Are idioms surprising?

J. Nathanael Philipp
Serbski Institut
August-Bebel-Straße 82
03046 Cottbus
Leipzig University
Augustusplatz 10
04109 Leipzig
nathanael@philipp.land

Michael Richter
Erik Daas
Leipzig University
Augustusplatz 10
04109 Leipzig
mprrichter@gmail.com
erik.daas.uni@outlook.de

Max Kölbl
Osaka University
1-5 Yamadaoka, Suita
565-0871 Osaka
max.w.koelbl@gmail.com

Abstract

This study focuses on the identification of English Idiomatic Expressions (IE) using an information theoretic model. In the focus are verb-noun constructions only. We notice significant differences in semantic surprisal and information density between IE-data and literals-data. Surprisingly, surprisal and information density in the IE-data and in a large reference data set do not differ significantly, while, in contrast, we observe significant differences between literals and a large reference data set.

1 Introduction

The aim of this study is the identification of English Idiomatic Expressions (IE) with an information theoretic model (Shannon, 1948). We focus solely on verb-noun constructions (VNC) such as *kick the bucket*, *make scene*, *blow whistle* or *take heart*. As in a study from Peng et al. (2018), we look at VNC which can be used either idiomatically or literally. In this study, we restrict ourselves to IE in English because we had manually annotated data available in which sentences are labelled as "idiomatic" or as "literal". We assume that the amount of information in general and the *Flow of Information* (FoI) in IE and literals differ from each other. We operationalise FoI as information density (see below subsection *Information Density*). Information density is calculated from the change of information over time in linguistic units such as sentences and utterances. The principle of *Uniform Information Density* in language production postulates the smallest possible information changes in a linguistic unit (preferably no steep information peaks and no deep information troughs) in order not to threaten the processing of the message by the receiver (Levy and Jaeger, 2007). In this study, we utilise contextualised information that is *surprisal* (Tribus, 1961; Hale, 2001; Hale et al., 2015; Levy, 2008)[1]. Surprisal represents the amount of certainty / uncertainty, i.e., it measures the deviation between what the language processor expects to occur and what actually occurs in a linguistic unit. We expect that idioms will cause a different amount of surprisal (over time) than literals do because we assume that literal meaning is the expected case, while IE is a deviation from that and will provide surprisal. That is, the information jumps in the sentence should be more pronounced with IE than with literals. In particular, we use *semantic surprisal* as the feature of words since it is derived from the topics in the environment of the target word, and to this end, we employ the *Topic Context Model* (TCM) (Kölbl et al., 2020, 2021; Philipp et al., 2022). TCM indicates how surprising a word is given its distribution in topics and given the distribution of topics in its environment, which can for instance be a document or the entire corpus. We motivate the use of the TCM to distinguish IE and literals by the assumption that the distributions of topics in either case differ which will cause significant differences in surprisal.

IE are far less subject to the principle of compositionality than literal expressions (Espinal and Mateu, 2019; Nunberg et al., 1994). IE are stable linguistic constructions, mostly with specific syntax as in *loose face* or *blow whistle*, a feature referred to as *(In)flexibility* (Espinal and Mateu, 2019; Nunberg et al., 1994). This feature also means the impermeability of IE, i.e., grammatical transformations, extractions and insertions lead to ungrammaticality. To understand an IE touches on conventional-

[1] For empirical evidence of surprisal, see i. a. (DeLong et al., 2005; Bentum, 2021)

Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), pages 149–154
September 18–22, 2023. ©2023 Association for Computational Linguistics

ity in language, since its meaning has evolved through specific language usage and convention. Espinal and Mateu (2019) emphasise that *[t]he meaning of IE involves metaphors, hyperboles, and other kinds of cognitive figure.*

2 Selected work on automatic detection of idiomatic expressions

To the best of our knowledge there is no work on IE within the framework of information theory. However, the following two studies described take the approach that is also taken in the present study, that the occurrence of IE is a semantic deviation from the expected. Peng et al. (2018) report an unsupervised classification of IE that is based on topic detection. The authors show that words that are highly relevant in the main topic of the discourse are not very likely to occur in IE, that is, IE are semantically distinct from the main topic of the discourse. In their point of departure, Peng et al. (2018) follow an earlier study by (Feldman and Peng, 2013) in which the authors state that IE are semantic outliers in a given context. This approach is also pursued in Zeng and Bhat (2021) where a BiLSTM-neural network is employed for the prediction of a token as idiom or literal. Basis are static and contextualised embeddings. To the former, additional information such as PoS-tags is added, and the enriched static embeddings are further combined with the contextualised embeddings. If a contextualized representation is semantically compatible with its context, is classified as literal, else it is an idiom. In both studies, IE classification is successful which is indicated by high precision, recall and accuracy values.

