Time flies like an arrow and fruit flies like a banana Parsing multiword constructions with *DepVis*

Seongmin Mun [1][2] Ilaine Wang [2] Guillaume Desagulier [3]
Gyeongcheol Choi [1] Kyungwon Lee [1]

[1] Ajou University, Suwon, South Korea

[2] MoDyCo (UMR 7114), CNRS, Paris Nanterre

[3] MoDyCo (UMR 7114), Paris 8, CNRS, Paris Nanterre & Institut Universitaire de France

ICCG 10
Université Sorbonne Nouvelle – Paris 3
July 17th, 2018

outline

1. Introduction
2. review
3. methods
4. results
5. discussion
6. conclusion

multiword expressions (MWEs)

minimal working definition

- a string of 2+ lexemes
- idiomatic in some respect

MWEs are frequent

reference	share of MWEs	corpus
Sag et al. (2002)	41%	WordNet 1.7
Graça Krieger and Finatto (2004)	70%	specialized corpus
Ramisch (2009)	50%-80%	scientific biomedical abstracts
Ramisch et al. (2013)	51.4% (nouns) 25.5% (verbs)	English WordNet

multiword expressions (MWEs)

A vast inventory Sag et al. (2002)'s pain-in-the-neck typology

institutionalized phrases and clichés

(1) love conquers all

idioms

(2) sweep under the rug

fixed phrases

(3) by and large

compounds

(4) frequent-flyer program

verb-particle constructions

(5) eat/look/write up

light verbs

(6) a. have a drink/$^{?}$an eat
b. make/*do a mistake

named entities

(7) Oakland A's, Oakland, the A's

lexical collocations

(8) a. telephone box/booth/*cabin
b. emotional baggage/*luggage

etc.

multiword expressions (MWEs)

NLP vs linguistics

As a term, 'MWE' has a strong NLP flavor since Sag et al. (2002) and under the influence of the the Stanford MWE project (`http://mwe.stanford.edu/`)

- NLP → capture, recognition, and comprehension
  - task-oriented (computer-mediated lexicography, OCR, morphological and syntactic tagging, information retrieval, computer-mediated translation, L2 teaching, word-sense disambiguation, etc.)
- linguistics → acquisition and usage
  - linguistic competence & performance

Looks like we don't have the same priorities

MWEs in linguistics

corpus linguistics

working definitions for MWEs

Firth (1957)

- you shall know a word by the company it keeps
- mutual expectations

collocations & phraseological units

(Sinclair 1991)

- the open-choice principle
- the idiom principle

MWEs in linguistics

Meaning-Text Theory (Mel'čuk)

3 kinds of entries in the Explanatory Combinatorial Dictionary (Mel'čuk and Polguere 1987; Mel'čuk and Polguère 1995)

- full phrasemes ($\approx$ non-compositional collocations)
  e.g. long time no see
- semi-phrasemes (partially compositional collocations)
  e.g. `Magn(rain) = {heavy, torrential}`
- quasi-phrasemes (semantically-constrained, partially-compositional collocations)
  e.g. *bus stop* is the place where the bus stops and passengers get in, not the moment when it stops (at a traffic light)

MWEs in linguistics

generative linguistics & CxG

- 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 should have no place in the grammar proper because they are lexical. Because the lexicon consists of words or morphemes, it does not include MWEs because they are phrasal*

- Fillmore et al. (1988, p. 504)

  > *the descriptive linguist needs to append to this maximally general machinery [. . .] knowledge that will account for speakers' ability to construct and understand phrases and expressions in their language which are not covered by the grammar, the lexicon and the principles of compositional semantics, as these are familiarly conceived. Such a list of exceptional phenomena contains things which are larger than words, which are like words in that they have to be learned separately as individual whole facts about pieces of the language, [. . .]*

MWEs in linguistics

generative linguistics & CxG

3 classes (Fillmore et al. 1988)

- unfamiliar pieces unfamiliarly combined
  e.g. the X-er, the Y-er
- familiar pieces unfamiliarly combined
  e.g. all of a sudden
- familiar pieces familiarly combined
  e.g. once every blue moon

MWEs in linguistics

generative linguistics & CxG

- "phrasal lexical items" (i.e. "lexical items larger than X^0") should be part of the lexicon (Jackendoff 1997, chapter 7)
- MWEs are part of the 'constructicon' (Goldberg 2006, p. 64)

MWEs in linguistics

language acquisition

computational simulations of acquisition models

- Joyce and Srdanović (2008)
- Rapp (2008)

studies on specific MWEs

- verb-particle constructions (Villavicencio et al. 2012)
- nominal compounds (Devereux and Costello 2012)
- light-verb constructions (Nematzadeh et al. 2013)
- multiword terms (Lavagnino and Park 2010)

