Automated Phraseology Extraction and Cultural Factors: An Experiment

Jean-Pierre Colson

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

Abstract*:* This paper reports the results of an experiment with the *Parseme* 1.1. dataset for English. While the Parseme initiative represented a breakthrough in computational phraseology, it also raised a number of theoretical and practical issues. In this experiment, an attempt is made to improve the results obtained for English, by having recourse to external resources, in the form of a large web corpus. At the same time, attention is paid to the subtle interaction between linguistic tradition, culture and the manipulation of linguistic data in a supervised model for the automatic extraction of verbal multiword expressions. The results show that our algorithm, relying on an open track with external linguistic data, scores better in terms of recall, while deep learning systems yield a better precision. At various stages of the supervised model, the experiment shows that cultural factors play a crucial role.

Keywords: phraseology, automatic extraction, algorithm, culture

1. Introduction

→ Phraseology, the study of set phrases in the broad sense or *phrasemes* (Burger et al., 2007) is devoted to a broad palette of linguistic structures (e.g. collocations, formulas, partial idioms, idioms, clichés, proverbs). Formulaic language as defined by Wray (2008) includes formulas in the broadest sense, many of which have a communicative function. Finally, construction grammar or CxG (Croft, 2001; Goldberg, 2006, Hoffmann and Trousdale, 2013) considers that language as a whole consists of constructions in the sense of (partly) arbitrary pairings of form and meaning, at various degrees of abstraction and schematicity. In particular, *constructional idioms* such as constructions with *let alone* or with *just because* may also be included in phraseology in the broad sense.

→ For phraseology, formulaic language or construction grammar, however, a major theoretical issue is posed by the notion of *gold set*. Indeed, it is of paramount importance, in computational linguistics, to be able to test the performance of any model or algorithm against a gold set, whether you wish to segment a language automatically, or to extract any type of specific linguistic information.

→ For phrasemes, formulas or constructions, the trouble is that establishing any gold set is a particularly complex task. Relying on dictionaries offers only a partial view on the total number of phraseological structures: their description in the dictionary may be incomplete, or some of them may even be absent, particularly in the case of formulas or constructional idioms.

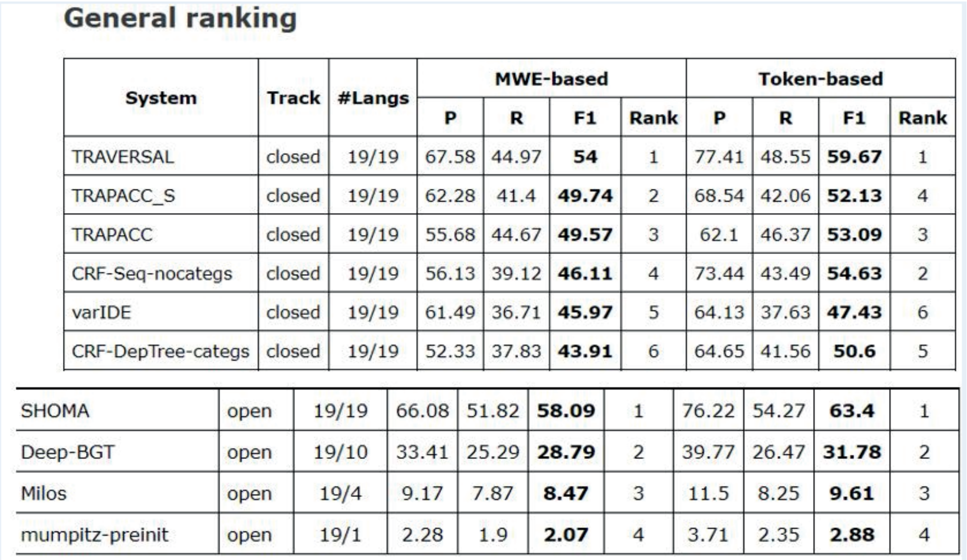
→ Producing a list of phrasemes from a dictionary is one thing, but extracting all cases of phraseology in the broad sense from a running text is a daunting challenge. A number of comparable texts have to be selected in different languages, a team of native speakers has to manually annotate all phraseological units and an acceptable degree of interpersonal agreement must be reached between them. As stated above, there is no agreement on the precise boundaries of phraseology, in particular with respect to formulas and constructional idioms. Paradoxically, many researchers, language teachers or students use the notion of phraseology, but looking for all phraseological units in a running text is partly subjective, because of the different views prevailing on what phraseology should include or not.

→ The PARSEME project (PARSing and Multiword Expressions, https://typo.uni-konstanz.de/parseme/), funded as a European COST Project from 2013 to 2017, was the first large-scale attempt to build a partial gold set of multiword expressions, roughly corresponding to phrasemes in the broad sense. It organised two shared tasks for the automatic identification of verbal multiword expressions, based on gold sets and training lists: edition 1.0 on the occasion of EACL 2017 (Markantonatou et al., 2017) and edition 1.1 on the occasion of COLING 2018 (Savary et al., 2018).

