

[https://doi.org/10.53360/3080-3861-2026-3\(7\)-3](https://doi.org/10.53360/3080-3861-2026-3(7)-3)
IRSTI: 16.21.41

S. Zh. Zhanzhigitov

L.N. Gumilyov Eurasian National University,
Z01C0T6, Republic of Kazakhstan, Astana, 11 Kazhymukan Str.
ORCID: 0000-0002-7814-1378
e-mail: syrymphd@gmail.com

ANALYSING WORD FORMS IN KAZAKH LEGAL VOCABULARY USING VECTOR MODELS

Abstract. Distributional similarity between word forms does not establish that a model distinguishes morphological form from grammatical function. This study investigates that distinction in Kazakh legal vocabulary, treating *Word2Vec* and *FastText* as analytical instruments. A corpus of 2,188 documents and 12,327,449 tokens supports a reference set of 200 pairs, eight morphological relations and 4,800 analogies. Retrieval, relation-vector coherence, cross-domain transfer and error diagnostics are combined with a new comparison of identical versus different surface allomorphs. *FastText* achieves higher overall exact retrieval than *Word2Vec* (35.21% versus 28.98%), but genitive and derivational results qualify any general superiority claim. Among 195 target pairs admitting both allomorph conditions, target-matched same-allomorph advantages average 5.61 and 2.32 percentage points, respectively. For accusative *FastText* analogies, an apparent unmatched advantage of 11.93 points becomes -0.68 after target matching, demonstrating lexical-composition confounding. The models preserve similar inter-relation geometry despite different local coherence and error profiles. The contribution is a linguistically constrained account of morphological regularity: shared spelling, lexical retrieval and grammatical interpretation must be evaluated separately. Findings concern one specialized corpus and a rule-based reference resource, not an exhaustive model of Kazakh grammar.

Keywords: legal vocabulary, morphology, vector analysis, word formation, distributional semantics, *Word2Vec*, *FastText*.

Introduction

Kazakh legal vocabulary poses a specific problem for distributional morphology: the same grammatical relation can have different surface exponents, while a formally recognizable suffix can participate in different lexical or grammatical functions. A successful prediction of an extended word form therefore need not demonstrate knowledge of its grammatical status. Conversely, a poor prediction need not establish irregularity in the language. It may reflect the composition of the reference set, competition among related forms, or the specialized contexts represented in a legal corpus. The present study separates these possibilities rather than treating analogy accuracy as a direct measure of morphological competence.

Kazakh and related Kypchak languages have been addressed through explicitly morphological resources, including the finite-state transducers of Washington, Salimzyanov and Tyers [1]. Such resources motivate keeping the identity of a morphological relation distinct from the spelling of its realization. In Kazakh philological scholarship, Samekbayeva's discussion of Shaiakhmetuly's legacy highlights the boundary between word formation and inflection and the multifunctionality of affixes [2]. These distinctions matter when a computational category receives a grammatical name: a list of words ending alike is not automatically a homogeneous class of adjectives, case forms or derivatives.

Register adds a second constraint. Mamyrbek's account of historical Kazakh official-business language describes differentiated administrative genres and recurrent formal expression [3]. It provides a historical context for examining specialized distributions, but does not establish the causes of particular frequencies in a contemporary corpus. Mursalimov's analysis of agreement likewise foregrounds grammatical relations beyond isolated word shape [4]. Here, legal texts define the lexical environment; they do not license conclusions about syntactic agreement, legal interpretation or the complete morphology of general Kazakh.

Word-vector analogies offer a measurable way to compare relations across lexical items [5]. *Word2Vec* assigns representations to vocabulary items, whereas *FastText* shares information through character n-grams [6]. This contrast can reveal the contribution of repeated spelling

patterns. However, Linzen shows that analogy solutions also depend on candidate exclusion and lexical neighbourhoods [7], and Fournier and colleagues distinguish regularity of offsets from successful analogy retrieval [8]. Morpheme-aware modelling of Turkish and English [9] and morphology-specific evaluation of French embeddings [10] further caution against equating one aggregate embedding score with a linguistic explanation.

The theoretical framework has two connected components. *Form–function morphology* distinguishes an inflectional extension, a derivational relation and a contextual lexical use. *Distributional morphology* tests whether pairs assigned to those categories behave similarly in vector space. Their connection is operational: *written exponents* define the allomorph contrast; *target pairs* define lexical controls; *offset coherence* measures consistency within a designated relation; *concordances* test whether a grammatical label is justified by actual usage. Neither component substitutes for the other. A formal string rule is not semantic annotation, and a contextual interpretation does not validate a numerical model by itself.

The unresolved issue is therefore not whether subword information improves an overall score, but whether the apparent advantage reflects a relation that generalizes across surface variants or the particular lexical items entering each comparison. A balanced number of examples per relation does not balance allomorph frequencies. Comparing questions with identical and different suffix strings can consequently confound allomorph identity with target difficulty. Holding the target base and expected form constant offers a more informative, although still observational, diagnostic.

The aim is to explain the relation-specific distributional regularity of Kazakh legal word forms without collapsing orthographic, lexical and grammatical evidence. RQ1 asks how retrieval and offset coherence differ across six case relations, plurality and *-lq* derivation. RQ2 asks whether identical-allomorph advantages persist when the target pair is held constant. RQ3 asks what cross-domain retrieval, concrete predictions and error locality reveal about the limits of grammatical interpretation. The contribution combines existing item-level model evidence with a target-matched allomorph analysis and source-localized interpretation of derivational ambiguity.

Materials and Methods

Materials, analytical units and linguistic categories

The material comprises a frozen Kazakh legal corpus, a morphological reference resource and archived predictions from two Skip-gram models. The corpus contains 2,188 documents, 873,953 training segments and 12,327,449 tokens. Its source registry identifies texts from the Adilet legal information system. The analyses use the stored 200-pair reference set, 4,800-item analogy file, 9,600 prediction rows, 300-dimensional relation vectors, transfer outputs and corrected statistical tables. No models were retrained. Source-document concordances provide supplementary qualitative evidence, rather than an additional frequency sample. Table 1 records the principal controls.

