Pythonによる翻字システムの開発と課題の一例
―中央アジアのチュルク諸語を題材として―

日髙 晋介
(日本学術振興会特別研究員PD／新潟大学)

2023/4/19 (水) Luncheon Linguistics
@ 東京外大研究講義棟 419 語学研究所 + Zoom

自己紹介

学歴：

2006年4月 外国語学部 日本課程 日本語専攻 入学

2020年3月 『ウズベク語における形動詞と動名詞による従属節について』博士学位取得＠東京外国語大学

研究テーマ：

中央アジアにおけるチュルク諸語のモダリティ体系の解明

経歴など：

日高 晋介 (Shinsuke Hidaka) - マイポータル - researchmap

最近の関心

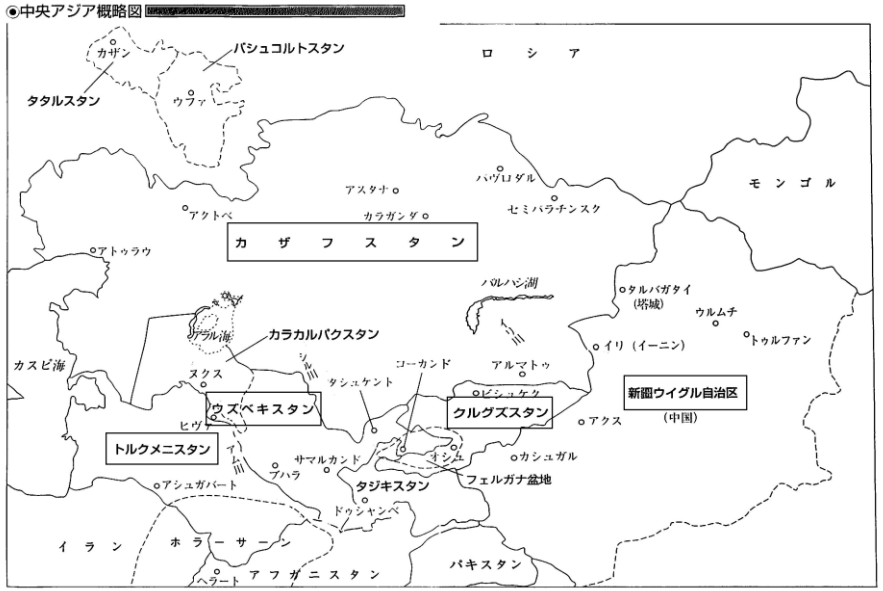


図 1: 中央アジアの概略図
(宇山編 2010: 2)

Johanson (1998: 82-83, 2021: 21-23)による
系統的分類：

南西 (オグズ) 語群：
Trk. = トルクメン語 (トルクメニスタン; 約666万)

北西 (キプチャク) 語群：
Kr. = キルギス語 (クルグズスタン (キルギス); 約513万)
Ka. = カザフ語 (カザフスタン; 約1270万)

南東 (カルルク) 語群：
Uz. = ウズベク語 (ウズベキスタン; 約2800万人)
Uy. = 現代ウイグル語 (新疆ウイグル自治区; 約1041万)

※ 話者数は、Ethnologue: Languages of the World (https://www.ethnologue.com/) の
各言語のページのTotal users in all countriesを参照

最近の関心

中央アジアの五つのチュルク諸語は、系統的には、南西 (オグズ) 語群・北西 (キプチャク) 語群・南東 (カルルク) 語群のいずれかに分類される。
ただし、系統的な分類を超えた共通の特徴を持つことも指摘されている。

-GAn-turkic (Schönig 1999: 72)

北西語群・南東語群に属する言語、および南シベリアのチュルク諸語とハラジ語が含まれ、南西語群に属するトルクメン語も -GAn-turkic のいくつかの特徴を持つとされている。

最近の関心

(1)「私が昨日**買ってきた**本」
a. トルクメン語 (南西語群; 風間 2022: 480):
*Men düýn **[sat-yn al]-an** kitab-ym*
I yesterday buy[trade-CVB.SIM take]-PTCP.PFV book-1SG
b. キルギス語 (北西語群; 風間・アクマタリエワ 2022: 668):
*Kečéě **sat-ïp al-gan** kiteb-im*
yesterday [sell-CVB.PFV take]buy-PTCP.PFV book-1SG.POSS
c. カザフ語 (筆者による作例、母語話者確認済):
*keše **sat-ïp al-ɣan** kitab-ïm*
yesterday sell-CVB.SEQ take-PTCP.PAST book-1SG.POSS
d. ウズベク語 (南東語群; 風間・日髙 2022: 714):
*(Men) kecha **sot-ib ol-ib kel-gan** kitob*
I yesterday buy[sell-CVB.PFV take]-CVB.PFV come-PTCP.PFV book
e. 現代ウイグル語 (南東語群; 風間・新田 印刷中):
*Men tünügün **al-ǧan** kitab*
I yesterday take-PTCP.PF book

