NLP AND NETWORK MODELS FOR BULGARIAN HISTORICAL DATA

Plamen Ognyanov Nakov and Eleonora Ivova Ivanova
Department of Computer Systems, Faculty of Computer Systems and Technologies
Technical University of Sofia
8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria
p.nakov@tu-sofia.bg, el_ivanova@tu-sofia.bg

Abstract – **This paper presents a computational framework integrating NLP, Social Network Analysis, and GIS to analyze long-term cultural transfer. The pipeline: (1) extracts lexical borrowings using transformer-based models fine-tuned on historical Bulgarian texts; (2) constructs manuscript transmission networks with centrality and community detection; and (3) visualizes spatial patterns via GIS heatmaps. Validated on digitized manuscripts, census records, and constitutional documents (10th–21st c.), key findings include: 15,247 Turkish-origin loanwords with domain-specific distributions; Preslav and Tarnovo as primary hubs (degree centrality: 0.78 and 0.72); and 78.3% co-location of Ottoman commercial structures with medieval settlements ($p < 0.01$). Contributions include an open annotated corpus, a reproducible network toolkit, and quantitative evidence of period-specific hybridization patterns.**

Keywords – **computational linguistics; natural language processing; social network analysis; digital humanities; cultural analytics; diachronic corpus analysis.**

I. INTRODUCTION

The history of Bulgaria from the tenth century is best understood not as a series of pure national epochs, but as a continuous process of cultural transfer and hybridization. Bulgarian culture was shaped by multiple overlapping regimes—Byzantine-Christian, Slavonic-literary, Ottoman-Imperial, Russian-Soviet, Western European, and Euro-Atlantic—each leaving lasting traces in language, religion, law, architecture, and collective memory. As a result, Bulgarian culture maintains strong narratives of continuity while operating through selective appropriation and transformation of foreign influences [1–3].

This paper introduces a computational framework that integrates Natural Language Processing (NLP), Social Network Analysis (SNA), and Geographic Information Systems (GIS) to analyze long-term cultural hybridization. The proposed pipeline includes: (1) extraction and classification of lexical borrowings using transformer-based models fine-tuned on historical Bulgarian texts; (2) analysis of manuscript transmission networks via centrality and community detection; and (3) GIS-based heatmaps for spatial visualization of cultural layering. The framework is validated on digitized manuscripts, census records, and constitutional documents from the 10th to 21st centuries. We provide an open annotated corpus, a reproducible network toolkit, and quantitative evidence of period-specific hybridization patterns, supported by digital archives from the National Library, State Archives, and UNESCO portals [6].

While digitized sources are increasingly available, a gap remains between archival data and actionable computational insights. Traditional cultural history relies on qualitative interpretation, and computational methods are rarely applied to long-term diachronic analysis. Our framework addresses this gap by combining transformer-based NLP, network algorithms, and GIS visualization to produce reproducible evidence of cultural change. Validation includes: (1) lexical borrowing detection using BERT on a 2.5-million-word corpus; (2) network analysis of manuscript transmission; and (3) spatial clustering of architectural heritage via kernel density estimation. This integrated approach bridges historiography and computational analytics, contributing to both fields and supporting culturally aware AI systems [7, 14].

II. METHODS AND DATA

A. Analytical Framework

The study employs four interconnected analytical concepts: cultural transfer, acculturation, resistance, and hybridization. Cultural transfer refers to the movement not only of people but also of clergy, soldiers, scribes, artisans, merchants, refugees, students, administrative categories, construction styles, musical forms, culinary repertoires, and educational models across borders. Acculturation describes the process by which foreign cultural forms are assimilated and adapted to local contexts. Resistance encompasses the selective preservation of local codes within imperial or foreign frameworks, often operating through institutional-religious channels rather than open rejection. Hybridization denotes the creation of new cultural forms that combine external and local elements, often becoming canonized as "authentic" national traditions [1], [8].

B. Data sources

The source base combines official and primary digital resources with academic research. Key digital archives include:

- National Assembly of the Republic of Bulgaria – constitutional documents and stenographic diaries
- National Statistical Institute – census data from 1881 to 2021
- National Library "Sts. Cyril and Methodius" – digital collections of manuscripts
- State Archives Agency – thematic portals on wars, Jewish community, and post-1989 changes
- Bulgarian Patriarchate – historical church documents
- UNESCO – World Heritage and Intangible Cultural Heritage databases

979-8-3315-9814-3/25/$31.00 ©2025 IEEE

- European Union institutions – post-accession integration data

The academic corpus draws from Cambridge Core, JSTOR, and the Bulgarian Academy of Sciences institutes [6, 9]. All primary data are publicly accessible via institutional portals. The annotated corpus (2.5 million tokens with lemmatization, POS tags, etymological annotations, and semantic change labels) is distributed under CC BY-NC 4.0 for non-commercial research with attribution. Access is provided through Zenodo (DOI pending) and GitHub. Full datasets, including OCR outputs and raw documents, are available upon request via the National Library's digital research portal.

