[Overleaf] - https://www.overleaf.com/2578565887mkmtvjbrpnyk#b27c61
Paper: Copula, Turklang2025 - [Association for Computational Linguistics (ACL) conference]

Phenomena

ң, ү, ө.

Bare predicates

Kyrgyz: Айша — доктур. (*Ayşe is a doctor.*) *link*
Tatar: Айшә — табиб.
Sakha: Айша — доктор.
Turkish: Ayşe doktor.
Uzbek: Oyisha – doktor.

Kyrgyz: Бул ким? (*Who is this?*)
Kyrgyz: Бул мен, Мурат. (*It's me, Murat.*)

Azerbaijani: Deniz harda? (Where is Deniz?)
Kyrgyz: Дениз кайда? (Where is Deniz?)
Uzbek: Deniz qayerda? (Where is Deniz?)

Azerbaijani: Deniz evdə. (Deniz is at home.) Link
Kyrgyz: Дениз үйдө. (Deniz is at home.)
Uzbek: Deniz uyda. (Deniz is at home).

Negated bare predicate (değil/эмес)

- The fruits are bad.
- I know the fruits are not good.

Turkish: Meyvelerin güzel olmadığını bilmiyorum.

meyve-ler-in	güzel	ol-ma-dığ-ı-nı	bil-iyor-um
fruit-PL-GEN	good	be-NEG-VN-POSS.3-ACC	know-PRES-1SG

Kyrgyz: Дүкөндөгү жемиштер жакшы эмес экенин билем.

жакшы же жакшы эмес экенин билбейм.
Azerbaijani: Bazarda mivə var amma yaxçı dəyil. Link

Kyrgyz: Дүкөндө жемиш бар, бирок жакшы эмес. (*There's fruit at the store, but it's not nice.*) *link*

Tatar: Кибеттә җимеш бар, ләкин яхшы түгел.

Sakha: Маҕаһыыҥҥа фрутка баар, ол гынан баран үчүгэйэ суох. (*There's fruit at the store, but it's not nice.*)
Sentence #9 in link
Oh... is it then: Маҕаһыыҥҥа фрутка баар, ол гынан баран үчүгэй буолбатах?

Uzbek: Bozorda meva bor, ammo yaxshi emas.

Azerbaijani: Ayşə düktürdü(r). Link
Kyrgyz: Айша — доктурдур. (*Ayşe is a doctor.*)
Tatar: Айшә — табибдыр. (an inferential meaning 'Ayşe is apparently/probably a doctor.'; there are possibly also dialectal differences)
Turkish: Ayşe doktordur.
Uzbek: Oyisha doktordir.

Kyrgyz: Бул менмин, Муратмын. (*It's me, Murat.*)
Sakha: Бу мин, Мураппын.
Uzbek: Bu menman, Murotman.

Sentences with forms of defective verb i-/ə-/etc.

Azerbaijani: Böyük nənəsı pərəstarıdı. *(His/her grandmother was a nurse.)* Link
Kyrgyz: Анын таенеси мед айым эле. (*Their grandmother was a nurse. Also: Their grandmother is just a nurse.*)
Tatar: Әбисе шәфкать туташы иде.
Äbise şäfqät tutaşı ide.

Sakha: Кинилэр эбээлэрэ медсестра этэ. (*Their grandmother was a nurse.*)
Sentence #6 in link

Uzbek: Buvisi hamshira edi.
Kyrgyz: Дениз үйдө эле. (Deniz was at home.)

	1st person	2nd person	3rd person
copula эле (“was”)	Мен студент элем.	Сен студент элең.	Ал студент эле.
adverbial эле (“just”)	Мен студент элемин.	Сен студент элесиң.	
auxiliary эле	Мен студент болот элем.	Сен студент болот элең.	Ал студент болот эле.

Tatar: Дениз өйдә иде.
Deniz öydä ide.

Uzbek: Deniz uyda edi.

Sakha: Дениз дьиэҕэ (баар) этэ.

Sentences with forms of "become" verb (in "be" meaning) бол-/ol-/буол-

Kyrgyz: Дениз доктур болот, билип кой! (*Deniz will be/become a doctor, know this!*)

Kyrgyz: Дениз доктур болот. (*Deniz will be/become a doctor.*)
Tatar: Айшә — табиб була. (CHECK THAT THIS IS "BE" MEANING)
Turkish: Deniz doktor olacak. (*Deniz will be/become a doctor.*)

Kyrgyz: Дениз үйдө болчу. (*Deniz was at home.*)
Uzbek: Deniz doktor bo'ladi. (Deniz will become a doctor)

Sentences with forms of "become" verb (in "become" meaning) бол-/ol-/буол-

Azerbaijani: Deniz düktür olacax, bilmiş ol! *(Deniz will be a doctor, know this!)*
Link

Sakha: Дениз доктор буолуо, ону бил! (*Deniz will be a doctor, know this!*)
Sentence #10 in link
Kyrgyz: Дениз доктур болот, билип кой! (*Deniz will be/become a doctor, know this!*)
Tatar: Дениз табиб булачак, бел шуны!

Kyrgyz: Дениз доктур болот. (*Deniz will be/become a doctor.*)
Tatar: Айшә — табиб була. (CHECK THAT THIS IS "BECOME" MEANING)

Uzbek: Deniz doktor bo'ladi. (Deniz will become a doctor)

Sentences with forms of "become" verb (in "happen" meaning) бол-/ol-/буол-

Kyrgyz: Бир нерсе болду/болуптур. (*Something seems to've happened.*)

Turkish: Bir şey oldu.

Uzbek: Bir narsa bo'ldi/bo'libdi.
оддо
Embedded copula: verbal adverb forms

Kyrgyz: Өсүмдүк узун болуп өстү. (*The plant grew tall.)*

Embedded copula: verbal noun forms of бол-/ol-/буол-

Azerbaijani: Deniz tətil olduğunu yadınnan cıxardıp, mədrəsəyə getmişdi. (*Deniz (evidentially) went to school (because) she/he had forgotten that it was holiday.)*
Link

Tatar: Дениз каникул (таътил) булганны онытып мәктәпкә баргандыр.
Deniz kanikul (ta'til) bulğannı onıtıp mäktäpkä barğandır. (haven't been confirmed by a native speaker)
Uzbek: Deniz ta'til bo'lganini unutib, maktabga ketibdi. (non-firsthand information)

Embedded copula: verbal noun forms of э-/i-

Kyrgyz: Дениз каникул экенин унутуп мектепке барыптыр. (*Deniz apparently went to school forgetting that it was a holiday.*)

Tatar: Дениз каникул (таътил) икәнен онытып мәктәпкә баргандыр.
Deniz kanikul (ta'til) ikänen onıtıp mäktäpkä barğandır. (haven't been confirmed by a native speaker)
Uzbek: Deniz ta'til ekanini unutib, maktabga boribdi. (non-firsthand information)

What's the line between copula forms of verbal nouns and forms that we treat as just verb forms?

Kyrgyz: Китеп жазуудамын. (hypothetical/fancy)
Китеп жазган жокмун.
Китепти бүтүрө турган болом.
Китепти бүтүргөн болом.
Китеп бүтүргөнмүн.
Бүтүрсө болот/болбойт.
Бүтүрүүгө болот/болбойт.
Бүтүргөнгө болот/болбойт.
Turkish: Kitap yazmaktayım.
Kitap bitirmek üzereyim.
Yapıyor olacağım.
Bitirmiş olacağım.
Olur. / Olmaz.
Uzbek: Kitob yozmoqdaman.
Kitob yozganim yoʻq.
Kitobni bitirgan boʻlaman.

Annotation questions

1. i-/e- vs. bol-/ol- (become *sometimes*)

*Deniz evde **ol**acak.* (Deniz will be at home.) *Deniz evde.* (Deniz is at home.)
Deniz ev-de ol-acak Deniz ev-de

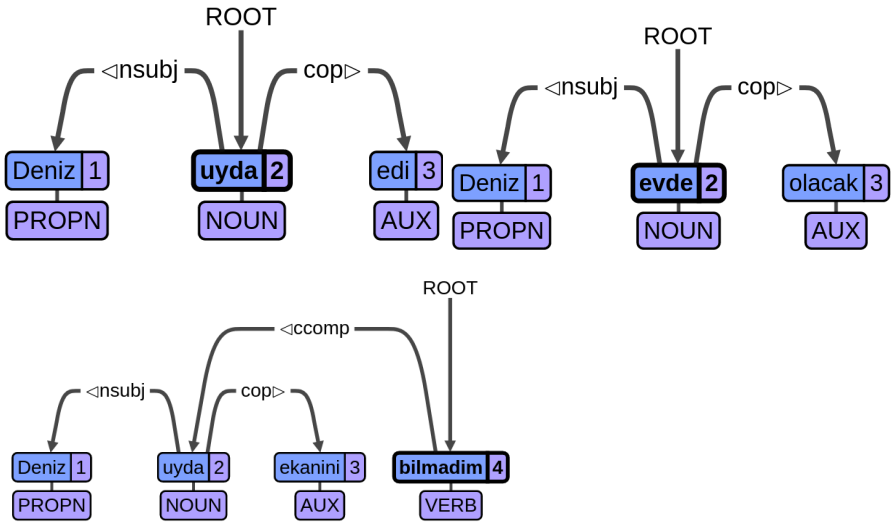
PROPN house-LOC be-FUT PROPN house-LOC

*Bir şey **ol**du.* (Something happened.)
bir şey ol-du
one thing be-PST

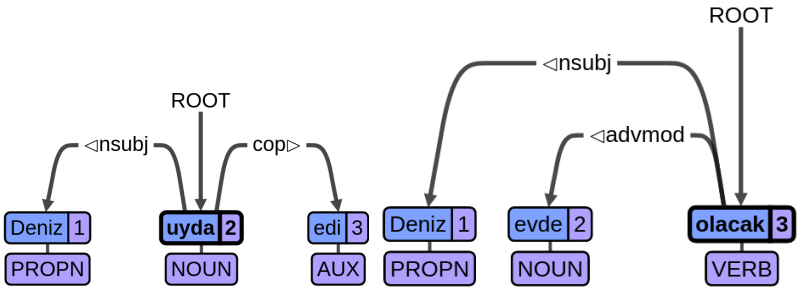
Дениз(дин) үйдө экенин билген жокмун.
*Deniz'in evde **ol**duğunu bilmiyorum.* (I don't know that Deniz was at home.)
Deniz-'in ev-de ol-duğ-u-nu bil-m(A)-iyor-um
Deniz(-GEN) house-LOC COP-VN-POSS.3-ACC know-NEG-PRES-1SG

