Contextualizing the Present and Future of Old Turkish Annotation within Universal Dependencies for Turkic Languages

An Amalgamation for Dialogue

mehmetoguzderin 20230908

Agenda

Sections

- Universal Dependencies for Old Turkish
- Principled Morphosyntactic Boundaries for Turkic

Universal Dependencies for Old Turkish

Agenda
Overview

- Publication Universal Dependencies for Old Turkish
- Future of Universal Dependencies for Old Turkish

Universal Dependencies for Old Turkish

Abstract

We introduce the first treebank for Old Turkic script Old Turkish texts, consisting of 23 sentences from Orkhon corpus and transliterated texts such as poems, annotated according to the Universal Dependencies (UD) guidelines with universal part-of-speech tags and syntactic dependencies. Then, we propose a text processing pipeline for the script that makes the texts easier to encode, input and tokenize. Finally, we present our approach to tokenization and annotation from a crosslingual perspective by inspecting linguistic constructions compared to other languages.

Introduction

Instances like splits of adpositions from words with consistent colon separator require UD treebanking of Old Turkish language to take an alternative approach compared to existing Turkic language treebanks. Therefore, we adopt the recent short-unit word perspective in other languages to help them cross-linguistically annotate better. But Old Turkic script treebanking brings further challenges like normalizing sentences and having consistent sentence segmentation.

	Unround		Round	
	Back	Front	Back	Front
Open	𐰀 a	𐰀 e	𐰆 o	𐰇 u
Closed	𐰃 w	𐰃 i	𐰆 o	𐰇 u

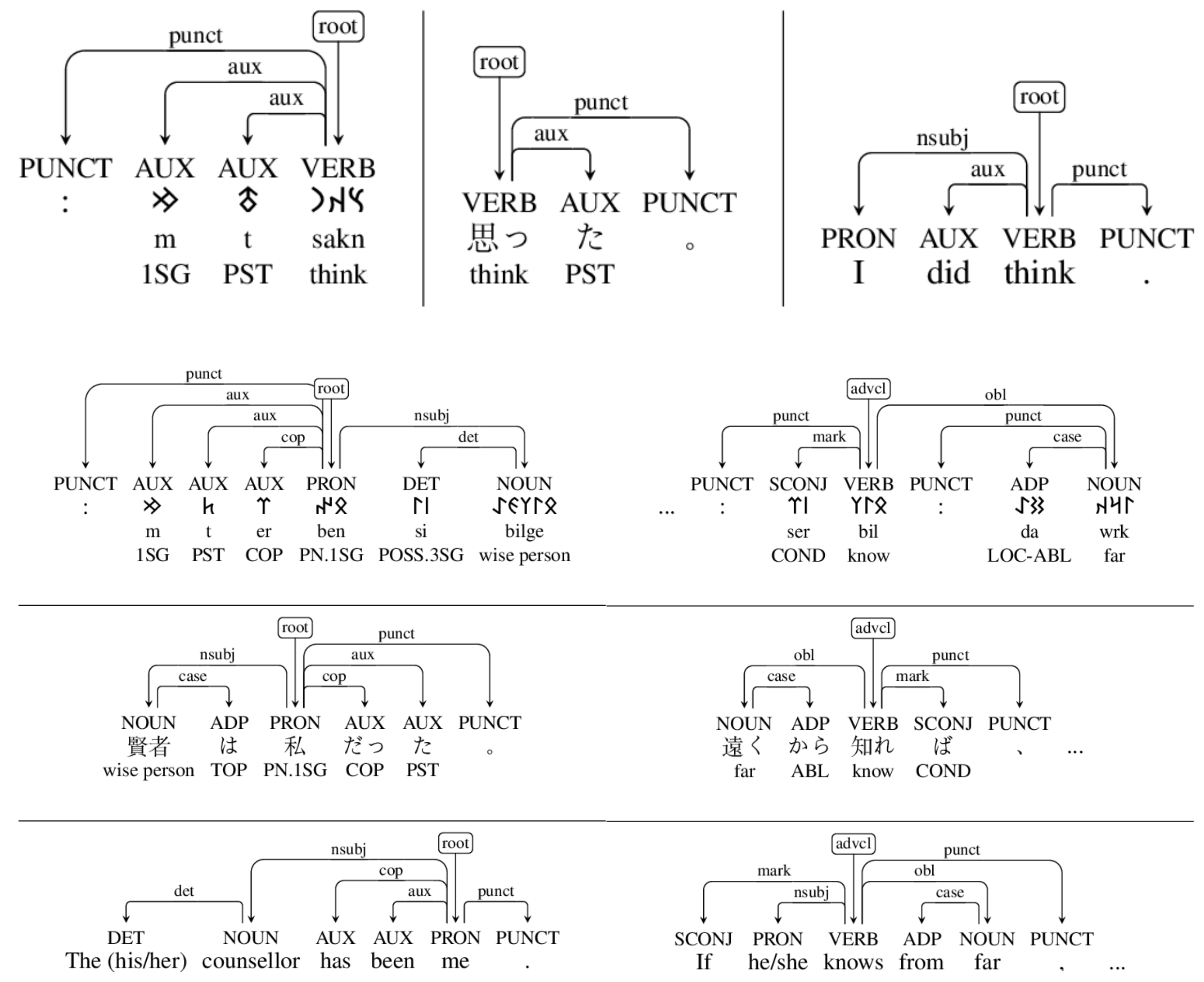
Occlusive				Fricative				Nasal		Vibrant		Approximant	
Voiceless		Voiced		Voiceless		Voiced		Voiced		Voiced		Voiced	
Back	Front	Back	Front	Back	Front	Back	Front	Back	Front	Back	Front	Back	Front
𐰴 ak	𐰚 ek	𐰍 ag	𐰏 eg					𐰭 aq	𐰶 eq				
𐰲 ac	𐰲 ec			𐱁 ax	𐱂 ex			𐰪 aj	𐰪 ej			𐰖 ay	𐰘 ey
𐱃 at	𐱅 et	𐰑 ad	𐰓 ed	𐰽 as	𐰾 es	𐰔 az	𐰔 ez	𐰣 an	𐰤 en	𐰺 ar	𐰼 er	𐰞 al	𐰠 el
𐰯 ap	𐰯 ep	𐰉 ab	𐰋 eb			𐰉 av	𐰋 ev	𐰢 am	𐰢 em				

