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-mlv.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"

Several features as a single criterion

When a classification surmises two features to be equivalent and uses them as a single criterion ([1]:279, Fig. 12.1), it takes the risk of misclassifying the entries for which they conflict. This compromises computational usage, since a major function of a classification is to ensure that the members of each class have the corresponding defining features.
As long as all properties are not securely established for all entries, it is a good practice to specify each criterion precisely. Such practice leads to individuating a number of features, and to specifying which lexical entries have which features, like in Lexicon-Grammar tables of idioms [3].

Selection of features

How to select the features to be used for classification, and therefore to be investigated in priority?

- linguistic intuition
- reproducibility, a technical criterion

Reproducibility of feature observation

Reproducible observations are those inherently susceptible to high inter-judge agreement during manual description of lexical entries. Features are inherently not equal in this respect.

- Existence of cleft constructions of a MWE is judged by checking the grammaticality of some sentences.
- Semantic decomposability in the sense of [8] and [1] relies on pure semantic intuition: in the case of the verbal idiom *se mettre le doigt dans l'œil* "bark up the wrong tree", how else to arbitrate between speakers for which *le doigt* stands for an element of meaning like `understanding(x)` and *dans l'œil* for false, and those that do not share these impressions?

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 ([4]: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 scientificity

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 scientificity. 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.

Impact of reproducibility on computational applications

Reproducibly observable features are often potentially useful in language-processing applications, especially when they determine the possibility of occurrence of actual forms, such as the cleft variants of an idiom. This is essential to automatically recognising such MWEs. In contrast, semantic decomposability is relevant to psycholinguistics, but less likely to be useful in computational applications.

A little-known feature

Possibility of anaphoric reference to a component of a MWE is an interesting feature for applications. The idioms above do not admit such reference:

> **J'ai mis toutes les chances de mon côté. Elles me donnent de l'espoir*
>
> *I did not take any chances. Those make me hope

Many technical expressions like *citer un témoin* "call as a witness" are different:

- applying the criterion of distributionally frozenness, they are idiomatic, because the verb *citer* has this meaning only with this noun;
- 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.

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 of 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-mlv.fr/`). Some lexicon-grammars of MWEs are used in parsing now [2].

References

[1] T. Baldwin & S.N. Kim, Multiword Expressions, in *Handbook of NLP*, 267-292, CRC Press, Boca Raton, USA, 2010.

[2] M. Constant, A. Sigogne, J. Le Roux, 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, 2013.

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

[4] M. Gross, On the relations between syntax and semantics, in E.L. Keenan (ed.), *Formal Semantics of Natural Languages*, 389-405, Cambridge University Press, 1975.

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

[6] M. Gross, The Lexicon-Grammar of a Language: Application to French, In R.E.Asher (ed.), *The Encyclopaedia of Language and Linguistics*, vol. 4, 2195-2205, Oxford/NewYork/Seoul/Tokyo: Pergamon, 1994. 2nd edition, 2005.

[7] J. Labelle, *Étude de constructions avec opérateur* avoir *(nominalisations et extensions)*, Thèse de troisième cycle, Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7, 1974.

[8] G. Nunberg, I.A. Sag, Th. Wasow, Idioms, *Language* 70, 491-538, 1994.

PARSEME meeting, 10-11 March 2014, Athens, Greece