Non Idiomatic Conventionalised Expressions: A New Pain in the Neck?

Ganesh Katrapati and Manish Shrivastava
International Institute of Information Technology Hyderabad
ganesh.katrapati@research.iiit.ac.in m.shrivastava@iiit.ac.in

Abstract

Multiword Expressions (MWEs) are linguistic units that are fixed, conventionalised, and function as single semantic units. While idioms have received considerable attention in NLP, another class of conventional expressions like "as far as I know", have remained underexplored. Unlike idioms such as *kick the bucket*, which are characterised by complete non-compositionality, these expressions can often be interpreted compositionally. However, this does not capture their meaning fully. We introduce the term *Non-Idiomatic Conventionalised Expressions (NICEs)* to describe this category of expressions that are largely compositional but retain crucial non-compositional elements. NICEs play important roles in discourse and pragmatics, making their systematic study essential.

We automatically discover and manually refine candidate NICEs, validating their distinct position on the compositionality spectrum through lexical substitution tests. Results show that they differ from regular constructions and that machine translation often renders them inaccurately.

1 Introduction

Multi Word Expressions (MWE) are linguistic forms spanning word boundaries that may have idiosyncratic interpretations (Jackendoff, 1996; Nunberg et al., 1994; Sag et al., 2002). *MWEs* are also generally fixed with little or no variation in their form (Moon, 2002, Calzolari et al., 2002) making them entrenched in language and *conventionalised* in their usage. *Idioms*, such as *"kick the bucket" or "spill the beans"*, are a sub-class of *MWEs* for which the meaning of the whole cannot be predicted from their component words alone i.e. they are semantically *non-compositional*.

Compositionality and Conventionalisation are non-binary and MWEs can be found across both spectra. At one end of the compositionality spectrum, we have completely idiomatic expressions (*Idioms*) while near the other end are fully conventionalised *MWEs* where all the words contribute to the meaning of the expression (Baldwin et al., 2003, Reddy et al., 2011). For instance, *Conjunct Verbs* like *"take a bath"* or Named Entities like "International Institute of Information Technology" which is perfectly compositional but acts as a single unit nonetheless.

Expressions such as *"as far as I know"*, *"as a matter of course"*, *"at one time or another"* are fixed and occur in predictably specific contexts. They are frequently used by proficient speakers and are considered to be important for successful participation in a linguistic community (Coulmas, 1979, Wray, 2000, Yorio, 1989, Edmonds, 2014). They also show a high degree of conventionalisation - the words in such expressions are not easily substitutable. In contrast to *Idioms*, they usually have a compositional interpretation but, while being seemingly semantically transparent, the usage of these expressions in language cannot be fully predicted from their words alone.

Consider the following sentence :

(a) Her father raised his hand once, hardly sparing ***a second glance*** .

Here, the expression *"a second glance"* is not considered an idiom (Haagsma et al., 2020), and indeed, one could interpret it compositionally but that would result in an incomplete understanding.

We categorise such expressions as *Non Idiomatic Conventionalised Expressions (NICE)* and situate them in the spectrum of compositionality. We draw inspiration from the Cognitive and Construction Grammar traditions (Langacker, 2008, Tyler, 2005, Wulff, 2013) which define all constructions as conventionalised grammatical patterns.

In the following sections, we briefly look at related work (2). We then describe our approach to automatically discover and extract candidate *NICEs* (3) and the filtering process (4). To validate our claim, We present two measures based on substitution tests (5) which show that conventionalised expressions exhibit a distinct behaviour when compared to regular compositional constructions. Finally, we demonstrate that machine translation systems frequently fail to capture NICEs accurately, leading to partial or misleading renderings.

2 Related Work

Most recent research on *MWEs* primarily centers on *Idioms*, particularly on determining whether a given *Potentially Idiomatic Expression* is used literally or idiomatically in context (Haagsma et al., 2019, 2020; Sporleder and Li, 2009a; Fazly et al., 2009a). These approaches typically depend on pre-defined lexical resources or idiom dictionaries (Moon, 1998; Cowie, 1993), thereby avoiding any automatic identification of new or unseen *MWEs*. Nonetheless, a subset of studies has addressed the problem of *MWE discovery*, which is of particular relevance to this paper.

There are two main approaches for MWE discovery. One approach relies on the *fixedness* property of MWEs and estimates the degree of association and co-dependency between its component words using Pointwise Mutual Information(PMI) (Church and Hanks, 1989) and its modified formulations (Fazly and Stevenson, 2006; Bouma, 2009) along with other statistical measures (Pecina and Schlesinger, 2006).