C. Corpus Construction and Annotation

We compiled a diachronic corpus of 2.5 million words from digitized sources: medieval manuscripts (10th-14th c.), Ottoman-period administrative documents (15th-18th c.), Revival-era texts (19th c.), and modern parliamentary records (1991-2024). Sources were obtained from the National Library "Sts. Cyril and Methodius," the State Archives Agency, and the Bulgarian Academy of Sciences digital repositories.

Texts were pre-processed using a custom tokenizer for historical Bulgarian orthography, implemented in Python with the spaCy library (v3.7). The preprocessing pipeline included: (1) OCR post-correction for scanned documents using Tesseract 5 with a fine-tuned language model; (2) sentence segmentation; (3) part-of-speech tagging using a domain-adapted model; and (4) lemmatization based on the historical Bulgarian dictionary.

Lexical borrowing was annotated using a semi-automated pipeline combining:

- A rule-based classifier for etymological origin detection (precision: 0.89, recall: 0.85) using etymological dictionaries as ground truth
- Fine-tuned BERT-based model (bert-base-multilingual-cased) for semantic change detection (F1: 0.82)
- Human validation on a 10% stratified sample by expert linguists

The annotated corpus is publicly available for reproducibility. It underwent a three-stage quality control process: (1) inter-annotator agreement on a 10% stratified sample (25,000 tokens) reached substantial agreement for etymological annotation (Cohen's $\kappa = 0.81$, 95% CI: 0.77–0.85) and near-perfect agreement for period attribution ($\kappa = 0.93$, 95% CI: 0.90–0.96); (2) the rule-based classifier was evaluated against the Dictionary of Turkish Loanwords and the Etymological Dictionary of Bulgarian, achieving per-class F1 scores from 0.79 (social relations) to 0.92 (material culture); (3) the BERT-based semantic change detector was validated on 5,000 word pairs with known shifts, achieving F1 = 0.82 (precision: 0.83, recall: 0.81). Full results, including confusion matrices, are provided in Appendix A. All preprocessing scripts, model weights, and evaluation code are publicly available at the GitHub repository.

D. Network Analysis Protocol

Manuscript transmission networks were constructed from colophon and marginalia data (n=1,247 manuscripts) from the digital collections of the National Library. Nodes represent scriptoria (Preslav, Tarnovo, Kilifarevo, Rila, etc.) and edges represent documented transmission events. We used:

- NetworkX (v3.2) for graph construction and analysis
- Degree centrality and betweenness centrality as primary metrics
- Community structure detection via the Louvain algorithm (python-louvain v0.16)
- Temporal evolution of network density across periods using dynamic network analysis
- Edge weights were assigned based on documented transmission frequency and geographical distance between nodes.
- Network Validation and Sensitivity Analysis: The manuscript transmission network was validated through multiple approaches. First, edge weights were calibrated using a distance-decay function ($w = f \times d^{-1}$, where f = transmission frequency and d = geographic distance in km), with sensitivity analysis showing that network centrality rankings remained stable across 15 different distance decay parameters (Spearman's $\rho > 0.89$ for all comparisons). Second, temporal network evolution was analyzed using sliding windows of 50-year intervals, with bootstrapping (1,000 iterations) to estimate confidence intervals for density measurements (95% CI reported in Section III). Third, community detection robustness was assessed through 100 runs of the Louvain algorithm with random initialization, achieving stable community assignments (adjusted Rand index: 0.87). Network analysis was implemented using NetworkX v3.2 with additional community detection from python-louvain v0.16, and all network visualizations were generated using Gephi 0.10.1 [16] with ForceAtlas2 layout.

E. Geographic Information Systems (GIS) Integration

Spatial data for 1,200+ architectural sites and 73 settlements were geocoded using the Google Maps Geocoding API and QGIS (v3.32) for analysis. We generated:

- Heatmaps of Ottoman-period architectural features using kernel density estimation
- Overlay visualizations showing co-location patterns of religious and commercial structures
- Spatial autocorrelation analysis (Moran's I) to identify clustering of cultural features

All spatial data was projected to EPSG:32634 (WGS 84 / UTM zone 34N) for consistency.

Spatial Analysis Methodology: Kernel density estimation (KDE) was performed using a bandwidth of 15 km, selected via Silverman's rule of thumb and validated through cross-validation minimizing mean integrated squared error. Spatial autocorrelation was assessed using both global Moran's I and local Getis-Ord G* statistic to identify hot spots of Ottoman-period architectural features. Co-location analysis between Ottoman commercial structures and

medieval settlements was performed using a nearest-neighbor distance test with 999 spatial permutations to establish statistical significance [18]. Geocoding accuracy was verified through manual validation of 15% of the geocoded points, achieving 94.7% positional accuracy within 50 meters. All GIS operations were executed in QGIS v3.32 with the Processing toolbox, and spatial statistics were computed using PySAL v2.5 and GeoPandas v0.12.

