Annotation guidelines for multi-word expressions

Agnès Tutin, LIDILEM & LIG-GETALP

24 april 2015
version 2 - draft

This guide presents the annotation scheme for multiword expressions (hereafter MWEs) in the framework of the Franco-Brazilian project AIM-WEST. It extends a first experiment conducted by two interns in the LIDILEM and LIG-GETALP labs (may-july 2014) (Justine Reverdy and Manolo Iborra), that we wish to thank here.

The annotation scheme is based on several typologies of MWEs, mainly Heid (2008) and Tutin (2010). It is widely inspired by the Explanatory Combinatorial Lexicology model (Mel'čuk et al., 1995). This annotation scheme includes several kinds of MWEs, including **discontinuous expressions** and **overlapping expressions**. In order to annotate discontinuous MWEs and to deal with different kinds of tokenizers in NLP systems, each part of the MWE is annotated.

Several parameters are taken into account in this annotation scheme:

- The **type of the MWEs**, e.g. function word, collocation, proverb or "full phraseme", e.g. *in order to* is a function word.
- The **grammatical class of the MWEs**., e.g. adverb, noun, adjective …
- The **POS of the components of the MWEs**, e.g. *in order to* is decomposed as Prep + Noun + Prep

1) Delimitation of MWEs

We include in the class of MWEs multiword elements including several graphical units, separated by blanks or hyphens, or separated by several other words not included within the MWE: i.e. *pomme de terre, couch potatoes, rendez-vous, to take this problem into account* …

The first step in the annotation process is the delimitation of the elements of MWEs and this is not a trivial process. Each lexical unit (and hyphen) of the MWE is annotated with an element <mwe> and the same numerical identifier (num="en"). Only core lexical units within the MWE are annotated. Pure grammatical words which are not core elements of the MWE are excluded from the annotation:

Auxiliaries are not included in MWEs. For example, in the following sentence, only the elements *pris, en,* and *charge* are annotated. The auxiliaries *ont* and *été* are not taken into account.

1. Ainsi, la mécanisation et l'automatisation des procédés de travail dans l'industrie manufacturière ont été <mwe num="e1">prises</epl> <mwe num="e1">en</epl> <mwe num="e1">charge</epl> par la production à la chaîne, le taylorisme et le fordisme.

Similarly, grammatical words depending on nominal, adjectival or verbal MWEs are not annotated. In (1) for example, the preposition *par*, which depends on the verb, is not taken into account. In (2), the preposition *d'* wich depends on the collocation *tirer profit* is not included as a <mwe> element :

2. Par conséquent, on augmente considérablement les chances de <mwe num="e1">tirer</epl> <mwe num="e1">profit</epl> d'un tel changement si on l'ajuste avec l'ensemble de ces éléments avec lesquels il interagit.

Conversely, prepositions are included, as *en* in the MWE *prendre en charge* because when they are mandatory parts of the MWE. As regards determiners, they are omitted when any of them can occur within the MWE (for example, in the collocation *soulever* DET *question*).

3. L'utilisation des technologies <mwe num="e1">soulève</epl> des <mwe num="e1">questions</epl> comme la qualité …

Determiners are annotated when they are frozen and specific to the MWE as in the following example where *la* cannot be replaced by any other determiner.

4. Parallèlement à cet "éclatement" des organisations, on assiste à ce qui apparaît comme une certaine dissolution des structures de ces dernières, qui <mwe num="e1">cède</epl> <mwe num="e1">la</epl> <mwe num="e1">place</epl> aux processus (dynamiques) comme mécanismes organisationnels de cohésion.

Apostrophes are included in the annotation of the MWE. Contracted determiners (e.g. *au, du*) are not decomposed.

2) **Typology of MWEs**

Our annotation scheme does not only aim at helping evaluation of NLP processes, such as translation and recognition of MWEs, but also aims at providing a better understanding of MWEs in linguistics. This is why we propose a varied typology of MWEs, mainly based on Heid (2007), Tutin (2010) and widely inspired by Explanatory Combinatorial Lexicology model (Mel'čuk et al., 1995).

Different kinds of MWEs are distinguished[1]:

[1] For reasons of readability, we only include in the examples the attributes under discussion.

- **Function words (type="F")** include grammatical words such as conjunctions (*even if*), determiners (*a large bunch of*), prepositions (*in front of*), pronouns (*quelque chose*), and discursive, modal, intensive and negation adverbs (*on the one hand, to a large extent, not at all*). They are characterized by a vague - and mainly functional - meaning. These MWEs are very numerous in corpora.

