Post-predicate elements in Iranian and neighbouring languages

Explanations and guidelines for text coding
Revised version 06.01.2020

Geoffrey Haig

1. General principles for compiling a data set

1. Each data set **is a text in a particular variety** ('doculect'), locatable to a particular place, and assignable to some language group.
2. Texts should be spoken rather than written, i.e. ideally reliable transcriptions (or in a generally accepted orthography) of recorded speech events, and not translations from a standardized language.
3. We aim at minimum of 500 coded tokens (see below) for each data set.
4. The data set has the **form of a table, and it should ideally be comprehensive in the sense that the entirety of the text is contained in the file**, including those sequences that contain no usable tokens (see below). This is important to enhance the interpretability of the text, but also in the interests of re-usability for later research questions.
5. The text is then broken up into 'utterance units', approximately corresponding to a simple sentence, which ideally contains one predicate. Just how you decide to segment your text is a matter of preference and practicability; technically it is easiest to adopt those units which are distinguished in your source text (for example, "sentences" defined by full stops in a text, or lines in a Toolbox transcription, or intonation units, as in the Kumzari data set), because these can simply be directly imported (semi-)automatically into Excel. Remember, relevant tokens need to be identified (see #6), and each provided with their own line; it will be easiest if most utterance units contain one token, but this will never work out perfectly, so you will need to manually duplicate, or ignore, some lines as you start coding.
6. We then identify the relevant **token**(s) (the actual unit coded), in each utterance unit, which will be a NP or PP of some sort, or in the case of predicatively-used adjectives, may be an adjectival phrase

7. Each token **gets its own line** in the table. If one utterance unit contains more than one relevant token, the utterance unit needs to be duplicated (i.e. the line is copied and inserted into its own line), so that each codable token has its own line; this is illustrated below; the utterance unit contains two relevant tokens (‘my wife’ and ‘with my sword’, hence is duplicated in G and H, so that each token has its own line [iran_baloch_turkmen1, 39-40]

G	H	I	J	K	L	M	N	O	P
mnī ǰinēnā gōmē mnī zāmā nikē kan u bzūr u byār.	marry my wife with this sword of mine, (then) take her and bring her back”.	mnī ǰinēnā	my wife		hum	2	do-def	case	0
mnī ǰinēnā gōmē mnī zāmā nikē kan u bzūr u byār.	marry my wife with this sword of mine, (then) take her and bring her back”.	gōmē mnī zāmā	with my sword		inan	2	other	prep	0

8. As a general rule, only those NP’s are counted which are referential, that is, have some reference to an entity, real or imagined, definite or indefinite. The referentiality criterion is a rough guide that is helpful when dealing with things like light verb constructions, for example the following from Zaza:

   (1) *ez **mêrdi** nêkena* ‘I don't want to marry’ (lit. **‘I man** not.do’).

   The word *mêrdi* ‘man’ does not refer to any particular man, and in this context it could normally not be the antecedent for an anaphoric pronoun. Furthermore, the semantics of the expression are not compositional; we cannot guess the intended meaning ‘marry (said by a woman)’ from the individual items. We therefore treat the entire expression *merdî nêkena* as a single lexicalized predicate, and thus we have no non-subject constituents in the clause, hence no relevant tokens in this particular sequence.
9. We **do not code subject constituents**. If a clause only contains a subject (e.g. **‘The child** was asleep’), then the clause is counted as ‘not classifiable’, and "1" is entered in column Q. The utterance unit nevertheless remains in the spreadsheet so that we have a full record of the text (see #4 above)
10. 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 (see #9); instead we code the ‘wanted’ or ‘possessed’ constituent (if overtly present in the clause); see N and O under §4, detailed explanations below.
11. Any utterance that **contains no clearly codable token** receives a "1" in column Q, indicating that this line is to be skipped in the analysis, and you simply carry on with the next line.

