Semantic Information: A difference that makes a difference

Anonymous ACL submission

Abstract

In this paper, we introduce a new notion of semantic information for natural language. The key idea is as follows: a linguistic sign carries semantic information (SemI) about a document if it reduces a language processor's (LP) processing effort. We consider two systems, an informed one and an uninformed one, and describe SemI in their terms. Processing effort is quantified via surprisal where the informed system is 'aware' of the linguistic sign and the uninformed one is not. High surprisal means high cognitive effort for the language processor. On an English fairy tale corpus and on two German news corpora, we tested successfully the prediction that if the linguistic sign in question carries semantic information about the document, we measure lower surprisal, that is, higher values of SemI.

1 Introduction

When Claude Shannon (Shannon, 1948) published his remarkable theory of communication, his main goal was to account for the noise that distorts messages by using statistical methods. This makes Shannon information (SI) a type of information concerned with *form* rather than meaning. Hence, it is usually classified as a type of *syntactic information* (SynI), which is contrasted against *semantic information* (SemI). In fact, Shannon himself made sure to clarify that his work treats information purely quantitatively and makes no claims about the semantics of a sign.

This, however, did not stop others from using SI in an effort to come up with a theory of SemI. There are SI-backed theories of semantic information in formal logic (Carnap et al., 1952), philosophy/epistemology (Dretske, 1981; Floridi, 2004), and recently statistical physics (Kolchinsky and Wolpert, 2018). In this paper, we present and test a new notion of semantic information (SemI) for natural language which we contrast against SI. The basic idea is thus: Consider a system consisting of a document (such as a text, a corpus, or another form of linguistic production), and a language processor (LP). Then a token carries SemI about the document if it reduces the effort the LP has to expend while processing the document. Hence, we always consider two systems: an *informed* one and an *uninformed* one whose respective processing efforts we compare.

To quantify this, we rely on *Surprisal Theory* (Hale, 2001) which forms a subset of Shannonian Information theory. Hale's surprisal model deals with parser surprisal.

Here, the SI of a word given a context is interpreted at the *amount of surprise/surprisal* an LP experiences when encountering that word. Surprisal has been shown to be an empirically confirmed entity (e.g. DeLong et al. (2005); Bentum (2021); Heilbron (2022)): it correlates with processes and states in the brain which makes it a concept of psycholinguistics and the *philosophy of mind*. In (Levy and Jaeger, 2007), a proportional relationship between surprisal and cognitive effort is formally defined. All this taken together makes it sense to use surprisal to model processing effort.

For the calculation of surprisal in informed systems, we employ a variant of the *Topic Context Model* (TCM) Kölbl et al. (2020, 2021); Philipp et al. (2022, 2023a,b) (see Section 5).

TCM computes the *topic vector* of a given document and uses that to assign a surprisal value to every word. The type of TCM we use in this study is based on *latent dirichlet allocation* (LDA) (Blei et al., 2003). In this case, the topic vector assumes that there is a finite number of topics and that every topic is to some percentage present in any given document. Then, every word is to some percentage associated to every topic. For example, the words *AI* and *neural networks* have high probabilities in a topic like *computer science* but low probabilities in topics like *cooking*. Thus top-

ics (and consequently topic vectors) carry semantic significance. For the calculation of surprisal in uninformed systems, we use *lexical surprisal*, which is purely based on word frequencies (infrequent words are more surprising than frequent ones) based on large corpora without any regard for the semantics of a given document. Surprisal, which we derived from TCM, is what we call *semantic surprisal*, while unigrams provide *lexical surprisal*. Totion of semantic information that we utilise is based on the difference between semantic and lexical suprisal (see 3).

Our approach falls into the category of *optimisation-based* notions of SemI. To the best of our knowledge, this is the first endeavour of this type in the field of natural language processing. We test it on the basis of three corpora in English and German in the genres of prose and news.

