Discovery of Multiword Expressions with Loanwords and Their Equivalents in the Persian Language

Katarzyna Marszałek-Kowalewska
YUKKA Lab AG / Berlin, Germany
k.marszalek.kowalewska@gmail.com

Abstract

This paper presents an attempt at multiword expressions (MWEs) discovery in the Persian language. It focuses on extracting MWEs containing lemmas of a particular group: loanwords in Persian and their equivalents proposed by the Academy of Persian Language and Literature. In order to discover such MWEs, four association measures (AMs) are used and evaluated. Finally, the list of extracted MWEs is analyzed, and a comparison between expressions with loanwords and equivalents is presented. To our knowledge, this is the first time such analysis was provided for the Persian language.

1 Introduction

Today, almost 19 years after the seminal paper "Multiword Expressions - A Pain in the Neck for NLP" by Sag et al. (2002), multiword expressions (MWEs) are still an interesting and challenging aspect of many Natural Language Processing (NLP) tasks, which is reflected in the number of papers addressing this phenomenon as well as the number of people contributing to, and attending workshops, conferences, and initiatives such as SIGLEX-MWE[1] or PARSEME.[2] MWEs are very frequent in language and range over a number of different linguistic constructions, from idioms, e.g., *to pay an arm and a leg*, to fixed expressions, e.g., *rock and roll*, light verb constructions, e.g., *take a shower*, to noun compounds, e.g., *golf club*. Biber et al. (1999) claim that the number of MWEs in spoken English is 30% – 45% and 21% in academic prose. Jackendoff (1997) suggests that the number of MWEs in a speaker's lexicon is the same as simple words. Nevertheless, if we take into consideration the domain-specific lexicons, this number seems to be an underestimation (Sag et al., 2002). Indeed, the research conducted by Ramisch (2009) suggests that the MWEs ratio can be between 50% and 80% in a corpus of scientific biomedical abstracts. Research by Krieger and Finatto (2004) estimate that MWEs can constitute more than 70% of the specialized lexicon.

MWEs have received considerable attention in recent years and it has been suggested (Sag et al., 2002) that their proper treatment could make a significant improvements in a number of NLP tasks, e.g., lexicography (Church and Hanks, 1990; Gantar et al., 2018; Fellbaum, 2016), word sense disambiguation (Finlayson and Kulkarni, 2011), part-of-speech tagging and parsing (Baldwin et al., 2004), information retrieval (Newman et al., 2012), language learning (Christiansen and Arnon, 2017), machine translation (Carpuat and Diab, 2010) or sentiment analysis (Berend, 2011; Williams et al., 2015).

The research on MWEs in Persian has so far focused mainly on verbal multiword units and light verb constructions (LVCs) in particular. Taslimipoor et al. (2012) adopted a method originally proposed by Fazly et al. (2007) for identifying LVCs in Persian. They extended existing statistical measures of the acceptability of English LVCs, used semantic classes of nouns, and proved that semantic class information is useful for LVC acceptability of new combinations. Salehi et al. (2012) used bilingual parallel corpus (Persian-English) and investigated the usefulness of several linguistically-informed features for automatic identification of Persian LVCs. Persian (among 18 other languages) and the analysis of its verbal MWEs have also been addressed as part of the PARSEME shared task on automatic identification of verbal MWEs (Savary et al., 2017). The best system for Persian in the task obtained an outstanding F-score, which exceeds 0.9. The reasons for such a high F-score can be

[1] http://https://multiword.org

[2] https://typo.uni-konstanz.de/parseme/

Proceedings of Recent Advances in Natural Language Processing, pages 918–928
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_105

perceived in two factors: the density of light verbs is exceptionally high in Persian, and the information about LVCs was contained in morphological companion files. Salehi et al. (2016), on the other hand, do not focus on any particular MWE type but rather try to cover the whole spectrum of MWE types. Their model is trained on a treebank with MWE relations of a source language applied to a corpus of a surprise language to identify its MWE construction types.

This paper presents an evaluation of four association measures used for the extraction of Persian multiword expressions with loanwords and their native counterparts. *Farhangestan-e zaban va adab-e farsi* ('Academy of Persian Language and Literature') is an official body responsible for the Persian language, its resources, and reforms. One of the tasks of the Academy is to propose Persian equivalents for borrowed terms. So far, the Academy has been successful in issuing thirteen lists of "Collection of Terms Approved". These are, as the name suggests, terms that the speakers of Persian should use. The total number of approved terms is more than 45,000. They are also available on Academy's website (Dabir-Moghaddam, 2018). This study aims at 1) applying AMs to extract MWEs and 2) comparing ten loanwords and their equivalents to evaluate their potential to form MWEs.

The rest of this paper is organized as follows. Section 2 presents information on work related to association measures used for MWEs discovery. Methodology is presented in Section 3. It describes the corpus used in this study, lemmas selected as initial seeds, and the four association measures. The results and their evaluation are addressed in Section 4. Finally, the conclusion and plans for future work are presented in Section 5.