5. Il y a moyen, croyons-nous, de faciliter le changement et d'améliorer <mwe type="F">non</epl> <mwe type="F"> seulement</epl> l'efficacité de l'utilisation des technologies, <mwe type="F">mais</epl> <mwe type="F"> aussi</epl> celle des processus de travail et la qualité de vie qui l'entoure.

- **Full phrasemes (type="PH")** include MWEs which **1) are not compositional**, i.e. the meaning of the MWE cannot be deducted from the meaning of the parts (i.e. the meaning of *couch potato* cannot be deducted from the meaning couch and potato) and/or **2) words, mainly nouns, which refer to a specific referent**. For example, the expression *death penalty* is quite easy to decode and partly compositional, but it refers to a specific fact and can thus be considered as a full phraseme.
  Metaphoric idioms such as *spill the beans* and *cross the line* can also be considered as full phrasemes. Nominal and adjectival compounds are also considered as full phrasemes. Example 6 provides an overview of different kinds of full phrasemes.

6. Afin de situer plus clairement notre propos, commençons par décrire à quoi pourrait ressembler la <mwe num="e1" type="PH">mise</epl> <mwe num="e1" type="PH">en</epl> <mwe num="e1" type="PH">oeuvre</epl> d'un changement technologique en <mwe num="e2" type="PH">faisant</epl> <mwe num="e2" type="PH"> appel</epl> à une organisation fictive.

Voir les cas comme *établissement thermal ….*

- **Collocations or semi-phrasemes (type="C")** include frequent compositional expressions, mainly binary expressions, where a word, the base, keeps its usual meaning, while the other word is more unpredictable (in translation, this is often not a literal translation). For example, for the expression *heavy smoker, smoker* is the base while *heavy* is the collocate. As a matter of fact, *heavy* is not literally translated intoits French equivalent *lourd* but into *gros*. Collocations can be analysed as predicate-argument constructions (Cf. Tutin 2008). Most collocations can be described with the help of the syntagmatic Lexical Function Collocations (lexical functions such as: intensifier, causative, light verb constructions, …)(Mel'čuk et al., 1995). Collocations can be difficult to distinguish from free combinations and the use of statistical measures can be useful to facilitate the annotation process. Collocations, and especially light verb constructions, are

very numerous in corpora. The following example provides an example of several kinds of collocations.

7. De plus, pour bien <mwe num="e1" type="C">faire</epl> leur <mwe num="e1" type="C">travail</epl>, les pigistes <mwe num="e2" type="C">ont</epl> <mwe num="e2" type="C">besoin</epl> d' <mwe num="e3" type="C">avoir</epl> rapidement <mwe num="e3" type="C">accès</epl> à une information très diversifiée et ce, pour deux <mwe num="e4" type="C">raisons</epl> <mwe num="e4" type="C">principales</epl>.

- "Strong collocations" : à mi-chemin entre collocations et phrasème. mouche bleue, médecin militaire, patineur artistique … EPL : hyponyme du nom. Ex : un N ADJ est un N. Les N Adj font partie de la classe des N.

  Type=**"SC"**

- Named entities of countries and cities include determiners if they are fixed determiners (e.g. *La France*, *Le Sappey*), that is, no other determiner can be used (except in the case of antonomasia).

- **Named entities (type="NE")** include MWE proper names, dates, events, places, organizations.

8. L'auteur tient à remercier ses collègues du programme Travail et technologies, <mwe num="e1" type="NE">David</epl> <mwe num="e1" type="NE">Tippin</epl>, <mwe num="e2" type="NE"> Richard</epl> <mwe num="e2" type="NE">Lavoie</epl> …

Named entities of countries and cities include determiners if they are fixed determiners (e.g. *La France*, *Le Sappey*), that is, no other determiner can be used (except in the case of antonomasia).

- **Phrasal verbs (type="PV")** are typical of Germanic languages and are absent from romance languages such as French and Portuguese. Some particles are separable, some are inseparable.

9. It could be helpful to <mwe type="PV">set</epl> <mwe type="PV">up</epl> **a se**cure framework for the mobility of minors.

- **Proverbs (type="PROV")**

10. "Nobody should refuse a taxi, if you have a queue system, it's <mwe type="PROV">first</epl> <mwe type="PROV">come</epl>, <mwe type="PROV">first</epl> <mwe type="PROV">served</epl>. "The points system is about refusing fares and a queue ...

- **Routine formulae (type="R")** are typical routines in sublanguages. They often are compositional but are not predictable.

11. Il y a moyen, <mwe type="R">croyons</epl> <mwe type="R">-</epl> <mwe type="R">nous</epl>, de faciliter le changement et d'améliorer non seulement l'efficacité de l'utilisation des technologies,

- **Complex terms (type="T")** can be considered as a subtype of full phrasemes. They are mainly nominal full phrasemes typical of specialized corpora (scientific or professional corpora).

