タタール語におけるサブワード単位の言語識別を加味したキリル文字からラテン文字への翻字システムの開発

坂井 優介 * 田口 智大 * 渡辺 太郎
奈良先端科学技術大学院大学
{sakai.yusuke.sr9,taguchi.chihiro.td0,taro}@is.naist.jp

概要

現代のタタール語におけるキリル文字からラテン文字への翻字は困難を極める．タタール語はロシア語との単語内でのコードスイッチングが含まれている影響で，単純な翻字規則だけではタタール語の翻字を行うことはできない．またタタール語は低資源言語であり，特にラテン文字で書かれたタタール語の文章の入手は極めて困難であるため，深層学習などの統計的手法の適用は現実的ではない．

本研究では，タタール語におけるサブワードレベルの言語識別に基づいた翻字システムを提案する．サブワードごとにタタール語かロシア語かを判別する言語識別器を用意し，それに基づき特定された言語に応じて該当する翻字規則を適用することで翻字を行う．提案手法では既存のタタール語の翻字システムより高い精度で翻字が行えることを示した．

1 はじめに

現代のタタール語にはキリル文字とラテン文字の2種類の正書法が存在している．タタール語由来の単語のみで構成された文の場合，簡単な翻字規則に従って相互に書き換え可能であるが，現代のタタール語には大量のロシア語の借用語が含まれており，文中でロシア語のフレーズ等に切り替わることがあるため，単純なタタール語の翻字規則のみでは正しく翻字することはできない[1)]．よってタタール語とロシア語の単語にはそれぞれ異なる翻字規則を適用しなければならないが，ロシア語の単語にタタール語の接尾辞が付いている場合等において単語内でコードスイッチング（intra-word CS）が発生するためトークンごと或いは形態素ごとに言語を特定する必要があり，形態素解析等が必要となる．また現代のタタール語のテキストにはすでにロシア語の借用語が大量に含まれており，言語識別器等を作成するための純粋なタタール語のテキストを得ることは容易ではない．以上の理由からタタール語におけるキリル文字からラテン文字への翻字は困難を極める．

既存のタタール語の翻字システムはタタール語のみの翻字規則を適用しているか [1]，ロシア語の借用語のための膨大なルールベースに基づいて対処している [2]．このため，5 節の実験結果でも述べているように，前者のタイプの翻字システムではロシア語の単語をサポートしていないため翻字精度が低い．また，後者のタイプではルールベースによりロシア語由来の単語についても翻字が行えるため翻字精度が高いが，依然としてある程度のロシア語由来の単語を正しく翻字できていない．よって厳密なルールベースの手法では全ての単語に対して逐次的・網羅的にルールを追加するための継続的な更新が必要であり，明らかに非現実的で非効率的である．

本研究ではタタール語におけるキリル文字からラテン文字へのシンプルで高精度な自動翻字システムを提案する．具体的にはタタール語とロシア語のそれぞれについて単純な翻字規則を用意し，文中の各サブワードに対してタタール語かロシア語か判別するための言語識別器を学習し，各サブワードに対して識別された言語の翻字規則を適用することを繰り返すことでキリル文字からラテン文字に翻字する．

その結果，本研究の提案手法は既存手法より高い翻字精度を達成した．また，提案手法はタタール語の翻字規則のみを用いた場合より高い精度を示していることから提案手法がロシア語の形態素をある程度正しく予測し，翻字できることを示している．

1.1 タタール語でのコードスイッチング

タタール語におけるコードスイッチングの例として (1) にラテン文字での翻字と訳文を示す．下線部の класс（*klass*「学級，学年」）はロシア語からの借

* 共同責任者
1) このように2つ言語が文中で切り替わることをコードスイッチング（CS）という．

用語であり，ロシア語の音韻で発音されるため，ロシア語の翻字規則が必要となる．また，ロシア語からの借用語にタタール語の接尾辞が付くこともあり，単語内でのコードスイッチング（intra-word CS）が発生するため，この例ではタタール語の位置指定の接尾辞 -та（*-ta* 「〜で」）が付いている．