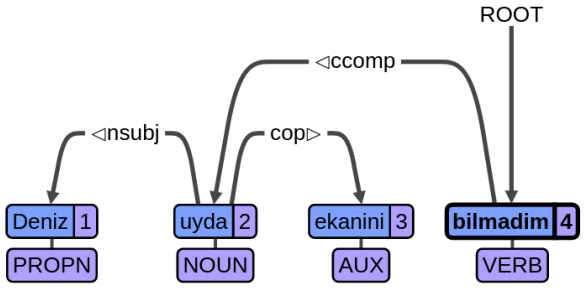
Four options:

1. i-/e- is separate "copula", бол-/ol- is separate copula (ALL COP OPTION - POLICE STATE)

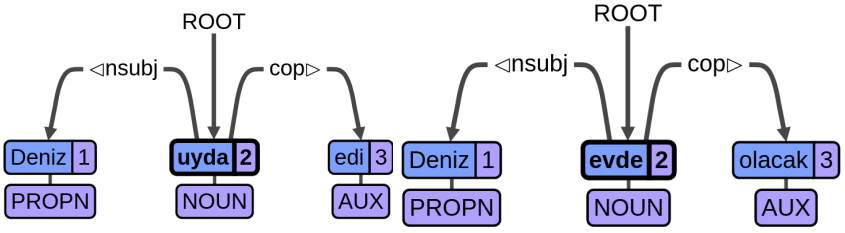


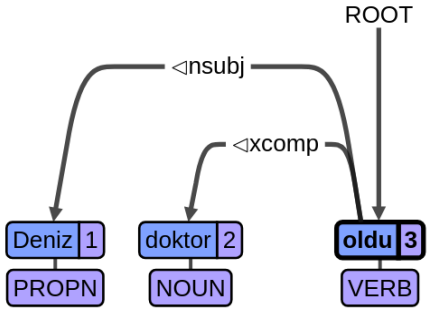
2. i-/e- is separate "copula", бол-/ol- is distinct verb (Jonathan's & Nikolett's approach) → let's go with this one

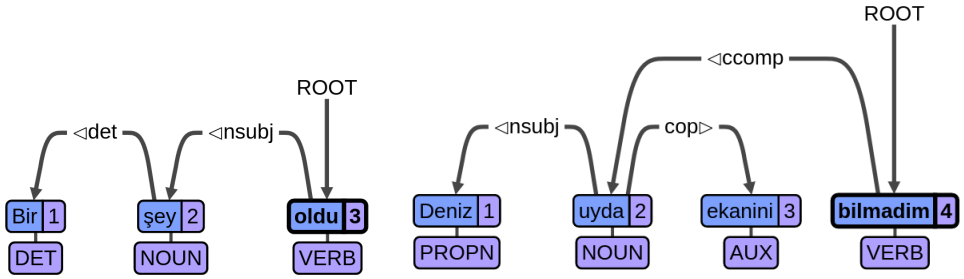




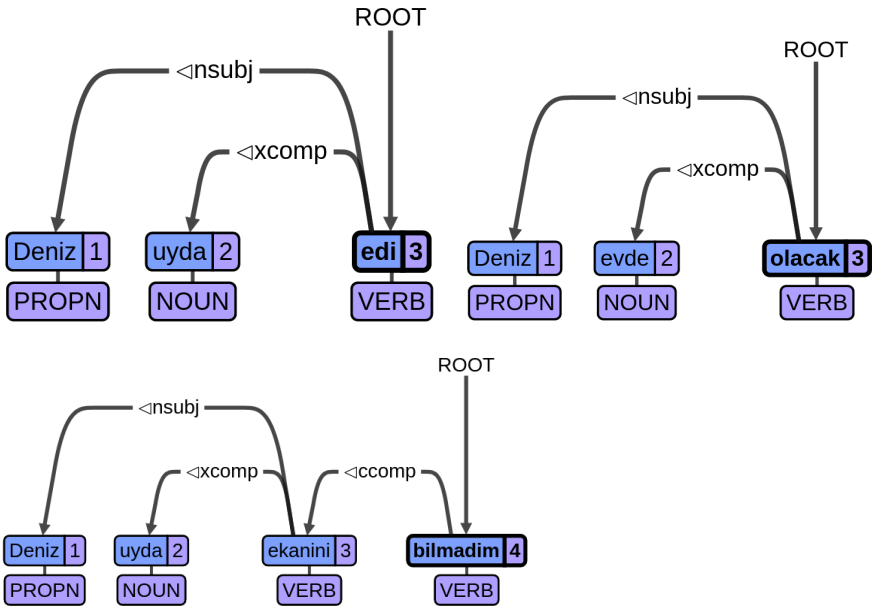
3. i-/e- is separate “copula”, бол-/ol- is distinct verb depending on whether it means “be” or not → (2025-02-21) **choose this with fallback of #2 when annotator is not sure** (but not necessarily all cases of ambiguity)





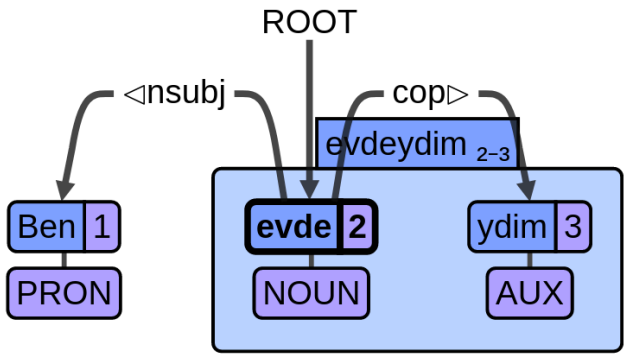


4. i-/e- is distinct verb, бол-/ol- is distinct verb (NO COP OPTION - DEFUND THE POLICE)



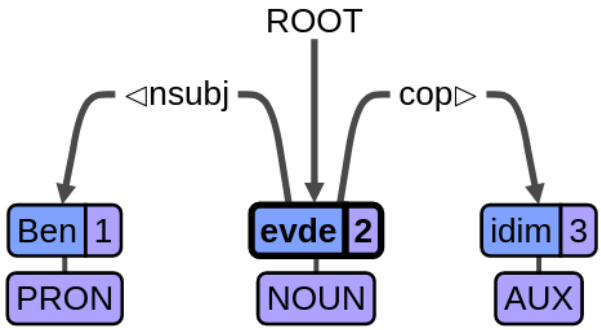
- Re 4, when copula is null (3(sg) ~present) having it as ROOT would be a problem
- Re 3, it's hard to know what was intended, and interpretation may rely too much on translations into other languages
- Re 1, ol-/бол- seems like a lexical verb that sometimes behaves as a copula, easier to treat as a lexical verb all the time (see "Re 3")
- Re 2, easy to annotate, and consistent; theoretically speaking, the copula is treated as a supportive element (although this becomes not ideal when embedded)

2. Tokenisation

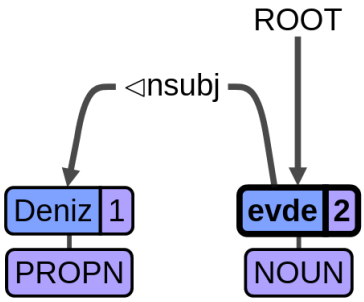


1 Ben _ PRON _ _ 2 nsubj _ _
2-3 evdeydim _ _ _ _ _ _ _ _
2 evde _ NOUN _ _ 0 root _ _
3 ydim i AUX _ _ 2 cop _ _

Ben evdeydim. (I was at home.)
ben ev-de=y-di-m
me house-LOC=COP-PAST-1SG

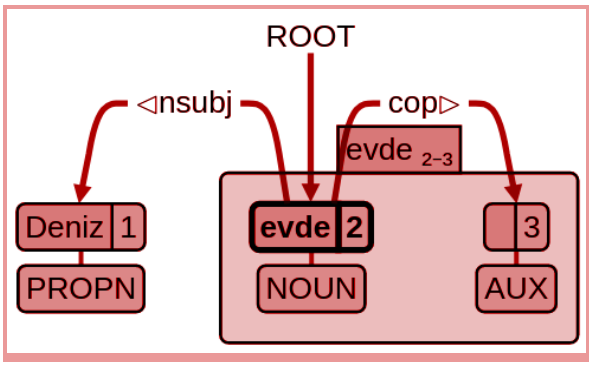


Ben evde idim. (I was at home.)
ben ev-de i-di-m
me house-LOC COP-PAST-1SG



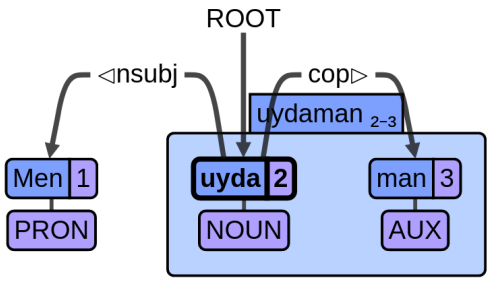
1 Deniz _ PROPN _ _ 2 nsubj _ _
2 evde _ NOUN _ _ 0 root _ _

Deniz evde. (Deniz is at home.)
Deniz ev-de
PROPN house-LOC



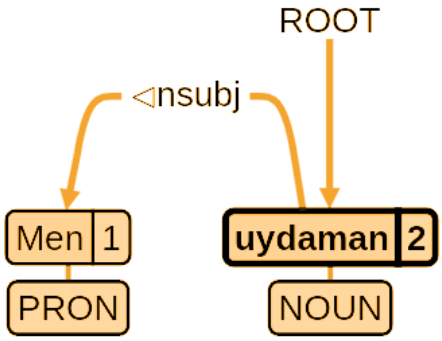
CAN'T DO THIS ^
No empty tokens
Can just leave out token 3

Preferable:



