Universal Dependencies Treebank for Tatar

Incorporating Intra-Word Code-Switching Information

Chihiro Taguchi[1,2] Sei Iwata[1] Taro Watanabe[1]

[1]Nara Insitute of Science and Technology
{taguchi.chihiro.td0, iwata.sei.is6, taro}@is.naist.jp

[2]University of Edinburgh

Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)
June 20, 2022

Abstract

Aim

Build language resources for code-switching minority languages with the case study in Tatar

- Universal Dependencies Tatar NMCTT Treebank
- Introducing annotation for intra-word code-switching
- Evaluating the usefulness of the intra-word code-switching annotation in UD

Table of Contents

Motivation

Why Tatar?

- NLP has centered around only a handful of “big” languages out of approx. 6,000 languages in the world
- Minority languages have been marginalized because they are low-resource
- Minority languages often exhibit code-switching (CS)
- Tatar: low-resource language with intra-word CS
- A case study on such a language enriches and encourages corpus building for marginalized languages

Backgrounds: Tatar

- Spoken in Tatarstan, Russia by 5–7 mil. people
- Tatar $<$ Kipchak (Northwestern) $<$ Turkic
- SOV, head-final
- asymmetric bilingualism[1]: CS with Russian, language shift
- Cyrillic ortography (+ Latin)

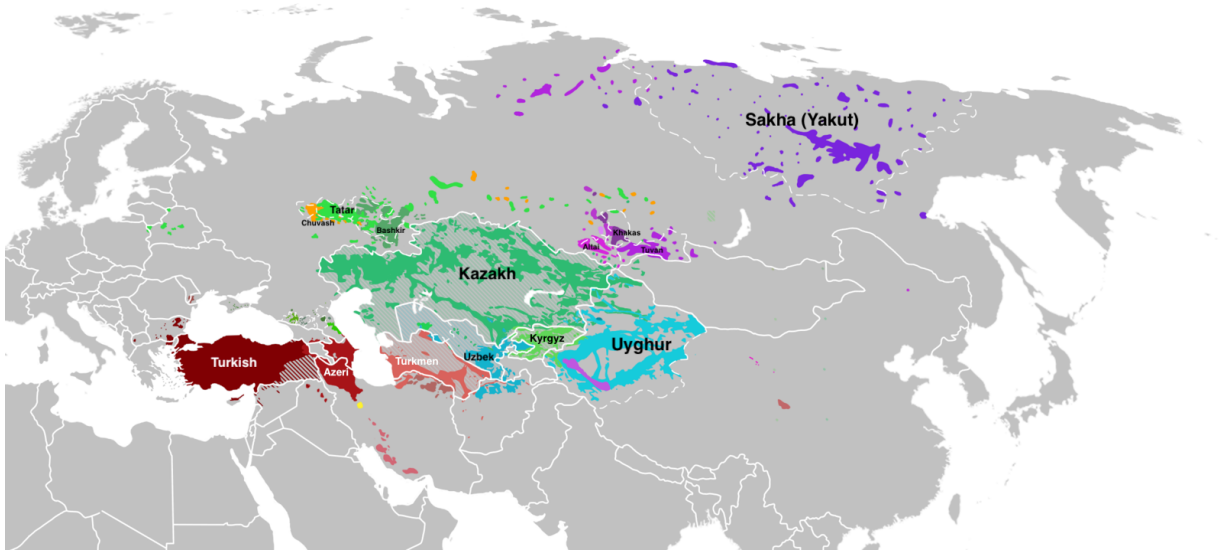

Figure: Distribution of the Turkic languages. Light green in the northwest indicates the Tatar-speaking area. Map modified from the original one by Wikimedia Commons.

[1] Safina (2020) “Bilingualism in the Republic of Tatarstan: language policy and attitudes towards Tatar language education”

Backgrounds: Code-Switching (CS)

Code-Switching (CS)[2]

A sociolinguistic phenomenon where a speaker switches a language to another, typically from a local, minority language variety to a prestigious, majority one

Intra-word CS: CS inside a word, often found in morphologically-rich languages

- Tagalog – English (Philippines)
  - *magtextan* ('to text each other')
- Huichol – Spanish (Mexico)
  - *pecansadoxi* ('you are tired') [3]
- **Tatar – Russian** (Russia)
  - прививканы ('vaccination'- ACC[4])
    privivka-nı (transliteration)

[2] Alvanoudi (2017) "Language contact, borrowing and code switching: a case study of Australian Greek"

[3] Mager, Çetinoğlu, and Kann (2019) "Subword-Level Language Identification for Intra-Word Code-Switching"

[4] ACC: accusative.

