Anotation guidelines in the PARSEME shared task on automatic detection of verbal MWEs (VMWEs)

AIM-WEST/PARSEME-FR meeting
3-4 October 2016
IMAG, Grenoble, France

PARSEME

Challenges from verbal MWEs

Semantic non-compositionality

The VMWE's meaning cannot be deduced on the basis of its syntactic structure and of the meanings of its components.

Incorrect word-for-word translation

- (EN) *to* ***take the cake*** $\not\equiv$ (FR) *prendre le gâteau*

Both idiomatic and literal readings

- (EN) *to* ***take the cake*** 'to be outstanding'
- (FR) ***retourner*** *sa* ***veste*** lit. *to return the jacket* 'to change one's opinion in an opportunist way'

Implications for NLP

- VMWEs central **challenges in machine translation**.

Challenges from verbal MWEs

Discontinuity in texts

- (EN) *a **mistake** was frequently **made***,
- (EN) *to **turn** it **off***.

Implications for NLP

- VMWE identification requires **syntactic analysis** rather than sequence labeling.

Challenges from verbal MWEs

Unpredictable lexical or syntactic constraints

- (EN) *to **throw** sb **to the lions**, #to fling sb to the lions*,
- (EN) *to **take** a **decision**, *he took the committee's decision*.

Regular variability

- (EN) *he was **thrown to the lions***,
- (EN) *he **took/made** important **decisions***.

Implications for NLP

- The description of VMWEs can be limited neither to the level of the **lexicon** nor of the **syntax** only.

Challenges from verbal MWEs

Both idiomatic and literal readings

- (EN) *to **take the cake*** 'to be outstanding'
- (FR) ***retourner** sa **veste*** lit. *to return the jacket* 'to change one's opinion in an opportunist way'

Same syntactic structure and lexical choices, different VMWE categories

- (EN) *to **make** a **mistake*** is an **LVC** (light-verb construction),
- (EN) *to **make** a **meal** out of sth* 'to spend more time or energy doing something than needed' is an **idiom**.

Syntactic ambiguity

- (EN) *to **get** **up** a petition* - *up* is a particle,
- (EN) *to get up the hill* - *up* is a preposition.

Implications for NLP

- VMWE identification and categorization can **not** be based on **solely syntactic patterns**.

State of the art

MWE annotation in treebanks [13]

- Survey on 17 treebanks for 15 languages,
- Most frequently annotated MWE categories:
  - named entities,
  - continuous MWEs such as compound nouns, adverbs, prepositions and conjunctions
- Few VMWE annotations and for selected categories only,
- Heterogeneous annotation practices.

MWE annotation in UDs [11]

- 3 relations to signal MWEs: `compound`, `mwe` and `name`; inconsistent annotation practices [6],
- 3 different conventions for the LVC annotation [12]
  - no distinction for LVCs,
  - LVC annotation by structure (complements attaced to predicative nous rather than to light verbs),
  - LVC annotation by explicit dependency labels,

State of the art

VMWE datasets in English

- 50 English Wikipedia articles annotated for MWEs, including several VMWEs types [19],
- 2,162 sentences from the BNC with positive and negative examples of LVCs with 6 verbs (*do*, *get*, *give*, *have*, *make*, and *take*) [15],
- Crowdsourced corpus of positive and negative VPC examples for 6 verbs [16],
- VPCs corpus by [1],
- SZPFX - English–Hungarian parallel corpus with LVC annotations [17].

State of the art

VMWE datasets in other languages

- (DE) **V PrepNP** combinations: a database [10] and annotations in the TIGER corpus [5],
- (CS) identifying monosemic subtrees in the deep syntactic layer and replacing them by single nodes [3] linked to a MWE lexicon, which unifies different **morphosyntactic variants** of the same MWE [4],
- (ES) database and corpus of VMWEs for **particle verbs** [8, 9],
- (HU) database and corpus of VMWEs for **LVCs** [18],
- (TR) unique dependency label for all MWEs [7],
- (AR) dataset of **verb-noun** and **verb-particle** MWEs [2],
- (FA) **lvc** dependency labels used in a Persian treebank [14].

VMWE annotation

Motivations

- Unify the terminology on VMWEs in different domains (NLP, lingustics) and languages.
- Account for both universal and language-specific phenomena.
- Create best practices for VMWE annotation.

