EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki

Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Evaluation of Verb Multiword Expressions discovery measurements in literature corpora of Modern Greek

Stamou V.[1], Malli M.[2], Takorou P.[2], Xylogianni A.[2] , Markantonatou S.[1]

[1] *Institute for Language and Speech Processing, Greece*
[2] *Department of French Language Literature, Greece*

Abstract

We report on issues concerning the use of association measures and linguistic knowledge (Part-of-Speech sequences) with the environment MWETOOLKIT (Ramisch et al. 2010) for discovering all types of verb multiword expressions (VMWE) in corpora of Modern Greek (MG) literature. "MWE discovery" refers to detecting new MWEs in a corpus for lexicographic purposes (Constant et al. 2017). We are interested in boosting lexicographic work with (semi-)automatic facilities, in particular, the development of the VMWE database IDION (Markantonatou et al. 2019).

Keywords: verb multiword expression discovery; association measures; lexicography

1 Introduction

We discuss the behaviour of 5 association measures as regards the discovery of verb multiword expressions (VMWEs) on literature corpora of Modern Greek (MG). We used the environment MWETOOLKIT3 for this purpose. MWETOOLKIT was selected because (i) it can be applied and adapted to any language provided that lemma and pos tag information are available, (ii) the level of linguistic analysis can be specified by the user, (iii) it is a complete pipeline (includes also evaluation module), (iv) it contains lexical association measures, (v) it has been applied to several languages: English (Ramisch et al. 2010), French (Dubremetz & Nivre 2014), Brazilian Portuguese (Duran et al. 2011), Latvian and Lithuanian (Mandravickaite & Krilavičius 2017). The tool exploits either n-grams extracted from a corpus or morphosyntactic patterns defined using regular expressions in order to produce a list of candidate MWE phrases that are subsequently filtered into a significance list with the use of AMs (association measures).
MWETOOLKIT3 has already been used for the Modern Greek language to discover nominal MWEs (Linardaki et al. 2010) and light verb constructions (Stripeli et al. in prep). Both MWE types are typically bigrams. Neither approach discusses the behaviour of the AMs in relation to the grammatical patterns used probably because the short length of the particular MWEs reduces the number of possible patterns.
Some scepticism has been expressed in the literature concerning the appropriateness of AMs for discovering strings longer than bigrams (Constant et al. 2017). VMWEs quite often contain more than two words and, perhaps this is the reason why studies on the behavior of AMs as regards all types of VMWEs are sparse.
In Section 2 below we present a short review of the literature on the evaluation of the AMs with respect to MWE discovery and/or identification and then move to present our work: the corpus we developed and used is described in Section 3, Sections 4 and 5 report on two experiments and, finally, in Section 6 we summarize our conclusions.

2 A brief review of the relevant literature

Most of the literature related to the evaluation of the AMs used either for identification or for extraction points is about bigrams and attributes the observed differentiations to corpora diversity (size, type, etc.) or to the type and characteristics of the studied MWEs. This implies that no AM always achieves best performances and that a combination of AMs (Pecina and Schlesinger 2006) is required to guarantee best results. Furthermore, many studies emphasize the importance of manual validation as regards the exploration of the functionality of AMs in MWE discovery/identification tasks.
In Krenn and Evert (2001) AMs, namely *mutual information (MI), dice coefficient, x^2 measure, log-likelihood and t-score*, are utilized for detecting PP-verb collocations in the German language (newspaper and newsgroup corpus); *t-score* yielded the best precision score, while the remaining AMs yielded results below the baseline (random selection) and the simple co-occurrence frequency. Evert and Krenn (2001) evaluate lexical association measurements in German corpora against a gold set of manually detected cases. They worked with *Adj+Nouns* and *Preposition+Noun+Verb* sequences; the later sequences actually form triples, but AMs were implemented as P(reposition)N(oun),V(erb). The following AMs were used: *MI, LL, t-test and χ^2-test.* For *Adj+Nouns*, *log-likelihood* and *t-test* were observed to receive higher precision values on the sets of 100 and 500 candidates, while the *MI* measure obtained the lowest precision value. In addition, co-occurrence frequency, which was also examined outperformed both *χ^2-test* and *MI.* In the case of PN,V pairs, *t-score* and *frequency* were better than *LL*. Next, they considered different frequency thresholds (low-high values) and observed that *t-score* achieves the best results in high frequency data for the PN,V pairs, while for the low frequency data *LL* was found to get higher values. Summarizing the findings of this study, it was shown that the AMs' functionality in filtering phrases highly depends on

the MWE type, and the role of word frequency in the resulting candidate list was highlighted. English prepositional verbs (e.g. '*come across'*) are discussed by Baldwin (2005), in terms of extraction and methodology evaluation. Among the proposed methods for accepting a verb preposition sequence as a PV or not, along with linguistic tests, statistical criteria were used as well as *the co-occurrence frequency, the dice coefficient, the PMI, χ^2 and the LL*. The AMs were tested on Brown corpus, Wall Street Journal and BNC; the *dice* coefficient performed best while the other statistical criteria made better guesses than selection based on pure frequency.