1 Men _ PRON _ _ 2 nsubj _ _
2-3 uydaman _ _ _ _ _ _ _ _
2 uyda uy NOUN _ _ 0 root _ _
3 man e AUX _ _ 2 cop _ _
→ This is our preferred approach

Undesirable:



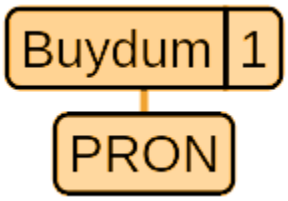
Ben evdeyim. (I'm at home.)
ben ev-de-yim
me house-LOC-[COP.PRES]1SG

ben ev-den-im
me house-ABL-[COP.PRES]1SG

ben ev-den-miş-im
me house-ABL-[COP]PST-1SG

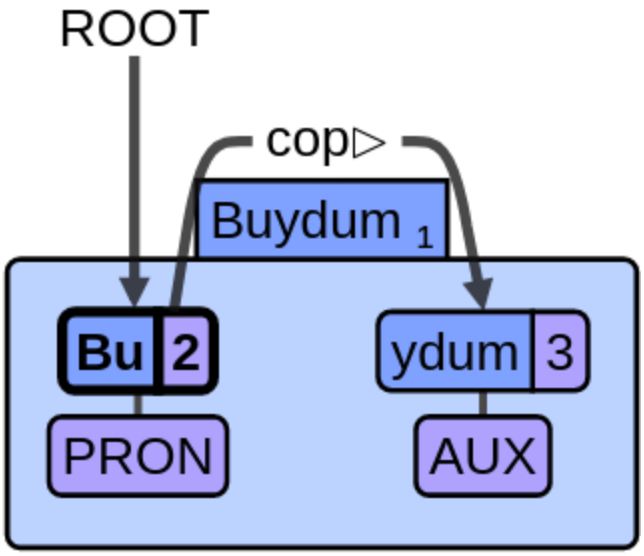
ben ev-den i-miş-im
me house-ABL COP-PST-1SG

2.
Pers=3
Num=1
Pers=1
Num=1



Pers=3
Num=1
Pers=1
Num=1
Tense=Past

It was me.
Bu-y-du-m
this-COP-PAST-1SG



→ This is our preferred approach

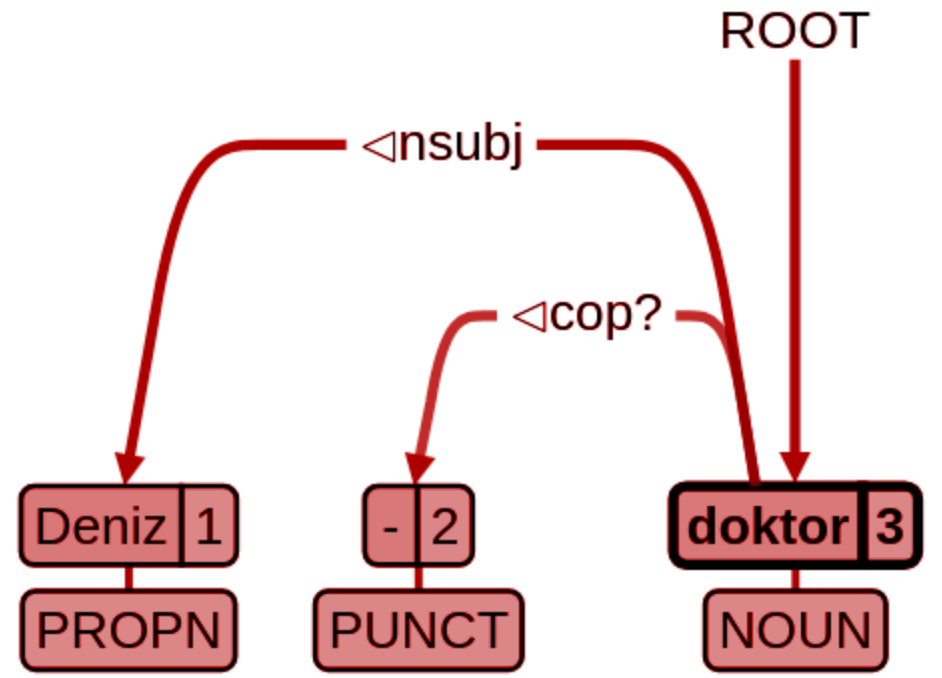
1. Keeps structures parallel
2. Avoids tense(etc.) features on nouns
3. Avoids conflicting person/number/etc. Features
4. Treats it syntactically as two words (“subwords”)

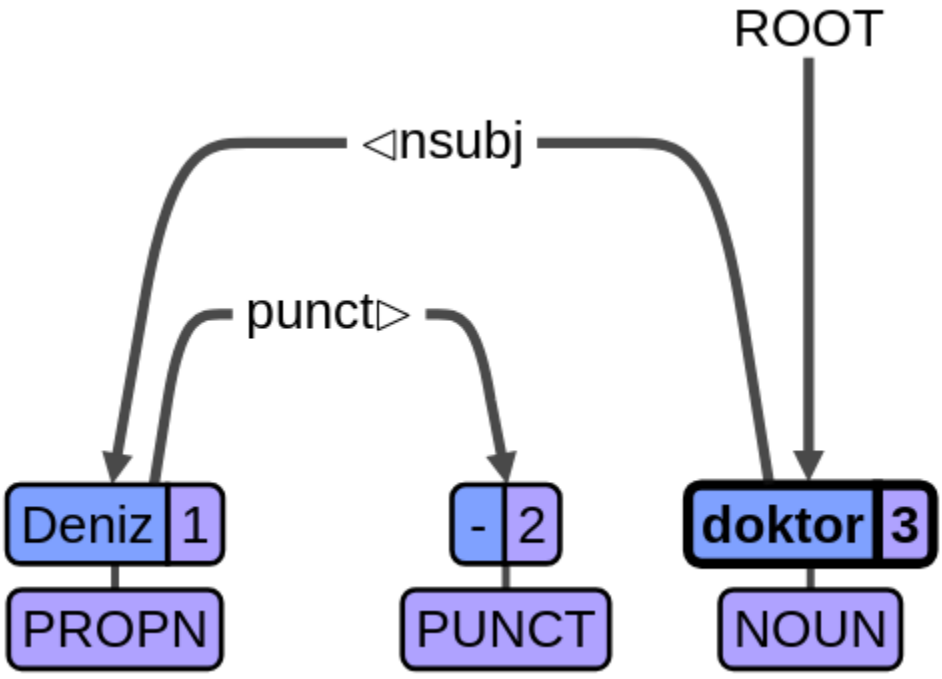
3. Dash/hyphen

—

- There's a sense that this indicates copula, although only orthographically (maybe including prosodic pause, as with other punctuation)
- However, it seems like more of an orthographic convention to help with reading than a syntactic element indicating copula (cf. "I like cooking, family and friends.").
- Even in Russian when — is used in copula sentences, the actual word indicating a copula (это) can also be used (" X — это Y"). You wouldn't want to annotate both as `cop`.

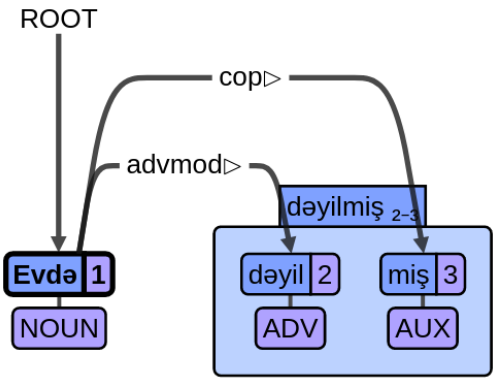
Thus, `punct` is preferred.

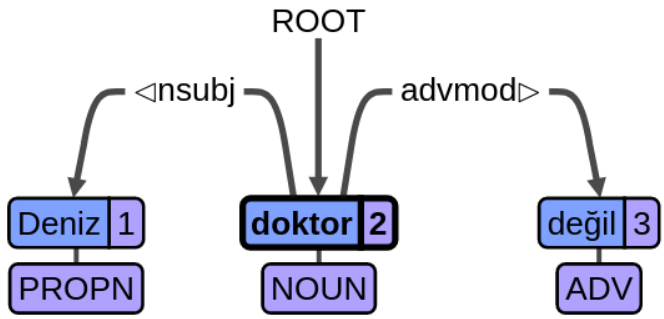




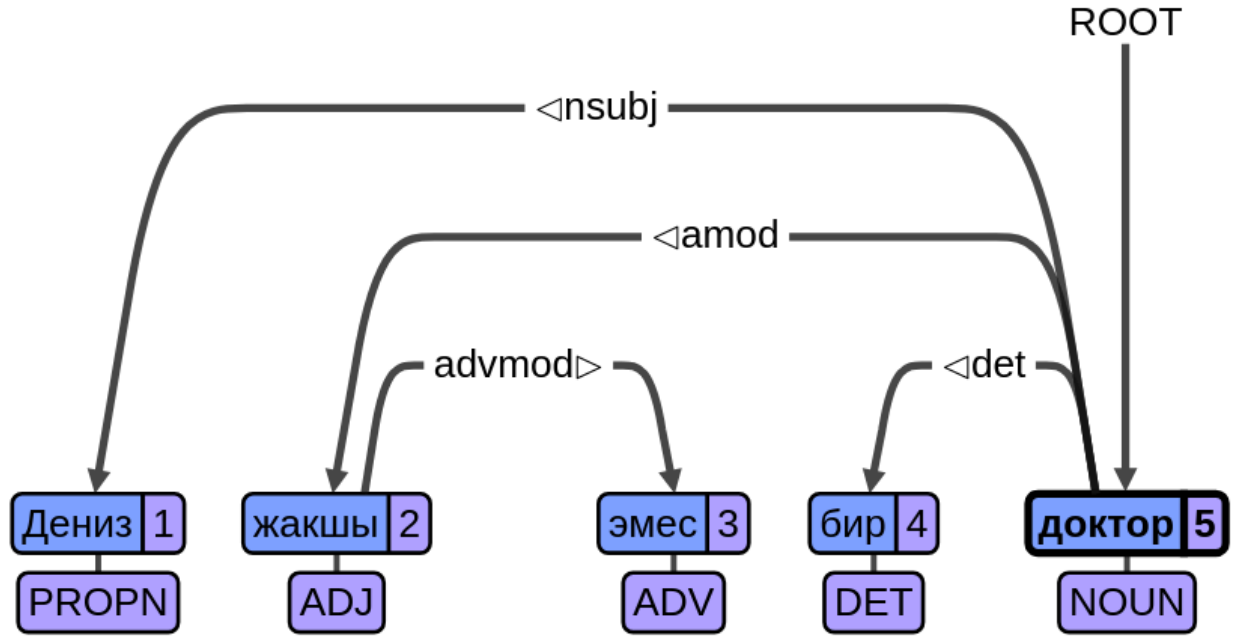
4. Değil/эмес

His/her health was not well.
Sağlığı iyi değil idi.
Ден соолугу жакшы деле эмес болчу.