MWEs in NLP

some landmark studies

- Choueka (1988): collocation extraction based on *n*-gram statistics
- Smadja (1993): `Xtract`, a tool for collocation extraction based on simple POS filters + μ and σ of word distance
- Dagan and Ken Church (1994): `Termight`, a semi-automatic tool to help translators and terminologists identify technical terms and their translations (candidate terms selected with a variant of PMI, as found in Kenneth Church and Hanks 1990)
- Lin (1998a,b) use of syntactic dependencies to extract candidate collocations restricted to a specific category
- the Stanford MWE project (early 2000s), (see Sag et al. 2002)
- Ramisch (2014)

MWEs in NLP

scientific events

MWEs are featured at major conferences

- COLING
- ACL (SIGLEX, EACL, NAACL)
- LREC

and courses/tutorials

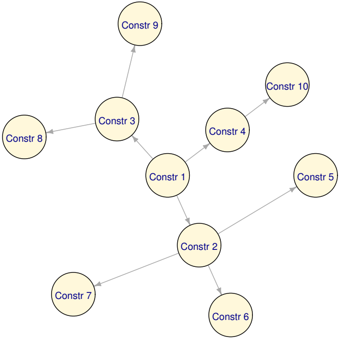
- ACL
- SemEval
- PARSEME

but again, task-oriented

MWEs in GxC

the constructicon

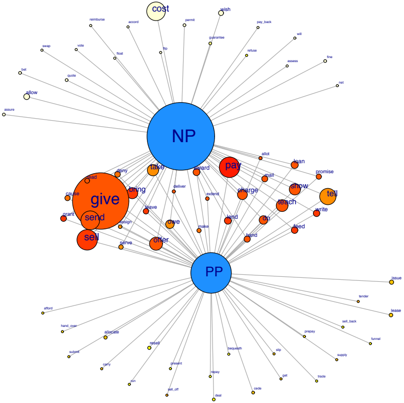
Tenet 7. The totality of our knowledge of language is captured by a network of constructions a 'construct-i-con'. (Goldberg 2003)



MWEs in GxC

the constructicon

An undirected graph based on corpus data (after Bresnan et al. 2007) and the `languageR` dataset

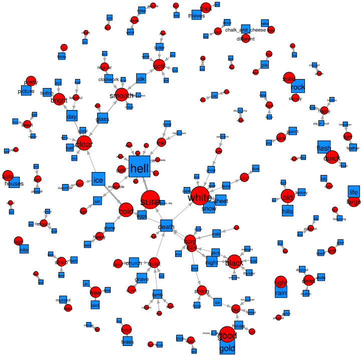


the dative alternation

MWEs in GxC

the constructicon

A directed graph based on corpus data and association measures (Desagulier 2015)



A as NP

MWEs in GxC

the constructicon

our long-term objective

- simulate the construction network from a very large corpus

how

- collect a large database of existing MW-constructions
- train an algorithm so that it can detect them
- vectorize the constructions by means of artificial neural networks
- plot the results by means of a graph
- further train the algorithm so that it can detect new constructions

MWEs in GxC

the constructicon

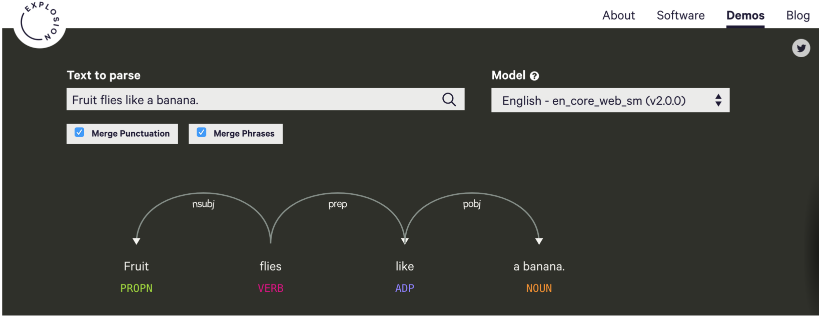
problems

- ambiguity
- polysemy
- homonymy
- long-distance dependencies
- etc.

Most, if not all the issues listed in Sag et al. (2002) are still unresolved today

When things go surprisingly wrong

AI



Text to parse

Fruit flies like a banana.

