Morphological glossing in the ATMO project

Arienne M. Dwyer
(v.0.5 2016-02-16, v. 0.9 2017-06-04)

Table of Contents

1 Introduction

The *Annotating Turki Manuscripts from the Jarring Collection Online* project (ATMO, Henry Luce Foundation, 2015–2017, Arienne M. Dwyer and C.M. Sperberg-McQueen, PIs) grew out of the co-PIs' interest in making transcriptions and basic linguistic, cultural, and historical analyses of late eastern Chaghatay (Turki) texts available to the public for further use. These texts date primarily from the 17th to the mid-20th century. The current document describes the linguistic annotation chosen and its motivations.

The ATMO project was preceded by the *Uyghur Light Verbs* (UyLVs) project (2011–2015, NSF-BCS1053152, Arienne M. Dwyer, PI). The current morphological glossing scheme developed from one for the UyLVs project. That prior project focused on the typology of complex predicates (verbal and nominal light verb constructions), and looked backwards in time from modern Uyghur; the

current project instead looks forward in time from mid to late Chaghatay up until the mid-20th c., and aims for a broad account of morphology.

The morphological glossing tagset and this document were prepared by Arienne Dwyer, in consultation with Claus Schönig (May 2011, June 2014, and May 2015). It is the third iteration of a tagset originally prepared for the UyLVs project in 2011.

- *Version 1* (2011): Dwyer first compiled a list of Modern Standard Uyghur (MSU) grammatical morphemes in 2009 for a textbook (Engesæth et al. 2009/2010), augmenting the list with items from Tömür (1987) and Sugawara and Osmanov (2007), and referring to the Turkic Terminology draft 4 (AATT 2004), the Turkish Treebank (Oflazer et al. 2003), Kornfilt (1997), and Friedrich (2002). That first iteration in 2011 numbered approximately 250 items, including derivational grammatical morphemes and grammatically-relevant inflecting lexemes.
- *Version 2* (2014): After several years of tagging on the UyLVs project (mostly by Gülnar Eziz and Travis Major, who suggested the addition of about ten items), Dwyer revised the tagset, adding many items hand-culled from UyLVs annotated texts, especially premodern Uyghur and Chaghatay (again in consultation with Schönig). The tagset had grown to over 300 items.
- *Version 3* (2015–2017): The current tagset represents a major revision to focus on Chaghatay (unlike versions 1 and 2) and still includes modern Uyghur (both standard and non-standard). Items and glosses are primarily from the UyLVs/ATMO corpus (with the identification of tagging errors and inconsistencies greatly facilitated by a March 2015 "anomaly editor" by co-PI Sperberg-McQueen) and from Schönig (1997), Eckmann (1966), with Old Turkic comparative material from Erdal (2004).

The current version must be considered a draft until the ATMO team and others can evaluate and improve it by identifying gaps and inconsistencies. The tagset would also benefit from an automated extraction of grammatical morphemes from the existing UyLVs and ATMO corpora, a task which is likely beyond the scope of the current project.

2 Scope

The ATMO corpus focuses on manuscripts in late **Chaghatay** (ISO 639-3:chg) and **premodern to early modern Uyghur** (uig); the language of the latter period is also known as Turki. Temporally, the ATMO corpus spans **ca. 17th to mid-20th century**. At a minimum, the tagset must account for the Chaghatay and premodern Uyghur forms during that period. Ideally, the current tagset would also cover a much longer timeline, namely the earlier Chaghatay period through modern Uyghur (i.e. the 14th or 15th century to the 21st century). Since both forms and meanings of morphemes change, the ATMO tagset must account for several stages of the language. For details, see section 5 "Rationale" below, and the comments to the individual tags in the tagset itself.

The MORPHOLOGICAL GLOSSING of the ATMO project is intended to capture the part of the linguistic system that is the set of morphemes, including affixes and stem morphemes, primarily that of inflectional morphology. A speech community uses classes of lexical words in its linguistic system, not just stems but idioms, clitics and particles. These are classified into the parts of speech (which are distinguished by the syntactic criteria outlined in 7.1 below).

LINGUISTIC ANNOTATION is the association of linguistic information with each segment (here, morpheme) of the transcribed data. Here the linguistic information is primarily morphological, and the

segments are morphemes. The process of linguistic annotation entails at least SEGMENTATION (here, identifying each segment as a morpheme, word, text line, and sentence) and morphosyntactic annotation (here, about the segments in the primary data, e.g., a morphosyntactic annotation in which a part of speech and lemma are associated with each segment in the data). To make this morphosyntactic annotation useful for different users, we provide PART OF SPEECH information (a lexical category tag) and INTERLINEAR GLOSSES (brief labels for a single meaning or sense of a linguistic form).

An interlinear glossed text commonly consists of some or all of the following, usually in this order, from top to bottom (adapted from Lehmann 2004):

- An electronic version of the original orthography
- A conventional transliteration into a Latin alphabet
- A phonetic transcription
- A morphophonemic transliteration, where morphemes within a word are separated
- A morpheme-by-morpheme part of speech gloss, using all-caps abbreviated grammatical tags
- A morpheme-by-morpheme gloss, with substantives glossed in a translation language and grammatical categories marked with all-caps abbreviations of grammatical tags
- A translation, which may be literal or free. (It "may be placed in a separate paragraph or on the facing page if the structures of the languages are too different for it to follow the text line by line" (Lehmann 2004)).
- Commentary (linguistic, textual, cultural, historical, etc.)

MINIMALLY ACCEPTABLE LEVEL OF ANNOTATION:

Documentary linguists have found that at a minimum, linguistic annotation should include at least (1) some kind of rendered text (orthographic or phonemic transcription), (2) a free translation, and (3) "any contextual commentary that is essential for the interpretation of the communicative event in question by outsiders" (Schultze-Berndt 2006:248). For documenting endangered languages, the VW-DOBES found that to make those data interpretable, in practice the "contextual commentary" in (3) should not merely be socially-situated metadata, but should also best include interlinear glossing (Dwyer 2000).

Those familiar with the language may well require less annotation than linguists. The UyLVs team, for example, found that searching only on transcribed text with associated metadata (without interlinear glossing) was sufficient for preliminary analysis.

The target users of the ATMO project are diverse; at a minimum, linguistic annotation should include an orthographic transcription and metadata. The ATMO project includes a portion of the newly-scanned manuscripts to be fully annotated (including all of the above bullet points), and another group of manuscripts to be presented in the minimal format, with only orthographic transcription and metadata.

