VMWE discovery: a comparative analysis between Literature and Twitter Corpora

Vivian Stamou[1], Artemis Xylogianni[2], Marilena Malli[2],
Penny Takorou[2] and Stella Markantonatou[1]
Institute for Language and Speech Processing (ILSP / "Athena" R.C.)[1]
{vivianstamou, stilianimarkantonatou}@gmail.com
Department of French and Language Literature, University of Athens[2]
{artemis.xylo, mallimariaeleni, pennytak07}@gmail.com

Abstract

We evaluate manually five lexical association measurements as regards the discovery of Modern Greek verb multiword expressions with two or more lexicalised components using mwetoolkit3 (Ramisch et al., 2010). We use Twitter corpora and compare our findings with previous work on fiction corpora. The results of LL, MLE and T-score were found to overlap significantly in both the fiction and the Twitter corpora, while the results of PMI and Dice do not. We find that MWEs with two lexicalised components are more frequent in Twitter than in fiction corpora and that lean syntactic patterns help retrieve them more efficiently than richer ones. Our work (i) supports the enrichment of the lexicographical database for Modern Greek MWEs 'IDION' (Markantonatou et al., 2019) and (ii) highlights aspects of the usage of five association measurements on specific text genres for best MWE discovery results.

1 Introduction

Most prior work on MWE discovery exploits lexical association measures (AMs). MWE discovery experiments on Twitter corpora has investigated alternatives to AMs. Daoud et al. (2015), who study Arabic tweets, generate candidate MWEs by storing bi-/tri-grams and their frequencies, filter them on the basis of the frequency of relevant query responses and ensure filtering validity with a confidence ratio measure that takes into account the number of distinct tweets in which a MWE appears. Londhe et al. (2016) propose a language agnostic, graph based method for multilingual MWE extraction (bigrams) from Twitter corpora, evaluate it on various datasets and yield promising results in a number of languages; in the case of English, however, AMs outperform their method.

Prior work on verb MWEs (VMWEs) has focused on expressions with two lexicalised components (Markantonatou et al., 2018) namely to particle verbs and light verb constructions (Stevenson et al., 2004; Constant et al., 2017). Reported work on Modern Greek VMWE discovery with the use of AMs includes Stripeli et al. (2020), who work with light verbs, and Stamou et al. (2020), who work with all types of VMWEs in fiction corpora. To the best of our knowledge, this is the first work on mining VMWEs with 2 or more lexicalised components in Modern Greek Twitter corpora (with lexicography as a goal).

AM performance has been found to be subject to corpus characteristics (size, type etc.) and to the type of target MWEs (see the AMs' evaluation literature with regards to VMWEs, e.g., German and English Verb+Preposition MWEs (Krenn et al., 2011; Baldwin et al., 2005), Verb particle MWEs (Hoang et al., 2009), English, Portuguese and Spanish Verb+Object MWEs (Garcia et al., 2019). Since different AMs seem to achieve best results on different corpora, MWE types and frequencies (Garcia et al., 2019), combinations of AMs have been found to be more efficient for discovery purposes (Pecina and Schlesinger, 2006). We use the mwetoolkit3 (Ramisch et al., 2010), and discuss the application of five AMs, namely Dice, Log likelihood (LL)[1], MLE (Maximum Likelihood Estimation), PMI (Pointwise Mutual Information) and T-score, on Twitter corpora. The tool offers a complete pipeline for MWE extraction

[1] It should be noted that LL is implemented only for bigrams.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 66–72
Barcelona, Spain (Online), December 13, 2020.

where candidate phrases are created with n-grams or predefined linguistic patterns applied on corpora annotated for lemma and PoS tags.

In Section 2 we present the details reported in Stamou et al. (2020) that are important to the work discussed in this paper. In Section 3 we present and evaluate three experiments. In Section 4, the differences observed between the two datasets, namely Twitter and fiction corpora, are discussed. In Section 5 we present our conclusions and comment on the use of the AMs for lexicographic purposes.

2 Experiments with fiction corpora of Modern Greek

Stamou et al. (2020) used `mwetoolkit3` (Ramisch et al., 2010) to discover VMWEs in fiction corpora tagged and lemmatised with the ILSP tools (Papageorgiou et al., 2000). A Gold Standard (GS) was defined from these corpora by three expert annotators. Two experiments were conducted, one with simple and one with enriched linguistic patters. In the first experiment, six syntactic patterns were applied comprising the most frequent PoS sequences in the GS. The results were evaluated both manually and automatically. Manual evaluation was applied to identify True and False positives in the first 3000 candidates returned by each of the following AMs: Dice, LL, MLE, PMI, T-score. In the second experiment, richer patterns were used featuring double prepositional phrases and conjunction structures, as shown in Table 1 (brackets denote optionality).