☑ Merge Punctuation ☑ Merge Phrases

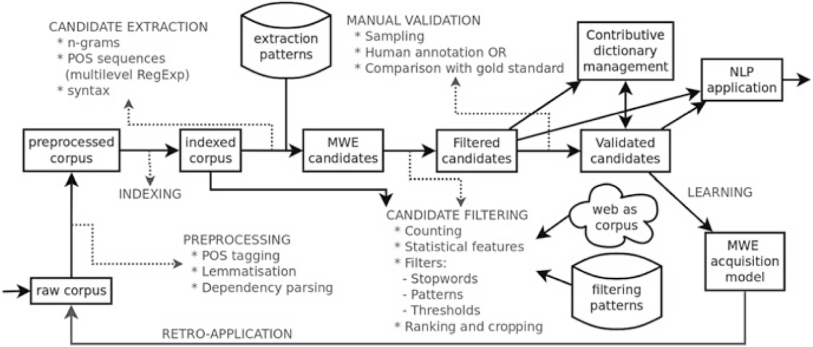
Model

English - en_core_web_sm (v2.0.0)

displaCy (`https://demos.explosion.ai/displacy/`)

previous work

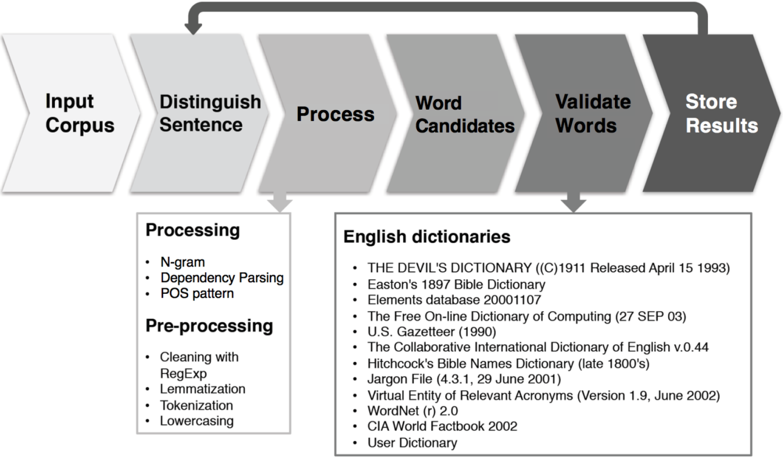
mwetoolkit (Ramisch 2014)



Framework for MWE extraction with *mwetoolkit*

data processing

overview



Input Corpus

Distinguish Sentence

Process

Word Candidates

Validate Words

Store Results

Processing

- N-gram
- Dependency Parsing
- POS pattern

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

English dictionaries

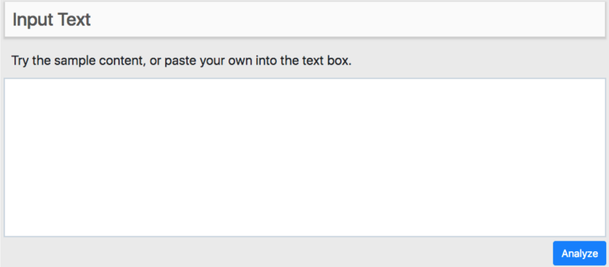
- THE DEVIL'S DICTIONARY ((C)1911 Released April 15 1993)
- Easton's 1897 Bible Dictionary
- Elements database 20001107
- The Free On-line Dictionary of Computing (27 SEP 03)
- U.S. Gazetteer (1990)
- The Collaborative International Dictionary of English v.0.44
- Hitchcock's Bible Names Dictionary (late 1800's)
- Jargon File (4.3.1, 29 June 2001)
- Virtual Entity of Relevant Acronyms (Version 1.9, June 2002)
- WordNet (r) 2.0
- CIA World Factbook 2002
- User Dictionary

data processing

step 1

Input Corpus

✓ **Interface of Input text**

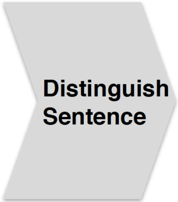


Input Text

Try the sample content, or paste your own into the text box.

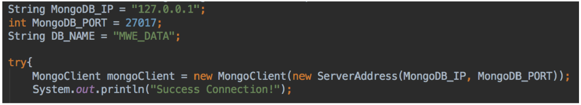
data processing

step 2



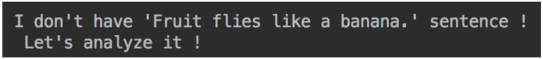
Distinguish Sentence

✓ **MongoDB & JAVA**



```
String MongoDB_IP = "127.0.0.1";
int MongoDB_PORT = 27017;
String DB_NAME = "MWE_DATA";

try{
    MongoClient mongoClient = new MongoClient(new ServerAddress(MongoDB_IP, MongoDB_PORT));
    System.out.println("Success Connection!");
```

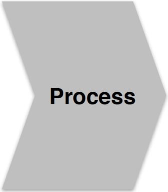
✓ **Out Put**



```
I don't have 'Fruit flies like a banana.' sentence !
 Let's analyze it !
```

data processing

step 3



Process

✓ **N-gram**

N-gram method is a contiguous sequence of ***N*** items from a given sequence of text.

