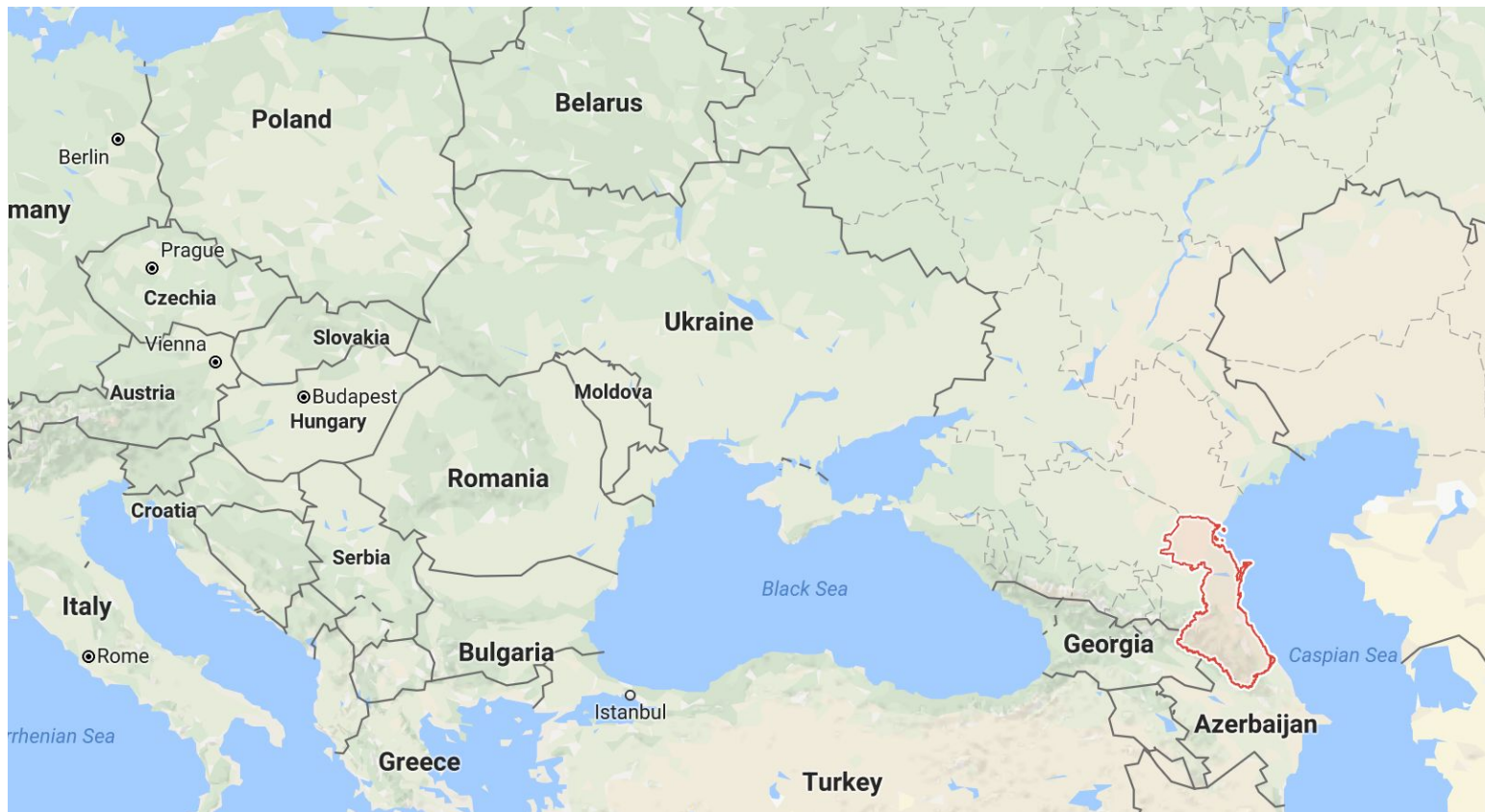


Universal Dependencies for Mehweb

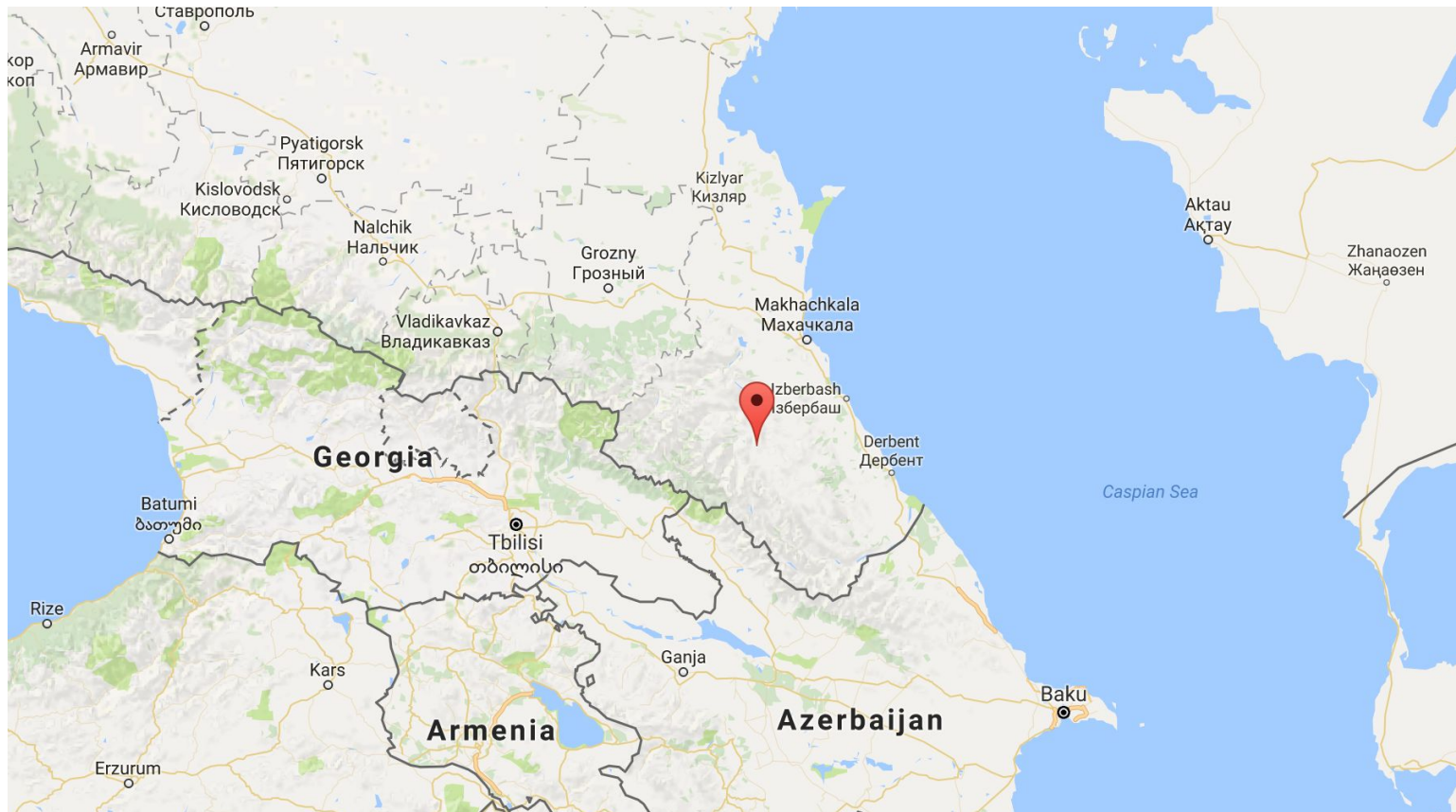
17.10.17

O. Lyashevskaya • A. Kozhukhar

Mehweb: Location



Mehweb: Location



Mehweb: General Info

- East Caucasian language family, Dargwa group
- ~800 native speakers
- One-village idiom surrounded by Laks and Avars
- Non-literate, Avar orthography is used
- Ergative alignment
- Agglutinative morphology
- SOV-order
- Rich system of locative cases

	Sg	Pl	
M	<i>w</i>	<i>b</i>	HPL
F	<i>d</i>		
F1	<i>d-r</i>		
N	<i>b</i>	<i>d-r</i>	NPL

Mehweb: General Info

- rasuj-ni muḥammad-la k^wih.me ar-d-uk-ib
rasul.OBL-ERG Muhammad-GEN sheep.PL(ABS) away=NPL=lead:PFV-AOR
'Rasul took Muhammad's sheep.'
- surat aqi-le le-b ba^ɕhi-ze-b
picture(ABS) up-ADVZ be-N wall.obl-INTER=N(ESS)
'A picture is hanging on the wall.'