最近の関心

- 系統的な分類を超えた共通の特徴を持つ
  例: -GAn-turkic (Schönig 1999: 72)
  →他にどんな現象なのか？その現象に理由はあるのか？

- 博士論文：
  『ウズベク語における形動詞と動名詞による従属節について』
  ⇒PDではより範囲を広く
  「ウズベク語」から「中央アジアのチュルク諸語」へ
  「非定形動詞（動名詞・形動詞）」から「モダリティを表す諸表現」へ

- PDでの研究課題：
  「中央アジアにおけるチュルク諸語のモダリティ体系の解明」

最近の関心

論文：

- 日髙晋介 (2023)「ウズベク語における推定・可能性を表す分析的表現の差異：*V-sa kerak*と *V-(i)sh mumkin* に注目して」『北東アジア諸言語の記述と対照』3: 79–98.
- 日髙晋介 (2023)「中央アジアのチュルク諸語におけるモダリティ対照の試み」『言語の普遍性と個別性』 14: 109–35.
- 日髙晋介（2023)「ウズベク語における小詞 *=chi* の機能」『北方言語研究』13: 17–38.

最近の関心

日本言語学会 第166回大会 6/17-18 @専修大学

- **6/17（土）口頭発表 A会場 14:55－15:25**

  「ウズベク語における命題的モダリティを表す分析的表現の相互承接―主観性に注目して―」

- **6/18（日）ワークショップ B会場 10:00-12:00**

  「チュルク諸語の副動詞にまつわる諸問題 ―節連結・副詞句・複雑述語―」

  企画者：日髙　晋介

  司会者：アクマタリエワ・ジャクシルク

  コメンテーター：江畑　冬生

成果発表のために速くミスなく翻字したい

現時点で主に用いられている文字

（※ ロシア語やアラビア語の表記のために使用されている文字とは、それぞれ少し異なる）

トルクメン語：　キリル文字からラテン文字に完全移行

カザフ語：　キリル文字（カザフスタン）、アラビア文字（中国）

※ ラテン文字化については議論中？

キルギス語：　キリル文字

ウズベク語：　キリル文字からラテン文字にほぼ移行

現代ウイグル語：　アラビア文字（文字事情は菅原 2012に詳しい）

ウズベク語ラテン文字アルファベット			
立体	斜体	筆記体	対応するウズベク語キリル文字
A a	*A a*	𝒜 𝒶	А а
B b	*B b*	𝐵 𝒷	Б б
D d	*D d*	𝒟 𝒹	Д д
E e	*E e*	ℰ 𝑒	Е е / Э э（語頭のみ）
F f	*F f*	ℱ 𝒻	Ф ф
G g	*G g*	𝒢 𝑔	Г г
H h	*H h*	ℋ 𝒽	Ҳ ҳ
I i	*I i*	ℐ 𝒾	И и
J j	*J j*	𝒥 𝒿	Ж ж
K k	*K k*	𝒦 𝓀	К к
L l	*L l*	ℒ 𝓁	Л л
M m	*M m*	ℳ 𝓂	М м
N n	*N n*	𝒩 𝓃	Н н
O o	*O o*	𝒪 ℴ	О о
P p	*P p*	𝒫 𝓅	П п

立体	斜体	筆記体	対応するウズベク語キリル文字
Q q	*Q q*	𝒬 𝓆	Қ қ
R r	*R r*	ℛ 𝓇	Р р
S s	*S s*	𝒮 𝓈	С с
T t	*T t*	𝒯 𝓉	Т т
U u	*U u*	𝒰 𝓊	У у
V v	*V v*	𝒱 𝓋	В в
X x	*X x*	𝒳 𝓍	Х х
Y y	*Y y*	𝒴 𝓎	Й й
Z z	*Z z*	𝒵 𝓏	З з
O‘ o‘	*O‘ o‘*	𝒪̃ ℴ̃	Ў ў
G‘ g‘	*G‘ g‘*	𝒢̃ 𝑔̃	Ғ ғ
Sh sh	*Sh sh*	𝒮𝒽 𝓈𝒽	Ш ш
Ch ch	*Ch ch*	𝒞𝒽 𝒸𝒽	Ч ч
ng	*ng*	𝓃𝑔	нг
’	’	’	Ъ ъ（分離記号）