→ It should be pointed out that the Parseme shared tasks represent a real breakthrough in computational phraseology, as previous studies (summarised in Gries, 2013) were mainly limited to binary collocations, which were extracted from existing lists or dictionaries. In the case of the Parseme shared tasks, the research teams taking part in the task not only received a fully parsed file containing the test corpus in various languages, but also a training set in order to train their algorithm. The systems were accepted in a 'closed track' (i.e. by using just the training data, as in deep learning) or in an open track (allowing for the use of external resources). For all their merits, the Parseme shared tasks display a number of theoretical and practical issues that are debatable. In this paper, we present the results of an original experiment with the Parseme 1.1. data, in which we attempt to identify linguistic and cultural factors that may contribute to future research in the area of computational phraseology.

2. Some problematic aspects of the Parseme 1.1. results

→ The 2018 edition (1.1) of the Parseme shared task reached interesting results, though somehow disappointing, especially if we consider the

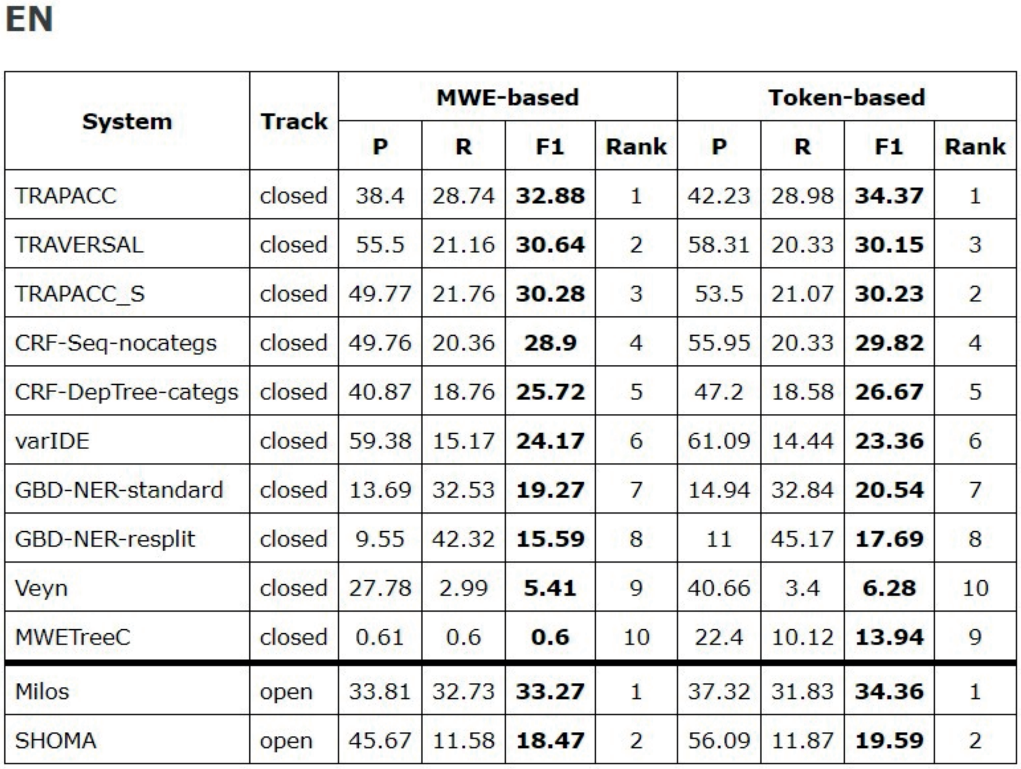


General ranking

System	Track	#Langs	MWE-based				Token-based			
			P	R	F1	Rank	P	R	F1	Rank
TRAVERSAL	closed	19/19	67.58	44.97	**54**	1	77.41	48.55	**59.67**	1
TRAPACC_S	closed	19/19	62.28	41.4	**49.74**	2	68.54	42.06	**52.13**	4
TRAPACC	closed	19/19	55.68	44.67	**49.57**	3	62.1	46.37	**53.09**	3
CRF-Seq-nocategs	closed	19/19	56.13	39.12	**46.11**	4	73.44	43.49	**54.63**	2
varIDE	closed	19/19	61.49	36.71	**45.97**	5	64.13	37.63	**47.43**	6
CRF-DepTree-categs	closed	19/19	52.33	37.83	**43.91**	6	64.65	41.56	**50.6**	5
SHOMA	open	19/19	66.08	51.82	**58.09**	1	76.22	54.27	**63.4**	1
Deep-BGT	open	19/10	33.41	25.29	**28.79**	2	39.77	26.47	**31.78**	2
Milos	open	19/4	9.17	7.87	**8.47**	3	11.5	8.25	**9.61**	3
mumpitz-preinit	open	19/1	2.28	1.9	**2.07**	4	3.71	2.35	**2.88**	4

Table 1. Results (general ranking) from the Parseme 1.1 shared task

data for English. Table 1 below displays the main results from the general ranking obtained for all languages (closed track and open track), and Table 2 the results for English[1].