Hoang et al. (2009) attempt to group 82 AMs discussed in Pecina & Schlesinger (2006) into two classes in order to better understand and justify similarities or differences among the ranking scores and guess which AMs fit better to specific MWE classes. Class I (e.g. *PMI, t-score*) was defined by considering the tendency of a phrase to form a semantic unit rather than being a random word combination, and Class II (e.g. *entropy, distance metrics*) by taking into account the context and therefore non-compositionality. The MWE examined were Verb particle constructions (only bigrams) and Light verb constructions extracted from the Wall Street Journal corpus. Ranking similarity (or 'rank equivalence') on the candidate lists allowed to characterize groups of AMs based on their average precision (AP) values. Interestingly, *MI* and *LL* were found to be included into the same group, while in total 5 groups of AMs were detected. Subsequently, the authors proposed a replacement of more complex AMs by simpler ones in the case they return the same AP scores. As regards the MWE type examined, they concluded that AMs belonging to Class II were not appropriate for identifying both MWE types due to the small corpus size.

In Antunes and Mendes (2014), MWEs of different types (14,000) belonging to the COMBINA-PT lexicon generated from a subcorpus of Contemporary Portuguese are used to extract example cases (n-grams:2-5). The example cases were subsequently selected and sorted according to their *MI* values. The authors focused on candidate MWE phrases with *MI* values around 8 and 10, given the previous research findings (Pereira & Mendes 2002). They checked the lists manually considering several criteria (both linguistic and quantitative). Interestingly, often phrases (n-grams) accepted in terms of their linguistic properties received very low *MI* values. *MI* was compared with other AMs such as *t-test* and *LL* with respect to their ranking preferences. Differentiations were observed for the cases ranked in the middle but not for the cases ranked high or low. In general, *LL* was found to be more similar to *MI*.

More recently, Garcia et al. (2019) evaluated twelve AMs (including *raw frequency*) on three types of dependency-based collocations, namely *adjective noun, verb object and nominal compounds* pairs for English, Portuguese and Spanish. The following AMs were included in the study: *ll, t-score, z-score, MI(PMI), MI2, Dice, log-likelihood,* and χ^2. Precision and Recall values were computed against gold corpora annotated with 1,394 unique collocations. *Frequency* and *t-score* achieved the best performance, while *PMI* obtained the lowest values. Similar average results were obtained for the three languages. Although *frequency* was found to be amongst the best measures for collocation extraction, the authors discuss cases missed due to their low frequency of occurrence and conclude that there is a need of "applying specific AMs for different relations and frequency folds" (Garcia 2019: 56).

3 The Corpus

The corpus we developed for our experiments consists of five novels (Table 1), all of them containing contemporary colloquial language rich in VMWEs. The scanned text was corrected manually and tagged/lemmatized with the ILSP tools (Papageorgiou et al. 2000).

Subcorpora	Tokens	Lemmas	Number of sentences	Number of MWEs
Maratos (2007)	84,172	5,937	5,931	1,136
Markaris (1995)	105,575	5,916	8,457	1,877
Markaris (2016)	84,172	4,86	5,931	1,175
Papadaki(2001)	82, 084	6,281	10,311	912
Tachtsis (1970)	104,215	6,701	6,700	1,554

Table 1: The five subcorpora.

Each novel in the corpus was annotated for VMWEs. The annotators read the texts and listed the text extracts they considered as instances of verbal MWEs (VMWEs), in other words, the annotators performed manual VMWE identification. The lists derived from each subcorpus were annotated by a third expert. Interannotator agreement (IA) between the corpus annotators and the third annotator was calculated with the Fleiss kappa coefficient (Table 2). A Golden Standard (GS) of about 3500 VMWEs corresponding to ~2400 (lemmatized) types was formed that contains the text extracts that were considered MWEs by both the annotators; crucially, GS does not necessarily contain all the VMWEs in the corpus.