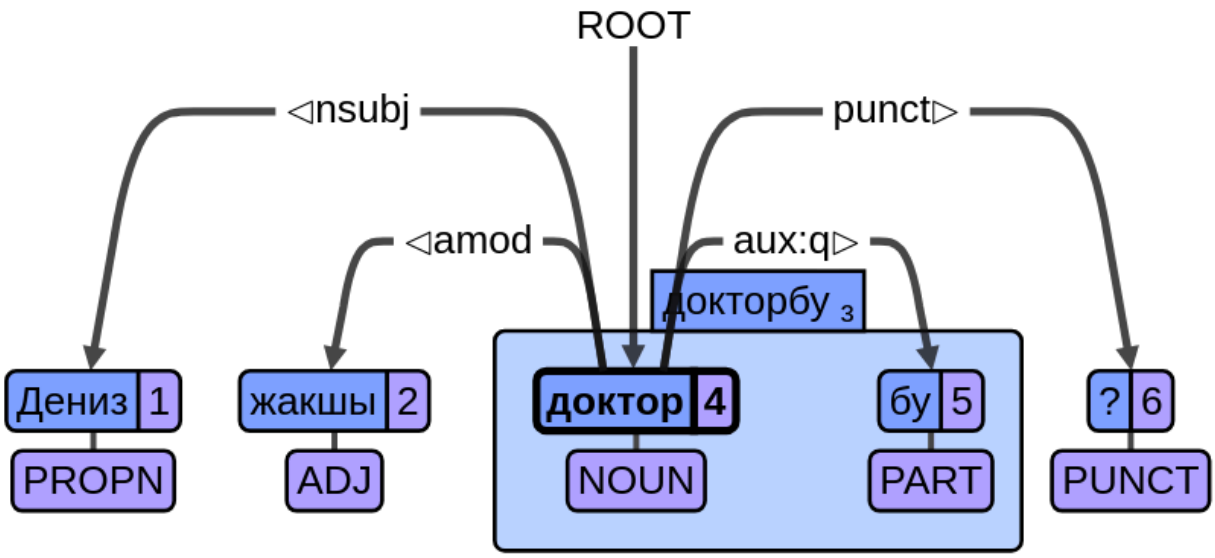


Okay in some languages (Kazakh and Kyrgyz, but not Turkish or Uzbek):

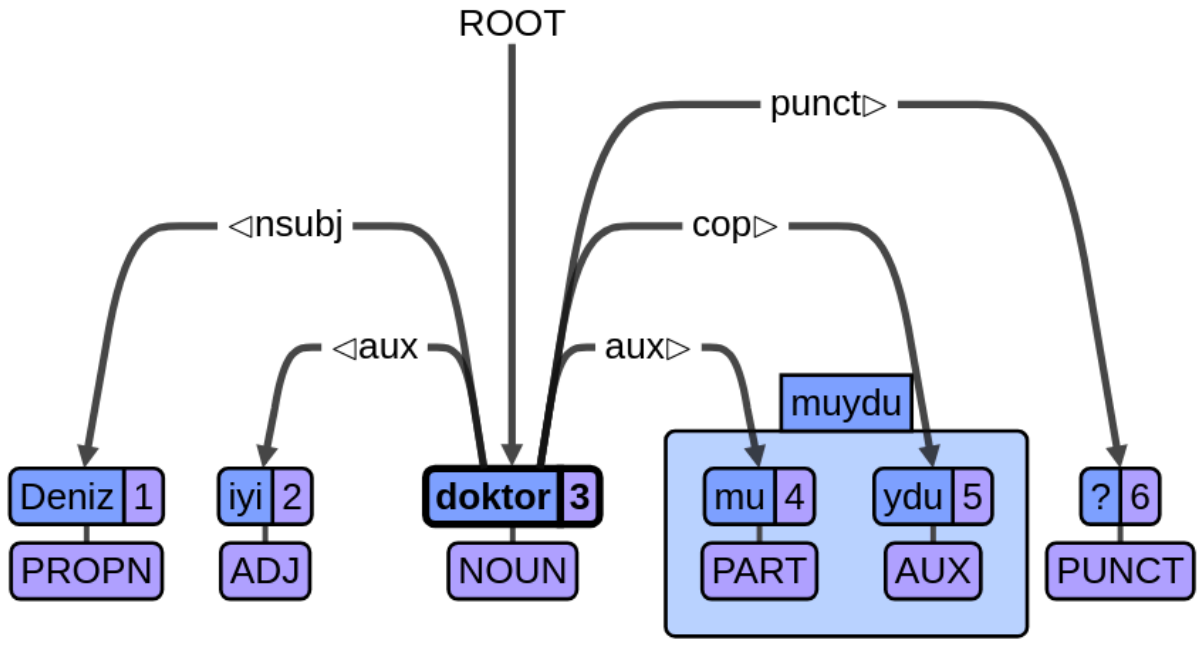


“Deniz is a not good doctor.”

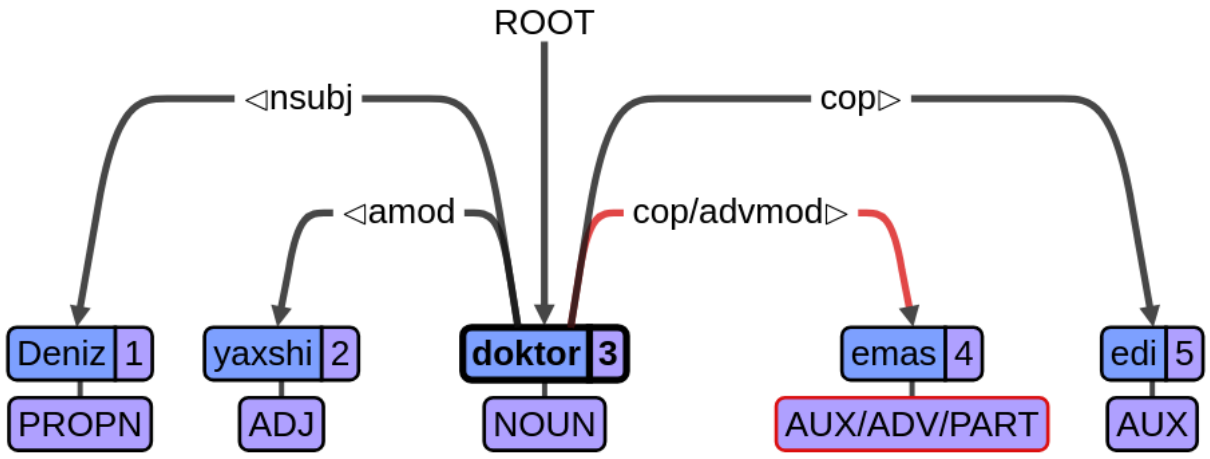
(What PART is for:



)

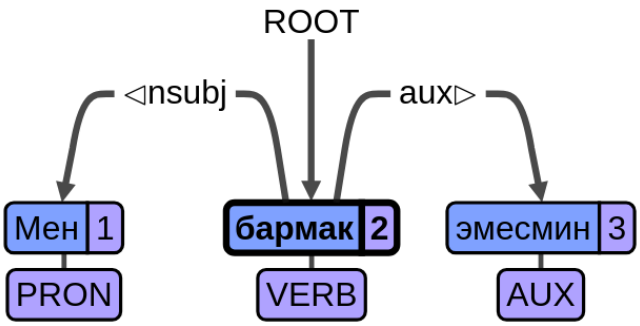


Other considered approaches with emas/deyil:



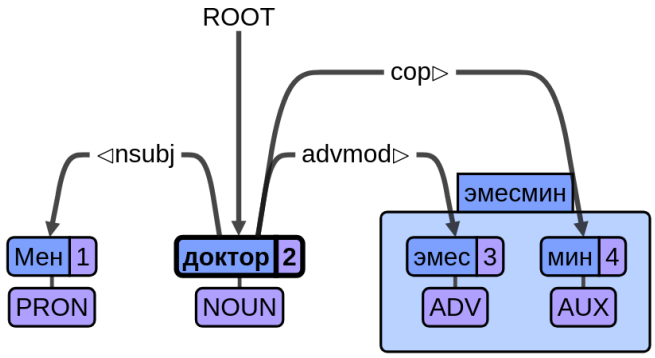
Emes as a negator of finite verb forms:

Treat them as auxiliary?