Table 1 – Corpus, model and evaluation controls

Resource / setting	Recorded value
Documents; training segments	2,188; 873,953
Corpus tokens; types	12,327,449; 101,901
Explicit vocabulary; evaluation forms	44,202; 337 (100% covered)
Architecture; dimensions	Skip-gram; 300
Window; min_count; negative samples	5; 5; 10
Epochs; training seed	10; 42
FastText character n-grams	3-6
Reference pairs; analogy questions	200; 4,800
Primary / sensitivity transfer questions	260 / 694

Corpus preparation retained word forms rather than lemmatizing away their endings. The preparation documentation records normalization of Latin *i* to Cyrillic *і*, removal of remaining mixed-script tokens and exclusion of metadata segments. The shared explicit vocabulary contains 44,202 types: 43.3774% of the corpus types but 99.1622% of token occurrences. All 337 distinct evaluation forms occur explicitly in both vocabularies. Consequently, the comparison does not

reward FastText for constructing representations of absent words; it concerns the representation of attested vocabulary items.

The reference pipeline reduced 19,248 candidate pairs to 7,830 high-stability candidates, a 400-pair review pool, 314 eligible pairs and finally 25 pairs per relation. Selection used concatenative form compatibility, morphophonological rules and paradigm support. The reference set was constructed by formal rules; the author reviewed all pairs and linguistic examples, including analyses and translations. No independent multi-annotator validation was performed. Its eight codes cover *PLURAL*, *GENITIVE*, *DATIVE*, *ACCUSATIVE*, *LOCATIVE*, *ABLATIVE*, *INSTRUMENTAL* and *RELATIONAL_ADJECTIVE*. The last is interpreted here as *-lq* derivation, a cover notation for the recorded suffix family, because lexicalized and nominal formations prevent treating every pair as adjective formation.

The unit is an attested base extended-form pair, not necessarily a lemma inflection pair. A base may already be inflected, as in *тұлғалар* → *тұлғалармен* (*tūlğalar* → *tūlğalarmen*, «persons → with persons»). Combining each pair with the other 24 pairs in its relation gives 600 analogies per relation and 4,800 overall. In A:B::C:D, all four forms are distinct. The 2,400 reciprocal dyads exchange the source and target pairs, A:B::C:D and C:D::A:B; they do not reverse suffixation into suffix removal. This distinction prevents a claim of bidirectional morphological testing.

Retrieval, geometry and inferential scope

The archived retrieval procedure uses unit-normalized word vectors and the normalized query $v(B) - v(A) + v(C)$, ranked by cosine against the shared explicit vocabulary with A, B and C excluded. Top-1 measures exact expected-form retrieval; Top-5 and Top-10 retain wider candidate sets; mean reciprocal rank (MRR) averages $1/\text{rank}(D)$. These are lexical retrieval measures, not judgments of grammatical acceptability. The comparison follows the distinction between analogy behaviour and relation structure in [7, 8], without claiming to implement every diagnostic proposed there.

For geometry, each unit offset is the normalized difference between the target and source word vectors. Coherence is the mean cosine of distinct offsets within a relation. Leave-one-out classification assigns each offset to the most similar relation centroid after excluding that offset from its own centroid. The two models are compared through their internal eight-by-eight centroid-cosine matrices, using the 28 upper-triangle values rather than raw coordinates from separately learned spaces. Recalculation from saved unit offsets reproduces the reported coherence and classification counts.

Statistical inference is based on five archived paired comparisons: overall Top-1 and MRR over 200 target-pair units; transfer Top-1 and MRR over 44 target-domain–relation units; and coherence over eight relations. The stored procedure combines Holm-adjusted Wilcoxon tests, corrected sign-flip tests and 100,000-draw paired bootstrap intervals. Sign-flip enumeration is exact for the transfer Top-1 comparison with 15 nonzero differences and for eight coherence differences; larger comparisons use 100,000 Monte Carlo draws. Aggregation reduces question-level pseudoreplication, but shared source pairs and relations can leave residual dependence. Inference is conditional on these resources and single training instances, not on independently sampled models or all Kazakh lexemes.

Allomorph comparison and target-matched control

The allomorph analysis is an additional descriptive analysis of the archived predictions. Every reference extension preserves its base as an exact initial string. The appended substring is therefore extracted as $s = B[\text{len}(A):]$ and $t = D[\text{len}(C):]$. Within the same relation, $s = t$ defines the same-surface-allomorph condition; $s \neq t$ defines the cross-allomorph condition. This operationalization identifies written exponents, not underlying phonological representations or independently annotated morpheme boundaries. It does not test suppletion, stem alternation or all possible Kazakh allomorphs.

First, correct predictions and denominators are tabulated for both conditions. Second, for each fixed target pair C:D, mean cross-allomorph Top-1 accuracy is subtracted from mean same-allomorph accuracy. Each eligible target pair contributes one equally weighted difference. The target-matched estimate is the arithmetic mean of these differences; it holds target spelling, lexical

identity and corpus frequency constant. It does not match source frequency, source contexts or neighbourhood density and is not interpreted causally.

Unequal allomorph group sizes make this control necessary even in a complete pair-combination design. If an allomorph has k reference pairs within a relation of 25 pairs, each target using it has $k - 1$ same-allomorph sources and $25 - k$ cross-allomorph sources. Question-weighted averages consequently give different weights to different targets. In instrumental, the 19, five and one pairs yield $19 \times 18 + 5 \times 4 = 362$ same-allomorph questions. The target-matched analysis removes this particular weighting difference but deliberately leaves the original benchmark and model outputs unchanged.

The overall target-matched contrast weights all 195 eligible targets equally, rather than averaging eight relation estimates equally. Relation-specific contrasts retain their own denominators. Because exclusion depends on allomorph availability, the matched population is slightly narrower than the original benchmark; its estimate should not replace the 4,800-question model comparison.

