Guidelines for Annotating Irish MWEs

These guidelines have been partially adapted from PARSEME Annotation Guidelines.

As a note, embedded MWEs are not annotated in this pass. For example, the **VPC** *cuir suas* in the **IAV** *cuir suas le* would be annotated as **IAV**, to match the outermost MWE label.

Similarly, when annotating MWEs in isolation (e.g. in a lexicon), it is important to note that where MWEs may be part of a larger MWE, these guidelines are intended to be applied to the text as is, rather than potential MWEs: e.g. *in eolas* occurs in verbal MWEs *cuir in eolas* and *bí in eolas*, but should be annotated as FE where not included in those contexts.

Category	**Example**	**Tag**
Nominal Compounds	*mac tíre*	NC
Named Entity	*Baile Átha Cliath*	NE
Institutionalised Phrase	*domhan uile*	IP
Copular Construction	*is maith le*	CC
Light Verb Constructions	*déan obair*	LVC
Inherently Adpositional Verb	*bain le*	IAV
Verbal Idiom	*an lá atá inniu*	VID
Fixed Expression	*in aghaidh*	FE
Verb Particle Construction	*leag amach*	VPC

Nominal Compound (NC)

Nominal compounds (**NCs**) are compound noun phrases that consist of a head noun and a dependent noun or adjective, where the compound can be said to be semantically semi-compositional or non-compositional.

Also included in this category are technical or specialised language terms, such as those included in the Tearma corpus.

Test NC.1: Noun phrase is non-compositional

Does the combination of noun and noun, or adjective and noun, lend a non-compositional or semi-compositional meaning to the noun?

If yes, annotate as **NC**.

mac tíre
mí na meala

If not, continue to test NC.2.

nósanna imeachta
cothrom an lae

Test NC.2: Specialised term

Can the construction be considered a technical term or does it have a specific meaning within a certain domain, even if the meaning is compositional?

NB: Included in these specialised terms are collective nouns such as *saithe beach* “swarm of bees” and *scuaine lachan* “flock of ducks”

If yes, annotate as **NC**.

uiscebhealaí intíre
aigéad sulfarach
scuaine lachan

If not, continue to test NC.N.

grúpa cannon
taifead ábhartha

Test NC.N: Numbers

Does the construction form a number?

If yes, annotated as **NC.N**

Ceathrar déag
Fiche a haon

If not, do not annotate.

Named Entities (NE)

Named entities are all proper noun phrases that make up a recognisable entity; these include but are not limited to proper names, organisations or agencies, place names, compound months and foreign titles.

The category is similar to the MWE flat relation in the UD guidelines for the IDT, however there are some differences. While the flat relation is always used for syntactically flat constructions (i.e. there is no internal syntax or hierarchy), named entities may include syntactically structured constructions such as official committee titles (e.g. A*n Ghníomhaireacht Eorpach chun Comhar Oibríochtúil a Bhainistiú ag Teorainneacha Seachtracha Bhallstáit an Aontais Eorpaighle*). Also, syntactically flat constructions that do not form a recognisable entity, such as dates, are annotated with the flat relation, but are not considered here as named entities, due to their productivity.

Test **NE**.**1**: Noun phrase forms an entity

Does the construction take the form of a noun phrase that represents some recognisable entity?

If yes, annotate as **NE**.

Deireadh Fómhair
Na Náisiún Aontaithe
Comhairle Cathrach Bhaile Átha Cliath

If not, do not annotate.

Fixed Expressions (FE)

Fixed expressions (**FE**) are those that have no internal modification permitted, and do not inflect for any grammatical change. This category closely aligns with the fixed label in the UD annotation guidelines, but is expanded slightly to include certain semi-fixed constructions, which closely align with other expressions in this category.

Test **FE.1**: Does not contain content words

Does the construction contain content words which contribute specific semantic meaning to the phrase as a whole?

If yes, continue to test **NVID.**

Chomh críonna le sionnach

If no, or unsure, continue to test **FE.2**.

Faoi láthair
Ar leith

Test **FE.2**: Unit of language

Does the construction form a linguistic or semantic unit, i.e. all the words in the construction contribute towards a **non-productive constituent** whose words cannot be replaced by others of the semantic class, and in which the entire construction modifies the sentence in some particular way.