Code-Switching vs. Loanwords

How is code-switching different from loanwords (in Tatar)?

- Switch from the Tatar phonology to the Russian phonology
- "Contact-induced speech behavior that occurs extensively in the talk of bilinguals"[5]
- Language-specific merit: transliteration from Cyrillic to Latin[6]

[5] Alvanoudi (2017) "Language contact, borrowing and code switching: a case study of Australian Greek"

[6] Taguchi, Sakai, and Watanabe (2021) "Transliteration for Low-Resource Code-Switching Texts: Building an Automatic Cyrillic-to-Latin Converter for Tatar"

Backgrounds: Universal Dependencies (UD)

Universal Dependencies (UD)[7]

A project for building multilingual annotated treebanks in the unified CoNLL-U format

- Covering more than 200 treebanks of 130 languages
- No treebanks for Tatar before this study

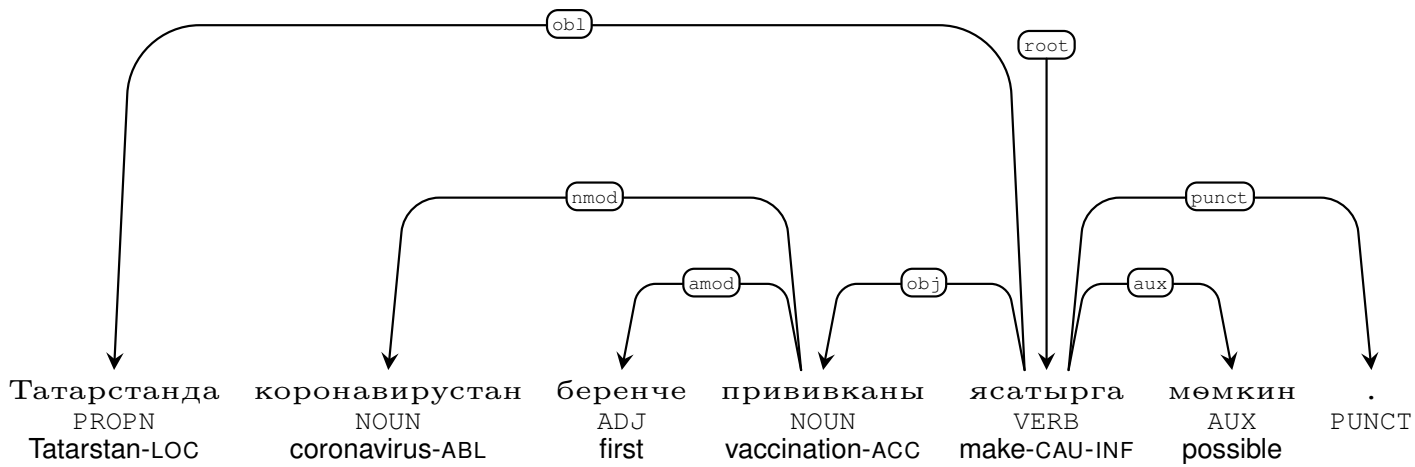

Figure: Example of a visualized CoNLL-U tree

[7] Nivre et al. (2020) "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection"

Backgrounds: Universal Dependencies (UD)

