Universal Dependencies for Old Turkish

Mehmet Oguz Derin
Morgenrot, Inc.
Turkey
mehmetoguzderin@mehmetoguzderin.com

Takahiro Harada
Morgenrot, Inc.
Advanced Micro Devices, Inc.
USA
takahiro.harada@amd.com

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 cross-lingual perspective by inspecting linguistic constructions compared to other languages.

1 Introduction

Old Turkish[1] (ISO 693-3[2]: otk) was a pluricentric[3] Turkic language with different dialects spoken across Eurasia between the 7th and 14th centuries CE[4], written with different scripts, including Old Turkic script, and its corpora consist of three groups (Ağca, 2021). The modern descendant languages of Old Turkish, a subset of the Turkic language family (Glottocode[5]: comm1245), have more than a hundred million speakers. Some of these languages are classified as endangered by UNESCO[6]. The corpora represent the first sizable record of Turkic languages. Thus, the language and corpora are essential for research into the Turkic language family as they provide clues about stages with scarce data (Savelyev and Robbeets, 2020).

Old Turkic (ISO 15924[7]: Orkh) was a script used to write Old Turkish between the 7th and 10th centuries that reflects characteristics of Turkic languages, including vowel harmony, the binary distinction of non-nasal consonants, letters that sound of the object they depict, and its inventory consists of texts on stelae, papers, and other items including seals, bowls (Erdal, 2004). Materials written with Old Turkic are the very first texts in Old Turkish corpora. The Old Turkic Unicode Range (10C00–10C4F) makes it possible to digitize a subset of the script in a standards-compatible way (The Unicode Consortium, 2021). Such feasibility of maintaining a digital Old Turkic corpus provides an opportunity to compare later Old Turkish corpora texts with prior ones using a unified encoding.

Despite the extensive growth of research and print literature around the Old Turkish language and the Old Turkic script in recent decades, the digitization efforts for Old Turkic script Old Turkish

[1]The language itself still goes by different names in research, we call it by the Old Turkish name to stay consistent with ISO, reserve Old Turkic name for the script, and avoid naming either language or script as Orkhon since it stands better as the name of the corpus. An apparent name clash could be the study of the precedent of the modern Turkish language. However, it almost ubiquitously goes by Old Anatolian Turkish name and not as Old Turkish.

[2]iso639-3.sil.org/code_tables/639/data?title=otk

[3]The rigorous work by Ağca (2021) verifies the observation by Erdal (2004), as "The differences within Old Turkic are by no means greater than, e.g., within Old Greek", that differences could fit in dialectology; hence we briefly call as pluricentric.

[4]For periodization, we adopt the convention by Ağca (2021) since it is a data-based study of texts. The recent work by Johanson (2021) starts the period from the fifth and sixth centuries CE, but we avoid including these centuries where the amount of tangible text is minuscule and extrapolating the grammar as found in Old Turkish corpora texts directly could be misleading.

[5]glottolog.org/resource/languoid/id/comm1245

[6]unesco.org/languages-atlas/, also see endangeredlanguages.com/lang/search/#/?q=Turkic

[7]unicode.org/iso15924/iso15924-codes.html

inscriptions to produce reusable, standards-compatible, open-access computational resources and data are scarce, and advanced tooling for NLP does not exist. Therefore, we have developed the first Universal Dependencies (UD) (de Marneffe et al., 2021) (Nivre et al., 2020) treebank with part-of-speech tags and syntactic annotation for Old Turkic script Old Turkish texts and its tooling to start the NLP applications' building process. We chose the UD scheme because it provides guidelines for consistent annotation of typologically different languages and has extensive adoption and active community, making it possible to validate annotations by specification, data, tools, and discussion. We also built tooling for the treebank to establish a workflow for further research. This small, manually annotated treebank and associated tooling is the first step towards a larger-scale, potentially automated analysis of Old Turkic script encoded Old Turkish texts for UD.

We organized the remainder of this paper as follows. First, in Section 2, we briefly summarize digital or printed related work. Then, in Section 3, we provide an overview of the Old Turkish language and the Old Turkic script. In Section 4, we describe the texts and tools used in building the treebank. In Section 5, we discuss issues with tokenization and sentence segmentation before presenting an approach that makes further annotation consistent, and then, in Section 6, we explain the annotation process for part-of-speech tagging and dependencies in a cross-lingual perspective. Finally, in Section 7, we conclude the paper and contemplate future work.