ORTHOGRAPHY:

The tagset is represented provisionally in a minimally-extended Uyghur Latin script (ULY), to facilitate typing. Uyghur Latin represents /e/ and /ɛ/ as *é* and *e*, respectively (rather than the *e* and *ä* of European Turkology); it also represents /j/ as *y*, /y/ as *ü,* and /ɣ/ as *gh*. The extension here includes back *i*, represented as *ï* = IPA /ɨ/ and vowel length (represented by doubled vowels, e.g. *dunyaa* = IPA /dunja:/). The morpheme forms, tags, and allophones here are presented in this minimally extended ULY form. Since Modern Standard Uyghur (MSU) does not represent *ï* or vowel length, the modern Uyghur *examples* here do not represent these two features. (As this is new to ATMO, there are no doubt many inconsistencies, even in the current document, to address.)

3 Outside Scope

The tagset is not (yet) intended to account for syntax, grammatical relations, or discourse features.

It also does not account for derivational morphology (e.g. *ishchi* is composed of the noun *ish* 'work' and the agentive denominal suffix *+chi;* in our project, *ishchi* is always glossed as a noun and never segmented.)

When morphemes have both derivational and inflectional characteristics, then we provide morpheme glosses for them. Examples of the latter type include voice suffixes such as the causative voice (*kör-* 'see' glossed as Vt, *kör-set-* 'show' glossed as Vt-CAUSSET), the diminutive *+KInA,* the abstract nominalizer *+lIk* (in e.g. *emeslik, qiliwatqanlighi*) etc.

4 Goals of tagset

The tagset is designed to represent parts of speech for written and spoken Chaghatay and its descendants.

1. Form is prioritized over function; thus the morphological gloss should succinctly reference a formal property of the morpheme if possible, rather than its function. (Otherwise, one would be compelled to assign many different functional tags to the same synchronic morpheme.)
2. The tagset should express the basic grammatical contrasts in the language, e.g. anteriority and non-anteriority, directivity and indirectivity (cf. "expressive adequacy," Ide et al. 2004).
3. Parsimony: Tagging is a tradeoff between time and available resources; generally, the more linguistic distinctions accounted for, the more time annotation consumes.
4. The tagset should be easily recognizable and interpretable to its presumed users (primarily Turkologists and linguists); therefore, this tagset takes into consideration (1) the the so-called Leipzig Glossing Rules (2008–2015); (2) Draft of Turkic Terminology (AATT 2004); and (3) the glossing scheme of the Turkish Treebank (Oflazer et al. 2003).
5. The part of speech annotation (POS) contains only POS tags found in the tagset. The interlinear glossing tier (ILG) contains (1) a terse literal English gloss of lexemes (N, V, AV, AJ etc), and (2) POS tags for grammatical morphemes (identical to those in the POS tier).

Goals 1–3 above are somewhat at odds with each other; basic grammatical contrasts are difficult to distinguish without some reference to function, (in)directivity being a perfect example. Without being parsimonious, one could assign form to one tier and function to another (for an extreme version, see e.g. the thought experiment of Lieb and Drude 2000). But such an approach is impractically time-consuming in the extreme, and no one to our knowledge has ever implemented such a scheme.

In the previous UyLVs project, POS and ILG were aligned at the sentence level. In the ILG, certain grammatical morphemes were glossed with different tags in the POS and ILG tiers, to facilitate searching for sub-types and functions. For example, light verbs were tagged in POS either as a nominal or verbal light verb (LVN, LVV), and the ILG tier spelled out in capital letters an archiform of that verb (e.g. *QIL, BAR, QOY* etc.). To reduce complexity, the ATMO project has eliminated this POS/ILG split, so that nominal light verbs are tagged LVN in both POS and ILG, and verbal light verbs, LVV.

5 Rationale for resolution of problematic issues

5.1 Mergers and splits

The tagset has to deal with two kinds of mergers and splits. In the tagset itself, there are two ways to identify these: (1) referring to the (column) spl/merg [splits and mergers]; (2) by comparing the (ATMO columns) morph_new and ilg_new with the (UyLVs columns) POS_old and ILG_old.

5.1.1 Representing multiple synchronic states

In accounting for several hundred years of related languages (here, Chaghatay and early modern Uyghur), due to diachronic change, we are in effect accounting for successive synchronic "time slices" of the language(s). One common result of language change is grammaticalization. For example:

- Change of form, function, and semantics: The Old Turkic and early Chaghatay converbial construction -A *tur(ur)* (expressing a durative imperfective, tagged as -CNV LVV in POS and -CNV TUR in ILG) grammaticized to the late Chaghatay verbal clitic -*adur*- (tagged as PRS), to the modern Uyghur verbal suffix -*idi/-idu* (tagged as PRS, expressing primarily present tense). Since the ATMO corpus contains several forms, -A *tur(ur)* is tagged as -CNV LVV and -CNV TUR, -*adur*- is tagged as PRS, whereas (MSU) -*idu* is tagged PRS.3s2.
- Change of form: The Old Turkic and early Chaghatay copular auxiliary *érken* was in Uyghur grammaticized into the auxiliary particle *iken*. Both are tagged XINDIR.
- Change of semantics: The Old Turkic and Chaghatay *plural* personal pronoun, *siz* 'you (pl.)' is in Modern Uyghur a singular formal form. Therefore, *siz* is tagged as plural (PN.P2p) for Chaghatay texts, and singular formal for Modern Uyghur texts (PN.P2sf).

5.1.2 From the UyLVs tagset to the ATMO tagset

1. The earlier UyLVs tagset primarily accounted for modern Uyghur, with the addition of nonstandard and early modern forms (below, SEG is the segmentation tier, POS is the POS tier, ILG the ilg tier).
2. The current ATMO tagset attempts to account for at least 300 years of Chaghatay and Uyghur.
2. In adopting a new tagset, the project is converting legacy material from the UyLVs to the ATMO tagset. Some tags have been retained, some have been changed, and some are new.

- Retained tags - no action required.
- For the UyLVs tags changed in ATMO :
    - Most map 1:1 onto the new ATMO tag.
        - For example, the zero morpheme on the second person imperative in e.g. *ber! kél*! was in UyLVs tagged IMP.zero in POS is in ATMO tagged 2si.IMP. (Seg is e.g. *ber-0)*
        - In ATMO all POS grammatical tags are repeated in the ILG tier; since in UyLVs, the ILG tier sometimes had different functional tags, there will be a 1:1 substitution, e.g. -*(I)p két*- was tagged CNV LVV in POS and CNV KET in ILG. In ATMO, both tiers would be LVV. Another example: *ber!* was tagged IMP.zero in POS and 2si.IMP in ILG; in ATMO both tiers are now 2si.IMP.