EN

System	Track	MWE-based				Token-based			
		P	R	F1	Rank	P	R	F1	Rank
TRAPACC	closed	38.4	28.74	**32.88**	1	42.23	28.98	**34.37**	1
TRAVERSAL	closed	55.5	21.16	**30.64**	2	58.31	20.33	**30.15**	3
TRAPACC_S	closed	49.77	21.76	**30.28**	3	53.5	21.07	**30.23**	2
CRF-Seq-nocategs	closed	49.76	20.36	**28.9**	4	55.95	20.33	**29.82**	4
CRF-DepTree-categs	closed	40.87	18.76	**25.72**	5	47.2	18.58	**26.67**	5
varIDE	closed	59.38	15.17	**24.17**	6	61.09	14.44	**23.36**	6
GBD-NER-standard	closed	13.69	32.53	**19.27**	7	14.94	32.84	**20.54**	7
GBD-NER-resplit	closed	9.55	42.32	**15.59**	8	11	45.17	**17.69**	8
Veyn	closed	27.78	2.99	**5.41**	9	40.66	3.4	**6.28**	10
MWETreeC	closed	0.61	0.6	**0.6**	10	22.4	10.12	**13.94**	9
Milos	open	33.81	32.73	**33.27**	1	37.32	31.83	**34.36**	1
SHOMA	open	45.67	11.58	**18.47**	2	56.09	11.87	**19.59**	2

Table 2. Results (English) from the Parseme 1.1 shared task

→ These tables should read as follows. Both table show the results (in decreasing order of best performance) obtained by the various systems which submitted their results to the shared task. Those results were computed automatically from a script developed by the workshop organisers. The script precisely relied on a gold set of phrases that had to be identified. The gold set was produced by one or two native speakers, depending on the language. It should be pointed out that the agreement between the annotators was not checked against the whole set of data, but in a limited portion of it.

→ In spite of those shortcomings inherent to the organisation of an extraction task on a wide scale, the results displayed by Table 1 are rather convincing: the best overall system for all 20 languages was TRAVERSAL in the closed track (no use of external resources), with an F1 score[2] or 59.67 (token-based), while the best score in the closed track went to Shoma (F1 = 63.4). It is worthy of note, however, that all systems shown in Table 1 display a better value for precision than for recall. For instance, the best overall system, Traversal, "tells the truth" in more than 77 percent of the cases (i.e. the identified MWEs are indeed MWEs according to the gold set), but it failed to detect more than 50 per-

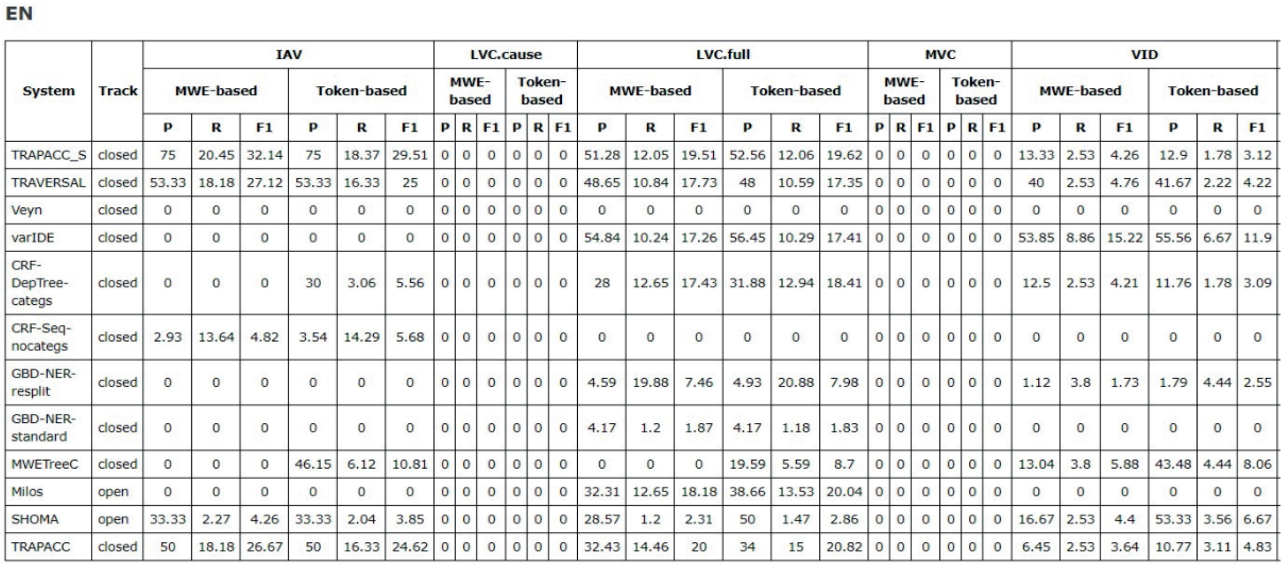
[1] Source: (http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_50_Shared_task_results)

[2] The F1 score corresponds the average between Precision (P in the table) and Recall (R in the table). For the extraction of phraseology, the recall score (expressed as a percentage or as a score from 0 to 1) corresponds to the proportion of items (phraseological units) that should have been extracted on the basis of the gold set and that were indeed extracted by the methodology. Precision corresponds to the percentage of items that were identified as positives (in this case phraseological units) and that were indeed positive. A careful algorithm may have a high precision because its results are mainly correct, but may miss out on many items that should also have been included, and therefore receive a low recall score.

cent of the MWEs that had to be extracted, as the recall score is just 48.55 percent (token-based).