2 Related Work

Despite the lack of Old Turkish treebanking, there is an increasing body of academic work for treebanking of Turkic languages and studies of the Old Turkish language and the Old Turkic script. Inside the Universal Dependencies project, there are treebanks for Turkish, besides others, by Sulubacak et al. (2016) and Uyghur by Eli et al. (2016) with more than 10K tokens. Despite not living inside the Universal Dependencies project (still, some components of Universal Dependencies take part in the paper), the recent Turkish treebanking approach by Kayadelen et al. (2020) represents a landmark in the consistent annotation that presents guidelines akin to the Short-Unit Word perspective of Japanese treebanking in Universal Dependencies (Omura and Asahara, 2018) (which we use as a convention in our cross-lingual comparisons, besides EWT for English by (Silveira et al., 2014)). Another essential reference for Turkish NLP is the comprehensive work by Oflazer and Saraçlar (2018). Although we mention the Turkish NLP works due to their size compared to other existing languages in the Turkic language family, it is crucial to note that Old Turkish has critical differences from Turkish, not only phonetically but also grammatically. The recent encyclopedic work on the Orkhon corpus by Ercilasun (2016) makes extensive use of literature to provide a methodic reading of the script and the interpretation of the language. The recently published dictionary by Wilkens (2021) provides an essential contribution with its open-access model and focuses on the Old Uyghur corpus. A recent, comprehensive survey of Turkic languages by Johanson (2021) also adopts the open-access model and provides an essential resource for our work. On historical dictionaries that have compilation near Old Turkish period, the renditions made in the last decade on historical Karakhanid bilingual dictionary by Ercilasun and Akkoyunlu (2014), and later historical Old Uyghur by Yunusoğlu (2012), Khwarezmian by Kaçalin and Poppe (2017), and Cuman by Argunşah and Güner (2015) bilingual dictionaries remain as primary references. The grammars by Tekin (1965), Erdal (2004), and Eraslan, Kemal (2012) are comprehensive works with different scopes that help check grammar points. The grammar by Erdal (2004) is especially helpful as it is written in the English language (a feature that eases the correspondence-finding process further when used in tandem with the recent work, which puts the concepts found in the book into a cross-lingual perspective) and includes comparisons that assist with evaluation inside Universal Dependencies context, such as pronominal copula as found in Hebrew. The comparative grammar by Serebrennikov and Gadžieva (2011) provides a bridge between works inside the language family. The textbook treatises by Tekin and Ölmez (2014) and Ölmez (2017) cover a variety of topics in a comprehensively indexing way. The textbook by Ölmez (2017) also includes a word-by-word breakdown of sentences with further morphological analysis, which is the closest work that we can find to anything resembling tagging of sentences. However, by its textbook nature, it does not provide full coverage. For the delimitation of

the corpora, the works of Yıldırım (2017), Aydın (2017), Aydın (2018), and Aydın (2019) provide a comprehensive account of Old Turkic script texts, whereas the recent work of Ağca (2021) provides a detailed analysis of Old Turkish corpora's boundaries with special attention on Old Uyghur corpus. For the digitalization of Old Turkic texts, the essential precedents are the often-cited web portal bitig.kz[8] by Abuseitova and Bukhatuly (2005), which does not use the Unicode Old Turkic block to encode the text due to lack of it at the time of its establishment and does not cover the recently found texts, and the atalarmirasi.org[9] by International Turkic Academy (2017) which provides a listing with more brief coverage of their content. An important Turkic language family digitalization work is Chagatai 2.0[10] (Amat et al., 2018), which includes per-sentence annotations with glossing, but it does not cover Old Turkish period.

3 Background

To provide a background for the rest of the paper, in this section, we provide a very brief overview of key features of the Old Turkish language and Old Turkic script.

3.1 Old Turkish Language

As a historical language, Old Turkish belongs to the Turkic family of languages. The three groups of Old Turkish corpora define the language's three main dialects: Orkhon, Old Uyghur, and Karakhanid. Since it represents some of the earliest attestations inside the Turkic language family and to the extent of material the corpora covers, it bears an essential value for studying the languages that are direct descendants of it and the ones branched earlier (such as Chuvash or Sakha), and stands as a bridge for the under-resourced, endangered Turkic languages which preserve archaic features like anticipating numerals (Zhong, 2019). Following are some general characteristics of the Old Turkish language and Turkic languages, which are also present in languages like Japanese and Korean (Han et al., 2020):

1. Dominant word order is subject-object-verb, but rich morphology allows for out-of-order constructions, especially for translated material.
2. Preference for postpositions (suffixing) and verbal endings.
3. Head-final language in which the embedded clause precedes the main clause.

Besides these, the following are some distinguishing features of Old Turkish that separate it within the Turkic language family:

1. Preservation of the /d/ and /ɲ/[11] phonemes inside and at the end of the words.
2. Use of the locative  “at, from”[12] also as an ablative.
3. Presence of er “to be” as a fully conjugated copula.

3.2 Old Turkic Script

As a phonetic script, Old Turkic consists of more than 40 characters, counting variants. Most of these characters represent a single phoneme, and except for five characters, they hint about the backness of vowels between consonants, while some denote a specific consonant cluster or a specific consonant

[8] bitig.kz

[9] atalarmirasi.org/en

[10] uyghur.ittc.ku.edu

[11] When we write phonetic values between forward slashes, we use International Phonetic Alphabet (IPA) per International Phonetic Association et al. (1999) to denote the value instead of the custom phonetic schemes of our reference work.

