Nominal MWEs - Annotation Guidelines

Version 0.1 (in progress – pilot annotation due)

Contributors: Agata Savary, Voula Giouli, Carlos Ramisch, Stella Markantonatou, Sara Stymne

1st UniDive Training School

Annotation process and decision diagram

We propose the following methodology for MWE annotation:

- **Step 1** - identify a candidate, that is, a combination of words which could form a MWE. Recall that a candidate can be composed of only one token if it contains several words (cf. the MWT tests). If the candidate has the structure of a meaning-preserving variant, find the corresponding canonical form. This is non-trivial only for VMWEs, since for all other categories, any meaning-preserving variants are considered canonical forms. The following steps should be applied to this canonical form. This step is largely based on the annotators' linguistic knowledge and intuition after reading this guide.
- **Step 2** - determine which components of the candidate (or of its canonical form) are lexicalized, that is, if they are omitted, the MWE does not occur anymore. Corpus and web searches may be required to confirm intuitions about acceptable variants.
- **Step 3** - apply the generic decision diagram to the candidate's canonical form, in its reformulated context (looking at its lexicalized components). Notice that your intuitions used in Step 1 to identify a given candidate are not sufficient to annotate it: you must confirm them by applying the tests in the guidelines.

The decision tree below indicates the order in which tests should be applied in **step 3**. The decision trees are a useful summary to consult during annotation but contain very short descriptions of the tests. Each test is detailed and explained with examples in the following sections.

Decision Tree for Nominal MWEs

Test NMWE.3 - [SPECIF_REF] - Specific referent: Is the candidate used in context to refer to a single specific entity in the discourse world?

⇒ ***Yes*** => Apply test NMWE.4

- (en) ***John Smith*** *showed up unexpectedly.*
- (en) *He used the* ***cold weapon*** *hidden under his coat →* ***cold weapon*** *refers to a specific weapon*
- (en) *The* ***theory of relativity*** *was proposed by Einstein.*
- (en) *The* ***UN Secretary-General*** *visited Greece.*

⇒ ***No*** => Apply test NMWE.6

- (en) *Many* ***Johns Smiths*** *live in London*.
- (en) *A* ***cold weapon*** *is a weapon that does not involve fire or explosions →* ***cold weapon*** *is used generically, i.e. refers to all instances of a class*
- (en) ***Cold weapons*** *are prohibited on a plane →* ***cold weapon*** *is used generically, i.e. refers to the whole class*
- (en) *The two* ***cold weapons*** *were found at the place of the crime →* ***cold weapon*** *refers to several instances of the class*

Test NMWE.4 - [CONCEPT_NAMING_CONV] - Concept naming convention. Does the naming convention between the candidate c and an entity e refer to all instances of a whole concept, i.e. can c refer to another entity e' based on the properties of e', with no need of an extra naming convention?

⇒ ***Yes*** ⇒ Apply **test (NMWE.5)**

⇒ ***No****, because there could be no other e' =>* Go to the next test (NMWE.5)

(en) *The* ***UN Secretary-General*** *visited Greece* → There is currently no other UN Secretary-General

(en) *quantum physics, president of the US*

⇒ ***No*** *=> this is most likely a proper name,* ***exit***

Test NMWE.5 [**SEM_TYPE]** Is the entity e referred to by c a PERSON, ORGANIZATION, LOCATION, HUMAN PRODUCT or EVENT?

⇒ ***Yes*** ⇒ *this is most likely a proper name,* ***exit***

⇒ *No* ⇒ *Go to the next test*

Test NMWE.6 - [CRAN: Cranberry word?] Does the candidate contain a cranberry word?

⇒ ***Yes***, annotate as a **NID**

(en) ***cran|berry*** – ***cran*** is a cranberry word
(en) ***status quo*** – foreign words like 'status' and 'quo' are considered cranberry words
(el) ***δούρειος ίππος*** durios ipos 'Trojan horse'- ***δούρειος*** is a cranberry word (obsolete word for wooden)
(el) ***άρες μάρες κουκουνάρες*** ares mares kukunares lit. ares mares pine-cones 'nonsense' - both ***άρες*** and ***μάρες*** are cranberry words that simply rhyme with κουκουνάρες kukunares 'pine-cones'
(sv) ***körsbär*** Lit. *körs*|berry Trans. 'cherry' – ***körs*** is a cranberry word
⇒ ***No***, go to the next test
(en) *eager beaver* – both 'eager' and 'beaver' are stand-alone words
(el) *αιχμή του δόρατος echmi tu doratos* both αιχμή and δόρυ are stand-alone words
(sv) *blåbär* Trans. 'blueberry' – both 'blå' and 'bär' are stand-alone words