→ The situation may therefore be considered as somehow imperfect for the 20 languages involved in the Parseme 1.1. shared task, but taking a look at the specific situation of English also speaks volumes. In spite of the fact that English is the most important international language, and is certainly the most documented language of the world and a constant object of study for linguistics, including corpus linguistics, the results are incredibly poor. Thus, as shown by Table 2, the best overall system, Traversal, just reached for English a precision score of 58.31 (token-based) and a recall of no much than 20.33 percent. In other words, almost 80 percent of the MWEs that were identified in the gold set were completely ignored by the best overall system.

→ Another very surprising point in the Parseme 1.1. results is the even poorer scores obtained for English verbal idioms (e.g. *spill the beans*, *paint the town red*). This category of very idiomatic constructions has received a lot of interest in linguistic studies, and English is in addition, as mentioned above, the most documented language in the world. As shown in Table 3, no system taking part in the Parseme 1.1. shared task, however, reached acceptable precision and recall scores for English verbal idioms.


EN

System	Track	IAV						LVC.cause						LVC.full						MVC						VID					
		MWE-based			Token-based			MWE-based			Token-based			MWE-based			Token-based			MWE-based			Token-based			MWE-based			Token-based		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TRAPACC_S	closed	75	20.45	32.14	75	18.37	29.51	0	0	0	0	0	0	51.28	12.05	19.51	52.56	12.06	19.62	0	0	0	0	0	0	13.33	2.53	4.26	12.9	1.78	3.12
TRAVERSAL	closed	53.33	18.18	27.12	53.33	16.33	25	0	0	0	0	0	0	48.65	10.84	17.73	48	10.59	17.35	0	0	0	0	0	0	40	2.53	4.76	41.67	2.22	4.22
Veyn	closed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
varIDE	closed	0	0	0	0	0	0	0	0	0	0	0	0	54.84	10.24	17.26	56.45	10.29	17.41	0	0	0	0	0	0	53.85	8.86	15.22	55.56	6.67	11.9
CRF-DepTree-categs	closed	0	0	0	30	3.06	5.56	0	0	0	0	0	0	28	12.65	17.43	31.88	12.94	18.41	0	0	0	0	0	0	12.5	2.53	4.21	11.76	1.78	3.09
CRF-Seq-nocategs	closed	2.93	13.64	4.82	3.54	14.29	5.68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GBD-NER-resplit	closed	0	0	0	0	0	0	0	0	0	0	0	0	4.59	19.88	7.46	4.93	20.88	7.98	0	0	0	0	0	0	1.12	3.8	1.73	1.79	4.44	2.55
GBD-NER-standard	closed	0	0	0	0	0	0	0	0	0	0	0	0	4.17	1.2	1.87	4.17	1.18	1.83	0	0	0	0	0	0	0	0	0	0	0	0
MWETreeC	closed	0	0	0	46.15	6.12	10.81	0	0	0	0	0	0	0	0	0	19.59	5.59	8.7	0	0	0	0	0	0	13.04	3.8	5.88	43.48	4.44	8.06
Milos	open	0	0	0	0	0	0	0	0	0	0	0	0	32.31	12.65	18.18	38.66	13.53	20.04	0	0	0	0	0	0	0	0	0	0	0	0
SHOMA	open	33.33	2.27	4.26	33.33	2.04	3.85	0	0	0	0	0	0	28.57	1.2	2.31	50	1.47	2.86	0	0	0	0	0	0	16.67	2.53	4.4	53.33	3.56	6.67
TRAPACC	closed	50	18.18	26.67	50	16.33	24.62	0	0	0	0	0	0	32.43	14.46	20	34	15	20.82	0	0	0	0	0	0	6.45	2.53	3.64	10.77	3.11	4.83

Table 3. Results (English verbal MWEs) from the Parseme 1.1 shared task

→ As illustrated by Table 3, the category of English verbal idioms (VID) actually reached the lowest scores across all systems: the best precision score (55.56, token-based) was obtained by the varIDE system in the closed track, but with a recall of only 6.67, which means that 93.33 percent of all verbal idioms present in the gold set were not recognised by the system. In addition, this poor recall score of 6.67 is the best one for English verbal idioms. In comparison, the best system across all languages (TRAVERSAL) reached only a recall score of 2.22 for English verbal idioms, which is certainly food for thought.

