Words in Old Turkish: Analysis and Evaluation

Universal Bindings, Syntactic Words, and Old Turkish Corpora in 2025H1

A Holistic Overview of Language Universals, Old Turkish, Bound Morphemes, and Universal Dependencies in 2025H1 for boosting multilingual interpretability in the age of neural models and AI

Oguz (Mehmet Oguz Derin), 2025年April月29日, UniDive WG2 Task 2.1.1

Overview

- The challenges of defining syntactic words, universals, and Old Turkish
- Limitations in current definitions of bound morphemes
- Evidence from historical manuscripts
- Case for a more defined delineation
- Implications for cross-linguistic analysis and AI
- Q&A Discussion

Deep thanks to the organizers for this opportunity, and I appreciate your understanding in advance due to the current busy schedule leading to a crammed presentation, which might be harder to view with mobile devices without zoom and pan.

Please find references at the bottom of the slide pages. If any specific reference seems to be missing, the preceding slides already include it, as the listed ones are not subject to repetition. Q&A at the end. Slides will be accessible online at
https://github.com/mehmetoguzderin/ud-2025h1-otk
soon after the meeting. Thank you for your understanding.

Motivation

- Universal Dependencies is now making progress in unifying syntactic word definition across languages with Haspelmath's publication
- Despite Haspelmath's work providing a great basis for most aspects, the definition of bound morphemes is limited in helping where analysis needs it the most, currently:
  - "attach to both verbal and nominal" or
  - "suffix is when only attaches to single class"
  - Examples from typologically similar languages
  - Creates a blocker for further documentation since consistent delimitation and treatment as words is a significant boiling point toward having such downstream steps in a principled, systematic, and cross-linguistically comparable manner
- Suggestion: Better define the role of varying conjugation strategies in agglutinative languages and others
- Aim: Boost multilingual interpretability in the age of neural models and AI.
- Scope: Hence, this presentation will focus on orthographic attestations, linguistic governing rules, and similar considerations for delineating bound morphemes rather than aspects that build upon them, such as head directionality, further annotations, or (pseudo-)compounds.

- Are "words" important for grammatical dependency analysis? (2025)
- Defining the word (2023)
- https://github.com/UniversalDependencies/docs/issues/986
- Interpretability and Analysis in Neural NLP (2020)
- Compound and pseudo-compound words in Orkhon Inscriptions (2016)

Linguistic Universals: Deep Roots & Comparability

- Language approximates concepts for communal or internal understanding as a brain-expression interface.
  - This practical function shapes universals and variations.
- Human language builds on foundational structures and necessities.
  - Lower Paleolithic: Emergence of word basics, G1 level.
  - Middle Paleolithic: Maturity in root syntax layer, G2 level.
  - Upper Paleolithic: Necessity of conjunctions, G3 level, emergence of mythology.
  - Mesolithic Holocene: Complex expression surfaces form and diversification at proto-language-families level.
  - Attempts to teach apes show an association of first the goal (object or location) and then the action (verb), an important clue.
- Tracking basic universals (e.g., pronouns requiring perspective shift, which apes can not acquire well) is relatively straightforward.
  - *ben, me, ware, wo* (1st person)
  - Cross-linguistic patterns (e.g., m and n for first) reflect anatomical/genetic constraints and embodied cognition.
- Later constructs (tech/culture related) are more challenging.
  - Language expanded to accommodate new concepts, layering constructions on universal foundations.
  - Deeper functional comparisons reveal parallels despite surface diversity (~7000+ languages).
- Goal: Highlight universality and parallels via considerate delineations for better cross-linguistic comparison and AI interpretability.