Rich grammatical information

- `FORM`: Word form (token)
- `LEMMA`: Lemma
- `UPOS`: Part-of-speech
- `FEAT`: Morphological information
- `HEAD`: Syntactic head
- `DEPREL`: Dependency relation
- `MISC`: Miscellaneous information (+ language tag)

```
# sent_id = 5853693_15
# link = https://tatar-inform.tatar/news/bu-atnada-tatarstanda-hava-temperaturasy-23s-ylylyktan-2-graduska-kadar-tosacak-5853693
# genre = ecology
# text = Яңгыр яву һәм яшен булу ихтималы саклана.
1	Яңгыр	яңгыр	NOUN	_	Case=Nom|Number=Sing	2	nsubj	_	LangID=TT
2	яву	яу	VERB	_	Case=Nom|Number=Sing|VerbForm=Vnoun	6	nmod	_	LangID=TT
3	һәм	һәм	CCONJ	_	_	5	cc	_	LangID=TT
4	яшен	яшен	NOUN	_	Case=Nom|Number=Sing	5	nsubj	_	LangID=TT
5	булу	бул	VERB	_	Case=Nom|Number=Sing|VerbForm=Vnoun	2	conj	_	LangID=TT
6	ихтималы	ихтимал	NOUN	_	Case=Nom|Number=Sing|Person[psor]=3	7	nsubj	_	LangID=TT
7	саклана	сакла	VERB	_	Person=3|Tense=Pres|VerbForm=Fin|Voice=Mid	0	root	_	LangID=TT|SpaceAfter=No
8	.	.	PUNCT	_	_	7	punct	_	LangID=OTHER
```

Figure: Example of UD annotation in Tatar.

Related Work: CS and Language Resources

- First language resources for CS:
  - Lyu et al. 2010: Mandarin–English
  - Li, Yu, and Fung 2012: Mandarin–English
- CS language resources in UD:
  - UD Turkish-German SAGT (Çetinoğlu 2016) (intra-word CS)
  - UD Hindi-English HIENCS (Bhat et al. 2018) (word-level CS)
- Language resources of Tatar:
  - The Corpus of Written Tatar (Saykhunov et al. 2021): 356 million tokens
  - The Tatar National Corpus (Suleimanov et al. 2013): 180 million tokens
  - No manually annotated treebank of Tatar

Related Work: Available Turkic Treebanks

- Turkish-German SAGT (Çetinoğlu 2016): 37K tokens
- Kazakh KTB (Makazhanov et al. 2015): 10K tokens
- Old Turkish Tonqq (Derin and Harada 2021): 221 tokens
- Uyghur UDT (Eli et al. 2016): 40k tokens
- Yakut YKTDT (Merzhevich and Gerardi 2021): 271 tokens
- + 9 Turkish treebanks (733K tokens in total)

Turkic languages in UD are overall low-resourced except Turkish.

Related Work: Language Processing for CS

- Earliest work: Joshi 1982 (Marathi–English)
- CS point prediction: Solorio and Liu 2008
- POS tagging: Anastasopoulos et al. 2018 (Griko–Italian)
- Intra-word CS:
  - Mager, Çetinoğlu, and Kann 2019 (Wixarika–Spanish) with SegRNN[8]
  - Sabty et al. 2021 (Arabic–English)
  - Taguchi, Sakai, and Watanabe 2021 (Tatar–Russian) with BPE[9]

[8] Lu et al. (2016) "Segmental Recurrent Neural Networks for End-to-end Speech Recognition"

[9] Sennrich, Haddow, and Birch (2016) "Neural Machine Translation of Rare Words with Subword Units"

Overview of NMCTT

UD Tatar-NMCTT: The first UD treebank of Tatar

- First release: November 2021
- Latest version: UD v2.10 (May 2022)
- Original text: Online news media (Tatar-Inform)
- Annotator: Chihiro Taguchi
- Size: 1,458 tokens
- Repository: `https://github.com/UniversalDependencies/UD_Tatar-NMCTT`

NMCTT: tokenization

Tokenization by splitting at spaces and punctuation

- Locative adjectivizer: different token?
  - SAGT: different token (cf. 1)
  - NMCTT: same token (cf. 2)

ID	FORM	LEMMA	UPOS
1-2	Berlin'deki	_	_
1	Berlin'de	Berlin	PROPN
2	ki	ki	ADP

Table: Tokenization and tags in SAGT.

ID	FORM	LEMMA	UPOS
1	Берлиндагы	Берлин	PROPN

Table: Tokenization and tags in NMCTT (transliterated).