Of 200 target pairs, 195 admit both conditions. Five are excluded from this contrast because their suffix occurs in only one reference pair: accusative *-ны* (*-ny*), genitive *-дің* (*-diñ*) and *-тің* (*-tiñ*), instrumental *-бен* (*-ben*), and dative *-ке* (*-ke*). With the identical pair unavailable as its own source, no same-allomorph question exists for those targets. The allomorph tabulations were calculated with Python 3.13.5, pandas 2.2.3 and NumPy 2.3.5, then checked by an independent loop over target pairs. No significance tests are claimed for these exploratory contrasts.

Transfer, errors and qualitative interpretation

Transfer uses a nine-domain legal sampling framework. Auditing retained 74 domain-anchor pairs; the primary design selects one anchor per available domain–relation cell and yields 260 questions across seven relations. Instrumental lacks sufficient domain coverage. The primary set is balanced by anchor selection, not by equal relation counts: ablative and locative have 12 questions each, accusative and dative 72 each, genitive 30, plural 42 and derivation 20. A 694-question set supplies sensitivity evidence. Transfer applies a source-pair offset to a base from another legal domain; it does not involve an unseen corpus or newly trained model.

Errors were classified into five string- and suffix-based diagnostic categories defined in the archived analysis. A shared initial string is treated as a prefix-family proxy, not an expert-confirmed morphological family. Close readings examine the expected form, actual highest-ranked competitor, segmentation and plausible function. They are selected to explain aggregate patterns and exceptions, not to estimate their prevalence. The Civil Code’s archived Kazakh text supplies two localized checks on the nominal use of derivational forms: Articles 43(3) and 329(1) [11]. These examples neither recode the complete corpus nor establish a new part-of-speech distribution.

Examples retain Cyrillic spelling, followed by practical transliteration and an analytical English gloss. The transliteration distinguishes *қ/қ*, *ә/ғ*, *һ/һ*, *ө/ө*, *ұ/ұ*, *ү/ү*, *ш/ш*, *ж/ж*, *ы/ы* and *і/і*; *y* is rendered *u* and *ю yu*. It is a consistent reading aid, not a phonetic transcription or a claim about an official Latin alphabet. English glosses are analytical translations prepared for this study. The study analyzes public legal texts and stored model outputs without participant recruitment or experimental intervention.

Generative AI disclosure. ChatGPT (GPT-6 Astra Pro, OpenAI) supported language editing, translation, transliteration, example presentation and analytical drafting in the Introduction, Materials and Methods, Results and Discussion, Conclusion and abstracts, including code preparation for the additional allomorph tabulation. The corpus, model training and archived predictions predate this assistance. The author reviewed the linguistic examples and remains responsible for the manuscript.

Results and Discussion

Relation-specific regularity rather than uniform morphological success

FastText retrieves 1,690 of 4,800 expected forms at rank 1 (35.21%), compared with Word2Vec’s 1,391 (28.98%). Top-5 accuracy is 62.15% versus 49.79%, Top-10 73.67% versus 55.96%, and MRR 0.4733 versus 0.3852. These totals show an advantage for subword

representations, but Table 2 demonstrates that the linguistic relation determines its direction and magnitude. Genitive is strong in both models and favours Word2Vec at Top-1; instrumental shows the largest FastText gain. The *-lq* class is difficult for both and favours different models depending on the metric.

Table 2 – Retrieval by morphological relation: 600 questions per relation

Relation	FT Top-1, %	W2V Top-1, %	Δ , pp	FT MRR	W2V MRR	Median rank FT / W2V
ABLATIVE	9.6667	3.1667	+6.5000	0.2445	0.0875	7 / 105
ACCUSATIVE	63.6667	57.3333	+6.3333	0.7492	0.6866	1 / 1
DATIVE	38.0000	37.3333	+0.6667	0.5387	0.5553	2 / 2
GENITIVE	76.3333	81.5000	-5.1667	0.8306	0.8665	1 / 1
INSTRUMENTAL	37.3333	10.1667	+27.1667	0.5007	0.1995	3 / 27.5
LOCATIVE	14.6667	4.0000	+10.6667	0.2642	0.1139	9 / 75
PLURAL	40.0000	33.1667	+6.8333	0.5714	0.4910	2 / 2
<i>-lq derivation</i>	2.0000	5.1667	-3.1667	0.0869	0.0811	29 / 545.5

The genitive example *тұлға : тұлғаның :: орган : органның* (*tūlğa : tūlğanıñ :: organ : organnyñ*; «person : person.GEN :: body : body.GEN») receives rank 1 from both models. The appended *-ның* (*-nyñ*) is identical, but this example alone cannot explain the whole category. Genitive includes six recorded allomorphs, and Word2Vec remains stronger in overall exact retrieval even though FastText has more coherent genitive offsets. Formal consistency and lexical accessibility are therefore distinct findings, not interchangeable definitions of grammatical regularity.

In *тұлғалар : тұлғалармен :: бантар : бантармен* (*tūlğalar : tūlğalarmen :: baptar : baptarмен*; «persons : with persons :: articles : with articles»), the operation adds instrumental *-мен* (*-men*) to plural bases. FastText retrieves *бантармен* first; Word2Vec ranks it 4,422 and instead predicts *бөліктері* (*bölikteri*, «its/their parts»). The competitor is a plausible legal noun form, yet it does not preserve either the intended lexical base or instrumental extension. The result illustrates how contextual association can compete with the specified morphological relation.

A similar contrast occurs in *қызмет : қызметтен :: бюджет : бюджеттен* (*qyzmet : qyzmetten :: byudjet : byudjetten*; «service : from service :: budget : from the budget»). FastText gives rank 1, whereas Word2Vec gives rank 2,610 and selects *қаражатының* (*qarajatynyñ*, «of its/their funds»). The budget–funds association makes the error lexically interpretable, but the expected ablative is displaced by a genitive-possessive form. This is evidence about the competition observed in these vectors, not proof that legal semantics caused the error.