(1)Безнең класста кызлар сигезенчедән эчә башлаган иде.
translit.: Bezneñ klassta qızlar sigezençedän eçä başlağan ide.
“私たちのクラスでは，女子は八年生から飲み始めていた．”

2 関連研究

2.1 コードスイッチング（CS）

CS は口語的な言語現象であり CS がテキストに記録されることはほとんどないため, リソース不足が大きな難点となっている．文献 [3] に出版時点での利用可能な CS データセットを列挙している．データセットの入手の容易さと整備済みのタスク等の観点からヒンディー語–英語 [4, 5]，スペイン語–英語 [6, 7]，アラビア語の方言（アーンミーヤ）から現代標準アラビア語 [8] が CS の研究のトレンドとなっている．intra-word CS の翻字研究はドイツ語–トルコ語，スペイン語–ウイチョル語 [9]，オランダ語–リンブルフ語 [10]，トルコ語–英語 [11] が存在しており，これらの手法は本研究と同様のアプローチをとっている．本研究の手法との違いは Mager ら [9] は単語分割と言語識別に SegRNN [12] を採用していること，Nguyen ら [10] は形態素ごとの分割に Morfessor [13] を採用していることである．しかしサブワードごとの言語識別と翻字を組み合わせた本研究のタスクは行われていない．

2.2 タタール語の翻字

現時点でタタール語のキリル文字からラテン文字への翻字には以下のサービスが利用できる．The Tatar Transcription Tool（TTT）[1] は Mari Web Project の一環としてウィーン大学がオンラインで公開している翻字サービスである．speak.tatar[2] は匿名で開発された翻字サービスである．Aylandirow [2] はタタール語についての膨大なルールベースを基に翻字を行うシステムであり，タタール語由来の単語だけでなくロシア語由来の単語も広くカバーしている．

2） https://speak.tatar/en/language/converter/tat/cyrillic/latin

3 提案手法

タタール語の翻字は単語を跨いだ翻字規則が無いため単語ごとに処理を行う[3]．

はじめに，タタール語には intra-word CS が含まれていることを考慮し，各サブワードに対してロシア語かタタール語のどちらかを判別する言語識別器を作成する．言語識別器を作成するためにタタール語とロシア語の 2 つの単言語コーパスを用意した．現代のタタール語のテキストには CS を含まない純粋なタタール語のテキストは存在しない[4]．そのため言語識別器の学習にはタタール語のテキスト内のロシア語の形態素のノイズを避ける目的でロシア語の借用語を含まない 1912 年に翻訳されたタタール語のコーラン[5]（19,691 語，重複あり）とエジプトの宗教財産（Awqaf）省 が翻訳したロシア語版コーラン[6]（21,256 語，重複あり）を採用した．

言語識別器の学習プロセスは以下の通りである．まずデータセット内の単語を Byte Pair Encoding（BPE）[14] で文をサブワードに分割する．BPE は SentencePiece [15] 等のサブワード手法とは異なり，短いサブワード列の結合によって長いサブワードを形成していく性質から，全サブワードの集合に文字単位のサブワードが含まれており，全ての文をサブワード化できるので，OOV 問題を解決できる [16]．今回のケースでは学習用の単言語データが乏しいため，BPE を採用して OOV 問題を回避した．またロシア語とタタール語の判別をしやすくするため，「長いサブワードはよりその言語の特徴を捉えたサブワードとなっている」という仮説のもと，BPE のアルゴリズムで結合できなくなるまでサブワードの結合操作を繰り返した．得られたサブワードを基に fastText[7] [17] を用いてサブワードごとの埋め込み表現を得る．得られたサブワードの埋め込み表現に言語ラベルを付与し，fastText [17] で提供されている教師あり分類器 [18] で言語識別器を作成する[8]．

