Knowledge Graphs: Semantics, Machine Learning, and Languages
M. Acosta et al. (Eds.)

doi:10.3233/SSW230006

Perplexed by Idioms?

J. Nathanael PHILIPP [a,b,1], Max KÖLBL [c], Erik DAAS [b], Yuki KYOGOKU [b] and Michael RICHTER [b]

[a] *Serbski institut*
[b] *Leipzig University*
[c] *Osaka University*

ORCiD ID: J. Nathanael Philipp https://orcid.org/0000-0003-0577-7831, Max Kölbl https://orcid.org/0000-0001-7460-4139, Erik Daas https://orcid.org/0009-0007-0879-0878, Michael Richter https://orcid.org/0000-0001-7460-4139

Abstract. The aim of this study is to identify idiomatic expressions in English using the measure *perplexity*. The assumption is that idiomatic expressions cause higher perplexity than literal expressions given a reference text. Perplexity in our study is calculated based on n-grams of (i) PoS tags, (ii) tokens, and (iii) thematic roles within the boundaries of a sentence. In the setting of our study, we observed that no perplexity in the contexts of (i), (ii) and (iii) manages to distinguish idiomatic expressions from literals. We postulate that larger, extra-sentential contexts should be used for the determination of perplexity. In addition, the number of thematic roles in (iii) should be reduced to a smaller number of basic roles in order to avaiod an uniform distribution of n-grams.

Keywords. perplexity, idiomatic expressions, literal expressions, information theory

1. Introduction

Idiomatic expressions, or phraseologies, (in the following **IE**) such as the verb-noun compounds (**VNC**) *kick the bucket*, *hit the fan*, *blow whistle*, *hit the sack* or *lose face* are ubiquitous in the English language. The aim of this study is to automatically identify IE in English texts using an information theoretic framework [1].

IE are far less subject to the principle of compositionality than literal expressions [2,3] since in most cases, the meaning of an expression cannot straightforwardly be derived from the meaning of its parts. The interpretation of idioms thus is not reductionist. So, to unlock the meaning of the IE for instance in the sentence *the shit has hit the fan at our house* it is not sufficient to know the meaning of *shit*, *hit* and *fan* alone. Understanding an IE touches on conventionality in language, since its meaning has evolved through specific language usage and convention. [2] emphasise that *[t]he meaning of IE involves i.a. metaphors and hyperboles*, and the meaning of the constituents of IE is overruled. IE are stable linguistic constructions, mostly with specific syntax as in *lose face* or *blow whistle*, a feature referred to as *(In)flexibility* [2,3]. This feature also means the imper-

[1] Corresponding Author: J. Nathanael Philipp, nathanael@philipp.land.

meability of IE, i.e., grammatical transformations, extractions and insertions lead to ungrammaticality, as the sentences in (1) exemplify. ('*' indicates ungrammaticality):

(1)a * The fan at our house was hit by the shit. (passive transformation: the object becomes the grammatical subject)
b * It was the fan that the shit has hit. (clefting)
c * The shit has hit the wanky fan at our house. (insertion)

In this study, we investigate whether it is possible to distinguish IE from literal expressions by the divergence measure of *perplexity* which is close related to the information theoretic feature of *surprisal* [4,5,6,7]. So, we ask whether IE cause higher perplexity than literals given a reference data set. The amount of perplexity which is exactly defined in the Eqs. 1 and 2, represents the amount of surprisal which is contextualised information and is to be interpreted as the deviation from the expected. The higher the surprisal of a linguistic unit, and thus the higher the perplexity it causes, the greater the cognitive effort is to process it [5,7] which is an empirically proven claim (see amongst others [8,9]).

In our study, we measure the divergence of IE and literals from reference texts, and the assumption is that these divergences should differ. In the data set of the study (see Section 3), the VNC can be understood both literally or as IE.