Plurality demonstrates successful transfer across distinct allomorphs: *тұлға : тұлғалар :: орган : органдар* (*tūlğa : tūlğalar :: organ : organdar*; «person : persons :: body : bodies») maps *-лар* (*-lar*) to *-дар* (*-dar*). FastText retrieves the plural first; Word2Vec ranks it third and predicts *уәкілетті* (*uäkiletti*, «authorized»). Dative *тұлға : тұлғаға :: орган : органға* (*tūlğa : tūlğaға :: organ : organға*; «person : to a person :: body : to a body») is solved by both models. These cases show why a relation-specific account must accommodate both allomorphic generalization and strong lexical competitors.

Allomorph identity and lexical-composition confounding

The reference set balances relations but not written exponents. Instrumental contains 19 *-мен* (*-men*), five *-пен* (*-pen*) and one *-бен* (*-ben*) pairs; dative contains 17 *-ға* (*-ğa*), five *-ге* (*-ge*), two *-қа* (*-qa*) and one *-ке* (*-ke*) pairs. Allomorph inventories are also incomplete: the ablative sample contains no *-тан* (*-tan*). These imbalances are properties of the selected reference resource, not estimates of Kazakh suffix frequencies. They directly affect which source and target lexemes appear in the two comparison conditions.

Across 1,476 same-allomorph questions, FastText solves 632 (42.82%) and Word2Vec 450 (30.49%). Across 3,324 cross-allomorph questions, the corresponding counts are 1,058 (31.83%) and 941 (28.31%). The unmatched differences are therefore 10.99 and 2.18 percentage points. However, those conditions contain different mixtures of target lexemes. Table 3 distinguishes the

raw question-level pattern from the target-matched contrast rather than attributing the complete raw difference to suffix identity.

Table 3 – Surface-allomorph conditions and target-matched Top-1 contrasts

Relation	Questions same / cross	FT % same / cross	W2V % same / cross	Match ed target s	Matched Δ FT, pp	Matched Δ W2V, pp
ABLATIVE	110 / 490	22.73 / 6.73	3.64 / 3.06	25	+13.35	+0.11
ACCUSATIVE	182 / 418	71.98 / 60.05	61.54 / 55.50	24	-0.68	+0.02
DATIVE	294 / 306	41.50 / 34.64	42.52 / 32.35	24	+4.51	+7.01
GENITIVE	154 / 446	77.27 / 76.01	81.82 / 81.39	23	+0.57	+4.91
INSTRUMENTAL	362 / 238	42.54 / 29.41	10.77 / 9.24	24	+7.82	+1.09
LOCATIVE	172 / 428	20.35 / 12.38	3.49 / 4.21	25	+4.14	-0.17
PLURAL	86 / 514	47.67 / 38.72	36.05 / 32.68	25	+12.27	+4.10
-llyq derivation	116 / 484	4.31 / 1.45	6.03 / 4.96	25	+2.30	+1.71

Equal weighting of the 195 eligible target pairs reduces the FastText same-allomorph advantage to 5.61 percentage points; Word2Vec's estimate is 2.32 points. The clearest change concerns accusative. Its raw FastText accuracies are 71.98% for identical suffix strings and 60.05% for different strings, a difference of 11.93 points. When each of the 24 eligible target pairs is compared with itself across source conditions, the average difference becomes -0.68 points. The aggregate contrast thus cannot support an accusative-specific claim that matching the allomorph improves retrieval.

The retained contrasts are heterogeneous. Target-matched FastText differences reach 13.35 points for ablative and 12.27 for plural, but only 0.57 for genitive. Dative favours same-allomorph sources more strongly in Word2Vec (7.01 points) than in FastText (4.51). This does not refute the usefulness of subword information; it identifies where its apparent effect survives a basic lexical control and where it does not. The control extends the warning in [7] from analogy neighbourhoods to the composition of morphological reference pairs. It is also consistent with [10]: an evaluation must operationalize the morphological property it claims to measure.

The comparison still does not isolate phonological knowledge. Identical written suffixes share character sequences with the model's subword units, while source lexemes in the two conditions can differ systematically in frequency and context. The persistent ablative and plural contrasts are consequently associations requiring further controlled explanation. The accusative reversal is a methodological result about aggregation, not a causal claim that different allomorphs are intrinsically easier. This qualification is essential when corpus evidence is used to interpret linguistic categories.

Word formation, ambiguity and the limits of a category label

The derivational analogy *тауар : тауарлық :: шекара : шекаралық* (*tauar : tauarlyq :: şekara : şekaralyq*; «goods : goods-related :: border : border-related») is solved by Word2Vec at rank 1. FastText ranks *шекаралық* eighth and instead selects *шекараның* (*şekaranyñ*, «of the border»). Here, the base is preserved but the relation changes from derivation to genitive inflection. Greater orthographic locality is therefore not equivalent to selecting the appropriate grammatical operation. The example connects the computational error directly to the inflection–word-formation distinction emphasized in [2].

The original *RELATIONAL_ADJECTIVE* class also includes *орта → орталық* (*orta → ortalyq*, «middle → centre»), *өкіл → өкілдік* (*ökil → ökildik*, «representative → representation/representative office») and *кепіл → кепілдік* (*kepil → kepildik*, «pledge/guarantor → guarantee»). These forms can involve nominal or lexicalized meanings; their ending does not make all occurrences relational adjectives. Each target contributes 24 questions, none solved at Top-1 by either model. The examples expose heterogeneity within the existing class. Its original membership and scores are retained, keeping grammatical interpretation separate from retrieval measurement.

A source-localized check strengthens this interpretation. Article 43(3) of the Civil Code states: «Филиалдар мен өкілдіктер заңды тұлғалар болып табылмайды» (*Filialdar men ökindikter zaңdy tұлғalar bolyp tabylmaydy*; «Branches and representative offices are not legal persons») [11]. The plural *өкілдіктер* (*ökindikter*) denotes institutional entities. Article 329(1) begins «Кепілдік бойынша кепілдік беруші» (*Kepildik boyynsha kepildik беруші*; «Under a guarantee, the guarantor») [11]. In that construction, *кепілдік* supplies the nominal designation of a legal arrangement. These attestations establish available nominal uses, not the proportion of such uses in the training data.