Universal Dependencies: Motivation



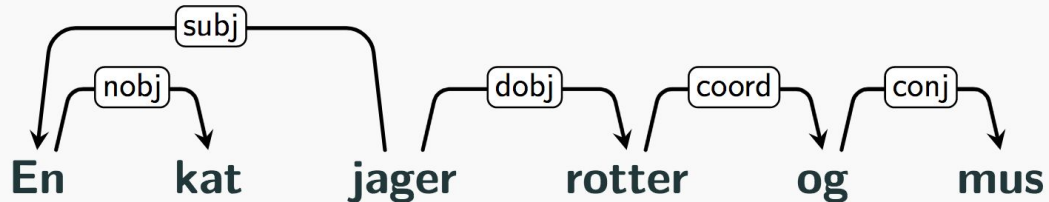
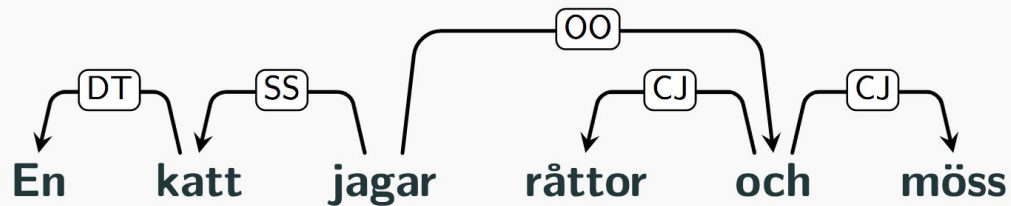
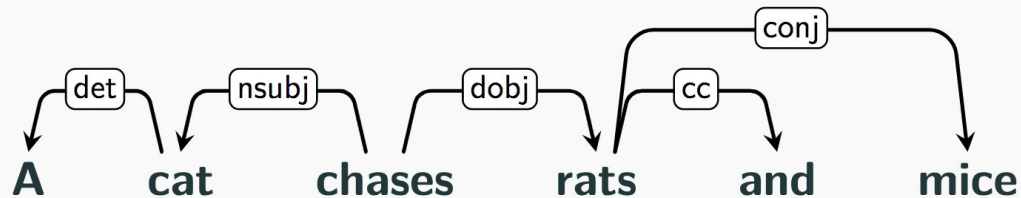
Increasing interest in multilingual NLP:

- Multilingual evaluation campaigns to test generality
- Cross-language learning to support low-resource languages

Increasing awareness of methodological problems:

- Current NLP relies heavily on annotation
- Annotation schemes vary across languages

Universal Dependencies: Motivation



Universal Dependencies: Motivation

For theory:

- Hard to compare across languages empirical data collected in the field
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to evaluate cross-language learning

For practice:

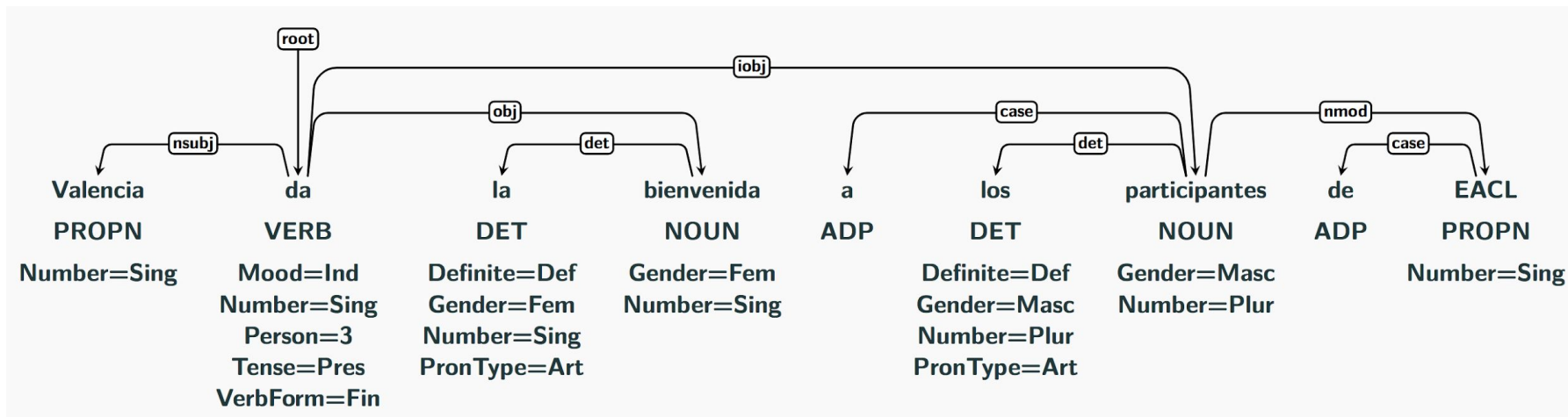
- Hard to usefully do cross-language structure transfer
- Hard to build and maintain multilingual systems
- Hard to make progress towards a universal parser