If yes, or unsure, continue to test **FE.3**.

Ceart go leor
Chomh maith
Go dtí

If no, do not annotate.

Ar an
Bhain sé

Test **FE.3:** Syntactically fixed

Can the construction undergo inflection or internal modification?

If yes or unsure, continue to test **FE.SEMI.**

ina dhiaidh

If not, annotate as **FE**.

Tar éis
Le linn
Ar chor ar bith

Test **FE.SEMI**: Syntactically semi-fixed

Is the modification or inflection that can occur only to show grammatical information such as number or person? If yes, annotate as fixed.

ina dhiaidh → i mo/do dhiaidh

Otherwise, do not annotate.

Institutionalised Phrases (IP)

Institutionalised Phrases (**IP**) are expressions whose meaning is not idiomatic, but the lexical items become fixed through conventions of language use. MWEs in this category may be similar to those annotated as fixed expressions, but where the former are non-productive, shorter constructions that modify the sentence in some way, institutionalised phrases can take the form of longer, more complex constructions.

Test **IP.1**: Meaning of the lexicalised terms is not idiomatic

Do any of the lexicalised items have a meaning that could be understood to be idiomatic or semantically opaque in some way?

If yes, continue to test **NVID**.
Idir dhá thine Bealtaine

If no, continue to test **IP.2**.
Idir dhá roghanna

Test **IP.2**: Fixed lexical usage

Does conventional language use tend to favour a certain selection of lexical items, rather than a semantically equivalent lexical item?

If yes, annotate as **IP**.
Gruth agus meadhg
Seomra is cistin

If no, do not annotate.
Greim docht

Non-Verbal Idioms (NVID)

This category defines idiomatic expressions with no verbal head. These constructions differ from fixed expressions in that they include content words and are not limited to short functional units of language, and differ from institutionalised phrases as they are semantically idiomatic.

(e.g. *idir dhá thine Bhealtaine*)

Light verb constructions (LVC)

Light verb constructions (**LVC**) are formed by a verb v and a (single or compound) noun n, which either directly depends on v (and possibly contains a case marker or a postposition), or is introduced by a preposition.

Bain triail as
Cuir lúchair ar
Déan iarracht ar
Cuir fuil-shrón le
Déan dearmad ar
Tabhair faoi deara
Déan dreas cainte

The (single or compound) noun n is predicative and refers to an event (e.g. decision, visit) or a state (e.g. fear, courage). Predicative nouns are nouns that have semantic arguments, that is, they express predicates whose meaning is only fully specified by their semantic arguments.

Test **LVC.1**: Noun is abstract

Is the noun abstract (i.e. does it denote a quality, state, or idea)?

If yes, continue to test **LVC.2.**

dearmad
trial
fearg

If no (i.e. denotes a concrete object), do not annotate.

cathaoir
lámh
leabhar

Test **LVC.2**: Noun is predicative

Does the noun n have at least one semantic argument, implying that it is a predicative noun?

If yes, continue to test **LVC.3**.

Tabhair cuairt ar → event with two arguments: the visitor and the visitee

If no, do not annotate.

Cuir tuí

Tabhair peann ar

Test **LVC.3**: Verb's subject is noun's semantic argument

Is the subject of the verb a semantic argument of the noun? In other words, is the verb linking the predicative noun to one of its semantic arguments that occurs as the subject of the verb?

If yes, continue to test **LVC.4**
Chaith Seán vóta → Seán is the subject of the verb and a semantic argument (the voter) of the noun

If no, do not annotate.
Chomhair Séan na vóta → Vote does not have a semantic argument of the counter

Test **LVC.4**: Verb with light semantics

Is v semantically light, that is, is the semantics that v adds to n restricted to: (i) what stems from its morphological features (e.g. future, plural, perfective aspect, etc.), (ii) pointing at the semantic role of n played by v's subject?

If yes, continue to test **LVC.5**.
Rinne mé iarracht ar m'obair bhaile→rinne adds no meaning to *iarracht* except performing an activity

If not, go to test **LVC.6**
Thosaigh mé iarracht ar m'obair bhaile → *thosaigh* has an added aspectual meaning

Test **LVC.5:** Verb reduction

Try to build an NP without the verb, in which v's subject s becomes n's dependent. You might need to test several prepositions, possessives, case markers, etc, as long as you use no verb. Can this verbless NP refer to the same event or state as the candidate v+n construction does? (This is a test using an ownership construction).