島田志津夫 (2019)『大学のウズベク語』pp. 20-21

成果発表のために速くミスなく翻字したい

ブラウザ上で使えるWEBコンバーター

- **キリル文字**から**ラテン文字**

  トルクメン語

  ウズベク語

  カザフ語

  （キルギス語はなし？）

- **アラビア文字**から**ラテン文字**

  現代ウイグル語

成果発表のために速くミスなく翻字したい

問題

- カザフ語、現代ウイグル語は複数のラテン文字体系が錯綜して存在するので、あるコンバータでどの文字体系を使用しているのかがパッと見で判断できない。
- 任意の文字体系で翻字できない。

要望

- ある文字で書かれたテキストを別の文字（任意の文字）に瞬時に置換できればなあ…

※「もっと効率的な方法があるのにな…」「もっとこうしたら使えるのにな…」などご助言・ご要望ありましたら、ぜひともお願いします。

Pythonでコンバーターを作る

Python (パイソン)とは？

　1990年代初頭ごろから公開されているプログラミング言語で、わかりやすく、実用的な言語として、広く使われ続けています。Pythonはプログラムの「読みやすさ・わかりやすさ」をとても重視していて、Pythonを知らない人でも、理解しやすいようにデザインされています。

　もちろん、読みやすさ一辺倒ではなく、実用的で、高い拡張性も備えています。読みやすさ・習得しやすさと、実用性のバランスが、Pythonの大きな魅力といえるでしょう。

（プログラミング言語 Pythonの紹介 - python.jp より引用）

Pythonでコンバーターを作る

Google Colaboratory (公式紹介ページ)

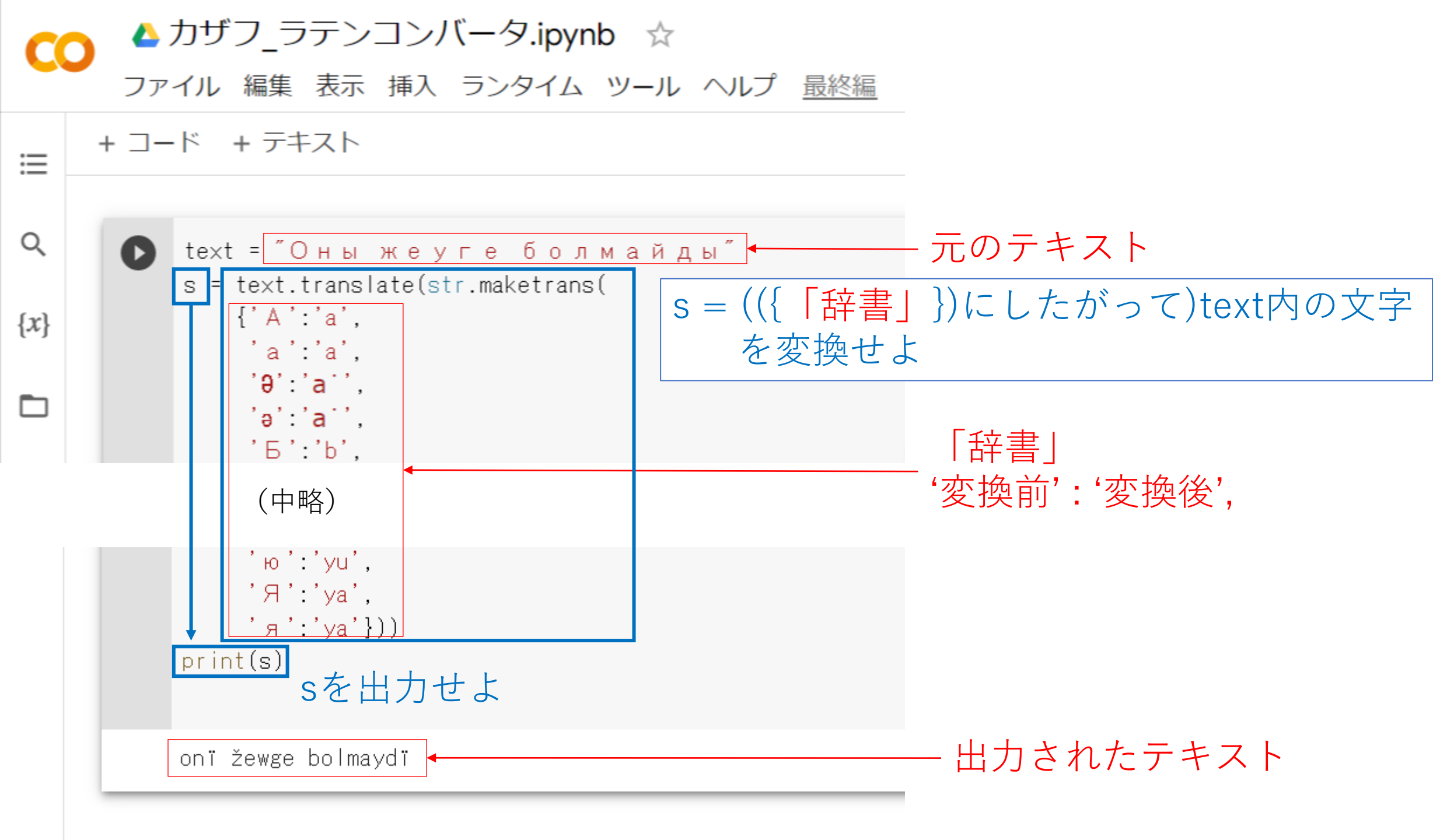
ColabはPythonの環境構築を一切せずに、サクッとPythonコードを書いて実行できる、便利なサービス

