UDC 004.912:81'322

A SEMANTIC LINKING MODEL FOR TURKIC LANGUAGES: A FAIRNESS-CONSTRAINED CROSS-LINGUAL APPROACH FOR LOW-RESOURCE SETTINGS

Akhmedjanova D.A.
daxmadjonova@gmail.com
Digital Technologies and Artificial Intelligence Development Research Institute,
17A, Buz-2, Tashkent, 100125 Uzbekistan

We propose a semantic linking model for the Turkic language family, most members of which are considered low-resource from the computational linguistics viewpoint. The theoretical foundation of the model is Liddy's layered scheme of natural language processing, compressed here into four levels: graphemic-morphological, syntactic, semantic, and cross-lingual. Rather than treating expert-built knowledge resources as hard constraints, we introduce them as a noisy supervision signal weighted by a per-language reliability coefficient π_α; this construction subsumes both classical retrofitting ($\pi_\alpha \to 1$) and purely distributional models ($\pi_\alpha \to 0$) as special cases of a single scheme. The optimisation problem is augmented with a fairness constraint bounding the gap between the best and worst per-language quality, which prevents the model from drifting toward resource-rich languages at the expense of low-resource ones. A two-level algorithm is derived that combines Adam updates for embeddings, a closed-form orthogonal Procrustes step for cross-lingual mappings, and softmax reparametrisation for the language-weight simplex with a projected subgradient step for the dual variable; a standard $O(1/\sqrt{T})$ bound is stated for the outer loop. The evaluation protocol includes five baselines that isolate the contribution of each model component, intrinsic metrics (Spearman correlation on SimRelUz, Precision@1 and Precision@5 for bilingual lexicon induction) and extrinsic metrics (chrF++ for downstream machine translation), with 95% bootstrap confidence intervals and Holm–Bonferroni correction for multiple comparisons.

Keywords: Turkic languages, cross-lingual embeddings, low-resource languages, noisy supervision, fairness constraint, Procrustes problem

Citation: Akhmedjanova D.A. 2026. A semantic linking model for Turkic languages: a fairness-constrained cross-lingual approach for low-resource settings. *Problems of Computational and Applied Mathematics.* 4(74): 171-181.

DOI: 10.71310/pcam.4_74.2026.11

1 Introduction

The Turkic language family, spoken by more than 200 million people and comprising several official state languages with a written tradition spanning over a millennium, is nevertheless classified as *low-resource* for the majority of its members from the standpoint of computational linguistics. Weitzman and Hartmann [18] report that the Turkic languages of Central Asia — Kazakh, Uzbek, Kyrgyz, and Turkmen — share the typical difficulties of low-resource settings, namely limited data availability, scarce linguistic resources, and slow technological development; the same authors note, however, that significant progress has been achieved recently through the release of language-specific datasets and the development of models for applied tasks. Uzbek, in particular, holds the second position among Turkic languages by number of speakers, yet the volume of digital

linguistic resources available for it is several times smaller than that available for major world languages [18].

Natural language processing systems traditionally rely on a sequence of analysis layers. Such a stratified model is common ground in computational linguistics; its most frequently cited formulation is that of Liddy [9], who distinguishes seven levels of linguistic analysis — phonetic, morphological, lexical, syntactic, semantic, discourse, and pragmatic — and emphasises that higher levels depend on the outputs of the lower ones, with the share of strict rules diminishing and the share of statistical regularities growing as one moves upwards. The same stratification is reflected in the structure of the classic textbook on computational linguistics by Jurafsky and Martin [8], which is organised sequentially into parts on "Words" (phonetics, morphology), "Syntax", and "Semantics and Pragmatics"; the same ordering appears in the fundamental exposition of statistical NLP by Manning and Schütze [11].

In the present work, Liddy's seven-level model is compressed into four layers: the phonetic level is omitted because the object of the study is written text; the lexical level is subsumed into the semantic layer, since word-meaning modelling in vector form is precisely the subject of this study. The resulting hierarchy is as follows. The first layer, *graphemic-morphological*, is responsible for tokenisation, word-form normalisation (lemmatisation, stemming), part-of-speech tagging, and identification of morphological categories. The second layer, *syntactic*, produces the structural analysis of the sentence, represented as a tree or graph. The third layer, *semantic*, provides a computational representation of textual meaning, converting form-level information into content-level information. Finally, the fourth layer, *pragmatic-discourse*, covers speaker intent, context-dependent interpretation, and relations that ensure text coherence.