On the other hand, Baldwin et al. (2003); Katz and Giesbrecht (2006); Kiela and Clark (2013); Salehi et al. (2015) base their work on the *non-compositionality* of *MWEs*. They use distributional semantic models to compute the difference between *MWE* representation and the representations of its component words as a measure of the degree of its non-compositionality. Our work blends these two approaches for discovering candidate expressions.

While several types of Multi Word Expressions such as Verb-Noun constructions, Verb-Particle constructions, Noun Compounds and Idioms have been explored, non idiomatic conventionalised expressions (NICE) are relegated to the generic class of *collocations* (Sag et al., 2002) which can include *any* combination of words with a high enough degree of association.

Phrasal collocations also called *Phraseological Units*, *Lexical Bundles* are loosely defined as multi word collocations at the phrase level. They too have been mined from raw corpora using association measures (Colson, 2017, Hyland, 2008). However, these collocations are not differentiated by their compositionality and the categorisation of NICEs simply as collocations or lexical bundles does not capture the fact that these expressions are not completely compositional in nature.

3 Approach

We use the *British National Corpus (BNC)* (BNC, 2007) to extract candidate expressions. We find it ideal because of its size, variety and availability. It is also widely used in the area of idiom discovery and processing (Fazly et al., 2009b, Sporleder et al., 2010, Sporleder and Li, 2009b, Salton et al., 2017, Haagsma et al., 2019).

To extract candidate expressions, we follow a PMI-based approach (Church and Hanks, 1989; Fazly and Stevenson, 2006), scanning the corpus sequentially and merging word pairs (w_a, w_b) if their PMI exceeds a threshold θ [1]. Each merge creates a new candidate $(w_a\_w_b)$, and the algorithm proceeds greedily.

Since this over-generates candidates, we prune them using a *Minimum Description Length (MDL)* criterion, which balances data likelihood and model size. *MDL* has been widely used in unsupervised morphological segmentation (Creutz and Lagus, 2002), clustering (Li and Abe, 1992), and text pattern mining (Wu et al., 2010). Our pruning iteratively evaluates candidates and retains only those offering a cost benefit.

3.1 Ranking

We consider *NICEs* as constructions that are largely compositional but retain some degree of non-compositionality, which gives them distinct distributional patterns. To capture this, we segment the *BNC* so that each candidate expression forms a single token[2]. We then train a *word2vec* model (Mikolov et al., 2013) on the segmented corpus and compute a *compositionality score*: the cosine distance between a candidate's embedding and the normalised sum of its component embed-

[1] We set $\theta = 10$

[2] Words are joined with underscores

dings. Candidates with higher scores (lower compositionality) are ranked and selected for further analysis.

It is worth noting that when we tried this approach using Contextualised Embeddings (Devlin et al., 2019), the resultant score did not reflect the compositionality of the candidates in any way. Our results echo those of Pickard (2020); Cordeiro et al. (2016); Nandakumar et al. (2018) in this matter and this merits further observation.

4 Filtering

From the ranked candidate list, we removed Named Entities, Nominal Compounds, Conjunct/Compound Verbs, and Idioms (Haagsma et al., 2020), as they are not within our scope. Three computational linguists proficient in English then filtered the list manually. An expression was marked as *NICE* if substitutions made it less natural (e.g., *"a moment or two"* vs. *"a moment or three"*) or altered its meaning entirely (e.g., replacing *little* with *less* in *"with little difficulty"*). As we could not perform an exhaustive filtering of all expressions, the top *300* ranked expressions were selected; out of these, *90* expressions were tagged as *NICEs* by all the validators [3] . *We submit that this is not at all a comprehensive list of NICEs but perhaps sufficient to illustrate the topic of our study.*

5 Measures

Here, we describe two measures to test the results of our approach against our definition of *NICEs*. Both the measures are based on substitution tests designed to compare *NICEs* with regular language usage.

For both the measures, we randomly selected a list of (a) Adjective-Noun (AN) pairs, (b) Verb-Determiner-Noun (VN) combinations and (c) NGrams (Random) of length 3, and compare them with the *NICEs*.

