Constructing a Corpus of Verbal MWEs in English

Abigail Walsh — ADAPT Centre, DCU

Claire Bonial — US Army Research Laboratory

Kristina Geeraert — University of Alberta

John P. McCrae — Insight Centre, NUIG

Nathan Schneider — Georgetown University

Clarissa Somers — Georgetown University

The **PARSEME Shared Task 1.1**, based on the previous ST 1.0[1], aims to

1. Identify VMWEs in running texts across a variety of languages
2. Establish a consistent set of guidelines for the annotation of these VMWEs

Addition of English in the new shared task resulted in a number of modifications to the annotation guidelines.

Annotation Timeline

Pilot Annotation 1

June - August 2017:
Discussion of guidelines and annotation issues

October 2017:
Guidelines 1.1 released (based on annotation feedback)

Pilot Annotation 2

November - December 2017:
Discussion of guidelines and annotation issues

Pilot Annotation 3

December 2017 - January 2018:
Final discussion of annotation practises

February 2018:
Annotation begins

Annotation Period

March 2018:
Annotation ends, consistency checks performed

4th of April 2018 Corpus released

April 2018:
Doubly annotated portion of corpus submitted for IAA study

PARSEME ST Annotation Guidelines

Updates Based on Pilot Annotations

Light Verb Constructions

Distinctions made between:
Fully light verbs (LVC.full): Verb adds no additional semantics
*She **has** a terrible **headache***

Caustitive light verbs (LVC.cause): Verb carries semantics of causation
*The buzzing radio **gave** him a **headache***

Additional guidance on abstract vs. concrete nouns:
*The scholarship plan would **provide** federal **contributions** to each […] school equal to $1500...*

Additional notes on productive nature of LVC.Cause
*A certain **vagueness** may also be **caused** by tactical appreciation of the fact...*

Verb Particle Constructions

Distinctions made between:
Fully non-compositional particles (VPC.full): verb meaning significantly changes
***Check in** on arrival*

Semi non-compositional particles (VPC.semi): verb meaning does not significantly change, but aspectual or other information added
*the Senate **passed** the bill **on** to the House*

Prioritising tests for particle vs. adposition:
*to **set aside** the privilege resolution*

About the Corpus

Labels Used
- VPC: Verb Particle Constructions (Full and Semi)
- LVC: Light Verb Constructions (Full and Cause)
- VID: Verbal Idioms
- MVC: Multi Verb Constructions
- IAV: Inherently Adpositional Verbs

Final Corpus
- **7437 sentences** from
  - English Web Treebank
  - LinES parallel corpus
  - PUD treebank
- **832 MWEs** were categorised,
- 4 dialects represented: Irish E, British E, American E, Canadian E
- Mixed genre dataset, including user-generated content, news and Wikipedia articles, and literary text

Pilot Annotations
200 sentences taken from the Brown corpus

IAA Study
804 sentences from the final corpus were annotated by each of the four annotators, and Kappa, Kappa-cat and F-score measures were calculated from this data.

Conclusion
Agreement between two annotators who completed all three pilot tasks (A3 and A4) is higher than between annotators who did not (A1 and A2), when annotating span of the VMWEs (F-score and Kappa), indicating that training improves identifying VMWEs, though not necessarily categorisation (Kappa-cat).

Results

Jointly Annotated Corpus

Category	Pilot 1	Pilot 2	Pilot 3	Final	Ann. 1	Ann. 2	Ann. 3	Ann. 4
VPC.full	40	33	49	297	27	41	62	41
VPC.semi	0	25	25	45	17	3	9	23
LVC.full	37	43	82	244	77	32	43	42
LVC.cause	0	21	44	43	28	2	5	11
VID	38	19	30	139	13	14	25	41
MVC	0	2	1	4	4	0	0	1
IAV	0	15	34	60	22	9	9	17
Total	115	158	265	832	188	101	153	176

IAA Study

Kappa Scores from IAA Study

Pair	F-score	Kappa	Kappa-cat
1x2	0.436	0.396	0.661
1x3	0.452	0.402	0.647
1x4	0.478	0.427	0.635
2x3	0.480	0.446	0.773
2x4	0.513	0.479	0.636
3x4	0.529	0.487	0.625

References

[1]Savary, Agata; Ramisch, Carlos; Cordeiro, Silvio Ricardo; Sangati, Federico; Vincze, Veronika; QasemiZadeh, Behrang; Candito, Marie; Cap, Fabienne; Giouli, Voula; Stoyanova, Ivelina; Doucet, Antoine; 2017, The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions, In *Proc. of the 13th Workshop on Multiword Expressions (MWE 2017),* pages 31-47, Valencia, Spain, April.

This work is funded by the Irish Government Department of Culture, Heritage and the Gaeltacht under the GaelTech Project, and is also supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) at Dublin City University