Table 1: Rich VMWE patterns used by Stamou et al. (2020)

Patterns
(Pn)+(Vb)+**Vb**+(Ad)+(At)+(Aj)+(At)+**No**+(Pn)+(Aj)
Vb+(At)+(No)+**Cj**+(Pn)+(At)+(No)
Vb+**Cj**+(Pt)+(Pt)+**Vb**
(Pn)+Pn+(Pt)+(Pt)+(Vb)+**Vb**+(Pn)+**No**+(Pt)+(Pn)+(Vb)+**Vb**
(Pn)+(Vb)+**Vb**+(At)+(No)+(Ad)+**AsPp**+(Aj)+(At)+**No**
(Pn)+(Vb)+**Vb**+(Cj)+(Ad)+(At)+(No)+**AsPp**+(Ad)+(At)+**No**+(At)+(No)
Vb+**AsPp**+(At)+**No**+ **AsPp**+(At)+No
(Pn)+(At)+(No)+(Pt)+**Pn**+(Vb)+**Vb**
(Pn)+(Vb)+**Vb**+**Ad**+(Ad)+(Pn)

The automatic evaluation of both experiments against the GS returned the following order of scores: T-score, MLE, LL, Dice, PMI while the manual evaluation showed Dice as the most reliable AM; PMI scored last in the first experiment and LL in the second. Reliability was expressed as interanotator agreement computed with Fleiss $\varkappa$ values. Similar observations can be found in Linardaki et al. (2010) and Gurrutxaga and Alegria (2011) for nominal Modern Greek MWEs and Basque VMWEs respectively.

In the first experiment, T-score, MLE and LL shared about 850 out of the first 3000 phrases returned by each one of these AMs. Dice and PMI promoted less frequent VMWEs not included in the other AM results. These facts suggested that annotators tended to select hapax legomena, in contrast to automatic evaluation that relies on an, inevitably, incomplete GS.

Summing up Stamou et al. (2020) work: (i) rich syntactic patterns enhanced VMWE discovery results (ii) manual evaluation supported the discovery of less frequent VMWEs (iii) Dice (the most reliable AM in both experiments) and PMI (it returned VMWEs not found by the other AMs) should be applied for a more efficient VMWE discovery procedure (Church and Hanks, 1989; Pereira and Mendes, 2002).

3 Experiments on a Twitter corpus of Modern Greek

1M tweets (13.531.036 tokens, 253.230 Types & 1.160.036 sentences) were preprocessed to remove mentions, https links, hashtags and emoticons and were tagged and lemmatised with the ILSP tagger. In these experiments, we paid special attention to food language; a subcorpus was created by querying dish names such as πατάτες τηγανητές 'french fries', πιπεριά Φλωρίνης 'Florinis pepper'. These tweets formed a 10-15% of our tweet corpus. Throughout our study we tried to see whether this subcoprus behaves differently from the remaining tweet corpora but in no step we found some remarkable difference. Therefore, the results reported here concern the whole corpus, including the tweets related to food.

3.1 First experiment: the Baseline

We used the enriched syntactic patterns of Stamou et al. (2020) (Table 1) to obtain a baseline and manually checked the top 3000 lemmatized phrases per AM. Only 310 unique lemmatised phrases were judged as True positives out of 15,000 candidates. The AM order by decreasing reliability in Fleiss ϰ scores was: MLE (0.82), T-score (0.79), LL (0.70), Dice (0.61), PMI (0.50). When the amount of discovered VMWEs by each AM is considered, PMI scores first (107 VMWEs) followed by LL (106 VMWEs), MLE (74 VMWEs), T-score (71 VMWEs) and Dice (60 VMWEs). As in the case of fiction corpora, we observe that the T-score, LL and MLE sets overlap significantly (Figure 1).

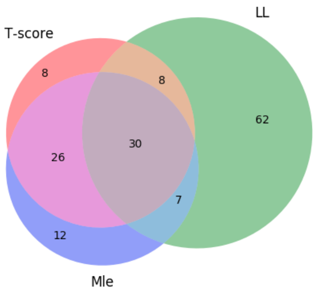


Figure 1: Intersection among T-score, LL and MLE True positives.