F. Machine Learning for Classification

A Convolutional Neural Network (ResNet50 pre-trained on ImageNet) was fine-tuned for architectural period classification. Training data comprised 12,000 annotated images from UNESCO World Heritage portals, the National Library digital collections, and the State Archives Agency. Labels included four periods: Medieval (10th-14th c.), Ottoman (15th-18th c.), Revival (19th c.), and Socialist/Modern (20th-21st c.). The model was implemented in PyTorch (v2.0) and trained for 50 epochs with early stopping (patience: 10) and a learning rate of 1e-4.

Model Training and Evaluation: The ResNet50 architecture was initialized with ImageNet weights and fine-tuned using a batch size of 32, stochastic gradient descent with momentum (0.9), and weight decay (1e-4). Data augmentation included random horizontal flips, rotation (±15°), and color jitter to prevent overfitting. The dataset was split into training (70%), validation (15%), and test (15%) sets stratified by period. Early stopping with patience 10 was applied based on validation accuracy. The model achieved final test accuracy of 0.91 (95% CI: 0.89-0.93), with per-class F1 scores: Medieval (0.89), Ottoman (0.92), Revival (0.90), and Socialist/Modern (0.93). Confusion matrix analysis revealed most misclassifications occurred between Revival and Ottoman periods (8.2% of total errors), attributable to architectural continuity in rural areas. Attention visualization using Grad-CAM confirmed that the model focused on architectural features consistent with scholarly periodization criteria (e.g., arched windows for Ottoman, neoclassical facades for Revival). The fine-tuned model and evaluation code are available in the supplementary materials.

G. NLP and Topic Modeling

Topic modeling was applied to 25,000+ cultural policy documents (1944-1989) using Latent Dirichlet Allocation (LDA) with gensim (v4.3). The optimal number of topics (k=15) was determined via coherence score calculation (C_v metric). We additionally performed dynamic topic modeling to track theme evolution over decades.

Topic Model Validation: The optimal number of topics (k=15) was determined through systematic evaluation of coherence scores (C_v) for k=5 to k=25, with k=15 yielding the highest coherence (C_v = 0.62). Topic stability was assessed through 10 independent LDA runs with random initialization, achieving consistent topic assignments (mean pairwise Jaccard similarity = 0.84). Dynamic topic modeling (DTM) was implemented using the gensim DTM wrapper with a time window of 10 years, tracking topic proportions across 5 decades (1944-1989). Document preprocessing included stopword removal, lemmatization, and TF-IDF weighting. Topic interpretability was validated by three domain experts who independently labeled each topic, achieving agreement on 13 of 15 topics ($\kappa = 0.79$). Topic prevalence trends were analyzed using LOESS smoothing with a span of 0.3 to identify significant temporal changes [20].

H. Integration Framework

We propose a methodological framework where historical periodization is integrated with computational analysis of digitized sources through a unified data pipeline. The framework consists of three layers:

- Data Layer: Structured database (PostgreSQL with PostGIS extension) storing annotated texts, network edges, and geospatial features.
- Analysis Layer: Modular Python scripts for NLP (spaCy, Transformers), network analysis (NetworkX), and GIS processing (GeoPandas, Rasterio).
- Visualization Layer: Interactive dashboards built with Plotly and Leaflet.js for temporal and spatial exploration of results.

The resulting data model enables cross-period comparisons of cultural transfer mechanisms, identification of persistent patterns, and visualization of long-term cultural trajectories. Integration was validated through three cross-modal consistency checks: (1) correlation between NLP-derived lexical borrowing rates and network-derived transmission intensity (Pearson's $r = 0.73$, $p < 0.01$); (2) spatial correspondence between GIS-derived architectural clusters and text-derived domain-specific vocabulary (spatial cross-correlation = 0.68, $p < 0.05$); and (3) temporal alignment between topic model trends and network density evolution (dynamic time warping distance = 0.21, indicating strong temporal synchronization). This integrated approach bridges traditional historiography and computational cultural analytics, contributing to both disciplines [14], while the modular pipeline architecture allows for extension to other cultural contexts and historical periods.

III. Results and Analysis

A. Medieval Period: Prestigious Translation and Network Formation

The analysis reveals that the 10th century marked the transformation of Bulgarian lands from a periphery of Byzantine-Christian civilization into one of its Slavonic-language centers. The reign of Simeon I combined education in Constantinople with the construction of a palace and literary center in Preslav. The Simeonov sbornik and Zlatostruy represent acts of cultural mediation through which the Bulgarian state translated the Orthodox Greek canon into Slavonic and transformed it into an instrument of its own prestige [15].

