Choosing features for classifying multiword expressions

Éric Laporte
Université Paris-Est, Laboratoire d'informatique Gaspard-Monge CNRS UMR 8049, F77454
Marne-la-Vallée, France

A major requirement for a classification of MWEs is that it be explicitly based on features of the MWEs. Designing a satisfactory classification involves choosing such features. Accordingly, the resulting classification may be more or less fruitful for computational use.

1. Equivalence between features

Distinct features[1] are equivalent if they are observed in the same lexical entries. For example, 'decomposable [verbal idioms] tend to be syntactically flexible, in a manner predicted by the nature of the semantic decomposition; non-decomposable [verbal idioms], on the other hand, tend not to be syntactically flexible' (Baldwin & Kim, 2010:277).[2] Such an equivalence invites you to use the features for classification: assigning an entry to a class implicitly specifies both features at the same time. Thus, Baldwin & Kim (2010:279, Fig. 12.1) subdivide verbal idioms into two subclasses: non-decomposable idioms 'with hard restrictions on word order and composition'; decomposable, syntactically flexible idioms.

However, the temptation should be resisted until a systematic investigation confirms the intuition of equivalence between the features. In the case of decomposability and syntactic flexibility, Baldwin & Kim (2010) do not make use of comprehensive lexical databases of MWEs. An examination of facts in the lists of French verbal idioms (available at http://infolingu.univ-mlv.fr/) by Gross (1982) readily reveals counter-examples. *Trouver chaussure à son pied* 'find the perfect match' seems decomposable as *find*(*x*, *partner*), but does not admit syntactic variations. Conversely, *mettre toutes les chances de son côté* 'not take any chances' is hardly semantically decomposable, but admits the passive form: *Au moins, toutes les chances sont mises de mon côté* 'At least, I am not taking any chances'. Freckleton's (1985) Lexicon-Grammar tables show similar facts in English.

When a classification assumes features are equivalent and uses them as a single criterion, it takes the risk of misclassifying the entries for which they diverge. This compromises computational usage: a classification should ensure the members of each class have the corresponding defining features. Thus, it is a good practice to specify each criterion precisely and specifying which entries have which features, like in Lexicon-Grammar tables.

2. Reproducibility of features

Among the features that discriminate MWEs, which should be used for classification, and therefore investigated in priority? Of course, linguistic intuition plays a prominent role in this selection, but reproducibility is relevant too. Reproducible observations are those inherently susceptible to high inter-judge agreement. Features are not equal in this respect.

[1] By 'feature' we mean a property which can differ between MWEs, and therefore be used to assign them to different classes. For example, the French verbal idiom *guérir le mal par le mal* 'fight fire with fire' admits a cleft construction, as in *C'est par le mal qu'on guérit le mal* 'It is with fire that you fight fire', whereas *rater un éléphant dans un couloir* 'be unable to hit the broad side of a barn' does not: **C'est dans un couloir que tu raterais un éléphant*.

[2] Baldwin & Kim's (2010:270) notion of 'decomposability' is about the existence of a mapping between parts of the literal and the non-literal meaning of the expression. They equate this notion to Nunberg *et al.*'s (1994:496) notion of being an 'idiomatically combining expression'. Nunberg *et al.* (1994, 496-497) contrast it with 'transparency', which is about speakers' ability to guess why an expression with some literal meaning is used to convey a given non-literal meaning.

For example, the existence of cleft constructions of a MWE is judged by checking the grammaticality of some sentences, which is relatively factual; semantic decomposability relies on pure semantic intuition.[3] This language-independent contrast between features is also observed in English and in Italian. Reproducibility of observation is classically improved by adjusting feature definition, and in particular by resorting to formal or syntactic criteria. Among semantic features, differential semantic evaluation is more reproducible than absolute semantic evaluation (Gross, 1975:391).

Features with high reproducibility of observation, such as cleft construction and passivization, are a good basis for classification.[4] Existing classifications and tables of idioms (e.g. Gross, 1982; Freckleton, 1985) produced by users of the Lexicon-Grammar method (Gross, 1994) prioritize two types of features particularly easy to observe: syntactic variations and selectional restrictions. Other interesting features could be encoded too, e.g. the possibility of anaphoric reference to a component of a verbal idiom.[5]

Reproducibly observable features are often useful for language processing when they determine the possibility of occurrence of actual forms, such as a cleft variant, or the anaphoric form above: this is essential to automatically recognising such MWEs.

3. Cost and benefit of intensive description of the lexicon

The examples above illustrate the benefits of intensive description of the lexicon. It is costly, this might explain why so many NLP researchers shun it. However, such work is not unfeasible: lexicon-grammars of MWEs with representation of individual features have been published as early as 1974 (Labelle) for support-verb constructions in French, and 1985 (Freckleton) for verbal idioms in English (both available at http://infolingu.univ-mlv.fr/). Some are used in parsing now (Constant *et al.*, 2013).

References

Baldwin, Timothy, Su Nam Kim. 2010. Multiword Expressions, in *Handbook of Natural Language Processing*, 267-292. Boca Raton, USA: CRC Press.

Constant, Matthieu, Anthony Sigogne, Joseph Le Roux. 2013. Combining Compound Recognition and PCFG-LA Parsing with Word Lattices and Conditional Random Fields. *ACM Transactions on Speech and Language Processing* 10 (3), 8.1-8.24.

Freckleton, Peter. 1985. Sentence idioms in English, *Working Papers in Linguistics*, 153-168 + appendix (196 p.). University of Melbourne.

Gross, Maurice. 1975. On the relations between syntax and semantic, in *Formal Semantics of Natural Languages*, 389-405. Cambridge University Press.

Gross, Maurice. 1982. Une classification des phrases "figées" du français. *Revue Québécoise de Linguistique* 11.2, 151-185. Montréal: UQAM.

Gross, Maurice. 1994. 2nd edition, 2005. The Lexicon-Grammar of a Language: Application to French, in *The Encyclopaedia of Language and Linguistics*, v. 4, 2195-2205. Pergamon.

Labelle, Jacques. 1974. *Étude de constructions avec opérateur* avoir *(nominalisations et extensions)*. Thèse de troisième cycle, LADL, Université Paris 7.

Nunberg, Geoffrey, Ivan A. Sag, Thomas Wasow. 1994. Idioms. *Language* 70, 491-538.

[3] For *se mettre le doigt dans l'œil* 'bark up the wrong tree', how else to arbitrate between speakers who associate *le doigt* with the semantics of 'opinion(*x*)' and *dans l'œil* with 'false', and those that do not?

[4] Baldwin & Kim (2010) claim 'the exact form of syntactic variation [of verbal idioms] is predicted by the nature of their semantic decomposability', but it would not be effectual to infer their syntactic variation from a description of their decomposability, because the former is more reproducibly observable than the latter.

[5] For example, applying the criterion of distributional frozenness, *citer un témoin* 'call as a witness' is idiomatic, because the verb *citer* has this meaning only with this noun. Still, anaphoric reference to the noun is possible: *La défense a cité un témoin. Il vient de s'exprimer* 'The defence called a witness. He has just testified'. This property can be observed in a reasonably reproducible way, due to the formal criterion involved in its definition. It may be correlated to semantic decomposability, but it is more definite.