3） ソースコード: https://github.com/naist-nlp/tatar_transliteration. デモンストレーションサイト: https://yusuke1997.com/tatar.
4） 今回用意した評価用データセットには 8,466 語のうち 1,009 語に少なくとも 1 つのロシア語の形態素が含まれていた．
5） https://cdn.jsdelivr.net/gh/fawazahmed0/quran-api@1/editions/tat-yakubibnnugman.json
6） https://cdn.jsdelivr.net/gh/fawazahmed0/quran-api@1/editions/rus-ministryofawqaf.json
7） https://github.com/facebookresearch/fastText
8） ハイパーパラメータとして次元数 16，文字 n-gram の最小値は 2，最大値は 4 である．損失関数には Hierarchical softmax を用いた．なおハイパーパラメータ等の設定は次の記事を参考にした．https://fasttext.cc/blog/2017/10/02/blog-post.html

	BLEU	LCS F-score	ACC	# error word
speak.tatar	0.869	0.953	0.952	1,747
TTT	0.879	0.956	0.946	1,505
Aylandirow	0.971	**0.994**	0.991	526
tt-based	0.968	0.989	0.989	552
tt-ru hybrid	**0.981**	**0.994**	**0.993**	**332**

表 1 検証用データ 700 文中，重複なし 5,261 語における翻字精度の比較結果

作成した言語識別器を用いて各サブワードの言語ラベルを予測する．ここで隣接する同じ言語ラベルのサブワードは可能な限り結合してより長いサブワードにする．例えば 2 つの連続したサブワードの言語ラベルが両方ともタタール語である場合，それらは結合され，1 つのタタール語の言語ラベルが付与される．最後にそれぞれのサブワードは予測された言語ラベルの翻字規則に基づいてラテン文字へと変換され，1 つの単語にまとめられて出力される．

なお Mager ら [9] とは異なり本研究は深層学習によるアプローチを行っていない．Mager ら [9] のタスクがマルチラベリングなのに対して，本研究における言語識別タスクは低リソースの訓練データを用いた二値分類器であり，このタスクの達成のために深層学習の使用は高コストと判断したためである．

4 実験設定

性能評価のために Corpus of Written Tatar [19] からランダムに抽出した 700 文[9]を手動でラテン文字のタタール語に翻字し，ネイティブスピーカーがチェックしたものを検証用データとして用意した．

また，コーパス内のロシア語の形態素がタグ付けされるように，キリル文字のテキストにアノテーションを施した．その結果，検証用データにはロシア語の形態素が含まれる単語が 1,009 語（重複あり），intra-word CS が含まれる単語が 598 語（重複あり）含まれていた．評価指標は単語内の各文字に対して BLEU と LCS（Longest Common Sequence）F-measure，さらに単語全体の翻字精度（ACC）[10]を用いた．また単語として誤翻字した単語数も計測した．LCS F-measure と ACC の計算は Chen ら [20] を基にしている．実験では提案手法（tt-ru hybrid）だけでなく，タタール語のみの翻字規則のみを用いた評価（tt-based）も行った．また既存手法との比較に

9） 重複ありで 8,466 語，重複なしで 5,261 語

10） ACC は文字誤り率（CER）を用いて算出され，CER と ACC は相補的関係にある（$ACC = 1 - CER$）．

set (total 5,261 words)	# words
$V(T)$	552
$V(H)$	332
$V(T) \cap V(H)$	216
$V(T) \setminus V(H)$	336
$V(H) \setminus V(T)$	116

表 2 tt-based と tt-ru hybrid の誤翻字した単語数の比較（重複なし）．$V(T)$ は tt-based で誤翻字した単語の集合，$V(H)$ は tt-ru hybrid で誤翻字したの集合，$V(T) \cap V(H)$ は両方とも誤翻字した単語の集合，$V(T) \setminus V(H)$ は tt-based のみ誤翻字した単語の集合，$V(H) \setminus V(T)$ は tt-ru hybrid のみ誤翻字した単語の集合である．

	ru words		CS words	
	#	accuracy	#	accuracy
speak.tatar	**805**	**0.798**	455	0.752
TTT	258	0.256	175	0.283
Aylandirow	738	0.731	461	0.762
tt-based	471	0.467	294	0.486
tt-ru hybrid	788	0.781	**464**	**0.767**