Prominent models of SemI, such as the ones introduced in (Carnap et al., 1952; Dretske, 1981; Floridi, 2004; Venhuizen et al., 2019a), are based on the *Correspondence Theory of Truth* (CTT): propositions are classified as 'true' or 'false' by a characteristic function, given a model. These models, however, are not constituted from real world text data, but rather they are small 'toy models'. Thus the models of SemI known so far do not model the natural use of language. This is the link between our understanding of the concept SemI and linguistic meaning: because our approach is not within the framework of CTT, and our concept of SemI does not capture the relation of a linguistic sign to its referent in the world, we instead quantify the processing effort of an LP in language comprehension, which is a semantic task, but one pertaining to *relevance* rather than truth.

It should now have become clear that the relationship between SI and the meaning of linguistic units is not straightforward. As Chalmers (Chalmers, 1997, p. 261) points out, *Shannon was not concerned with a semantic notion of information, on which information is always information* about something. *Rather, he focused on a formal or syntactic notion of information, where the key is the concept of a state selected from an ensemble of possibilities.* Chalmers then describes the concept of an *information space* (Chalmers, 1997, p. 262) which comprises a number of *information states* and *difference relations between those states*, however there is [...]*not any interpretation of these states*. This is the reason why our concept of SemI is approaching the concept of meaning only indirectly. Our point of departure is that it can be measured how the processing effort is affected when the LP processes the output of a semantically informed system. However, meaning itself cannot be quantified. [1]

2 Points of departure and relevant work

In addition to the above-mentioned models by Kolchinsky and Wolpert (2018) and Hale (2001), further inspirations for our study are Dretske (1981), and Floridi (2004, 2009).

Dretske's and Floridi's approaches handle information differences between two distinct systems and distinguish meaningful and meaningless information. Floridi (2004, 2009) terms this difference *strongly semantic information*. Inspired by these works are, among others, the studies of (Feldman and Peng, 2013; Peng et al., 2018; Rubino et al., 2016; Venhuizen et al., 2019b) on idiom detection, translation-classification and predictive language comprehension, respectively. These studies have in common that information differences represent qualitative differences between a baseline condition and a special, surprising condition which are interpreted as a representation of a semantic difference. In (Feldman and Peng, 2013; Peng et al., 2018; Philipp et al., 2023a) for example, the baseline condition includes sentences that can be understood literally, while the surprising, deviant condition comprises idiomatic sentences.

3 Measuring semantic information

We start by defining a *system* as a language processor (LP) in an *environment* (a text/corpus/etc.). Following ideas by Hale (2001), we base our definition of SemI around the effort of the LP to process a text. The idea is thus: the more information (SemI) an LP has about a text, the less effort, i.e., SynI will be involved in parsing it. In the terms of surprisal theory, SemI reduces the amount of uncertainty for a languahe processor. Another term for SynI in this context is *surprisal*; a term that suggests that the quantity at hand measures how surprising it is for the LP to encounter a word.

In this view, SynI exists *within* an environment, while SemI is information *about* it. In particular,

[1] As indicated above and further explained below (Section 2 2), difference is a keyword in information theory. The title of our paper refers to a quote from Gregory Bateson. The condensate of his model of information is the claim that information is *a difference that makes a difference*

SemI is information about a text which helps *reduce* the surprisal that is measured while an LP parses the text. To be able to measure SemI thus requires us to compare two LPs: an informed and an uninformed one. We assess their respective processing efforts by first considering the probability space of context-prediction pairs they can process. This could be, for example, bigrams where one word is known to the LP and the other one is its prediction based on that. On this space, we get three different distributions: the (un-)informed distributions by the two LPs, and the *actual* distribution directly taken from the text.

