Fruit flies like a banana : Parsing multiword constructions with *DepVis*

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

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

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

[3] Ajou University, Suwon, South Korea

Hankuk University of Foreign Studies
December 14th, 2018

modyco Modèles, Dynamiques, Corpus
UMR 7114
Sciences du langage.

institut
universitaire
de France

Université
Paris Nanterre

outline

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

multiword expressions (MWEs)

- Words in a text corpus include features and information
- Words can be broadly divided into two categories

multiword expressions (MWEs)

- "With profound gratitude and great humility, I accept your nomination for the presidency of the United States."(Barack Obama's presidential speeches)

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.

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 linguistics

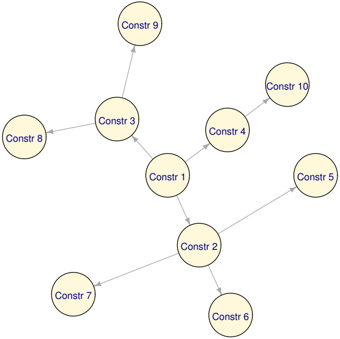
generative linguistics & CxG

- “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 GxC

the constructicon

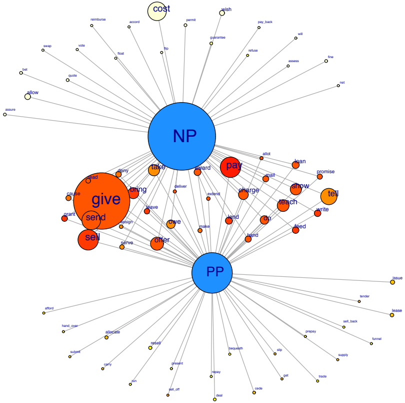
Tenet 7. The totality of our knowledge of language is captured by a network of constructions a ‘construct-i-con’. (Goldberg 2003)



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

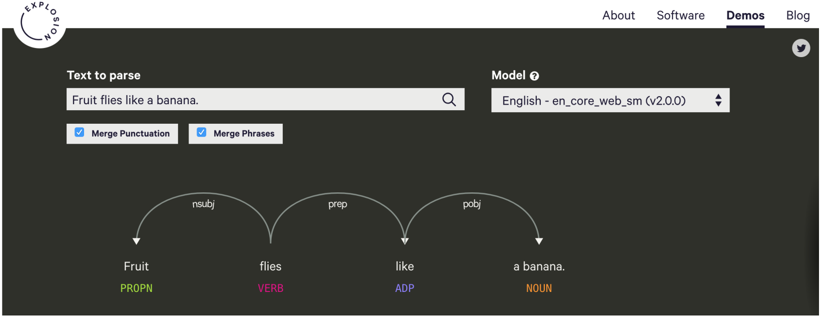
problems

- 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.

Model

English - en_core_web_sm (v2.0.0)