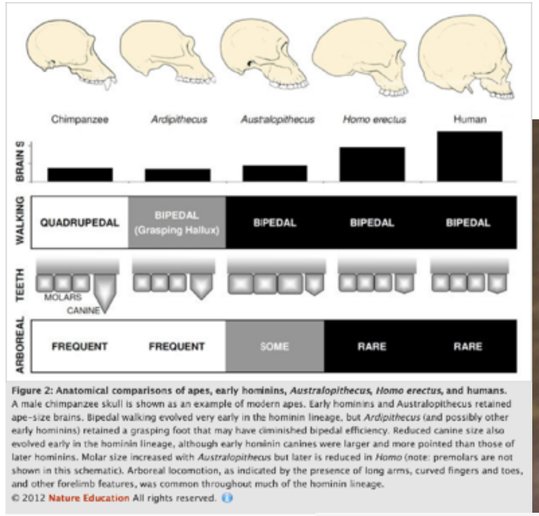


Figure 2: Anatomical comparisons of apes, early hominins, *Australopithecus*, *Homo erectus*, and humans. A male chimpanzee skull is shown as an example of modern apes. Early hominins and Australopithecus retained ape-size brains. Bipedal walking evolved very early in the hominin lineage, but *Ardipithecus* (and possibly other early hominins) retained a grasping foot that may have diminished bipedal efficiency. Reduced canine size also evolved early in the hominin lineage, although early hominin canines were larger and more pointed than those of later hominins. Molar size increased with *Australopithecus* but later is reduced in *Homo* (note: premolars are not shown in this schematic). Arboreal locomotion, as indicated by the presence of long arms, curved fingers and toes, and other forelimb features, was common throughout much of the hominin lineage.

- Speech components were already in place before great apes split from lesser apes.
- Premodern language capacity can be associated with *early Homo erectus sensu lato*.
- Modern language capacity emerged in *pre-archaic Homo sapiens sensu lato*.
- Root syntax is a basic layer and corresponds to conceptual representations.

So, how can we know when distinctively human language was first used? The archaeological record is invaluable in this regard. Roughly 100,000 years ago, the evidence shows, there was a widespread appearance of symbolic activity, from meaningful markings on objects to the use of fire to produce ochre, a decorative red color.

Like our complex, highly generative language, these symbolic activities are engaged in by people, and no other creatures. As the paper notes, "behaviors compatible with language and the consistent exercise of symbolic thinking are detectable only in the archaeological record of *H. sapiens*."

- Overview of Hominin Evolution - Nature (2012)
- When did human language emerge? - MIT News (2025)
- Linguistic capacity was present in the Homo sapiens population 135 thousand years ago (2025)
- Semiotics and the Origin of Language in the Lower Palaeolithic (2020)
- On how "early syntax" came about (2023)
- How Much Does a Human Environment Humanize a Chimp? (1999)
- Technological Features of Decorated Ivory Artifacts in the "Classic" Collection from the Mal'ta Site (2017)
- Acoustic and language-specific sources for phonemic abstraction from speech (2024)
- How did language evolve in the lineage of higher primates? (2021)

Why Old Turkish is a Critical Data Point for Bound Morphemes

- Agglutinative languages challenge tokenization, especially regarding bound morphemes.
- Old Turkish (OTK, approx. 7th-13th c.) is crucial due to:
  - Earliest extensively documented Turkic language, insights into development.
  - Data for early grammatical elements from words to morphemes.
  - Preservation in diverse scripts (Runiform, Uyghur, Manichaean, Sogdian, and more) with varying orthographic conventions.
  - **Lack of consistent delimiters (esp. Old Turkic script)** necessitates higher-order linguistic analysis for tokenization.

- An Etymological Dictionary of Pre-Thirteenth Century Turkish (1972)
- A Grammar Of Old Turkic (2004)
- A Grammar of Orkhon Turkic (1968)
- Orhon-Uygur Hanlığı Dönemi Moğolistan'daki Eski Türk Yazıtları (2021)
- Sibirya'da Türk İzleri (2019)
- On the Reduplication är- bar- in the Orkhun Inscriptions (2019)
- Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family (2020)
- Irk Bitig ve Orhon Yazılı Metinlerinin Dili (2017)

Important Peculiarities in Old Turkish Analysis

- **Homonyms:** Distinguishing different bound morphemes written identically.
- **Allomorphs:** Recognizing variations of the *same* morpheme (avoid treating as different).
- **Vowel Harmony:** Does not reliably delimit elements (e.g., affects separate words like modern 'mu'; genitive '-ki' ignores it).
- **Verbal-Nominal Interface:** Transitions have diverse strategies, and boundaries are not always clear-cut.
- **Terminology:** Due to loosely coupled academic circles in different geographies and lack of linear evolution in the scholarly use of terms such as particles, clitic, suffixes, and others, studies and their translations subject the terms to overloaded usage.