→ In the same way, light-verb constructions (e.g. *take a look*, *make a point*), were also problematic for most systems taking part in this shared task: as shown in Table 3, the best results for the LVC.full category were a precision of 56.45 but a pretty low recall of 10.29.

3. The experiment: methodological issues

→ In order to better understand the somehow intriguing results yielded by the Parseme 1.1. shared task for English, we specifically designed a new extracting system for the English Parseme dataset.

→ Our system is based on the *cpr-score* (Colson 2017): this score, derived from metric clusters used in information retrieval, measures the average distance between the component grams

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# source_sent_id = . . 2739
# text = When a user connects to the SQL Server database through a Microsoft Access project, the connection is enabled
through a Windows NT user account.
1    When      when      SCONJ  _         _         4         mark      _         _         *
2    a         a         DET    IND-SG _  3         det       _         _         *
3    user      user      NOUN   SG-NOM    _         4         nsubj     _         _         *
4    connects connect    VERB   PRES      _         19        advcl     _         _*
5    to        to        ADP    _         _         9         case      _         _         *
6    the       the       DET    DEF       _         9         det       _         _         *
7    SQL       sql       PROPN  SG-NOM    _         8         compound  _         _*
8    Server    server    NOUN   SG-NOM    _         9         compound  _         _*
9    database database NOUN   SG-NOM    _         4         obl       __*
10   through   through   ADP    _         _         14        case      _         _         *
11   a         a         DET    IND-SG _  14        det       _         _         *
12   Microsoft           microsoft          PROPN   SG-NOM    _14       compound*
13   Access    Access    PROPN  SG-NOM    _         12        flat      _         _         *
14   project   project   NOUN   SG-NOM    _         9         nmod      _         SpaceAfter=No   *
15   ,         ,         PUNCT  Comma _   4         punct     _         _         *
16   the       the       DET    DEF       _         17        det       _         _         *
17   connection          connection         NOUN    SG-NOM    _         19        nsubj:pass*
18   is        be        AUX    PRES-AUX  _         19        aux       _         _*
19   enabled   enable    VERB   PASS      _         0         root      _         _         *
20   through   through   ADP    _         _         25        case      _         _         *
21   a         a         DET    IND-SG _  25        det       _         _         *
22   Windows             window PROPN     SG-NOM    _         25        compound  _         _*
23   NT        nt        PROPN  SG-NOM    _         22        flat      _         _         *
24   user      user      NOUN   SG-NOM    _         25        compound  _         _*
25   account   account   NOUN   SG-NOM    _         19        obl       _         SpaceAfter=No*
26   .         .         PUNCT  Period    _         19        punct     _         _         *
```

Table 4. Example of a UDPipe output text (Parseme 1.1., English dataset)

of an n-gram, given a maximal window that is experimentally set on the basis of the language and the corpus (for a full explanation, see Colson, 2018). In other words, the semantic links are considered as a space model, in which metric distance between the elements is an indication of semantic attraction.

→ An experimental tool based on the *cpr-score* is available at https://idiomsearch.lsti.ucl.ac.be. The tool offers the possibility of entering an input text in English, Spanish, French or Chinese and to receive a corpus-based analysis of the main phraseological units present in the source text. In addition, another function of the tool can measure the *cpr-score* for any n-gram (from bigrams to 10-grams) in English, German, Spanish, French, Dutch and Chinese). As pointed out in Section 1, it is particularly difficult to measure the efficiency (in terms of precision and recall) of any general extraction method for phraseology, because native speakers will never totally agree on any gold set. In Colson (2018), we have therefore proposed to measure extraction methods on a closely related linguistic task, for which there is much documentation and a much higher rate of inter-individual agreement, namely Chinese word segmentation. As explained in Colson (2018), our *cpr-score* obtained for Chinese word segmentation reached the average degree of inter-individual agreement between native speakers, namely a recall of 70 %, measured automatically from an established gold set.

→ In an extraction task such as Parseme 1.1., it is not sufficient to have a statistical score, a training set and a set of data to be analysed. It is also crucial to choose between an unsupervised and a supervised model. We chose for the latter, because we precisely wished to check which changes to the algorithm were able to improve the results, and for what precise reason.

→ Contrary to most systems taking part in the Parseme 1.1. shared task, our methodology used the "open track": a recourse to external linguistic data and no use at all of the training data. We actually used a huge web corpus for English (the ukWaC corpus, 1.4 billion tokens) which can be downloaded from the web[3].

→ The Parseme data are already parsed on the basis of a sophisticated tool, UDPipe (http://ufal.mff.cuni.cz/udpipe). The output produced is exemplified in Table 4, the first sentence from the English dataset of Parseme 1.1.

→ The UDPipe parser uses for text annotation the CoNLL-U Format: sentence receive, vertically, the following fields (https://universaldependencies.org/format.html):

→ Horizontally, each line will display an entry for each of the vertical fields. Thus, at ID 4, *connects* is the FORM field, whereas *connect* is the corresponding LEMMA. In addition to this tagging, the sentence is parsed and all dependencies are indicated. For instance, at ID 24, the FORM *account* receives as its HEAD (column 7) the figure 19, which is the ID of *enabled.* This indicates that *account*, in *enabled through a Windows NT user account*, depends syntactically on *enabled*.

→ From a computational point of view, each sentence is parsed as shown by Table 4, which creates a multidimensional table or *hash*. Most object-oriented programming languages can cope easily with such tables, by having recourse to objects such as the Java HashMap of the Py-