        - In ATMO, all word boundaries are delineated by whitespace. In UyLVs, the close bond of certain syntactic units (incl. light verbs to their preceding converbs) were marked with a pound sign (#). These are replace 1:1 with whitespace in ATMO. So e.g. *-(I)p két-* was tagged (in POS) -CNV#LVV- in UyLVs but -CNV LVV- in ATMO.
    - Some of the UyLVs tags map 1:2 onto ATMO tags, when:
        - A morpheme ignored in UyLVs is marked, e.g. the 3rd person form imperfective/aorist *-Ur* in UyLVs was tagged as AOR, in ATMO the zero morpheme following the suffix *-Ur* is marked (as *Ur-0* in seg) and as IPFV-3 in POS/ILG.
        - One OT form and/or grammatical category develops two forms or senses over time. For example, *néme* in Chaghatay is an indefinite pronoun (PN.INDEF) 'something, anything, thing' only; in MSU, it has become primarily an interrogative pronoun (PN.INTER) 'what?', though both functions occur in MSU. So, UyLVs tagging for néme was PN.INTER; ATMO has PN.INTER and PN.INDEF.
    - Some map 2:1 , for example the passive morpheme (used also for the reflexive -(I)l) in UyLVs was glossed both REFX and PASSL; in ATMO, both are glossed PASSL.
- New tags were introduced for Chaghatay forms that are not present in modern standard Uyghur, e.g. the tag PREP for *taa*, in e.g. *lut.fungnï könggülge <u>taa</u> qïyaamat yétkür* 'grant your favors to the heart until the resurrection' (Eckmann 1966:133)

5.2 Zero morphemes

We mark zero morphemes (those morphemes without a surface realization) with a hyphen and zero (*-0*). These include the following:

kel-0! -IMP-2si Second person singular informal imperative

kél-di-0 -ANT.DIR-3 The 3rd person zero affix following past/perfect direct affix -di is unmarked for number.

kel-se-0 -COND-0 The 3rd person zero affix following the conditional -sA is unmarked for number.

bar-idu-0 -PRS-0 The 3rd person zero affix following the modern Uyghur present -idu is also unmarked for number. Schönig has argued that -idu should be considered an allomorph of -i- (PRS), since as we know diachronically, the PRS morpheme is idu(r) (<ADUr), whether 3, 2, or 1st person. e.g. *bar-i-men < bar-Adur-men.*

NB:

- Except for the 2nd person informal imperative (*kel-0*), we did not consistently mark these in the UyLVs project (often erroneously as IMP.2si, PST.DIR.3s, COND, and -PRS or PRS.3s, respectively).
- Much of traditional Turkology conflates the tense-aspect with person marking, and so considers *-di* to be a "3rd person past suffix"; ditto the zero imperative. Much of traditional Turkology also usually ignores 3rd person marking on the conditional, and and analyzes *-idu* as a "3rd person present suffix".)
- Nominative case is zero-marked, but to date we have not marked it with a zero morpheme, e.g. (MSU) *siz* PN.P2f, not **siz*-0 PN.P2f-NOM.

5.3 Word (word stem) spans more than one morpheme

For example, *kör-* 'see' glossed as Vt, *kör-set-* 'show' glossed as Vt-CAUSSAT in UyLVs POS, but in ILG, as see-CAUSSAT-, even though the causative form 'cause to see' is best glossed as 'show'. (Another possible solution: SEG: *körset-* POS: Vt.CAUSSAT. ILG: show. This solution makes searching for the stem and the causative morphemes harder.)

Another example type: compounds, e.g.:

- Composed numerals, orthographic *on ikki* 'eleven' if represented in SEG as *on ikki* would result in the erroneous ILG of a sequence of two independent numeral, *'ten one'. So in SEG, 'eleven' must be represented as *onikki. Cf. beš yüz*, 'five hundred'.
- Compound words, such as (orth-ULY) *ata-ana* 'parents' (composed of 'father'-'mother'), <ipa>atʰaʔana</ipa> <w>ata-ana</w> <gloss>parents</gloss>
- Approximate numerals (via juxtaposition) are treated as separate words: *Ikki üch adem keptu.* 'Two or three (~a few) people came.' *ikki üch* NU NU <w>two</w> <w>three</w>.

5.4 What counts as a clitic?

A morpheme which takes on the phonological and prosodic properties of a preceding host (usually by assuming the host's vowel and consonant harmony, while the clitic itself is unstressed).

May be auxiliary or particle. Examples of clitics in Chaghatay and modern Uyghur:

- chg: epistemic copula *=dur* interrogative *=mi* complementizer *=ki*
- MSU: *=ken, =(i)mish;* sentential *=chu, =ghu*, interrogative *=mu*

The defective copular auxiliaries in Chaghatay (*=dur, e(r)-*) have largely become suffixes in MSU. True postpositions (see 7 below) should be treated as clitics, pseudo-postpositions should not (Schönig p.c. 2015), thus *uning=bilen* but *uning ich-i-da*. See Clitics in 6.1 below.

5.5 Redundancy: representing composed forms

In the presentation of the tagset, given that Turkic is agglutinative, we've chosen to list many of the composed forms (annotators may expect to see them composed, and in some cases the composed glosses are different than the sum of the separate glosses. E.g. there are separate entries for *-GAn, bol-* and *-GAn bol-*; for *-DI-0, -GU, dé-,* and *-DI-[]ghu deymen.*

Due to the diachronic nature of the tagged materials, if a morpheme significantly changes its form to the extent that it is **segmented differently** in chg than uig, the tagset also redundantly presents both forms. For example, Chaghatay *-(X)p tur(ur)* vs. Uyghur *-(X)ptu(r)*. (Morphemes that do not change their form and segmentation are not listed twice.)

5.6 How much form and function to gloss?

Given the principles of glossing form over function, and of parsimony (as set out in 4 above), the ATMO project will have a single morph tag in both (the equivalent of) the POS and ILG tiers. In the UyLVs project, for certain morphemes of particular interest, we assigned different POS and ILG glosses, such as verbal light verbs (e.g. *qal-* as a light verb was glossed LVV in POS and QAL in ILG), to facilitate the extraction of all variants of *qal-*. Another example from UyLVs is the participle *-GAn*, which we marked PRTC.PST when functioning as a plain participle, but PRTC.RZR when functioning as a relativizer. This approach was prone to annotator error; further, how much functional detail to capture is rather arbitrary. Further, that approach failed to distinguish finite and non-finite *-GAn*.