3 Dataset, concepts and technique of analysis

The dataset in the recent study comprises 1,997 sentences that are labelled as idioms and 535 sentences labelled as literals.[2] The sentences have been extracted from *British National Corpus* (BNC) and, in addition, from COCA, COHA and GloWbE[3] and served as data basis in Peng et al. (2018). For the determination of a VNC as IE or as a literal expression, Peng et al. (2018) used the list in Cook et al. (2008); Fazly et al. (2009). Peng et al. (2018) treated idiomacity as a binary and explicitly not as a gradual property (Pradhan et al., 2018), and this dichotomy is maintained in the present study.

3.1 Topic Context Model

TCM (Kölbl et al., 2020, 2021; Philipp et al., 2022) is an extended topic model, since it outputs surprisal based on genuine topic models. In this study, we employ *Latent Dirichlet Allocation* (Blei et al., 2003) (LDA).

TCM is built within the framework of *Surprisal Theory* (Hale, 2001; Jaeger and Levy, 2007). It calculates semantic surprisal of a word w given the distribution of topics its non-local environment, for instance a corpus, or in its local environments, for instance documents and paragraphs. Surprisal is defined as the negative log-conditional probability of w, as given in Formula 1.

$$\textit{surprisal} = \log_2 P(w|\text{CONTEXT}) \qquad (1)$$

We define the context as a topic calculated by LDA and calculate the average surprisal for each word, see Formula 2, where n is the number of topics of the LDA. We fixed this at 100 topics. The calculation is given in Formula 2.

$$\overline{\textit{surprisal}}(w_d) = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(w_d|t_i) \qquad (2)$$

The term $P(w_d|t_i)$ is the probability of a word w_d given a topic t_i in a document d, which is calculated according to Formula 3. $c_d(w)$ is the frequency of a word w given a document d, $|d|$ is the total number of words in the document d, WT is the normalized word topic distribution of the LDA[4], and $P(t_i|d)$ is probability of a topic t in a document d given by the LDA.

[2] The data were made available for us by Jing Peng and Anna Feldman.

[3] http://corpus.byu.edu, https://github.com/bondfeld/BNC_idioms

[4] model.components_ / model.components_.sum(axis=1)[:, np.newaxis] as suggested by https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

$$P(w_d|t_i) = \frac{c_d(w_d)}{|d|} WT_{w_d,t_i} P(t_i|d) \qquad (3)$$

We trained the LDA on a compilation of an English news corpus (from 2020) and an English Wikipedia (from 2016) corpus, with with 1M sentences each. Both corpora are taken from the *Wortschatz Leipzig* (Goldhahn et al., 2012)[5]. This compilation of two corpora forms the reference data set.

3.2 Information Density

We compare the flow of information in IE and literals utilising the concept of information density.

Formula 4 defines *local Uniform Information Density* (Collins, 2014) (UID, also termed *wordwise* Information density (Scheffler et al., 2023)) as the average of the squared change in surprisal from word-to-word in sentences. In Formula 4, it is not distinguished between increases and decreases in surprisal.

$$UID_{LOCAL} = -\frac{1}{n}\sum_{i=1}^{n}(id_i - id_{i-1})^2 \qquad (4)$$

UID_{LOCAL} is per definition negative (Jain et al., 2018), and therefore a UID_{LOCAL} value close to zero indicates a high uniformity of the information density distribution. A high UID value is close to zero and thus expresses, on average, small changes in surprisal in the flow of information in sentences.