2 Related Work

The assumption that MWEs stand out, i.e., they exhibit some sort of sailence, allows to extract (or discover) them automatically from texts. This salience is also the reason why especially statistical measures have been so popular when it comes to the discovery of multiword expressions.

Many studies indicate that words that tend to co-occur more frequently than by a pure chance are good candidates for MWEs and propose detecting this statistical significance by measuring the association strength between these words (Manning and Schütze, 1999; Pecina and Schlesinger, 2006; Constant et al., 2017). Such statistical metrics that can estimate the relationship strength between words in a corpus, based on these words' co-occurrence count and their individual word counts, are known as association measures (AMs). Since MWEs are characterized by strong collocational behavior, statistical association measures have been widely used in MWEs discovery. The number of proposed association measures over the years has been impressive. More than 30 AMs were described by Evert (2005), Pecina (2008) presented a list with over 80 and new measures as well as their variants are constantly being proposed (Evert, 2008a). However, although numerous studies propose and experiment with association measures performance, there is no consensus on which metric is best for extracting MWEs. Evert (2008a) mentioned that although some measures are more popular and have become standards (e.g., pointwise mutual information, log-likelihood, or t-score), the choice of a suitable metric depends on the particular task as each measure focuses on a different aspect of collocation strength. Since different measures capture various aspects of MWEs, Pecina and Schlesinger (2006) proposed combining some of them and showed that when in combination, AMs can generate better results for MWE discovery than if used in isolation.

The most widely used association measure for MWE discovery is the pointwise mutual information (PMI) proposed by Church and Hanks (1990) for terminology discovery. It is derived for bigrams directly from the mutual information between two random variables, using the log-ratio between the observed co-occurrences of the sequence and the individual words to determine how much the co-occurrence is due to mutual preference. The reported issue with PMI is that it is biased towards infrequent events (Ramisch and Villavicencio, 2018; Villavicencio and Idiart, 2019). Therefore, as observed by Bouma (2009) a moderately associated low-frequency bigram might obtain a better score than a highly associated high-frequency bigram.

Another popular group of AMs used for MWE discovery is based on hypothesis testing. Assuming the null hypothesis that words are independent, their observed and expected counts should be the same. Large values indicate that the candidate words are not independent and can potentially form a MWE. Examples of hypothesis-based AMs are t-score and z-score. They are both based on the assumption of normal distribution, and they work well for frequent events. However, their usage is not recommended for low-frequency pairs. The z-score test is also not suited for small corpora (Seretan, 2008).

AMs based on contingency tables record the marginal frequencies of the words in an n-gram and the probability of their non-co-occurrence. One such measure is Pearson's chi-squared test (χ^2) which overcomes the normal distribution problem as it makes no data assumptions. However, χ^2 is again not recommended for small corpora (Manning and Schütze, 1999), and it also tends to prefer common events (Kilgarriff, 1996). Another example is log-likelihood ratio (Dunning, 1993) - a well-known association measure for collocation extraction. It performs well with both frequent and rare events as well as different corpora sizes (Dunning, 1993). However, its reliability is affected by low values of expected frequencies in the contingency table (Pedersen, 1996).

Although AMs have a long history and their utility have been sometimes questioned (e.g., Stubbs, 2002), they are still sucessfully used in extraction systems, e.g., Evert et al. (2017), Uhrig et al. (2018), Garcia et al. (2019). They also remain an important part of other approaches to MWEs discovery, e.g., Squillante (2014), Tsvetkov and Wintner (2014) or Farahmand and Henderson (2016).

3 Multiword Expressions Discovery Methods

3.1 Definition

The definition adopted in this paper is the one presented by Baldwin and Kim (2010) (following Sag et al., 2002): "Multiword expressions (MWEs) are lexical items that: a) can be decomposed into multiple lexemes and b) display lexical, syntactic, pragmatic and/or statistical idiomaticity." It is one of the most frequently used definitions of MWEs, and it describes the phenomenon this paper focuses on, i.e., multiword constructions displaying some sort of idiomaticity.

3.2 Corpus

The corpus used in the study was sampled from MirasText (Sabeti et al., 2018) corpus - an automatically generated text corpus for Persian. It is one of the largest available Persian corpora, containing 2.8 million documents and over 1.4 billion tokens. The corpus size is 15GB. Each data point is provided with the following information: content, title, content summary and keywords, base website, and exact URL of the webpage.

The content of the MirasText corpus was generated from 250 web pages selected from a wide range of fields to ensure the diversity of data, e.g., news, economy, technology, sport, entertainment, or science.

