilbilgisi.* İstanbul: Vate.

Mankier (2017). *Hunspell.* https://www.mankier.com/5/hunspell. (Accessed: 2026-05-09)

Mann, O., & Hadank, K. (1932). *Die Mundarten der Zâzâ, hauptsächlich aus Siverek und Kor.* Leipzig: Verlag der Preussischen Akademie der Wissenschaften. (Nachlass Oskar Mann; herausgegeben von Karl Hadank)

Müller, F. (1865). Beiträge zur kenntniss der neupersischen dialekte. iii. Zaza-dialekt der kurdensprache. *Sitzungsberichte der Kaiserlichen Akademie der*

Wissenschaften, Philosophisch-Historische Classe, *48*, 227–245, (Vorgalent in der Sitzung vom 12. October 1864)

Oflazer, K. (2022). Morphology. R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (2nd ed., pp. 29–48). United Kingdom: Oxford University Press.

Ooms, J. (2024). The Hunspell package: High-performance stemmer, tokenizer, and spell checker for R [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/hunspell/vignettes/intro.html (R package version 3.0.4)

Pakray, P., Gelbukh, A., Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, *31*(2), 183–197,

Paul, L. (1998). *Zazaki: Grammatik und versuch einer dialektologie* (Vol. 18). Wiesbaden, Germany: Ludwig Reichert Verlag.

Paul, L. (2009). Zazaki. G. Windfuhr (Ed.), *The Iranian languages* (pp. 545–586). London: Routledge.

Samvelian, P. (2018). Specific features of Persian syntax: The Ezâfe construction, Differential Object Marking and Complex Predicates. A. Sedighi & P. Shabani-Jadidi (Eds.), *The Oxford handbook of Persian linguistics* (pp. 226–269). Oxford/New York: Oxford University Press. (First published in press 2017)

Selcan, Z. (1998). *Grammatik der Zaza-sprache: Nord-dialekt (Dersim-dialekt)*. Berlin: Wissenschaft und Technik Verlag.

Spylls Project (2024). *data.aff: contents of .aff file.* https://spylls.readthedocs.io/en/latest/hunspell/data_aff.html. (Accessed: 2026-05-09)

Todd, T.L. (1985). *A grammar of Dimili (also known as Zaza)* (Ph.D. dissertation). University of Michigan, Ann Arbor, MI.

Todd, T.L. (2008). *A grammar of Dimili also known as Zaza* (2. ed.). Ann Arbor: Iremet Förlag. (Original work published 1985, Electronic publication 2008)

Toosarvandani, M., & Van Urk, C. (2014). The syntax of nominal concord: What ezafe in Zazaki shows us. *Proceedings of the 43rd annual meeting of the North East Linguistic Society (NELS 43)* (Vol. 2, pp. 209–220). Amherst, MA: GLSA.

Uí Dhonnchadha, E., Ní Pháidín, C., van Genabith, J. (2003). Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, *18*(3), 173–193, https://doi.org/10.1007/s10590-004-2475-4 Retrieved from https://doi.org/10.1007/s10590-004-2475-4

UNESCO (2010). *Atlas of the world's languages in danger* (3rd ed.). Paris: UNESCO Publishing. Retrieved from http://www.unesco.org/languages-atlas/ (Edited by Christopher Moseley. Cartography by Alexandre Nicolas. Also available in French and Spanish.)

Vylomova, E., White, J., Salesky, E., Mielke, S.J., Wu, S., Ponti, E.M., ... Cotterell, R. (2020). SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. *Proceedings of the 17th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology* (pp. 1–39). Online: Association for Computational Linguistics.

Werner, E. (2012). *Morphological sketch of Southern Zazaki: Spoken in the area of Çermik, Çüngüş, Siverek and Gerger*. İstanbul: Tij Yayınları/SIL International. (This morphological sketch was published in Turkish language in the Dictionary of Roşan Hayig and Brigitte Werner (2012).)

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212, https://doi.org/10.1080/01621459.1927.10502953 Retrieved from https://doi.org/10.1080/01621459.1927.10502953

Zilan, B. (2020). Vate Çalışma grubu ve Kırmanccayı (Zazacayı) standartlaştırma Çalışmaları. S. Vali (Ed.), *Yazınsal varoluştan akademik literatüre Kırmancca (Zazaca): M. Malmîsanij'a armağan*. İstanbul: Vate Yayınevi.

Zuhra, F., & Khan, A. (2009). A corpus-based finite state morphological analyzer for Pashto. *Proceedings of the conference on language & technology (CLT09)* (pp. 21–25). Lahore, Pakistan: National Language Authority.

Appendix A Inflectional Inventories and Full Paradigms

Tables A1 and A2 present the nominal and verbal affix inventories; Tables A3 and A4 give the full transitive and intransitive conjugation paradigms on separate rotated pages.