Universal Dependencies: Motivations

- The treebanks were mostly converted from their specific annotation schemas – need for annotation from scratch
- The mainstream UD annotation schemas were developed primarily for Indo-European treebanks, ie. oriented on IE categories and constructions
- Need for morphological and syntactic diversity

Universal Dependencies: Overview

<http://universaldependencies.org>



- Part-of-speech tags
- Morphological features
- Syntactic dependencies

Universal Dependencies: Goals

- Cross-linguistically consistent grammatical annotation
- Support multilingual NLP and linguistic research
- Build on common usage and existing standards
- Complement language-specific schemes
- Open community effort

Universal Dependencies: Principles

Maximize parallelism

- Don't annotate the same thing in different ways
- Don't make different things look the same
- Don't annotate things that are not there

Universal taxonomy with language-specific elaboration

- Languages select from a universal pool of categories
- Allow language-specific extensions

Universal Dependencies: Morphological annotation

Le
le
DET

Definite=Def
Gender=Masc
Number=Sing

chat
chat
NOUN

Gender=Masc
Number=Sing

chasse
chasser
VERB

Mood=Ind
Number=Sing
Person=3
Tense=Pres
VerbForm=Fin

les
le
DET

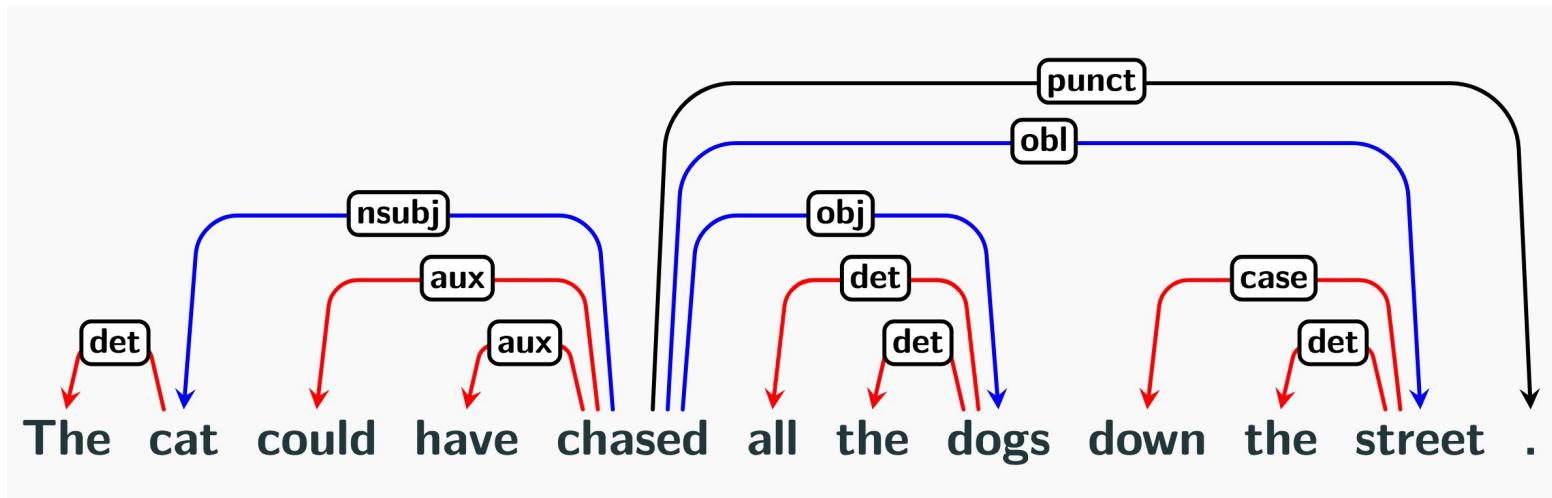
Definite=Def
Gender=Masc
Number=Plur

chiens
chien
NOUN

Gender=Masc
Number=Plur

.
.
PUNCT

Universal Dependencies: Syntactic annotation



- Content words are syntactically related
- Function words attach to the content word they modify
- Punctuation in most cases attaches to head of phrase or clause



















Universal Dependencies: CONLL-U markup format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	adajni	adaj	NOUN	-	-	4	nsubj	-	-
2	mas'inkalic'u	mas'inka	NOUN	-	-	4	obl	-	-
3	muc'ur	muc'ur	NOUN	-	-	4	obj	-	-
4	berc'ur	#erčur	VERB	-	-	0	root	-	-
5	.	.	PUNCT	-	-	4	punct	-	-

- How should we define lemmas for verbs?
- Should we mark ABS?