Since the corpus data was obtained via crawling, it seemed necessary to perform certain cleaning and normalization tasks. Articles containing clipped content were excluded from the final corpus used in this study. The whole corpus data was normalized with Parsivar (Mohtaj et al., 2018) - a tool for processing the Persian language. These steps led to obtaining the final corpus of 50 million tokens, which was used to discover multiword expressions.

3.3 Lemmas

In order to discover Persian multiword expressions with loanwords and their equivalents proposed by the Academy of Persian Language and Literature, a list of 10 pairs (loanword-equivalent pair) was prepared.[3] There were two conditions for choosing these particular lemmas:

1. The Persian lemma is officially proposed as an alternative to the loanword by the Academy of Persian Language and Literature.

2. Lemmas should be part of everyday language, thus belong to general discourse.

Table 1 presents all 20 lemmas (both loanwords and their Persian equivalents) that served as initial seeds to discover MWEs. This table contains the following information: 1) meaning of a lemma, 2) its type, 3) information about lemma's ambiguity, e.g., lemma ماشین (māšin) apart from *machine*, can also mean *engine* or *motor*;[4] 4) information about other possible spelling variations of a lemma, e.g.,

[3] The motivation behind targeting MWEs with loanwords lies in the language policy in Iran, which actively proposed native Persian equivalents for borrowed elements. For more details on Iranian language policy see, e.g., Marszałek-Kowalewska (2011) or Moghaddam and Moezzipour (2017).

[4] Information about other possible meanings come from the following dictionaries:

- online dictionary including a number of Persian monolingual dictionaries (https://www.vajehyab.com)
- online dictionary and thesaurus *Abadis* (https://dictionary.abadis.ir)
- online Persian glossary based on dictionary of Dehkhoda (https://www.parsi.wiki)

فنّاوری (fannāvari) and فن آوری (fanāvari) for the word *technology*, 5) lemmas' raw frequency in a 50 million token corpus, and 6) information about the date the equivalent was proposed by the Academy of Persian Language and Literature.

3.4 Association Measures

In order to extract MWE candidates with loanwords and their equivalents, statistical association measures were used. For every lemma, its bi-grams and tri-grams were extracted from 50 million token corpus using the following association methods:[5]

- PMI
- log-likelihood
- t-score
- χ^2 test

These particular AMs were chosen as they are the most popular ones used for the discovery of MWEs (Evert, 2008a; Seretan, 2008; Wahl and Gries, 2018; Villavicencio and Idiart, 2019).

For each association measure, its top 100 bi- and tri-grams per lemma were extracted. This resulted in 1487 unique MWE candidates.

4 Results

4.1 Candidates Filtering

Since association measures produce ranked lists of MWE candidates, their evaluation is usually done through gold standard corpus or manual validation by trained experts.

The outcome of employing AMs to discover Persian MWEs with loanwords and equivalents is a list with 1487 unique MWE candidates. In order to evaluate individual AM performance, all candidates were assessed by external annotators from a crowdsourcing platform. These annotators were linguistically trained native speakers of Persian. The total number of workers contributing to this project was 18, and the inter-annotator agreement (IAA) was calculated with *Fleiss' Kappa* - a statistical measure used to evaluate the agreement between three or more raters (Fleiss, 1971).

To ensure the highest quality of annotators' work, the main part of MWE candidates filtering task was preceded by a trial run on a small gold test set. The IAA on this gold test set was 82%. All contributors' performance on the gold test set was taken into account, and only annotators with the best performance were invited to perform the main task. Therefore, the final IAA was 87% which indicate that *almost perfect agreement* was achieved.[6]

Annotators were provided detailed guidelines, which included an operational definition of MWEs (as presented in 3.1) and several examples showing true and false MWEs. Each MWE candidate was evaluated by at least three annotators who answered the question: Is the following sequence a valid multiword expression? Possible answers include: YES, NO, and UNABLE TO DETERMINE.[7]

4.2 Association Measures Evaluation

One of the two objectives of this study was to apply and evaluate association measures used to discover MWEs for ten loanwords in Persian and their Persian equivalents proposed by the Academy of Persian Language and Literature. The outcome of applying AMs is a list of MWE candidates. Out of 1487 MWE candidates, 389 turned out to be true MWEs. Figure 1 shows the performance of the four selected association measures when it comes to the discovery of true MWEs.

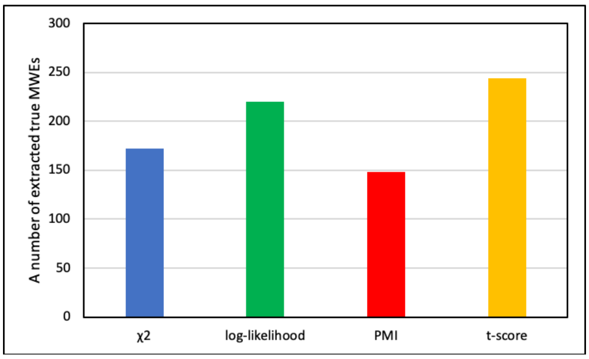

Figure 1: A number of true MWEs extracted via particular association measures.