The processing effort can then be measured by comparing the LPs' distributions with the third one. In the practical part of this paper, we used the *Kullback Leibler Divergence* (KL) (Kullback and Leibler, 1951), but other types of metrics, such as perplexity, may also work. The interpretation of KL is that it measures how well a distribution P is approximated by another distribution Q. In terms of surprisal, it can be interpreted as the *average surprisal* an LP will experience working in a P-distributed environment while expecting a Q-distributed one. We denote it $KL(P, Q)$. Values yielded by that operation are always non-negative and reach 0 when they are identical. It is in general not symmetric. We thus get the two values $KL(T, U)$ and $KL(T, I)$ where U and I denote the uninformed and informed distributions respectively, and T denotes the 'text' or 'true' distribution. We can compare them using either Formula 1 or Formula 2.

$$\begin{aligned} SemI &= -log_2 \frac{KL(T, I)}{KL(T, U)} \\ &= \log_2 KL(T, U) - \log_2 KL(T, I) \end{aligned} \tag{1}$$

$$SemI = KL(T, U) - KL(T, I) \tag{2}$$

We deliberately give two formulae as possible alternatives because they measure slightly different things, each with the potential of having a unique merit. Both formulae measure how effectively semantic information has influenced the system. However, Formula 1 can be regarded as the relative version of Formula 2. Both can be interpreted giving the average reduction in the cost of text processing per word. However, in Formula 2 we measure the exact number of reduced bits from the average encoding length of a word, whereas in Formula 1, we consider the reduced amount of information relative to the average code lengths given the uninformed distribution.

In stark contrast to Shannon information, in the present setting there may be instances where the information content is negative. This happens when the informed LP has to expend more effort into processing than the uninformed one. In such cases, we may speak of 'disinformation' or 'deception', wherein a misleading expectation complicates the processing.

We emphasise that the specifics of how an LP operates are not rigidly defined. This is on purpose since different applications may call for different notions of uninformedness or different modes of operation of the LP.

4 Data

To test our prediction, we used three corpora:

(i) an English fairytale corpus from INESC-ID Human Language Technology Lab [2](Lobo and De Matos, 2010) with 111 stories and a total of 83,845 unique words. The average number of words per fairytale is 270. Preprocessing includes removing of all punctuation and converting them to lowercase, the words were already lemmatised. We split the texts into 300 training texts and 110 test texts.

(ii) the *Heise* tech news (Philipp et al., 2022) corpus in German language consisting of 5,322 articles and a total of 449,609 unique words with an average of 280 words per document. Preprocessing included conversion to lowercase, removing all punctuation, and lemmatising. It was done using spaCy.

(iii) the *Frankfurter Allgemeine Zeitung* (FAZ) newspaper corpus in German language consisting of 20,924 articles and a total of 605,681 unique words with an average of 470[3] words per document. Preprocessing was identitcal to that of the heise corpus.

[2] https://www.hlt.inesc-id.pt/w/Fairy_tale_corpus

[3] While it is suggestive that the fact that all three word averages per document are powers of ten, we rounded only to the nearest integer which happens to be a power of ten every time.

5 Probability distributions and Workflow

5.1 The distributions

For every text, we need a total of three distributions: an *uninformed* one, an *informed* one, and the *actual* one. The uninformed distribution U has to be independent of the text, the informed one I has to depend on an informative token extracted from the text, and the actual one T is the real distribution of words in the text.

For the uninformed distribution, we choose for the probability function the relative frequency of every word in the training corpus. Before normalising however, we add 10^{-17} to every word, including those that do not make an appearance in the training corpus, so as to prevent a division by 0 when the KL-divergence is computed. Hence, the distribution is given by Formula 3.

$$P_U(w) = \frac{N + 10^{-17}}{\sum_{w \in \text{training and test corpus}(N+10^{-17})}} \quad (3)$$

Where N is the number of occurrences of w in the training corpus. For the calculation of surprisal in the informed distribution, we make use to the *Topic Context Model* (TCM) (Kölbl et al., 2020, 2021; Philipp et al., 2022, 2023a,b)[4]. The TCM is an extended topic model, calculating the probability of a word w given a distribution of *topics* for the text or corpus the word appears in. TCM utilises in this study the topic detection model *Latent Dirichlet Allocation* (Blei et al., 2003) (LDA). We initialised LDA with $n = 100$ topics and train it on the training corpus. This gives us for each topic a probability distribution $P(w_i|t_i)$ that indicates the probability a word is associated to a specific topic. We can define the *topic space* as the simplex $\{(x_1, x_2, \ldots, x_n) \in [0,1]^n | \sum x_k = 1\}$. Then for each document d, its *topic vector* v_d is an element of the topic space whose coordinates are given by the probabilities $P(t_i|d)$ that any given word in d is associated to topic t_i. Now the informed distribution for a word w given the topic vector v_d of a document is given by Formula 4.