表 3 ロシア語の単語と intra-word CS の単語における正しく翻字できた単語数の比較．ru words は検証用データ内のロシア語 1,009 語のうち，正しく翻字できた数と割合．CS words は 605 語のうち正しく翻字できた数と割合．

speak.tatar，TTT，Aylandirow を用いた．

5 実験結果

表 1 に示した実験結果より，提案手法の tt-ru hybrid が全ての評価指標において既存手法よりも高い精度となっている．ロシア語の翻字規則をサポートしていない手法（特に speak.tatar，TTT）とロシア語の翻字規則をサポートしている手法（Aylandirow，tt-ru hybrid）の間には評価値に差があり，特に BLEU スコアでは 0.1 ポイントの差があることがわかる．なおタタール語のみの翻字規則で作成した tt-based は同じタタール語のみの翻字規則をサポートした 2 つ翻字手法（speak.tatar，TTT）より全体的に高いスコアであった理由は 6 節で説明する．また同じロシア語の翻字規則をサポートした Aylandirow と比較すると tt-ru hybrid は BLEU で 0.01，ACC は 0.002 ポイント上回り，LCS F-score は同等であった．さらに誤翻字した単語数を比較すると Aylandirow は 526 語なのに対して tt-ru hybrid は 332 語と大幅に改善した．

6 分析

tt-based はロシア語の翻字規則を適用していないのにも関わらず，タタール語のみの翻字規則をサポートした他の 2 つの手法より高いスコアである．これは，既存手法が現在のタタール語の翻字規則

Source	РФ Закон чыгаручылар советы президиумы утырышында катнашты.
tt-based	RF Zaqon çığaruçılar sowetı prezidiumı utırışında qatnaştı.
tt-ru hybrid	RF Zakon çığaruçılar sovetı prezidiumı utırışında qatnaştı.
Source	Һәр юлчы туктап, аның хозурлыгына сокланып китә.
tt-based	Här yulçı tuqtap, anıñ xozurlığına soqlanıp kitä.
tt-ru hybrid	Här yulçı tuqtap, anıñ xozurlığına soklanıp kitä.

表 4 tt-ru hybrid が翻字に成功した例と失敗した例．下線部の単語はロシア語の借用語における誤翻字の単語である．上の文は tt-ru hybrid が下線付きの単語を正しく翻字できているのに対して，下の文は tt-ru hybrid が下線の付いた単語を誤翻字し，tt-based が正しく翻字できた例である．なお，各手法における出力例の比較は補足の表 6 に示しておく．

（2013Latin）に従っていないことに起因している[11)]．例えば тәнкыйть（批判）という単語では，2013Latin では *tänqit* と訳されるべきところ，TTT では *tänqit'* と訳される．実際，ラテン語の正書法を規定する機関が事実上存在していないため，インターネット上では様々な綴り方が見受けられる．従って，tt-based の高いスコアは正書法の一貫性に起因している．また，この点を考慮すると，tt-based と tt-ru hybrid の間でスコアが向上したことは，ロシア語の借用語の翻字に対応するためにサブワード単位で言語識別を行ったことが成功していることを意味している．

表 2 に tt-based と tt-ru hybrid について翻字ミスとなった単語数を示している．表 2 から tt-based で観測された誤りの単語のうち 336 語が tt-ru hybrid で正しく翻字されたことがわかる．表 4 に tt-ru hybrid と tt-based を比較した際に tt-ru hybrid が翻字に成功した例と失敗した例を載せる．表 4 の上段の文中の закон と совет について，tt-based では *zaqon* と *sowet* に変換されて誤りとなっているが，tt-ru hybrid では *zakon* と *sovet* となり翻字に成功している．これらの単語はロシア語からの借用語であるため，言語をうまく識別できていることがわかる．しかし，表 2 に示したように，tt-based で正しく翻字できていた 116 単語については，tt-ru hybrid ではむしろ誤った翻字が行われていた．表 4 の下段では，tt-ru hybrid では下線部の *soklanıp* が正しく翻字できていない一方で，tt-based では *soqlanıp* と正しい翻字結果になっている．これは，言語識別器が該当箇所のサブワードをロシア語として誤認識したためにロシア語の翻字規則が適用されたことが原因となっている．