Universal Dependencies

- First guidelines launched in October 2014
- Treebank releases every six months
- Version 2 in December 2016 (guidelines) and March 2017 (treebanks)
- 50 languages
- 70 treebanks
- Next release in November (v2.1)

▶		Afrikaans	49K		–				
▶		Ancient Greek	202K				✓		
▶		Ancient Greek–PROIEL	211K		–		✓		
▶		Arabic	242K		–		✓		
▶		Arabic–NYUAD	629K		–		✓		
▶		Arabic–PUD	20K		–				 W
▶		Basque	121K				✓		 
▶		Belarusian	8K		–		✓		
▶		Bulgarian	156K			 ✓	✓		   
▶		Buryat	10K		–				  
▶		Catalan	530K			 ✓	✓		

Mehweb: Morphological annotation

POS mapping

POS	Mehweb	POS	Mehweb
ADJ	+	CCONJ	-/+
ADV	+	DET	+
NOUN	+	NUM	+
VERB	+	PART	+
ADP	+	SCONJ	-
AUX	+	PRON	+
INTJ	+	PROPN	+

Mehweb: Morphological annotation

POS mapping

98 times in 12 texts

- CCONJ is present only in Magometov's texts, additive or non-finite verb forms are used instead:

- (1) xaj-ja ursi-li-ʔini il-i-ce aqu ik'w-es sik'al b=uxib
khan.OBL-GEN son-OBL-ERG that-OBL-SUPER to put.on:PF-INF something N=bring:PF-AOR
- wa**
and
- sune-la-l quli d=uk-ib.
SELF-GEN-ATR home F=lead:PF-AOR
- 'Khan's son brought her a dress **and** took her to his house.'

- SCONJ is irrelevant, non-finite verb forms are used instead:

- (2) nab b-ik-ib ʔali w-ebk'-i-**le** ile
I(DAT) N=happen:pf-aor Ali(ABS) M=die:PF-AOR-**CONV** COMP
- 'I thought (it occurred to me) **that** Ali was dead.'

Mehweb: Morphological annotation

Feature mapping

Feature	Relevant Values	Feature	Relevant Values
Animacy	Hum, Nhum	Person	1, 2
Aspect	Imp, Perf	Polarity	Pos, Neg
Case	Abs, Erg, Com, Dat, Gen, Voc	Reflexive	Yes
Foreign	Yes	Tense	Fut, Past, Pres
Gender	Fem, Masc, Neut	VerbForm	Fin, Inf, Part, Conv, Vnoun
Mood	Ind, Imp, Cnd, Pot, Jus, Opt, Prp	Voice	Cau
NumType	Ord, Card	PronType	Dem, Rcp, Int, Neg, Ind
Number	Sing, Plur	Evident	Fh

Mehweb: Morphological annotation

Feature mapping

- More than one verbal nominalization:

matrix verb	-ri (masdar)	deš-nominalization
(d)iges ‘want’	+	+
qumartur ‘forget’	+	
urhes ‘tell’	+	
alhes ‘know’	+	(only with copula)
uhes ‘can’		

- Meaning of the person marker depends on the polarity of the sentence:

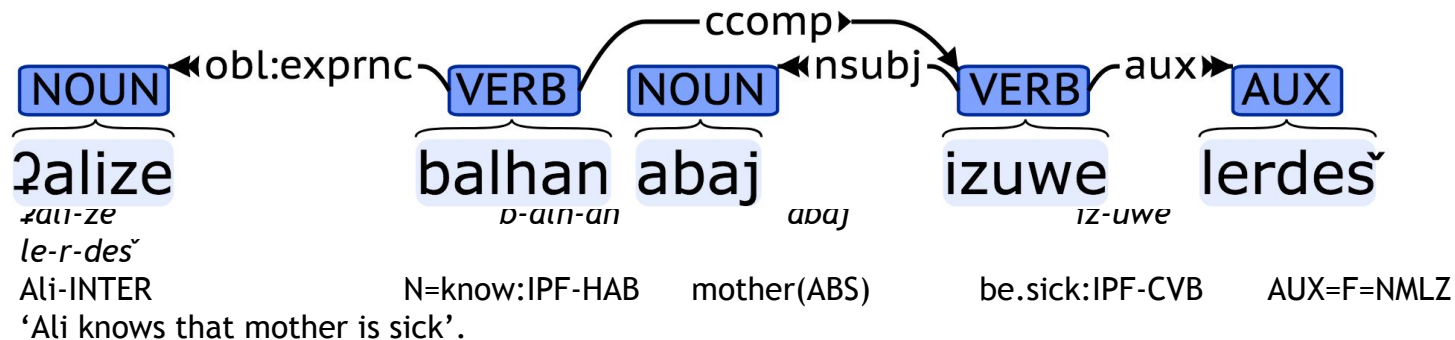
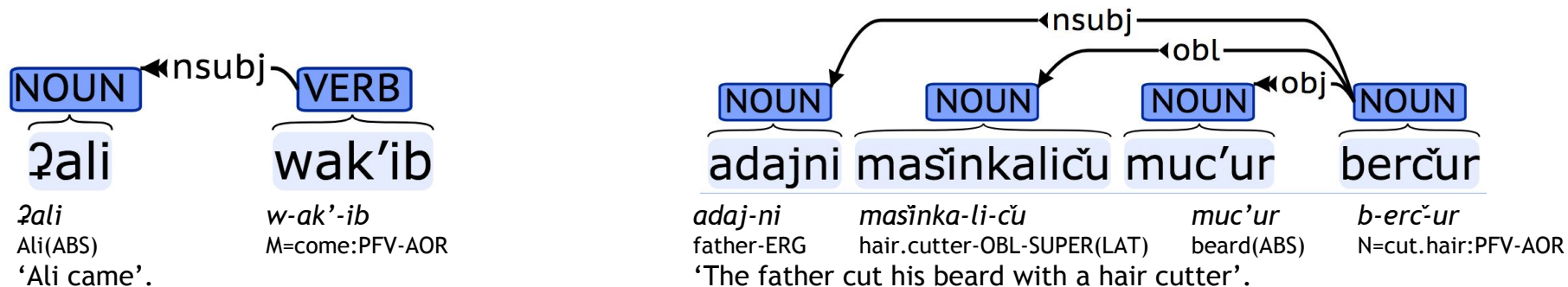
nu quliw lew-**ra**. ‘I am at home.’
ñu quliw lew. ‘You are at home.’