2. Overview of the coding scheme

A-F	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.
G-J	The actual forms, and translations (some coders have not provided a translation of the actual token, because it is usually inferrable from H (context translation); this is a matter of discretion. Minimally, you will therefore need G, I and J
K-O	What we are mostly interested in: The independent variables for structure (K-O). **CONSULT EXPLANATIONS in §4-5 for details of these columns.**
P	'Position relative to the verb', either before "0", or after "1". This is the main dependent variable
Q	'non-classifiable': enter "1" here in any line that does not contain a codeable token
R	Any comments on individual items, or additional language-specific coding that analysts may wish to add

3. Detailed explanations for columns A-J

	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 Excel files
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	Contains the individual utterance units of the text, in the original transcription/orthography, some of which will be duplicated (see §1, #5 above), and some of which may not contain a relevant token.
H	context translation	Translation of context (not always provided; depends on the nature of the source text, and how much time the coder has)
I	token	Actual form of the constituent, with relevant flagging (if adpositional)
J	token translation	Translation of the token

4. Overview of all available values for the columns K-O

Note:

- YOUR FILES SHOULD **ONLY CONTAIN VALUES FROM THIS** LIST IN THE RELEVANT COLUMNS!!
- ADDITIONAL LANGUAGE-SPECIFIC TAGS CAN BE ADDED IN COLUMN R (comments)
- AVOID EMPTY CELLS, EXCEPT IN COLUMN K!!
- ALL OF K-O (EXCEPT M) OFFER THE OPTION "other" IF YOU ARE UNSURE OF HOW TO CODE THAT PARTICULAR CATEGORY

K (pro)	L (anim)	M (weight)	N (role)	O (flag)
[empty]	adj	1	abl	bare
1	adv	2	addr	case
2	anim	3	becm	circ
3	bp	4	becm-c	circ-relnoun
4	hum		ben	lvc-poss
1-bound	inan		com	other
2-bound	other		cop	postp
3-bound			cop-loc	postp-relnoun
4-bound			do	prep
other			do-def	prep-relnoun
refl			goal	relnoun
refl-1 (optional)			goal-c	relnoun-case
refl-2 (optional)			instr	
refl-3 (optional)			loc	
wh			other	
			poss	
			rec	
			rec-ben	
			stim	

5. Detailed description of coding categories in K-O

5.1 COLUMN K (pro)

In this column we capture aspects of those tokens which we assume to be pronouns, rather than full lexical NP. This is the first decision that has to be made:
'is item X pronominal or not?' If yes, then you need to fill out this column; if not, this cell **remains empty** > proceed to column L.

In many cases, the answer is obvious, but there are quite a number of cases where the answer is not so simple (expressions such as 'person' may exhibit properties typical of pronouns, or of lexical items). These are decisions that each analyst needs to make based on his or her overall assessment of the grammar of the language.

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 it is part of the verb, and therefore cannot be meaningfully coded as 'before' or 'after' the verb.
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 as bound, and then treated as any other argument.

K (pro)	Explanations
[empty]	leave cell empty if the token is not one of the pronoun types below
1	first person pronoun, singular or plural
2	second person pronoun, singular or plural
3	third person pronoun, human referent, singular or plural
4	third person pronoun, non-human referent, singular or plural
1-bound	same as "1" but prosodically dependent on a word that is not the predicate
2-bound	same as "2" but prosodically dependent on a word that is not the predicate
3-bound	same as "3" but prosodically dependent on a word that is not the predicate
4-bound	same as "4" but prosodically dependent on a word that is not the predicate
other	pronominal, but does not match any of the categories provided (for example mixed human and non-human referents); also available for various kinds of indefinite pronouns ('someone', 'no one' 'anything' etc.)
refl	reflexive pronoun, regardless of person value
refl-1 (optional)	
refl-2 (optional)	
refl-3 (optional)	
wh	interrogative pronoun

5.2 COLUMN L (animacy)

This column captures primarily semantic features of the token, but is also used if the token falls into the class of (certain) adjectives and adverbs.

L (anim)	Explanations
adj	only applicable when the predicate itself is an adjective in the language concerned, e.g. a clause corresponding to English the teacher was young; in the 'role' column you would therefore expect to have 'cop' or 'becm' (see below)
adv	a somewhat messy category for items such as 'here', 'upstairs' etc. which could be construed as having referential sense. Note that we do not code manner or modal adverbs.
anim	the NP refers to a living, but non-human, entity (e.g. 'horse')
bp	the NP refers to a human body part
hum	the NP refers to a human
inan	the NP refers to an inanimate entity ('house', 'moon')
other	any referential expression not covered by the above, or which is not unambiguously classifiable

5.3 COLUMN M (weight)

This is an opportunity to capture - very roughly - the syntactic and phonological weight of the token. There are a host of problems associated with this category; after trialling various options (how to deal with function words, clitics etc.) in the end we have opted for the smallest common denominator, namely that item which is recognized as a 'word' by the respective analysts.