3.2 Second experiment: Leaner patterns

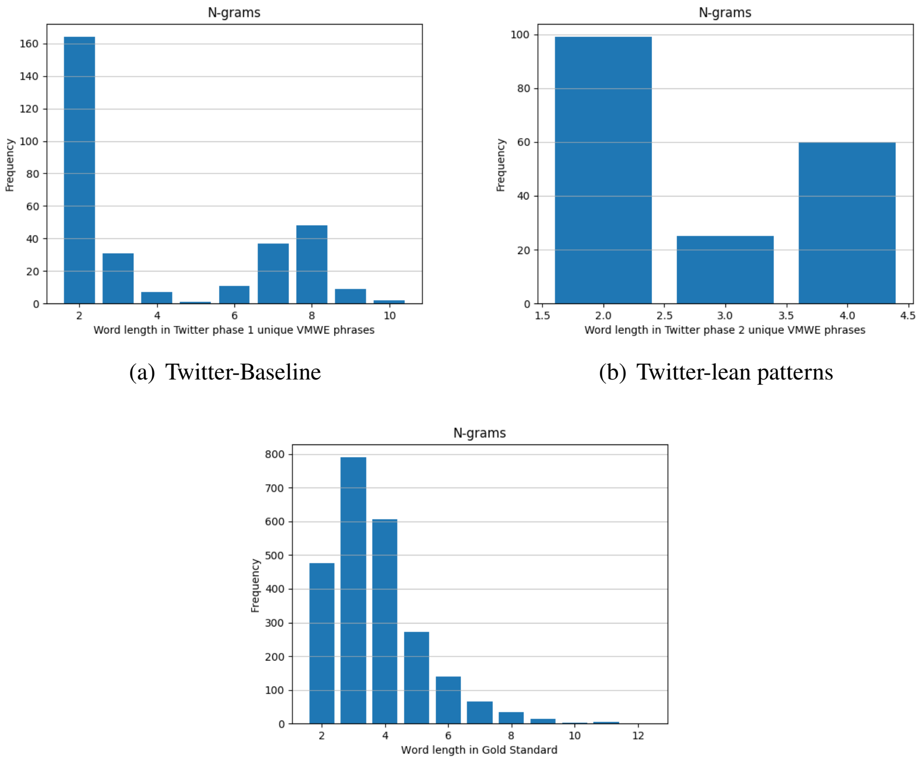
We used leaner patterns because the Twitter corpora returned VMWEs with few lexicalised components (Section 4). The leaner syntactic constructions contained at most four PoS tag sequences per pattern and were simplified versions of the rich patterns but not identical with the lean patterns used by Stamou et al. (2020). Again, the annotators checked the top 3000 phrases returned by each AM. Dice was found the most reliable AM and PMI the less reliable one. Dice returned 86 MWEs, T-score 83, MLE 72, LL 66 and PMI 60 (these results are an indication that PMI performs better with rich patterns). In total, 184 unique VMWEs were extracted. Again, the T-score, MLE and LL sets were found to share a considerable number of phrases (47 phrases), while Dice and PMI, the best and the worse scores respectively, had only one common phrase. The results of the two experiments (with rich and lean patterns) had only 25 phrases in common; the leaner patterns helped to discover 159 new phrases.

3.3 6000 candidates: Evaluation of the applied AMs

The Baseline returned results similar to those obtained from fiction corpora (first experiment): LL, MLE and T-score overlapped significantly but Dice and PMI did not intersect. These facts suggest that for VMWE discovery purposes the results of one of/the intersection of LL, MLE and T-score should be evaluated as well as the results of Dice and PMI. To test this idea, we evaluated more candidate phrases (+3000, total 6000 phrases per AM) because our Twitter corpora are twice in size as compared to the fiction corpora. The LL, MLE and T-score set received a high Fleiss ϰ value (0.79). In the additional 3000 phrases, Dice and PMI received low kappa values (ϰ=0.58 and ϰ=0.45 respectively) and returned 62 and 65 True positives respectively. Again it was observed that PMI retrieves low frequency VMWEs with more than two lexicalised components. If the 6000 candidates are considered, the set LL-MLE-T-score returned 137 True positives, PMI 65 + 107 = 172 and Dice 60+62=122. The total amount of identified VMWES was 431; the improvement is not impressive given the amount of annotation effort required. We estimate that we received per AM a 2% of True Positives (60 VMWEs) out of the 3000 candidates; this estimation illustrates the significant overlap among the LL, MLE and T-score (we obtained 137 MWEs when the "expected" ones were 180).