NMCTT: Part-of-Speech

Disproportional distribution of CS w.r.t. POS tags

Class	UPOS	Total	Russian	Mixed
Open	NOUN	413	21	62
	PROPN	79	34	8
	VERB	169	0	1
	ADJ	117	8	0
Closed	AUX	18	0	0
	DET	9	0	0
	ADV	40	0	0
	SCONJ	8	0	0
	ADP	35	0	0
	CCONJ	26	0	0
	PRON	26	0	0
	NUM	12	0	0
Other	PUNCT	167	0	0

Table: The distribution of UPOS tags in the treebank with respect to language code. The first column specifies whether the UPOS tag is an open class or a closed class.

NMCTT: Morphology

Policy

Different morphemes with different semantics have different morphological features.

Morphological annotation of Turkic treebanks has been inconsistent

- Example: Converbs (Haspelmath 1995)
  - Annotation in NMCTT in Table
  - See Table 5 of the paper for the variation of converb annotation among Turkic treebanks

Converb Form	Semantics	Features
-п	Associated event	`VerbForm=Conv`
-а/ә	Simultaneous event	`Aspect=Prog|VerbForm=Conv`
-гач/гәч	Completed event	`Aspect=Perf|VerbForm=Conv`
-ганчы/гәнче	Unaccomplished event	`Aspect=Imp|VerbForm=Conv`

Table: Tatar converbs and their morphological features in NMCTT

NMCTT: Syntactic Dependency

NMCTT follows other UD treebanks' policies for syntactic dependency.
Language-specific treatments:

- Light Verb Construction: `compound:lvc`
- Verbs grammaticalized into auxiliaries: `aux`

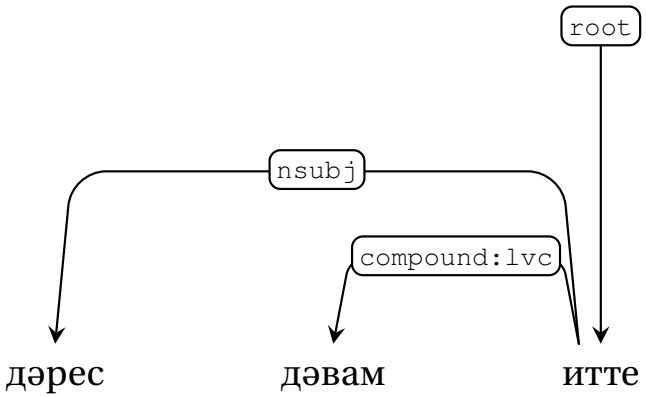

Figure: An example of `compound:lvc`.

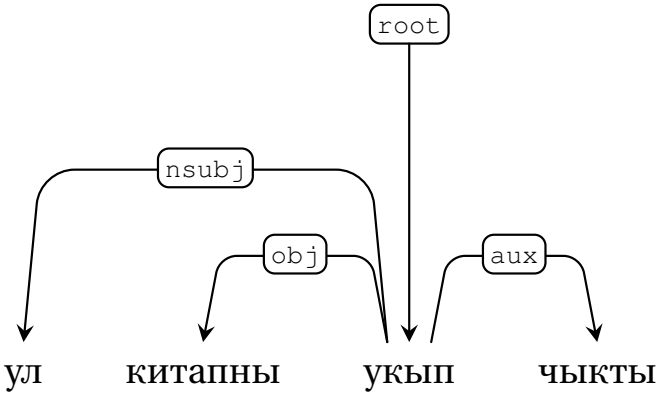

Figure: An example of a grammaticalized auxiliary `aux`.

NMCTT: Language Tags `LangID=`

Tatar CS with Russian can be within a word (intra-word CS):

- Гыйбәтдиновка (Russian in red)
  Ğıybätdin-ov-qa
  'To Gibatdinov'

Annotation of intra-word CS proposed in the treebank:

- `CSPoint=`Гыйбәтдин§ов§ка`|LangID=MIXED[TT§RU§TT]`
- `CSPoint=` indicates the point(s) of language switch
- `LangID=` is tagged as `MIXED`, and the details are specified in brackets with ISO language codes

Experiment

Issue

Does the intra-word CS annotation in UD contribute to correct prediction of CS segmentation and language identification?