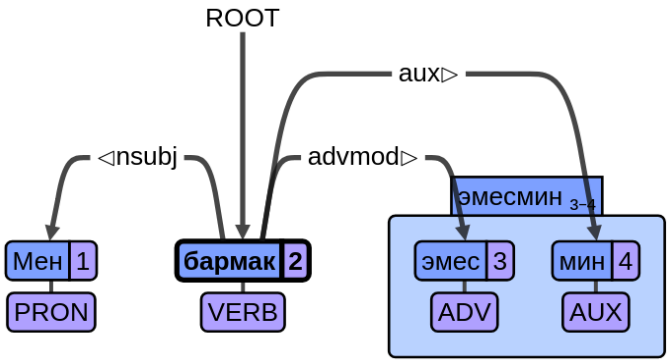


Better than ADV with agreement

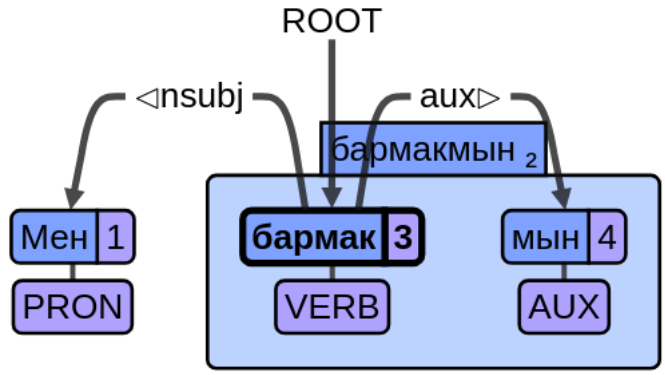
Similar to Uralic negative auxiliaries. (Those take different tenses; эмесмин looks like ADV + COP, e.g. the following, like “isn’t”):



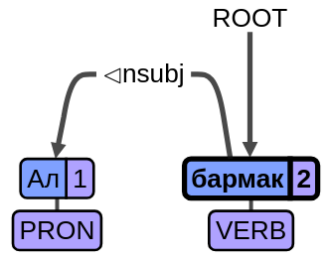
So a better approach could be this:



“I wouldn’t have gone”
By keeping a similar structure, we can keep it parallel:
(aff: Мен бармакмын. “I would’ve gone”)

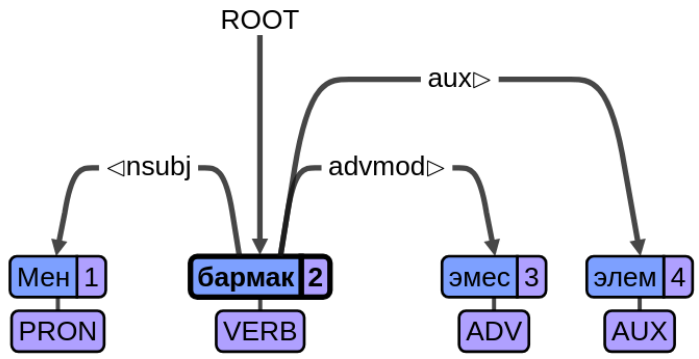


For 3rd person we want to avoid an empty token:



“He/she would’ve gone.”

Means that when there’s a copula used as auxiliary that’s a separate word, it has the same structure too:



5. Non-finite verb forms with copula

verbal noun with locative case and copula

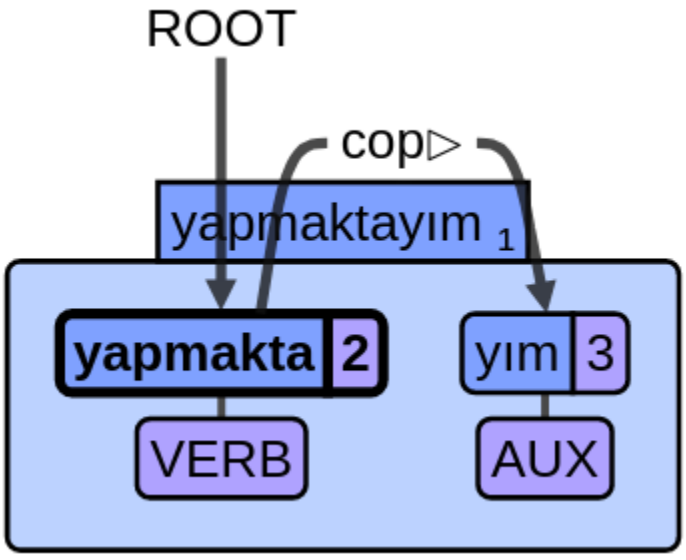
- yapmaktayım (I'm in the process of doing ...), gerçekleştirilmektedir (it's being realized)
- кылуудамын, пландалууда, мерчемделүүдө, үйрөнүүдөмүн
- қылудамын, пландалуды, ..., үйренудемін
  → these we could analyse as VN.loc + copula (with 2 subtokens)

  Lemma=yap, Form=VerbalNoun, Case=Loc

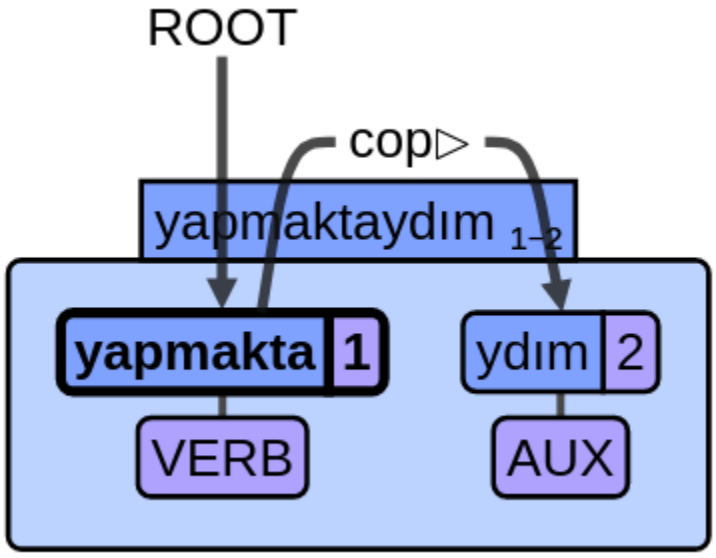
non-finite forms with copula

- yapmış olacağım (I'll have done ...), yapıyor olacağım (I'll be doing ...)
- кылганмын (I have done), кылган болом (I will have done (?)), кылган болчумун (I had done (?)), кыла турган болом (I will prepare to do (??)), кыла турган болдум (I decided to do)
- жүдөтүп бүтмөй болду
  → let's just say these are all auxiliary constructions and forget about it

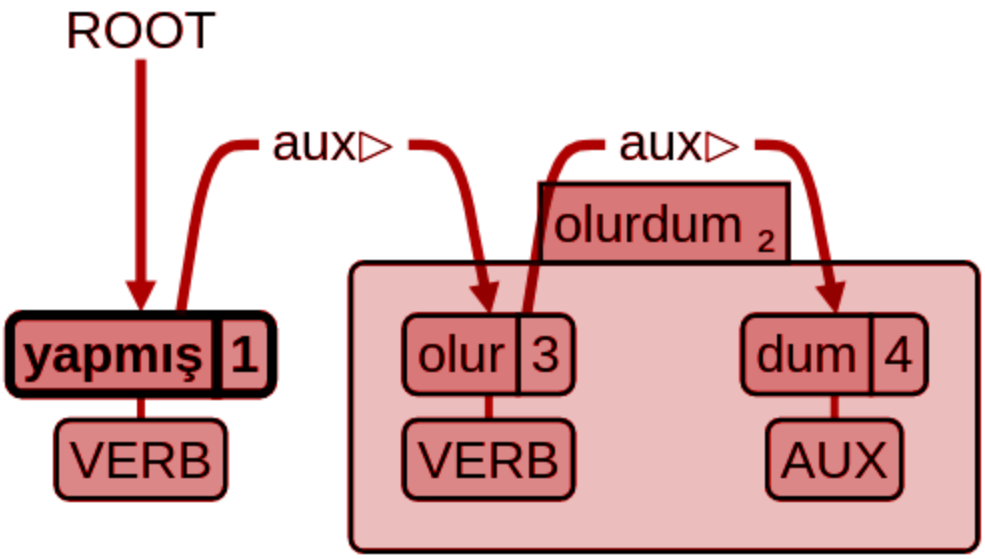
- Доктор болот болуш(ум) керек
  - Doktor olacak olmalı(yım).
  - Doktor olacak olsa gerek.
  - Doktor olsa gerek. / Doktor olacak. (She must be a doctor.) Uz: Doktor bo‘lsa kerak.
- Finite and non-finite verb forms with эле:
  - Болду беле / болду эле
  - Калган эле
  - Болчу беле
  - Айтат белең
  - Ошондой болмок беле?

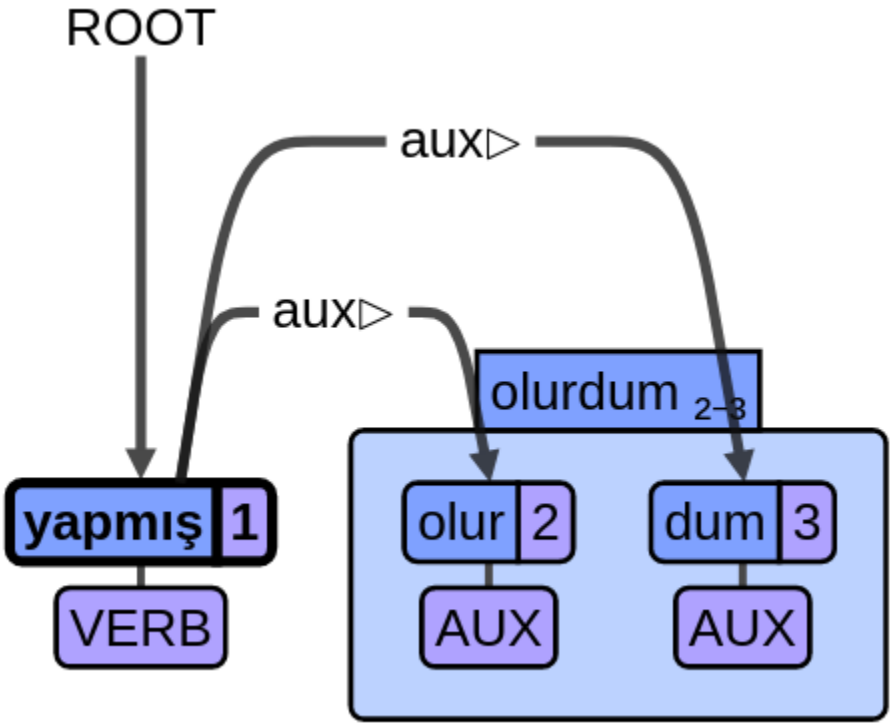


VerbForm=Vnoun Tense=Pres
Case=Loc



Tense=Past



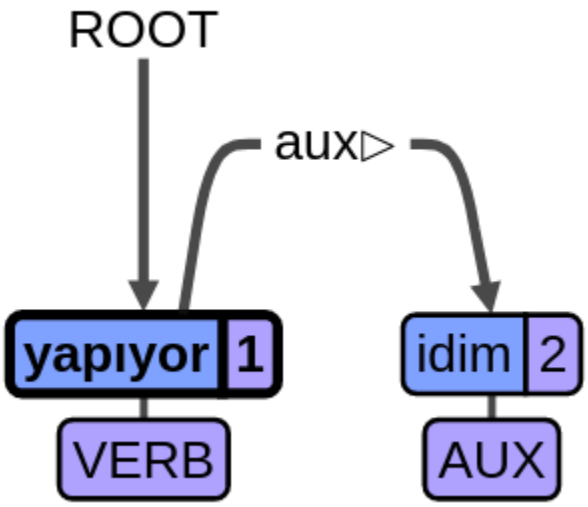


yapmış [yapmış adamı gördüm / yapmış[]ı gördüm] → normally Vform=Part
Lemma=yap
Vform=Inf

olur []
POS=AUX
Lemma=ol

Vform=Inf

dum
POS=AUX
Lemma=i
Vform=Fin



(1)