4 Results

First, we run *Welch* tests (Welch, 1938) to check whether there are significant mean differences between the data for surprisal. A Welch test does not assume homogeneity of variances in the dataset that are compared. The sizes of the data sets vary considerably: the News-Wikipedia data set comprises 41, 284, 165 surprisal values, the IE set hat 48, 500, and the data set with literals has 11, 655 surprisal values. We observe significant differences of means between IE ($M = 30.26$, $SD = 8.12$) and literals ($M = 30.10$, $SD = 8.19$), i.e., $t = 2.19$, $p = .029$ and between the News and Wikipedia-training and reference data set ($M = 30.25$, $SD = 7.94$) and literals, i.e., $t = 2.23$ $p = 0.025$. Not significant is the difference of means between the News and Wikipedia data set and IE ($t = -0.40$, $p = 0.69$). Despite of a number significant mean differences as described above, the effect sizes that we determined by *Cohen'sd* (Cohen and Cohen, 1988) are consistently small in these cases. That is to say, idiomacity has not a strong effect on the information density: *Cohen's d* for the pair IE and literals yields .022, and for the pair News-Wikipedia and IE it yields .021.

Figure 1 depicts the distribution of the UID_{LOCAL}-values in complete sentences. Values close to zero represent small surprisal jumps in sentences. The x-axis gives the UID_{LOCAL}-values, the y-axis gives the normalised relative frequency of each value, and the area under each curve should be 1.

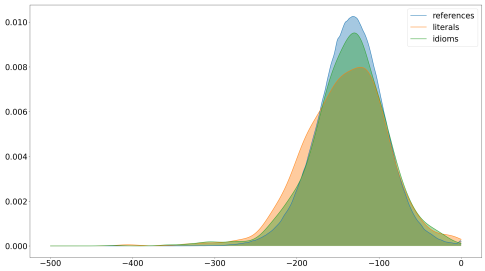

Figure 1: The density of the average surprisal change per word (UID) and **sentence** in the datasets. The x-axis depicts the average surprisal change, the y-axis depicts normalised frequencies of UID-values.

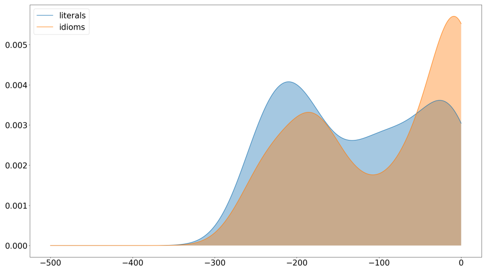

Figure 2: The density of the average surprisal change per word (UID) and **VNC** in the datasets. The x-axis depicts the average surprisal change, the y-axis depicts normalised frequencies of UID-values.

[5] https://wortschatz.uni-leipzig.de/de

The plots show that the distribution of the IE data takes the middle position between the distributions of the news wiki data and the literals. The News-Wikipedia training set forms the steepest peak and the most even distribution. In contrast, a strongly flattened peak and a distribution that buys out more to the left and right can be observed in the literals, while the IE data set occupies the middle position. As a next step, we focus solely on the VNC in IE and Literals data, in particular on the VNC-list of 12 constructions in Peng et al. (2018). The resulting data sets comprise 793 (IE, $M = 30.40$, $SD = 7.61$) and 637 (literals $M = 31.22$, $SD = 8.07$) surprisal values. A Welch test discloses a significant difference $t = -1.955$, $p-value = 0.05$. Cohen's d is now higher than in the comparisons above, that is .104. The corresponding UID_{LOCAL} density plots are given in Figure 2. The density peak of IE is closer to 0 than the one of the literals whose density is evenly distributed, indicating that information jumps tend to be smaller in IE.

5 Conclusion and discussion

Our study provides first evidence for differences in surprisal between IE and literals. This is reflected in the differences average level of the surprisal values and also in the flow of information (flow of surprisal), the determination of which we operationalised through the measurement of information density (UID_{LOCAL}). We conclude therefore that semantic surprisal from our TCM is a discriminating feature that distinguishes IE from literals. Our study is comparable with the precursor study (Philipp et al.): Here, surprisal was derived from POS tags and thematic roles which did not result in any differences between IE and literal expressions.[6]

Our study has the same point of departure as (Peng et al., 2018): we as well assumed that IE are deviations from the semantically expected, and so it seemed to be plausible to predict that sentences with IE deviate stronger than literals from the reference set w.r.t. the total amount of surprisal and the sentential information density.

