Published Annotation Guidelines (English Version)

These guidelines are intended to assist with manually assigning compositionality and additional metric scores to noun compound candidates in Irish, in order to gain perspective on the definition of the Irish noun compound, which remains an open question (McGuinness, Phelan, et. al, 2020). Compositionality is a known, but difficult to define, feature of language. For the purposes of these guidelines, we consider the compositionality of the construction to be the degree to which the components of the construction adhere to normal language usage and meaning.

Currently, the data is uploaded to INCEPTION in CoNLL-U format and each noun compound candidate is annotated with the five metrics listed below. The guidelines are actively being developed as part of ongoing work on Irish multiword expressions at the ADAPT Centre, Dublin City University, and will be updated as the project develops.

Until the precise definition of Irish noun compound has been determined, we use noun compound candidate (NCC) for all syntactically viable constructions annotated.

Annotation Steps

1. Extracting Noun Compound candidates (NCCs) from the sentence

The noun compound candidates we currently consider are two-word NCCs consisting of a noun and a noun or a noun and an adjective. Idiomatic constructions, terminology, and institutionalised phrases are considered as NCCs for the purposes of this annotation work. Named entities are also extracted, but are not annotated as NCCs, rather, are given a NE label.

1. Begin by reading the sentence and noting the noun phrases
    1.1. In the case of an embedded noun phrase, identify each nested noun phrase as a separate candidate, and select the most granular noun phrase from this construction
        1.1.1. E.g. *Ollscoil na hÉireann* ***Maigh Nuad***, ***láithreán gréasáin*** *na* ***Seirbhíse Cúirteanna****, Baile* ***Átha Cliath***

1.2. The noun phrase must not be partly composed of a compound construction (e.g. compound preposition) where the attachment to the other elements of the compound construction is “tighter” than the attachment between the nouns
    1.2.1. E.g. *ceisteanna* ***ar nós*** *rochtain* ***d'***airgead
1.3. Potential NCCs may appear to be viable through ambiguous word placement. By consulting the dependencies provided by the treebank annotations or UDPipe, these ambiguities may be resolved*
    1.3.1. E.g. *bhí sé i* ***riocht éan*** *a thainic san ubhallghort a caitheadh* > *Riocht éan* should not be extracted as an NCC

2. If the noun phrase appears to be a **named entity,** exit the tests and go to test 6.
    2.1. E.g. *Maigh Nuad*
3. Locate the head word in each noun phrase
    3.1. E.g. *lá breá*
4. Identify what words are modifying the head word, and select the modifier that most tightly adheres to the head noun
    4.1. E.g. *comórtas drámaíochta scoile* (*comórtas drámaíochta* rather than *comórtas scoile*)
5. The head noun can be a common or proper noun modified by either a noun (common or proper) or an adjective (see special cases)
        5.1.1. E.g. *siopa earraí, cailín rua*
6. The modifying word must immediately follow the head noun (e.g. only whitespace separates the words)
    6.1. This discounts non-contiguous compounds
        6.1.1. E.g. *cúig lá déag*
    6.2. The Part of Speech of the modifying words provided by treebank annotations or UDPipe* can help to clarify in cases where the POS is not clear
        6.2.1. E.g. *madra eile* (*eile* modifies *madra* as a demonstrative determiner, not an adjective)

Special cases

1. Where foreign words are used as part or all of the compound which otherwise follows the rules above (two words separated by a whitespace, acting as a noun phrase, that are not attached more tightly to another word), these can be extracted as NCCs
    1.1. E.g. *fear* **off-license**, **big fish**, **President** Biden
    1.2. Where these NCCs form NEs, they should be treated as other NEs, i.e. not annotated as NCCs
2. Numerals are not considered as valid heads or modifiers.
    2.1. This applies to all instances of numerals, including figures (e.g. *15 teidil*, *sos 4r.n.*, *Buiséad 2022*), count nouns and cardinal numbers (e.g. *dhá rogha, an triúir acu, duine amháin*), and ordinal numbers (*céad lá, dara leath)*
3. Demonstratives and quantitative are not considered as valid heads or modifiers.
    3.1. E.g. *Tuilleadh eolais*, *roinnt airgid*, *iomarca oibre* > do not extract as NCC
    3.2. E.g. *An ghort seo, An seomra siúd, cén háit* > do not extract as NCC