Table A1 Nominal inflection table (number, definiteness, ezafe, case)

	Noun			**Adjective**		
	Masculine	**Feminine**	**Plural**	**Masculine**	**Feminine**	**Degree**
Definite	{-∅}, {-ê}, {-o}, {-î}, {-anê}, {-an}, {-ano}	{-∅}, {-a}, {-ê}, {-î}, {-anê}, {-êno}	{-∅}, {-ê}, {-anê}, {-an}, {-ênê}, {-ên}	{-∅}, {-î}, {-an}	{-e}, {-î}, {-an}	{-êr}, {-tir}
Indefinite	{-êk}, {-êkê}, {-êko}, {-êkî}, {-êr}, {-ên}	{-êke}, {-êka}, {-êr}, {-ên}	*tayê* (lexical)	—	—	—

[a] In the plural, indefiniteness is expressed lexically with *tayê* 'some' rather than by a suffix.

Table A2 Verbal inflection table (tense, gender, number, voice)

Suffixes	**Prefixes**	**Copula**	**Possession**
{-ê}, {-o}, {-a}, {-î}, {-o}, {-a}, {-î}, {-n/-en}, {-e}, {-ê}, {-enê}, {-an}, {-ane}, {-na}, {-îya}, {-nîya}	{bi-}, {nê-}, {me-}	{-a}, {-î}, {-o}, {-ê}	est + {-a}, est + {-î}, est + {-o}, est + {-ê}

Table A3 Comprehensive transitive conjugation of the verb *kerdene* (to do) in the main tenses and aspects, illustrating the split ergative alignment of Zazaki.

Tense	Subject	Object	*Surface form*	Translation
Pres.	1 Sg.	Any	*ez kena*	I do (it)
	2 Sg. (M)	Any	*ti kenî*	You do (it) (m.)
	2 Sg. (F)	Any	*ti kena*	You do (it) (f.)
	3 Sg. (M)	Any	*o keno*	He does (it)
	3 Sg. (F)	Any	*a kena*	She does (it)
	1 Pl.	Any	*ma kenê*	We do (it)
	2 Pl.	Any	*şima kenê*	You (pl.) do (it)
	3 Pl.	Any	*ê kenê*	They do (it)
Past	1 Sg.	M	*mi kerd*	I did (it, m.)
	1 Sg.	F	*mi kerde*	I did (it, f.)
	1 Sg.	Pl.	*mi kerdî*	I did (them)
	2 Sg.	M	*to kerd*	You did (it, m.)
	2 Sg.	F	*to kerde*	You did (it, f.)
	2 Sg.	Pl.	*to kerdî*	You did (them)
	3 Sg. (M)	M	*ey kerd*	He did (it, m.)
	3 Sg. (M)	F	*ey kerde*	He did (it, f.)
	3 Sg. (M)	Pl.	*ey kerdî*	He did (them)
	3 Sg. (F)	M	*aye kerd*	She did (it, m.)
	3 Sg. (F)	F	*aye kerde*	She did (it, f.)
	3 Sg. (F)	Pl.	*aye kerdî*	She did (them)
	1 Pl.	M	*ma kerd*	We did (it, m.)
	1 Pl.	F	*ma kerde*	We did (it, f.)
	1 Pl.	Pl.	*ma kerdî*	We did (them)
Impf.	1 Sg.	Any	*mi kerdêne*	I was doing (it)
	2 Sg.	Any	*to kerdêne*	You were doing (it)
	3 Sg. (M)	Any	*ey kerdêne*	He was doing (it)
	3 Sg. (F)	Any	*aye kerdêne*	She was doing (it)
	1 Pl.	Any	*ma kerdêne*	We were doing (it)
	2 Pl.	Any	*şima kerdêne*	You (pl.) were doing (it)
	3 Pl.	Any	*înan kerdêne*	They were doing (it)
Past	2 Pl.	M	*şima kerd*	You (pl.) did (it, m.)
	2 Pl.	F	*şima kerde*	You (pl.) did (it, f.)
	2 Pl.	Pl.	*şima kerdî*	You (pl.) did (them)
	3 Pl.	M	*înan kerd*	They did (it, m.)
	3 Pl.	F	*înan kerde*	They did (it, f.)
	3 Pl.	Pl.	*înan kerdî*	They did (them)
Perf.	1 Sg.	M	*mi kerdo*	I have done (it, m.)
	1 Sg.	F	*mi kerda*	I have done (it, f.)
	1 Sg.	Pl.	*mi kerdê*	I have done (them)
	2 Sg.	M	*to kerdo*	You have done (it, m.)
	2 Sg.	F	*to kerda*	You have done (it, f.)
	2 Sg.	Pl.	*to kerdê*	You have done (them)
	3 Sg. (M)	M	*ey kerdo*	He has done (it, m.)
	3 Sg. (M)	F	*ey kerda*	He has done (it, f.)
	3 Sg. (M)	Pl.	*ey kerdê*	He has done (them)
	3 Sg. (F)	M	*aye kerdo*	She has done (it, m.)
	3 Sg. (F)	F	*aye kerda*	She has done (it, f.)
	3 Sg. (F)	Pl.	*aye kerdê*	She has done (them)
	1 Pl.	M	*ma kerdo*	We have done (it, m.)
	1 Pl.	F	*ma kerda*	We have done (it, f.)
	1 Pl.	Pl.	*ma kerdê*	We have done (them)
	2 Pl.	M	*şima kerdo*	You (pl.) have done (it, m.)
	2 Pl.	F	*şima kerda*	You (pl.) have done (it, f.)
	2 Pl.	Pl.	*şima kerdê*	You (pl.) have done (them)
	3 Pl.	M	*înan kerdo*	They have done (it, m.)
	3 Pl.	F	*înan kerda*	They have done (it, f.)
	3 Pl.	Pl.	*înan kerdê*	They have done (them)
Fut.	1 Sg.	Any	*ez do bikera*	I will do (it)
	2 Sg.	Any	*ti do bikerê*	You will do (it)
	3 Sg. (M)	Any	*o do bikero*	He will do (it)
	3 Sg. (F)	Any	*a do bikera*	She will do (it)
	1 Pl.	Any	*ma do bikerê*	We will do (it)
	2 Pl.	Any	*şima do bikerê*	You (pl.) will do (it)
	3 Pl.	Any	*ê do bikerê*	They will do (it)