Methodology

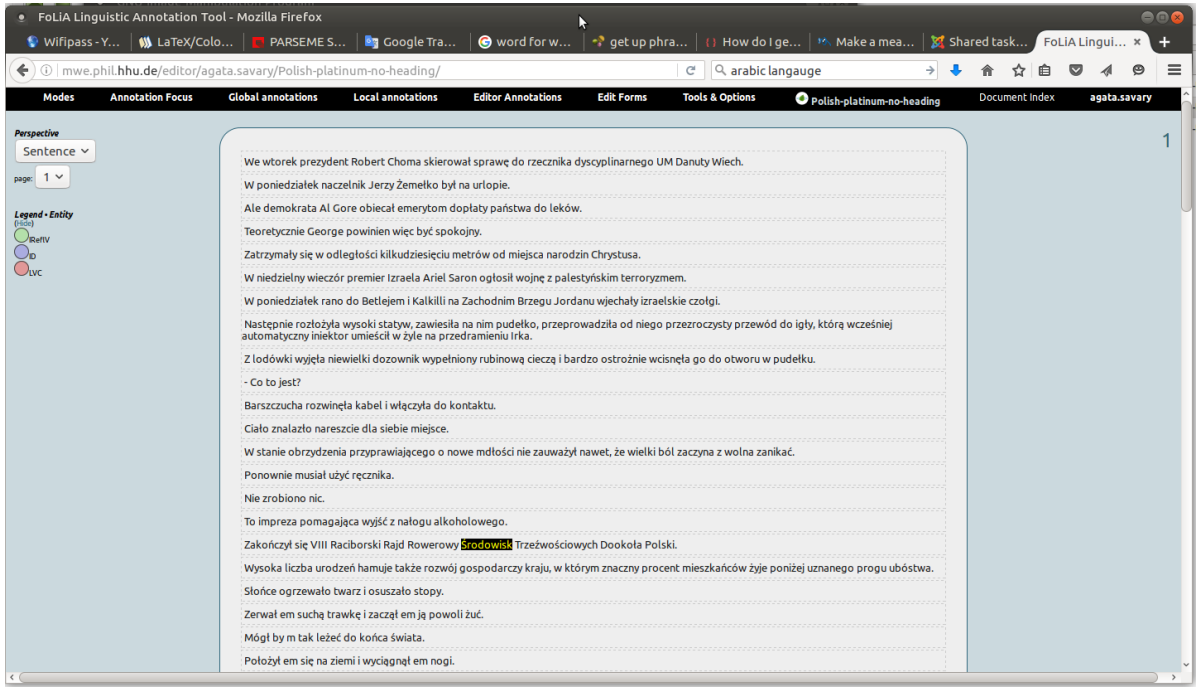
- 21 languages (4 non Indo-European), 4 language groups (Balto-Slavic, Germanic, Romance, other),
- roles:
  - 4 project managers,
  - 2 technical experts,
  - 5 language group leaders,
  - 21 language leaders.
- 2 pilot annotations (200 sentences per language),
- gradual enhancement of the guidelines (now v6).

CoNLL-U-like annotation format

1-2	Wouldn't				
1	Would				
2	not				
3	questioning				
4	colonial				
5	boundaries				
6	**open**			1	ID
7	a				
8	dangerous				
9	**Pandora**	nsp	A	1	
10	**'**	nsp	A	1	
11	**s**		A	1	
12	**box**	nsp		1	
13	?				

1	They					
2	were					
3	**letting**		1	VPC	2	VPC
4	him					
5	**in**		1			
6	and					
7	**out**				2	
8	.	nsp				

FLAT – on-line annotation tool (Radboud University, NL)



- Access rights according to roles.
- Discontinuous and overlapping word sequences.
- Language-specific tagset.

Basic definitions

Words vs. tokens

- Word = token, e.g. (EN) *take, astonishment*
- Multi-token words (MTWs): (EN) *Pandora|'|s*
- Multi-word tokens (MWTs): (IT) *della = de la*, (DE) ***aus|machen*** 'to open'

Multi-Word Expressions

Continuous or discontinuous sequences of words which:

- Show orthographic, morphological, syntactic or semantic **idiosyncrasy**. Collocations (sequences withs statistical idiosyncrasy only) are disregarded.
- Include a **head** word and at least one other **syntactically related** word.
- At least two components are **lexicalized**.

VMWEs

VMWEs

MWEs whose syntactic **head** in the prototypical form is a **verb**.