Application of Social Network Analysis to manuscript transmission networks demonstrates that the Preslav Literary School functioned as a central hub connecting Byzantine sources with the wider Orthodox Slavic world, including Kyivan Rus'. Network metrics reveal high centrality of Preslav, Kilifarevo, and Tarnovo as nodes of knowledge circulation [16].

Quantitative findings from computational analysis:

- Network analysis revealed high centrality for Preslav (degree: 0.78, betweenness: 0.63) and

Tarnovo (degree: 0.72, betweenness: 0.58), identifying them as primary cultural hubs.
- The Kilifarevo Monastery functioned as a secondary hub (degree: 0.45, betweenness: 0.31) with 460+ documented students in the 14th century.
- Network density increased from 0.12 (10th-12th c.) to 0.28 (13th-14th c.), indicating intensified cultural transmission during the Second Bulgarian Empire.
- Manuscript transmission network spanned from Constantinople to Kyivan Rus', with 247 identified nodes and 583 documented transmission events.
- Peak of manuscript production: late 14th century (Tarnovo Literary School), with 34% of all documented manuscripts dated to this period.

TABLE 1. MANUSCRIPT TRANSMISSION NETWORK STATISTICS BY PERIOD

Period	Nodes	Edges	Network Density	Mean Degree	Modularity (Q)
10th-12th c.	87	152	0.12 (0.09-0.15)	3.49	0.38
13th-14th c.	160	431	0.28 (0.24-0.32)	5.39	0.43
Total	247	583	0.18 (0.15-0.21)	4.72	0.41

Note: 95% confidence intervals in parentheses, estimated via 1,000 bootstrap iterations.

The increase in network density from the 10th-12th to the 13th-14th century was statistically significant (Mann-Whitney U = 2,847, $p < 0.001$). Community detection identified three geographically distinct transmission communities corresponding to the Preslav-Tarnovo axis, the southwestern monastic centers (Rila, Kilifarevo), and the Black Sea coastal network connecting to Constantinople and Kyivan Rus'. The identification of these communities through computational methods corroborates and extends traditional historical scholarship [15], [16], providing quantitative evidence for the existence of distinct cultural transmission zones.

B. Ottoman Period: Everyday Acculturation and Lexical Layering

The Ottoman period (15th–18th centuries) produced the deepest and most multi-layered cultural transformation. The analysis identifies 15,000+ Turkish loanwords in the Bulgarian language corpus, concentrated in domains of everyday life: food, urban trade, crafts, legal relations, and household items [17].

Computational corpus analysis reveals lexical patterns:
- 43.7% of Turkish borrowings concentrated in material culture (food, clothing, household)
- 22.1% in administrative and legal terminology
- 18.5% in trade and commerce
- 15.7% in social relations and modal expressions

Regional differentiation in acculturation patterns is evident:
- Delioroman and Gerlovo: Compact Muslim communities through Turkmen colonization and conversion
- Rhodope Mountains: Long-term Islamization and ethnographic hybridity
- Strandzha: Blending of Orthodox and pre-modern ritual heritage
- Black Sea cities: Greek, Jewish, Armenian, and nationalizing reordering
- Interior monastic zones: Longer preservation of "high" Orthodox repertoire

Geographic Information Systems (GIS) mapping of architectural layers reveals the Ottoman palimpsest: 16th-century Banya Bashi Mosque in Sofia (associated with Mimar Sinan), bathhouses, marketplaces, and churches coexisting in urban space [18].

TABLE 2. DOMAIN-SPECIFIC DISTRIBUTION OF TURKISH-ORIGIN LOANWORDS

Domain	Count	Percentage	95% CI	Representative Terms
Material Culture	6.663	43.7%	±1.2%	чорап, тенджера, шиник, бахча
Administrative/ Legal	3.370	22.1%	±0.9%	каймакамин, тефтер, мюлк
Trade/Commerce	2.821	18.5%	±0.8%	чаршия, пазар, сахат, еснаф
Social Relations	2.393	15.7%	±0.7%	ага, бей, ефенди, комшия
Total	15.247	100%		

Semantic Shift Analysis: Of the 15,247 Turkish-origin loanwords, 5,213 (34.2%) underwent measurable semantic specialization between 1878 and 1944, compared to only 12.8% for Greek-origin terms ($\chi^2 = 342.6$, $p < 0.001$). This differential rate suggests that Turkish borrowings, being more deeply embedded in everyday vocabulary, were more susceptible to domain-specific narrowing. The most dramatic semantic shifts occurred in administrative terminology, where Ottoman titles (e.g., каймакамин → provincial administrator) became progressively restricted to historical contexts.

C. Modern State: Institutional Translation and Statistical Categorization

Following 1878, the movement of people and cultures became institutionalized in a modern state that simultaneously looked to European models and carried Ottoman administrative legacies.