Note. The alignment categories (Nom-Acc / Erg-Abs) are not shown in the columns to save space, but they follow the structural split criterion: present/future constructions show Nom-Acc alignment, while past/perfect constructions follow Erg-Abs alignment based on the gender/number of the object.

Table A4 Comprehensive intransitive conjugation of the verb *kewtene* (to fall) in the main tenses and aspects (nominative–absolutive alignment).

Tense	Subject	Gender	*Surface form*	Translation	Tense	Subject	Gender	*Surface form*	Translation
Pres.	1 Sg.	M/F	*ez kewena*	I fall	**Perf.**	1 Sg.	M/F	*ez kewta*	I have fallen
	2 Sg.	M	*ti kewenî*	You fall (m.)		2 Sg.	M	*ti kewtê*	You have fallen (m.)
	2 Sg.	F	*ti kewena*	You fall (f.)		2 Sg.	F	*ti kewta*	You have fallen (f.)
	3 Sg.	M	*o keweno*	He falls		3 Sg.	M	*o kewto*	He has fallen
	3 Sg.	F	*a kewena*	She falls		3 Sg.	F	*a kewta*	She has fallen
	1 Pl.	M/F	*ma kewenê*	We fall		1 Pl.	M/F	*ma kewtê*	We have fallen
	2 Pl.	M/F	*şima kewenê*	You (pl.) fall		2 Pl.	M/F	*şima kewtê*	You (pl.) have fallen
	3 Pl.	M/F	*ê kewenê*	They fall		3 Pl.	M/F	*ê kewtê*	They have fallen
Past	1 Sg.	M/F	*ez kewta*	I fell	**Fut.**	1 Sg.	M/F	*ez do bikewa*	I will fall
	2 Sg.	M	*ti kewtî*	You fell (m.)		2 Sg.	M/F	*ti do bikewê*	You will fall
	2 Sg.	F	*ti kewta*	You fell (f.)		3 Sg.	M	*o do bikewo*	He will fall
	3 Sg.	M	*o kewt*	He fell		3 Sg.	F	*a do bikewa*	She will fall
	3 Sg.	F	*a kewte*	She fell		1 Pl.	M/F	*ma do bikewê*	We will fall
	1 Pl.	M/F	*ma kewtî*	We fell		2 Pl.	M/F	*şima do bikewê*	You (pl.) will fall
	2 Pl.	M/F	*şima kewtî*	You (pl.) fell		3 Pl.	M/F	*ê do bikewê*	They will fall
	3 Pl.	M/F	*ê kewtî*	They fell					
Impf.	1 Sg.	M/F	*ez kewtêne*	I was falling					
	2 Sg.	M/F	*ti kewtêne*	You were falling					
	3 Sg.	M/F	*o/a kewtêne*	He/she was falling					
	1 Pl.	M/F	*ma kewtêne*	We were falling					
	2 Pl.	M/F	*şima kewtêne*	You (pl.) were falling					
	3 Pl.	M/F	*ê kewtêne*	They were falling					

Note. With intransitive verbs the subject remains in the nominative/direct case in all tenses, and the verb always agrees with the subject in person, number, and (in the second and third person singular) gender. The first person singular simple past and perfect surface forms are identical in many regional varieties. The alignment is consistently Nom-Abs in all columns.