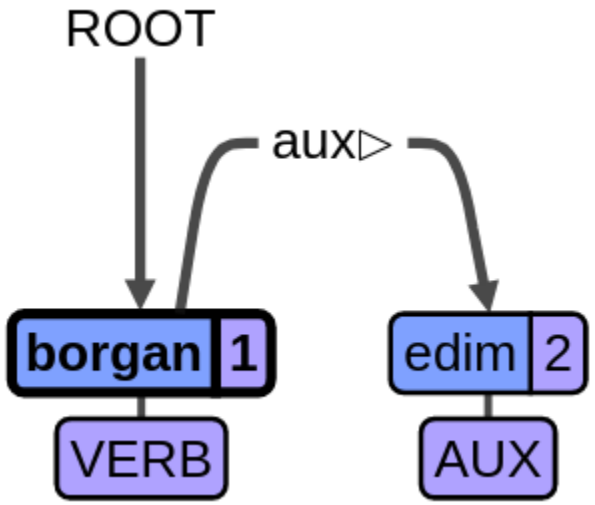
yapıyor
Vform=Inf

idim
deprel=aux
POS=AUX
Lemma=i
VerbForm=Fin

(2)

yapıyor
VerbForm=Vnoun
Case=Nom

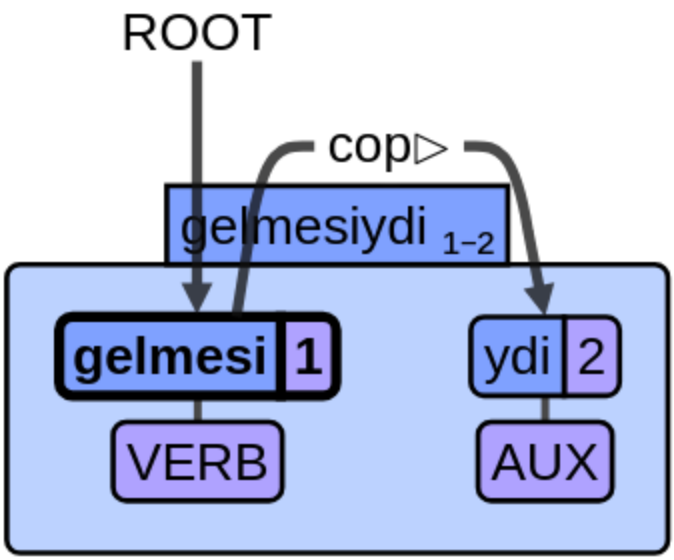
idim
deprel=cop

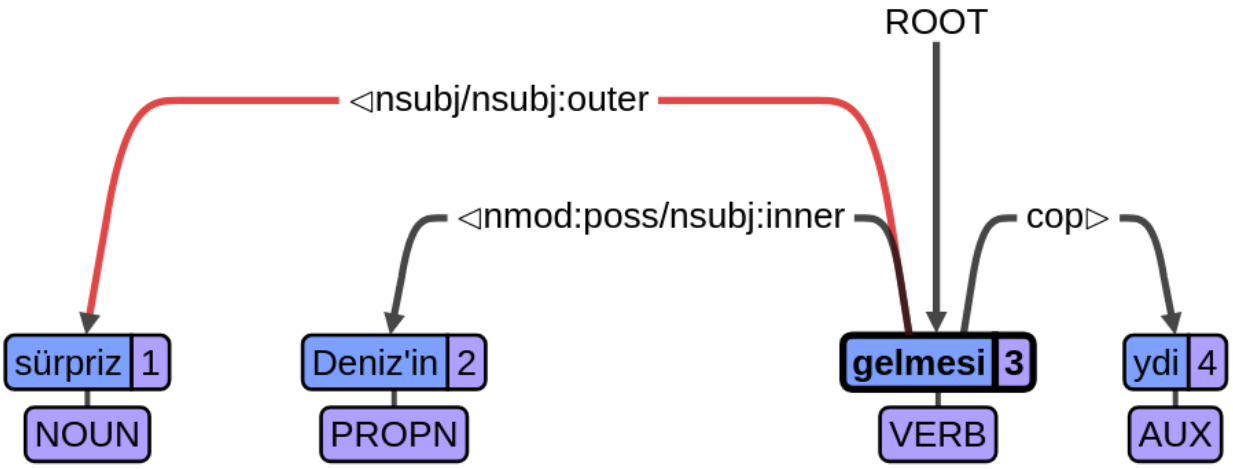


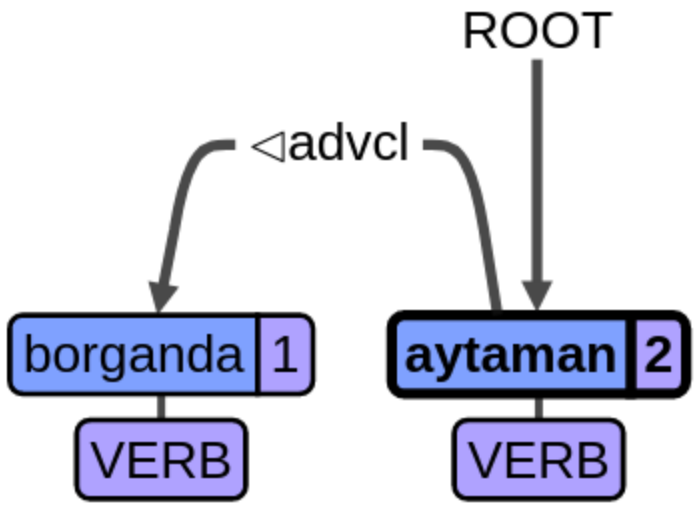
(Kyrgyz: барган элем)
(Uyghur: Barghan idim)
(Turkish: gitmiş idim / gitmiştim)

borgan	edim
VerbForm=Inf	POS=AUX (deprel: aux)
	VerbForm=Fin
	Person=1

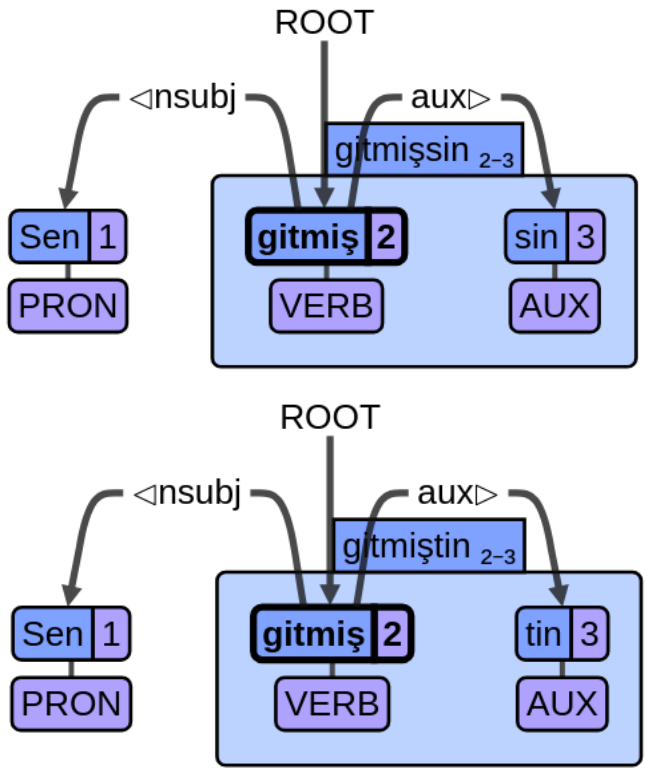
borgan	edim
VerbForm=Vnoun	POS=AUX (deprel: cop)
Case=Nom	

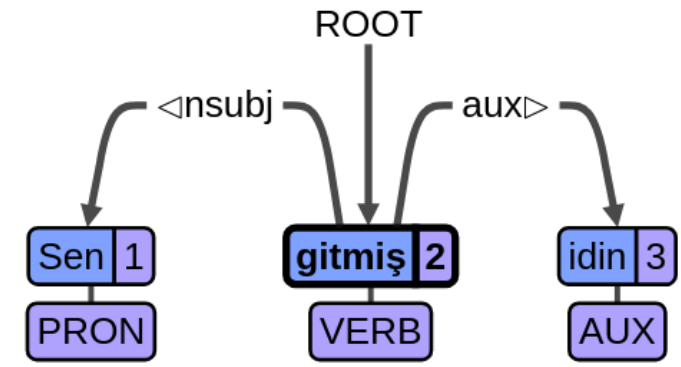

ROOT
cop▷
gelmesiydi 1–2
gelmesi 1
ydi 2
VERB
AUX

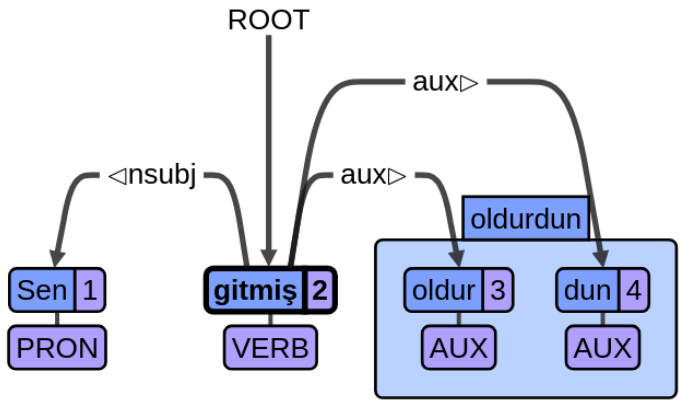

ROOT
◁nsubj/nsubj:outer
◁nmod:poss/nsubj:inner
cop▷
sürpriz 1
Deniz'in 2
gelmesi 3
ydi 4
NOUN
PROPN
VERB
AUX