12. Ainsi, la mécanisation et l'automatisation des <mwe num="e1" type="T">procédés</epl> <mwe num="e1" type="T">de</epl> <mwe type="R">travail</epl> dans l' <mwe num="e2" type="T">industrie</epl> <mwe num="e2" type="T">manufacturière</epl> ont été prises en charge …

- **Pragmatemes (type="PRAG")** (see Mel'čuk et al., 1995) are MWEs associated with a specific pragmatic function (generally in dialogues), specific to spoken language (to reply to thanking for example : *you're welcome* (Eng.) *il n' y a pas de quoi. De rien. Je vous en prie* (Fr.).

13. - Merci pour vos renseignements !
- <mwe num="e1" type="PRAG">De</epl> <mwe num="e1" type="PRAG">rien</epl>, <mwe num="e2" type="PRAG">au</epl> <mwe num="e2" type="PRAG">revoir</epl>!

3) **Grammatical class of the MWE**

The annotation indicates the grammatical class of the MWE as a whole **(mcat="A|N|…")**. Distributional properties must be taken into account to choose the appropriate grammatical class and not only morphological properties.

Abbreviation	Grammatical class	Example
A	Adjective	*Un livre bon marché* *Whisky on the rocks*
ADV	Adverb	*On the one hand, we think* *We wish to debate that, from time to time*
CC	Coordinating conjunction	*Mais aussi …*
DET	Determiner	*A good deal of work has been done* *Beaucoup de chercheurs …*
N	Common Noun	*A couch potato*

O	Other	
P	Preposition	*Contrary to us, he …* *In order to come, I …*
PN	Proper Noun	*L'auteur remercie David Tippin,*
PRO	Pronoun	*Lui-même, tout le monde …*
S	Sentence	*Il y a moyen, croyons-nous, de faciliter*
SC	Subordinating Conjunction	*Insofar as …* *Même si nous pensons ….*

Here is an example of the annotation of the grammatical class of MWEs.

```
14. <mwe mcat="P">Parallèlement</epl> <mwe mcat="P">à</epl> cet
    "éclatement" des organisations, on assiste à ce qui apparaît
    comme une certaine dissolution des structures de ces dernières,
    qui <mwe mcat="V">cèdent</epl> <mwe mcat="V">la</epl> <mwe
    mcat="V">place</epl> aux processus (dynamiques) comme mécanismes
    …
```

4) **Parts of speech of the elements of the MWEs**

Within the MWEs, the POS of the elements are annotated (pos="A|ADV|N|P…". Sometimes, the elements do not exist as single autonomous words (e.g. *fur* in *au fur et à mesure*). In this case, they are annotated with a specific label (X).

Grammatical categories for parts of speech

Abbreviation	**Part of Speech**	**Example**
A	Adjective	*bon* in *bon marché*
ADV	Adverb	*widely* in *widely open*
CC	Coordinating conjunction	*and* in *and so on*
DET	Determiner	*la* in *prendre la mouche*
MORPH	Morphological element	*entretien semi-dirigé*
N	Noun	*marché* in *bon marché*
NUM	Numeral	*14 juillet*
P	Preposition	*in* in *in order to*
PDET	Contraction of preposition + determiner	*au* in *au-dessusde …*
PRO	Pronoun	*lui* in *lui-même*
PUNCT	Punctuation mark	- in *rendez-vous*
SC	Subordinating conjunction	*if* in *even if*
V	Verb	*cross* in *cross the line*
X	No independent category (foreign words or archaic	*priori* in *a priori* *fur* in *au fur et à mesure*

	expressions)	

5) **Overlapping MWEs**

Lexical units can be involved in several MWEs. There are mainly two cases:

- **There is a partial overlapping** between two MWEs, for example, there are two collocations with the same base, for example, in *pay close attention.* In this case, we have two collocations with the same base, *attention* : *pay* ***attention*** and *close* ***attention***. This case is very frequent.

  In this case, we will duplicate attributes (id2, type2, mcat2, id3 ..), as in the following example:

```
15. In education, we should <mwe num="e1" type="C" mcat="V" pos="V">pay</epl> <mwe num="e2" type="C" mcat="N" pos="A">close</epl> <mwe num="e1" id2="2" type="C" type2=="C" mcat="V" mcat2="N" pos="N">attention</epl> to such developments and trends.
```

- **There is an inclusion of one MWE into another MWE**. The same principle is adopted. For example, the collocation *réduire* ***au minimum*** includes a function word *au minimum*. It will be annotated as follows:

```
16. Afin de <mwe num="e1" type="C" mcat="V" pos="V">réduire</epl> <mwe num="e1" id2="2" type="C" type2="F" mcat="V" mcat2="V" pos="PDET">au</epl> <mwe num="e1" id2="1" type="C" type2="F" mcat="V" mcat2="V" pos="N">minimum</epl> cet effort d'ajustement, la direction a demandé au groupe de travail…
```

References

Mel'čuk, I. A., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. De Boeck Supérieur.