5.1 Naturalness

As noted in (4), substitutions within a conventionalised expression reduce its *naturalness*. In a corpus, such valid substitutions are therefore rare. To measure this, we take the nearest neighbours of each word in an expression, generate leave-one-out substitutions, and compute their occurrence

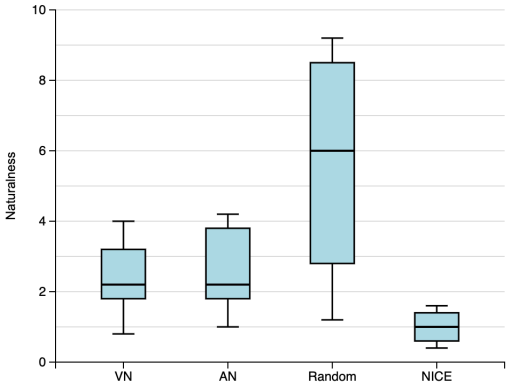

Figure 1: A box plot for naturalness metric across different types of expressions

counts along with the number of unique substitutions found.

$$Sub_N(E) = C(E) * CR(E) * MR(E) \quad (1)$$

$$CR(E) = \frac{\sum_{i \in S_E} C(s_i)}{C(E)} \quad (2)$$

$$MR(E) = \frac{\sum_{i \in S_E} min(1, C(s_i))}{S_E} \quad (3)$$

Where $C(E)$ is the frequency of expression E in the corpus, S_E is the set of possible substitutions, and $CR(E)$ is the ratio of all substitution counts to that of E (eq. 2). We also measure substitution variability $MR(E)$ (eq. 3) as the proportion of substitutions found in the corpus out of all possible ones, and finally normalise by the word length of E.

5.2 Semantic Shift

NICEs typically occur in contexts not shared by their substituted variants. To capture this, we use a transformer-based language model[4] to obtain sentence embeddings for both the *NICE* instances and their substitutions, and compute cosine distance as a measure of semantic similarity.

$$Sub_S(E) = \frac{1}{S_E} \sum_{i \in S_E} distance(\overrightarrow{V_E}, \overrightarrow{V_{S_i}}) \quad (4)$$

Where $\overrightarrow{V_E}$ and $\overrightarrow{V_{S_i}}$ are averaged sentence embeddings of the *NICE* (E) and the substituted expression S_i respectively. The measures show consistently lower means across all quartiles for conventionalised expressions compared to other types

[3] You can download all the data and the MT evaluations at https://github.com/neshkatrapati/nice-data

[4] We use https://www.sbert.net/ (Reimers and Gurevych, 2019)

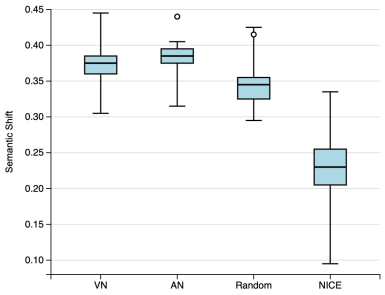

Figure 2: A box plot for semantic shift metric across different types of expressions

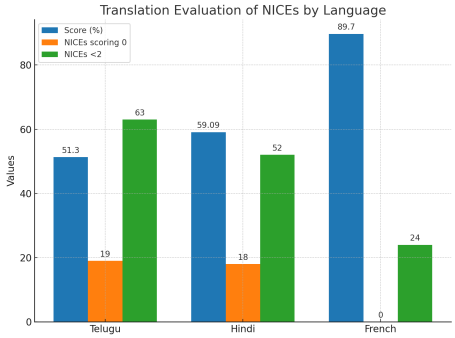

Figure 3: Translation Evaluation of NICEs by Language

(Figures 1, 2), supporting our claim that *NICEs* cannot be substituted without losing naturalness or altering meaning.

6 Impact on Machine Translation : A Preliminary Study

To investigate the effect *NICEs* may have on NLP applications, we chose to explore Machine Translation and performed a *preliminary study.*

We chose three languages, each with varying links to English. *French* for familiar syntax and shared cultural aspects, *Hindi*, an Indo-Aryan language which shares some structural features (same language family), and finally *Telugu*, a Dravidian language with marked diverge in morpho-syntax when compared to English.

We designed a translation evaluation task to assess how well NICEs are preserved across languages. Annotators were given English source sentences (taken from *BNC*) with a highlighted NICE, along with corresponding translations from an MT system [5]. Their task was to judge only the translation of the NICE , independent of the rest of the sentence. Each item was rated on a 0, 1, 2 scale: 0 for incorrect or literal translations, 1 for partially correct or unnatural ones, and 2 for correct and natural renderings. Each language was annotated by one annotator who was proficient in the respective language and also a linguist. The evaluation set consists of *220* sentences averaging *2 to 3* examples per *NICE*.