So for ATMO, we propose to use the same gloss for both (the equivalent of) the POS and ILG tiers. So, the light verb *qal-* would be glossed LVV (in both POS and ILG), and the non-finite *-GAn* would be PRTC.PFV (perfective participle, changed from UyLVs PRTC.PST and/or PRTC.RZR), while finite *-GAn* will be glossed PFV.

6 Segmentation scheme

6.1 Current segmentation scheme (ATMO)

The current project distinguishes three degrees of morphological boundedness: word, clitic, and affix.

- **Word** (canonically marked with whitespace)

Words can be monomorphemic or multi-morphemic, and mono- or multi-lexemic. A sequence of two or more morphemes or lexemes is distinguished as one "word" if the sequence (1) can be identified as a single part of speech and (2) is semantically interpretable as a single unit.

  - Multi morphemic example: the lexicalized MSU postposition *toghrïsïda* 'concerning, about' is considered one word (even though it can be segmented into *toghrï-sï-da*)
  - Multi-lexemic example: *Ijtima'i panler akademiyisi; til-yeziq komititi.* These words typically are dictionary headwords, except for proper nouns (such as *Yakup Tursun*, *Yengi Hisar*).

- **Clitic** (marked with equals sign =)

Clitics are are loosely bounded to a host stem (which may be inflected, and in Chaghatay precedes the clitic) Dwyer has observed three types of clitics to date: (1) clausal clitics (e.g. *=la, =mu*); (2) clausal/ sentential clitics (e.g. MSU *=ghu, =chu, =de*), and (3) true postpositions (POST e.g. Chg. *birle(n), burun*, MSU *bilen, burun, üchün, dek*). Turkic languages have true and pseudo-postpositions (POSTP, see 7.2 below); the former are non-inflecting and should be treated as clitics (Schönig p.c. 2015), i.e. as N=POST. (Pseudo-postpositions should be treated as free morphemes (N POSTP), e.g. Chg, MSU *ich, ust*, N-*ning toghri-si-da* N-GEN POSTP-POSS-LOC).

NB: (1) The UyLVs project did not segment *toghr-i-si-da*, but we now recommend that ATMO does so (Schönig agrees). (2) Clitics are not known in China, nor in Turkology; many term these "suffixes."

- **Affix** (marked with a hyphen -)

In Chaghatay and Uyghur inflectional morphology, these follow the stem. (In derivational morphology, Chaghatay has a number of Persian prefixes.)

6.2 Previous segmentation scheme (UyLVs)

Although no longer part of the ATMO morphological glossing, besides word, clitic, and affix boundaries, the UyLVs project also distinguished a fourth boundary type: some closely-bound syntactic units (marked by #). To be grammatical, these units could not be separated by any intervening material, including a pause. (There are some exceptions; within Tatar verbal light verb constructions at least *-gina* and *=dA* can be inserted (Schönig, p.c.).)

For nouns and adjectives, this marking was used for two or more nouns forming a larger syntactic (and sometimes semantic) unit, e.g. *öy#igi-si* home#master-POSS3 'head of household';

Mahmud#alKashgari 'Mahmud al-Kashgari'; *ap#aq* AJ.REDUPP#AJ 'very white', *chay#pay* AJ#AJREDUPP 'tea and snacks', but *ijirmijir* AJ 'jumbled, disorderly' (pseudo-reduplication, neither *ijir nor *mijir alone are in the lexicon).

Examples of verbs forming larger syntactic units are, first, nominal light verbs formed with a N/AJ + the verb 'do' (usually *qïl-*, less commonly *et-*), *bol-* 'be, become', and occasionally other verbs, e.g. *teyyer#qil-* N#LVN- preparation#do- 'prepare'; *hapa#bol-* N#LVN anger#become- 'be(come) angry'. Second, directional complements and verbal light verbs also were marked in this way: *bér-ip#kél-* go-CNV#come- 'go out and come back', *chüshendür-üp#kél-* understand-CNV#come- 'come to understand', *oqu-p#qoy-* read-CNV#QOY- 'look over, read cursorily'.

The ATMO project no longer uses the above ad hoc marking of syntactic boundedness with #.

6.3 Reduplication

Reduplication is a common Turkic feature in nearly all parts of speech, particularly adjectives (examples given in the ULY orthographic form). Since the reduplicated portion is syntactically and prosodically dependent on the host, we propose to treat the reduplicant as a clitic within a segmented word. In cases of pseudo-reduplication, the entire string is treated as one unsegmented word:

- Adjectives: *ap-aq* AJ.REDUPP=AJ 'very white' (cf. *aq* 'white'), *chay-pay* AJ=AJREDUPP 'tea and snacks' (cf. *chay* 'tea'), but *ijir-mijir* AJ 'jumbled, disorderly' (pseudo-reduplication, neither *ijir nor *mijir alone are in the lexicon).
- Nouns: (orthographic): *xilmu-xil* (seg): <w>*xil=mu=xil*</w> (pos) N=PRT=N.REDUP (gloss) 'all sorts of', from *xil* 'type, sort'; *yut-yutqa* 'from homeland to homeland' (pos) N=N.REDUP-DAT; (/yurt/, uig19561004_as4t23); *renga-reng* 'all sorts of colors' (uig1905_kg_HorseCamel1); *qïsm-qïsm* (pos) N=N.REDUP 'all kinds' (uig1905_kg207-ii14_garm3)
- Verbs: *Tal aynalur-aynalur* (pos) N V-PASSL-IMPV=V.REDUP-PASSL-IMPV 'the tree grows and grows' (uig19561118_yk5t48)
- Adverbs: *ayrim-ayrim chüshütö* (pos) AV=AV.REDUP V... '(They) descended separately.' (uig19561126_ht2t53)
- Interjections: *i-i henim anglang, bu sözné* INTJ=INTJ.REDUP (uig19561108_mr3t34)
- Measures: *deste-deste* (pos) M=M.REDUP 'bouquet upon bouquet' (uig19561108_mr2t33.xml); *térem-térem suyu var* (pos) M=M.REDUP ... 'There was trickle upon trickle of juice' (uig19561004_as10t29)

7 Tag set and definitions

For the tagset itself, refer to the spreadsheet UyMorphTags3.ods.