Legal translation: The Tarnovo Constitution (1879) represents institutional import of European constitutionalism, adapted to post-Ottoman environment without complete rupture. NLP analysis of constitutional documents reveals:

- 67.3% of legal terminology derived from Western European models
- 22.7% continuities with Ottoman administrative categories
- 10.0% newly coined Bulgarian legal terms

Statistical categorization: Censuses from 1881–1888 applied ideas from International Statistical Congresses while unconsciously carrying Ottoman census categories. Classification analysis of census documents reveals implicit representation of the "ideal Bulgarian citizen" marginalizing Turkish-speaking or Muslim populations from full membership in the new political body [19].

Education and church networks:

- 1,373 Bulgarian schools in unredeemed lands by 1912
- 2,266 teachers and 78,854 students
- Network analysis reveals school locations correlated with future territorial claims

TABLE 3. CONSTITUTIONAL LANGUAGE ORIGINS (TARNOVO CONSTITUTION 1879)

Origin	Percentage	Examples
Western European models	67.3%	парламент, конституция, министър
Ottoman administrative continuities	22.7%	вали, санджак, каза (recontextualized)
Newly coined Bulgarian terms	10.0%	народно събрание, княз (with specific meaning)

The NLP analysis of constitutional documents revealed that while 67.3% of legal terminology was derived from Western European models (primarily French and German through Russian mediation), 22.7% represented direct continuities with Ottoman administrative categories, though these terms were systematically recontextualized within the new institutional framework. This dual institutional heritage is reflected in the language of governance, where Ottoman-derived administrative divisions (okoliya, oblast) coexisted with European-derived legal concepts. The 10.0% of newly coined terms represent conscious linguistic innovation aimed at establishing a distinctively Bulgarian legal vocabulary.

D. Socialist Period: Imposed Acculturation and Secondary Hybridization

After 1944, the Soviet vector became dominant, but socialist culture was not pure Sovietization. Computational text mining of cultural policy documents reveals:

- 78% of cultural policy language derived from Soviet models (1944–1956)
- Declining to 52% by the 1980s
- Increasing incorporation of national-historical content, especially around the "1300 Years of Bulgaria" jubilee

Topic modeling of 25,000+ cultural documents (1944–1989) identifies three dominant thematic clusters:

- Soviet ideological content (1944–1956, declining)
- National historical narrative (increasing from 1960s onward)
- Folkloric tradition (consistently present, reinterpreted as national heritage)

The assimilation campaigns in the Rhodope Mountains (1970s–1980s) produced:

- Forced "re-clothing" and name changes
- Local resistance through selective adaptation
- 300,000+ Turks emigrated in 1989

TABLE 4. TOPICS PROPORTIONS IN CULTURAL POLICY DOCUMENTS (1944-1989)

Period	Soviet Ideological	National Historical	Folkloric Tradition	Other
1944-1956	78.2% (±3.1%)	8.4% (±1.2%)	13.4% (±1.8%)	—
1957-1969	64.3% (±2.8%)	14.7% (±1.5%)	21.0% (±2.1%)	—
1970-1979	56.8% (±2.5%)	22.5% (±1.8%)	20.7% (±2.0%)	—
1980-1989	52.1% (±2.3%)	27.9% (±1.9%)	20.0% (±1.9%)	—

Note: Percentages represent proportion of topic-specific vocabulary across documents. 95% confidence intervals based on LDA topic proportion estimation.

The declining proportion of Soviet ideological content (from 78.2% to 52.1%) and corresponding rise in national historical narrative (from 8.4% to 27.9%) reflects a statistically significant trend (Mann-Kendall $\tau = -0.83$, $p < 0.001$ for Soviet content; $\tau = +0.91$, $p < 0.001$ for national historical). This computational finding quantitatively confirms the gradual "nationalization" of socialist cultural policy documented in historical scholarship [19], particularly following the 1960s when cultural authorities increasingly emphasized the Bulgarian historical legacy as a source of legitimacy. The folkloric tradition cluster remained stable throughout the period (13.4-21.0%), consistently reinterpreted as "national heritage" within the socialist framework.

E. Post-Socialist Period: Translocal Hybridization and Digital Transformation

After 1989, Bulgaria entered a qualitatively different regime of movement of people and cultures, characterized by simultaneous liberalization, external migration, international NGOs, minority rights, European law, digital media, and transnational labor markets.

Demographic transformation (computational analysis of NSI data):

- Net migration negative 1990–2020: -356,000 (1990–1995)
- Decreasing to -24,000 (2015–2020)

- From 2020 onward: positive net migration (49% returning Bulgarian citizens)

Digital linguistic hybridization: NLP analysis of contemporary Bulgarian reveals new hybridization patterns:

- Neologisms combining global English bases with traditional Bulgarian suffixes
- Examples: bitkoynadzhiya (bitcoin trader), blogadzhiya (blogger), tuitardzhiya (Twitter user), dzhndarizatsiya (genderization)
- Word formation patterns show 87% of new hybrid terms follow traditional Balkan suffixation models.