[3] The WaCky corpora can be downloaded from https://wacky.sslmit.unibo.it/doku.php.

ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes (decimal numbers can be lower than 1 but must be greater than 0).

FORM: Word form or punctuation symbol.

LEMMA: Lemma or stem of word form.

UPOS: Universal part-of-speech tag.

XPOS: Language-specific part-of-speech tag; underscore if not available.

FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.

HEAD: Head of the current word, which is either a value of ID or zero (0).

DEPREL: Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.

DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.

MISC: Any other annotation.

thon Dictionary objects. As our project entailed complex word manipulations and interaction with a huge corpus, we chose for the Perl programming language, with the Data::Table object.

→ It should be stressed that checking collocations or idioms by means of a statistical score is one thing, but extracting phraseology from running text is another, and indeed a daunting task. In the case of the Parseme shared task, it is made even more complex by the fact that not only does it require to extract all verbal expressions, but also to label them according to the correct category that was chosen in the gold set. The full list of categories included: IAV (inherently adpositional adverbs, e.g. *to come across*), LVC.cause (light-verb constructions in which the verb adds a causative meaning to the noun, e.g. *to grant rights*), LVC.full (light-verb constructions in which the verb only adds meaning expressed by morphological features, e.g. *to give a lecture*), MVC (multi-verb constructions, e.g. *to make do*), VID (verbal idioms, e.g. *to go bananas*), VPC.full (fully non-compositional verb-particle constructions in which the particle totally changes the meaning of the verb, e.g. *to do in*), VPC.semi (semi-compositional verb-particle constructions, in which the particle adds a partly predictable but non-spatial meaning to the verb, e.g. *to eat up*).

→ From the point of view of corpus and computational linguistics, there are very diverse methods for reaching this goal. Our methodology consisted in using a supervised model in which one category of verbal expression was added at a time. A computer program (Perl script) automated a number of requests to the corpus, by checking all *cpr-scores* for the relevant category. For instance, in the case of the LVC.full category (*to give a lecture*), all verbs accompanied by an object and all nouns that were subject of passive verbs were checked on the corpus. If the association score was high enough, the combination was labelled as a light-verb construction. The next category of verbal expressions was then treated, and so on. After all categories were treated, the precision and recall were measured on the basis of the training data, and this is precisely where theoretical questions arise, in addition to purely technical issues.

→ While trying to improve the precision and recall of this extraction method, we sought to examine the following theoretical question: to what extent do cultural factors play a role in automatic phraseology extraction? One may indeed wonder if the technical aspects of multiword extraction, as illustrated by Tables 1 to 4, suffice to deal with the complex domain of phraseology, in which cultural factors play a central role.

3. Results and discussion

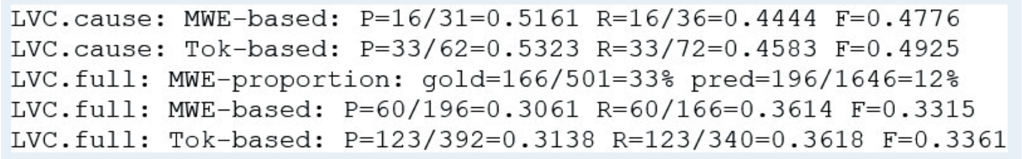
→ In our experiment with the English dataset of the Parseme 1.1. shared task, we first sought to improve the results for light-verb constructions. As shown in Table 3, the different systems taking part in the shared task did not make the difference between LVC.full and LVC.cause, as they were supposed to do. In addition, the best results for LVC.full are disappointing: the best overall system for all languages, TRAVERSAL, reached for the English LVC.full category an F1-score of 17.35 (token-based), with a fairly good precision at 48, but a poor recall of 10.59. The best score was reached by TRAPACC in the open track, with F1 at 20.82 (token-based), again with a recall (at 15) much lower than precision (at 34).

→ In order to extract the LVC.full category from the dataset, we actually extracted all verbs and their direct object, as well as all subjects followed by a passive verb. This was easy to achieve with the parsed data as shown in Table 4. The next step was simply to select the relevant combinations according to their association score (*cpr-score*). In our supervised model, each improvement was tested against the training data, which made it possible to check which technical improvement had positive results, and for what precise reason. Crucially in this case, the supervised model tested the introduction of external data.

→ Indeed, our hypothesis was that the category of light-verb constructions corresponds to a specific feature of English (and other European languages): in this construction, verbs carry little meaning (e.g. *take a walk*) and are therefore very generic verbs, with a high frequency. We therefore made a list[4] of the 100 most frequent English verbs and added it to the supervised model. When a relevant *cpr-score* (higher than 0.06) was found for a verb/direct object construction or for a subject/passive verb construction, the verb was first checked as a high frequency verb before the algorithm assigned the label "LVC" (light-verb construction). This line of reasoning also made it possible to make the difference between LVC.full and LVC.cause, as required by the Parseme gold set, in the following way. Before deciding whether the result should be classified either as LVC.full or as LVC.cause, an additional list of English causative verbs was checked, consisting simply of the following verbs: *bring, cause, draw, foster, generate, incite, occasion, perform, produce, provoke, raise, result, yield*. When the frequent verb was identified as a causative verb, the label LVC.cause was assigned, instead of LVC.full. The results obtained for the LVC category by our methodology were then the following (Table 5):

→ As illustrated by Table 5, our experiment made it possible to reach better scores for English light-verb constructions than those obtained by the other systems: the F1-score (represented as F in Table 5) was 33.61 percent (token-based)

[4] Lists of the most frequent English verbs are freely available on the Web, for instance at https://www.linguasorb.com/english/verbs/most-common-verbs/



