Investigating Idiomaticity in Multiword Expressions used in Cinematic Discourse

Marina Kogan & Arina Polyakova
Peter the Great Saint Petersburg Polytechnic University, Russia

The study-in-progress presents a corpus-based exploration of idiomatic expressions in contemporary film dialogue, focusing on their identification, classification, and potential application in both language pedagogy and NLP tasks. The academic community hasn't reached complete agreement on what exactly constitutes an idiom. In this study we considered that idioms are fixed or semi-fixed multiword expressions whose overall meaning is non-compositional, and whose grammatical structure is often frozen or resistant to syntactic variation (Halliday, Matthiessen, 2013). The research examines how idioms function at the intersection of lexis and grammar in authentic cinematic discourse.

The initial phase of the project involved the manual annotation of idiomatic expressions in the script of *The Trial of the Chicago 7* (2020), a film characterized by rich socio-political dialogue. This hand-annotated material forms the foundation of a mini-corpus of sentences containing multiword expressions accounting for idiomaticity of the characters' speech. The mini-corpus serves as a critical dataset for both qualitative analysis, basis for further NLP modeling and the development of methodology focusing on the recognition, grammatical structure, and lexical components of idiomatic expressions. A total of 72 idioms have been identified so far. To validate and expand the dataset, the idioms were checked in three corpora from English-corpora.org (Movies corpora, SOAP, TV\Movies and Spoken COCA sub-corpora), using a combination of PMI metrics, POS-tagging, and semantic similarity via BERT (Ramisch & Villavicencio, 2018; Baldwin et al., 2002). However, the available 1-million-word samples lacked sufficient context for reliable automatic identification. Retrieved expressions were repetitive across sub-corpora and rarely aligned with manually annotated data. The following expressions appeared with the biggest frequency: 'mess up', 'hang on', 'count on', 'break the law', 'give a shit', 'on behalf of'. This suggests that pre-trained models and standard corpora are currently inadequate for capturing low-frequency, context-dependent idioms, reinforcing the necessity of manual verification.

The study raises three central questions: 1) what are the limitations of current corpus infrastructure for idiom identification; 2) how can idiomatic expressions be classified based on their lexicogrammatical properties; 3) how to build the corpus around the selected list of film idiomatic expressions for fostering idiomaticity in language pedagogy and NLP applications such as machine translation and automatic subtitle generation (Forchini, 2018)?

Preliminary findings suggest that even a small set of high-frequency idioms exhibits significant structural and functional diversity, reinforcing the need to approach idioms not merely as lexical items, but as compound units that participate in genre-specific communicative practices (Fillmore et al., 1988; Villavicencio et al., 2005). A selected set of 72 idioms was grouped into five relevant categories: strongly figurative idioms ('get off the chest', 'trade a cow for magic beans'), light idioms and collocational verbs ('egg on', 'settle things down'), legal/institutional fixed expressions ('contempt of court', 'call the case'), metaphorical constructions ('bumper sticker patriots', 'double-lip talking'), and pragmatic colloquial phrases ('stoned', 'whatshisname'). This categorization not only underscores the lexical diversity of idioms but also their intricate grammatical structures. Even among fewer than ten core idioms, significant variation was observed in frequency, structural patterns, and genre-specific usage. This categorization revealed how idioms cut across traditional linguistic boundaries. Many items resist easy classification, blending features of fixed expressions, phrasal verbs, and metaphors. Their grammatical behavior, especially verb-preposition-noun combinations, emerge as a productive area for corpus-based modeling (Baldwin et al., 2002; Ramisch & Villavicencio, 2018). For example, expressions like 'get off the chest' or 'fight fire with fire' reflect both idiomatic meaning and stable grammatical form. Legal expressions such as 'call the case', 'sustained', and 'voir dire' show syntax bound to institutional contexts, making them valuable for legal-domain ESP and NLP applications (Calzolari et al., 2002; Forchini, 2018).

Ultimately, this paper positions idiomatic expressions as a valuable test case for examining such concepts as compositionality and idiomaticity. By highlighting the methodological challenges of idiom identification and corpus expansion, and by proposing practical applications for both language learning and language technology, the study contributes to ongoing discussions about idiomaticity, corpus methods, and how to capture complex patterns of meaning and grammar in real-life usage (Villavicencio et al., 2005; Baldwin & Kim, 2010). Specifically, the findings advance the field of lexicogrammar by providing empirical evidence of how idioms exemplify the integration of lexical and grammatical elements in authentic discourse which can be used to develop more accurate language models and teaching materials.

References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), Handbook of natural language processing (2nd ed., pp. 267–292). CRC Press.

Baldwin, T., Sag, I., Bond, F., Copestake, A., & Flickinger, D. (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), Mexico City (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45715-1_1

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002) (pp. 1934–1940). European Language Resources Association.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. Language, 64(3), 501–538. http://www.jstor.org/stable/414531

Forchini, P. (2018). The applicability of movies in legal language teaching: Evidence from multi-dimensional analysis. International Journal of Linguistics, 10(6), 245–262.

Halliday, M.A.K. & Matthiessen, Christian. (2013). Halliday's Introduction to Functional Grammar. https://doi.org/10.4324/9780203431269.

Ramisch, C., & Villavicencio, A. (2018). Computational treatment of multiword expressions. In R. Mitkov (Ed.) The Oxford Handbook of Computational Linguistics (2nd ed., pp. 649–678). Oxford University Press. DOI: 10.1093/oxfordhb/9780199573691.001.0001

Villavicencio, A., & Idiart, M. (2019). Discovering multiword expressions. Natural Language Engineering, 25(6), 715–733. https://doi.org/10.1017/S1351324919000270

Villavicencio, A., Bond, F., Korhonen, A., & McCarthy, D. (2005). Introduction to the special issue on multiword expressions: Having a crack at a hard nut. Computer Speech & Language, 19(4), 365–377. https://doi.org/10.1016/j.csl.2005.05.001