Bojanowski et al. [4] identify the disregard for word morphology — that is, the assignment of a separate vector to each word without taking its internal structure into account — as a critical limitation of well-known word-vector models, especially for morphologically rich languages. Mikolov et al. [12] explicitly acknowledge that their models "have no idea about word morphology" and associate further progress with the incorporation of information about word structure. The same authors partition their evaluation set into semantic and syntactic questions and demonstrate that different architectures behave differently on the two groups. Pennington et al. [14] extend this dichotomy, showing empirically that semantic and syntactic information depend on context in qualitatively different ways: syntactic information is obtained mainly from the immediately adjacent context and is strongly sensitive to word order, whereas semantic information is more non-local and requires a larger context window. This observation is later turned into a formal model component (see Section 4).

Ruder et al. [15] enumerate applications of cross-lingual embeddings following the same layer ordering: from part-of-speech tagging and dependency parsing (morphological-syntactic layer), through named entity recognition and semantic parsing (semantic layer), up to discourse analysis and dialogue state tracking (discourse-pragmatic layer); they note that discourse analysis is typically carried out within Mann and Thompson's rhetorical structure theory [10].

The layered approach is not without criticism. Collobert et al. [6] put forward their "Natural Language Processing (Almost) from Scratch" framework as an alternative to the pipeline of layers: instead of language-specific engineering for each task, a unified neural architecture is applied. In the era of large language models this tendency has become dominant, and the layers now exist not as explicitly separated modules but as

latent representations inside a single model. Nevertheless, the layered model retains its methodological value: it provides a convenient analytical framework for classifying tasks, delimiting the scope of a study, and comparing different approaches. In the present work, it is used precisely in this descriptive and classifying role.

Within this hierarchy, the semantic layer occupies a central and simultaneously most complex position. Its centrality has the following explanation: the outputs of the first two layers, although valuable in themselves, are in practice most often intermediate — the user of a system does not expect a morphological parse but an answer based on *understanding the meaning* of the text. By contrast, the outputs of the semantic layer feed directly into applied tasks such as machine translation, information retrieval, question answering, text classification, and summarisation.

The subject of the present study — semantic communication between Turkic languages — belongs to the *cross-lingual level* of semantic analysis. It should be emphasised, however, that the cross-lingual level requires the solution of tasks at the three preceding levels as a *necessary precondition*: to construct a shared semantic space of two languages, one must first construct high-quality monolingual semantic spaces. This is formalised in the survey by Ruder et al. [15]: the objective of a cross-lingual model consists of *monolingual loss terms* and a *cross-lingual regularisation term* that links them. This observation serves as the starting point for the mathematical formulation of the model in the next section.

2 Problem statement

2.1 Limitations of the knowledge-based approach

Historically, two fundamentally different approaches to semantic analysis have been formed. The *knowledge-based* approach represents semantics through an expert-designed structured resource — a thesaurus, ontology, or lexical database. The classical example is *WordNet*, developed at Princeton University under the leadership of G. A. Miller [13]. The basic unit of WordNet is a *synset* — a set of synonyms representing a lexicalised concept. Synsets are grouped by part of speech (noun, verb, adjective, adverb) and linked to one another by bidirectional pointers that express semantic relations such as synonymy, antonymy, hyponymy, meronymy, troponymy, and entailment [13].