Task Description

Predict a pair of language tag and span tag at the character level utilizing the corresponding POS tag

Dataset:

- UD Tatar-NMCTT (1,119 tokens)
- UD Turkish-German-SAGT (26,929 tokens) (Çetinoğlu 2016).

Model

Conditional Random Fields (CRFs)[10]: loose description

- Probabilistic graphical model
- Able to consider context
- Suitable for labeling sequential data where each data point depends on its previous data

Intuitions:

- CS depends on context
- CS depends on part-of-speech; open-class words (e.g., nouns, verbs etc.) are more likely to code-switch than closed-class words (e.g., pronouns, auxiliaries)

[10] Lafferty, McCallum, and Pereira (2001) "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data"

Model

Conditional Random Fields (CRFs)

To predict correct labels $\boldsymbol{y}$ (language tag and span tag) given a sequence of input $\boldsymbol{x}$ (character-level in this task), the CRF model is defined as

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\right\}$$

where:

- $Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\right\}$: a normalization function to ensure the sum of p is 1;
- λ_k: a parameter vector; and
- $\{f_k\}_{k=1}^{K}$: a set of feature functions.

Model

- K: the total number of features after combining transition features $f_{ij}(y, y', x) = \mathbf{1}_{\{y=i\}}\mathbf{1}_{\{y'=j\}}$ for each transition (i, j) and observation features $f_{io}(y, y', x) = \mathbf{1}_{\{y=i\}}\mathbf{1}_{\{x=o\}}$ for each state-observation pair (i, o).

Namely,

$$\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)$$

$$= \sum_t \left\{ \sum_{i,j} \lambda_{ij} f_{ij}(y_t, y_{t-1}, \boldsymbol{x}_t) + \sum_{i,o} \lambda_{io} f_{io}(y_t, y_{t-1}, \boldsymbol{x}_t) \right\}.$$

Feature table

We used features that might affect the model. In particular, in order to verify the usefulness of UD's grammatical information in the prediction task, the POS feature is added in the feature extraction function.

Feature	Example	Value
Character	`"M"`?	1
Character +1	`"a"`?	1
Character +2	`"r"`?	1
Character -1	`False`?	1
Character -2	`False`?	1
Word-initial?	`True`?	1
Word-final?	`False`?	1
Word in titlecase?	`True`	1
Character in uppercase?	`True`	1
Punctuation?	`False`	0
Number?	`False`	0
Word length	`4`	1
POS	`"PROPN"`	1
Word	`"Mars"`	1

Table: An example of a feature table for the character `"M"` in `"Mars"`. Features in red are omitted in ablation studies.

Results

Highest scores in the **Default** models

- [-POS]: without the POS feature
- [-word]: without the Word feature
- [-POS, -word]: without both the POS and Word features

Features	Precision	Recall	F1
Default	**90.9**	**90.0**	**88.9**
[-POS]	87.3	86.5	84.3
[-word]	86.4	86.5	84.9
[-POS, -word]	86.7	87.0	85.7

Table: Ablation study of features on NMCTT. Scores are calculated at the character level.

Features	Precision	Recall	F1
Default	**95.9**	**96.1**	**95.9**
[-POS]	95.9	95.8	95.6
[-word]	94.6	94.7	94.6
[-POS, -word]	93.7	93.9	93.8

Table: Ablation study of features on SAGT. Scores are calculated at the character level.

Results

- **Default** (w/ `POS`, `word`) performs the best both in NMCTT and SAGT
- Excluding `POS` information worsens the performance in both NMCTT and SAGT
- The results imply that `POS` is meaningful in CS segmentation and language identification

Concluding Remarks

This series of studies has shown:

- NMCTT: the first UD treebank for Tatar
- the new way to annotate intra-word CS
- the usefulness of the CS annotation in applied tasks

For future work:

- Enlarging the corpus: NMCTT is still too small to be used for wider applications (at least ~10k tokens)
- Encouraging further corpus building for low-resource languages with CS

Acknowledgments

- This study was a part of Creative and International Competitiveness Project (CICP) in 2021 and funded by NAIST.
- The right to exploit the news text data used in the NMCTT treebank is granted by Tatar-Inform.