☑ Merge Punctuation

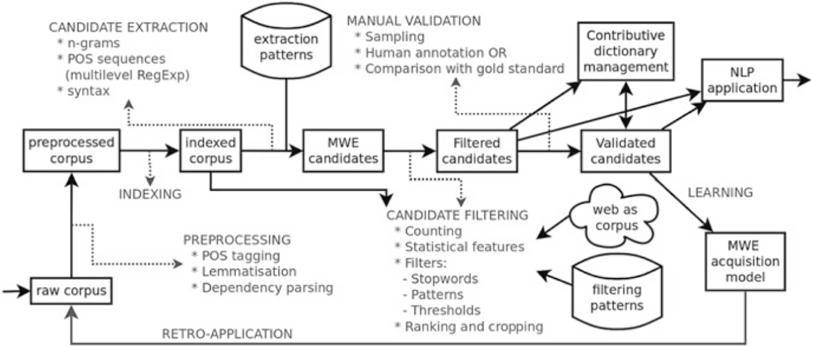
☑ Merge Phrases

Fruit	flies	like	a banana.
PROPN	VERB	ADP	NOUN

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

previous work

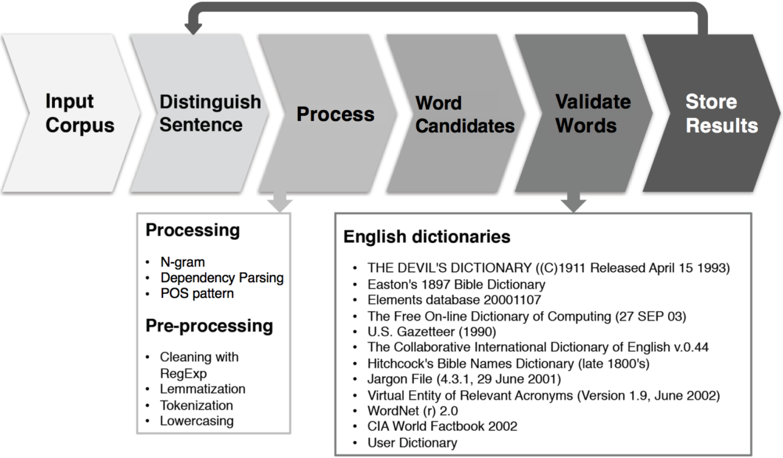
mwetoolkit (Ramisch 2014)



Framework for MWE extraction with *mwetoolkit*

data processing

overview



Processing

- N-gram
- Dependency Parsing
- POS pattern

Pre-processing

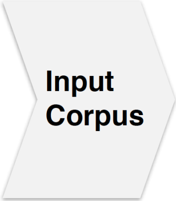
- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

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

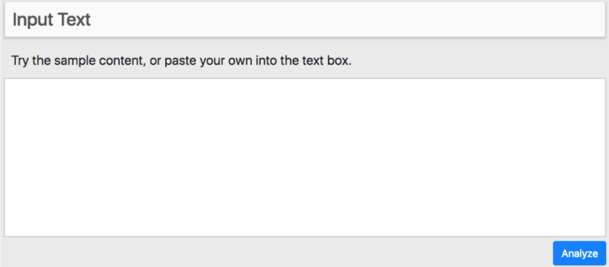
data processing

step 1



Input Corpus

✓ **Interface of Input text**



Input Text

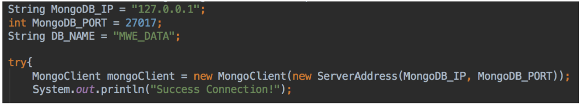
Try the sample content, or paste your own into the text box.

data processing

step 2

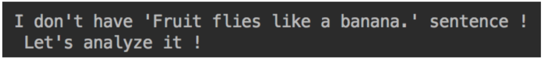
Distinguish Sentence

✓ **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!");
```

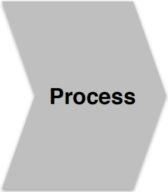
✓ **Out Put**



```
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

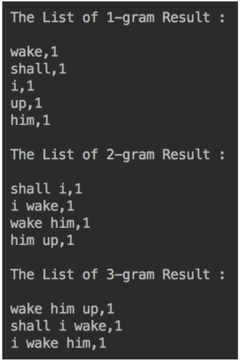
✓ **N-gram**

“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
```

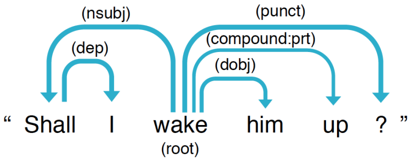
data processing

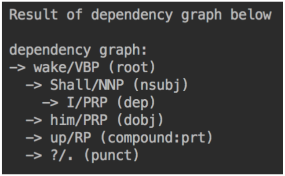
step 4

Words Candidates

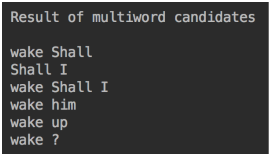
✓ **Dependency parser**

“Shall I wake him up?”





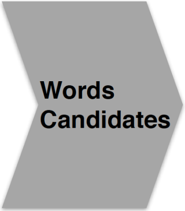
```
Result of dependency graph below

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 ?
```

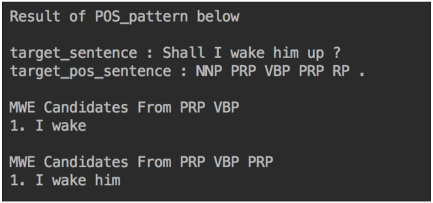
data processing

step 4



✓ **POS(Part Of Speech)**

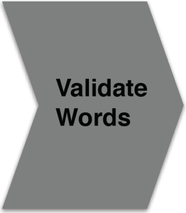
“ Shall	I	wake	him	up	? ”
(verb)	(pron)	(verb)	(pron)	(part)	(punc)