4. Definitive articles attached to either the head or modifying noun are **not** included in the noun compound candidates
    4.1. E.g. ***an*** *deireadh seachtaine* > extract *deireadh seachtaine*
    4.2. E.g. *Fear* ***an*** *phoist* > do not extract as NCC
5. Verbal adjectives are treated like regular adjectives
    5.1. E.g. *lámh breóidhte* > extract as NCC
6. Hyphenated words are treated as one word
    6.1. E.g. *crua-adhmaid* > Do not extracted as NCC
7. Similarly, where a whitespace is included, even if normally it wouldn't be included, treat the construction as two different words
    7.1. E.g. *Sráid bhaile* > Although commonly written as *sráidbhaile*, in this case, extract as NCC
8. In cases of words with mispellings, non-standard spellings, or dialectal differences, attempt to understand the most likely meaning of the phrase in it's context.
    8.1. E.g. *cúirt bréagh* > Likely a dialectal version of *cúirt breá,* not *cúirt bréag*, as indicated by the context

*Note: Dependencies generated by UDPipe have not been manually corrected and may contain inaccuracies

2. Annotating NCCs with compositionality score

Compositionality as a feature of language roughly equates to the **'expectedness' in semantic behaviour of the components of the NCC,** arising from Ferge's principle of compositionality (i.e. the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them).

Unexpected behaviour may appear as an NCC that contains implied real-word knowledge; an NCC whose meaning is idiomatic; where meaning of an NCC was once compositional, but the meaning of words have changed over time; an NCC's meaning is derived partially from one or both of the components, but there is an unexpected element to the meaning of the NCC as a whole.

Compositionality scores range from 0 (totally opaque/idiomatic/odd) to 5 (totally transparent/compositional/conventional). The following rules should help to determine this value.

1. When assigning a score, consider how **each word** in the compound contributes to the overall meaning

1.1. If the meaning of the whole is not determined by the head noun or the modifying word **alone**, it is non-compositional to some degree.
1.2. Very idiomatic (semantically odd) NCCs will have a meaning very different from either of the words in the compound noun
    1.2.1. E.g. *mac tíre*
1.3. Mostly compositional NCCs will have a meaning that is plausible from the component words, but may required extra information encoded by context or familiarity of use
    1.3.1. E.g. *tonn teasa*
1.4. Very compositional (conventional) NCCs are wholly interpretable from the elements of the compound, and the meaning of the NCC could be derived from the meaning of the individual elements
    1.4.1. E.g. *sraith cartún*
1.5. Consider how much external (real-world) knowledge you're applying to the meaning of one or both words.

2. Compositionality should be scored based on contemporary understanding of the construction and/or constituent words rather than historical understanding.
    2.1. E.g. *Maigh Nuad* (plain of-Nuad) Maynooth is no longer a plain owned by King Nuadha
    2.2. Names of people, places & groups that were originally named meaningfully, and the name has lost meaning over time, should be considered non-compositional/idiomatic/odd
        2.2.1. E.g. *Cathair Baile Átha Cliath* (does not mean City of the Town of the Hurdled Fords)
        2.2.2. E.g. Aisling Breathnach (this name likely does not intend to mean "dream of Welsh origin")
3. Consider the amount of contextual knowledge needed for some broad-meaning words (*cúrsaí*, *bealach, córas*)
    3.1. If these words retain their original meaning when combined with another word in an NCC, they are still considered compositional, even if their meaning changes depending on the context
        3.1.1. E.g. *Cúrsaí reatha* vs *cúrsaí ollscoile*
    3.2. Words of the same semantic class may combine productively with these broad meaning words
        3.2.1. E.g. *Lucht oibre*, *lucht féachana*, *lucht leanúna* (although each NCC represents a different context, lucht contributes the same semantic meaning for each NCC)
    3.3. To test this, consider words of a similar semantic class, and whether the broad meaning word retains the same semantic characteristics in that context
        3.3.1. E.g. *córas oideachas/sláinte/eacnamaíochta*
        3.3.2. BUT *Slí beatha*
4. When annotating NCCs that are composed of or include foreign words, consider the foreign word as contributing non-compositional meaning
        4.1.1. E.g. *Cor Ochtar*, cul de sac, *Tógra* EPOC III+