We determine perplexity as the divergence between n-gram-probabilities of linguistic patterns in two sorts of data sets on the one hand, i.e., one that contains IE and another set that does not, and on the other hand two large reference data sets.[2]

What is the practical interest in dealing with the identification of IE? While some state-of-the-art translation programs and chat bots are able to recognise IE and provide correct translations and interpretations, some others do not. Google Translator for instance translates *yesterday the shit has hit the fan at our house* into German as *gestern hat die Scheiße den Lüfter unseres Hauses getroffen* which is literally correct, but does not make any sense at all. DeepL and in particular OpenAI/ChatGPT do a much better job in this respect. However, the latter shows slight weaknesses in interpreting the German IE *nicht auf den Kopf gefallen sein* which means to be clever and smart. We had a dialogue in with ChatGPT from Open AI[3].

Question: *Wir sind nicht auf den Kopf gefallen. Was heißt das?* lit: 'We did not fall onto our heads. What does that mean?'

Answer: *Dieser Ausdruck bedeutet, dass man klug und clever ist und nicht dumm. Man ist in der Lage, kluge Entscheidungen zu treffen und ist nicht leicht hinters Licht zu führen.* 'This expression means that one is smart and clever and not stupid. One is able to make wise decisions and is not easily fooled.'

This is a really good answer, the chat bot recognises the IE and explains it well. Now consider the following dialogue:

Question: *Wir sind nicht vom Fahrrad auf den Kopf gefallen. Was heißt das?* 'we did not fall off our bikes onto our heads. What does that mean?'

[2] Doing so, we are inspired by the workflow in [10] on classification of translations by perplexity.
[3] https://chat.openai.com

Answer: Diese Redewendung bedeutet, dass man nicht dumm oder unerfahren ist, sondern dass man gut informiert und handlungsfähig ist. Es drückt auch die Zuversicht aus, dass man in einer bestimmten Situation erfolgreich sein wird. 'This idiom means that one is not stupid or inexperienced, but that one is well informed and able to act. It also expresses confidence that one will succeed in a given situation.'

This explanation however is not correct, since *vom Fahrrad* is inserted into IE whereby the expression lost its idiomatic character and is to be understood literally. ChatGPT/OpenAI however interpreted also this sentence as an idiomatic expression.

In the present study, the identification of IE is unsupervised and as a first step will be based on n-grams of PoS-tags [10] which yields perplexity expressing grammatical surprisal. Secondly, we will consider n-grams of tokens as a representation of the respective data sets which yields perplexity expressing lexical surprisal. Thirdly, we will consider sequences of *thematic roles*, i.e., semantic types of entities, processes and events as defined by [11,12] which yield perplexity expressing semantic surprisal.[4] In the due course of this paper, we will use the latter term. [15, p. 48] describes thematic roles as follows:

Thematic roles are generalisations among the arguments of a predicate in order to capture regularities between the semantic representation and the syntactic expression of that predicate.

Very elementary thematic roles are, for example, *agent* and *patient*, i.e., the participants in transitive scenes in the world and transitive linguistic constructions that represent those scenes. A sentence like *we eat a chocolate* could be assigned the thematic roles **Agent-Process-Patient**.

2. Previous work on automatic detection of idiomatic expressions

To the best of our knowledge there is no work on identification of idioms within the framework of information theory. A recent study on the automatic classification of phraseologies by [16] reports a unsupervised, classification of IE based on topic detection. The authors assume that words that are highly relevant in the main topic of discourses are not very likely to occur in IE. In other words, IE are assumed to be semantically distinct from the main topic of the discourse, and, in addition, the study brings to light that IE are associated with a higher level of affectivity. This was proposed already in an earlier study [17], in which the authors state that IE are semantic 'outliers' in a given context and thus cause surprisal. Identification of IE in [17] is carried out by the *Principal Component Analysis*.

The model in [18] first generates both static and contextualised word embeddings. Additional information such as PoS-tags is incorporated in the attention phase, and in the enriched static phase, embeddings are further combined with the contextualised embeddings. This is input to a BiLSTM-neural network: if the contextualised representation of a word is semantically compatible with its context, it is literal; if not, it is figurative.