If yes, annotate as **LVC.full**.
Déanann Mícheál cur síos ar →an cur síos do Mhícheál
Thug mé tacaíocht do Mháire → mo thacaíocht féin

If no, do not annotate.
Fuair Máire tacaíocht ó Sheán → an tacaíocht a fuair Máire

Test **LVC.6:** Verb's subject is noun's cause

Is the subject of the verb expressing the cause of the predicate expressed by the noun? In other words, does the verb bring an additional participant to the scene, representing the source or cause of the event or state referred to by the noun?

If yes, annotate as **LVC.cause**

> *Chuir Aoife áthas orm* → The happiness was inspired by Aoife

If not, do not annotate.

> *Chuir Aoife airgead sa bhanc* → The money was not caused by Aoife

Verb-particle constructions (VPC)

Verb-particle constructions (**VPCs**) are formed by a lexicalized head verb v and a lexicalized particle p dependent on v. Examples include constructions such as *tarraing anuas, cas as, tabhair amach*, etc.

The meaning of the **VPC** can be fully or partly non-compositional. In fully non-compositional VPC VPC.full the change in the meaning of v goes significantly beyond adding the meaning of p.

> *Cas as*
> *Tabhair amach*

In semi-non-compositional VPCs VPC.semi, p adds a partly predictable but non-spatial meaning to v.

> *Tabhair suas*
> *Glan suas*

Test **VPC.1**: Verb without the particle refers to the same event/state

Can a sentence without the particle refer to the same event/state as the sentence with the particle? Special care must be taken when the same construction might or might not be a valid **VPC** depending on its context.

If no, annotate as **VPC.full**

> *Cas as* → ‘put out' does not imply to turn

Tarraing anuas → draw down (i.e. bring up) does not imply to draw in a literal sense
Seas amach → stand out does not imply to stand

If yes, go to test **VPC.2**.
Scríobh síos → scríobh implies to write
Glan suas → glan implies to clean
Féach amach → féach implies to look
Éirigh amach → éirigh against implies to rise

Test **VPC.2**: Spatial particle

Is the particle spatial in the context of the verb, i.e. does it express direction or position?

If not, annotate as **VPC**.
Glan suas → suas is not directional here, but rather implies completely
Éirigh amach → amach implies 'out' in a rebellious way

If yes, do not annotate.
Seas le chéile
Féach amach
Test **VPC.3**: Spatial particle in a literal reading
Does the VPC candidate have a literal counterpart in which the particle is spatial, i.e. expresses direction or position?

If not, annotate as **VPC.semi**.
Éirigh amach éirigh amach an bosca
Glan suas glan suas an balla

If yes, it is not a **VPC**, exit.
Lig isteach

Inherently adpositional verbs (IAVs)

Inherently adpositional verb (**IAV**) is considered a special and experimental category. It consists of a verb or VMWE and an idiomatic selected preposition that is either always required or, if absent, changes the meaning of the verb of VMWE significantly. **IAV** constructions should be annotated only after annotating **LVC** or **VPC** constructions, since this category can overlap with these two. However, we do not consider this category as overlapping with either **CC** or **VID** categories, so they must be annotated after this category.

This definition of inherently adpositional verbs is a generalisation (applying to many languages) of the annotation guidelines of the English STREUSLE corpus, which define guidelines for annotating prepositional verbs.