[12] We use this notation of specialized original-form transliteration “translation” in the rest of the paper. For original-form, the direction is right-to-left, so in this instance, values of original-form would translate to , , , and . For transliteration, the direction is left-to-right, so in this instance, values of transliteration would translate to te, ta, de, and da, corresponding to the order of values in the original-form. Number of readings is combination of all options at all positions.

with a specific vowel. Although there are five characters for spelling vowels, texts do not write open unrounded vowels explicitly unless they are at the end of the word, and some instances omit vowels in non-initial syllables between consonants, reflecting the tongue root harmony of Turkic languages. An essential phenomenon in Old Turkic script texts is the representation of 𐰽 n /n/ followed by 𐰍 g /g/ or 𐰴 k /k/ by single 𐰾 q /ŋ/ at instances, often found when words that end with an alveolar nasal consonant have the dative case. The punctuation is mostly a colon separating words or word groups, sometimes also found as a dot with single color or a colon or dot with colored marking. Whitespace and line breaks also do not bear a meaning most of the time. The Unicode block for Old Turkic script does not include all characters and variants, thus making it infeasible to do one-to-one digitalization of the Old Turkic script corpus. However, it does include enough characters to represent non-included characters phonetically in a way that would allow for direct representation of all characters through conversion when the block expands. The dominant writing direction is right-to-left, but the layout varies between texts, and the writing material varies between different surfaces like paper, stone, mirrors. We provide tables (see Table 1) for vowels and consonants of our digital rendition of the script and transliteration[13]. Following are some essential characteristics of the Old Turkic script:

1. Open rounded vowels are implicit and not written, except for when they are final.
2. Some consonants have synharmonic variants that govern the realization of vowels.
3. Punctuation is very minimal, and its usage is sparse.

4 Corpus

In this section, we present current texts and explain our Old Turkic script and transliterated text encoding. Additionally, we introduce tools that ease the development process and help validate string conformance to our guidelines.

4.1 Texts

Our current Old Turkish treebank consists of 23 manually annotated sentences through treebank-specific tools. Twenty-one of these sentences come from the first face of the first stele of Tuŋukuk inscriptions, a personal account of events the author witness and partake, themselves being Old Turkic script at the source itself but encoded through our pipeline's character mapping scheme. The remaining two of these sentences come from two recently found as syntactically-analogous by Kurnaz (2009) (which resolved the questions about the latter sentence's ambiguous use of particle) poetic sentences to represent a marginal exemplary case of transliteration from later text into Old Turkic script. Our treebank currently contains 341 tokens, with 14.826 tokens per sentence on average.

4.2 Tools

When working with the Old Turkic script block of Unicode directly, there are not many tools other than tools suited for general text or Unicode purposes, and this lack is even present for problems like missing characters in the block. To not give up on directly encoding using Old Turkic script Unicode block, and to take advantage of the fact that a subset of the range can represent all of the vowels and consonants found in Old Turkish corpora (excluding foreign words in non-Old Turkic script texts), we first define a normalization of digital Old Turkic texts and develop a tool to facilitate automatic normalization.

The normalization does two transformations: reduce characters with Orkhon and Yenisei variants to Orkhon only and break up syllabic characters into multiple characters. We base interpretation of syllabic characters on the work of Ercilasun (2016), which presents a consistent, regular approach. We base the second transformation on corresponding instances found in Old Turkic script themselves, and our approach is available in the source code where we store these transformation rules in a JSON file and apply them through a Python script. The normalization also disallows characters other than

[13] Our reference works do not share a common transliteration scheme, and we do not include them due to the space constraints, only presenting ours, which fits the criteria of being representable in lowercase, ASCII only setting.

	Unround		Round	
	Back	Front	Back	Front
Open	𐰀 a	𐰃 e	𐰺 o	𐰽 u
Closed	𐰆 w	𐰇 i	𐰺 o	𐰽 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	𐰵 ek	𐰓 ag	𐰔 eg					𐰾 aq	𐰿 eq				
𐰚 ac	𐰛 ec			𐰜 ax	𐰞 ex			𐰋 aj	𐰌 ej			𐰉 ay	𐰊 ey
𐰢 at	𐰣 et	𐰖 ad	𐰗 ed	𐰍 as	𐰏 es	𐰑 az	𐰒 ez	𐰠 an	𐰡 en	𐰤 ar	𐰥 er	𐰨 al	𐰩 el
𐰀 ap	𐰁 ep	𐰃 ab	𐰄 eb			𐰃 av	𐰄 ev	𐰸 am	𐰹 em				

Table 1: Vowels and consonants in Old Turkic script on left followed by our transliteration (for consonants, preceding vowels are not included in transliterations of text, these are for backness notation in a context-free setting) on right. Phonetic regions (of which we omit the row labels for consonants due to space constraints) are figurative, as we serve the table only to facilitate understanding of our transliteration scheme, and we repeat characters that can represent multiple sounds in respective cells, see mentioned works on Old Turkish grammar and Turkic languages in Related Work section for more details about the realization of the sounds.

allowed explicitly like colons to avoid characters introduced by various tools such as right-to-left or left-to-right or redundant line feed or return markers to be part of the treebank through a sanitizer. The normalization results in 33 letters, excluding specifically allowed punctuation. There are 4 vowels, which we spell twice to form digraphs representing closer versions of these vowels consistently if desired, 5 neutral consonants, 24 synharmonic consonants, representing 12 consonants with varying influence on the realization of vowels. We were able to reduce development overhead by adopting this normalization scheme. Thereby, content encoded with Old Turkic in this paper assumes the normalization applied, and they do not graphically cover characters that are either out of our normalization range or not even in the Unicode block.