Test NMWE.7 - [MORPH] Does a regular morphological change lead to ungrammaticality or an unexpected meaning shift?
⇒ ***Yes***, annotate as a **NID**
(en) **bits and pieces** - #bit and piece
(el) **χωρικά ύδατα** chorika idata 'territorial waters' - χωρικό ύδωρ choriki idor # territorial water

⇒ ***No***, go to the next test
(en) light year - light years
(el) έτος φωτός etos fotos lit. year.SG light.GEN.SG light year - έτη φωτός eti fotos lit. year.PL light.GEN.SG light years

Test NMWE.8 - [IRREG_STRUCT] Does the candidate have an irregular internal syntactic structure?
⇒ ***Yes***, annotate as a **NID**
(en) ***secretary general***

⇒ ***No***, go to the next test
(en) general secretary

Test NMWE.9 [IRREG_STRUCT_DISTRIB] Does the candidate have an internal structure which is unexpected for its (external) distribution, i.e., it does not have the distribution of a nominal?
⇒ ***Yes***, annotate as a **NID**

(fr) **à-coup** – this expression functions as a noun, but has an internal PREP+N structure
⇒ ***No***, go to the next test

Test NMWE.10 [INSERT] Does regular insertion of modifiers (adjectives, relative clauses, adverbs, determiners, PPs, etc.) for internal components of the candidate result in ungrammaticality or unexpected meaning shift?
⇒ ***Yes***, annotate as a **NID**
(en) very **cold weapon**
⇒ ***No***, go to the next test

Test NMWE.11 - [SYNT: Does a regular syntactic change that would normally be allowed by general grammar rules lead to ungrammaticality or to an unexpected change in meaning?]
⇒ ***Yes***, annotate as a **NID**
(en) **state-of-the-art -** #art state
⇒ ***No***, go to the next test

Test NMWE.12 [COORD] Does coordination of the candidate with another candidate of the same head lead to ungrammaticality or unexpected meaning shift?
⇒ ***Yes***, annotate as a **NID**
(en) *****blue****- and black****berry***
⇒ ***No***, go to the next test –
(en) *car and boat traffic*

Test NMWE.13 - [ID:Is the semantic head h of the candidate c its hypernym, which can be reformulated by "is c a type of h"? Note that sometimes the semantic and syntactic heads do not coincide.]
⇒ ***No,*** annotate as a **NID**
(en) ***white elephant*** – It is not a type of elephant, it is a valuable possession
(en) ***white elephant*** – It is not a type of elephant, it is a valuable possession

(sv) ***jordgubbe*** Lit. earth|man Trans. strawberry – It is a berry, not a man
(fr) ***cordon bleu*** 'lit. cord blue' 'good cook' is not a cordon 'cord'
(ro) ***etaje climatice*** 'climate floors' - *etaj* 'floor' is not an existing term in the domain of climatology
⇒ *Yes*, go to the next test
(en) ***student teacher*** 'teacher-in-training' – a student teacher is both a student and a teacher, so technically passes the test
(el) ***κόκκινο δάνειο*** kokino danio lit. red load 'non performing loan' - a red loan is a loan
(sv) ***blueberry*** Lit. blue|berry Trans. blueberry – a blueberry is a berry
(en) *piece of land* - the syntactic head is *piece* but the semantic head is *land* and it is a hypernym of *piece of land*

Test NMWE.14 - [LEX: Does a regular replacement of one of the lexicalized components by related words taken from a relatively large semantic class lead to ungrammaticality or to an unexpected change in meaning?]
⇒ *Yes*, annotate as a **NID -** (en) **hot dog** / # hot cat
⇒ *No*, go to the next test

Test NMWE.15 - [DEVERBAL: Does the candidate contain a deverbal noun and can it be rephrased (in the given context) using a verbal expression which passed any of the VMWE tests?]
⇒ *Yes*, annotate as a **VMWENom**
(en) *She is a quick* ***decision maker****. => She* ***makes decisions*** *quickly.* - ***makes decisions*** is an LVC
No, **exit (it is not a MWE).**