7.1 Notational conventions

- Morphophonemic notation: In the tagset, capital letters comprise a whole set of vowels or consonants, which are generally harmonic variants in native stems, but sometimes regional variants. The forms below are given in IPA, with their European Turkological equivalents in parentheses, e.g. ɨ (ï). Allowed domain: This notation appeared only in the tagset in UyLVs; in ATMO, we could consider using the archiforms of grammatical morphemes in the seg tier,

which may aid querying, e.g. *toghri-si-DA, qil-(I)n-GAn, untu-(X)p qal-DI-0.*

- X - the vowels /i ɨ (ï) u y (ü) ø (ö)/ or some subset thereof.
- A - the vowels /a ɛ (ä)/
- I - the vowels /i ɨ (ï) /
- U - the vowels /u y (ü)/
- O - the vowels /o ø (ö) /
- G - the consonants /g k ɣ/ʁ (ġ ğ gh) q/
- Q - the consonants /k q/
- D - the consonants /t d/
- ( ) -phonologically conditioned, e.g. for *+(X)p*, the suffix occurs with a vowel represented by *X*, except when the stem is vowel-final.

- { } Curly braces { } enclose unclear, or partially or completely **illegible or inaudible material**. If completely illegible, then marked {illegible} (possibly also { }); otherwise, a transcriptionist's best guess appears between the curly braces, e.g. {fslt}. Domain: UyLVs orth tier; proposed for the ATMO lit tier.
- Curly braces { } have also enclosed **conversational repair** in the UyLVs project (frog stories) in the orth tier, with the tag REP in POS. Domain: orth. E.g.: {meˀ} men.... (POS: REP PN1s)
- Square brackets [ ] in the tagset are an abbreviation place-holder for person agreement. For example, in *-GAn emes* (tagged in POS/ILG as PFV=[ ] XIPFVN), *-GAn* is followed by person agreement markers -men/sen/siz/la/0; thus, *-GAnmen emes* would be tagged PFV=1s1 XIPFVN.
- FOR indicates non-analyzed **foreign strings**. Domain: seg, pos, ILG.
- A slash / has (in the UyLVs project) marked line breaks. Domain: orth tier. Deprecated for the AMTO project.
- A pipe | has (in the UyLVs project) marked **pauses** in speech or text; in text, the following punctuation is assumed to correspond to a pause: , . ; : ! ? - (the hyphen only marks a pause if surrounded by whitespace (so e.g. *U mu'ellim iken - dédi Nesreddin Epeni* would be marked with | in the IPA tier, but *ata-ana* would not). Domain: IPA tier.
- An asterisk * precedes **ungrammatical** sentences (at least in the UyLVs project), e.g. * *U kelmidim*. Domain: orth tier. For ATMO, marking ungrammaticality is probably unnecessary, but if we were to mark it, it would be convenient to have a separate grammaticality element, which would be filled by default "Y" (yes - grammatical), which annotator could change to "*" (or "N" - ungrammatical) if needed.

7.2 Part of speech definitions

Here we define some major lexical categories for Chaghatay and Uyghur, grouped into nominals, verbs and verb-like categories, adjuncts, and interjections. Tags follow in parentheses, e.g. Noun (N). Definitions hold for chg and uig unless otherwise specified.

Nominals

Can take case suffixes and serve as head of NP; nouns and adjectives share many properties.

- **Noun (N)**: Takes nominal morphology (incl. case, plural, possessive, delimiter *=la* (the latter only occurs on nouns and numerals)); cannot host comparative *+rAK*; head of an NP; verbalized with *+lA*. Nominal, Adjectival predicates are negated by *e(r)mes* (uig: *U on yil burun muellim emes idi*. 'Ten years ago, s/he wasn't a teacher.') Subtypes:
  - Proper nouns (Npr), Toponyms (Ntop), Organizational Nouns (Norg) - not normally pluralized or possessable.

- **Adjective (AJ)**: Takes certain nominal morphology; also comparative. Nominalized with *+liK*. Generally describes a quality.
  - Tests: (1) (uig) Takes intensifier *+rAK* (2) (chg, uig) AJ N is NP, N AJ is AJL; (3) (uig) *teximu/eng* AJ (any AJ **-rAK*?) (4) (chg, uig) AJ-*dek* 'seem AJ-ish' *U mashina qizildek kördüm* 'That car seemed reddish.'
  - AJ used as N **must** have prior N in discourse, and generally take possessive *+(s)I*:
    - uig: (*Qaysi restaurantqa barimiz?*) *Yéqin**i**gha barimiz*. 'Let's go to the near (one).' [Noun elided].
    - uig: *Chong**i**ni alay!* "I'll take the big one" (Uyghur linguists see this AJ as as a N)

- **Pronoun (PN)**: Free. Takes person (s/p), number suffixes (1, 2, 3). Subtypes:
  - Personal (PN.P): incl. register (informal (i), formal/polite (f), honorific(h)).
  - Demonstrative (PN.DEM): based on *bu+, ol*.
  - Interrogative (PN.INTER): largely formed with *qa-* and *ne- (nä-)*
  - Indefinite (PN.INDEF): (1) formed from interrogative PNs; (2) semantically indefinite, e.g. uig: *palan, palanchi*
  - Reciprocal (PN.RECP): *öz*.

- **Numeral (NU)**: can take certain nominal morphology, including possessive (creating a partitive) and plural (creating collective nouns), and collective, e.g. uig: *Mahire **ikkeylen** bazarghe berip keldi* 'Mahire went to the market with him/her' (person in previous discourse). *Mahire **ikkeylen** bazargha bérip kel**duq***. "Mahire and I went to the market"
  - Can serve as determiner: *bir kishi* 'a/one person'
  - Can serve as predicate, rarely: *Kala besh* 'The cows are five.'
  - Subtypes: cardinal (NU, otherwise unmarked) ordinal (NU-ORD)

- **Quantifiers (QNT)**: Serve as determiners for Ns: *bashqa, bezi, eng, pütün, her, hich* etc. Can be compounded: *biraz, herbir*, etc.

- **Measures (M)**: between numeral and noun. Minor class, usually non-native, e.g. *tsun* 'inch', *sheng~shing* 'liter'. Temporary measures **(Nmeas)** often Turkic and formed with deverbal nominalizer *+(X)m*, e.g. *bir tutum tuz* 'a handful of salt'.

Verbal constructions:

- **Verb (V)**: takes verbal morphology (voice, TAM, abilitative, gerund etc):
  - Transitive (Vt), Intransitive (Vi).
  - Verbal constructions determine argument structure through valence and case assignment.