Heid, U. (2008). Computational phraseology. An overview. *Phraseology. An interdisciplinary perspective.* Amsterdam, Benjamins, 337-360.

Tutin, A. (2008) . For an extended definition of lexical collocations. *Proceedings of Euralex.* Barcelona, 15-19 juillet 2009. 1453-1460.

Tutin, A. (2010). Sens et combinatoire lexicale: de la langue au discours. *Unpublished Dossier en vue de l'habilitation à diriger des recherches). Grenoble: Université de Stendhal.*

Appendix: example of annotated text

<s num="es255"> <mwe type="F" num="e1" pos="P" mcat="P">À</epl> <mwe type="F" num="e1" pos="N" mcat="P">travers</epl> l'exemple d'INC Inc., on a pu constater que ce contexte particulier est <mwe type="F" num="e2" pos="PRO" mcat="PRO">lui</epl><mwe type="F" num="e2" pos="PUNCT" mcat="PRO">-</epl><mwe type="F" num="e2" pos="ADV" mcat="PRO">même</epl> un <mwe type="PH" num="e3" pos="N" mcat="N">milieu</epl> <mwe type="PF" num="e3" pos="A" mcat="N">social</epl>, résultat d'un cheminement historique, des cultures occupationnelles des acteurs qui y participent, <mwe type="F" num="e4" pos="ADV" mcat="SC">ainsi</epl> <mwe type="F" num="e4" pos="SC" mcat="SC">que</epl> d'un environnement social et économique. </s>

<s num="es256"> C'est <mwe type="F" num="e1" pos="P" mcat="ADV">en</epl> <mwe type="F" num="e1" pos="N" mcat="ADV">partie</epl> <mwe type="F" num="e2" pos="P" mcat="P">par</epl> <mwe type="F" num="e2" pos="N" mcat="P">rapport</epl> <mwe type="F" num="e2" pos="P" mcat="P">à</epl> un tel contexte que s'élaborent des attentes, des stratégies (explicites ou non) et des <mwe type="C" num="e3" pos="N" mcat="N">prises</epl> <mwe type="C" num="e3" pos="P" mcat="N">de</epl> <mwe type="C" num="e3" pos="N" mcat="N">position</epl> <mwe type="F" num="e4" pos="P" mcat="P">par</epl> <mwe type="F" num="e4" pos="N" mcat="P">rapport</epl> <mwe type="F" num="e4" pos="P" mcat="P">à</epl> des phénomènes comme le <mwe type="T" num="e5" pos="N" mcat="N">changement</epl> <mwe type="T" num="e5" pos="N" mcat="N">technologique</epl>. </s>

<s num="es257"> La <mwe type="PH" num="e1" pos="N" mcat="N">mise</epl> <mwe type="PH" num="e1" pos="P" mcat="N">en</epl> <mwe type="PH" num="e1" pos="N" mcat="N">oeuvre</epl> de changements sera souvent guidée par la volonté soit d'<mwe type="C" num="e2" pos="V" mcat="V">assurer</epl> <mwe type="C" num="e2" pos="D" mcat="V">la</epl> <mwe type="C" num="e2" pos="N" mcat="V">pérennité</epl> de ce contexte en l'adaptant à un environnement en mouvement, soit de rompre avec lui pour assurer le renouvellement de l'organisation (Bergquist, 1993). </s>

<s num="es258"> Si les organisations sont soumises aux contraintes du poids de leur histoire, de leur environnement social, économique et culturel, <mwe type="R" num="e1" pos="PRO" mcat="S">il</epl> <mwe type="R" num="e1" pos="PRO" mcat="S>en</epl> <mwe type="R" num="e1" pos="V mcat="S>va</epl> <mwe type="R" num="e1" pos="p" mcat="S>de</epl> <mwe type="R" num="e1" pos="ADV" mcat="S>même</epl> pour les technologies. </s>

<s num="es259"> Il faut réaliser que ces dernières sont conçues et appliquées en <mwe type="PH" num="e1" pos="V" mcat="V">faisant</epl> <mwe type="PH" num="e1" pos="N" mcat="V">référence</epl> à des cultures occupationnelles et sociales, donc selon les exigences de <mwe type="PH" num="e2" pos="N" mcat="N">secteurs</epl> <mwe type="PH" num="e2" pos="P" mcat="N">d'</epl><mwe type="PH" num="e2" pos="N" mcat="N">activité</epl> particuliers. </s>