We find that **word-class-definition-oriented tokenization** can **consistently annotate Old Turkish language** with better **cross-linguistic syntax properties.** Also, we build **tools** to validate that a subset of the **Unicode** for **Old Turkic script** can store texts in a **unified representation.**

ACL Anthology

Keyman

Conclusion



Our tokenization scheme results in a workflow that respects the utility of words and maps non-computational research of Old Turkish language and Old Turkic script in a generalizing manner. In addition, this approach produces constructions that compare better across languages and better surface the known typological similarities between Turkic languages and others. Besides opening up a prototype scale data of our paper's approach, we contributed the Keyman implementation of our transliteration scheme to the centralized Keyman repository which requires less trust towards our distributions and produces more modular output for the benefit of study in this specific context. Unfortunately, our data is too small to train a pipeline. Still, this work is an essential step towards the long line of tasks towards better understanding and applying computational methods to the Old Turkish corpora.

Mehmet Oguz Derin - Takahiro Harada

Future of Universal Dependencies for Old Turkish

A Summary

- After publication, future work was made priority for better value proposition:
  - Influential treebanks of typologically similar languages didn't annotate MWE
  - Better preservation of original character sequences
  - Ability to better bound the words despite phonetic peculiarities of Old Turkic script
- Repository currently on pause and only contains archetypical constructed sentences
- Hope to discuss whether we can have better alignment across Turkic before improving treebank with the aforementioned qualities

Universal Dependencies for Turkic Languages: A Principled Approach for Syntactic Words

Defining Morphosyntactic Boundaries to Maximize Linguistic Consistency and Computational Efficiency

mehmetoguzderin 20230622

Agenda
Overview

- Problem
- Proposal
- Impact
- Examples
- Conclusion
- Discussion

Problem

Challenges & Ambiguities in Current Annotation Methodologies

- Existing inconsistencies in defining syntactic word boundaries
- Interpretational ambiguities in Universal Dependencies (UD) guidelines
- Limitations of current UD guidelines in representing diverse morphosyntactic phenomena

Proposal

Principled Approach: Standardizing Syntactic Words

- Principles underpinning our approach: closed class promotion, morphosyntactic tagging, preserving word boundaries, and aligning with typologically similar languages
- Proper UPOS tag coverage by mapping bound particles into words without sacrificing morphology
- Preserving fully conjugated copula er by attestations in Old and Middle periods of the languages
- Maintaining official word boundaries through Multi-word Expressions

Impact

Advantages: Precision, Consistency, and Computational Efficiency

- Enhanced cross-linguistic comparability and deeper tree structures for robust graph comparisons
- Improved alignment with other languages through well-defined distinction between morphology and syntax
- Consistent annotations across temporal and spatial contexts, allowing for the respect of linguistic evolution