Resources of the WordNet type capture semantic relations with high accuracy, but their nature imposes three limitations. *First*, the creation of such a resource is a manual, labour-intensive process. Church et al. [5] analyse strategies for increasing the coverage of such resources and show that the rate of manual expansion systematically lags behind the growth rate of language corpora. *Second*, coverage is principally bounded: new, domain-specific, and rare lexicon remain outside the base; by the time the expert has updated the dictionary, the language corpus has already changed. This is a systematic gap that arises from the static nature of the resource and the dynamic nature of the language. *Third and most importantly*, such resources are almost non-existent, or are of insufficient quality, for low-resource languages. For Uzbek, the largest available lexical-semantic base — UzWordNet — covers 28 140 synsets, 64 389 senses, and 20 683 words, with an expert-evaluated accuracy of 75.98% [1]. For comparison, the Princeton WordNet contains more than 117 000 synsets [13]. For Kazakh, Kyrgyz, and Turkmen, thesauri of comparable depth are still at an early stage of development.

An important methodological conclusion follows: to adopt a knowledge-based resource as a mandatory input to a cross-lingual model is to bound the entire model by its weakest link. In the case of Uzbek, this means that approximately one in four items of the input signal is erroneous, and that the number of synsets is about four times smaller than in

the Princeton WordNet. For this reason, we treat the hand-crafted resource not as a mandatory input but as a *noisy supervision signal*, and we introduce its reliability into the model as an explicit parameter.

2.2 Limitations of the distributional approach

The alternative — the *distributional* (statistical) approach — represents word meaning through its contextual distribution in the form of a dense vector. This approach requires no manual annotation, but when applied to Turkic languages it faces two serious obstacles. *The first* is morphological richness: models that assign a separate vector to each word disregard its internal structure [12], which constitutes a serious limitation for morphologically rich languages [4]. In an agglutinative language, the number of word forms derivable from a finite set of morphemes is practically unbounded; as a result, each form occurs rarely in the corpus and its vector is estimated unreliably. *The second* is corpus scarcity: the corpus size demands reported by Church et al. [5] — of the order of 10^6 tokens for basic statistics and 10^8 tokens for high-quality vector representations — are not met for the majority of Turkic languages. Thus both the distributional and the knowledge-based approach hit the *same resource barrier*, only from different sides.

2.3 The classical form of cross-lingual models and its assumptions

The objective of cross-lingual models consists of monolingual loss terms and a cross-lingual regularisation term that links them [15]:

$$\mathcal{J}(\Theta) = \sum_{\alpha=1}^{A} \mathcal{L}^{\alpha}_{\text{mono}}(\Theta_\alpha) + \lambda\, \Omega_{\text{cross}}(\Theta_1, \ldots, \Theta_A). \tag{1}$$

Classical models of the form (1) rely on three assumptions, henceforth referred to as A1–A3. A1: for each language α there exists a monolingual corpus of size sufficient for a high-quality estimate of $\mathcal{L}^{\alpha}_{\text{mono}}$. A2: for the construction of Ω_{cross} there exists a large seed dictionary or a parallel corpus. A3: languages are equal partners — resource quality is uniformly distributed across all α. As shown in Subsections 2.1–2.2, for the Turkic language family A1 is partially satisfied, A2 is unsatisfied for the majority of language pairs, and A3 is almost never satisfied.

The third assumption deserves particular attention. In (1) the sum is taken without weights, so the optimisation process naturally *drifts toward resource-rich languages*: the minimum of the overall functional is attained at the expense of the large-corpus language, while the low-resource one is sacrificed in terms of quality. This is precisely the point at which the classical form must be superseded.

2.4 Formal problem statement

Notation. Let $\mathcal{L} = \{l_1, \ldots, l_A\}$ be the set of languages; C_α the monolingual corpus of l_α with $|C_\alpha| = T_\alpha$ tokens; M_α the morpheme (subword) vocabulary of l_α; $\Theta_\alpha = \{\mathbf{z}_m \in \mathbb{R}^d : m \in M_\alpha\}$ the set of morpheme vectors; $\mathcal{K}_\alpha$ the set of intra-synset pairs taken from the knowledge resource available for l_α; $\pi_\alpha \in (0, 1]$ the expert-evaluated accuracy of that resource (for Uzbek $\pi_{\text{uz}} = 0.7598$, and $\pi_\alpha = 0$ for languages with no such resource).