[19] propose a model that is characterised by syntagmatic and context features, and, in addition, by other features such as the number of words in a collocation. For the syntagmatic feature, both count-based and predictive models are proposed (for the two

[4][13] notes that frame semantics has two origins: a linguistic origin from Fillmore's case grammar and an origin from 'Artificial Intelligence' in [14].

models, see [20]). The effect of each feature varies according to the characteristics of the datasets.

[19] conclude that the context feature contributes to detecting semantically dissimilar words, while the count-based measure contributes to assessing the fixedness of collocations.

3. Dataset, concepts and technique of analysis

The data on which this study is based have been extracted from the *British National Corpus* (BNC)[5] and were already used in [16][6]. The dataset comprises 1997 sentences with idioms and 535 literal expressions. To avoid any bias, we split the data into five data sets of almost even size: four sets with about 500 sentences each which contain idioms and one set with 535 sentences containing literal expressions. As in [16], only verb-noun constructions (VNC) are in the focus of the present study. For the labeling of VNC as IE or literal expression, [16] used the list in [21,22]. [16] treated idiomacity as a binary property and explicitly not as a gradual property [23].

In order to determine the perplexity of IE and literals, we use two types of reference data sets, namely a news corpus with 1M sentences from the *Wortschatz Leipzig* (eng_news_2020_1M) corpora collection[7], and a Wikipedia corpus also with 1M sentences, taken from the same source (eng_wikipedia_2016_1M).

For PoS-tagging, we employ the *spaCy*[8] parser which assigns 15 PoS tags. Thematic roles are assigned by *LOME*, a *system for multilingual information extraction* [24]. For entity-assigns makes use of about 2000 thematic roles[9]. Specifically, we used for the entity-type parsing the program *span-finder*[10].

3.1. Perplexity: Measure of surprisal

Perplexity (PP) is a measure of how well a probability distribution in a statistical language model predicts a data sample[11]. It is defined as an two to the power of the entropy H of a probability distribution as exponent, as given in Eq. 1; the lower the perplexity, the better the model.

$$PP = 2^H \tag{1}$$

In this study, H is the *conditional entropy*: expressed in terms of information theory, the conditional entropy is a measure of the quality of a model for a probability distribution q, given a true distribution p. This is reminiscent of the idea behind the Kullback-Leibler divergence, and like this, conditional entropy is not symmetrical. PP uses condi-

[5] https://github.com/bondfeld/BNC_idioms
[6] The data were made available by Jing Peng and Anna Feldman.
[7] https://wortschatz.uni-leipzig.de/de
[8] https://spacy.io
[9] https://framenet.icsi.berkeley.edu/fndrupal/frameIndex
[10] https://github.com/hiaoxui/span-finder
[11] https://en.wikipedia.org/wiki/Perplexity

tional entropy to indicate the degree perplexity, when a model's prediction is compared with a data sample. For the calculation of entropy, we use Eq. 2:

$$H = -\sum_{w,c} q(w,c) \log_2 p(w|c) \tag{2}$$

w denotes a linguistic unit and c its context. For the PoS-tags, each dataset is represented as a combined distribution of bi- to heptagrams, for the tokens, each dataset is represented as a combined distribution of the bi- and trigrams, and for the thematic roles a combined distribution of the bi- to tetragrams is used. PP expresses the divergences between these probabilities.

4. Results

Table 1 gives the perplexity between the reference data sets and the BNC data, i.e., four data sets with IE (*Idioms* 1–4) and one data set with exclusively literal expressions (*Literals*).

Reference dataset	Idioms 1	Idioms 2	Idioms 3	Idioms 4	Literals
Wikipedia	6.06	6.11	6.11	6.18	6.06
News	5.92	6.00	6.00	6.03	5.93

Table 1. Perplexity based on PoS-tag probabilities between reference data sets, literals and idioms.

Table 1 shows approximately the same perplexity for all IE and literals: for n-grams and POS tags, IE and literals are each distributed similarly to the reference data.**Table** 2 gives the perplexity between the reference data and the BNC data based on bi- and trigrams of tokens.