Literature Texts	κ
Μαράτος	0.94

Μάρκαρης (Offshore)	0.87
Μάρκαρης (Νυχτερινό Δελτίο)	0.94
Παπαδάκη	0.77
Ταχτσής	0.97

Table 2: Interannotation agreement (Fleiss κ) for each subcorpus.

The annotators used the following criteria to identify VMWEs:

- Native speakers' knowledge of the language.
- The construction does not demonstrate an established sense of the verb head.
- Frequent occurrence of an expression in the corpus.
- Frequent use of the construction in the web.
- Distinguishing between MWEs and literal occurrences, which is always a matter of context (Savary et al. 2019).
- Whether the idiomatic meaning of a potential VMWE was preserved when the construction was translated literally in English (Constant et al. 2017; Salehi et al. 2018).

Figure 1 shows the distribution of VMWE lengths in the corpus. Most MWEs contain more than two fixed words. This is a potential problem for the AMs used in MWE detection that normally perform on bigrams. Actually, a large part of the literature on VMWE detection is about particle verbs and light verb constructions that are naturally modelled as bigrams (Stevenson et al. 2004). Instead, we try to identify the best method of discovering MG VMWEs of any length.

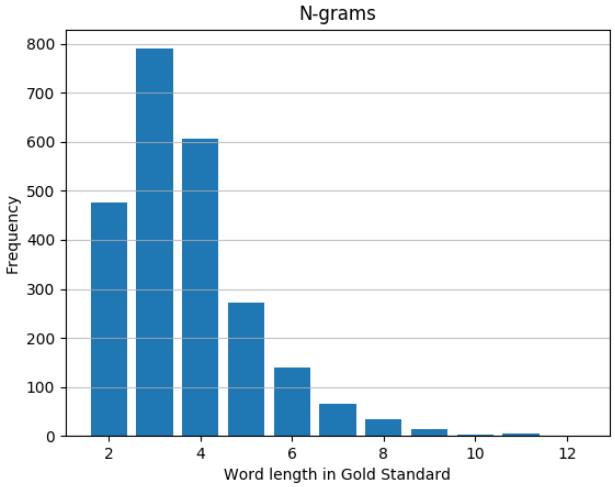

Figure 1: Distribution of VMWE lengths in terms of words in the GS.

4 The First Experiment

We have already mentioned that MWETOOLKIT3 (Ramisch et al. 2010) allows for both n-grams and syntactic patterns. We defined 6 syntactic patterns, shown below as Parole (Labropoulou et al. 1996) tag sequences (Table 3), aiming at modelling a large portion of the VMWE phrase types in the corpus. We also used flat ngrams (nmin:2grams-nmax:5grams).

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

Table 3: VMWE patterns (first experiment).

In Table 3, brackets indicate optionality. The first pattern practically captures verb plus noun sequences, the second pattern subordination and the next two patterns noun plus verb sequences and the next one pronoun plus verb sequences. The last pattern captures verb plus prepositional phrase sequences. Several tenses in Modern Greek are formed periphrastically with auxiliary verbs and this fact is captured with the (Vb)+Vb sequences. Pronouns at the beginning of the strings capture the

very frequent use of dative genitives.

Candidate strings were filtered with the following AMs: *dice coefficient*, *log likelihood*, *relative frequency (mle), PMI (pointwise mutual information)* and *t-score*. We report on the manual and automatic evaluation of a list of candidate VMWEs composed of the top 3000 highest ranking candidates proposed by each AM, in total 15,000 candidates; in this list, all the expression tokens sharing the same lemma form were merged under a single entry (the lemma).
We followed the same manual evaluation method as with the corpus annotation. The ranking of the AM scores by the annotators in terms of decreasing reliability is shown in Table 4.

Association Measures	**κ**
Dice	0.72
Log likelihood	0.69
Mle	0.67
T-score	0.60
Pmi	0.53

Table 4: AM scores ranked (by Fleiss κ).

Figure 2 shows the intersection among the sets of the top best ranking 3000 expressions returned by *ll*, *mle* and *t-score*: *mle* and *t-score* overlap significantly, a phenomenon also observed in Linardaki et al. (2010).