(Kyrgyz: барганда айтам)
borganda
VerbForm=Vnoun
Case=Loc







6. Light verb constructions (?) with ol-/бол-

With ol-/бол-

- нерв болдум (borrowing), пайда бол- (only used in this construction), шор бол- (only used in this construction), айран-таң бол- (only used in this construction), ык бол-, каза бол- ‘die’ (only used in this construction and with тап- ‘be killed’)
- Berhüdar ol! (Be blessed!; only used with ‘ol-’ but it’s obvious that ‘berhüdar’ is adj)
- sağ ol,
- Китебим оомин болду. (My book came to an unfortunate end / met an unfortunate fate.)
    - To be PROPN-ed.
- kayb-olmak (to be lost). Kitabım kayboldu. (My book is lost.)
    - Onu kaybettik. (We lost him/her. [He/she’s passed away.])
    - mahv-olmak

- "N as indefinite object of verb: annotate these as compound, unless noun can take morphology or noun and verb can be separated by adverb" - but hard to tell with бол-/ol-

  Many appear to be adj/n+be:
  - memnun oldum, how is this different from “güzel oldu”

  Maybe just predicate adjective
  - şok ol-

Also attested with other verbs (yap-/кыл-)
- yürüyüş yapıyorum
- гимнастика кылып атам

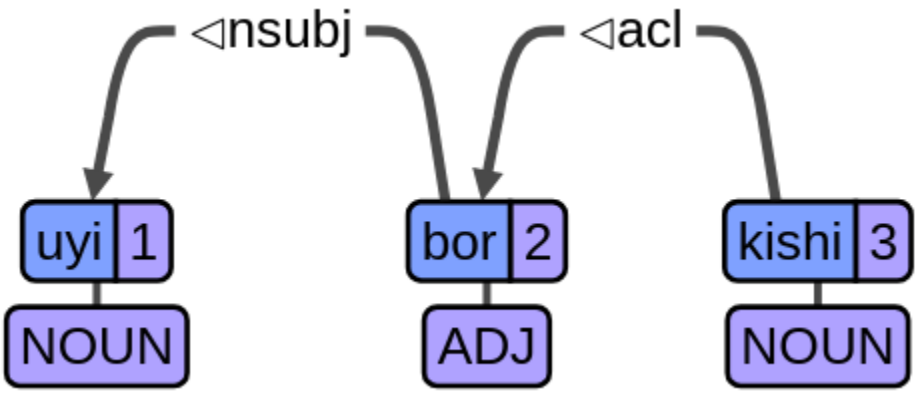
Forms with эт-/et- (lvc)
- звантет-
- ziyaret / tebrik etmek (visit / congratulate)
- "word limited to use in compound - annotate these as compound (or fixed?) "

Үйгө нанды [алып кел].	Нанды үйгө [алып кел]. (**SVCs** allow scrambling)
Досум менен нан алып сүйлөштүк. "	%Нан досум менен алып сүйлөштүк. (**productive converbs** don't allow scrambling)
Досумду нан алып үйүнө жеткирдим.	***Нан досумду алып үйүнө жеткирдим.
Мышыкка нанды алып тамак-аш бердим. → I got the bread and gave the cat food → I got the bread for the cat and gave it food	Нанды мышыкка алып тамак-аш бердим. → *** I got the bread and gave the cat food → ??? I got the bread for the cat and gave it food
	?Нан үйгө ал. (could show that ал isn't ditransitive;

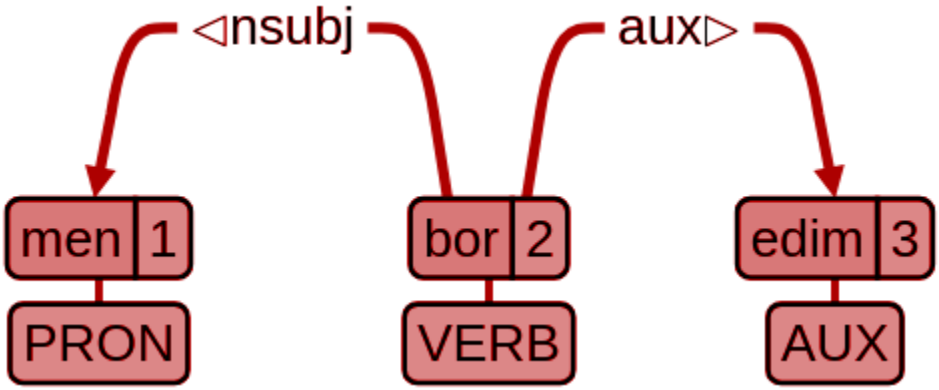
	Dative is licensed by кел, not ал-)
Үйгө нан ала кел.	?Нан үйгө ала кел.
Eve ekmek al.	Elma değil, ekmek eve al.

7. Existentials var/yok, бар/жок, bor/yo'q

- Adjective-like (function as predicates with non-verbal agreement and copula support), but lack many properties of adjectives (can't directly modify nouns IN TURKISH/AZƏRBAYCANI), but in Central Asian Turkic this is okay:



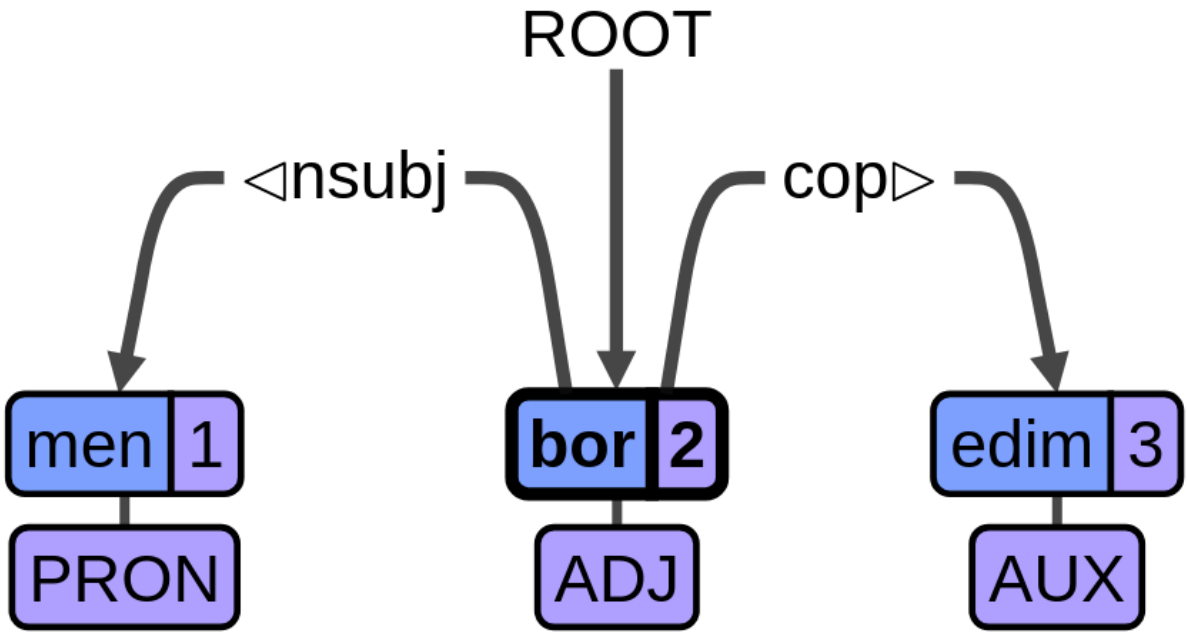
- Another approach:



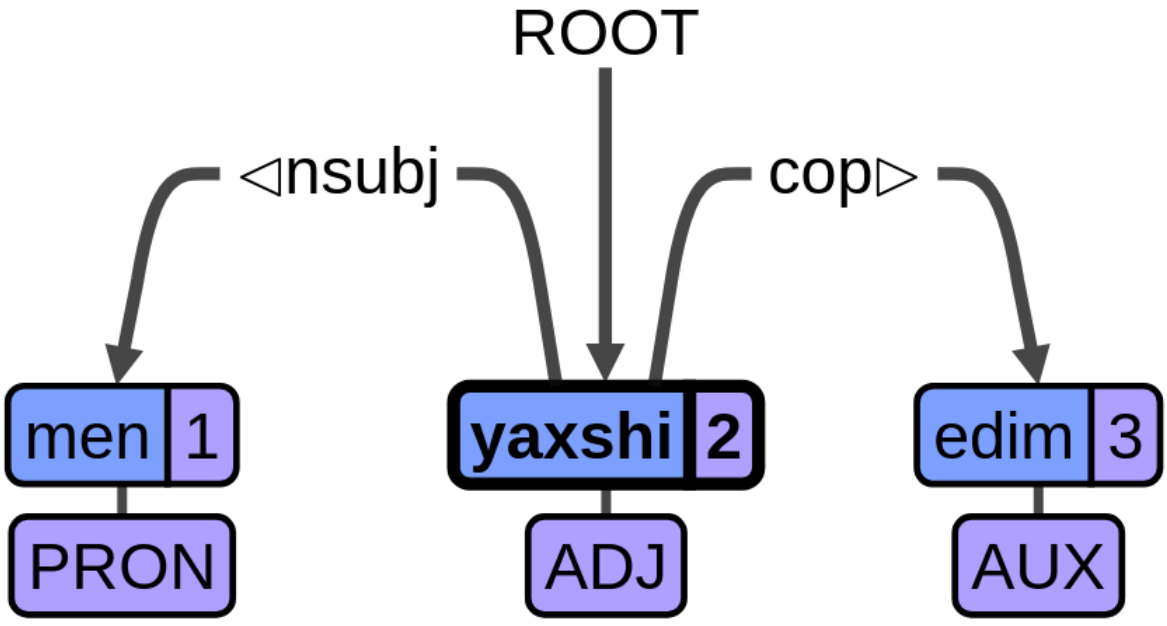
  - The problem with this approach is that existentials are not verb-like: no verb morphology (tense, person, etc.).