For the generation of text identifiers and storage in places where only alphabetic lowercase ASCII (potentially with underscore or dash) is allowed, we developed a simple, rule-based bidirectional transliteration scheme that can represent all consonants and vowels alongside currently present punctuation. We also developed a reverse transliterator from ASCII to Old Turkic script that is more permissive to be compatible with manual transliterations that read better. To store the consonants in the lookup table, we precede the ASCII consonant with a closed vowel at the start, and if the consonant is front, we make the preceding vowel an always front vowel, and if the consonant is back, we make it a back vowel, if neither, we make it both front and back vowel, while single space always represents a backness neutralizer, to provide the users with the backness information. We use this transliteration table and usage of the Old Turkic letter that we do not preserve with the second transformation for choosing the backness of consonant as control key to deduce a Keyman[14] keyboard project that allows us to input Old Turkic script in the normalized form across many devices to develop the treebank and surrounding material.

As we lack an automated tokenization module, we store manually annotated token ranges and annotations in a JSON file and use a Python script to extract tokenized and annotated CONLL-U files from our CSV texts, which need a text column present. However, we do not restrict the presence of other columns that might use an extended range of Unicode blocks or define their features which could allow for easier identification of inscriptions with similar names through embedding GeoJSON of the location of the inscription in a column or other means. Furthermore, in the future, we intend to check for duplicate

[14] keyman.com/14

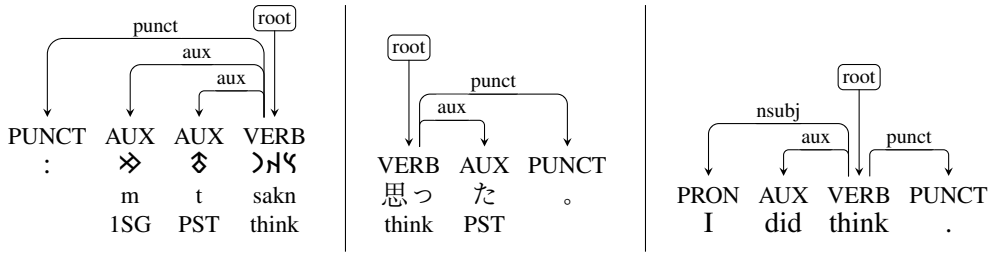

Figure 1: Tokenization and annotation of "I did think." sentence in Old Turkish with comparison to the annotation of figurative translations in Japanese and English. This example highlights the annotation of auxiliaries, especially the person marker, which derives from possessive markers.

sentences before storing them into the CONLL-U file as some inscriptions share verbatim sentences to avoid duplicate content inside the treebank, of which we currently have none.

We also have minor tools that depend on UDAPI products (Popel et al., 2017), such as denoting font automatically for exported TikZ graphs to ease authoring and experimental Anki[15] deck generator from features found in the treebank to demonstrate an edge-case utilization of the Universal Dependencies scheme. We distribute these tools in the not-to-release folder of our treebank.

5 Tokenization and Sentence Segmentation

Tokenization and sentence segmentation of Old Turkic script Old Turkish texts is a challenging task. The script lacks regular punctuation or whitespace for splitting into tokens and a marker for splitting into sentences. Another aspect that makes tokenization harder is letters representing multiple phonemes, but our text processing pipeline eliminates this issue in the resulting output.

5.1 Tokenization

Tokenization requires context-dependent decisions with Old Turkic script Old Turkish texts. Line breaks do not act as tokenizers, especially in limited-space texts, sometimes splitting even the base morpheme. Thus we ignore them in the process of tokenization. Character flipping due to synharmonism also does not act as a consistent tokenizer, and existing treebanking approaches for other Turkic languages also ignore it (for example, they always tokenize question particle despite its second vowel acting according to the harmony). Commonly found colon (or dot in some cases) does meaningful splits, sometimes into words and adpositions, into words, into phrases containing more than one word in other times (and not always consistently, e.g., separating an adjective and a proper noun in some cases while not in other cases). Thus, we always delimit tokens by this punctuation class before further splits. Generalization of such delimitations leads us to treat primarily inflectional morphemes such as possessive markers, case markers, auxiliaries, converbs, tense-aspect-modality-evidentiality (TAME) markers (including personality markers that derive from possessive markers) as tokens to preserve consistency across all of the Old Turkic script texts. We also tokenize particles outlined in the Universal Dependencies guidelines, such as the question particle and other similarly behaving particles in the Old Turkish language, including negation and intensifier particles. If bound morphemes act as nominalizers, resulting in a word that we treat as either noun or pronoun in the Universal Dependencies analysis, we do not split them into tokens and treat them as a single word. We do not treat verbalizer morphemes that impact voice or produce commonly lexicalized verbs while not violating previous steps as individual tokens. This direction results in an approach that provides a rich syntactical analysis and is similar to the recent Turkish treebanking work by Kayadelen et al. (2020) and some Universal Dependencies works like the Japanese language with Short Unit Words perspective (Omura and Asahara, 2018), also a recent highlight in cross-lingual perspective by de Marneffe et al. (2021), and the Shipibo-Konibo language (Vasquez et al., 2018). It is important to note that our guidelines only match with recent work by Kayadelen et al. (2020), and our treebank is the first to adopt this approach in Universal Dependencies Turkic family treebanks.