```
LVC.cause: MWE-based: P=16/31=0.5161 R=16/36=0.4444 F=0.4776
LVC.cause: Tok-based: P=33/62=0.5323 R=33/72=0.4583 F=0.4925
LVC.full: MWE-proportion: gold=166/501=33% pred=196/1646=12%
LVC.full: MWE-based: P=60/196=0.3061 R=60/166=0.3614 F=0.3315
LVC.full: Tok-based: P=123/392=0.3138 R=123/340=0.3618 F=0.3361
```

Table 5. Results obtained in the experiment for English light-verb constructions

for the category LVC.full, as opposed to the best score of 20.82 in the other systems, and the additional category of LVC.cause, not treated by the other systems, reached an F of 49.25 percent. More importantly, our results display no significant difference between precision and recall. For the LVC.full, category, the recall is even slightly better than precision, as opposed to the situation prevailing in the results from the other systems.

→ While there is obviously always a margin for improvement in those results, it should be reminded that the gold set according to which the results are measured was just assembled by two native speakers, and that there are some inconsistencies in their judgment, in addition to the sometimes subjective choice between one category of verbal expressions instead of another. From a theoretical point of view, it is noteworthy that our supervised model in the open track (as there is a recourse to an external corpus and to some lexicographic data) reaches better results when cultural elements are added. Indeed, the category of light-verb constructions could only be extracted with some success from the data when a specific list of high frequency English verbs was provided. This means that our methodology does not validate the existence of such constructions per se. The algorithm was only able to extract them when a crucial, culture-specific part of the construction was explicitly added to the model, viz. high frequency verbs. For light-verb constructions, the interaction between externals factors (related to culture in the broad sense) and the automatic extraction of phraseology is therefore clearly demonstrated by our experiment.

→ For lack of space, we will limit the presentation of our other results to one central category, that of verbal idioms (VID). As shown in Table 3 and as mentioned above, the results of the Parseme 1.1. shared task for English verbal idioms were extremely poor, with the best overall recall at 6.67 percent (token-based). Table 6 displays our results for this category.