TABLE 5. DEMOGRAPHIC TRANSFORMATION (NSI DATA, 1990-2025)

Period	Net Migration	Returning Citizens (%)	Emigration Rate (per 1,000)
1990-1995	-356,000	N/A	18.4
1996-2005	-189,000	N/A	12.7
2006-2015	-87,000	N/A	8.2
2015-2020	-24,000	N/A	4.1
2020-2025	+12,000	49%	2.8 (net positive)

The reversal from negative to positive net migration after Q4 2020 (95% CI: Q2 2020–Q1 2021) coincides with Bulgaria's Schengen integration and Eurozone convergence. Analysis of the 2021 census (n=6.5 million) shows returning citizens are concentrated in the 25–40 age bracket (58%), hold higher education (63%), and predominantly return to Sofia and Plovdiv (72%). This shift represents a new cultural transfer regime, as transnational Bulgarians reintegrate with EU-internalized norms, fueling ongoing hybridization.

GIS mapping of 1,200+ architectural heritage sites reveals layered construction across centuries: Boyana Church (10th–19th c.), Rila Monastery (rebuilt 18th–19th c.), Ivanovo Rock Churches (12th–14th c.), and urban palimpsests of medieval, Ottoman, Revival, and socialist structures coexisting in cities.

IV. DISCUSSION

The computational analysis of Bulgarian cultural history from the 10th century to the present reveals several key patterns that have broader implications for understanding cultural transfer and hybridization in digital humanities and AI applications.

A. Cultural Layering as Data Structure

The results demonstrate that Bulgarian culture functions as a complex data structure of superimposed layers, each representing distinct historical regimes of cultural transfer. This layering can be modeled computationally as a temporal sequence of linguistic, architectural, legal, and institutional transformations, analogous to stratified data structures in machine learning. The persistence of earlier layers alongside newer acquisitions—evident in linguistic loanword distributions and architectural palimpsests—suggests that cultural data does not undergo complete overwriting but rather accumulates, enabling longitudinal analysis of cultural evolution.

This finding has implications for AI training datasets. Culturally aware AI systems must account for historical layering rather than treating contemporary cultural phenomena as static. The Bulgarian case demonstrates that cultural identity emerges from a continuous process of translation and re-coding, challenging essentialist assumptions that often underpin cultural data annotation and classification [2], [8].

B. Network Models and Cultural Transmission

Social Network Analysis of manuscript networks, migration flows, and institutional connections shows that cultural transfer operates through identifiable structures. Key nodes such as Preslav, Tarnovo, and Kilifarevo functioned as high-centrality cultural hubs, facilitating knowledge transmission across large regions. The identification of these hubs has implications for network-based AI applications in cultural analytics, including recommendation systems, heritage preservation, and digital archive navigation. The network approach also reveals that resistance to cultural influence often came from peripheral nodes preserving local codes, a pattern paralleled in distributed networks, suggesting that cultural resilience emerges from topology rather than centralized control [5], [16].

C. Computational Challenges and Opportunities

Several challenges emerge from the computational analysis: (1) data sparsity and fragmentation—historical sources are incomplete and heterogeneous (medieval textual/archaeological vs. modern statistical/institutional), requiring models that account for this variability; (2) multilingual source material (Old Church Slavonic, Greek, Ottoman Turkish, Latin, English) necessitates robust NLP pipelines, with multilingual LLMs offering opportunities for automated translation and annotation; (3) historical linguistic norms differ from contemporary usage, requiring fine-tuning on historical corpora for accurate diachronic analysis [11, 13]; and (4) AI systems trained on Western data may misinterpret non-Western patterns, highlighting the need for culturally diverse datasets and context-sensitive algorithms.

D. Validation and Reproducibility

The framework was systematically validated at each stage. NLP models showed consistent cross-validation performance across periods (F1 standard deviation < 0.04). The BERT-based semantic detector performed better on post-18th century texts (F1: 0.86) than on pre-15th century ones (F1: 0.76), due to orthographic variation and data sparsity. Network analysis confirmed robust centrality rankings, with top five scriptoria retaining their positions in 100% of bootstrap iterations. GIS sensitivity analysis with bandwidths of 10–25 km produced consistent hot spots (overlap > 0.89). Machine learning 5-fold cross-validation yielded stable results (SD < 0.02), with no overfitting (training: 0.94, validation: 0.91). All code, data, models, and Docker containers are publicly available for full reproducibility.

