Post-predicate elements in Iranian and neighbouring languages
Explanations and guidelines for text coding, version 15.01.2020

1. The basic unit (token) for analysis is the **CONSTITUENT**, which will be a NP of some sort, so each constituent counted **gets its own line** in the table.
2. We code only those **non-subject constituents** that fall under the broad headings given in Column N.
3. Subjects, and other kinds of constituent not covered in N can be ignored, which means that some clauses don't contain any constituents that need to be coded (e.g. ’the woman left’ would contain no item that needs to be coded).
4. Other clauses might contain **more than one constituent that needs to be coded** - see the sample file for examples.
5. If you cannot easily decide on how to code a particular constituent, it is better not to code it: enter "1" into the 'nc-Column' (Q) and move onto the next clause.
6. For clauses with various kinds of non-canonical subjects, e.g. experiencer expressions, expressions of desire, or expressions of possession, which take some kind of additional case marking: These **non-canonical subjects are treated - for the purposes of this investigation - as subjects**, hence not coded; instead we code the ‘wanted’ or ‘possessed’ constituent (if overtly present in the clause); see N and O under detailed explanations below.

Overview of the coding scheme

Letters A-R refer to letters used in the Excel template

A affiliation1	B affiliation2	C location1	D location2	E text ID	F token ID	G token in context	H context trans-lation	I token	J token trans-lation	K pro	L anim	M weight	N role	O flag	P pos-ition	Q nc	R comments

GREY	A-F General information on the language, the text, and the position in the text: You only need to fill out the first lines in A-E, we copy this content into the other cells later: F is a space for a tokenID, i.e. an indication of where, in your text, a given example is found. In the sample Excel file, I have adopted the paragraph numbering of the original source for tokenID
YELLOW	The actual forms, and translations. In the sample Excel file, column H (context translation) is empty (too much work at the moment, and the entire text translation is available in the published sources); this is a matter of discretion. Minimally, you will therefore need G, I and J
GREEN	What we are mostly interested in: The independent variables for structure (K-O), the dependent variable 'position relative to predicate' (P), and the Q for empty / uninterpretable cells. CONSULT EXPLANATIONS BELOW
WHITE	any comments on individual items, or additional language-specific coding that analysts may wish to add

Detailed explanations for the individual column

	Labels	**Description of content, abbreviations**
A	affiliation1	Highest-level groups: Iranian, Armenian, Kartvelian, Turkic, Semitic, Nakh-Daghestanian etc.
B	affiliation2	relatively flexible, at your discretion; intermediate level grouping, language name, dialect (e.g. West Iranian, Central Kurdish, Mukri)
C	location1	Place of the speaker's socialization. This can be entered as a place name here in the meantime, later we enter latitude coordinates
D	location2	can remain empty, later for longitude
E	textID	Identifier for the particular text (remains the same for all examples taken from that text), see sample
F	tokenID	Position of token within the text (e.g. if the clauses of the text are numbered, then the clause number etc.). The degree of precision will vary according to source and text, but this is necessary if we need to locate a particular example at later stages.
G	token context	The actual object form of the constituent, together with relevant context
H	context translation	translation of context (empty in the sample Excel file)
I	token	actual form of the constituent, with relevant flagging (if adpositional)
J	token translation	translation of the token
K	pronoun	A value is inserted here **only if the token is a pronoun**. Otherwise this column is empty. Special **case clitic pronouns**: If the token is a clitic pronoun, it can be indicated with e.g. "1-bound" = first person pronoun, clitic or generally prosodically dependent form. If this form is expressed through an affix or clitic that is attached to the verb (as is often the case in Iranian languages), then it is ignored, because if is part of the verb, and therefore cannot be meaningfully coded as 'before' or 'after' the verb; two examples of this are found in 18 and 42 in the Sample Excel file of Central Kurdish. You may of course add your own, corpus-specific comment on this in the comments column R if you wish to note the presence of these elements. If it is bound to some element outside the verb (e.g. an adposition), it can be coded in the same way as other arguments. Note that in the ANALYSIS, we will analyse these bound elements separately from the free argument constituents. The following values are available:
		1 first person 2 second person 3 third person, human 4 third person, non-human wh wh=any kind of interrogative pronoun, including internally complex interrogatives 'for what = why'; refl refl=any kind of reflexive or reciprocal pronoun (could be optionally extended with a person value, e.g. "refl-1") other for example indefinite pronouns, as in Iranian *hēč, kas, yakī*, etc.