- Finite/Nonfinite. Non-finite verbs (Vnfin) take limited verbal morphology (derivational, voice, and limited TAM marking but no finite verb morphology or person marking, e.g. *U bolghan bol-sa*, 'If s/he were there', *bol-sa* Vi-COND only, no tense, aspect or person marking); finite verbs (Vfin) take full verbal morphology and person marking (*bol-di-m* Vi-ANT.DIR-1s with person marking).

- **Auxiliary (X)** - a verb-like element, prototypically copular and usually in finite position, which takes only limited verbal morphology and serves as a carrier for tense-aspect-mood (TAM) marking. Most auxiliaries may follow a participle or gerund (verbal noun) and in this case form a matrix clause. Auxiliaries differ from light verbs in that they do not follow converbs. (This definition differs from that of the traditional Turkology, which holds that 'auxiliaries' are what we would term 'light verbs'.) Under the current definition, there are two types of auxiliaries:
  - Copular auxiliaries: defective copular constructions from OT/Chg *er-* 'be' and *bol-* 'be, become'
    - *er-*:
      - Chg *é(r)-di* X-ANT.DIR-3 *é(r)-mish* X-INFR, cf. MSU *idi, imish*
      - Chg *eken*, MSU *iken*
      - chg *erur-* XIPFV *er-mish* X-INFR etc.
      - chg *ermes*, MSU *emes* XIPFVN, *emes-lik* XIPFVN-ABS
    - *bol-*:
    - (copula-like) *dur/tur(ur)*
  - Adjectival auxiliaries *kérek, zorur, lazim, mümkin*. (XAJ)
    - In MSU, these constructions are preceded by a nominalized (*-(X)sh* or *-mAq*) complement clause, e.g. *Bu ishni qilishim kérek.* 'I need to finish this task', cf. *Bu ishni qilishim kérek idi*. 'I was supposed to finish this task.' !!*Bu ishni eng awwal putturimiz zorur-rek~-dek turidu*. (*rek/dek* + *tur-* always) *lazim* (*-dek* ok). Such adjectival auxiliaries are negated by *emes* – which in turn can be followed by auxiliaries (e.g. needed).
    - In MSU, these auxiliaries can be used attributively with *+lXG* (XAJ-ABS): *Gülnar bilen Travis manga kéreklik ademler*.

- **Light Verb (LV)** - Syntactically, LVs are a closed set of full verbs forming a complex monoclausal predicate. Semantically, LVs modulate the meaning of the main verb. All light verbs may also function as full independent verbs. There are nominal and verbal light verbs.
  - **Nominal light verbs** (LVN) have the structure nominal - light verb, e.g. uig: *teyer qïl-* 'prepare' (from *teyer* 'readiness' + *qïl-* 'do').
  - **Verbal light verbs** (LVV) have the structure main verb - converb light verb, e.g. uig: *untu-p qal-* (to forget (with a lasting result)' cf. *untu-* 'forget', *qal-* 'remain'. Light verbs are fully inflectional as finite verbs (unlike auxiliaries), and they are not normally negated (unlike auxiliaries).
  - Light verbs are distinct from auxiliaries in syntax, morphology, and semantics.

Adjuncts - not closely related to predicate meaning; optional (i.e. not obligatory within a sentence).

- **Adverb (AV)**: precedes predicate; part of VP and "modifies" V; Disallows intervening material between AV and predicate, except for the uig intensifier clitic *=mu [bek=mu chirayliq*] (should find out if any other material allowed).

- **Postposition (POST, POSTP)**: Requires a (preceding) NP, some of which require case marking, and others which are free case.
  - There are two types, true and pseudo-postpositions:
    - True postpositions (POST): are primarily non-inflecting; are clitics. N=POST *bilen, burun, üchün, dek*, etc.
    - Pseudo-postpositions (POSTP): usually take case, often require GEN (should be marked as lexical words with whitespace) e.g. *ich, ust,* etc.
  - Segmentation of postpositions: N-*ning toghri-si-da* (and if *ich-i-de* is segmented, then *toghriside* should be segmented)

- **Preposition (PREP)**: precedes an NP, only Persian loans.

- **Conjunction (CONJ)**: lexemes coordinating clauses, such as *ve, ya, hem*. Orthographically surrounded by whitespace. also known as sentence connectives (Kornfilt 1997).

- **Particle (PRT)**: Not stress bearing, primary word accent precedes them (many not genuine clitics, they have no corresponding free versions, and have no full, independent lexical meanings). Typically cliticized to preceding clause, and may undergo vowel/consonant harmony according to the features of the preceding word.

- **Interjection (INTJ):** An expression of emotion, sentiment, or a pause-filler. Often a single word or non-sentence phrase; in MSU often preceded and followed by punctuation (esp. - and !, respectively). We take *exclamations* to belong to the same category (although we sometimes tagged INTJ as EXCL in UyLVs).

- **Complementizer (CZR)**: Follows a complement clause, which becomes the subject or object of the matrix clause, e.g. *ki, kim*; *dép.*

7.3 Tag set headers - column glosses

Headers (uig=Uyghur; chg=Chaghatay; MSU = Modern Standard Uyghur)

- **Infl_type**: Inflectional type (Infl [inflectional]; lex [lexical]; xref [cross-reference], transconv [transcription convention])
- **POS:** Part of Speech category (if inflectional, of stem; if lexical, of lexeme), e.g. N, V, Vfin (finite verb), etc.
- **Category**: grammatical category (e.g. voice, tense-aspect, participle, light verb, particle
- **type**: subordinate to Category (e.g. causative [voice], verbal [light verb], informal [person ending] etc.
- **subtype**: (optional) subordinate to Type (e.g. 2nd person singular [person ending], distal [demonstrative pronoun]
- **verbose** The category, type and subtype in verbose, human-readable prose.
- **gloss_Uy** The MSU term for this morpheme (only if a lexeme; incomplete; may not align with our current analysis (e.g. a *chetilma rewishdash* 'limiting adverbial' is what ATMO would call an aspectual/actional light verb). Very incomplete at present.
- **(neg)** For verbal affixes only: optional negation