As can be seen, the highest number of MWEs were extracted with t-score (248 MWEs), followed closely by log-likelihood (220). Surprisingly, the popular PMI method obtained the worst results, extracting only 148 true MWEs.

In order to further evaluate AMs, precision and recall were computed for all *n* candidates and plotted as a precision-recall curve. The precision-recall

[5] For more detailed information about formulas used for these particular AMs see Appendix A.

[6] For interpretation see Landis and Koch (1977).

[7] MWE candidates annotated as UNABLE TO DETERMINE by at least 3 annotators were treated as NO.

lemma	transliteration	meaning	type	ambiguity	variation	freq	source
کامپیوتر	kāmpyuter	computer	loanword	no	no	10527	2004 (Vol. 1)
رایانه	rāyāneh	computer	equivalent	yes	no	7720	2004 (Vol. 1)
ماشین	māšin	machine	loanword	yes	no	9606	2004 (Vol. 1)
دستگاه	dastgāh	machine	equivalent	yes	no	27627	2004 (Vol. 1)
تکنولوژی	teknoloži	technology	loanword	no	no	12727	2004 (Vol. 1)
فناوری	fannāvari	technology	equivalent	no	yes	32448	2004 (Vol. 1)
پاسپورت	pāsport	passport	loanword	yes	no	1075	2008 (Vol. 5)
گذرنامه	gozarnāmeh	passport	equivalent	yes	no	2599	2008 (Vol. 5)
برند	brand	brand	loanword	no	no	21216	2012 (Vol. 9)
نمانام	namānām	brand	equivalent	no	no	45	2012 (Vol. 9)
آنلاین	ōnlāin	online	loanword	no	no	16497	2005 (Vol. 2)
برخط	barkhat	online	equivalent	no	yes	1646	2005 (Vol. 2)
پانوراما	pānourāmā	panorama	loanword	no	no	445	2015 (Vol. 12)
سراسرنما	sarāsarnāmā	panorama	equivalent	no	no	6	2015 (Vol. 12)
اکولوژی	ekoloži	ecology	loanword	no	no	859	2004 (Vol. 1)
بومشناسی	bumšenāsi	ecology	equivalent	no	yes	72	2004 (Vol. 1)
اسپرت	espourt	sport	loanword	no	no	3587	2008 (Vol. 5)
ورزش	varzeš	sport	equivalent	yes	no	26901	2008 (Vol. 5)
سمپوزیوم	simpouzium	symposium	loanword	yes	no	831	2004 (Vol. 1)
همنشست	hamnešast	symposium	equivalent	yes	yes	338	2004 (Vol.1)

Table 1: Lemmas used as seeds for multiword expressions discovery task.

curve is used to visualize the tradeoff between precision and recall for different thresholds (as proposed by Evert and Krenn, 2001), and it allows for direct comparison of different AMs. The precision-recall curve for the four selected AMs is presented in figure 2. For example, the red line shows that a ranking according to PMI achieves a recall of 11% at a precision of 37%. The same recall achieves 28% precision in case of log-likelihood (green line), 27% in case of χ^2 (blue line), and 33% in case of t-score (yellow line). A high area under the curve represents both high recall and precision. It is visible from the graph that t-score comprises the most significant area under the curve, achieving 17% precision at 63% recall, while for the remaining AMs, it is significantly lower.

The ratio of MWEs with loanwords and equivalents is shown in figure 3. Among all true MWEs, more cases were extracted with loanwords (55%) than with Persian equivalents (45%).

Finally, figure 4 shows the ratio of MWEs extracted with loanword and equivalent according to selected AMs. In all cases, MWEs with loanwords constitute a bigger group, with best results achieved by t-score (57%), followed by χ^2 and log-likelihood (both 54%). The best results for MWEs with equivalent were achieved by PMI (47%). Shared MWEs, i.e., MWEs that occur with both loanwords and equivalents, constitute almost 16% of all true MWEs (when counting shared MWEs only once).

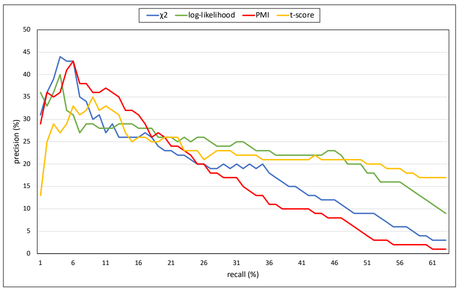

Figure 2: Precision-recall graphs for selected association measures evaluated against a final list of true MWEs.