[15] apps.ankiweb.net

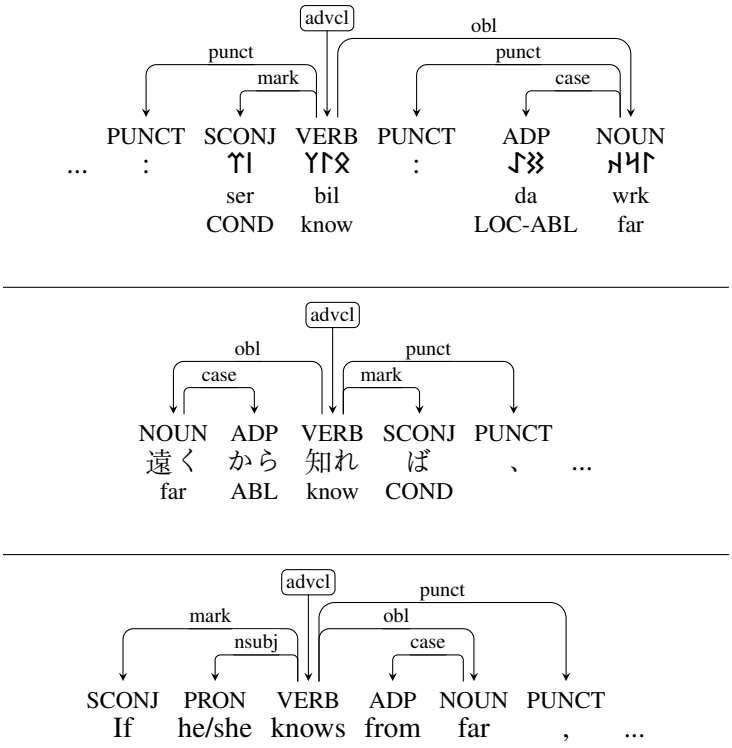

Figure 2: Tokenization and annotation of "If he/she knows from far.." clause in Old Turkish with comparison to annotation of figurative translations in Japanese and English. This example highlights the annotation of conditionals.

Our tokenization guidelines produce entries with characteristics that map into Universal Dependencies guidelines for both tags and dependencies, sitting at balance for cross-lingual perspective inside UD.

5.2 Sentence Segmentation

Sentence segmentation has to be done per the interpretation of the text due to the lack of any punctuation for this matter in Old Turkic script, and not even line breaks act as a regular means of sentence segmentation. After tokenization, we first work through detecting clauses, conjunctions, and finally roots of sentences to do sentence segmentation. We avoid producing parataxis constructions unless found in reported speech, favoring treatment as conjunctions if not fit for sentence split. Our sentence segmentation guidelines produce a set of delimiters that are the union of proposed sentence delimiters in the referenced work for the analyzed text.

6 Annotation

In this section, we go over our application of the Universal Dependencies for annotating parts of speech and syntactic dependencies in a cross-lingual perspective. Currently, our treebank does not have lemma or morphological annotations, and as such, we do not present any guidelines for them, and we only utilize miscellaneous for SpaceAfter=No annotation to all tokens since Old Turkic script texts, as far as we cover, do not contain spaces as a means of separating tokens.

6.1 Part-of-Speech Tagging

We adopt Universal Part-of-Speech (UPOS) tagset as the only convention in our treebank. After tokenization, challenging ones are the bound morphemes and pronominal copulas. We tag possessive (or person) markers as determiners (DET) if they are bound to a noun, but if they act as the only pronominal component of a phrase in a head-final position, we tag them as pronouns (PRON). We tag case markers as adpositions (ADP). We tag verbal endings or converbs that make adverbial clauses subordinating conjunction (SCONJ) or coordinating conjunction (CCONJ) depending on their function.