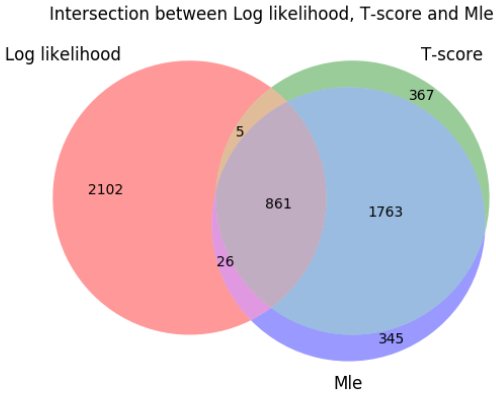

Figure 2: Intersection among the sets returned by *log-likelihood, mle* and *t-score*.

Qualitatively speaking, *dice* and *pmi* returned VMWEs not found by the other AMs. The plots of the frequency values for the top 3000 phrases returned by *pmi* and *dice* (Figures 3 & 4) show that most of the retrieved phrases were hapax legomena.

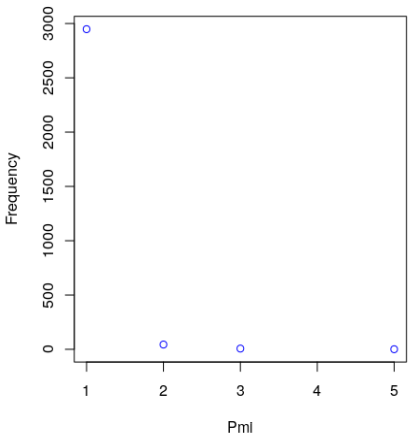

Figure 3: Histogram for the top 300 candidates of *pmi*.

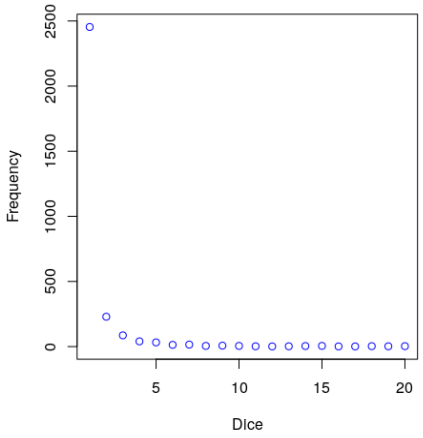

Figure 4: Histogram for the top 3000 candidates of *dice*.

Despite the annotators' agreement on *dice*'s superiority, the automatic comparison of the AM best scores against the GS returned the following decreasing list of predictions: *t-score* (302), *mle* (282), *log likelihood* (221), *dice (*172) and *pmi* (42). The Precision score with GS was 2%. Precision P is defined as P= TP/TP+FP where TP=true positives, FP=false positives. We also used n-grams in order to have a back-off mechanism for cases that might not be captured by the 6 patterns. We ran each n-gram (2:5) separately and evaluated the output against the GS. The best guesses were made by the *t-score* for 2grams (44 phrases) and 4grams (71 phrases). In the second experiment we did not use n-grams.
We employed a sequence matching technique using the python class *SequenceMatcher* to compare strings in the GS with the 829 manually identified "True Positives" (TP), i.e., the lemmatized VMWEs in the AMs output that were shared by all the annotators who annotated each set of results.
We applied additional computations on the results because of the limitations of the lemmatizer, for instance it misses verbal types with apostrophes (example 1) and because of the non-fixed words in a VMWE such as the possessive pronouns that do not turn up in the lemmatized version of the VMWE (example 2); e.g.

(1) του *΄ψηνε* το ψάρι στα χείλη

Lit. to.him roasted.3rd the fish on.the lips

'he made his life extremely difficult'

(2) χάνω το χρώμα *μου*

Lit. lose.1st the colour mine

'I become pale'.

We checked manually the common phrases with a ratio above 0.8 and identified 408 phrases. Therefore, an enriched GS2 was created from GS with the addition of 421 new VMWEs (TPs were 829 in total). GS2 contained ~3000 lemmatized phrases.
An automatic evaluation of the AM results against GS2 returned an improved Precision score of 4,76%. Recall values are not reported since we do not know the precise size of the VMWE population in our corpus.

5 The Second Experiment

In the hope of improving the results of the fist experiment, we increased the amount of linguistic knowledge and adjusted the patterns to include adverbs, adjectives, double prepositional phrases and more complex conjunction patterns. The enriched patterns are shown in Table 5 where the boldfaced PoS indicate the additions to the patterns of Table 3. We followed the same evaluation method.

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)

Table 5: VMWEs patterns (second experiment).