The analysis of MWE candidates rejected by annotators revealed that sequences not labeled as true MWEs tend to belong to one of the following groups: 1) expressions with comparative adjectives, e.g., ماشین ارزان تر *cheaper car*, 2) expressions with adjectives describing nationalities, e.g., گذرنامه ایرانی *Iranian passport*, 3) expressions with intensifying adjectives, such as *super*, e.g., سوپر اسپرت *super sport* or 4) expressions containing

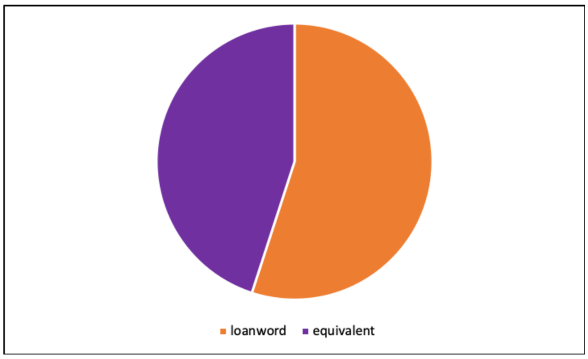

Figure 3: Comparison of the number of MWEs with loanword and equivalent.

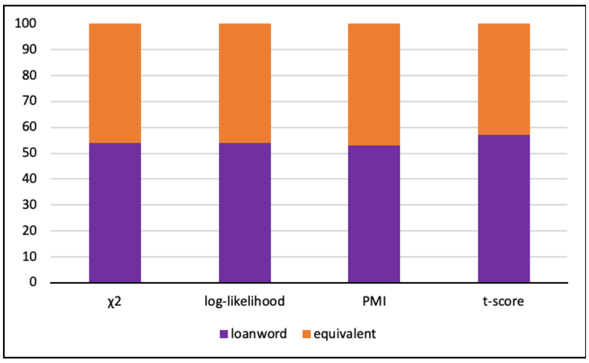

Figure 4: Relative ratio of true MWEs with loanword and equivalent according to selected association measures.

numbers, e.g., سومین سمپوزیوم *third symposium*.

This shows that there is still room for improvement, e.g. by expanding the stopword list or tagging the corpus with part-of-speech information.

4.3 Loanwords and Equivalents Evaluation

All analyzed loanwords turned out to form MWEs, while for 3 of the proposed equivalents (PANORAMA, BRAND, and SYMPOSIUM), no MWEs were found in the present data. The average number of MWEs per loanword is 21 and 17 per equivalent.

Interestingly, the more detailed analysis of extracted MWEs shows that in many cases, loanwords were not only not replaced by the equivalents proposed by the Academy, but the two (loanword and its equivalent) evolved to form distinct MWEs or even MWEs clusters. Pairs that, apart from sharing a substantial number of MWEs, have separate MWEs are: COMPUTER (33% shared MWEs), ONLINE (17% shared MWEs), TECHNOLOGY (18% shared MWEs) and PASSPORT (24% shared MWEs).

Figure 5 presents semantic network for lemma COMPUTER.[8] Both loanword and equivalent share a big number of MWEs. Among all MWEs, two main topics can be distinguished: computer types and computer parts. In case of computer types, apart from common MWEs (*quantum computer*, *pocket computer*), there are also MWEs that occur only with loanword, e.g., *gaming computer*, *minicomputer* or *all-in-one computer*, and ones that appear only with its Persian equivalent, e.g., *pentium computer* or *tablet*. The topic of computer parts can be found among shared MWEs, e.g., *computer monitor* or *computer keyboard* and MWEs with equivalent, e.g., *computer hard disc*, *computer mouse* or *computer processor*. It seems that the loanword does not have distinct MWEs related to computer parts. An interesting observation is related to attacks on computers. The loanword forms MWEs related to malware programms, e.g., *computer worm* and *computer virus* whereas the equivalent tends more to form MWEs referring to the activity itself, e.g., *infected computer* or *computer hacking*.

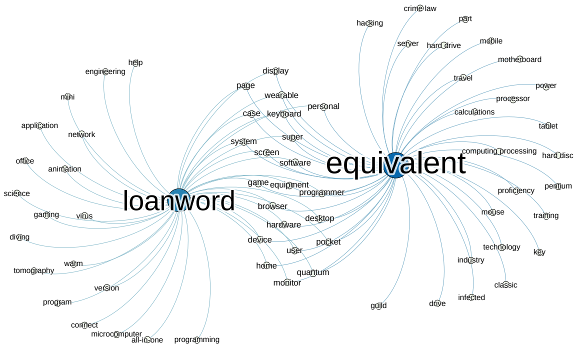

Figure 5: COMPUTER.