- E.g., чыгам / чыкпа from чык-,
  but *жогом / *жокпо from жок
- çıkayım / çıkma, but *yokayım / *yokma

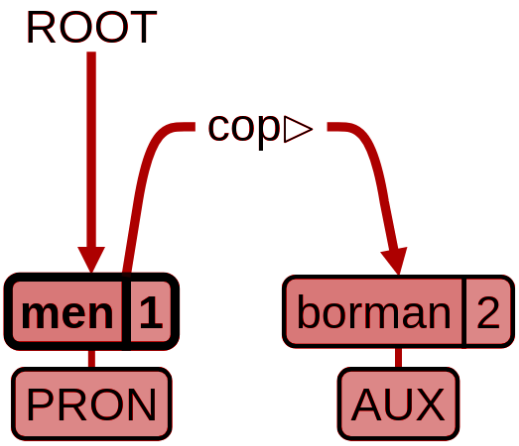
- As adjectives, they can be glossed in English as PRESENT, ABSENT
- Instead:



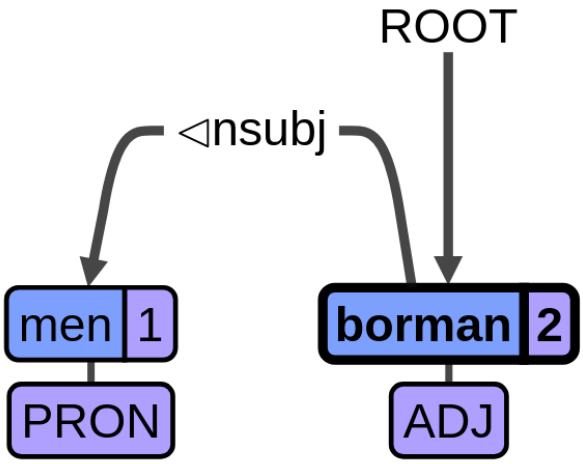
- Instead:



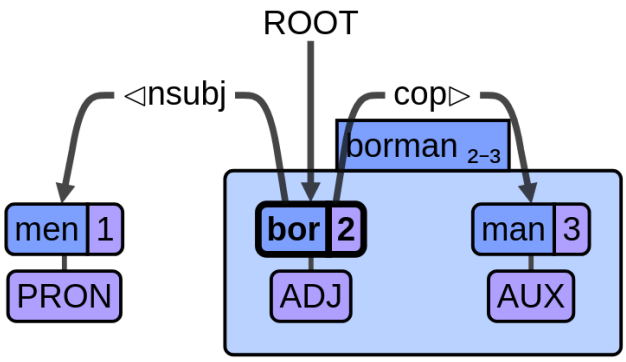
- If we treat existentials as "copulas", then this is what it looks like:



- Instead:



- Or more accurately:



- How do we know they're ADJ and not NOUN

        - ADJ and NOUN can both be predicates (доктормун, жакшымын)
        - Morphologically speaking, жокчулук/барчылык argue for noun status
        - жоктук/бардык argue for adjective status
        - most adjectives can be used as nouns in Turkic
        - as adjectives, can't be considered gradable (*en var, *эң жок, *жогураак)
        - as adjectives, in TURKISH/AZ, can't be attributive
            - This is like predicate-only adjectives in English: awake, alone, alive, alert, asleep (*an asleep person)
            - It's completely fine for adjectives to appear only in predicates
- It's easier to treat existentials as ADJ
    - seem to be no arguments against
    - might be some arguments for NOUN
        - would be fine in TURKISH/AZ
        - is difficult for Central Asian Turkic, where existentials can be attributive

сары - **ADJ**	сары - **ADJ** used as NOUN	сары - **NOUN**
сары гүлдөр (attr) менин гүлүм сары (pred)	сарылар мага жакты сарысы эң жакшысы экен (nsubj) sarı [araba] güzel	сары жакшы өң sarı güzel renk
(жок гүлдөр)	жогу ...	

NOUN	**NOUN** used as ADJ	ADJ
erkekler geldi	—	erkek çocuk öğretmenim erkek
demir iron	demir yumruk?	demir kapı iron gate/door
алма	алма дарак	—
ерлер келді - the men came	ер кісі - male person	ер кісі - brave person

жашым ___	—	жаш киши
жумурткanын сарысы (yolk) (сарысы смузиси)		сары (yellow)

QA-Note: kir: ’үй тапшырма - üy tapşırma (eng.: home work)’, does it work as **NOUN** used as ADJ? Or, is it just ‘үй <-compound - тапшырма’? nmod vs. compound

8. Auxiliaries/verbs used in copula-like ways

Дениз үйдө.
Дениз үйдө жүрөт.
(feels like үйдө *болуп* жүрөт, as if there’s a non-finite copula, and жүрөт is adding aspect)
Дениз үйдө отурат.

In Kazakh:

1	2
	Дениз үйде.
Дениз үйде жүреді.	Дениз үйде жүр.
Дениз үйде жатады.	Дениз үйде жатыр.
Дениз үйде тұрады.	Дениз үйде тұр.
Дениз үйде отырады.	Дениз үйде отыр.
Мен үйде жүремін.	Мен үйде жүрмін.

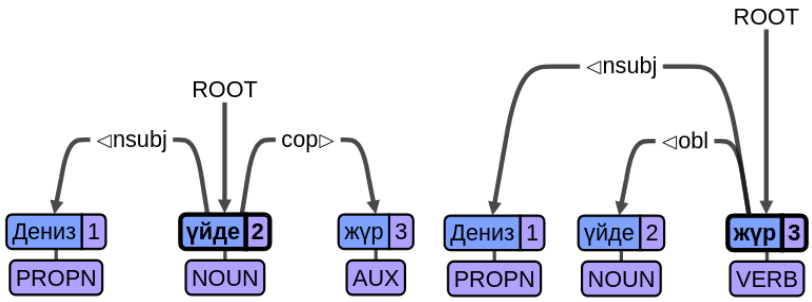
First sentences (1) are ~more lexically “goes, lies, stands, sits” — probably main verbs
Second sentences (2) are “is” with different aspectual information.

Question is: are the second sentences copulas or main verbs?

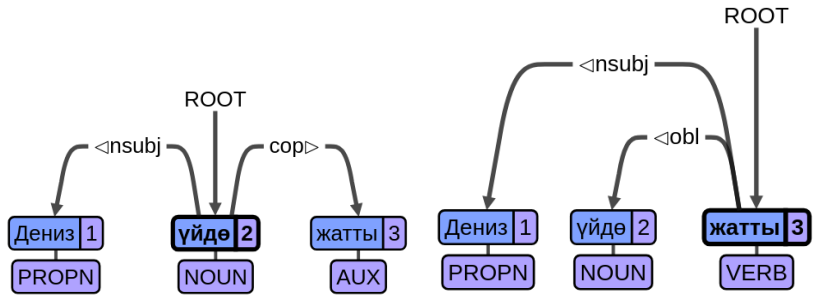
Like Kyrgyz, this is ambiguous in the past in Kazakh as well:
Дениз үйде жатты. - (1) Deniz lay down at home, OR (2) Deniz was at home (for a while).
(1) Deniz evde yattı/ (2) Deniz evdeydi / evde idi.

In **Kyrgyz** you get the (1) readings only in some cases
Дениз жумушсуз үйдө турат/жатат/отурат/жүрөт.
But жүр is more broadly usable: Дениз үйдө жүрөт.

Which of these two would be the “correct” annotation:



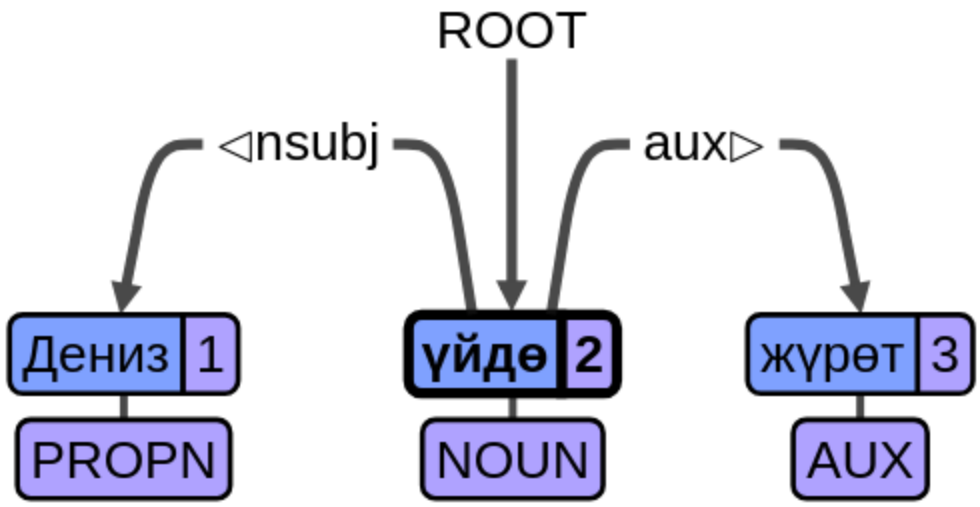
In Kyrgyz both could be possible depending on the meaning:
(1) (2)



(1) Denis <u>was</u> at home (doing nothing) - Deniz evdeydi.
(2) Deniz <u>lay</u> at home - Deniz evde yattı.

In Kazakh it’s unambiguous ONLY in the present tense.
In Kazakh and Kyrgyz, it’s ambiguous in all (other) tenses.

OPTION 3.
Дениз үйдө жатты. ≈ Дениз үйдө болуп жатты. ? (feels like an -ip form of э-)
Бат-баттан келип турат. - where тур is an auxiliary with similar aspectual meaning as in the copula use.
Auxiliary construction with elided non-finite copula (gap in copula paradigm?)



1	Андыктан	_	PRON	_	_	4	nmod	_	_
2	мен	_	PRON	_	_	4	nsubj	_	_
3	эч	_	ADV	_	_	4	advcl	_	_
4	барбай	_	VERB	_	_	0	root	_	_
5	коё	_	AUX	_	_	4	aux	_	_
6	алмак	_	AUX	_	_	4	aux	_	_
7	эмесмин	_	_	_	_	_	_	_	_

“I had no choice but to go”
“I couldn’t choose not to go”