For the three derivational targets just discussed, failure depth also varies substantially. Median expected ranks for *орталық* (*ortalyq*) are 19,905 in FastText and 26,039.5 in Word2Vec; for *өкілдік* (*ökindik*), 83.5 and 1,301.5; for *кепілдік* (*kepildik*), 223.5 and 4,741.5. The observations support different levels of retrieval difficulty within the same suffix family. They do not establish that lexicalization alone explains the differences, because contextual distributions and the availability of competing forms have not been independently controlled. Treating the entire class as one grammatical operation would hide this distinction.

The implication is a distinction between a form-based benchmark and a function-based grammatical analysis. The benchmark validly tests retrieval of designated *-llq* extensions, but its label alone cannot support conclusions about adjective formation. Likewise, case-labelled homographic forms may require contextual disambiguation before every token receives that grammatical analysis. This is where philological interpretation contributes something that model ranking cannot provide: it identifies the lexical category and contextual function actually being claimed, rather than accepting an automated label as its own validation.

Local coherence, transfer and failure structure

FastText's mean within-relation offset cosine is 0.125475, versus 0.069838 for Word2Vec, and it is higher in all eight relations. Leave-one-out classification assigns 166/200 offsets correctly (83.0%), versus 139/200 (69.5%). Yet the 28 inter-centroid similarities correlate strongly across models: Pearson $r = 0.901243$ and Spearman $\rho = 0.903667$ (Figure 1). These findings distinguish tighter local form relations from the broader arrangement of the selected classes. They do not show that a vector space reproduces a language-wide grammatical system, nor that the centroid structure is independent of the chosen vocabulary.

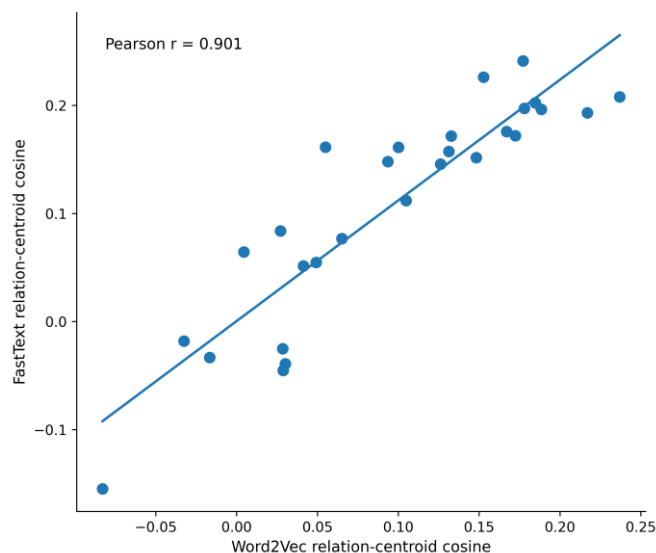


Figure 1 – Correspondence of within-model relation-centroid similarities

Note. Each point represents one of 28 pairs among eight benchmark relations. The display compares internal cosine values, not aligned coordinates.

Primary cross-domain transfer is harder: FastText retrieves 42/260 forms (16.15%) and Word2Vec 20/260 (7.69%), with MRR 0.3041 and 0.1963. In the 694-question sensitivity set, Top-1 is 22.05% versus 10.23%. The direction persists, but neither set supports uniformly

successful transfer. In particular, both models have zero primary locative Top-1 accuracy, and plural Top-1 favours Word2Vec. The shared corpus and uneven domain–relation coverage also mean that this is lexical transfer within a legal register, not generalization to a new register.

In *жауапкершілік : жауапкершіліктен :: мүлік : мүліктен* (*jauapkerşilik : jauapkerşilikten :: mülük : mülükten*; «liability : from liability :: property : from property»), FastText ranks the answer first and Word2Vec third. In *шағым : шағымды :: үкімет : үкіметті* (*şağym : şağymdy :: ükimet : ükimetti*; «appeal : appeal.ACC :: government : government.ACC»), ranks are 1 and 116. The second example transfers an accusative relation from a procedural term to a state-domain term while changing *-ды (-dy)* to *-мі (-ti)*. It shows that cross-allomorph success is possible, not that the models resolve these nouns' functions in complete sentences.

Table 4 presents the five principal statistical comparisons. The overall Top-1 difference remains 0.062292 after target-pair aggregation. Primary transfer has a question-weighted difference of 0.084615, but an equally weighted 44-unit difference of 0.097348. These are different estimands, not inconsistent results. All reported intervals exclude zero; nevertheless, their conditional support should not be expanded into an architecture-wide causal claim. The corrected permutation columns, rather than superseded values elsewhere in the archive, supply the reported *p*-values.

Table 4 – Paired statistical evidence from the archived corrected analysis

Test / metric	Unit; n	FT – W2V	Bootstrap 95% CI	Wilcoxon Holm p	Permutation Holm p	Rank-biserial
T1: Top-1	Target pair; 200	+0.062292	[0.036667; 0.088125]	0.00003621	0.000120	0.4092
T2: MRR	Target pair; 200	+0.088123	[0.066868; 0.109694]	8.899×10^{-10}	0.000050	0.5245
T3: Transfer Top-1	Domain–relation; 44	+0.097348	[0.022159; 0.175379]	0.019124	0.032898	0.6417
T4: Transfer MRR	Domain–relation; 44	+0.114374	[0.055427; 0.177677]	0.000814	0.001560	0.6303
T5: Coherence	Relation; 8	+0.055637	[0.035761; 0.079820]	0.015625	0.015625	1.0000

Rank tails clarify the practical scale of failure. Expected-answer ranks exceed 1,000 in 96/4,800 FastText cases (2.00%) and 688/4,800 Word2Vec cases (14.33%). Median ranks are 3 and 6, but mean ranks are 169.22 and 1,552.88 because of the long tail. Within-relation frequency quartiles show FastText Top-1 advantages of 8.56, 8.94, 5.30 and 1.74 points from the lowest to the highest stratum. These are relative frequency strata, not evidence that all lower-quartile forms are rare. Frequency and morphological category cannot be treated as experimentally manipulated causes.