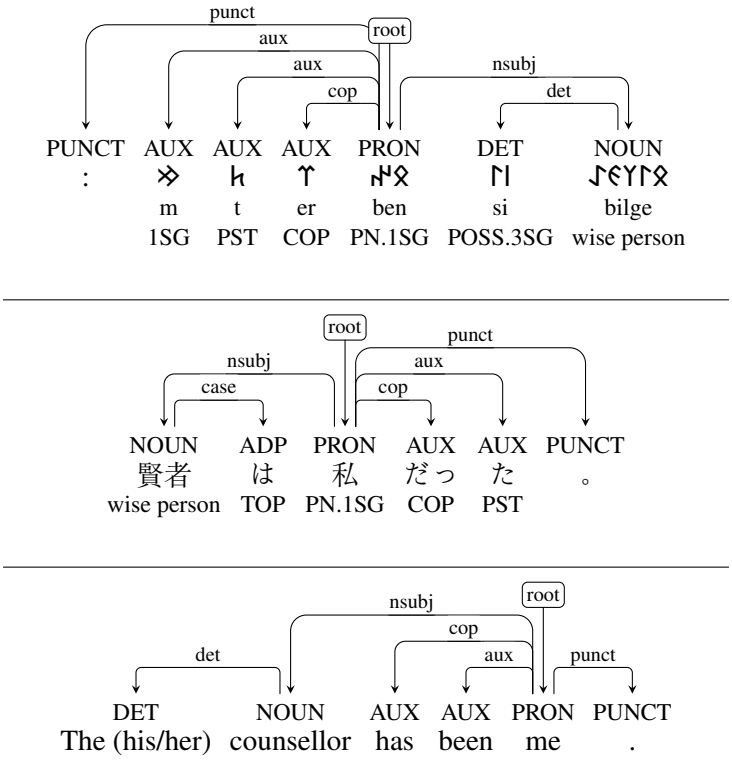

Figure 3: Tokenization and annotation of "The (his/her) counsellor has been me." sentence in Old Turkish with comparison to annotation of figurative translations in Japanese and English. This example highlights the annotation of possessive markers as determiners and auxiliaries.

We tag possessive marker derivative person markers, TAME markers, converbs that act as auxiliary along with a following auxiliary verb, the copula, and verbs that function as auxiliary as auxiliaries (AUX). Per tokenization, auxiliaries are not joined into a single word but instead kept separate units. We do not tag the verbs other than the fully-conjugated copula as an auxiliary (AUX) if they are the clause's predicate. We tag pronominal copulas found at the end of clauses as determiners (DET) per the recommendation of Universal Dependencies guidelines. We tag the regular punctuation as punctuations (PUNCT). Due to their usage in Old Turkish corpora, we treat the word which means "none, no, not, nothing", and the word which means "all, yes, is, everything" to be pronouns (PRON) as non-interrogative indefinite collective pronouns, a choice shared by the study of Lithuanian Karaim too (Robbeets and Savelyev, 2020), and also in the more recent study of Turkic languages (Johanson, 2021), or similar to other pronouns as determiners (DET) if they act as pronominal copulas. We always tag numbers as numbers (NUM). The rest of the tags map trivially to UPOS by the reference works we use. The treebank currently utilizes 15 tags, leaving out SYM and X. We expect to utilize SYM in the future due to texts containing pictograms. Our tagging approach produces a closed-class for all the tags denoted as such in the Universal Dependencies guidelines.

6.2 Syntactic Annotation

We use universal syntactic relations without subtypes or language-specific relations in our treebank. Out of 37 features, we explicitly avoid using the indirect object (iobj) relation as case markers, such as dative, always follow indirect objects, we use oblique (obl) in such cases, adopting the convention of some Uralic (Partanen and Rueter, 2019) and Japanese (Omura and Asahara, 2018) treebanks for the cross-lingual consistency of annotation. Direct objects (obj) also sometimes have case markers, especially genitive, but we do not treat them as oblique (obl) as they fulfill the core object function. Our treebank currently lacks instances of the clausal subject (csubj) and adnominal clauses (acl) dependencies due to the small data size, and their exact treatment requires special care with head-final characteristic Old Turkish in consideration, bearing challenges similar to Japanese, which we plan to address in future. Out