- **seg** The segmentation type (blank=whitespace: word; - : affix; = : clitic)
- **morph_new** Proposed ATMO project morphological gloss (= UyLVs <pos> tier).
- **ilg_new** Proposed ATMO project ILG gloss (=UyLVs <ilg> tier).
- **pos_old** In the UyLVs project, the tag appearing in the <pos> tier.
- **ilg_old** In the UyLVs project, the tag appearing in the <ilg> tier if a grammatical morpheme, or if a lexeme, its English gloss. The grammatical morpheme is the same as in the <pos> tier (except for light verbs, which are marked LVN, LVV, or Vdirc in <pos>, and then with the capital-letter archiform in <ilg> for the first two, in an English gloss for the last. E.g.
  - *teyer#bol-* (POS: N#LVN; ILG: preparation#BOL)
  - *untu-p#qal-* (POS: Vi-CNV#LVV; ILG: forget-CNV#QAL)
  - *chiq-ip#bar-* (POS: Vi-CNV#Vdirc; ILG: emerge-CNV#out)
- **bdry** In the UyLVs project, the morpheme boundary. Changed in ATMO. (Possible values: whitespace, hyphen, equals sign.)
- **archiform** The canonical, phonemicized form of the morpheme, that in the UyLVs project appeared in the <seg> tier.
- **allomorphs** Allomorphs of the seg_form. Given exhaustively for MSU; some but undoubtedly not all non-standard modern and premodern variant forms given here. (Dwyer has been adding them as she encounters them, but they need to be systematically harvested.)
- **examples_chg_uig** Examples of these forms in phrases or sentences, in chg and MSU and nonstandard Uyghur (NS), if relevant.
- **gloss** If a lexeme, the English gloss
- **UyTxtb** The chapter number in which the form is discussed in the Engesæth, Yakup Dwyer 2009/2010 textbook, *Greetings from the Teklimakan* (Possible values: number (chapter number), n/a (does not appear), [blank] (not yet looked up, info incomplete))
- **Comments** (AMD) - on usage and forms in OT (Old Turkic), Chaghatay, and Modern Uyghur. Some comparative info on Turkish and the Turkish Treebank.
- **tagging pitfalls n tips** - tips for annotators when tagging similar morphemes
- **OT_form** - the form of the morpheme in Old Turkic (if absent then marked "n/a"; if unknown, then left blank)
- **Erdal04_pg** - the page number(s) in Erdal's 2004 *A Grammar of Old Turkic* (Brill) for the morpheme.
- **Chag_morph** The (equivalent) morpheme in Chaghatay (Possible values: alphabetical string (Chaghatay morph), n/a (does not appear), [blank] (not yet looked up, info incomplete))
- **Chag_glossing** Comments about the characteristics of this morpheme in Chaghatay
- **domain1** **chg** (is present in Chaghatay) **0** (is absent in Chaghatay) [blank] (not yet looked up, info incomplete)
- **domain2** **uig** (is present in modern Uyghur) **0** (is absent in modern Uyghur) [blank] (not yet looked up, info incomplete))

8 Morphotactics

Native Turkic words are head-final and suffixing; prefixing occurs in words of Persian origin.

8.1 Nominal suffix order

	N	PL	POSS	Case
example	*at*	*+lAr*	*Im*	*+GA*

In a series of nouns, only the last one will be inflected.

8.2 Verbal suffix order

Derivational suffixes generally precede inflectional.

finite

		voice	voice	voice	voice	infl	infl	infl	inf
	V	Refl	Recip	Caus	Pass	(Neg)	Abil	Tense	Person Endings
example		*tonu-*	*(I)sh*	*DUr etc.*	*ul*	*mA*	*(y)Ala*	*y* etc.	*men* etc.

nonfinite

		voice	voice	voice	voice			
	V	Refl	Recip	Caus	Pass	Abil	(Neg)	Converb or Gerund or NZR
example		*tonu-*	*(I)sh*	*DUr* etc.	*ul*	*(y)Al*	---	*p, GAn* etc.
							mA	*y, GAn*, etc.
							mas	*(lIK)*

A double causative is possible, e.g. *qil-dur-ghuz-; men ularni kör-üsh-**tür-güz**-düm*.

8.3 Person Endings (Verbal agreement suffixes)

Agreement is required on all finite verbs, with the exception of *-GAn* (originally a Verbal N, still termed a *süpetdash* 'adjectival' in Uyghur linguistics), e.g. *körgen emes.* There are two main paradigms for subject agreement suffixes on finite verbs, pronominal and possessive types, which are derived from personal pronouns and nominal possessive suffixes, respectively. A third paradigm is only used for imperative forms.

- **Paradigm Type I: Pronominal type** (a.k.a. *Z*-series)
  - Stress: prestressing
  - Widest distribution: all simple tenses (except the definite anterior): present progressive, imperfective (aorist), indirect (reported, unwitness) perfective, future, necessitative and with the copula (as a nominal, adjectival, or participial). (This list may not be not exhaustive!)
- **Paradigm Type II: 'Possessive' type** (a.k.a. *K*-series)
  - Stress: stressable
  - Used with: Direct anterior (perf.), Conditional, Projection participle *-GU*

- **Paradigm Type III: Imperative type**
  - Stress: stressable
  - Used with: Imperative (2nd person) and volitional/hortative (1st, 3rd person).

In the second person, Chaghatay agreement suffixes, like its personal pronouns, follow the Old Turkic pattern of a single register, distinguishing for number only: agreement suffixes for the second person singular vs. plural. Modern Standard Uyghur, in contrast, distinguishes three different registers of second person personal pronouns and verbal agreement suffixes: singular (*birlik*) informal (2si, known in MSU as *addiy türi*), singular formal (2sf, *sipaye türi*), and singular honorific (2sh, *hörmet türi*); plural informal (2pi), plural honorific (2ph), and plural deferential (*setlime türi*).

The two sets of verbal agreement suffixes for MSU and Chaghatay (tags follow in parentheses):

Chaghatay (Eckmann 1966:152-3) *verbal agreement paradigm* (w/vowel *X* for C-final stems)

	Type 1	**Type 2**	**Type 3** (VOL/IMP)
1.sg.	-men (1s1)	-(X)m (1s2)	-(A)y(In) (1s.VOL)
2.sg.	-sen (2s1)	-(X)ng (2s2)	-∅ (2si.IMP) -gïl~gïn (2si.IMP)
3.sg.	-∅ (3) ~ -Dur (3) ~ -Dur-ur (3-IPFV)	-∅** (3)	-sun (3s.VOL), -dek (3s.VOL)
1.pl.	-biz (1p1)	-(X)q/k (1p2)	-(a)lï(ng) (1p.VOL)
2.pl.	-siz (2p1) -sizler (rarely) (2p1)	-(X)ngïz* (2**p**2) ~ -(X)nglar (2p2)	-(X)ng (2p.IMP) -(X)nglar (2p.IMP)
3.pl.	-(Dur)lar (3p)	-lArï, -sïlar (3p2)	-sunlar, -dekler (3p.VOL)

*Eckmann 1966:152 has *-nguz* / **Eckmann has *-(s)ï* and (for 3 pl) *-sïlar* "only for the categorical future"

While the modern Uyghur paradigm closely resembles that of Chaghatay, some Chaghatay plural forms have been redeployed as singular polite forms, such as OT/chg 2nd person *plural* formal *-ngiz* (2**p**f2) corresponding to MSU *singular formal/polite* 2nd person (2**s**f2). Similarly, *-lArï* is a *third* person *plural* form in Chaghatay (3**p**2), and (as *-liri*) a *singular honorific* 2nd person form in MSU (2**sh**2).