The paired outcome counts make the model comparison transparent. Both models answer 1,167 general questions correctly and miss 2,886; FastText alone succeeds on 523 and Word2Vec alone on 224. In primary transfer, the corresponding counts are 11, 209, 31 and nine. Thus, the differences do not mean that every FastText success replaces a Word2Vec failure or that the two systems err on disjoint vocabulary. Their large common failure set is especially important for linguistic interpretation: the comparison locates relative strengths within a difficult reference task rather than demonstrating reliable automatic morphological analysis.

The error denominator differs by model because their numbers of successful predictions differ. Among 3,110 FastText failures, 2,091 share the target's prefix-family proxy (67.23%); among 3,409 Word2Vec failures, 1,451 do so (42.56%). The corresponding transfer proportions are 128/218 (58.72%) and 64/240 (26.67%). Table 5 reports the distribution across all five diagnostic categories. A same-prefix prediction can still select a different case or derivative, as *шекараның* (*şekaranyñ*, «of the border») versus *шекаралық* (*şekaralyq*, «border-related») demonstrates. The proxy measures orthographic locality, not grammatical correctness or expert-confirmed stem identity.

Table 5 – Error distributions among incorrect Top-1 predictions, %

Context / model	Errors (n)	Prefix-family proxy	Expected suffix, wrong lexeme	Other relation suffix	Relation suffix, wrong lexeme	Unclassified lexical
General / FastText	3,110	67.2347	3.4727	3.5048	8.7138	17.0740
General / Word2Vec	3,409	42.5638	3.8721	16.1631	6.6882	30.7128
Transfer / FastText	218	58.7156	1.3761	6.4220	9.1743	24.3119
Transfer / Word2Vec	240	26.6667	1.6667	20.4167	7.9167	43.3333

Linguistic contribution and limitations

The findings connect two levels of analysis. Distributional measurements show which attested form relations are accessible to each representation; morphological interpretation determines what those relations mean linguistically. Character n-grams can strengthen local form associations [6], but they are not identical to morphemes [9]. Coherence and retrieval can diverge, as expected from the distinction made in [8]. The present additional contribution is that even an explicitly morphological same-allomorph comparison can be misleading when its target lexemes are not matched.

Recent Kazakh philological studies help distinguish the grammatical and register-specific implications of these findings. Samekbayeva's account of multifunctional affixes [2] informs the reinterpretation of the mixed derivational class. Mamyrbek's register perspective [3] supports treating legal vocabulary as situated material rather than as an unqualified sample of the language. Mursalimov's attention to grammatical relations [4] marks the boundary of the evidence: suffix retrieval does not establish agreement, government or sentence-level function. No historical continuity or syntactic mechanism is inferred directly from the model scores.

Several limits remain. One corpus and one training instance per architecture do not separate architecture effects from training variability. The rule-based reference set is small, concatenative and uneven in allomorph coverage; author review does not constitute independent multi-annotator validation. Target matching controls target identity but not source distributions. Archived vectors permit checking geometry, whereas absent model binaries prevent rerunning vocabulary-wide retrieval from scratch. The original environment versions and all statistical seeds are not fully documented. Shared lexical material limits inferential independence, and no external-register or downstream evaluation was conducted.

Practical use should therefore be diagnostic. For corpus lexicography, ranked competitors can identify forms that need concordance inspection and contextual disambiguation. For morphological resource design, the results motivate recording inflectional layers, derivational meaning and allomorph coverage separately, with expert review of ambiguous pairs. Such applications follow from the observed evidence; improved legal search, teaching effectiveness or automatic grammatical correction have not been tested. A larger context-annotated resource, matched source conditions and repeated model training would be necessary to evaluate those extensions.

Conclusion

The study used distributional retrieval, offset geometry, target-matched allomorph comparisons and close analysis of Kazakh legal forms to distinguish morphological regularity from orthographic similarity. RQ1 is answered by a heterogeneous relation profile: FastText improves aggregate retrieval and local coherence, while genitive and derivational exceptions prevent a universal model ranking. The similar centroid geometry does not erase these differences or establish a complete grammatical representation.

For RQ2, same-allomorph advantages persist on average after target matching, but their interpretation changes materially. Across 195 eligible targets, the FastText and Word2Vec contrasts are 5.61 and 2.32 percentage points. The accusative FastText contrast shifts from an unmatched +11.93 points to -0.68, showing why target composition must be controlled before an apparent allomorphic preference receives a linguistic explanation.

For RQ3, transfer examples and actual competitors show partial preservation of form relations alongside lexical and grammatical mismatches. The *-llq* analysis demonstrates that a suffix-based category can contain nominal or lexicalized uses, while prefix-local errors can still select an unintended relation. The principal contribution is therefore an evidence hierarchy: written exponent, target lexeme, distributional relation and contextual grammatical function must be distinguished. The resulting diagnostics support more precise corpus and morphological-resource analysis, within the limitations of the available legal corpus and reference set.

Data availability

The reference pairs, item-level predictions, relation vectors and statistical tables are deposited in Zenodo (<https://doi.org/10.5281/zenodo.21409221>). The additional allomorph comparisons are derived from these archived materials using the procedure described above. Original legal documents are identified by the Adilet source registry. Trained model binaries are not included in the archived analytical package.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP25793567, «Pedagogical model for the development of legal professionals' linguoconflictological competence (within the framework of the 'Listening State' concept)».