nu quliw lew? ‘Am I at home?’
ñu quliw lew-**ra**? ‘Are you at home?’

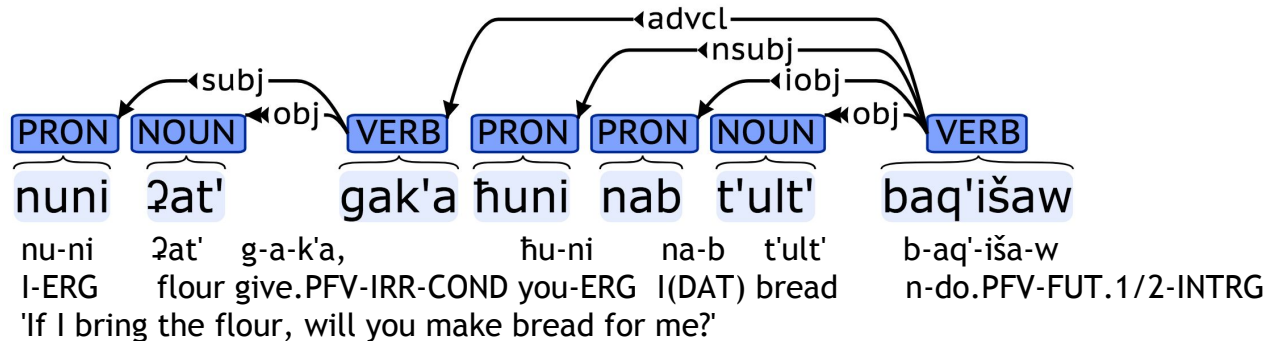
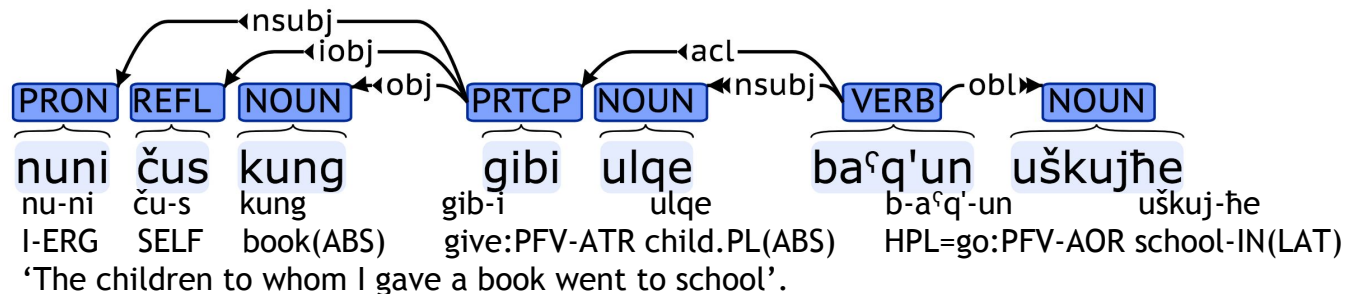
Mehweb: Syntactic annotation