6.1 Results

French translations performed best with 89.7%, which is expected given shared linguistic and cultural structures with English. In contrast, results dropped sharply for Indic languages: Hindi, despite being in the same language family, scored only 59%, and Telugu performed worst at 51.3%. This highlights the challenge of preserving NICEs across languages with less structural and cultural overlap. This cannot be attributed solely to data scarcity. English-Indic language pairs have considerable parallel corpora (Kunchukuttan et al., 2018; Ramesh et al., 2022), more so, in the case of Hindi. It should also be noted that *NICEs* are not exactly rare phenomena.

Moreover, as you can see in Fig.3, a number of expressions are unsatisfactorily translated and still some others entirely missed or mis-translated. Even French, which has a high overall score, had 24 out of 90 expressions with less than perfect score.

7 Conclusion

Our findings highlight that *NICEs* cannot be simply subsumed under broader categories such as *MWEs*, *Lexical Bundles*, or *Phrasal Units*. Their partially non-compositional meaning requires explicit recognition and treatment as a distinct linguistic phenomenon. Ignoring this dimension risks overlooking essential aspects of their semantics, which can undermine the performance of downstream NLP tasks as evidenced by our MT results.

Limitations

Our study has several limitations. The set of *NICEs* examined is relatively small and not exhaustive, and validation relied on only a few annotators, with limited language coverage. Nonetheless, the consistency of results across different languages and annotators suggests that the observed patterns are robust and worth pursuing in future, larger-scale studies.

[5]We used Google Translate which is one of the most widely used MT system

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. pages 89–96. Association for Computational Linguistics.

BNC. 2007. British national corpus, XML edition. Oxford Text Archive.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. 16(1):76–83.

Jean Pierre Colson. 2017. The idiomsearch experiment: extracting phraseology from a probabilistic network of constructions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10596 LNAI:16–28.

Silvio Cordeiro, Carlos Ramisch, Marco A. Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997. ACL.

Florian Coulmas. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics*, 3(3-4):239–266.

A. P. Cowie. 1993. *Oxford Advanced Learners Dictionary of Current English*, 5th edition. Oxford University Press, Oxford.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. pages 21–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Amanda Edmonds. 2014. Conventional expressions. *Studies in Second Language Acquisition*, 36(1):69–99.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009a. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009b. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 337–344.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, (May):279–287.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a Wide Net: Robust Extraction of Potentially Idiomatic Expressions. pages 1–34.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Ray Jackendoff. 1996. The architecture of the language faculty.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis. MWE '06, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1427–1432, Seattle, Washington, USA. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english–hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ronald Langacker. 2008. *Cognitive Grammar: A Basic Introduction*, volume 12. Oxford University Press.

Hang Li and Naoki Abe. 1992. Clustering Words with MDL Priciple. *Journal of Natural Language Processing*, 4(2):71–88.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint*.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press, Oxford.

Rosamund Moon. 2002. *Fixed Expressions and Idioms in English: A Corpus-Based Approach (review)*, volume 78.

Anoop Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. How well can we predict multiword expression compositionality using embedding-based methods? In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 73–81. ALTA.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

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.

Thomas Pickard. 2020. Comparing word2vec and glove for automatic measurement of mwe compositionality. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE+ELex 2020)*, pages 95–100, Barcelona, Spain (Online).

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. *IJCNLP 2011 - Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *CICLing*.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 977–983.

Giancarlo Salton, Robert Ross, and John Kelleher. 2017. Idiom type identification with smoothed lexical features and a maximum margin classifier. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 642–651, Varna, Bulgaria. INCOMA Ltd.

Caroline Sporleder and Linlin Li. 2009a. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009b. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Caroline Sporleder, Linlin Li, Philip John Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 639–646.

Andrea Tyler. 2005. Cognitive grammar. *Studies in Second Language Acquisition*, 27:650 – 651.

A Wray. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4):463–489.

Ke Wu, Jiangsheng Yu, Hanpin Wang, and Fei Cheng. 2010. Unsupervised text pattern learning using minimum description length. *2010 4th International Universal Communication Symposium, IUCS 2010 - Proceedings*, pages 161–166.

Stefanie Wulff. 2013. Words and idioms.

Carlos A. Yorio. 1989. *Idiomaticity as an indicator of second language proficiency*, page 5572. Cambridge University Press.