Note that when quantifying weight, we do not count adpositions as part of the weight, which we consider to be part of the flagging (Column O). Thus an English constituent like 'he lived **in the village**') would be counted as "2" in this column (and "cop-loc" in column N below).

Note also that the presence of adpositions is noted in Column O 'flagging'.

M (weight)	
1	Consists of a single phonological word (might also be a single word plus one or more clitics)
2	two phonological words
3	three phonological words
4	four or more phonological words

Note that we can also fall back on another measure of weight, namely number of segments in the transcribed form of the token; this is a question to be explored at the analysis stage.

5.4 COLUMN N (Role)

This column includes a set of primarily semantically-defined roles. The main exception to this is the direct object category (do) which is defined in language-specific structural terms, and which overrides the semantics when the two are in conflict.

N (role)	
abl	source of motion ('she came **out of the house**')
addr	addressee of a verb of speech ('they spoke **to him**, asked **her**, begged **the King** etc.')
becm	'become' i.e. the final state of a change-of-state (inchoative) predicate, such as 'become X','turn into X'
becm-c	final state of a caused change-of-state predicate ('they made him **King**', she turned him **to stone')**
ben	benefactive; a person who benefits, or is disadvantaged, by an event without being directly impinged on by the action
com	comitative; a person who accompanies another participant in some action, or state ('I went to the market **with my father**')
cop	complement of a copular expression ('they were **farmers**')
cop-loc	locational complement of a copular expression ('she was **in the car**')
do	direct object, which needs to be identified on language-specific criteria such as typical case marking properties
do-def	definite direct object (which will include most pronouns), i.e. an item whose identity is recoverable from the context through previous mention or assumed deictic reference ('she took **that cup**')
goal	endpoint or destination of a verb of motion ('it fell **on the table**')
goal-c	endpoint or destination of a verb of caused motion ('he put it **on the table**')
instr	instrument for carrying out an action
loc	static location (with no implication of movement) of a participant or event
other	none of the available categories
poss	possessed in a clause expressing possession 'she had two brothers', unless the language has a HAVE verb and expresses the possessed in the same way as a direct object (do)
rec	recipient of a theme in an event of transfer, typically GIVE
rec-ben	recipient-benefactive. This is included for contexts in which it is unclear whether a particular token is the recipient, or a benefactive of an action ('he bought the apples **for us**' - recipient or benefactive?
stim	stimulus, typically of verbs of emotion, perception, desire - if they are not coded as direct objects (English 'she was afraid **of the snake**' (stim), but not 'she hates **snakes**' (coded as "do")).

5.5 COLUMN O (Flagging)

This column indicates the nature of any overt morphology that primarily serves the purpose of indicating the token's relation to the predicate.
It is probably the most difficult column to code, because this is the realm of greatest cross-linguistic variability. In order to avoid inflation of tags, it was necessary to simplify many aspects; if in doubt, please use the "other" option. The category ‘relnoun’ is outlined in §5.5.1.

O (flag)	
bare	the NP concerned is not overtly marked for case in any (obvious) way
case	the NP concerned is case-marked. We do **not** count ‘case’ that is assigned by an adposition; in such examples, the coding would be prep or postp. Similarly, genitive or other NP-internal case can be ignored
circ	circumposition (e.g. Kurmanji Kurdish *ji NP ra* ‘for NP’
circ-relnoun	item that could be analysed either as a circumposition or a relational noun (→§5.5.1)
lvc-poss	light verb complement, possessor. In some languages, an argument may be structurally incorporated into the clause as a ‘possessor’ of a light verb complement, something like ‘he did acceptance-**of the suggestion**’ = ‘accepted the suggestion’. The token here is 'suggestion', and it is expressed as a possessor of the light verb complement where ‘acceptance’ is the complement of a light verb ‘do, make’
other	none of the available categories
postp	postposition
postp-relnoun	item that could be analysed either as a postposition or a relational noun, see explanations above (→§5.5.1)
prep	preposition
prep-relnoun	item that could be analysed either as a preposition or a relational noun (→§5.5.1)
relnoun	relational noun (→§5.5.1)
relnoun-case	relational noun that carries additional case marking (→§5.5.1)

