Choosing features for classifying multiword expressions

Éric Laporte
Université Paris-Est, Laboratoire d’informatique Gaspard-Monge, CNRS UMR 8049, 77454 Marne-la-Vallée, France

Features
A feature is an observable property which can differ from a MWE to another, and therefore be used to assign them to different classes:

- guérir le mal par le mal “fight fire with fire”
  C’est par le mal qu’on guérit le mal “It is with fire that you fight fire”
- rater un éléphant dans un couloir “be unable to hit the broad side of a barn”
  *C’est dans un couloir que tu raterais un éléphant

Equivalent features
Two features are equivalent if they are observed in the same lexical entries. Example:

- semantic decomposability ⇔ syntactic flexibility ([1]:277)
  Intuition often overestimates the degree of correlation between features. Counterexamples from the lists of French verbal idioms (available at http://infolingu.univ-miv.fr/) by [5]:
  - rater un éléphant dans un couloir “be unable to hit the broad side of a barn”
  - decomposable: miss(x, easy-target)
  - no syntactic variations, not even omission of the prepositional complement
  - mettre toutes les chances de son côté “not take any chances”
  - hardly semantically decomposable
  - passive form: Au moins, toutes les chances ont été mises de mon côté “At least, I am not taking any chances”

Factors of reproducibility

- Grammaticality judgment vs. semantic intuition. Grammaticality can be observed in a more reproducible way: it is more factual and can be backed by corpus attestations in some cases. The reproducibility of observation of a feature is classically improved by adjusting the definition of the feature, and in particular by resorting to formal or syntactic criteria rather than semantic evaluation.
- Among semantic features, differential semantic evaluation is more reproducible than absolute semantic evaluation ([8]:391).

Existing classifications and tables of idioms (e.g. [5], [3]) produced by users of the Lexicon-Grammar method ([6]) prioritize two types of features particularly easy to observe:

- syntactic variations, such as omissions and passive;
- selectional restrictions on arguments.

Impact of reproducibility on scientficity
Low reproducibility casts a doubt on what exactly a feature is. Features with more reproducibility of observation are a better basis for classification with an ambition of stability and scientficity. When [1] proposes that ‘the exact form of syntactic variation [of verbal idioms] is predicted by the nature of their semantic decomposability’, this suggests the syntactic variation of verbal idioms would not be worth describing, since it could be deduced from a description of their semantic decomposability. Such a suggestion is questionable for two reasons:

- no improvements of the definition of semantic decomposability seem to be at hand: the description of semantic decomposability would give hazardous results because of low reproducibility, and inference on syntactic variation would consequently yield shaky results, while syntactic variation can be described directly through more reproducible processes;
- the alleged rules of prediction are unknown, and formalizing them would be a challenge that no one has taken up yet.

Cost and benefit of intensive description of the lexicon
Intensive description of the lexicon is costly, but it deepens knowledge of how correlated two features are, and how reproducibly they can be observed;

- it is crucial to selecting features for classification, and therefore to the quality of classification;
- it provides examples and counter-examples which are useful to test hypotheses and proposals;
- it is complementary to corpus annotation, which deepens awareness of context-related issues;
- it is not unfeasible: comprehensive repositories (lexicon-grammars) of MWEs with representation of individual features have been published as early as 1974 [7] for support-verb constructions in French, and 1985 [3] for verbal idioms in English (both available at http://infolingu.univ-miv.fr/). Some lexicon-grammars of MWEs are used in parsing now [2].

References