UD	Mehweb	UD	Mehweb
nsubj	+	acl	+
obj	+	advcl	+
iobj	+	advmod	+
csubj	not attested	aux	+
ccomp	+	cop	+
xcomp	+	mark	-
obl	+	nmod	+
vocative	+	det	+
expl	-	clf	-

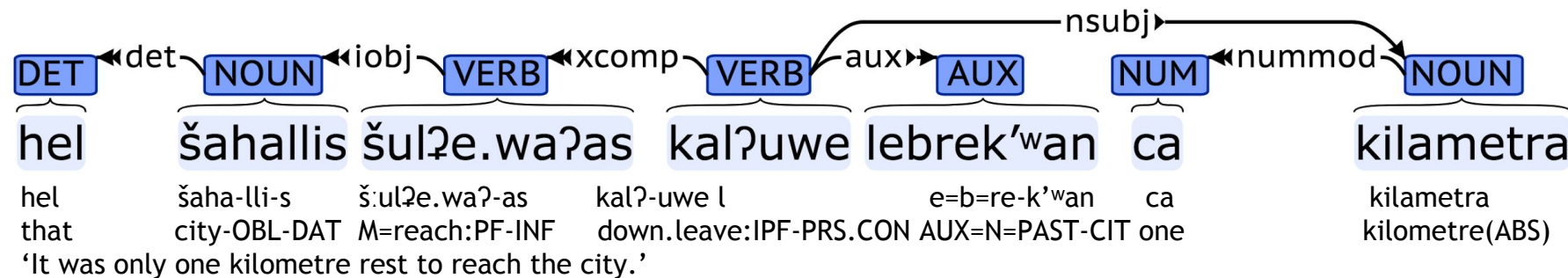
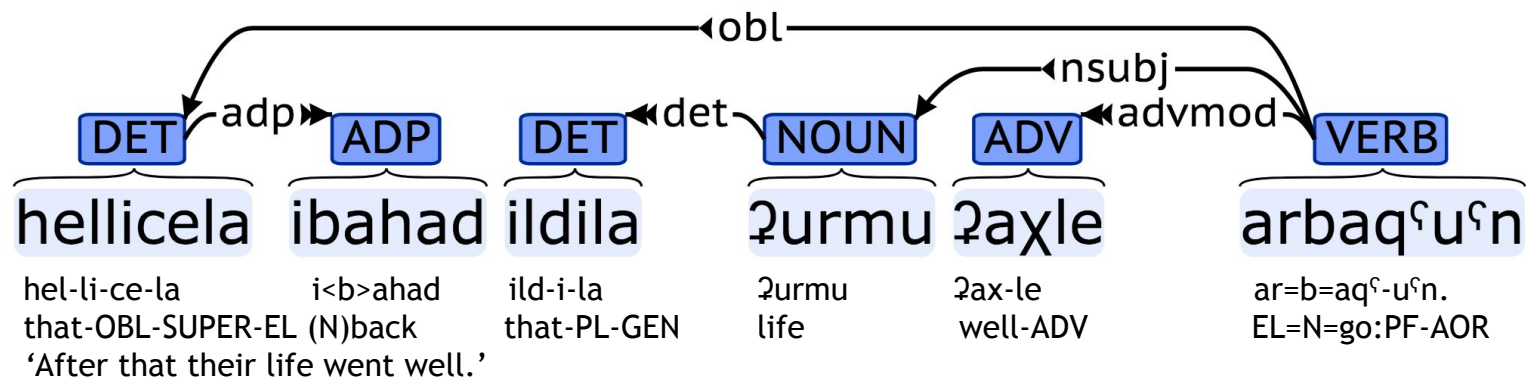
Syntactic annotation: Examples