5.5.1 Relnoun and related issues

The most problematic category is what we call "relnoun" = ‘relational noun’. These are elements transparently related to nouns referring to body parts (‘head’ ‘back’ ‘tail’) or local nouns such as ‘side’ ‘top’ etc. In many languages, these are used to indicate spatial roles, and the nouns may grammaticalize to become adpositions. However, they may remain indeterminate between nouns and adpositions, making coding decisions difficult. Consider the following from Northern Kurdish:

(2) *dike* ***hundur-ê*** ***sandiq-ê***
he.put inside-EZ.M chest-OBL ‘(he) put (her) inside the chest’
[iran_kurd_northern_mush_mcnk0103, 68]

This example (and many like it) raise the following questions:
Should we treat *hundur* as the head of the NP (the *inside* of the chest), or as some kind of adposition, or something else?
For this kind of example, we introduce the label ‘reln’ as a coding option for the category of ‘flagging’. Let’s consider how the three options for coding the token in (2) would work:

The first option: *hundur* is a noun, and the head of the construction, and *sandiq* is a possessor. Thus we have a complex NP with two words, hence ‘2’ for weight. Complete coding would then be as follows:

I	J	K	L	M	N	O	P
hundurê sandiqê	inside the chest		inan	2	goal-c	bare	1

The second option: *hundur* is a preposition, thus part of flagging and not counted for weight, and *sandiq* is the relevant token, Then it will be coded as follows:

I	J	K	L	M	N	O	P
hundurê **sandiqê**	inside the chest		inan	1	goal-c	prep	1

The third option: *hundur* is a relational noun, a special type of flagging that will count for weight:

I	J	K	L	M	N	O	P
hundurê sandiqê	inside the chest		inan	2	goal-c	relnoun	1

I actually opted for option 1 in my data set, because *hundur* carries a regular ezafe, like any other nominal head, linking it to *sandiq*; but I could have taken option 3, given what I know of Northern Kurdish and the way this particular item *hundur* behaves (namely as indeterminate between an adposition, noun, and sometimes adverbial).

Note that relational nouns may precede or follow a noun, and this can also be (optionally) specified, as for example the following:

(3) *tɛ* *bætundi* *inta=rɛ* *bævɛri* ***tæš-e*** ***kenar****?*
you can this=ACC carry fire-GEN side
'can you carry this beside the fire'
[iran_mazand_borjian, 27]

This has been coded as follows:

I	J	K	L	M	N	O	P
tæš-e kenar	beside the fire		inan	2	goal-c	postp-relnoun	1

Coding decisions are ultimately a question of degree of grammaticalization of the relational noun in this kind of construction. There are many diagnostics for this, and it is up to the discretion of the individual annotator to decide.

A further quite common issue in this respect is when relational nouns are used together with pronominal elements, which are attached as clitics or affixes. The following example from Turkic is typical:

(4) *došov* *dökærdelær* ***ičinæ*** *...*
grape_syrup pour.HAB.PST.3PL into_it
'(they) would pour grape syrup **into it**'
[turk_bayat_#86]

The bold-face form *ičinæ* is based on the relational noun *ič* 'interior', to which a 3sg possessive marker is attached (*-i*), followed by a dative case marker *-næ*.
In this kind of example, the referential item (token) we are interested in is actually **the possessive marker, which we treat as a bound pronominal element**. Thus the coding of this example in the columns K-P is as follows:

The complete glossing of the token in (4) is as follows:

K: 4-bound (pronominal, inanimate, bound)
L: inan (inanimate)
M: 1 (a single phonological word)
N: goal-c (goal of a verb of caused motion ('pour'))
O: relnoun-case (there is a combination of relnoun and case marking)
P: 1 (position after the predicate 'pour')

In sum, relational nouns pose considerable challenges for coding, and we do not expect to find a simple solution that will cover all eventualities, in all languages. The options that are available should, however, cover most of what you encounter; if you cannot decide, opt for 'other', or leave the token out entirely ('1' in Q), or make a note in R 'comments', collect all your disputable examples, and discuss with us. Don't waste time agonizing about the coding of individual tokens!