This research would have never been as it is now without these supports.

Bibliographical References I

Alvanoudi, Angeliki (2017). “Language contact, borrowing and code switching: a case study of Australian Greek”. In: pp. 1–42.

Anastasopoulos, Antonios et al. (2018). “Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2529–2539. URL: https://aclanthology.org/C18-1214.

Bhat, Irshad et al. (2018). “Universal Dependency Parsing for Hindi-English Code-Switching”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 987–998. DOI: 10.18653/v1/N18-1090. URL: https://aclanthology.org/N18-1090.

Çetinoğlu, Özlem (2016). “A Turkish-German Code-Switching Corpus”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4215–4220. URL: https://aclanthology.org/L16-1667.

Derin, Mehmet Oguz and Takahiro Harada (2021). “Universal Dependencies for Old Turkish”. In: *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 129–141. URL: https://aclanthology.org/2021.udw-1.11.

Eli, Marhaba et al. (2016). “Universal dependencies for Uyghur”. In: *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 44–50. URL: https://aclanthology.org/W16-5206.

Haspelmath, Martin (1995). “The converb as a cross-linguistically valid category”. In: *Converbs in Cross-linguistic Perspective: Structure and Meaning of Adverbial Verb Forms — Adverbial Pariticples, Gerunds —*. Ed. by Martin Haspelmath and Ekkehard König, pp. 1–55.

Bibliographical References II

Joshi, Aravind K. (1982). “Processing of Sentences With Intra-Sentential Code-Switching”. In: *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*. URL: https://aclanthology.org/C82-1023.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1558607781.

Li, Ying, Yue Yu, and Pascale Fung (2012). “A Mandarin-English Code-Switching Corpus”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2515–2519. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/964_Paper.pdf.

Lu, Liang et al. (2016). “Segmental Recurrent Neural Networks for End-to-end Speech Recognition”. In: *CoRR* abs/1603.00223. arXiv: 1603.00223. URL: http://arxiv.org/abs/1603.00223.

Lyu, Dau-Cheng et al. (2010). “SEAME: a Mandarin-English code-switching speech corpus in South-East Asia”. In: *INTERSPEECH*.

Mager, Manuel, Özlem Çetinoğlu, and Katharina Kann (2019). “Subword-Level Language Identification for Intra-Word Code-Switching”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2005–2011. DOI: 10.18653/v1/N19-1201. URL: https://aclanthology.org/N19-1201.

Makazhanov, Aibek et al. (2015). “Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report”. In: *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pp. 338–350.

Merzhevich, Tatiana and Fabrício Ferraz Gerardi (2021). *UD Yakut YKTDT*. https://github.com/UniversalDependencies/UD_Yakut-YKTDT.

Bibliographical References III


Nivre, Joakim et al. (2020). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497.


Sabty, Caroline et al. (2021). "Language Identification of Intra-Word Code-Switching for Arabic–English". In: *Array* 12, p. 100104. ISSN: 2590-0056. DOI: https://doi.org/10.1016/j.array.2021.100104. URL: https://www.sciencedirect.com/science/article/pii/S2590005621000473.


Safina, Kamila (2020). "Bilingualism in the Republic of Tatarstan: language policy and attitudes towards Tatar language education". In.


Saykhunov, M. R. et al. (2021). *Corpus of Written Tatar*. URL: https://www.corpus.tatar/en.


Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "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)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.


Solorio, Thamar and Yang Liu (2008). "Learning to Predict Code-Switching Points". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 973–981. URL: https://aclanthology.org/D08-1102.


Suleimanov, Dz. et al. (2013). "National Corpus of the Tatar Language "Tugan Tel": grammatical annotation and implementation". In: *Procedia - Social and Behavioral Sciences* 95, pp. 68–74.

Bibliographical References IV

Taguchi, Chihiro, Yusuke Sakai, and Taro Watanabe (2021). “Transliteration for Low-Resource Code-Switching Texts: Building an Automatic Cyrillic-to-Latin Converter for Tatar”. In: *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Online: Association for Computational Linguistics, pp. 133–140. DOI: 10.18653/v1/2021.calcs-1.18. URL: https://aclanthology.org/2021.calcs-1.18.