Reference dataset	Idioms 1	Idioms 2	Idioms 3	Idioms 4	Literals
Wikipedia	19.30	18.63	18.63	17.74	18.27
News	19.54	18.66	18.66	18.19	18.52

Table 2. Perplexity based on bi- and trigram probabilities of tokens between reference data sets, literals and idioms.

Again, there is hardly any difference in the perplexity caused by idioms and literals. We observe that the perplexity values are much higher here than for PoS-tags which is due to the fact that the set of PoS-tags is considerably smaller than that of the tokens. Recall Eq. 1: high average information due to low probabilities of signs causes high perplexity and vice versa. N-grams of tokens are more informative than n-grams of PoS tags. **Table** 3 shows the perplexity of semantic surprisal from thematic roles.

Here, the same picture emerges as in the **Tables** 1 and 2: based on semantic surprisal of thematic roles, literals and idioms cannot be distinguished from each other by perplexity. We notice again that the amount of perplexity correlates with the number of linguistic units it is based on. The results in **Table** 3 are derived from (about) 2000 thematic roles, and thus the values are higher than those based on 15 PoS tags, but they are lower than the values derived from the large, entire set of tokens.

Reference dataset	Idioms 1	Idioms 2	Idioms 3	Idioms 4	Literals
Wikipedia	14.49	14.49	14.49	14.09	14.40
News	14.03	13.97	13.97	13.34	13.71

Table 3. Perplexity based on bi- and trigram probabilities of thematic roles between reference data sets, literals and idioms.

5. Discussion and Conclusion

In our study, IE could not be identified by any type of perplexity. The observation that perplexity from PoS tags as grammatical surprisal fails is plausible because all IE in the corpus are verb-noun compounds (VNC). They structurally correspond to the VNC of the literals. We can interpret perplexity of grammatical surprisal as a baseline, and moreover, this shows that perplexity is a suitable measure for our aims: when perplexity is based on grammatical structures, VNC in IE and literals should not exhibit large differences. This is what we observe. However, there is also no perplexity with lexical surprisal, and this is not what we expected. That is, n-grams of tokens do not give the language processor any clue to distinguish IE from literals. This is actually a semantic distinction. This finding is surprising because in the common and influential distributional language model [25], it is assumed that the meaning of a linguistic unit is represented in its context. Prototypically, contexts in this model mean co-occurrences of a linguistic unit in a predefined context window. Similar contexts indicate a similar meaning of a linguistic unit, while different contexts indicate a different meaning. We attribute the problem with the n-grams of tokens in our study to the small context sizes. Contexts of bi- to tetragrams cannot represent a larger conversational context, let alone a discourse. So, for example, if the word *ganache*, i.e., a soft chocolate filling, occurs in the conversational context of, say, quantum mechanics, it is presumably not possible to derive from a bigram-context the high surprisal / perplexity that *ganache* in this special context causes. The same effect emerges with semantic surprisal, since n-grams of thematic roles within the sentence boundaries do not yield higher perplexity with IE than with literals.

The restriction to local contexts within the sentence boundaries is due to our data basis, because literals and idioms are represented as isolated sentences here. However, this is a crucial shortcoming, since Levy's definition of surprisal [7] provides for extra-sentential contexts. According to this definition, surprisal is the deviation from the expected based on co-occurrence, but also from a larger context which can be the discourse, an entire corpus, or a paragraph, in short, any linguistic material that exceeds sentence boundaries. The perplexity scores that we calculated for lexical and semantic surprisal underpin Levy's definition, since we observed that inner-sentential contexts are not sufficient to distinguish IE from literals. A language processor needs thus a larger context to decide whether an expression can be understood literally or as IE. Derivation of surprisal / perplexity from extra-sentential contexts is thus a task of future work.