Both loanword and equivalent of lemma ONLINE share a substantial number of expressions. Many of the shared MWEs tend to center around the topic of trading, e.g., *online trading system*, *online trader* or *online stock trading*. Analysing all discovered MWEs, it can be observed that the shopping-related thema is quite predominant. Here again, apart from common MWEs (*online sale*, *online payment* and *online shopping*), loanword and its equivalent evolved to have their own MWEs: *online purchase*, *online transaction* and *online bill* in case of equivalent and more place-where-you-

[8] To check semantic networks for other pairs, please see Appendix B.

can-buy MWEs with loanword: *online shop*, *online store* and *online retail*. Loanword on its own has more negatively associated MWEs, e.g., *online harrasment*, *online attack* as well as expressions refering to gambling, e.g., *online gambling*, *online hazard* and *online casino*. Interestingly, both lemmas form MWE *online encyclopedia* but with two different Persian words for *encyclopedia*.

Loanword and its Persian equivalent TECHNOLOGY share a substantial number of MWEs, most of which represent different technology types, e.g., *information technology*, *face recognition technology* or *nano-technology*. Apart from common MWEs, both lemmas have specialized in certain types, i.e. loanword occurs in the following combinations: *infrared technology*, *quantum technology* or *LTE technology* whereas an equivalent can be found as part of *AI technology*, *HDR technology* and *Bluetooth technology*. What is more, the more positively associated MWEs are the ones with loanword, e.g., *technology upgrade*, *advances in technology* or *technology enthusiast*. The one negatively assoicated MWE - i.e., *outdated technology* - occurs with the equivalent.

Lemmas expressing PASSPORT differ in the number of MWEs: there are twice as many expressions with equivalent than with loanword. Main topics that can be distinguished here: different passport types, passport parts, passport-related activities and authentification. When it comes to types, there are common MWEs, e.g., *diplomatic passport* or *political passport* and MWEs with equivalent only, e.g., *biometric passport* and *electronic passport*. Passport parts apart from one common: *passport number*, form MWEs with loanword, e.g., *passport photo* and *passport cover*. The activity-related MWEs, except one shared (*issuance of passport*), are all formed with equivalent, e.g., *passport annulment*, *passport renewal* or *passport confiscation*. Finally, there is a cluster of MWEs related to passport authentification, e.g., *passport validity*, *fake passport* and *counterfeit passport*.

In case of MWEs with lemma SPORT, loanword refers more to sporty appearance (i.e. casual yet attractively stylish), e.g., *sporty look*, *sporty model* or *sporty design*. The meaning of sport as a physical activity is employed by MWEs with equivalent, e.g., *sport federation*, *to exercise sport*, *sport activity* or *professional sport*.

In the case of lemma ECOLOGY, there are no shared MWEs. In fact, for the Persian equivalent, only one MWE was found in the corpus, i.e., *ecological economics*.

For lemma MACHINE, there is only one MWE that both loanword and equivalent share: *smart machine*. The loanword tends to form constructions reffering to different types of machines, e.g., *washing machine*, *centrifugal machine* and *dishwasher (machine)*. Similar MWEs (also reffering to machine types) are found with the Persian equivalent, e.g., *X-ray machine* or *coffee machine*. Since the Persian equivalent is ambiguous, it occures also in expressions refering to body systems, e.g., *immune system*, *digestive system* and *respiratory system*.

Only MWEs with loanwords were found for the remaining three pairs: PANORAMA, BRAND, and SYMPOSIUM. This might be related to a quite late introduction of the equivalent by the Academy (in the case of BRAND and PANORAMA) and to a relatively low raw frequency in the corpus.

5 Conclusion

In this paper, an approach to the discovery of Persian MWEs was presented. We focused on a particular group of MWEs: constructions including loanwords in Persian and their native equivalents proposed by the official Iranian body responsible for language reforms - the Academy of Persian Language and Literature. The extraction of MWEs was performed with the use of four popular association measures. There were two goals of this study: 1) to evaluate the performance of association measures for the discovery of MWEs in Persian, and 2) to compare and analyze MWEs with loanwords and MWEs with equivalents.

The former goal was achieved for the four most popular association measures, with t-score performing best with loanword MWEs and PMI with equivalent ones. To our knowledge, it is the first time such analysis was carried out to discover Persian MWEs. The evaluation of MWEs with loanwords and their Persian equivalents was performed for ten pairs, providing information on shared MWEs as well as distinct ones. The complete list of extracted MWEs will be available for translators and students of the Persian language.

Future work includes exploiting a bigger number of association measures and other approaches to MWEs discovery. Moreover, we would like to investigate the impact of genres and context on forming distinct MWEs with loanwords and equivalents.

Acknowledgments

The author would like to thank the anonymous reviewers for their encouraging feedback and insights.

References

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Ltd., Essex, England.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen. Gunter Narr Verlag.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

Morten H. Christiansen and Inbal Arnon. 2017. More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, 9:542–551.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Mathieu Constant, Gülsen Eryigit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Mike Rosner, and Amalia Todirascu-Courtier. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43:837–892.

Mohammad Dabir-Moghaddam. 2018. Academy of Persian language and literature. In Anousha Sedighi and Pouneh Shabani-Jadidi, editors, *The Oxford Handbook Of Persian Linguistics*, pages 318–329.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.d thesis, University of Stuttgart.