As regards the manual evaluation, we computed IA agreement on the phrases proposed by each score. The ranking of the AMs by reduced reliability is: *dice*, *mle, pmi, t-score*, *log likelihood,* while the Fleiss κ value ranges from 0.72 to 0.86. The automatic evaluation against the GS2 resulted in a different order, as in the first experiment: *t-score* (22,9%), *mle* (21%), *ll* (12%), *dice* (7,6%) and *pmi* (0,4%).
A similar disagreement between manual and automatic evaluation is reported by Linardaki et al. (2010) for nominal MG MWES and Gurrutxaga et al. (2011) for Basque VMWEs. This differentiation could perhaps be attributed to the fact that native speakers can easily detect hapax legomena in the outputs of the AMs and promote AMs sensitive to hapax legomena, while the automatic evaluation relies on the contents of the definitely incomplete GS; therefore, AMs that bring more frequent VMWEs fare better.
True Positives for the new AM results were 1455, of which 619 in the GS2. GS3, containing ~4000 VMWE lemmata was obtained with the addition of 836 new phrases to GS2 and returned P=5,2%.

6 Conclusions

This work aimed at evaluating the contribution of AMs to VMWE discovery for lexicographic purposes, therefore our conclusions refer to the discovery of new VMWEs, which often contain more than two words.
So, if the lexicographic goal is the discovery of new VMWEs, our experiments have shown that as far as literature corpora of Modern Greek are concerned, human annotation is indispensable. However, the combination of linguistic knowledge with AMs can significantly improve the results received with human annotation: in our experiments human annotation returned 2400 VMWEs and the application of patterns and AMs increased their number to 4000 but only when the output of AMs was evaluated by human experts. Here, we underline the fact that automatic evaluations against a GS failed to discover less frequently used VMWEs as was discussed in Section 2.
Furthermore, patterns rich in linguistic knowledge have produced better results than leaner patterns or n-grams: experiment two used richer phrasal patterns than experiment one and returned about 26% more VMWEs.
As regards the AMs, our experiments indicated that not all the AM outputs have to be evaluated because there are significant overlaps. So, evaluation of the results of *pmi*, *dice*, *t-score* and *ll* by the annotators is necessary in order to identify interesting VMWEs because:

1. It has been observed that the results of *mle* and *dice* overlap significantly (Figure 2); therefore, only the results of one of the two AMs could be evaluated. We recommend the evaluation of the results returned by *dice,* because it fared better in both the manual evaluations (as we have already mentioned, manual evaluation of the AM results is strongly recommended for VMWE discovery).
2. *Pmi* and *dice* return VMWEs that are rare in corpora on which the AMs have been applied, as can be observed in Figures 3 & 4 (and has also been reported in literature, Church & Hanks 1990; Pereira and Mendes 2002).

Overall, if human effort has to be saved, the results of *dice* and *pmi* should be evaluated as *dice* scored always first in the IA agreement lists and *pmi* returned a greater number of rare VMWEs than any other AM.

As regards the automatic evaluation, *t-score* was found to perform best with Modern Greek VMWEs; similar findings have been reported for other languages as well (Linardaki et al. 2010; Gurrutxaga & Alegria 2011 among others). The role of the GS in the automatic evaluation can only be strongly emphasized. However, the development of an adequate GS is not an easy issue as it has been widely acknowledged and was shown by this work as well.

7 References in Greek

[Maratos (2007)] Μαράτος, Τ. (2007). Οι τυφώνες ήταν γένους θηλυκού. Βιβλιοπωλείον της "Εστίας", Ι.Δ. ΚΟΛΛΑΡΟΥ & ΣΙΑΣ.
[Markaris (1995)] Μάρκαρης, Π. (1995). Νυχτερινό Δελτίο, ΓΑΒΡΙΗΛΙΔΗΣ.
[Markaris (2016)] Μάρκαρης, Π. (2016). Offshore. ΓΑΒΡΙΗΛΙΔΗΣ.
[Papadaki (2001)] Παπαδάκη, Α. (2001). Βαρκάρισσα της χίμαιρας. Καλέντης.

[Tachtsis (1970)] Ταχτσής, Κ. (1970). Το τρίτο στεφάνι. ΕΡΜΗΣ.

8 References in English

Antunes, S. & Mendes, A. (2014). An evaluation of the role of statistical measures and frequency for MWE identification. I*n Proceedings of The Ninth International Conference on Language Resources and Evaluation–LREC 2014*, 26-31 May 2014. Reykjavik, Iceland, pp. 4046–4051.

Baldwin, T. (2005). Looking for prepositional verbs in corpus data. I*n Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in computational linguistics formalisms and applications, Colchester*, UK, pp. 180-189.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics,* 16(1), pp. 22–29.

Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. & Todirascu, A. (2017). Survey: Multiword expression processing: A Survey. In *Computational Linguistics*, 43(4), pp. 837-892.

Dubremetz, M. & Nivre, J. (2014). Extraction of nominal multiword expressions in French. *In Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, April. Association for Computational Linguistics. Gothenburg, Sweden, pp. 72–76.

Duran, M., Ramisch, C., Aluisio, S. & Villavicencio, A. (2011). Identifying and Analyzing Brazilian Portuguese Complex Predicates. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE2011)*. Association for Computational Linguistics, 23 June 2011. Portland, Oregon, USA, pp. 74-82.

Evert, S. & Krenn, B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics.* Toulouse, France, pp. 188-195.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (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 (MWEWN 2019)*. Association for Computational Linguistics, Florence, pp. 49-59.

Gurrutxaga, A. & Alegria, I. (2011). Automatic extraction of NV expressions in Basque: Basic issues on cooccurence techniques. In *Proceedings of the workshop on multiword expressions: from parsing and generation to the real word*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 2-7.

Hoang, H. H., Su N. K. & Kan M.-Y. 2009. A re-examination of lexical association measures. *In Proceedings of the Workshop on Multiword Expressions*, Singapore, pp. 31–39.

Krenn, B. & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. *In Proceedings of the ACL Workshop on Collocations,* Toulouse, France, pp. 39-46.

Labropoulou, P., Mantzari, E. & Gavrilidou, M. (1996). *Lexicon-morphosyntactic specifications: Language specific instantiation (Greek)*. PP-PAROLE, MLAP report, pp. 63–386.

Linardaki, P., Ramisch C., Villavicencio, A. & Fotopoulou, A. (2010). Towards the construction of language resources for greek multiword expressions: Extraction and evaluation. In *Proceedings of the international conference on language resources and evaluatio*n, May. Valetta, Malta, pp. 31-40.

Mandravickaite, J. & Krilavičius, T. (2017). Identification of multiword expressions for Latvian and Lithuanian: hybrid approach. *Proceedings of the 13th Workshop on Multiword Expressions (MWE '17)*, Valencia, Spain, Association for Computational Linguistics, Stroudsburg, PA, pp. 97-101.

Markantonatou, S., Minos, P., Zakis, G., Moutzouri, V., & Chantou, M. (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. Toulouse, France: Association for Computational Linguistics, pp.130-134.

Papageorgiou, H., Prokopidis, P., Giouli, V. & Piperidis, S. (2000). A unified POS tagging architecture and its application to Greek. *In Proceedings of the second international conference on language resources and evaluation (LREC'00), May 2000.* Athens, Greece: European Language Resources Association (ELRA), pp. 1455-1462.

Pecina, P., & Schlesinger, P. (2006) Combining association measures for collocation extraction. In *Proceedings of the 21th international conference on computational linguistics and 44th annual meeting of the association for computational linguistics (COLING/ACL 2006)*. Sydney, Australia, pp. 651–658.

Pereira, L. & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. *In Proceedings of the 10th International Congress of the European Association for Lexicography. Copenhagen,* Denmark, vol. II, pp. 841-849.

Ramisch, C., Villavicencio, A. & Boitet, C. (2010). mwetoolkit: a Framework for Multiword Expression Identification. In N. Calzolari et al. (Eds.), *Proceedings of LREC 2010.* Valetta, Malta: ELRA, pp. 662–669.

Salehi, B., Cook, P. & Baldwin, T. (2018). Exploiting multilingual lexical resources to predict MWE compositionality. In Stella Markantonatou, Carlos Ramisch, Agata Savary and Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, Berlin: Language Science Press, pp. 343–373.

Savary, A., Cordeiro, S.R., Lichte, T., Ramisch, C., Iñurrieta, U. & Giouli, V. (2019). Literal Occurences of Multiword Expressions: Rare Birds That Cause a Stir. In *The Prague Bulletin of Mathematical Linguistics*, 112, pp. 5-54.

Stevenson, S., Fazly, A. & North, R. (2004). Statistical measures of semi-productivity of light verb constructions. In *Proceedings of the workshop on multiword expressions: Integrating processing*. Barcelona, Spain: Association for Computational Linguistics, *pp. 1-8.*

Stripeli, E., Prokopidis, P. & Papageorgiou, H. (in prep.). Multiword expressions in Greek, deltio epistimonikis orologias ke neologismon. *Academy of Athens*, 15(4), pp. 75-95.