✓ **Dependency Parsing**

Dependency parser can provide a simple description of the grammatical relationships in a sentence.

✓ **POS pattern**

The POS pattern is a Boolean value that indicates whether the expressions used in the sentence has the same part of speech pattern as the canonical form.

data processing

step 4

Words Candidates

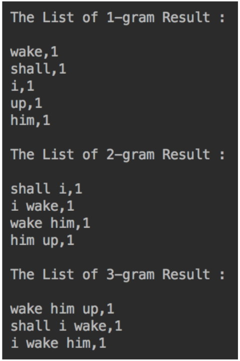
✓ **N-gram**

“Shall I wake him up?”

Unigram : Shall, I, wake, him, up.

Bigram : Shall I, I wake, wake him, him up.

Trigram : Shall I wake, I wake him, wake him up.



```
The List of 1-gram Result :

wake,1
shall,1
i,1
up,1
him,1

The List of 2-gram Result :

shall i,1
i wake,1
wake him,1
him up,1

The List of 3-gram Result :

wake him up,1
shall i wake,1
i wake him,1
```

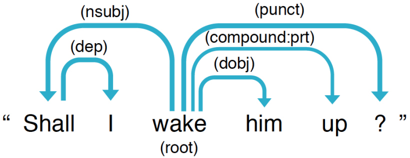
data processing

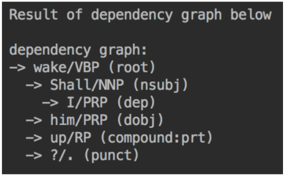
step 4

Words Candidates

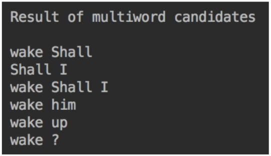
✓ **Dependency parser**

“Shall I wake him up?”





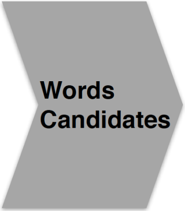
```
Result of dependency graph below

dependency graph:
-> wake/VBP (root)
  -> Shall/NNP (nsubj)
    -> I/PRP (dep)
  -> him/PRP (dobj)
  -> up/RP (compound:prt)
  -> ?/. (punct)
```



```
Result of multiword candidates

wake Shall
Shall I
wake Shall I
wake him
wake up
wake ?
```

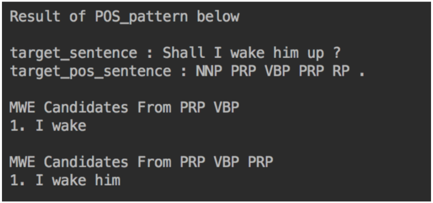
data processing

step 4


Words Candidates

✓ **POS(Part Of Speech)**

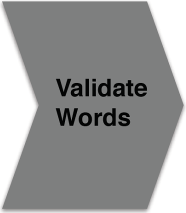
“	Shall	I	wake	him	up	?	”
	(verb)	(pron)	(verb)	(pron)	(part)	(punc)	



```
Result of POS_pattern below

target_sentence : Shall I wake him up ?
target_pos_sentence : NNP PRP VBP PRP RP .

MWE Candidates From PRP VBP
1. I wake

MWE Candidates From PRP VBP PRP
1. I wake him
```

data processing

step 5



Validate Words

✓ **English Dictionaries**

English dictionaries

- THE DEVIL'S DICTIONARY ((C)1911 Released April 15 1993)
- Easton's 1897 Bible Dictionary
- Elements database 20001107
- The Free On-line Dictionary of Computing (27 SEP 03)
- U.S. Gazetteer (1990)
- The Collaborative International Dictionary of English v.0.44
- Hitchcock's Bible Names Dictionary (late 1800's)
- Jargon File (4.3.1, 29 June 2001)
- Virtual Entity of Relevant Acronyms (Version 1.9, June 2002)
- WordNet (r) 2.0
- CIA World Factbook 2002
- User Dictionary