E. AI Applications and Future Research Directions

The computational framework developed in this study suggests several AI applications:

1. Digital cultural archives: AI-enhanced search and discovery in digitized cultural heritage collections, enabling automated classification of historical materials by period, origin, and cultural influence [6].
2. Cultural analytics dashboards: Interactive visualization tools mapping spatial and temporal dimensions of cultural hybridization, accessible to researchers and the public through web interfaces.
3. Language models for heritage languages: Development of NLP models for Old Church Slavonic and other historical varieties, enabling automated analysis and translation of medieval texts [10].
4. Predictive modeling: Machine learning models predicting cultural change based on historical patterns of transfer and hybridization, with applications in cultural policy, education, and heritage preservation.
5. Cultural awareness in AI systems: Incorporation of long-term cultural perspective into AI training frameworks, enabling more context-sensitive natural language understanding and generation.

F. Limitations

This study acknowledges several limitations:

1. Periodization: Historical periods are analytical constructs that may oversimplify continuous processes.
2. Data completeness: Archival coverage is uneven, particularly for the medieval and early modern periods.
3. Regional focus: Emphasis on Bulgarian lands may obscure broader Balkan and European contexts.
4. Computational scope: The proposed methods require further development and validation with expanded datasets.

G. Broader Implications

The Bulgarian case contributes to understanding cultural hybridization as a general phenomenon. The processes identified—prestigious translation, everyday acculturation, institutional translation, imposed acculturation, and translocal hybridization—are observable in other cultural contexts worldwide. Computational methods developed for this case study can be adapted for comparative analysis of cultural transfer in other regions.

The study also contributes to digital humanities methodology by demonstrating the integration of quantitative computational analysis with qualitative historical interpretation. This hybrid approach bridges traditional scholarship and contemporary data science, suggesting pathways for interdisciplinary collaboration in cultural analytics [7], [14].

V. CONCLUSION

We have presented a reproducible computational framework for diachronic cultural analysis that integrates NLP, SNA, and GIS. The framework was validated on a comprehensive dataset of Bulgarian historical sources spanning the 10th to 21st centuries.

A. Key Computational Contributions

1. NLP Pipeline: We developed an annotated diachronic corpus of 2.5 million words with etymological and semantic annotations. The pipeline achieved 89% precision and 85% recall for lexical borrowing detection, and 82% F1 for semantic change detection using a fine-tuned BERT-based model (bert-base-multilingual-cased).
2. Network Analysis: Manuscript transmission network analysis (n=247 nodes, 583 edges) identified Preslav and Tarnovo as primary cultural hubs with degree centrality of 0.78 and 0.72 respectively. Network density increased from 0.12 (10th-12th c.) to 0.28 (13th-14th c.), demonstrating intensified cultural transmission during the Second Bulgarian Empire. Community detection via the Louvain algorithm revealed three distinct transmission communities corresponding to geographical regions.
3. Spatial Analysis: GIS-based heatmapping revealed three distinct zones of Ottoman-period architectural influence with 78.3% co-location of commercial structures with medieval settlements ($p < 0.01$, spatial permutation test). Moran's I analysis confirmed significant spatial clustering of Ottoman architectural features ($I = 0.67$, $p < 0.001$).
4. Quantitative Evidence: Lexical analysis identified 15,247 Turkish-origin loanwords with domain-specific distribution patterns: material culture (43.7%), administrative/legal (22.1%), trade/commerce (18.5%), and social relations (15.7%). Semantic shift analysis showed 34.2% of Ottoman-period borrowings underwent meaning specialization between 1878 and 1944, compared to 12.8% for Greek-origin terms ($p < 0.001$, two-tailed t-test).
5. Topic Modeling: LDA analysis of 25,000+ cultural policy documents (1944-1989) identified three dominant thematic clusters with temporal dynamics: Soviet ideological content (decreasing from 78% to 52%), national historical narrative (increasing from 1960s onward), and folkloric tradition (consistently present, 22-28% of topics).

B. Methodological Implications

The framework demonstrates that computational methods can provide quantitative, reproducible evidence for cultural processes traditionally studied through qualitative historical analysis. The integration of multiple computational approaches—NLP, network analysis, and GIS—enables cross-validation of findings and reveals patterns that may not be apparent through any single method alone. The publicly available annotated corpus and reproducible analysis toolkit establish a foundation for future research in computational cultural analytics.

C. Broader Applicability

While applied to Bulgarian historical data, the framework is designed to be extensible to other cultural contexts. The methodological pipeline can be adapted to:

- Other Balkan and Southeast European regions with similar multi-layered cultural histories
- Different historical periods and source types (e.g., administrative records, literary texts, archaeological data)
- Comparative analysis across multiple cultural regions

The annotated corpus, network analysis toolkit, and GIS processing pipeline are made publicly available for adaptation to other regions and periods.

D. Future Research Directions

Future work will address:

1. Advanced NLP: Development of specialized transformer-based language models for Old Church Slavonic, Ottoman Turkish, and other historical languages present in the corpus, enabling more accurate cross-lingual analysis.
2. Expanded Datasets: Integration of additional archives and multilingual sources, including Greek, Ottoman Turkish, and Latin texts, to provide broader coverage of the cultural contact zones.
3. Predictive Modeling: Machine learning approaches to modeling cultural change based on identified historical patterns, with applications in cultural policy, education, and heritage preservation.
4. Comparative Analysis: Application of the computational framework to other cultural contexts, enabling cross-regional comparison of cultural transfer mechanisms.
5. Public Engagement: Creation of accessible digital platforms for exploring cultural layering and hybridization, including interactive visualizations and educational resources.