Another problem in calculating semantic surprisal concerns the large number of thematic roles that LOME assigns which makes the possibility quite high, that a given context will appear only once and finally, this would lead to uniform perplexity values. In future work, we will consider a reduction to 'basic' thematic roles, as initially introduced by Fillmore [11] such as *agent*, *patient*, *theme*, *instrument* and *source*.

References

[1] Shannon CE. A mathematical theory of communication. The Bell system technical journal. 1948;27(3):379-423.

[2] Espinal MT, Mateu J. Idioms and phraseology. In: Oxford Research Encyclopedia of Linguistics. Oxford University Press; 2019. .

[3] Nunberg G, Sag IA, Wasow T. Idioms. Language. 1994;70(3):491-538.

[4] Tribus M. Information theory as the basis for thermostatics and thermodynamics. Journal of Applied Mechanics. 1961 03;28(1):1-8. Available from: https://doi.org/10.1115/1.3640461.

[5] Hale J. A probabilistic Earley parser as a psycholinguistic model. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics; 2001. p. 1-8.

[6] Hale J, Lutz D, Luh WM, Brennan J. Modeling fMRI time courses with linguistic structure at various grain sizes. In: Proceedings of the 6th workshop on cognitive modeling and computational linguistics; 2015. p. 89-97.

[7] Levy R. Expectation-based syntactic comprehension. Cognition. 2008;106(3):1126-77. Available from: https://doi.org/10.1016/j.cognition.2007.05.006.

[8] McCusker L. Some determinants of word recognition: Frequency. In: 24th annual convention of the southwestern psychological association, fort worth, tx; 1977. .

[9] Smith NJ, Levy R. The effect of word predictability on reading time is logarithmic. Cognition. 2013;128(3):302-19.

[10] Bizzoni Y, Lapshinova-Koltunski E. Measuring Translationese across Levels of Expertise: Are Professionals more Surprising than Students? In: NoDaLiDa; 2021. p. 53-63.

[11] Fillmore CJ. l968. The case for case." by E. Bach and R. Harms, l-88. New York: Holt, Rinehart & Winston, Inc; 1968.

[12] Fillmore CJ, et al. Frame semantics and the nature of language. In: Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech. vol. 280. New York; 1976. p. 20-32.

[13] Hamm F. Frame semantics. The cambridge encyclopedia of the language sciences, Cambridge: CUP. 2009.

[14] Minsky M. A framework for representing knowledge. MIT, Cambridge; 1974.

[15] Butt M. Theories of Case. Cambridge Textbooks in Linguistics. Cambridge University Press; 2006.

[16] Peng J, Feldman A, Vylomova E. Classifying idiomatic and literal expressions using topic models and intensity of emotions. arXiv preprint arXiv:180209961. 2018.

[17] Feldman A, Peng J. Automatic detection of idiomatic clauses. In: Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I 14. Springer; 2013. p. 435-46.

[18] Zeng Z, Bhat S. Idiomatic expression identification using semantic compatibility. Transactions of the Association for Computational Linguistics. 2021;9:1546-62.

[19] Amin M, Fankhauser P, Kupietz M, Schneider R. Shallow context analysis for german idiom detection. Proceedings of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021. 2021.

[20] Fankhauser P, Kupietz M. Count-Based and Predictive Language Models for Exploring DeReKo. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10); 2022. p. 27-31.

[21] Cook P, Fazly A, Stevenson S. The VNC-tokens dataset. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Citeseer; 2008. p. 19-22.

[22] Fazly A, Cook P, Stevenson S. Unsupervised type and token identification of idiomatic expressions. Computational Linguistics. 2009;35(1):61-103.

[23] Pradhan M, Peng J, Feldman A, Wright B. Idioms: Humans or machines, its all about context. In: Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I 18. Springer; 2018. p. 291-304.

[24] Xia P, Qin G, Vashishtha S, Chen Y, Chen T, May C, et al. LOME: Large ontology multilingual extraction. arXiv preprint arXiv:210112175. 2021.

[25] Rubenstein H, Goodenough JB. Contextual correlates of synonymy. Communications of the ACM. 1965;8(10):627-33.