```
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/

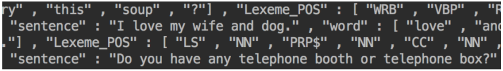
data processing

step 6

Store Results

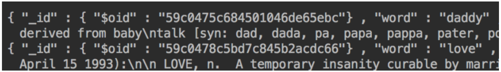
✓ **Data Base : MongoDB & JAVA**

✓ **Sentence Collection**



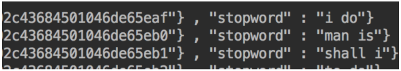
```
ry" , "this" , "soup" , "?"] , "Lexeme_POS" : [ "WRB" , "VBP" , "R
 "sentence" : "I love my wife and dog." , "word" : [ "love" , "and
."] , "Lexeme_POS" : [ "LS" , "NN" , "PRP$" , "NN" , "CC" , "NN" ,
 "sentence" : "Do you have any telephone booth or telephone box?"
```

✓ **Dictionary Collection**



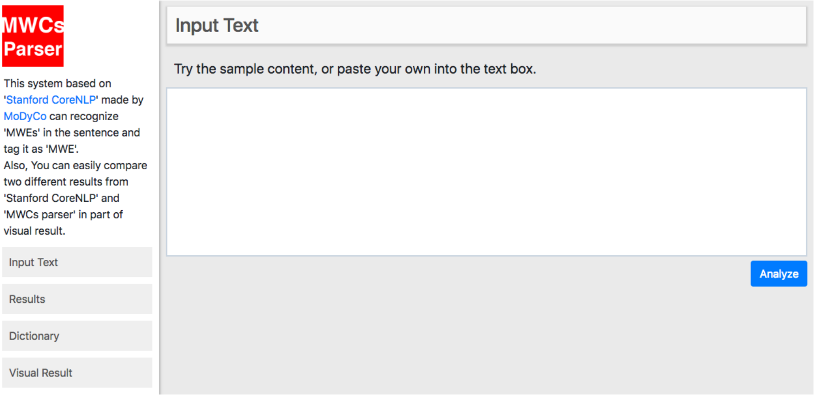
```
{ "_id" : { "$oid" : "59c0475c684501046de65ebc"} , "word" : "daddy"
  derived from baby\ntalk [syn: dad, dada, pa, papa, pappa, pater, po
{ "_id" : { "$oid" : "59c0478c5bd7c845b2acdc66"} , "word" : "love" ,
  April 15 1993):\n\n LOVE, n.  A temporary insanity curable by marri
```

✓ **Stopwords Collection**



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

MWCs parser

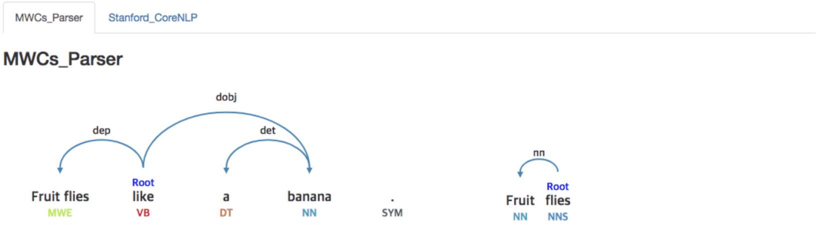


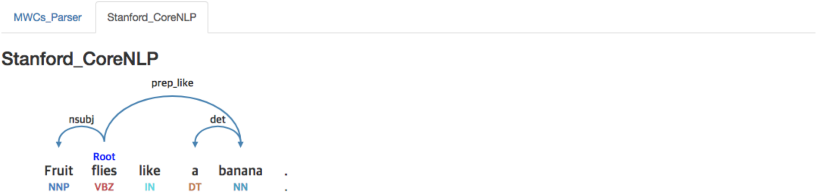
This system based on 'Stanford CoreNLP' made by MoDyCo can recognize 'MWEs' in the sentence and tag it as 'MWE'.
Also, You can easily compare two different results from 'Stanford CoreNLP' and 'MWCs parser' in part of visual result.

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

Video link (`https://www.youtube.com/watch?v=BddJ4kHDkxU`)

an ambiguous sentence





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

conclusion and perspectives

- MWCs parser is a syntactic parser taking MWCs into account which helps analyzing ambiguous sentences accurately
- MWCs parser can be improved collaboratively – access to the user dictionary + patterns database
- DepVis makes it possible to visualize MWCs both as (atomic) units (single POS in the sentence) AND as phrases (showing their internal syntactic structures)
- storing more sentences will improve the speed of the algorithm.
- storing more MWEs will allow the algorithm to recognize more MWEs.