References

1. Washington J.N., Salimzyanov I., Tyers F.M. Finite-state morphological transducers for three Kypchak languages // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). – Reykjavik: ELRA, 2014. – P. 3378-3385. – URL: <https://aclanthology.org/L14-1143/> (accessed: 13.09.2026).
2. Samekbayeva E.M. Q. Shaiahmetulynyn gylymi murasy: zamañauy til bilimindegi sabaqtastyq maselesi [The Scientific Legacy of K. Shayakhmetuly: The Issue of Continuity in Modern Linguistics] // Bulletin of Shakarim University. Philology Series. – 2026. – No. 2(6). – P. 16-26. – DOI: [https://doi.org/10.53360/3080-3861-2026-2\(6\)-2](https://doi.org/10.53360/3080-3861-2026-2(6)-2). (in Kazakh)
3. Mamyrbek G.M. XVIII-XIX gasyrlardagy qazaq resmi-iskeri tilinin qalyptasuy men funktsionaldyq tarmaqtary [The Formation and Functional Varieties of the Kazakh Official-Business Language in the 18th-19th Centuries] // Bulletin of Shakarim University. Philology Series. – 2026. – No. 2(6). – P. 26-36. – DOI: [https://doi.org/10.53360/3080-3861-2026-2\(6\)-3](https://doi.org/10.53360/3080-3861-2026-2(6)-3). (in Kazakh)
4. Mursalimov T.K. Qazaq tilindegi qıysu bailanysynyn grammatikalyq tabigaty [The Grammatical Nature of Agreement in the Kazakh Language] // Bulletin of Shakarim University. Philology Series. – 2026. – No. 1(5). – P. 53-61. – DOI: [https://doi.org/10.53360/3080-3861-2026-1\(5\)-7](https://doi.org/10.53360/3080-3861-2026-1(5)-7). (in Kazakh)
5. Mikolov T., Yih W.-t., Zweig G. Linguistic regularities in continuous space word representations // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Atlanta: Association for Computational Linguistics, 2013. – P. 746-751. – URL: <https://aclanthology.org/N13-1090/> (accessed: 13.09.2026).
6. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135-146. – DOI: https://doi.org/10.1162/tacl_a_00051.
7. Linzen T. Issues in evaluating semantic spaces using word analogies // Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. – Berlin: Association for Computational Linguistics, 2016. – P. 13-18. – DOI: <https://doi.org/10.18653/v1/W16-2503>.
8. Fournier L., Dupoux E., Dunbar E. Analogies minus analogy test: measuring regularities in word embeddings // Proceedings of the 24th Conference on Computational Natural Language Learning. – Online: Association for Computational Linguistics, 2020. – P. 365-375. – DOI: <https://doi.org/10.18653/v1/2020.conll-1.29>.

9. Üstün A., Kurfalı M., Can B. Characters or morphemes: How to represent words? // Proceedings of the Third Workshop on Representation Learning for NLP. – Melbourne: Association for Computational Linguistics, 2018. – P. 144-153. – DOI: <https://doi.org/10.18653/v1/W18-3019>.
10. Mickus T., Copot M. Stranger than paradigms: Word embedding benchmarks don't align with morphology // Proceedings of the Society for Computation in Linguistics 2024. – Irvine, 2024. – P. 173-189. – URL: <https://aclanthology.org/2024.scil-1.17/> (accessed: 13.09.2026).
11. Qazaqstan Respublikasynyn Azamattyq kodeksi (Zhalpy bolim): 1994 zhylygy 27 zheltoqsandagy kodeks [Civil Code of the Republic of Kazakhstan (General Part): Code of 27 December 1994] [Electronic resource]. – Art. 43, para. 3; Art. 329, para. 1. – «Adilet» aqparattyq-quyqytyq zhuiesi [«Adilet» Legal Information System]. – URL: <https://adilet.zan.kz/kaz/docs/K940001000> (accessed: 13.09.2026). (in Kazakh)

С.Ж. Жанжигитов

Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Z01C0T6, Қазақстан Республикасы, Астана қ., Қажымұқан к-сі, 11
ORCID: 0000-0002-7814-1378
e-mail: syrymphd@gmail.com

**ҚАЗАҚ ТІЛІНІҢ ҚҰҚЫҚТЫҚ ЛЕКСИКАСЫНДАҒЫ
СӨЗ ТҮЛҒАЛАРЫН ВЕКТОРЛЫҚ МОДЕЛЬДЕР АРҚЫЛЫ ТАЛДАУ**

Аңдатпа. Сөз тұлғаларының дистрибутивтік ұқсастығы модельдің морфологиялық тұлға мен грамматикалық қызметті ажырататынын өздігінен дәлелдемейді. Мақалада осы мәселе қазақ тілінің құқықтық лексикасы негізінде қарастырылып, *Word2Vec* пен *FastText* тілдік деректі талдау құралы ретінде пайдаланылады. Зерттеу материалы 2 188 құжат пен 12 327 449 токеннен тұратын корпусқа сүйенеді. Сегіз морфологиялық қатынас бойынша іріктелген 200 жұптан 4 800 аналогиялық тапсырма құрастырылған. Күтілетін сөз тұлғасын табу дәлдігі, қатынас векторларының үйлесімділігі, морфологиялық қатынастарды әртүрлі құқық салаларының лексикасына қолдану нәтижелері және қате жауаптар талданып, оларға бірдей және әртүрлі алломорфтарды нысаналы сөз жұбын өзгертпей салыстыру қосылды. *FastText* жалпы дәлдік бойынша *Word2Vec*-тен жоғары нәтиже көрсетті (35,21% және 28,98%). Алайда ілік септігі мен сөзжасамдық қатынастар бұл басымдықтың барлық санатқа бірдей таралмайтынын анықтады. Екі шартта да салыстыруға келетін 195 жұп бойынша бірдей алломорф қатысқан тапсырмалардағы орташа артықшылық 5,61 және 2,32 пайыздық тармақ болды. Табыс септігіне қатысты *FastText* нәтижесінде бастапқы 11,93 тармақтық айырма нысаналы жұптар сәйкестендірілгенде -0,68 тармаққа өзгерді. Демек, жиынтық көрсеткішке салыстырылатын сөздердің құрамы елеулі ықпал етеді. Жекелеген қатынастардың үйлесімділігі әркелкі болғанымен, екі модельде де санаттар арасындағы жалпы геометриялық құрылым ұқсас болды. Зерттеудің ғылыми мәні – жазылу ұқсастығын, сөз тұлғасын табуды және грамматикалық түсіндіруді дербес бағалау. Қорытындылар бір салалық корпус пен ережелер негізінде жасалған іріктеменің шегінде түсіндіріледі.