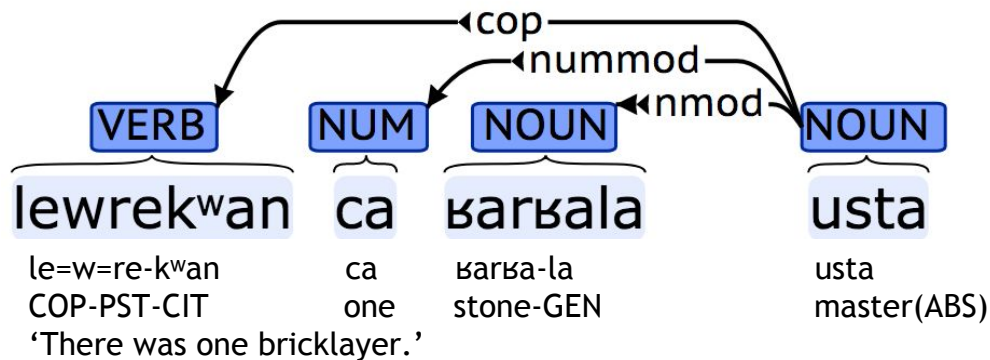
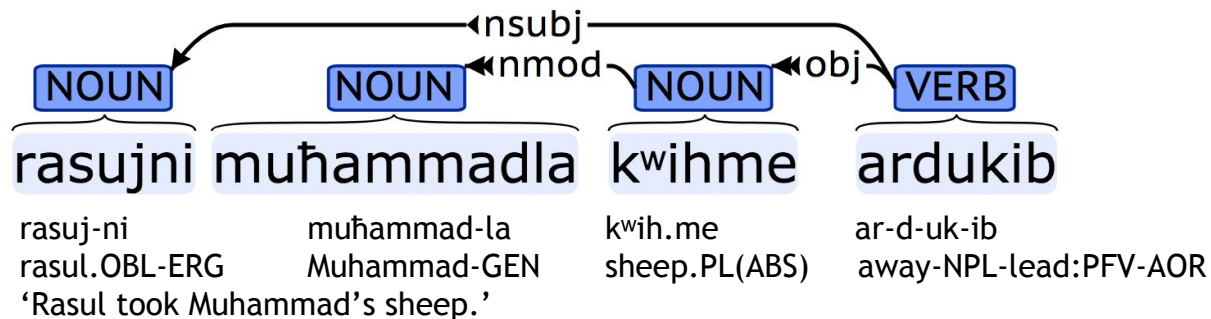
Syntactic annotation: Examples



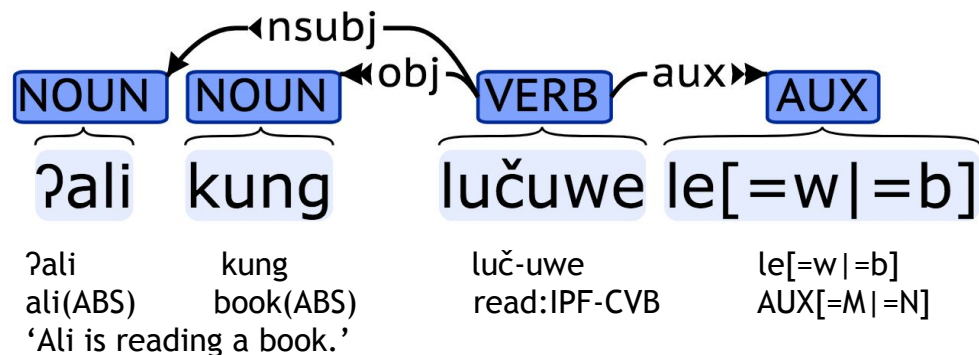
Syntactic annotation: Examples



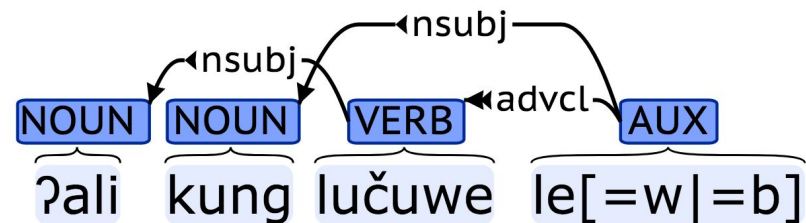
Syntactic annotation: Examples



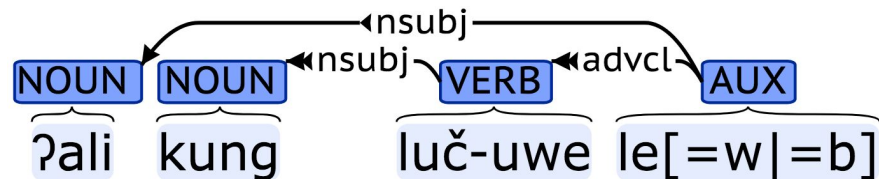
Mehweb: Biabsolutive (binominative) construction



unacceptable because
of the marking



both structures look OK



there are at least two other
possible structures

Additional issues

- We have texts from different time periods:
 - Magometov's texts from 1982
 - Contemporary texts from recent fieldtrips
- The texts are of different «type»:
 - elicited examples from the grammars (old & contemporary)
 - written texts (old)
 - peers-texts with audio
 - regular contemporary texts with audio

What's next?

- Getting ready for the UD 2.1 release on November 1st
- Trying out new visualization tool by Masha Sheyanova:
<https://maryszmary.github.io/ud-annotatrix/standalone/annotator.html>
- Documentation for the Mehweb treebank

Thank you!