※ 要Googleアカウント

YouTube：

【超便利！】Google ColaboratoryでPythonを書いて動かす方法 ～Pythonプログラミング学習入門・初心者向け～


ドライブ ①
ドライブで検索
新しいフォルダ
ファイルのアップロード
フォルダのアップロード
Google ドキュメント
Google スプレッドシート
Google スライド
Google フォーム
その他 ②
100 GB 中 23.58 GB を使用
保存容量を購入
ドライブ › Colab Notebooks
オーナー
最終更新
ファイルサイ
Untitled 自分 15:04 自分 306 バイト
ウイグル_ラテンコンバータ.ipynb 自分 2023/03/05 自分 1 KB
カザフ_ラテンコンバータ.ipynb 自分 2023/02/27 自分 2 KB
Google 図形描画
Google マイマップ
Google サイト
Google Apps Script
Google Colaboratory ③
Google Jamboard
アプリを追加


カザフ_ラテンコンバータ.ipynb
ファイル 編集 表示 挿入 ランタイム ツール ヘルプ 最終編
+ コード + テキスト
text = "Оны жеуге болмайды"
元のテキスト
s = text.translate(str.maketrans(
{'А':'a',
'а':'a',
'Ә':'a',
'ә':'a',
'Б':'b',
(中略)
'ю':'yu',
'Я':'ya',
'я':'ya'}))
s = (({「辞書」})にしたがって)text内の文字を変換せよ
「辞書」
'変換前':'変換後',
print(s)
sを出力せよ
onï žewge bolmaydï
出力されたテキスト


ウイグル_ラテンコンバータ.ipynb
ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべて
+ コード + テキスト
0秒
text = "مەن ئۇنىۋېرسىتېتنىڭ تۆتىنچى يىلىدا ئوقۇيمەن"
s = text.translate(str.maketrans(
{'ا':'a',
'ە':'ä',
'د':'d',
'ر':'r',
'ز':'z',
'ژ':'ž',
もちろんアラビア文字も翻字可能

(中略)


'م':'m',
'ھ':'h',
'ئ':'¥''}))
print(s)
män 'uniwėrsitėtniŋ tötinči yilliqida 'oquymän

Pythonでコンバーターを作る

colabへのリンク（自己責任での使用をお願いします）：

カザフ_ラテンコンバータ.ipynb

ウイグル_ラテンコンバータ.ipynb

課題：

より効率のよい（＝処理速度が速まる）方法があるかも

参考：文字の変換にはstr.translate()が便利 - Qiita

参考文献

風間伸次郎 (2022)「トルクメン語：特集補遺データ「他動性」「ヴォイスとその周辺」「受動表現」「アスペクト」「モダリティ」「情報構造の諸要素」「否定、形容詞と連体修飾複文」「所有・存在表現」」『語学研究所論集』26: 439–99.

風間伸次郎・アクマタリエワ、ジャクシルク (2022)「キルギス語：特集補遺データ「受動表現」「他動性」「連用修飾複文」「情報構造と名詞述語文」「情報構造の諸要素」「否定、形容詞と連体修飾複文」「所有・存在表現」」『語学研究所論集』26: 649–697.

風間伸次郎・日髙晋介 (2022)「ウズベク語：特集補遺データ「連用修飾複文」「情報構造と名詞述語文」「否定、形容詞と連体修飾複文」「所有・存在表現」」『語学研究所論集』26: 699–732.

風間伸次郎・新田志穂 (印刷中)「現代ウイグル語：特集補遺データ『「他動性」「ヴォイスとその周辺」「連用修飾複文」「受動表現」「アスペクト」「モダリティ」「情報構造と名詞述語文」「所有・存在表現」「否定、形容詞と連体修飾複文」「情報構造の諸要素」」『語学研究所論集』27: ページ未定.

Schönig, Claus (1999) The Internal Division of Modern Turkic and Its Historical Implications. *Acta Orientalia Academiae Scientiarum Hungaricae*. 52: 63–95.

菅原純 (2012)「試練に立つことば 「現代ウイグル語の歴史と現在」」中国ムスリム研究会編『中国のムスリムを知るための60章』76-80. 東京: 明石書店.