Principles

- **Observe language, do not impose meaning:** Aim is descriptive accuracy based on historical texts.
- **OTK ≠ Modern Turkic Languages:** Parts in 7th c. may fuse/disappear later (e.g., *ş* + *im* + *di* -> *şimdi*).
  - Furthermore, modern orthographies are neither stable nor remarkably consistent.
- Comparative assessment is made in additional consideration of the previously presented adjacent languages.
- **Harmonization ≠ Strict Molding:** Principled approaches acknowledging typology.
- Prioritize **corpus function & attestation** over pure hypotheticals.
- Avoid pitfalls:
  - **False Equivalences:** Distinguish borrowings and homonymous elements (e.g., Persian *ki* and Turkic *ki*).
  - **Focus:** General governing principles, not hypothetical edge cases or distractions from fundamental misunderstandings.
  - **Arrow of Time:** Different periods/dialects may need different (harmonized) approaches. Embrace polyphony from dialect, period, and script variations. Need for constructive, civil, inclusive discussion

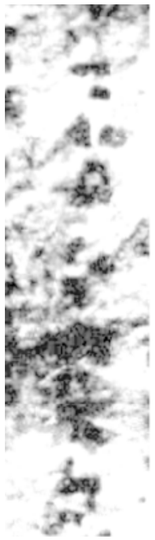
- UniDive Task 2.1.1 Meeting On Words in Other languages (fusional) (2025)
- A Study Of Uş Word (2012)
- An Introduction to Chaghatay: A Graded Textbook for Reading Central Asian Sources (2018)

A Brief Evaluation

- **Straightforward:** Pronominalized clitics (=m, =n), Copular clitics (=(i)di - treat as separate if spelled, empty node as ".ı = er-" if zero, but non-problem in OTK).
- **Affix:** Derivational morphemes (unless evidence suggests otherwise), deeply integrated and transitionary TAM markers.
- **More Argumentative:** Case markers, Converb markers, Plural markers, Possessive markers, Conjunctions.
- Current definitions (e.g., "attach to both verbal and nominal" = clitic; "suffix attaches to single class") are insufficient for agglutinative languages.
- Examples often from typologically similar languages.
- Underconsiders conjugation strategies crucial in languages like OTK.
- Hinders consistent cross-linguistic analysis (qualitative & quantitative).
- It is possible to reach further definition based on converging evidence from:
  - **Historical Orthography:** Scribal practices across multiple scripts.
  - **Linguistic Function:** Phrasal/clausal scope and cross-domain attachment.

Evidence I: Orthographic Practices (Old Turkic Script)