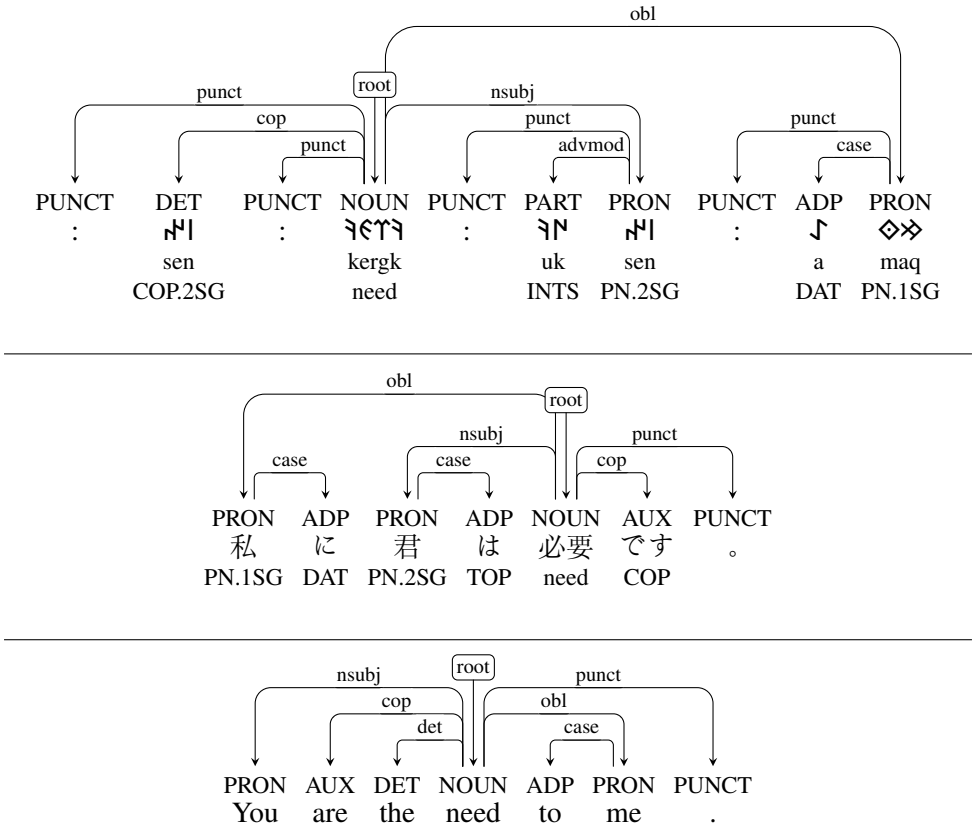

Figure 4: Tokenization and annotation of "I need you." sentence in Old Turkish with comparison to the annotation of figurative translations in Japanese and English. This example highlights out-of-order construction and pronominal copula.

of other currently unused relations, namely the vocative (vocative), the expletive (expl), the dislocated (dislocated), the classifier (clf), the fixed (fixed), the orphan (orphan), the goes with (goeswith), the reparandum (reparandum), the unspecified dependency (dep) dependencies, only expletive, classifier, and unspecified dependency are unlikely to be utilized in future. We annotate multi-word proper nouns using flat dependency. We annotate question and intensifier particles as adverbial modifiers (advmod). We annotate determiner (DET) tagged pronominal copulas with the copula (cop) relation. If not in proper clausal complement position, we treat reported speech and postposed, non-doubling, parenthetical elements (if we can not annotate as dislocated or appositional) as parataxis. As coordinating conjunction words can sometimes be present at the end of the sequence, we attach them to the element before as coordinating conjunction (cc), which provides a consistent annotation with analogous constructions like phrases formed with antonymy and parallelism markers. If a clausal complement has a null-subject, we annotate the dependency as a clausal complement (ccomp) rather than an open clausal complement (xcomp). We treat punctuations (punct) in line with guidelines while avoiding introducing non-projectivity. Treatment of punctuations might require improvement when treebank size grows as that combined with Universal Dependencies analysis can help further our understanding of punctuation in Old Turkic script texts. We annotate interjections as discourse. Some verbs like "to become, to have" can, depending on their usage, have either an object or a clausal complement attached to them, and we avoid annotating these as copula (cop), reserving the use of relation to the fully conjugated and pronominal copulas. Our tokenization and tagging choices lead to a consistent annotation of dependencies that allows for cross-lingual study.

7 Conclusion and Future Work

Using a cross-lingual perspective, in this paper, we presented the first application of Universal Dependencies to the Old Turkish language with Old Turkic script encoding. The characteristics of the Old Turkish language and the lack of tooling for both the language and the script pose significant challenges. However, as hinted by the extending body of traditional work for the language and recent work in NLP, we have argued through tokenization that it is crucial to define the word concept that creates analogies with other languages. Afterward, we have shown that we developed tooling and guidelines that allow for consistent tokenization, segmentation, tagging, and dependency annotation of the Old Turkish corpora through a finer-grained word definition. The treebank is currently, by its size, insufficient to cover all dependency types in Universal Dependencies or to train a pipeline (Straka et al., 2016) (Honnibal et al., 2020) (Qi et al., 2020), and the tooling does not live under a unified software package but as distinct modules, but it represents an important step towards the enlargement of both the encoded and the annotated text.

In the future, we plan to extend the data size, where we might prioritize using sentences matching recent work that study Old Turkish and its contemporaries in a comparative setting (Kasai, 2014) (Robbeets and Savelyev, 2020) (Lim, 2021) besides extending coverage over the oldest texts in the corpora. We also plan to add lemmas and features, which are crucial for automation due to their governance of how phrases act in a sentence and build additional tooling. As we provide tools for input, character normalization, transliteration, further work should encompass both improvements and extension towards tooling for more accessible span-based annotation of texts potentially through an extension of productive tools for Universal Dependencies (Tyers et al., 2017), automatic tokenization, sentence segmentation, lemmatization, part-of-speech tagging, dependency parsing, and coarser-grained normalization. Another critical area for future work is beginner-friendly guides and materials like dictionaries with references to cross-linguistic colexifications (Rzymski et al., 2020) for providing additional context to interpretations, encouraging people with a less technical background, and also for providing better visibility to the Universal Dependencies community, and if possible, creating avenues for bridging the disconnect in the study of Old Turkish between traditional (often restrained to the language of the work, less accessible towards non-speakers, and not always open-access or in a digitally accessible format) and computational works.