Modern Standard Uyghur:

	Type 1	**Type 2**	**3** (volitional VOL/ IMP)
1.sg. (1s)	-men (1s1)	-(I)m (1s2)	-Ay (1s.VOL)
2.sg. (2si) (2sf) (2sh)	-sen (2s1) -siz (2sf1) -la (2sh1)	-(I)ng (2s2) -(I)ngiz (2**s**f2) -liri (2sh2) -silA (2sd)	-∅ (2si.IMP) -gïl~gïn (2s.IMP)
3.sg. (3)	-∅ (3)	-∅ (3)	-sun (3VOL)
1.pl. (1p)	-miz (1p1)	-(I)q/k (1p2)	-(y)Aylï (1p.VOL)
2.pl. (2p) (2p(f)) (2ph)	-siler (2p1) -sizler (rarely) (2p1)* -la (2ph)	-(I)nglAr (2pi2) -(I)ngizlAr (rarely) (2pf2) -la (2ph)	-(I)ng(lar) (2**p.IMP**) -(I)ngiz(lar) (2pf2)
3.pl. (3)	∅ (3)	∅ (3)	-sun (3VOL)

*variant of siler, not necessarily formal (but formal for Turfan dialect (Tohti 1986 ms.))

- COND : singular **-sili** and the plural **-singizlar**.

coll. Numeral

1p1.COLL -(X)miz (AD: change in chart, right now "1p1COLL")
2pi.COLL -(X)ngiz (AD: change to 2pi2)

And add: 2pf2.COLL -(X)ngizlar
Cf also -Eylan COLL , PN.INDEF.COLL birev, -ev COLL, COLL(-POSS3) -la(si) (AD: change to (-3POSS))

9 References

AATT=American Association of Teachers of Turkic (Dwyer, Arienne, Gilson, Erika, Kornfilt, Jaklin, Onder, Sylvia, eds.). 2004. *Draft of Turkic Terminology*. Unpublished manuscript.

Boeschoten, Hendrik. 1998. Chaghatay. In Lars Johanson and Eva-Agnes Csato (eds). *The Turkic Languages*. Routledge. [NB: most examples are from Eckmann, without attribution.]

Dwyer, Arienne M. 2000. DOBES linguistic markup scheme: Towards a Minimal Annotation Standard for Encoding Linguistic Information. Unpublished DOBES technical paper. November. (referenced in Wittenburg, Peter, Ulrike Mosel, and Arienne Dwyer. 2002. *Methods of Language Documentation in the DOBES project*. Online: http://www.mpi.nl/lrec/2002/papers/lrec-pap-02b-dobes-talk-final.pdf)

Eckmann, János. 1966. *Chagatay Manual*. Bloomington: Indiana University Uralic and Altaic Series. (Vol. 60).

Engesæth, Tarjei, Mahire Yakup and Arienne Dwyer. 2009/2010. *Teklimakandin Salam: hazirqi zaman Uyghur tili qollanmisi / Greetings from the Teklimakan: a handbook of Modern Uyghur.* Lawrence, Kansas: University of Kansas ScholarWorks. ISBN 978-1-936153-03-9. Online: http://kuscholarworks.ku.edu/dspace/handle/1808/5624.

Friedrich, Michael (with Abdurishid Yakup). 2002. *Uighurisch Lehrbuch*. Wiesbaden: Dr. Ludwig Reichert Verlag.

Hahn, Reinhard (with Ablahat Ibrahim). 1991. *Spoken Uyghur*. Seattle: University of Washington Press.

Ide, Nancy, Laurent Romary, and Eric de la Clergerie. 2004. International Standard for a Linguistic Annotation Framework. DOI:10.3115/1119226.1119230, online: http://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf

Kornfilt, Jaklin. 1997. *Turkish.* London: Routledge.

Lehmann, Christian. 2004. Interlinear morphemic glosses. In Booij, Geert, Christian Lehmann, Joachim Mugdan, and Stavros Skopeteas (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband*. Berlin: de Gruyter. Online: http://www.folialinguistica.com/documents/Interlinearmorphemicglossing.pdf

Lieb, Hans-Heinrich and Sebastian Drude. 2000. Advanced Glossing, a Language Documentation Format. Unpublished DOBES working paper. Online: http://www.mpi.nl/DOBES/documents/Advanced-Glossing1.pdf

MPI-EVA Department of Linguistics and University of Leipzig Department of Linguistics. 2008-2015. Leipzig Glossing Rules. Online: http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Oflazer, Kemal, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish Treebank. Online: http://www.andrew.cmu.edu/user/ko/downloads/Papers/TurkishTreebank-Chapter.pdf (better Turkish Treebank ref needed)

Osmanov, Mirsultan. 1990. *Hazirqi zaman Uyghur tili di'alektliri*. Ürümchi: Shinjang Yashlar-Ösmürler neshriyati.
Schönig, Claus. 1997. *Finite Prädikationen und Textstruktur im Babur-name.* Wiesbaden: Harrassowitz Verlag.
Schultze-Berndt, Eva. 2006. Linguistic Annotation. In Gippert, Jost et al. *Essentials of Language Documentation.* Berlin: Mouton de Gruyter, pp. 213-252.
Sugawara Jun and Aysima Osmanov. 2007. *現代ウイグル語接辞索引*. Tokyo: Institute of the Languages of Asia and Africa.
Tömür, Xemit. 1987. *Hazirqi zaman Uyghur tili grammatikisi, morfologiye*. Beijing: Minzu.(=Tömür u). English translation 2003 [tr. Anne Lee]. *Modern Uyghur grammar: morphology*. İstanbul: Yıldız (=Tömür e).