API : http://services.aonaware.com/DictService/

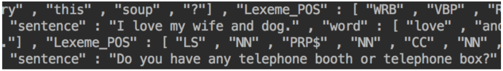
data processing

step 6

Store Results

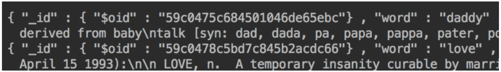
✓ **Data Base : MongoDB & JAVA**

✓ **Sentence Collection**



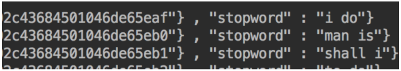
```
ry" , "this" , "soup" , "?"] , "Lexeme_POS" : [ "WRB" , "VBP" , "R
 "sentence" : "I love my wife and dog." , "word" : [ "love" , "and
."] , "Lexeme_POS" : [ "LS" , "NN" , "PRP$" , "NN" , "CC" , "NN" ,
 "sentence" : "Do you have any telephone booth or telephone box?"
```

✓ **Dictionary Collection**



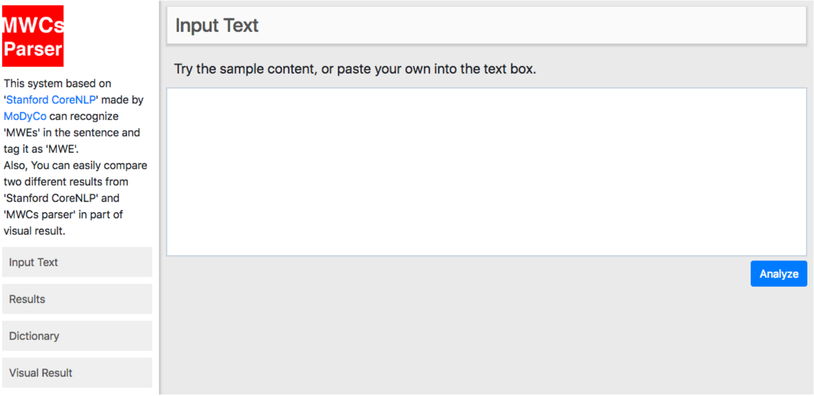
```
{ "_id" : { "$oid" : "59c0475c684501046de65ebc"} , "word" : "daddy"
  derived from baby\ntalk [syn: dad, dada, pa, papa, pappa, pater, po
{ "_id" : { "$oid" : "59c0478c5bd7c845b2acdc66"} , "word" : "love" ,
  April 15 1993):\n\n LOVE, n.  A temporary insanity curable by marri
```

✓ **Stopwords Collection**



```
2c43684501046de65eaf"} , "stopword" : "i do"}
2c43684501046de65eb0"} , "stopword" : "man is"}
2c43684501046de65eb1"} , "stopword" : "shall i"}
2c43684501046de65eb2"} , "stopword" : "to do"}
```

MWCs parser



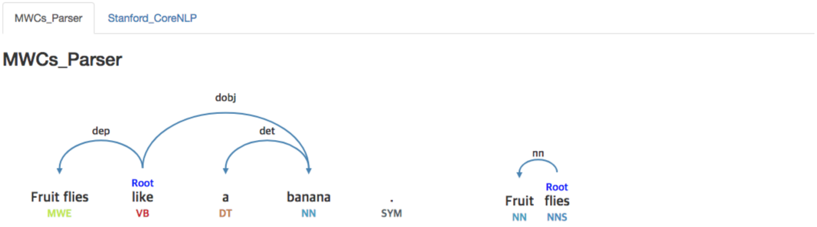
This system based on 'Stanford CoreNLP' made by MoDyCo can recognize 'MWEs' in the sentence and tag it as 'MWE'.
Also, You can easily compare two different results from 'Stanford CoreNLP' and 'MWCs parser' in part of visual result.

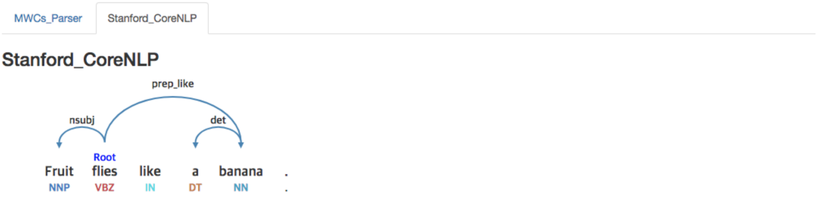
Input Text

Try the sample content, or paste your own into the text box.

Video link (`https://www.youtube.com/embed/meTWC5Nk9F4`)

an ambiguous sentence





DepVis link (`http://stat34.github.io/DepVis/`)

conclusion

- results show that our parsing algorithm recognize MWCs accurately, including in ambiguous sentences.
- storing more sentences will improve the speed of the algorithm.
- storing more MWEs will allow the algorithm to recognize more MWEs.