However, this is not what we observe: surprisingly IE and the reference dataset exhibit smaller differences in surprisal and UID_{LOCAL}, respectively, than literals and the reference dataset do. Even with significant mean differences, there is only a low effect strength of surprisal. We attribute this outcomes to the fact that surprisal and information density over the entire sentence lengths are compared, i.e., we used a *global measure*, so to speak. It is all the more remarkable, however, that between IE and literals differences nevertheless emerge. In contrast, the *local measure*, which we applied solely to VNC within sentences, increases the effect size of surprisal considerably which underlines the classificatory power of the surprisal feature.

The observation that IE and the reference dataset hardly differ in terms of surprisal and information density indicates that the reference-set has a certain idiomatic character. Our assumption that IE are semantic outliers given a reference dataset has thus to be revised, rather we conclude that the reference dataset seems to have a considerable amount of IE. One important question for future research is whether this conclusion could be generalised: Does language in general tend to be more idiomatic or literal?

Limitations

The News and Wikipedia corpora are only composed of single sentences. However, the TCM is designed to calculate semantic surprisal of words from large extra-sentential contexts, which the corpora do not offer. Future work should thus be based on longer, coherent texts and documents when calculating the surprisal in order to make full use of the possibilities of the TCM. The results could thus become more valid, to which larger corpora as data base will also contribute, especially in the case of literals. In addition, it would be desirable if the study could be extended to other languages and thus take on a comparative character. However, this requires annotated corpora in order to train classification models, which is a desideratum for the future.

[6] The comparison with the results in (Feldman and Peng, 2013) and (Peng et al., 2018) who took a completely different approach is hard because the evaluation measures there differ from ours.

References

Martijn Bentum. 2021. *Listening with great expectations: A study of predictive natural speech processing*. Ph.D. thesis, [Sl]:[Sn].

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jacob Cohen and Jacob Willem Cohen. 1988. *Statistical power analysis for the behavioral sciences*, 2. ed. edition. Erlbaum, Hillsdale, NJ [u.a.]. Literaturverz. S. 553 - 558.

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of psycholinguistic research*, 43(5):651–681.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22. Citeseer.

Katherine A DeLong, Thomas P Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117–1121.

M Teresa Espinal and Jaume Mateu. 2019. Idioms and phraseology. In *Oxford Research Encyclopedia of Linguistics*.

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

Anna Feldman and Jing Peng. 2013. 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*, pages 435–446. Springer.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

John Hale. 2001. 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*, pages 1–8. Association for Computational Linguistics.

John Hale, David Lutz, Wen-Ming Luh, and Jonathan Brennan. 2015. Modeling fmri time courses with linguistic structure at various grain sizes. In *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, pages 89–97.

T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2018. Uniform information density effects on syntactic choice in hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48.

Max Kölbl, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. 2020. Keyword Extraction in German: Information-theory vs. Deep Learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI,*, pages 459–464. INSTICC, SciTePress.

Max Kölbl, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. 2021. The semantic level of shannon information: Are highly informative words good keywords? a study on german. In Roussanka Loukanova, editor, *Natural Language Processing in Artificial Intelligence - NLPinAI 2020*, volume 939 of *Studies in Computational Intelligence (SCI)*, pages 139–161. Springer International Publishing.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849.

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

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2018. Classifying idiomatic and literal expressions using topic models and intensity of emotions. *arXiv preprint arXiv:1802.09961*.

J. Nathanael Philipp, Max Kölbl, Erik Daas, Yuki Kyogoku, and Michael Richter. Perplexed by idioms. (in press).

J. Nathanael Philipp, Max Kölbl, Yuki Kyogoku, Tariq Yousef, and Michael Richter. 2022. One step beyond: Keyword extraction in german utilising surprisal from topic contexts. In *Intelligent Computing*, pages 774–786, Cham. Springer International Publishing.

Manali Pradhan, Jing Peng, Anna Feldman, and Bianca Wright. 2018. Idioms: Humans or machines, it's 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*, pages 291–304. Springer.

Tatjana Scheffler, Michael Richter, and Roeland van Hout. 2023. Tracing and classifying german intensifiers via information theory. *Language Sciences*, 96:101535.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Myron Tribus. 1961. Information theory as the basis for thermostatics and thermodynamics.

Bernard L Welch. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.