References

M. Abuseitova and B. Bukhatuly. 2005. bitig.kz Turk Bitig.

Ferruh Ağca. 2021. *Dillik Ölçütlere Göre Eski Uygurca Metinlerin Tarihlendirilmesi*. Türk Dil Kurumu Yayınları.

Akbar Amat, Arienne Dwyer, Gülnar Eziz, Alexandre Papas, and CM Sperberg-McQueen. 2018. Annotated Turki Manuscripts from the Jarring Collection Online.

Mustafa Argunşah and Galip Güner. 2015. *Codex Cumanicus*. Kesit Yayınları.

Erhan Aydın. 2017. *Orhon Yazıtları*. Bilge Kültür Sanat Yayıncılık.

Erhan Aydın. 2018. *Uygur Yazıtları*. Bilge Kültür Sanat Yayıncılık.

Erhan Aydın. 2019. *Sibirya'da Türk İzleri*. Kronik Yayıncılık.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 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)*, pages 44–50, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Eraslan, Kemal. 2012. *Eski Uygur Türkçesi Grameri*. Türk Dil Kurumu Yayınları.

Ahmet Bican Ercilasun and Ziyat Akkoyunlu. 2014. *Kâşgarlı Mahmud Dîvânu Lugâti' t-Türk*. Türk Dil Kurumu Yayınları.

Ahmet Bican Ercilasun. 2016. *Türk Kağanlığı ve Türk Bengü Taşları*. Derĝah Yayınları.

Marcel Erdal. 2004. *A Grammar of Old Turkic*. Brill, Leiden, The Netherlands.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

International Phonetic Association, International Phonetic Association Staff, et al. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

International Turkic Academy. 2017. Heritage of the Ancestors: Multimedia Fund.

Lars Johanson. 2021. *Turkic*. Cambridge Language Surveys. Cambridge University Press.

Mustafa S. Kaçalin and Nicholas Poppe. 2017. *Mukaddimetü'l-Edeb, Moğolca-Çağatayca Çevirinin Sözlüğü*. Türk Dil Kurumu Yayınları.

Yukiyo Kasai. 2014. The Chinese Phonetic Transcriptions of Old Turkish Words in the Chinese Sources from 6th-9th Century: Focused on the Original Word Transcribed as Tujue. 29:57–135.

Tolga Kayadelen, Adnan Ozturel, and Bernd Bohnet. 2020. A gold standard dependency treebank for Turkish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5156–5163, Marseille, France, May. European Language Resources Association.

Cemal Kurnaz. 2009. Bana Seni Gerek Seni. *Atatürk Üniversitesi Türkiyat Araştırmaları Enstitüsü Dergisi*, 15(39):147–160.

An-King Lim. 2021. On Sino-Turkic verbal functional expressions. *International Journal of Chinese Linguistics*, 8(1):102–138.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Kemal Oflazer and Murat Saraçlar. 2018. *Turkish Natural Language Processing*. Springer.

Mehmet Ölmez. 2017. *Köktürkçe ve Eski Uygurca Dersleri*. Kesit Yayınları.

Mai Omura and Masayuki Asahara. 2018. UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125, Brussels, Belgium, November. Association for Computational Linguistics.

Niko Partanen and Jack Rueter. 2019. Survey of Uralic Universal Dependencies Development. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 78–86, Paris, France, August. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden, May. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.

Martine Robbeets and Alexander Savelyev. 2020. *The Oxford Guide to the Transeurasian Languages*. Oxford University Press.

Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(1):13, Jan.

Alexander Savelyev and Martine Robbeets. 2020. Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, 5(1):39–53, 02.

Boris A Serebrennikov and Ninel Z Gadžieva. 2011. *Türk Yazı Dillerinin Karşılaştırmalı-Tarihî Grameri*. Türk Dil Kurumu Yayınları.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Talat Tekin and Mehmet Ölmez. 2014. *Türk Dilleri*. BilgeSu Yayıncılık.

Talat Tekin. 1965. *A Grammar of Orkhon Turkic*. University of California, Los Angeles.

The Unicode Consortium. 2021. *The Unicode Standard, Version 14.0.0*. Mountain View, CA. ISBN: 978-1-936213-29-0.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. UD Annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17, Prague, Czech Republic.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium, November. Association for Computational Linguistics.

Jens Wilkens. 2021. *Handwörterbuch des Altuigurischen*. Universitätsverlag Göttingen, Göttingen.

Fikret Yıldırım. 2017. *Irk Bitig ve Orhon Yazılı Metinlerin Dili*. Türk Dil Kurumu Yayınları.

Mağfiret Kemal Yunusoğlu. 2012. *Uygurca-Çince İdikut Sözlüğü*. Türk Dil Kurumu Yayınları.

Yarjis Xueqing Zhong. 2019. *Rescuing a Language from Extinction: Documentation and Practical Steps for the Revitalisation of (Western) Yugur*. Ph.D. thesis, School of Culture History and Language, College of Asia and the Pacific, The Australian National University.