Aylandirow の LCS F-score は提案手法と同程度に高いスコアであるが，これは Aylandirow がロシア語の借用語のうち頻出するものをルールに記述しているためである．しかしこの手法は言語識別自体は行っていないため，全ての単語についてのルールを記述することはできない以上，カバレージ問題がつきまとう．例えば，Aylandirow ではロシア語の借用語である такси（*taksi*）がルールに含まれていないことで，*taqsi* へ誤翻字されることが確認された．

表 3 にロシア語との CS 単語に関する翻字性能の比較を示す．表 3 よりロシア語の形態素を含む単語については，tt-ru hybrid が既存手法より良い精度で翻字できたことが確認できる[12)]．特にロシア語の形態素に対して tt-ru hybrid は 78.1%，tt-based は 46.7% であることから，ロシア語を含む単語の翻字精度の向上が認められる．タタール語の中にロシア語の CS が恣意的に含まれることを考慮すると，ルールベースのシステムは数多く存在する未知のロシア語に対して柔軟な翻字を行うことができないが，言語識別器を用いた提案手法はどのロシア語の単語に対しても安定した性能を発揮することが期待できる．

7 まとめ

本研究ではタタール語におけるキリル文字からラテン文字への新しい翻字手法を提案した．ロシア語の借用語にはタタール語とは異なる翻字規則を適用されていることや intra-word CS が頻繁に起こっていることを考慮して提案手法ではサブワードごとに言語識別を行い，識別された言語ごとの翻字規則を適用して翻字を行った結果，既存の翻字サービスより優れた精度で翻字が行えた．本研究は言語情報に依存していないシンプルなアーキテクチャを採用しており，構文解析や POS タグ付けなどの詳細な解析を必要としないため，intra-word CS が存在する他の低リソース言語にも適用できる可能性がある．

11） 本研究で用いた正書法は 2013 年に制定された最新かつ主流の正書法である（補足 A.1 参照）．

12） speak.tatar がロシア語の単語に対して最も良いスコアを出しているが，これは単に翻字規則がロシア語の単語に対してうまく機能するように設計されているからであり，対象的にタタール語の単語に対する精度は低いことが表 1 よりわかる．

参考文献

[1] Jeremy Bradley. Tatar transcription tool, 2014. Available at: https://www.univie.ac.at/maridict/site-2014 [retrieved 28 March 2021].

[2] Dinar Korbanov. Aylandirow, n.d. Available at: http://aylandirow.tmf.org.ru [retrieved 29 March 2021].

[3] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, and J. P. McCrae. A survey of current datasets for code-switching research. In **2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)**, pp. 136–141, 2020.

[4] Abhishek Srivastava, Kalika Bali, and Monojit Choudhury. Understanding script-mixing: A case study of Hindi-English bilingual Twitter users. In **Proceedings of the The 4th Workshop on Computational Approaches to Code Switching**, pp. 36–44, Marseille, France, May 2020. European Language Resources Association.

[5] Pranaydeep Singh and Els Lefever. Sentiment analysis for Hinglish code-mixed tweets by means of cross-lingual word embeddings. In **Proceedings of the The 4th Workshop on Computational Approaches to Code Switching**, pp. 45–51, Marseille, France, May 2020. European Language Resources Association.

[6] Elena Alvarez-Mellado. An annotated corpus of emerging anglicisms in Spanish newspaper headlines. In **Proceedings of the The 4th Workshop on Computational Approaches to Code Switching**, pp. 1–8, Marseille, France, May 2020. European Language Resources Association.