Stefan Evert. 2008a. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

Stefan Evert. 2008b. A lightweight and efficient tool for cleaning web pages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France. Association for Computational Linguistics.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. 2017. E-view-alation – a large-scale evaluation study of association measures for collocation identification. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference*, pages 531–549, Leiden, The Netherlands.

Meghdad Farahmand and James Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 61–66, Berlin, Germany. Association for Computational Linguistics.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. 2007. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41:61–89.

Christiane Fellbaum. 2016. The treatment of multiword units in lexicography. In Philip Durkin, editor, *The Oxford Handbook of Lexicography*, pages 411–424.

Mark Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24, Portland, Oregon, USA. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2018. Multiword expressions: Between lexicography and NLP. *International Journal of Lexicography*, 32:138–162.

Marcos Garcia, Marcos García 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. Association for Computational Linguistics.

Ray Jackendoff. 1997. Twistin the night away. *Language*, 73:534–559.

Adam Kilgarriff. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In *AISB Workshop on Language Engineering for Document Analysis and Recognition*, pages 531–549, Sussex, UK.

Maria Krieger and Maria José Bocorny Finatto. 2004. *Introdução à terminologia: teoria & prática*. Contexto, Sao Paulo.

Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Katarzyna Marszałek-Kowalewska. 2011. Iranian language policy: a case of linguistic purism. *Investigationes Linguisticae*, 22:89–103.

Mostafa Morady Moghaddam and Farhad Moezzipour. 2017. Issues with language policy and planning in Iranian higher education. *Journal of English Language Teaching and Learning*, 9:187–221.

Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A language processing toolkit for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of COLING 2012*, pages 2077–2092, Mumbai, India. The COLING 2012 Organizing Committee.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech, Morocco.

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. Association for Computational Linguistics.

Ted Pedersen. 1996. Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*, Austin, USA.

Carlos Ramisch. 2009. Multi-word terminology extraction for domain-specific documents. Master thesis, Grenoble, France.

Carlos Ramisch and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslav Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, S.H.E. Mortazavi Najafabadi, and Amir Vaheb. 2018. MirasText: An automatically generated text corpus for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing-2002, pages 1–15, Mexico City, Mexico.

Bahar Salehi, Narjes Askarian, and Afsaneh Fazly. 2012. Automatic identification of Persian light verb constructions. In *Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2016. Determining the multiword expression inventory of a surprise language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 471–481, Osaka, Japan. The COLING 2016 Organizing Committee.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.d. thesis, University of Geneva.

Luigi Squillante. 2014. Towards an empirical subcategorization of multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 77–81, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Stubbs. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7:215–244.

Shiva Taslimipoor, Afsaneh Fazly, and Ali Hamze. 2012. Using noun similarity to adapt an acceptability measure for Persian light verb constructions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 670–673, Istanbul, Turkey. European Languages Resources Association (ELRA).

Yulia Tsvetkov and Shuly Wintner. 2014. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468.

Peter Uhrig, Stefan Evert, and Thomas Proisl. 2018. Collocation candidate extraction from dependency-annotated corpora: Exploring differences across parsers and dependency annotation schemes. In P. Cantos-Gómez and M. Almela-Sánchez, editors, *Lexical Collocation Analysis: Advances and Applications*. Springer International Publishing, Cham.

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733.

Alexander Wahl and Stefan Th. Gries. 2018. Multiword expressions: A novel computational approach to their bottom-up statistical extraction. In P. Cantos-Gómez and M. Almela-Sánchez, editors, *Lexical Collocation Analysis: Advances and Applications*, pages 85–109. Springer International Publishing, Cham.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42:7375–7385.

A Association Measures

Materials in this appendix section present information about association measures used and compared in this paper (as presented by (Evert, 2008b) and (Evert et al., 2017)).

	MWE	$\neg$ MWE
node	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$\neg$ node	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Table 2: Contingency table for MWE candidate pair: expected values (under the null hypothesis).

	MWE	$\neg$ MWE	
node	O_{11}	O_{12}	$= R_1$
$\neg$ node	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

Table 3: Contingency table for MWE candidate pair: observed values.

association measure	**formula**
PMI	$log_2 \frac{O_{11}}{E_{11}}$
log-likelihood	$2 \sum_{ij} O_{ij} log \frac{O_{ij}}{E_{ij}}$
t-score	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
χ^2 test	$\frac{N(\lvert O_{11}O_{22} - O_{12}O_{21}\rvert - \frac{N}{2})^2}{R_1 R_2 C_1 C_2}$

Table 4: Association measures compared in the study.