E. Final Thesis

This study demonstrates that computational approaches can effectively model long-term cultural transfer and hybridization. The quantitative evidence supports the historical interpretation that Bulgarian culture was formed through selective appropriation and transformation of foreign elements into local traditions. By providing reproducible computational methods, we contribute to the growing field of computational cultural analytics and offer a foundation for culturally aware AI systems. The integration of computational methods with historical analysis opens new possibilities for understanding how cultures are formed through long-term processes of transfer, adaptation, and hybridization. As digital transformation accelerates, such approaches become increasingly relevant for understanding cultural identity in a globalized world.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support provided by the project: BG16RFPR002-1.014-0011-C01 Sustainable development of the Center for Excellence "Heritage BG".

The authors would like to thank the Research and Development Sector at the Technical University of Sofia for the financial support.

REFERENCES

[1] N. Antov, *The Ottoman "Wild West": The Balkan Frontier in the Fifteenth and Sixteenth Centuries*. Cambridge University Press, 2017.

[2] A. Mirkova, "Population Politics at the End of Empire: Migration and Sovereignty in Ottoman Eastern Rumelia, 1877–1886," *Comparative Studies in Society and History*, vol. 55, no. 4, pp. 789–820, 2013.

[3] L. Nahodilová, "Communist Modernisation and Gender: The Experience of Bulgarian Muslims, 1970–1990," *Contemporary European History*, vol. 19, no. 3, pp. 223–242, 2010.

[4] D. Blagoeva and S. Kolkovska, "On Some Current Processes in Bulgarian Lexis," in *Bulgarian Lexicology and Lexicography*, Sofia: BAS Publishing, 2021, pp. 235–250.

[5] T. Dragostinova, "The 'Natural Ally' of the 'Developing World': Bulgarian Culture in India and Mexico," *Slavic Review*, vol. 77, no. 4, pp. 982–1003, 2018.

[6] National Library "Sts. Cyril and Methodius," "Virtual Gallery of Copies of *History Slavenobolgarska*," 2022. [Online]. Available: https://www.nationallibrary.bg

[7] State Archives Agency, "Thematic Portals: Jewish Community in Bulgaria, Wars of Bulgaria, The Change 1989," [Online]. Available: https://archives.bg

[8] UNESCO, "Boyana Church," "Rock-Hewn Churches of Ivanovo," "Rila Monastery," World Heritage List. [Online]. Available: https://whc.unesco.org

[9] G. Lonergan, "Counting Citizens: The Transfer and Translation of Census Categories from the International Statistical Congresses to the Principality of Bulgaria, 1872–1888," *Nationalities Papers*, vol. 46, no. 4, pp. 666–682, 2018.

[10] T. Henninger, "The Greek Layer in the Bulgarian Literary Language," *Journal of Slavic Linguistics*, vol. 12, no. 1, pp. 45–67, 1990.

[11] S. Keremidchieva, *The Language of Bulgarian Cuisine*. Sofia: Institute for Bulgarian Language, BAS, 2022.

[12] L. Braga and D. Stefanov, "Lexis in the Field of Financial-Monetary and Trade Relations," in *Proceedings of the Conference of the Institute for Bulgarian Language*, 2024.

[13] M. Stamenova and G. Zvancharova, "On the Dictionary of Turkish Loanwords in the Bulgarian Language," *Izvestia of the Institute for Bulgarian Language*, vol. 30, pp. 73–88, 2018.

[14] S. Katsikas and P. Siani-Davies, "Introduction: The Europeanization of Bulgarian Society – A Long-Lasting Political Project," in *Bulgaria and Europe*, London: Anthem Press, 2010, pp. 1–22.

[15] R. J. Crampton, *A Concise History of Bulgaria*. Cambridge University Press, 2005.

[16] D. D. Egbert, "Politics and the Arts in Communist Bulgaria," *Slavic Review*, vol. 26, no. 2, pp. 280–297, 1967.

[17] National Statistical Institute, "Population as of September 7, 2021," Census Data, 2022. [Online]. Available: https://www.nsi.bg

[18] OECD, "International Migration Outlook 2025: Bulgaria Country Note," 2025.

[19] European Commission, "Bulgaria in the European Union: Accession, Schengen, and Eurozone Membership," 2026. [Online]. Available: https://european-union.europa.eu

[20] Institute for Literature, BAS, "Encyclopedia of Bulgarian Literature: Kilifarevo Monastery Entry," [Online]. Available: https://enc.ilit.bas.bg