Time flies like an arrow and fruit flies like a banana; parsing multiword constructions with DepVis

Seongmin Mun [1, 2, @] , Ilaine Wang [1, @] , Guillaume Desagulier [3, *, @] , Gyeongcheol Choi [2, @] , Kyungwon Lee [2, *, @]
1 : Modèles, Dynamiques, Corpus (MoDyCo) - *Website*
Université Paris Nanterre : UMR7114, Centre National de la Recherche Scientifique : UMR7114
Université Paris 10 Bâtiment A - Bureau 402 A 200, avenue de la République 92001 Nanterre Cedex - France
2 : Ajou University - *Website*
3 : Modèles, Dynamiques, Corpus (MoDyCo) - *Website*
CNRS : UMR7114, Université de Paris X - Nanterre
Université Paris 10 Bâtiment A - Bureau 402 A 200, avenue de la République 92001 Nanterre Cedex - France
* : Corresponding author

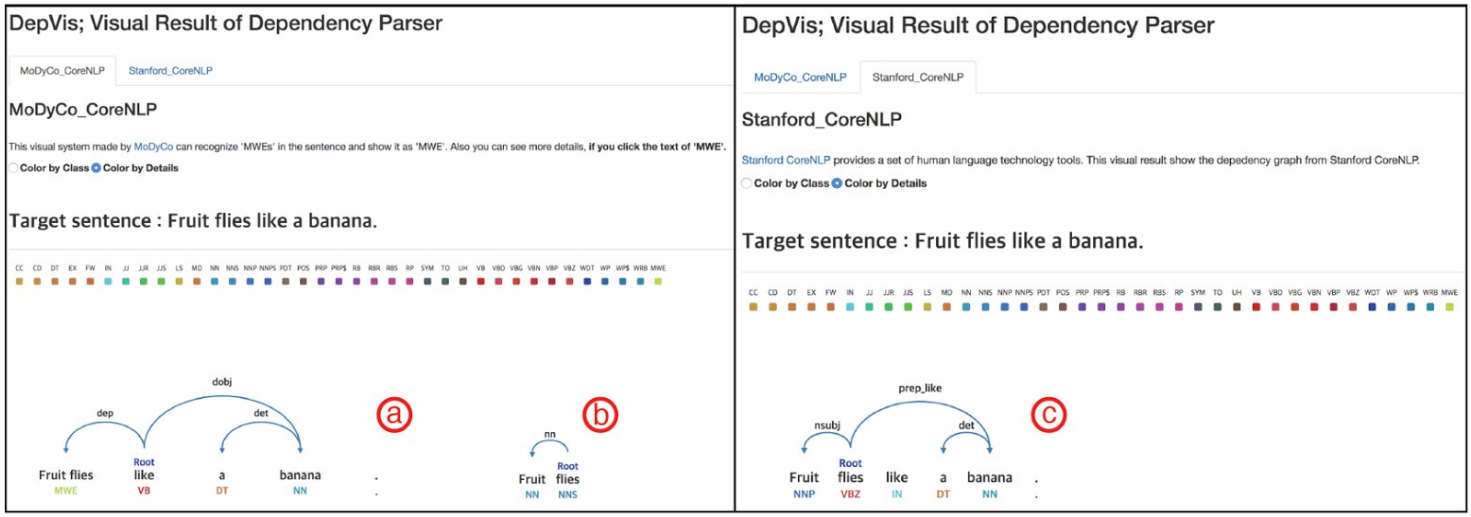

Figure 1. Interface of DepVis. The interface of the proposed visual system representing the ambiguous sentence 'Fruit flies like a banana.'

Abstract

Multiword expressions (MWEs) are strings of two or more lexemes that are idiosyncratic in some respect. Such complex strings are frequent. Sag et al. (2002) estimate that 41% of the entries in WordNet 1.7 are MWEs. MWEs assume a wide range of forms such as institutionalized phrases and clichés (*love conquers all*), idioms (*kick the bucket*), fixed phrases (*by and large*), compound nouns (*frequent-flyer program*), verb-particle constructions (*eat/look/write up*), light verbs (*have a drink/*an eat*), named entities (*Paris*), lexical collocations (*telephone box/booth/*cabin*), etc.

The grammatical status of MWEs has been an issue at least since the "rules vs. the lexicon" debate (Langacker 1987; Pinker 1999; Pinker and Prince 1988; Rumelhart and McClelland 1986). Because rules capture all the regularities in language, MWEs does have no place in the grammar proper because they are lexical. Because the lexicon consists of words or morphemes, it should not include MWEs because they are phrasal. Jackendoff (1997, chapter 7) advocates the inclusion of "phrasal lexical items" in the lexicon. An alternative, although related, solution inspired by construction grammar approaches delegates MWEs to a "constructicon" (Goldberg 2006, p. 64). In this paper, we treat MWEs as multiword constructions (MWCs).

The interpretation of MWCs poses a major challenge for NLP techniques due to their heterogeneous nature. We address two challenges: the automatic detection of MWCs from large corpora and the automatic resolution of ambiguities.

With respect to the first challenge, we present a parsing algorithm that combines *n*-gram processing and dependency analysis based on dictionaries. MWC candidates are extracted using one of the two methods and then compared to dictionary entries. If a MWC candidate matches at least one entry, the algorithm treats it as meaningful and stores it in the inventory of verified MWCs.

With respect to the second challenge, one common issue is the case where a MWC is ambiguous in a sentence (1).
(1) **Fruit flies like a banana.**

State-of-the-art dependency parsers such as Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/) or Universal Dependencies (http://universaldependencies.org/) fail to recognize that *fruit flies* is a compound NP and treat *flies* like a verb (Figure 1 c & d).

To fix this kind of problem, we built 'DepVis', a visual system that displays and compares the results from both the Stanford CoreNLP parser and our algorithm. With 'DepVis', users can visualize not only MWCs (Figure 1 (a)) but also their internal dependencies (Figure 1 (b)).

With the help of experiments and case studies on ambiguous sentences, we verify the effectiveness and usability of 'DepVis'. Results show that our parsing algorithm recognize MWCs quickly and accurately, including in ambiguous sentences. This is because it captures problematic expressions, compares them to the repository of verified MWCs, and outputs a correct representation. We believe our algorithm is a significant contribution to the understanding of the constructicon in construction grammar approaches to language.

Keywords

Multiword expressions, Natural Language Processing, Syntax, Semantics, Parsing, Dependency, Visualization, Web Application, Construction network

References

Goldberg, Adele E. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford & New York: Oxford University Press.

Jackendoff, Ray (1997). The Architecture of the Language Faculty. Cambridge, Mass. ; London: MIT Press.

Langacker, Ronald W. (1987). Foundations of Cognitive Grammar. Vol. 1. Stanford: Stanford University Press.
Pinker, Steven (1999). Words and Rules: The Ingredients of Language. New York: Basic Books.

Pinker, Steven and Alan Prince (1988). "On language and connectionism: Analysis of a parallel distributed processing model of language acquisition." In: Cognition 28.1-2, pp. 73–193.

Rumelhart, D. E. and J. L. McClelland (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2." In: ed. by David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group. Cambridge, MA, USA: MIT Press. Chap. On Learning the Past Tenses of English Verbs, pp. 216–271. isbn: 0-262-13218-4. url: http://dl.acm.org/ citation.cfm?id=21935.42475.

Sag, Ivan A et al. (2002). "Multiword expressions: A pain in the neck for NLP." In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 1–15.