Word vector. Following Bojanowski et al. [4], the vector of a word w is defined as the average of its morpheme vectors, which enables representation of forms unseen in the training corpus:

$$\mathbf{v}_w = \frac{1}{|G_w|} \sum_{m \in G_w} \mathbf{z}_m, \qquad G_w \subseteq M_\alpha, \tag{2}$$

where G_w is the set of morphemes composing w. **Monolingual term.** We take into account the qualitatively different context dependence of syntactic and semantic information noted in the introduction [14]: syntactic information is obtained from the immediate neighbourhood, whereas semantic information is captured by a wider window. Accordingly, the monolingual term is split by two window widths:

$$\mathcal{L}^{\alpha}_{\text{mono}} = \mathcal{L}^{\alpha}_{c_s} + \varkappa\, \mathcal{L}^{\alpha}_{c_m}, \qquad c_s = 2, \quad c_m = 10, \tag{3}$$

where $\varkappa > 0$ is a balancing coefficient (empirically chosen in the range $\varkappa \in [0.3, 3]$ by grid search). Each term is computed under the skip-gram with negative sampling scheme [12]:

$$\mathcal{L}^{\alpha}_{c} = -\sum_{(w,u)\in\mathcal{P}^{\alpha}_{c}} \Big[\log \sigma(\mathbf{v}_w^{\top}\mathbf{v}_u) + \sum_{j=1}^{k} \mathbb{E}_{w^-\sim P^{\alpha}_n} \log \sigma(-\mathbf{v}_w^{\top}\mathbf{v}_{w^-})\Big], \tag{4}$$

where $\mathcal{P}^{\alpha}_{c}$ is the set of (centre-word, context-word) pairs collected from C_α within a window of width c, $\sigma(\cdot)$ is the logistic function, P^{α}_{n} is the noise distribution (typically the unigram distribution raised to the power $3/4$), and k is the number of negative samples.

Knowledge term (noisy supervision). This is the key point of the model. Thesaurus relations enter not as hard constraints but as a reliability-weighted margin penalty:

$$\Omega_{\text{know}} = \sum_{\alpha=1}^{A} \pi_\alpha \sum_{(u,v)\in\mathcal{K}_\alpha} \Big[\, \gamma - \cos(\mathbf{v}_u, \mathbf{v}_v) + \cos(\mathbf{v}_u, \mathbf{v}_{v^-}) \,\Big]_{+}, \tag{5}$$

where $\gamma > 0$ is the expected margin between positive and negative pairs, v^- is a negative sample drawn uniformly at random from the vocabulary of language α that is not a member of the same synset, and $[\cdot]_+ = \max(0, \cdot)$ denotes the ReLU-type hinge. In the limit $\pi_\alpha \to 1$, the term reduces to classical retrofitting; in the limit $\pi_\alpha \to 0$, it vanishes and one recovers the purely distributional model. Both historical approaches thus emerge as special cases of the proposed scheme, which ensures theoretical consistency.

Cross-lingual term. To reduce dependence on parallel corpora, an orthogonal Procrustes mapping is used [2]:

$$\Omega_{\text{cross}} = \sum_{\alpha=2}^{A} \|W_\alpha X_\alpha - X_1\|_F^2, \qquad W_\alpha^{\top} W_\alpha = I, \tag{6}$$

where $X_\alpha \in \mathbb{R}^{d\times n_\alpha}$ is the matrix whose columns are the vectors of the words of l_α that appear in the seed pairs with the pivot language l_1, columns aligned in the order of the pairs; $\|\cdot\|_F$ is the Frobenius norm; and W_α is the mapping subject to the orthogonality constraint. The initial seed pairs are obtained automatically from cognate dictionaries [3], so no hand-crafted seed lexicon is required.