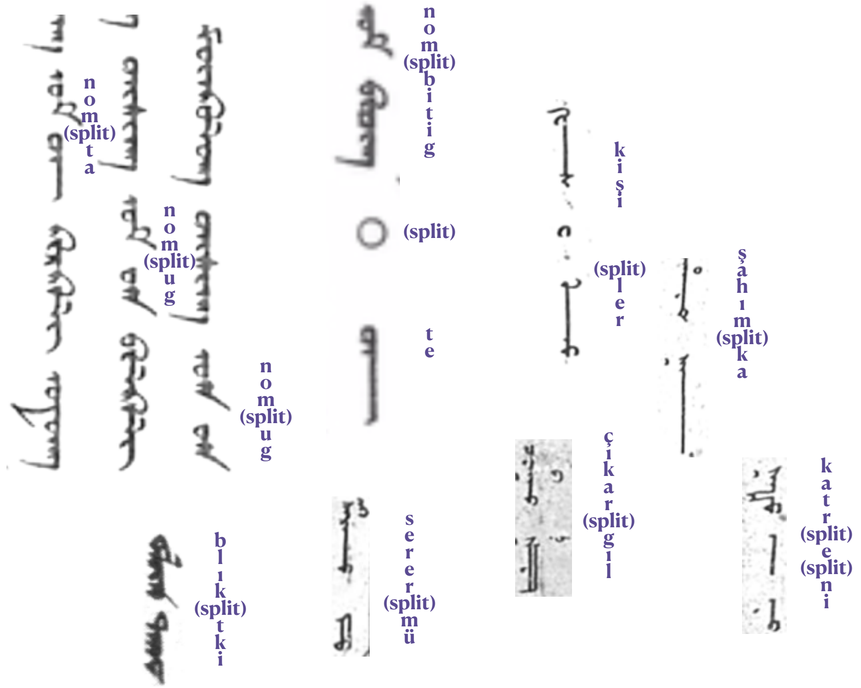
k
n
(split)
k
a



k
i
y
g
n
(split)
k
a

- Old Turkic script (largely 7th-9th c.) uses a **colon-like word separator (:)**.
- Examples include "ka" dative.
- These are not random splits or errors.
- Strongly suggests scribes perceived these markers as distinct units.
- However, the use of delimiter in Old Turkic script is not regular in general.
  - It is not rare (in fact, common) to see entire phrases written without any delimiter.
  - Nevertheless, except for a handful of cases, the delimiter does split meaningful units, so colons usually should not appear in the middle of a form of a token but rather act as a splitter.

Evidence I: Orthographic Practices (Old Uyghur, Manichaean, Sogdian, Brahmi, Arabic, and more)



- Other OTK manuscripts (Old Uyghur Script, Manichaean, Sogdian, and more, largely 7th-13th c.) often show separation:
  - Break in cursive connection before suffix.
  - Visible gap/space.
- Documented separation for *-dakï*, *-lar*, *-lug*, *-nïŋ*, tense *-gay* (like AUX), *-sar*.
- Consistency across different scripts indicates underlying reality, not just a peculiar convention or one-off typo.
- Similar to the Old Turkic script, splits are meaningful but, in this case, more regular, so it would be more faithful to preserve them as token boundaries at least.

to the preceding stems. In most East Old Turkic manuscripts, particularly older ones, case markers are written unconnected with the stem (Gabain 1950: 86).

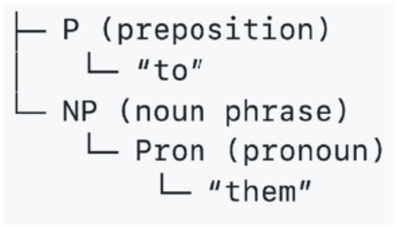
The ubiquitous interrogative particle *mU* appears, e.g., in *mini sävär mü siz* (KP 6,4-5) 'Do you love me?'. In Uygur writing as in this

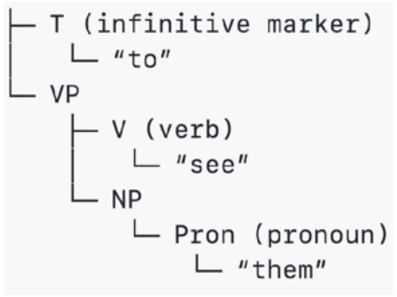
5. Suffixes are often written separately from their stem words, but it is difficult to predict when. Suffixes attached to nouns are more frequently separate, while those attached to verbs are generally joined.

- Eski Uygur Türkçesi Grameri (2012)
- Köktürkçe ve Eski Uygurca Dersleri (2017)
- The Paleographic Study on the Samarkand Copy of Atebetü'l-Hakâyık Written in Old Uyghur Script (2022)
- Turkic (Cambridge Language Surveys) (2021)
- Türkiye'deki Eski Uygurca Metin Neşirleri İçin Kullanılacak Harfçevrim ve Yazıçevrim Kılavuzu (2020)
- Das Alttürkische in sogdischer Schrift. Textmaterialien und Orthographie. (1991)
- Ein Hochzeitssegen uigurischer Christen. (1981)
- Alttürkische Handschriften. (2000)

Evidence II: Linguistic Analysis (Old Turkish)