Examples

Case Studies: Exemplifying Our Approach in Turkish, Kazakh, and Uzbek

- Application of our principles in Turkish to show establishment of closed classes for inherent syntax
  - Adpositions: evde yok => [MWE: ev + de] + [MWE: yok]
  - Coordinating: öğrenip geldi => [MWE: öğren + ip] + [MWE: gel + di]
- Illustrating how morphology still exists despite promoting many particles into syntactic words
  - Forms open class: -ki gibi (ADJ), -miş gibi (NOUN), -r gibi (VERB)
  - Pluralizers as morphology: -ler (sizler), -z (göz)
- Demonstrating scenarios like copula preservation and the handling of nominal + conditional suffixes
  - Copula: sevmişti => sev + miş + (i- +) di
  - Subordinating: yoksa => yok + (i- +) + se

Conclusion

Looking Ahead: A New Standard in Universal Dependencies

- Reduction in interpretational ambiguity
- Improved linguistic comprehension
- A firm groundwork for future research in Turkic languages and typology
- Fostering dialogue
- Communicating challenging cases

1. This new approach addresses the morphosyntactic complexity of Turkic languages for more accurate treebank compilation and syntactic word definition alignment.

2. The expansion of Universal POS (UPOS) coverage using this approach decreases interpretational ambiguity in Universal Dependencies (UD) guidelines.

3. Rather than merging particles into words as morphological constituents, the proposed approach maps bound particles into syntactic words, broadening the UPOS tag range.

4. The approach benefits from the work of existing Turkic treebanks, building a comprehensive linguistic model that preserves the unique intricacies of these languages.

5. This method enhances cross-linguistic comparability, producing data that's more applicable for in-depth linguistic analysis and comparisons across language families.

6. The approach generates more complex tree structures for better graph comparisons, while maintaining a balance between preserving language-specific word boundaries and standardizing syntactic word annotations.

7. The clear demarcation between morphology and syntax in this approach aligns with principles found in other languages, improving downstream task analysis capabilities and the applicability of direct source tokens.

8. This approach ensures annotation consistency across temporal and spatial contexts, facilitating linguistic evolution tracking and assessments of different Turkic languages.

9. A reduced lexical volume for finite state processing boosts computational efficiency, while ensuring that spelling does not influence the definition of syntactic words.

10. This approach allows for consistent guideline interpretation without conflicts with existing annotations for other languages.

11. Elements are mapped to open classes (e.g., noun, adjective, verb) based on their function (i.e., nominalizers to nouns, adjectivizers to adjectives, and verbalizers to verbs), reducing ambiguity.

12. This approach results in an increased number of closed-class UPOS tags, while maintaining all open-class tags, allowing for better closed-class distinction in a cross-linguistic lens.

13. The approach underscores the preservation and observability of the copula in Turkic languages, including in instances where it may appear hidden or fused with other elements.

14. The proposed approach acknowledges and annotates hidden verbs in modern languages, which helps reduce ambiguity and provides clarity in syntactic word definitions.

15. By recognizing hidden copulas and maintaining morphological markers, the approach manages to provide a comprehensive picture of language structures, even in challenging cases.

16. Nominal and verbal elements receive consistent treatment under this approach, which recognizes them as separate entities even when they appear together, ensuring the correct identification and treatment of plural forms.

17. This principled approach honors the complexity of Turkic languages while maximizing computational efficiency and reducing interpretational ambiguity.

18. The approach offers the advantage of maintaining a robust corpus of annotated data that is relevant for both linguistic research and computational tasks.

19. With this approach, we are able to strike a balance between maintaining the integrity of linguistic structures and making data more accessible and comprehensible for computational linguistics tasks.

20. This approach represents a significant advancement in computational linguistics by respecting the complexities of Turkic languages while maximizing computational efficiency and reducing interpretational ambiguity.

References

A cursory list for space, please refer to UDW publication for a more comprehensive list.
Any achievement stands on the shoulders of giants.

- Derin, M.O. and Harada, T., 2021, December. Universal Dependencies for Old Turkish. In Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021) (pp. 129-141).

- Johanson, L. (2021) Turkic. Cambridge: Cambridge University Press (Cambridge Language Surveys). doi: 10.1017/9781139016704.

- Taguchi, C., Iwata, S. and Watanabe, T., 2022, June. Universal Dependencies Treebank for Tatar: Incorporating Intra-Word Code-Switching Information. In Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference (pp. 95-104).

- Robbeets, M. and Savelyev, A. eds., 2020. The Oxford guide to the Transeurasian languages. Oxford University Press.

- 大村舞, 若狭絢 and 浅原正幸, 2023. 国語研長単位に基づく日本語 Universal Dependencies. 自然言語処理, 30(1), pp.4-29.

Discussion

Open Floor

- Insights
- Thoughts
- Steps

Treatment of Turkic Word in Universal Dependencies to

Balance the **etymology** and **practicality**

Increase **interpretability** and **generality**

Service **software** and **human**

Improve downstream application

As the optimal middle ground to reduce ambiguity of co-existing or competing divergent definitions of same set of terms