Final problem. Combining all terms, we obtain the following constrained optimisation problem:

$$\Theta^{*} = \arg\min_{\Theta} \Big[\sum_{\alpha=1}^{A} w_\alpha\, \mathcal{L}^{\alpha}_{\text{mono}} + \lambda_1 \Omega_{\text{cross}} + \lambda_2 \Omega_{\text{know}} + \lambda_3 \|\Theta\|_2^2 \Big] \tag{7}$$

subject to

$$\sum_{\alpha=1}^{A} w_\alpha = 1, \qquad w_\alpha > 0, \quad \alpha = 1, \ldots, A; \tag{8}$$

$$\max_{\alpha} Q_{\alpha}(\Theta) - \min_{\alpha} Q_{\alpha}(\Theta) \leqslant \varepsilon. \tag{9}$$

Here $Q_{\alpha}(\Theta) \in [0,1]$ is the semantic quality indicator for language l_{α}, measured by Precision@1 for cross-lingual lexical matching on a validation set (this metric is described in detail in Subsection 3.5); $\varepsilon \in (0,1]$ is the admissible imbalance. Constraint (9) directly compensates for the violation of assumption A3 identified in Subsection 2.3, and constitutes the main departure from the classical form (1).

3 Proposed solution

This section presents a module-based computational scheme for solving problem (7)–(9), together with a two-level iterative algorithm. We first describe the languages, corpora, and knowledge resources used (3.1); then the four modules corresponding to the layers of the hierarchy (3.2); the optimisation algorithm under the fairness constraint (3.3); its convergence analysis and computational complexity (3.4); and finally the experimental design — baselines, metrics, and statistical apparatus (3.5).

3.1 Corpora and knowledge resources

The languages, corpora, and knowledge resources adopted for the experiments are listed in Table 1. The reliability coefficient π_{α} is set to the expert-evaluated accuracy of the corresponding knowledge resource; for languages with no such resource, $\pi_{\alpha} = 0$, so that the term (5) vanishes for their contribution and the model does not penalise a language with no resource — a property consistent with the fairness constraint (9).

Table 1 Languages, corpus sizes, and knowledge resources used in the experiments.

Language	Corpus	Size (tokens)	Knowledge resource	π_{α}
Uzbek	UzWaC + National C.	$\sim 10^8$	UzWordNet	0.76
Turkish	TS Corpus + web	$\sim 10^9$	KeNet	to be filled
Kazakh	web corpus	$\sim 10^8$	—	0
Kyrgyz	web corpus	$\sim 10^7$	—	0
Azerbaijani	web corpus	$\sim 10^8$	—	0

3.2 Model modules

In strict correspondence with the hierarchy of the introduction, the system consists of four modules.

M1 — graphemic-morphological module. Unified transliteration (Latin-Cyrillic normalisation), tokenisation, and BPE segmentation [17]. For Turkic languages, a vocabulary size in the range 16–32 thousand tokens provides a suitable trade-off between morphological coverage and sparsity.

M2 — syntactic module. Context statistics with window $c_s = 2$; POS tags and dependency labels are used for the first term ($\mathcal{L}^{\alpha}_{c_s}$) of (3).

M3 — semantic module. Context statistics with window $c_m = 10$; the monolingual space is built via (2)–(5).

M4 — cross-lingual module. Orthogonal Procrustes alignment via (6); an initial matching is derived from cognate pairs, followed by self-learning iterations [2].

3.3 The optimisation algorithm

Problem (7)–(9) is non-convex, so a two-level scheme is used. The inner level updates the parameters Θ and mappings W_{α} while keeping the language weights w and the dual

variable μ fixed; the outer level adjusts w and μ based on validation qualities. The scheme is given in Algorithm 1.

Algorithm 1 Two-level optimisation under the fairness constraint.

1: **Input:** corpora $\{C_\alpha\}$, knowledge pairs $\{\mathcal{K}_\alpha\}$, reliabilities $\{\pi_\alpha\}$, tolerance ε, iterations T, initial step η_0
2: Initialisation: $\Theta^{(0)}$ random; $w_\alpha^{(0)} \leftarrow 1/A$; $\mu^{(0)} \leftarrow 0$
3: **for** $t = 0, 1, \ldots, T-1$
4: *Inner step:* $\Theta^{(t+1)} \leftarrow$ one Adam step on the functional (7) with $w^{(t)}$ fixed
5: Closed-form orthogonal Procrustes for W_α: $U\Sigma V^\top \leftarrow \mathrm{SVD}(X_1 X_\alpha^\top)$; $W_\alpha \leftarrow UV^\top$
6: *Outer step:* $Q_\alpha \leftarrow$ validation quality; $\bar{Q} \leftarrow \frac{1}{A}\sum_\alpha Q_\alpha$; $g^{(t)} \leftarrow \max\limits_\alpha Q_\alpha - \min\limits_\alpha Q_\alpha - \varepsilon$
7: $w_\alpha^{(t+1)} \leftarrow \mathrm{softmax}_\alpha\big(\log w_\alpha^{(t)} + \mu^{(t)} \cdot (\bar{Q} - Q_\alpha)\big)$
8: $\mu^{(t+1)} \leftarrow \max\big(0,\ \mu^{(t)} + \eta_t\, g^{(t)}\big), \quad \eta_t = \eta_0/\sqrt{t+1}$
9: **stopping:** $|\mathcal{J}^{(t+1)} - \mathcal{J}^{(t)}| < \delta$ and $g^{(t)} \leqslant 0$
10: **end for**
11: **Output:** Θ^*, w^*, μ^*