$$
\begin{aligned}
O_{ij} &= \text{contingency table of observed frequencies} \\
O_{11} &= \text{observed co-occurence frequency} \\
E_{ij} &= \text{contingency table of expected frequencies} \\
E_{11} &= \text{expected co-occurence frequency} \\
R_i &= \text{row sums of the contingency table} \\
R_1 &= \text{marginal frequency of node} \\
C_j &= \text{column sums of the contingency table} \\
C_1 &= \text{marginal frequency of collocate} \\
N &= \text{sample size}
\end{aligned}
$$

B Semantic networks

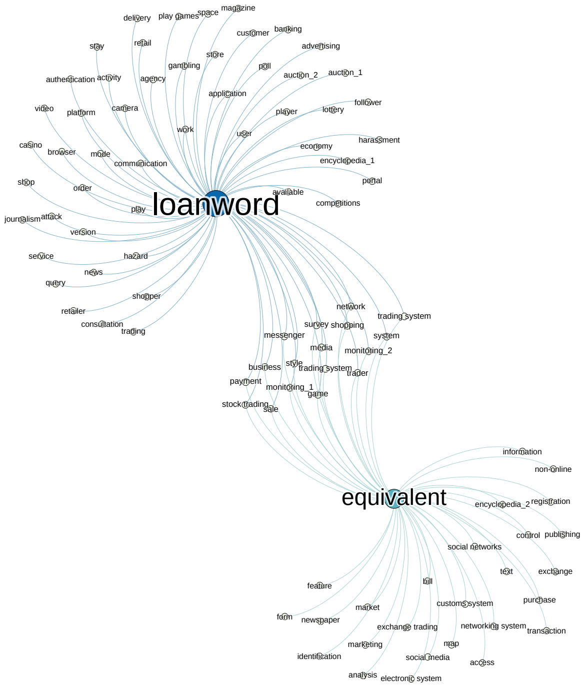

Figure 6: ONLINE.

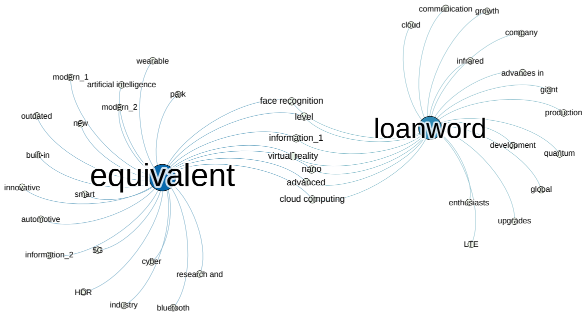

Figure 7: TECHNOLOGY.

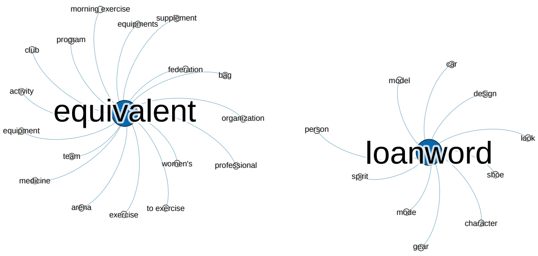

Figure 8: SPORT.

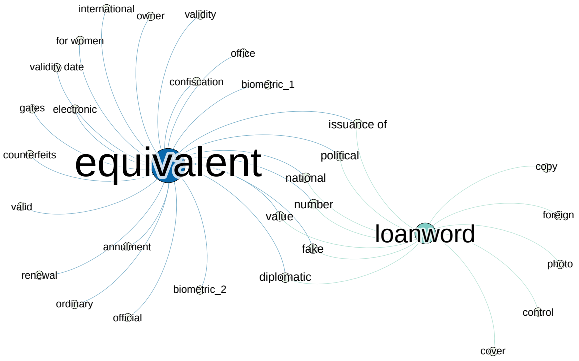

Figure 9: PASSPORT.

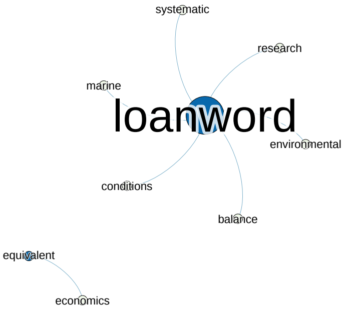

Figure 10: ECOLOGY.

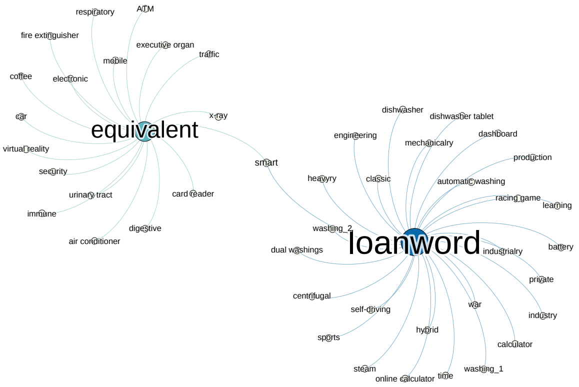

Figure 11: MACHINE.