$$P_I(w|v_d) = \sum_{i=1}^{n} P(w|t_i)P(t_i|d) \quad (4)$$

5.2 The informed distributions

To minimise the risk that our results are based on chance, we measured the informed distribution using four different topic vectors for each document. The first one is the **matching** vector, i.e., the vector TCM assigns to the document. The other three were the **fixed**, **random**, and **inverted** topic vectors which were chosen to deliberately 'mislead' the hypothetical LP. The fixed vector is the matching vector of one of the documents appearing in a corpus that is used indiscriminately for all documents. This vector was included because unlike the other two, it is a vector whose existence (and hence, plausibility) is established. The random vector is a randomly generated probability distribution over the topics. A different one is generated for each document. The inverted vector takes the matching vector of a document and reassigns the probabilities so that the n-th most likely topic becomes the n-th most *unlikely* topic. The prediction was that SemI calculated from the matching vector will be higher than the other three values since by assumption, the matching vector is the only one out of the four that prepares the LP with correct information about the topics.

5.3 Workflow

We compute P_U once at the beginning and then we compute for every document d in the test set four probability functions: P_T, $P_I^{(i)}$, $P_I^{(ii)}$, and $P_I^{(iii)}$. Here, P_T is the probability function of T. The other three are three different informed distributions, each computed with a different topic vector: $P_I^{(i)}$ uses v_d, i.e., the correct topic vector; $P_I^{(ii)}$ uses v_0, i.e., the topic vector of the first document in the test set; $P_I^{(iii)}$ uses a randomly generated element of the topic space; $P_I^{(iv)}$ uses the inverted probabilities in the topic vector in $P_I^{(i)}$.

Then we calculate $KL(T, U)$ and the four different versions of $KL(T, I)$. From these we calculate for each $KL(T, I)$ the pair of SemI measures given in Formulae 1 and 2.

6 Results

First, we wanted to know of whether there are significant mean differences in SemI between the groups 'matching', 'fixed', 'random' and 'inverted' (see above) in the three data sets. The question centres on whether the groups are taken from a common set or not. A brief explanation: the values express the amount of SemI of a group. In our model, positive values of SemI represent a reduction in surprisal, while negative values represent an increase. Figures 1, 2 and 3 illustrate that the

[4] https://github.com/jnphilipp/tcm

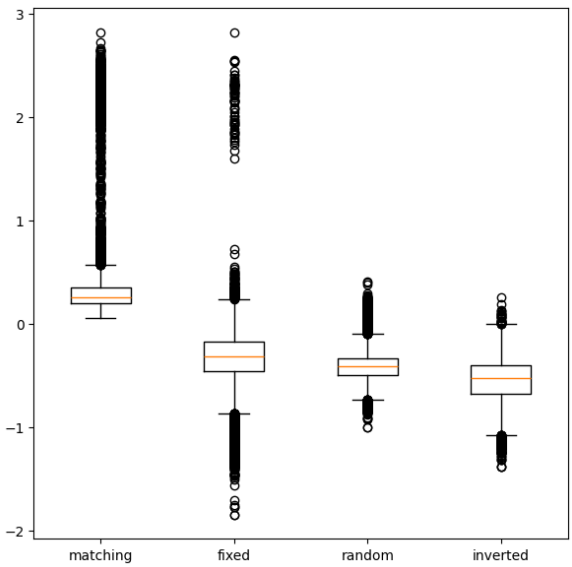

Figure 1: SemI calculated with Formula 1 for the FAZ corpus. The amount of SemI can be read off the y-axis.