Parametrisation of the weights. The softmax step on line 7 automatically enforces the constraint (8): $w_\alpha > 0$ and $\sum_\alpha w_\alpha = 1$ hold exactly at every iteration, so no separate projection operator is required. For a language with low quality ($Q_\alpha < \bar{Q}$) the argument is positive, and w_α increases — which directly serves the fairness objective.

Orthogonality. On line 5, the orthogonal Procrustes solution for W_α is found in *closed form* and no iterative convergence is required. By a classical result, the minimum of $\|WX - Y\|_F^2$ subject to $W^\top W = I$ is attained at $W = UV^\top$, where $U\Sigma V^\top$ is the singular value decomposition of $YX^\top$.

3.4 Convergence and computational complexity

The outer problem in μ is a maximum of finitely many linear functions, hence convex and Lipschitz-continuous. Under the step size $\eta_t = \eta_0/\sqrt{t+1}$, the standard subgradient bound applies:

$$\mathcal{J}(\mu^*) - \min_{t \leqslant T} \mathcal{J}(\mu^{(t)}) = O\!\left(\frac{1}{\sqrt{T}}\right). \tag{10}$$

The convergence of Adam under non-stationary conditions is not guaranteed in general; for this reason, a cosine learning-rate schedule is used on line 4 and the empirical convergence curves are reported in the results section.

Computational complexity. A single epoch requires

$$O\!\left(\sum_{\alpha=1}^{A} T_\alpha \cdot c_m \cdot k \cdot d\right)$$

operations, where k is the number of negative samples and d is the vector dimensionality. The Procrustes step is $O(A\,d^3)$, which is negligible for $d \leqslant 300$. The total memory footprint is $O\big(d \sum_\alpha |M_\alpha|\big)$.

Stability. The influence of the hyperparameters $\lambda_1, \lambda_2, \varkappa, \varepsilon$ on the results is quantified via Sobol sensitivity indices; the variation of the main quality metrics under $\pm 10\%$ perturbations is reported.

3.5 Experimental design

Implementation architecture. The system is implemented module by module as summarised in Table 2.

Table 2 Implementation modules and technologies.

Module	Task	Technology
`corpus/`	normalisation, transliteration, BPE	Python 3.11, `sentencepiece`
`embed/`	vectors via (2)–(3)	PyTorch
`know/`	term (5), thesaurus pairs	`networkx`
`align/`	Procrustes (6), self-learning	NumPy/SciPy (SVD)
`optim/`	Algorithm 1, fairness constraint	PyTorch + SciPy
`eval/`	metrics, bootstrap, ablation	`scikit-learn`, `statsmodels`

The code, configuration files, and experiment logs will be released in an open repository and archived on Zenodo with a DOI.

Baselines. To isolate the contribution of each model component, five baselines are considered. (B0) word2vec at the word level — isolates the contribution of the subword representation. (B1) subword model without the cross-lingual term — isolates the contribution of Ω_{cross}. (B2) knowledge term with $\pi_\alpha \equiv 1$ — isolates the contribution of reliability weighting. (B3) variant without the fairness constraint, $\varepsilon = \infty$ — isolates the contribution of (9). (B4) as an external comparison, pretrained multilingual representations from mBERT/XLM-R [7].