- The proposed definition or exemplification in Haspelmath's work ("attaches to N and V = clitic, else affix") does not suffice to consistently delineate due to underspecification, despite providing a great basis otherwise.
- A crucial factor (even in IE) is often attachment to phrases/clauses, not just word classes.
- OTK morphemes often attach to nominalized/non-finite constructs.
- Viewing from a phrase/clause perspective clarifies homonyms vs. allomorphs.
- Proposal: Incorporating **conjugation capabilities or further appropriate consideration** could enhance the definition.
- Examples:
  - Converb *-čä* / Equative case *-čä*: Same form, different scope (clause vs. NP).
  - Instrumental *-n* / Converb *-n* ("by doing"): Parallel functions.
- Categorization based on higher order context analysis would clearly delineate.





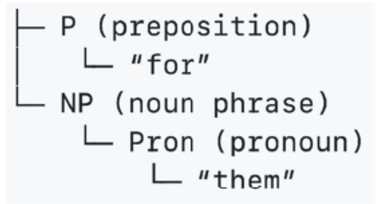
Definition 4: clitic
A clitic is a bound morph that is neither a root nor an affix.

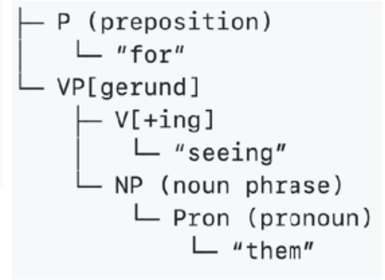
Examples of English clitics were given in (1b): *the, to, 's*. They are not roots because they are not contentful forms (forms denoting objects, actions or properties; see Definition 6 below), and they are not affixes because they combine with roots of different classes.[2]

Next, we need a definition of *affix*, because one of the most important ways in which composite words differ from combinations of words is by combining a root with an affix. Definition 5 serves to delimit affixes from clitics (Haspelmath 2021a, §6).

Definition 5: affix
An affix is a bound morph that is not a root, that must occur on a root, and that cannot occur on roots of different root classes.

Like clitics, affixes are bound morphs that are not roots, but in addition, they must always occur on roots of the same class, i.e., always on verb roots, noun roots, or adjective roots. For example, the German infinitival marker *zu-* (in *aus-zu-gehen* 'to





The third, and perhaps most widely accepted, hypothesis claims that the directive case had a verbal origin, such that in early Turkic languages many of the words

composite form, consisting of the vowel converb with the equative suffix; that would give the reading *bol-(u)+ča* in that passage and *yogur-(u)+ča* in Tuñ 26. *boluča* appears also in KT SW as completed

here mainly with Turkish units. As far as their development is concerned, the following diachronic path may be assumed: (1) lexical element > (2) enclitic particle > (3) enclitic suffix > (4) non-enclitic suffix.

D. Sinor çıkma hâli ekinin yapısını, +*tı* "zarf-fiil eki" ile,+*n* "vasıta hâli eki"nin birleşmesi şeklinde açıklamıştır.+*dın* / +*din* < +*dı* /+*di*+ *n*[431]

- The Oxford Guide to the Transeurasian Languages (2020)
- On the Origin of the Directive Case in Turkic (2002
- The Cambridge Grammar of the English Language (2002)

Some Ideas on Criteria for Syntactic Word Status

- **Cross-Domain Attachment:** If a bound morpheme attaches to both nominal bases AND verbal/phrasal/clausal constructs... -> Candidate for separate word.
- **Co-occurrence:** If it does not co-occur with independent tense/agreement... -> Candidate for separate word.
- **Orthographic Separation:** If scribal practice across scripts shows separation (colon, gap, extender)... -> Strong evidence for separate words.
- **Derivational Morphemes:** Unless clear evidence otherwise... -> Candidate for affix.
- **Subtree Morphology:** Morphological traits might belong to subtrees, not just words/features.

Implications

- Create more **cross-linguistically comparable** trees (measurable by depth, breadth, linking).
- Improve **multilingual interpretability** for AI / neural models.
- Maintain clearer distinctions between **closed** and **open** classes.
  - Unfortunately there are some treebanks that depend certain closed classes on NOUN or VERB through agglutination, causing them to be open classes.
- Principles (orthography, phrasal scope) may apply elsewhere.