Тірек сөздер: құқықтық лексика, морфология, векторлық талдау, сөзжасам, дистрибутивтік семантика, *Word2Vec*, *FastText*.

С.Ж. Жанжигитов

Евразийский национальный университет имени Л.Н. Гумилева,
Z01C0T6, Республика Казахстан, г. Астана, ул. Казымукана, 11
ORCID: 0000-0002-7814-1378
e-mail: syrymphd@gmail.com

**АНАЛИЗ СЛОВОФОРМ КАЗАХСКОЙ ПРАВОВОЙ
ЛЕКСИКИ С ПОМОЩЬЮ ВЕКТОРНЫХ МОДЕЛЕЙ**

Аннотация. Дистрибутивное сходство словоформ само по себе не доказывает, что модель различает морфологическую форму и грамматическую функцию. В статье исследуется это разграничение на материале казахской правовой лексики; *Word2Vec* и *FastText* используются как инструменты лингвистического анализа. Корпус включает 2 188 документов и 12 327 449 токенов. На основе 200 пар, представляющих восемь морфологических отношений, сформированы 4 800 аналогий. Оценка точности поиска, согласованности векторов отношений, междоменного переноса и ошибок дополнена сравнением одинаковых и разных

поверхностных алломорфов при фиксированной целевой паре. Общая точность FastText выше, чем Word2Vec: 35,21% против 28,98%, однако результаты родительного падежа и словообразовательной категории ограничивают универсальность этого преимущества. Для 195 пар, допускающих оба условия сравнения, средняя разность в пользу одинакового алломорфа составляет соответственно 5,61 и 2,32 процентного пункта. В винительном падеже исходная разность FastText в 11,93 пункта после сопоставления одинаковых целевых пар меняется на -0,68 пункта, выявляя влияние лексического состава выборки. При различиях локальной согласованности модели сохраняют сходную геометрию отношений между категориями. Научный вклад заключается в раздельной оценке сходства написания, поиска словоформы и её грамматической интерпретации. Выводы ограничены одним специализированным корпусом и эталонной выборкой, построенной по формальным правилам.

Ключевые слова: правовая лексика, морфология, векторный анализ, словообразование, дистрибутивная семантика, Word2Vec, FastText.

Information about the author

Syrym Zhanzhigitov – PhD, Acting Associate Professor, Department of Kazakh Linguistics, L.N. Gumilyov Eurasian National University, Republic of Kazakhstan, Astana, e-mail: syrymphd@gmail.com, ORCID: <https://orcid.org/0000-0002-7814-1378>.

Автор туралы мәлімет

Жанжигитов Сырым Жанмуратович – PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университеті қазақ тіл білімі кафедрасының қауымдастырылған профессор м.а., Қазақстан Республикасы, Астана, e-mail: syrymphd@gmail.com, ORCID: <https://orcid.org/0000-0002-7814-1378>.

Сведения об авторе

Жанжигитов Сырым Жанмуратович – PhD, и.о. ассоциированного профессора кафедры казахского языкознания Евразийского национального университета имени Л.Н. Гумилева, Республика Казахстан, Астана, e-mail: syrymphd@gmail.com, ORCID: <https://orcid.org/0000-0002-7814-1378>.

Received 14.07.2026

Accepted 27.09.2026

[https://doi.org/10.53360/3080-3861-2026-3\(7\)-4](https://doi.org/10.53360/3080-3861-2026-3(7)-4)
FTAMP: 16.21.55

М.Е. Әділов

Абай атындағы Қазақ ұлттық педагогикалық университеті,
050010, Қазақстан Республикасы, Алматы қ., Достық даңғылы, 13
ORCID: 0000-0001-9876-4516
e-mail: marlenadilov88@gmail.com

МАҒЖАН ШЫҒАРМАЛАРЫНДАҒЫ ТАРИХИ ТІЛДІК БІРЛІКТЕР: ҚҰРЫЛЫМЫ МЕН СЕМАНТИКАСЫ

Аңдатпа. Мақалада Мағжан Жұмабаев шығармалары тілінің қазіргі қазақ әдеби тілінен ерекшеленетін фонетикалық, лексика-семантикалық, морфологиялық және синтаксистік белгілері тарихи-лингвистикалық тұрғыдан қарастырылады. Зерттеудің мақсаты – ақын қолданысындағы өзгеше тұлғаларды жүйелеп, олардың қайсысы көне түркілік қабатпен, XX ғасыр басындағы жазба тіл нормасымен, ауызекі және аймақтық қолданыстармен, сондай-ақ өлеңнің ырғақ-буын құрылымымен сабақтас екенін анықтау. Материал ретінде М.Жұмабаев шығармаларының 2013 жылғы бірінші томы, көне және қазіргі түркі тілдеріне қатысты сөздіктер мен тарихи-лингвистикалық еңбектер пайдаланылды. Зерттеуде сипаттамалы, салыстырмалы-тарихи, контекстік және компоненттік талдау әдістері қолданылды. Нәтижесінде ертенді-кеш, бұ, тұмақ, ұжмақ, үзім, шүйінші, тәңірі, нажағай, есіт, ұсау тәрізді тұлғалардың бірқатары тарихи нұсқалармен сабақтас екені көрсетілді; адас, күлік, құртқа, там, ділдә, тамұқ, кешу, саяқ, жай тасы, орда, ноян, көбе сияқты лексемалардың көнерген немесе тарихи семантикасы айқындалды. -ын/-ін, -тұғын, -улі, дағы тәрізді формалар мен кейбір қос сөздердің құрылымы да тарихи және поэтикалық қызметі тұрғысынан сипатталды. Зерттеу ақын тіліндегі өзгешеліктерді кездейсоқ ауытқу емес, әдеби тіл тарихының қабаттарын сақтаған поэтикалық жүйе ретінде бағаланады.