Metrics. Semantic similarity: Spearman correlation on the SimRelUz dataset for Uzbek [16]; cross-lingual lexical matching: Precision@1 and Precision@5; worst-case per-language quality: $\min_\alpha Q_\alpha(\Theta)$ — this is the primary indicator of the model's fairness performance; downstream quality: chrF++ in machine translation.

Statistical apparatus. Each experiment is repeated with 5 random seeds; results are reported as mean ± standard deviation; 95% bootstrap confidence intervals are computed over 10 000 resamples; paired permutation tests are used; Holm–Bonferroni correction is applied for multiple comparisons. No numerical value is reported without a confidence interval.

Ablation. Grid search over $\varkappa$ in (3), λ_2 in (5), and ε in (9); each term is removed one at a time; π_α for each language is artificially varied over $[0, 1]$ to inspect the contribution of the knowledge term.

4 Results and discussion

This section is to be completed once the experimental programme is carried out. Full results for the baselines and metrics defined in Subsection 3.5 will be reported. The results will be structured as follows. First, the main quality indicators — Spearman correlation, cross-lingual Precision@1 and Precision@5 for all languages, and the minimum-language quality $\min_\alpha Q_\alpha$ — will be reported with confidence intervals and compared against baselines B0–B4. Second, an ablation study will quantify the contribution of every term and constraint of the model. Third, Sobol sensitivity indices for the hyperparameters and stability under ±10% perturbations will be presented. Fourth, downstream performance in machine translation will be measured by chrF++. Fifth, the practical effect of the fairness constraint — the dependence of the gap $\max_\alpha Q_\alpha - \min_\alpha Q_\alpha$ on the value of ε — will be plotted.

5 Concluding remarks

In this paper, we have proposed a cross-lingual model for semantic linking between Turkic languages. The three main contributions of the model are as follows. First, knowledge resources enter the model as a noisy supervision signal weighted by a reliability coefficient π_α, so that classical retrofitting and the purely distributional model become special cases of a unified scheme. Second, the fairness constraint (9) prevents the drift of the classical cross-lingual model (1) toward resource-rich languages and ensures uniform per-language quality for low-resource Turkic languages. Third, Algorithm 1 uses a closed-form orthogonal Procrustes solution and softmax reparametrisation, keeps the constraints exactly satisfied at every iteration, and provides an $O(1/\sqrt{T})$ convergence bound for the outer loop.

Open questions for future work include: automatic generation of a noisy supervision signal for languages without a knowledge resource (for example, through induction from cognate dictionaries), a formal proof of Adam's convergence under the non-stationary conditions imposed by the outer loop, and dynamic adjustment of the fairness tolerance ε based on validation qualities.

Declarations

Conflicts of interest. The author declares no conflict of interest.

Data availability. The data are available from the author on reasonable request.

References

[1] Agostini A. [et al.] 2021. UZWORDNET: A lexical-semantic database for the Uzbek language. *Proceedings of the 11th Global WordNet Conference (GWC-2021)*. Global WordNet Association. 8–19.

[2] Artetxe M., Labaka G., Agirre E. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Vol. 1. 789–798.

[3] Batsuren K., Bella G., Giunchiglia F. 2022. A large and evolving cognate database. *Language Resources and Evaluation*. 56: 165–189. doi: `https://doi.org/10.1007/s10579-021-09544-6`.

[4] Bojanowski P., Grave E., Joulin A., Mikolov T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 5: 135–146.

[5] Church K., Yuan X., Guo S., Wu Z., Yang Y., Chen Z. 2021. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*. 27(6): 763–778.

[6] Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*. 12: 2493–2537.

[7] Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. 8440–8451.

[8] Jurafsky D., Martin J. H. 2025. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models*. 3rd ed. Online manuscript. Available at: `https://web.stanford.edu/~jurafsky/slp3`.

[9] Liddy E. D. 2001. Natural language processing. *Encyclopedia of Library and Information Science*. 2nd ed. New York: Marcel Dekker.

[10] Mann W. C., Thompson S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text.* 8(3): 243–281.

[11] Manning C. D., Schütze H. 1999. *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press. 680 p.