Implementation: Treebanking & Tooling

- Significant update to OTK treebank WIP (pending decisions/tooling, not only UD aspects).
- Representing an extended set of ligatures and multi-consonant graphemes through existing UD sub-tokenization.
  - Enables decomposition-free normalization.
- Authoring chain: UDXML with Scheme -> Rust pipeline -> CoNLL-U.
- Rust chosen for processing speed (vs. JS benchmarks).
- Focus on **verifiable attestations** where reading/analysis is established.

```
<!-- Multi-word tokens example -->
<sentence id="2" text="vámonos al mar" text_en="let's go to the sea">
 <!-- Multi-word token with tokens -->
 <form id="1-2" lemma="_" upos="_" xpos="_" feats="_" head="_" deprel="_" deps="_" misc="_">
  vámonos
  <token id="1" lemma="ir" upos="VERB" xpos="VMM01P0" feats="Mood=Imp|Number=Plur|Person=1|VerbForm=Fin" head="0"
   deprel="root" deps="0:root" misc="_">
   vamos
  </token>
  <token id="2" lemma="nosotros" upos="PRON" xpos="PP1CP000" feats="Case=Acc|Number=Plur|Person=1|PronType=Prs"
   head="1" deprel="obj" deps="1:obj" misc="_">
   nos
  </token>
 </form>
 <form id="3-4" lemma="_" upos="_" xpos="_" feats="_" head="_" deprel="_" deps="_" misc="_">
  al
  <token id="3" lemma="a" upos="ADP" xpos="SPS00" feats="_" head="5" deprel="case" deps="5:case" misc="_">
   a
  </token>
  <token id="4" lemma="el" upos="DET" xpos="DA0MS0" feats="Definite=Def|Gender=Masc|Number=Sing|PronType=Art" head="5"
   deprel="det" deps="5:det" misc="_">
   el
  </token>
 </form>
 <form id="5" lemma="mar" upos="NOUN" xpos="NCMS000" feats="Gender=Masc|Number=Sing" head="1" deprel="obl" deps="1:obl"
  misc="_">
  mar
 </form>
</sentence>

<!-- Empty nodes example -->
<sentence id="3" text="Sue likes coffee and Bill tea" text_en="Sue likes coffee and Bill tea">
 <form id="1" lemma="Sue" upos="PROPN" xpos="NNP" feats="Number=Sing" head="2" deprel="nsubj" deps="2:nsubj" misc="_">
  Sue
 </form>
 <form id="2" lemma="like" upos="VERB" xpos="VBZ" feats="Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin"
  head="0" deprel="root" deps="0:root" misc="_">
  likes
 </form>
 <form id="3" lemma="coffee" upos="NOUN" xpos="NN" feats="Number=Sing" head="2" deprel="obj" deps="2:obj" misc="_">
```

- Reading Old Turkic runiform inscriptions with the aid of 3D simulation (2024)
- Augmenting parametric data synthesis with 3D simulation for OCR on Old Turkic runiform inscriptions (2024)
- Universal Dependencies for Old Turkish (2021)

Conclusion

- Orthographic and linguistic evidence strongly supports treating key OTK bound morphemes (case, plural, possessives, related converbs) as **separate syntactic words** in UD.
- This approach:
  - Respects historical evidence and orthography from multiple scripts.
  - Captures syntactic function accurately.
  - Provides more consistent annotation.
  - Reflects productivity of the language at the time of OTK better.
- Hope: UD reflects linguistic parallels across languages through comparable syntactic structures.
  - A more defined delineation of bound morphemes' classification would largely settle the case here.
- Lots of progress for infrastructure to improve digitalization of OTK, challenges besides UD.
- Would be happy to collaborate on UDXML with Scheme through Rust-based tooling if it resonates.

Special Acknowledgements

- Deep thanks to the workshop organizers for the opportunity.
- Deep gratitude to Dr. Erdem Uçar for particular orthography references.

Q&A Discussion

- Happy to participate in discussions and answer questions.