4 VMWEs in fiction and Twitter corpora

The plots of the number of lexicalised components in the VMWEs (Figure 2) reveal that VMWEs with two lexicalised components prevail in Twitter corpora and VMWEs with more than two lexicalised components in fiction corpora. This (not unexpected) fact may partly explain the results of our second experiment where lean patterns returned several new VMWEs.



(a) Twitter-Baseline

(b) Twitter-lean patterns

(c) Fiction (from Stamou et al.(2020))

Figure 2: Number of lexicalised components in the retrieved VMWEs.

The first two phrases below were extracted from the fiction corpora and the next two from the Twitter corpus:

(1) **μπαίνει** αμέσως **στο νόημα**
get.PRES.3SG immediately to the point.ACC
’getting the hang of it’

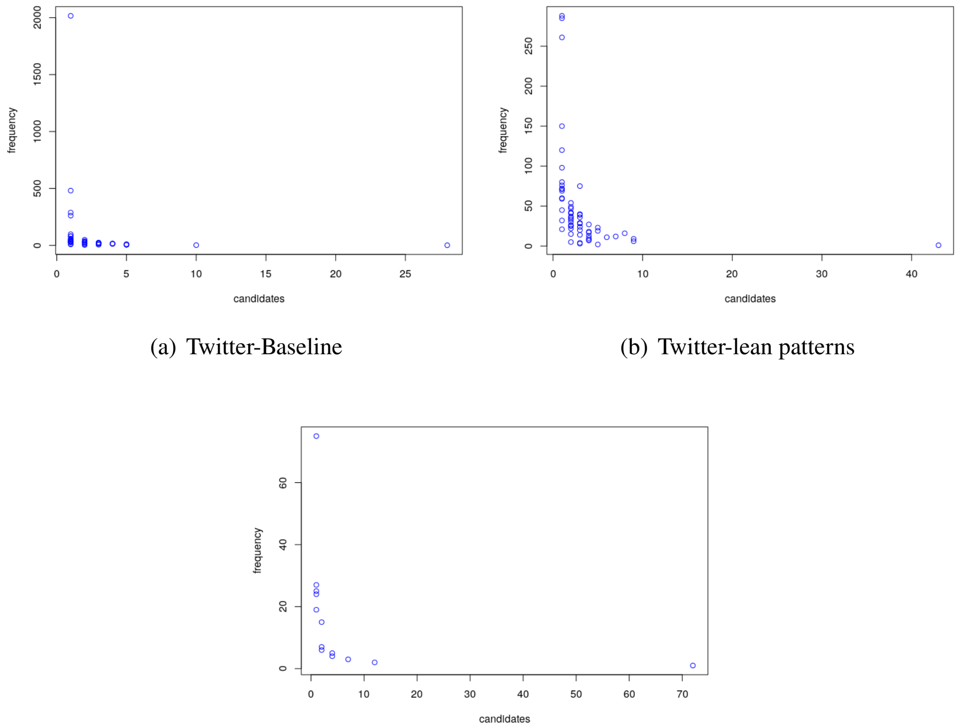
(2) **Έριξες άδεια για να πιάσεις γεμάτα**
throw.PAST.3SG. empty.ACC in order to catch.PAST.3SG. full.ACC
’to fish for information’

(3) **κατεβάζω ρολά**
put.down.PRES.1SG shutters.ACC
’to shutdown’

(4) **τρώω** ένα ωραίο **μπλοκ**
eat.PRES.1SG one.ACC nice.ACC block
’I was blocked in the social media’

Furthermore, the plots of the frequencies of the True positives per experiment with Twitter corpora (Figure 3), suggest that the VMWEs obtained with the leaner patterns were of higher frequency than the ones obtained with the Baseline (rich patterns). Again, this seems to be a reasonable result in the context of

Twitter corpora. At the same time it shows that patterns, frequency and number of lexicalised components of VMWEs may interact. Stamou et al. (2020) who also conducted separate experiments with lean and rich syntactic patterns on fiction corpora do not report a similar effect.



(a) Twitter-Baseline

(b) Twitter-lean patterns

(c) Twitter-6000 candidates

Figure 3: Frequencies of the retrieved VMWEs with the three experiments

5 Conclusions

Our experiments on Twitter corpora of Modern Greek, compared to experiments conducted by Stamou et al. (2020) on fiction corpora of this language, have shown that manual annotation of AM results better suits lexicographic purposes because it exploits more efficiently the output of PMI and Dice that tend to return hapax legomena and long VMWEs. Significant economy in evaluation effort can be achieved if the results of Dice and PMI are evaluated independently, because they hardly intersect, and only the intersection of LL, MLE and T-score results is evaluated. Furthermore, our experiments revealed that VMWEs with two lexicalised components prevail in Twitter corpora (but not in fiction corpora) and can be better identified with lean syntactic patterns rather than rich ones. This result could be of interest to lexicographers; it also indicates that there is an interaction among patterns used, number of lexicalised components and frequency of VMWEs in the corpus. We plan to accommodate the new VMWEs detected in the two corpora at the IDION database[2].