[7] Daniel Claeser, Samantha Kent, and Dennis Felske. Multilingual named entity recognition on Spanish-English code-switched tweets using support vector machines. In **Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching**, pp. 132–137, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Injy Hamed, Moritz Zhu, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. Code-switching language modeling with bilingual word embeddings: A case study for egyptian arabic-english, 2019.

[9] Manuel Mager, Özlem Çetinoglu, and Katharina Kann. Subword-level language identification for intra-word code-switching. **CoRR**, Vol. abs/1904.01989, , 2019.

[10] Dong Nguyen and Leonie Cornips. Automatic detection of intra-word code-switching. In **Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology**, pp. 82–86, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Zeynep Yirmibeşoğlu and Gülşen Eryiğit. Detecting code-switching between Turkish-English language pair. In **Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text**, pp. 110–115, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[12] Liang Lu, Lingpeng Kong, Chris Dyer, Noah A. Smith, and Steve Renals. Segmental recurrent neural networks for end-to-end speech recognition. In **Proceedings of Interspeech 2016**, Interspeech, pp. 385–389. International Speech Communication Association, September 2016. Interspeech 2016 ; Conference date: 08-09-2016 Through 12-09-2016.

[13] Mathias Creutz and Krista Lagus. Morfessor in the morpho challenge. In **PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes**, 2006.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[15] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[16] Tamali Banerjee and Pushpak Bhattacharyya. Meaningless yet meaningful: Morphology grounded subword-level NMT. In **Proceedings of the Second Workshop on Subword/Character LEvel Models**, pp. 55–60, New Orleans, June 2018. Association for Computational Linguistics.

[17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 135–146, 2017.

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.

[19] Mansur R. Saykhunov, R. R. Khusainov, T. I. Ibragimov, J. Luutonen, I. F. Salimzyanov, G. Y. Shaydullina, and A. M. Khusainova. Corpus of written tatar, 2019. Available at: http://www.corpus.tatar [retrieved 28 March 2021].

[20] Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. Report of NEWS 2018 named entity transliteration shared task. In **Proceedings of the Seventh Named Entities Workshop**, pp. 55–73, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[21] David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. **Ethnologue**. SIL International, Dallas, Texas, 24th edition, 2021. Available at: http://www.ethnologue.com [retrieved 12 March 2021].

[22] Guzel A. Izmailova, Irina V. Korovina, and Elzara V. Gafiyatova. A study on tatar–russian code switching (based on instagram posts). **The Journal of Social Sciences Research**, Vol. Special Issue. 1, pp. 187–191, 2018.

[23] National Council of the Republic of Tatarstan. О восстановлении татарского алфавита на основе латинской графики [on the restoration of the tatar alphabet based on the latin script], September 1999. Available at: http://docs.cntd.ru/document/917005056 [retrieved 28 March 2021].

[24] Boris Yeltsin. О языках народов Российской Федерации [on the languages of nations in the russian federation], 2020. First published in 1991, last amended in 2020. Available at: http://pravo.gov.ru.

[25] National Council of the Republic of Tatarstan. Об использовании татарского языка как государственного языка Республики Татарстан [on the use of the tatar language as the national language of the republic of tatarstan], January 2013. Available at: http://docs.cntd.ru/document/463300868 [retrieved April 26, 2021].

[26] Aynur Akhatovich Timerkhanov and Gulshat Rafailevna Safiullina. **Tatarça-inglizçä, inglizçä-tatarça süzlek: Tatar-English, English-Tatar dictionary**. G. Ibragimov Institute of Language, Literature and Art, Kazan, 2019.