```
* VID: MWE-proportion: gold=79/501=16% pred=269/1646=16%
* VID: MWE-based: P=22/269=0.0818 R=22/79=0.2785 F=0.1264
* VID: Tok-based: P=49/587=0.0835 R=49/225=0.2178 F=0.1207
```

Table 6. Results obtained in the experiment for English verbal idioms

→ In order to produce these results for the VID category, we selected all verbs and all their adjuncts or complements, which was easy to carry out by checking which tokens from the parsed data (see Table 4) were used with a verbal HEAD category. All resulting n-grams were then checked according to the *cpr-score* and associations higher than 0.69 were considered as idioms.

→ As shown by Table 6, our results for English verbal idioms could certainly be improved, but is should be borne in mind that the gold set was not very reliable for all idioms. A striking result is also that our methodology may be poor for precision, but it is far better than all other systems for recall, as the best recall score obtained for English verbal idioms (6.67 percent) does not compare with our figure of 27.85 percent. Most other systems used deep learning (with neural networks or conditional random fields or both). It turns out that such methodologies cannot cope very well with such sparse data as verbal idioms, because they are very unlikely to be present both in the test data and in the training dat. Suppose, for instance, that an idiom such as *spill the beans* is used in the test data and identified as such in the gold test. It is then very unlikely that this idiom will be used in the training data as well. The deep learning systems then have to rely on other characteristics of idioms to extract them, but the trouble is that these are very unpredictable and that idioms may even be ungrammatical (as in *long time no see*). All in all, our methodology with an open track (checking associations in a corpus) turned out to produce a far better recall.

→ Another aspect of our methodology for English verbal idioms confirms the results gained for the LVC category: it was necessary to refine the results by having recourse to a specific list of linguistic items. Indeed, some results (receiving a very high *cpr-score*, namely higher than 0.69) were actually verb-particle constructions according to the Parseme gold set. This choice also corresponds to a cultural factor. Nothing in the extracted data makes it possible to confirm that there should indeed be a difference, in the description of English verbal expressions, between an idiom and a verb-particle construction. If we want to take this cultural argument (as it refers to a given tradition) into account, we must actually add a specific list of English particles to the algorithm. This again confirms our previous remark that no automatic extraction of phraseology can be neutral: some choices in the gold set, in the numbers of categories or in the tokens and their parsing, will inevitably depend on a number of traditions that are specific to a given language or culture.

4. Conclusions

→ In our experiment, we have tried to improve the English results obtained by systems which submitted to the Parseme 1.1. shared task on the extraction of verbal multiword expressions. Contrary to most participating systems, we have chosen for an open track (having recourse to external data): the use of a large linguistic corpus, in combination with a new clustering algorithm.

→ In spite of the shortcomings of the English gold set used by Parseme 1.1., we have reached far better results for the extraction of two central categories of the phraseology around verbs, namely light-verb constructions and verbal idioms. Contrary to the results obtained by deep learning approaches, our methodology makes it possible to reach acceptable recall rates, which are often better than the precision rates. Deep learning approaches display an exactly opposite situation: the precision scores are good, while recall is very low. Our hypothesis is therefore that deep learning algorithms cannot treat verbal phraseology adequately, because of the sparse nature of structures such as verbal idioms. If a given verbal idiom is not present in the training data, deep learning algorithms will probably be unable to extract it in another set of data. On the contrary, our methodology is based on a large corpus, in which recurrent patterns can be measured, even for less common verbal idioms. As the deep learning approach reaches good precision scores, whereas our methodology performs better on recall, a conclusion of this experiment is also that a combination of both approaches may be fruitful in future studies.

→ The automatic extraction of phraseology from a running text is not just a technical issue. Some major theoretical questions are also at stake. Among these, the interaction between distributional semantics and training data is crucial. Indeed, deep learning algorithms attempt to reproduce human decisions that were made in a gold set of data. In distributional semantics, on the other hand, it is presumed that a large collection of linguistic data will display preferred statistical associations that are independent of any human judgment. In our results, we found that some distinctions between categories of verbal expressions that were made in the gold set, could only be reproduced by our algorithm if external linguistic data corresponding to a linguistic tradition were added to the system. This is clear, for instance, in the case of causative verbal expressions: the existence of such a category is a choice made by linguists on the basis of cultural factors, and the only way of reproducing this in our methodology was to add a specific list of causative verbs to the algorithm.

→ From a theoretical point of view, the results of our experiment therefore confirm that the debate on the best methods for automatic phraseology extraction is also one about the nature of semantics and semiotics: are there statistical associations that are naturally present in the data, or do wo only approximate meaning by adding specific, mainly cultural factors to our methodology?

Bibliography

BURGER, Harald / DOBROVOL'SKIJ, Dmitrij / KÜHN, Peter / NORRICK, Neal, eds. (2007), *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*, Berlin / New York, De Gruyter.

COLSON, Jean-Pierre (2017), "The Idiom Search Experiment: Extracting Phraseology from a Probabilistic Network of Constructions", in MITKOV, Ruslan (ed.), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596.* Cham, Springer International Publishing, pp. 16-28.

COLSON, Jean-Pierre (2018), "From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin" in Savary, A. et al. (eds.) (2018), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, Association for Computational Linguistics, 41-50.

CROFT, William (2001), *Radical Construction Grammar: Syntactic Theory in Typological Perspective*, Oxford, Oxford University Press.

GOLDBERG, Adele (2006), *Constructions at Work*. Oxford, Oxford University Press.

GRIES, Stefan (2013), "50-something years of work on collocations. What is or should be next…", *International Journal of Corpus Linguistics*, 18, 137-165.

HOFFMANN, Thomas / TROUSDALE, Graeme, eds. (2013), *The Oxford Handbook of Construction Grammar*, Oxford/NewYork, Oxford University Press.

MARKANTONATOU, Stella / RAMISCH, Carlos / SAVARY, Agata / VINCZE, Veronika, eds. (2017), *Proceedings of the 13th Workshop on Multiword Expressions (EACL 2017)*. Valencia, Association for Computational Linguistics.

SAVARY, Agata / RAMISCH, Carlos / HWANG, Jena D. / SCHNEIDER, Nathan / ANDRESEN, Melanie / PRADHAN, Sameer / PETRUCK, Miriam R.L., eds. (2018), *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, Association for Computational Linguistics.

WRAY, Alison (2008), *Formulaic Language: Pushing the Boundaries*, Oxford, Oxford University Press.

Biographic profile: Jean-Pierre Colson is professor and chairman of the Department of Translation and Interpreting at the University of Louvain (Louvain-la-Neuve, Belgium). He is also a member of the Board of the European Association for Phraseology (Europhras) and has published many papers on phraseology, translation studies and computational linguistics. In the last years his works are dedicated to the automatic processing of multi-word units in large electronic corpora.

e-mail: jean-pierre.colson@uclouvain.be