IAVs are verb+adposition combinations in which:

- the dependents of the adposition are not lexicalized
  - to stand for something is annotated as IAV because the object is not lexicalized, but in the ID to take something for granted, to take for cannot be annotated as IAV because granted is also lexicalized in the ID
- the adposition is integral, that is, "it cannot be omitted without markedly altering the meaning of the verb"
  - ‘to rely on’ → ‘to rely’ can never occur without the preposition on
  - ‘to count on’ → ‘to count’ can occur without the preposition, but it will never have a sense of ‘to depend/rely on’
- Note that idiomatic adpositional valency, in which the adposition opens a slot for a complement, should not be mistaken for verb-particle constructions. Tests distinguishing particles from prepositions can be used to disambiguate these categories.
  - ‘to wake up somebody’ cannot be annotated as IAV because ‘up’ is a particle, and not a preposition.
- Particles can occur after the object:
  - ‘to wake somebody up’,
- But prepositions cannot
  - ‘*to come a new restaurant across’
- Not only single verbs but also VMWEs may be inherently adpositional. This is why IAV annotation needs to be the last step, after all other VMWEs in a sentence have been identified and categorised. In case of overlap between another category and IAV, the whole VMWE annotation needs to be repeated with the addition of the lexicalized adposition, and the whole is annotated as an IAV
  - ‘to put up with’ bears 2 annotations:
    1. ‘to put up’ is annotated as VPC
    2. the whole sequence ‘to put up with’ is annotated as IAV

Éirigh as
Buail le

Test **IAV.1**: Circumstantial question with no adposition

Note: This is an adaptation of STREUSLE's guideline on prepositional verbs by Nathan Schneider and Meredith Green.

In response to a declarative sentence with the verb+adposition combination, is there a natural way to query the circumstances of the verbal event using the verb, but not the adposition?

If not, annotate as **IAV**.

Cuireann sé sin orm → Cén sort rud a gcuireann tú?→ Cuir ar is annotated as **IAV**

If yes, do not annotate.

Sheas mé ar an mbord → Cén fáth ar sheas tú ann? → Seas ar is not annotated as **IAV**

Note: This category requires further investigation, as directional adverbs in Irish have the property that they can inflect to indicate the number and person of the object. The third-person singular inflected form often appears the same as the uninflected preposition.

- i.e. *as* could be interpreted as the uninflected preposition 'out of' or the inflected preposition 'out of it'

As such, it is not certain whether the verb head in these constructions can be considered intransitive, and thus the relationship with IAV and VPC becomes more complex.

Idiomatic Copular Constructions (CC)

Idiomatic copular constructions are constructions formed with the copula and one or more arguments, where the construction has a meaning that is non-compositional from its component words.

Test **CC.1**: Non-compositional meaning

Is the meaning of the construction the same as the sum of its parts (i.e. do the components of the construction add meaning beyond their individual meanings to the construction?).

If yes, annotate as **CC**.

Is le → idiomatic construction indicating possession

Is maith le → idiomatic construction indicating enjoyment

If no or unsure, continue to test **CC.2**.

Test **CC.2**: Lexical inflexibility

Are each of the components of the construction lexically inflexible, so that replacing one token with another from the same semantic class would be incorrect?

If yes, annotate as **CC**

Is chóir →is ceartas ungrammatical despite the similar meaning of *ceartas* and *cóir*

Verbal Idioms (VID)

Verbal idioms constitute a universal category. A verbal idiom (**VID**) has at least two lexicalized components including a head verb and at least one of its dependents.

Is buí le bocht an beagán
Ag cur madraí i bhfuinneoga
Tá dhá thaobh ar an mbád
Déan cat is dhá eireaball ar
Gléasta go barr na méar
Moll an óige agus tiocfaidh sí

Idiomatic constructions with the copula are currently not annotated as **VID**, given that the syntactic head of the construction is not the verb in these cases:

Máire is ainm dom → head is noun *ainm*

Test **VID.1** - [CRAN] - Cranberry word

Does the candidate expression contain a cranberry word? If yes, annotated as **VID**, otherwise proceed to next test.

Test **VID.2** Fails other MWE tests

Does this construction fail as a different type of MWE (LVC, VPC, IAV, or CC)?

If yes, continue to test **VID.3**

Caith an phingin i ndiaidh an phuint
Bulla dall a dhéanamh de

If no, annotate as appropriate.

Test **VID.3**: Non-compositional meaning

Is the construction semantically non-compositional, i.e. the meaning cannot be derived entirely from the lexicalised components?

If yes, annotate as **VID**.

Bheith ar an bpláta beag → sense of being in jail is not evident from components
Cat a scaoileadh as an mála → sense of telling a secret not evident from components

If no, do not annotate.

Bheith ag obair

Madra a fheiceáil ar an mbóthar