A　タタール語の言語学的な背景

タタール語（ISO 639-1 Code: tt）は，トゥルク語族キプチャク語群に属する言語で，主にロシア国内のタタールスタン共和国で話されている．話者数は約 500 万人と推定されている [21]．一般的には語順は SOV の形をとるが，スクランブリングを伴う自由な語順を許容している．膠着語であり，格標示や動詞の屈折などは接尾辞によって導かれる．接尾辞は母音の後方性あるいは前方性が一貫して保たれる母音調和規則に従う．現代のタタール語は，ロシア語との言語接触の影響により，ロシア語の借用語を多く含んでいる．話者の多くはロシア語とのバイリンガルで，都市部に住む若い世代はタタール語よりもロシア語の方が得意な傾向にある．このようなバイリンガリズムは，特に口語での頻繁なコードスイッチング（CS）を引き起こしている [22]．

A.1　タタール語の正書法

現代のタタール語の正書法は，ロシア語の正書法の文字に 6 つの拡張キリル文字を加えたものが主流である．キリル文字のタタール語はロシア語圏に住むタタール人が主に使用しており，トルコやフィンランドなど他の地域に住むタタール人はラテン文字を使用している．1928 年まではタタール語はアラビア文字で書かれていたが，1920 年代から 30 年代にかけてソビエト連邦で行われたラテン語化計画に伴い，タタール語にラテン文字の正書法であるヤンガリフ（yañalif）が導入された．その後，1939 年に政治的な理由によってヤニャリフはキリル文字に変更され，現在のタタールスタン共和国で公式に使用されるに至る．しかしソ連崩壊後，ラテン文字正書法の復活を求める運動が再燃したため，1999 年にタタールスタン共和国法により新しいラテン語文字正書法が採用された [23]．しかしながら，2002 年にロシア連邦法に「ロシアのすべての民族言語はキリル文字で書かなければならない」という新たな項目が追加されたことにより，この法律の効力はすぐに失ってしまった [24]．現在のタタール語のラテン文字アルファベット（2013Latin）はテュルク語共通アルファベットをベースとしており，2013 年にタタールスタン共和国法で正式に採用され，タタール人のディアスポラコミュニティで使用されている [25]．本稿で使用するラテンアルファベットは 2013Latin とし，キリル文字からラテン文字への変換の詳細な翻字規則は文献 [26] に記載されている．また表 5 にも簡易的な対応表を示す．

Cyrillic	Latin (Tatar)	Latin (Russian)	Cyrillic	Latin (Tatar)	Latin (Russian)
а	a	a	у	u, uw, w	u
б	b	b	ф	f	f
в	w	v	х	x	x
г	g, ğ	g	ц	NA	ts
д	d	d	ч	ç	ç
е	e, ye, yı	e	ш	ş	ş
ё	NA	yo	щ	NA	şç
ж	j	j	ъ	—	—
з	z	z	ы	ı	ı
и	i	i	ь	—	—
й	y	y	э	e, ’	e
к	k, q	k	ю	yu, yü, yuw, yüw	yu
л	l	l	я	ya, yä	ya
м	m	m	ә	ä	NA
н	n	n	ө	ö	NA
о	o	o	ү	ü, üw, w	NA
п	p	p	җ	c	NA
с	s	s	ң	ñ	NA
т	t	t	һ	h	NA

表 5　タタール語におけるキリル文字とラテン文字の対応表．NA (not applicable) は対応する文字が存在しないことを意味している．「—」は その文字が翻字において無視されることを意味している．

original	РФ Президенты Владимир Путин Россия мөселманнарын изге Рамазан
gold (Latin)	RF Prezidentı Vladimir Putin Rossiyä möselmannarın izge Ramazan
speak.tatar	RF Prezidentı Vladimir Putin Rossiyä möselmannarın izge Ramazan
TTT	RF Prezidentı Wlädimir Pütin Rössiyä möselmannarın izge Ramazan
Aylandirow	RF Prezidentı Vladimir Putin Rossiä möselmännarın izge Ramazan
tt-based	RF Prezidentı Wladimir Putin Rossiyä möselmannarın izge Ramazan
tt-ru hybrid	RF Prezidentı Vladimir Putin Rossiyä möselmannarın izge Ramazan

表 6　各手法の翻字結果の例．上段 2 行はキリル文字で書かれた原文と翻字されたラテン文字のタタール語．中段 3 行は先行研究の手法，下段はタタール語のみの翻字規則を適用した tt-based と提案手法の tt-ru hybrid の翻字結果である．