Prototypical verbal phrase

A phrase whose syntactic head is a verb in finite form and its other components are dependents of the verbs (or conjunctions in case of coordinated head verbs).

- (EN) ***break*** *her* ***heart***,
- (EN) ***a little bird told*** *someone*,
- (EN) ***drink and drive***,
- (EN) ***the early bird catches the worm***.

Meaning-preserving variants

Infinitives

- (EN) *to **break** one's **heart***

Nominal groups with relative clauses

- (EN) ***hearts** which he **broke***

Gerunds

- (EN) ***heart breaking***

Nominal and adjectival groups with participles

- (EN) ***heart-breaking***

VMWE typology

Universal categories

- Light-verb construction (**LVC**)
  - (EN) *to **give** a **lecture***
- Idiom (**ID**)
  - (EN) *to **go bananas*** 'to get crazy'

Quasi-universal categories

- Inherently reflexive verb (**IReflV**)
  - (FR) ***se suicider*** lit. *suicide SELF* 'suicide'
- Verb-particle construction (**VPC**):
  - (EN) *to **do in*** 'to kill'

Language-specific categories

Other (**OTH**)

- (EN) ***to drink and drive***
- (EN) ***to voice act***

Annotation guidelines demo - by Carlos

Bibliography I

Timothy Baldwin.
Deep lexical acquisition of verb-particle constructions.
Computer Speech and Language, 19(4):398–414, October 2005.

Kfir Bar, Mona Diab, and Abdelati Hawwari.
Arabic multiword expressions.
In Nachum Dershowitz and Ephraim Nissan, editors, *Language, Culture, Computation. Computational Linguistics and Linguistics: Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday, Part III*, pages 64–81, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

Eduard Bejček and Pavel Straňák.
Annotation of multiword expressions in the Prague dependency treebank.
Language Resources and Evaluation, 44(1–2):7–21, 2010.

Eduard Bejček, Pavel Straňák, and Daniel Zeman.
Influence of Treebank Design on Representation of Multiword Expressions.
In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I*, volume 6608 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2011.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit.
Tiger: Linguistic interpretation of a german corpus.
Research on Language and Computation, 2(4):597–620, 2005.

Bibliography II

Koenraad De Smedt, Victoria Rosén, and Paul Meurer.

MWEs in Universal Dependency treebanks.
http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015,
2015.

Gulsen Eryigit, Kubra Adali, Dilara Torunoglu-Selamet, Umut Sulubacak, and Tugba Pamay.

Annotation and Extraction of Multiword Expressions in Turkish Treebanks.
In *Proceedings of NAACL-HLT 2015*, pages 70–76. Association for Computational Linguistics, 2015.

Heiki-Jaan Kaalep and Kadri Muischnek.

Multi-Word Verbs in a Flective Language: The Case of Estonian.
In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*, pages 57–64, Trento, Italy, 2006. ACL.

Heiki-Jaan Kaalep and Kadri Muischnek.

Multi-Word Verbs of Estonian: a Database and a Corpus.
In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 23–26, Marrakech, Morocco, 2008.

Brigitte Krenn.

Description of Evaluation Resource – German PP-verb data.
In *Proceedings of MWE 2008*, pages 7–10, Marrakech, Morocco, June 2008.

Bibliography III

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee.
Universal dependency annotation for multilingual parsing.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Joakim Nivre and Veronika Vincze.
Light Verb Constructions in Universal Dependencies.
http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015,
2015.

Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mitetelu.
A survey of multiword expressions in treebanks.
In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference*, Warsaw, Poland, December 2015.

Mojgan Seraji, Carina Jahani, Beáta Megyesi, and Joakim Nivre.
A persian treebank with stanford typed dependencies.
In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

Bibliography IV

Yuancheng Tu and Dan Roth.

Learning English Light Verb Constructions: Contextual or Statistical.
In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth.

Sorting out the Most Confusing English Phrasal Verbs.
In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 65–69, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Veronika Vincze.

Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus.
In *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

Veronika Vincze and János Csirik.

Hungarian corpus of light verb constructions.
In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1110–1118, Beijing, China, 2010. Coling 2010 Organizing Committee.

Veronika Vincze, István Nagy T., and Gábor Berend.

Multiword expressions and named entities in the Wiki50 corpus.
In *Proceedings of RANLP 2011*, Hissar, Bulgaria, 2011.