L	animacy	A value is inserted here for animacy **if the token is a full NP** (‘tree’, ‘big dog’, ‘my father’ etc.), rather than a pronoun (which would have an entry in K). The exception to this is pronouns with the value "4" (third person, non-human, see above), which can be additionally specified in L for the values "anim, bp, inanim, other" etc. The following values are available:	
		hum	human or anthropomorphized (e.g. speaking entities in folk-tales)
		anim	animate, non-human
		bp	body part (e.g. ‘hand’, ‘eye’ etc.)
		inanim	inanimate (including place names and expressions that are an NP, e.g. ‘at home’)
		adverb	’here’, ’there’, ’outside’, ’over there’, ‘yesterday’, ‘now’, ‘later’
		adj	adjectival complement of a copular expression (*the merchant was **very rich**)*
		other	none of the above (undecidable)
M	weight	For tokens with an entry in Column L, a value for weight can be added. We distinguish **three** levels:	
		[left empty]	- single word, e.g. "nails"
		1	- single word plus an additional independent word (*this nail, my mail, three nails, long nails, two-word proper names such as Ahmad(-i) Mahdi,* noun + REFL, noun + counting word). Note that languages with articles (if the latter are analysed as distinct words) will therefore have an overall higher rate of "1" and "2" than languages without articles. This needs to be kept in mind when comparing languages, but can be ignored for coding individual languages.
		2	- single word + more than one additional word (*this long nail, all my long nails, nails with flat heads*) IMPORTANT NOTE: IN COUNTING 'WORDS' WE DO NOT COUNT BOUND ELEMENTS SUCH AS INDEFINITENESS SUFFIXES, EZAFES, CLITIC PRONOUNS IN IRANIAN. NOR DO WE COUNT ADPOSITIONS HERE, AS THEY BELONG TO THE FLAGGING, RATHER THAN THE NP ITSELF (THE PRESENCE OF AN ADPOSITION IS NOTED ANYWAY IN COLUMN O)
N	role	**Values**	explanations, **Sub-types**
		do	direct object of a transitive verb; **do-def** if the constituent is identifiable within the discourse context (given); if in doubt, code with "do" as the default)
		goal	goal of verb of motion; **goal-r** (refined goal, e.g. on top of X, into the middle of X etc.), if in doubt code "goal")
		goal-c	goal of verb of caused motion (‘put, place, bring’ etc.) **goal-c-r** (refined caused motion goal)
		rec	recipient of verb of transfer (which involve a change of ownership)
		addr	addressee of verb of speech
		ben	benefactive (X does sth. in the interest of Y, which generally implies that Y is a sentient being)
		loc	location of event or participant
		instr	instrument
		abl	source of motion
		com	comitative (accompanying person)

		cop-c	copula complement noun e.g. she was **a teacher**)
		poss	possessed NP in a possessive expression with a non-canonical subject as possessor, i.e. *to-me is X*, X is coded with "poss"[1] when recipient cannot be distinguished from benefactive
		rec-ben	if the constituent is ambiguous between a recipient and benefactive reading ('I bought a bike for her'; show something to X)
		becm	the complement of a change-of-state verb (become, turn into etc. 'he became very weak', 'turned into a stone' etc.)
		stim	"stim" used for Georgian for the stimulus of a predicate of desire/wanting (added 07.08.2019)
		other	none of the above (temporal and space expressions; cause; reason (‘because of X’), elements coded lvc-poss in O)[2]
O	flagging	prep	
		circ	
		postp	
		case	optionally**: instr, abl, com** (if the language distinguishes these cases; otherwise just "case" as the default; same applies to coding of pronominal clitics, if they are analysed as being in some kind of non-nominative case)
		bare (0)	lack of any overt flagging; the form is indistinguishable from the form of an intransitive subject in the language
		poss-rec	(when e.g. a recipient is marked like a possessor, e.g. ‘I gave **his** money = I gave the money to him’)
		lvc-poss	possessor of a light verb complement (e.g. in Iranian ‘showing-EZ us do’ ‘show to us’)
		relnoun	used for words that are recognizably nouns with meanings like 'side', 'back' etc. and can inflect for e.g. case, but are frequently used in a similar manner to adpositions). **Subtype**: relnoun-case if the relnoun is additionally case marked
P	position	0	before the predicate
		1	after the predicate
Q	nc	1	If there is no relevant constituent, or if the clause / string is not analysable; otherwise leave this cell empty (see sample file)
R	comments		e.g. general comments, for example if the clause contains a light verb comlement (LVC), or other problems that may arise, can be noted here, similarly, I have noted the presence of auxiliary verbs, so that these can easily be identified for later investigations You can also define **language specific labels for special phenomena in your language** and include them in this column, where they can be easily searched.

[1] Not coded: possessor (*to me* is X), which is treated as a non-canonical subject, hence outside the scope of the investigation

[2] Not coded: any other elements, including non-verbal parts of complex predicates (if analysed as a complex predicate, otherwise they are most likely a "do" direct object.)
Not coded: Pronominal clitics which are part of the subject NP (*barādar-om* does XY ‘my brother does XY’).

S …	optional additional columns		Coders may be interested in other aspects of their language, not covered by the previous columns. They are free to introduce **additional coding columns for this kind of information**. The following have been suggested by Agnes Korn, based on her own data: VERB: contains values COP, motion, transfer, real CP, CP (free combination of LV and dO) NOUN: OBL, PL, EZ, SPC, DEM, counting word, *ham*; where any of these are present, but not otherwise counted RUNNING NO. for running number following order of text (in case one sorts data by various values and then wishes to return to order as in text; may be necessary because many sentences figure several times in table) If you introduce additional columns, please provide a note explaining the abbreviations etc.