[12] Mikolov T., Chen K., Corrado G., Dean J. 2013. Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013), Workshop Track.* arXiv:1301.3781.

[13] Miller G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM.* 38(11): 39–41.

[14] Pennington J., Socher R., Manning C. D. 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 1532–1543.

[15] Ruder S., Vulić I., Søgaard A. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research.* 65: 569–631.

[16] Salaev U., Kuriyozov E., Gómez-Rodríguez C. 2022. SimRelUz: Similarity and relatedness scores as a semantic evaluation dataset for the Uzbek language. *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022).* 199–206.

[17] Sennrich R., Haddow B., Birch A. 2016. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016).* Vol. 1. 1715–1725.

[18] Weitzman Y., Hartmann M. 2025. Recent advancements and challenges of Turkic Central Asian language processing. *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025).* Abu Dhabi: ACL. arXiv:2407.05006.

УДК 004.912:81'322

МОДЕЛЬ СЕМАНТИЧЕСКОГО СВЯЗЫВАНИЯ ТЮРКСКИХ ЯЗЫКОВ: ПОДХОД С ОГРАНИЧЕНИЕМ СПРАВЕДЛИВОСТИ ДЛЯ МАЛОРЕСУРСНЫХ УСЛОВИЙ

Ахмеджанова Д.А.

daxmadjonova@gmail.com

Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта,

100125, Узбекистан, г. Ташкент, м-в Буз-2, 17А

В работе предлагается модель кросс-язычного семантического связывания для семейства тюркских языков, большинство представителей которого относится к классу малоресурсных с точки зрения вычислительной лингвистики. Теоретической основой модели служит послойная схема обработки естественного языка Э. Д. Лидди, сжатая до четырёх уровней: графемно-морфологического, синтаксического, семантического и кросс-язычного. Вместо трактовки экспертных знаниевых ресурсов в качестве жёстких ограничений они вводятся в модель как зашумлённый сигнал контроля (noisy supervision), взвешенный по каждому языку коэффициентом надёжности π_α; эта конструкция включает классический retrofitting ($\pi_\alpha \to 1$) и чисто дистрибутивную модель ($\pi_\alpha \to 0$) как частные случаи единой схемы. Задача оптимизации дополняется ограничением справедливости, ограничивающим раз-

рыв между наилучшим и наихудшим по языкам показателем качества и тем самым предотвращающим смещение модели в пользу ресурсно-богатых языков. Разработан двухуровневый алгоритм, объединяющий шаги Adam для обновления вложений, замкнутое ортогональное решение задачи Прокруста для кросс-язычных отображений и softmax-репараметризацию для симплекса языковых весов с проекционным субградиентным шагом по двойственной переменной; для внешнего цикла приводится стандартная оценка сходимости $O(1/\sqrt{T})$. Программа экспериментов включает пять базовых линий, изолирующих вклад каждого компонента модели, внутренние (корреляция Спирмена на SimRelUz, Precision@1 и Precision@5 при индукции двуязычного словаря) и внешние (chrF++ при машинном переводе) метрики с 95%-ными бутстреп-доверительными интервалами и поправкой Холма — Бонферрони на множественные сравнения.

Ключевые слова: тюркские языки, кросс-язычные векторные представления, малоресурсные языки, зашумлённый контроль, ограничение справедливости, задача Прокруста

Цитирование: *Ахмеджанова Д.А.* Модель семантического связывания тюркских языков: подход с ограничением справедливости для малоресурсных условий // Проблемы вычислительной и прикладной математики. – 2026. – № 4(74). – С. 171-181.

DOI: 10.71310/pcam.4_74.2026.11

№ 4(74) 2026

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

AiRI
RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 4(74) 2026

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 71 263-41-98.
Э-почта: journals@airi.uz.
Веб-сайт: https://journals.airi.uz.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 25.08.2026 г.
Формат 60x84 1/8. Заказ № 4. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 4(74) 2026

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
Certificate of Registration No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 71 263-41-98.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIRI printing office.
Signed for print 25.08.2026
Format 60x84 1/8. Order No. 4. Print run of 100 copies.

Содержание

Contents