Acknowledgements

This work was supported by the project “GRE-Taste” (T1EDK-02015) that is funded by the action “RESEARCH-CREATE-INNOVATE” and co-financed by Greek National Funds and the European Union.

[2] The data can be reached at: http://idion.ilsp.gr/data.

References

Antton Gurrutxaga and Alegria, Iñaki. 2011. Automatic extraction of NV expressions in basque: Basic issueson cooccurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 2-7. Portland, Oregon, USA. Association for Computational Linguistics.

Brigitte Krenn and Stefan Evert. 2011. Can we do better than frequency? a case study on extracting PP-verb collocations. In *39 the Annua lMeeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39)* pages39–46,CNRS- Institut de Rechercheen Informatique de Toulouse, and Universite des Sciences Sociales,Toulouse,France,July.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Marian, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of LREC 2010*, Valetta, Malta. ELRA.

Draoud Draoud and Akram Alkooz, and Mohammad Daoud. 2015. Time-sensitive arabic multiword expressions extraction from social networks. *International Journal of Speech Technology* 19, 10.

Emilia Stripeli, Prokopis Prokopidis, and Haris Papageorgiou. 2020. Stella Markantonatou and Anastasia Christofidou (eds). 2020. Multiword expressions: Studies drawing on data from Modern Greek and other languages. Bulletin for Scientific Terminology and Neologisms (DEON), vol. 15. In *Academy of Athens*, pages 15(4):75-95, deltio epistimonikis orologias ke neologismon.

Evita Linardaki, Carlos Ramisch, Aline Villavicencio, and Angeliki/Aggeliki Fotopoulou. 2010. Towards the construction of language resources for greek multiword expressions: Extraction and evaluation. In Piperidis, S., Slavcheva, M., and Vertan, C., editors, *Proceedings of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May

Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A unified POS tagging architecture and its application to Greek. In *Proceedings of the Second International Conference on Languages Resources and Evalutation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

Hung Huu Hoang, Su Nam Kim, and Min-Yen Kan. 2009. A re-examination of lexical association measures. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, pages 31-39, Singapore. Association for Computational Linguistics.

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. *In 27th Annual Meeting of the Association for Computational Linguistics*, pages 76-83. Vancouver, British Columbia, Canada, June. Association for Computational Linguistics.

Luisa Alice Santos Pereira and Amalia Mendes. 2002. A electronic dictionary of collocations for european portuguese: Methodology, results and applications. In Anna Braasch and Claus Povlsen, editors, *Proceedings of the 10th EURALEX International Congress,* pages 841-849, København, Denmark, Center for Sprogteknologi.

Marcos Garcia, Marcos Garcia Salido, and Margarita Alonso-Ramos. 2019. A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the Joint Workshop on Multiword Expressions and Wordnet (MWE-WN 2019,* pages 49-59, Florence, Italy, August, Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit and Johanna Monti and Lonneke van der Plas and Carlos Ramisch and Michael Rosner and Amalia Todirascu. 2017. Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837-892.

Nikhil Londhe, Rohini Srihari, and Vishrawas Gopalakrishnan. 2016. Time-independent and language-independent extraction of multiword expressions from twitter. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2269-2278, Osaka, Japan, December. The COLING 2016 Organizing Commitee.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* pages 651-658, Sydney, Australia, July. Association for Computational Linguistics.

Stella Markantonatou, Panagiotis Minos, George Zakis, Vassiliki Moutzouri, Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. *In Proceedings of Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Workshop at ACL 2019,* Florence, Italy.

Stella Markantonatou, Calros Ramisch, Agata Savary, and Veronica Vincze. 2018. *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop.* Language Science Press.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing* pages 1-8, Barcelona, Spain, Association for Computational Linguistics.

Timothy Baldwin. 2005. Looking for prepositional verbs in corpus data. *In Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in computational linguistics formalisms and applications* Colchester, UK.

Vivian Stamou, Artemis Xylogianni, Marilena Malli, Penny Takorou, Stella Markantonatou. 2020. Evaluation of Verb Multiword Expressions discovery measurements in literature corpora of Modern Greek. *Proceedings of the Euralex XIX: Lexicography for Inclusion* Alexandroupolis, Greece (to appear).