5. As a final step, consider all the scores assigned to the NCCs annotated previously and make any adjustments to your scores as needed

Justification for 6-point scale

Using an even numbered Likhert scale allows for translation of compositionality score to a binary feature of the construction. The ability to encode compositionality as a binary feature can provide more options at the representation stage of word/token processing, and can provide useful information about the relationship between constituent word compositionality and construction compositionality. When compared to using a 5-point scale, the task was not found to be noticeably simpler. A four-point scale was found to be insufficient for capturing the difference between moderately compositional (4) and slightly compositional (3) NCCs, likewise with moderately non-compositional (1) and slightly non-compositional (2) NCCs, which meant many NCCs were allocated to the same compositionality score when the compositionality values were considered different by the annotators. Finally, using the same range of values [0-5] as Reddy et al., Cordeiro et al., and Garcia et al. allows for more direct comparison between Irish and these other languages.

3. Domain specificity scores

Domain specificity is a metric capturing how strongly the NCC is associated with a particular domain. The score ranges from 1 (general domain language) to 3 (domain specific language). When assigning a score, consider how much knowledge of particular domain is required to understand the meaning of the construction

1. Constructions annotated with 1 are those one would expect to encounter in general, daily life, and the meaning tends to be more accessible and pertaining to the aspects of life that are held in common with most people. A language learner could reasonably expect to encounter this NCC.
    1.1. E.g. *Tréimhse ama*
2. Constructions annotated with 2 are those associated with a particular topic and/or special knowledge or experience. A language learner may not be familiar with the term, but an fluent adult speaker of the language could reasonably expect to have encountered this NCC.
    2.1. E.g. *Teach scoite*
3. Constructions annotated with 3 are those one would expect to require special training or education to understand; the meaning tends to be more specialized and pertain to aspects of life or work that are not generally experienced or understood. A fluent adult speaker may not have encountered this NCC.
    3.1. E.g. *maitrís chóngarachta* (adjacency matrix)

4. Annotator familiarity scores

Annotator familiarity is a metric to indicate how much familiarity the annotator has with the NCC. The score ranges from 1 (not familiar at all with the NCC) to 3 (highly familiar with the NCC). When assigning a score, annotators must consider whether they have seen the NCC or its constituent words before and how often they see or use such NCCs.

1. Constructions annotated with 1 are those that the annotator have not seen before and do not recognise or those that the annotator is not sure of the meaning.
2. Constructions annotated with 2 are those that the annotator has seen or used rarely.
3. Constructions annotated with 3 are those that the annotator sees or uses on a regular basis and fully recognises.

5. Confidence scores

Confidence scores indicate how confident the annotator is with the extracted NCC and the scores assigned.

1. A high confidence score (3) means the annotator believes they have given the right score, even if another annotator might annotate the construction differently.
2. A confidence score of (2) means the annotator believes they have given the right score, but is aware they may have interpreted the guidelines incorrectly or misunderstood the meaning of the construction.
3. A low confidence score (1) means the annotator is unconfident about the scores assigned, because the meaning of the construction was not understood, or the annotation guidelines are not sufficiently clear on how to annotate this construction, or another reason.
4. Confidence scores can be reviewed after annotation.

6. Named entity

Named entities are treated separately to other NCCs, as they do not exhibit compositionality in the same way that other nouns do.

1. Annotate as named entity where the NCC refers to the name of a person, place, organisation, party, or other entity.
    1.1. E.g. Teresa Clifford, *Port Láirge*, Sinn Féin, *Contae Corcaigh*

2. Where the context of the sentence makes it clear that the referent is a particular entity, even when the NCC would not necessarily be a NE outside of that context, annotate as NE
    2.1. E.g. *Leabharlann Ceoil sa Lárleabharlann* > This title refers to a particular section of an identifiable Central Library, there is no other library that it can be referring to in this context