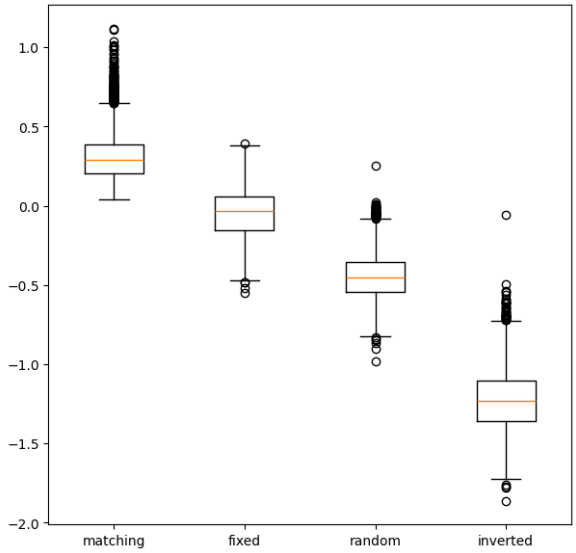

Figure 2: SemI calculated with Formula 1 for the Heise corpus. The amount of SemI can be read off the y-axis.

meaningful topic contexts of the group 'matching' cause a reduction of surprisal. In the Fairy tale and Heise corpora, the amount of SemI gradually decreases from 'fixed' to 'random' to 'inverted' whereby in the Fairy tale corpus, 'random' and 'inverted' show decrease of SemI. The FAZ corpus shows a dichotomy: only 'matching' yields SemI, while in the remaining groups, we observe a loss of SemI.

Because the four groups are not independent and, in addition, not normally distributed, we employed the non-parametric Friedman test for the comparison of means. The test statistic of the Friedman test has approximately a chi-square distribution. Table 1 displays the results which express high significant differences between the groups in all corpora: the six possible pairwise post-hoc test 'matching-fixed','matching-random', 'matching-inverted', 'fixed-random', 'fixed-inverted' and 'random-inverted' yielded high significant results ($p \approx 0$). The FAZ corpus has by far the highest chi square value, which indicates the strongest group differences.

Table 1: Results of the Friedman test.

Corpus	Chi-squared (χ^2)	df	p-value
Fairy tale	329.7	3	≈ 0
Heise	691.22	3	≈ 0
FAZ	45036	3	≈ 0

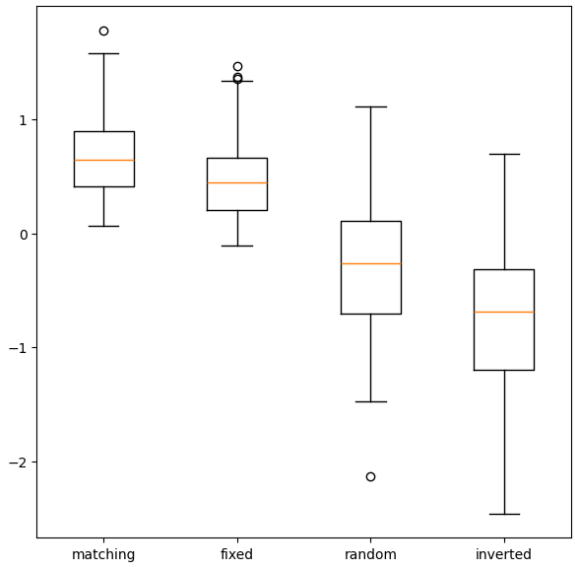

Figure 3: SemI calculated with Formula 1 for the fairy tale corpus. The amount of SemI can be read off the y-axis.

7 Discussion and conclusion

Across all corpora, we observe that the matching vectors carry the largest amount of semantic information, followed by the fixed, then random, and then inverted vectors. This shows that the matching vectors give the most accurate semantic information about its document. Fixed vectors being next in line can be explained by the fact that some documents are similar to one another. Hence, if a document is similar to the one the fixed vector is taken from, it will carry higher semantic information. The mysterious value gap observable in the FAZ corpus, however, cannot be explained by this and will be subject to future research.

In the case of the random vectors, it is a matter of chance whether or not they end up describing their respective documents well, but the inverted vectors are specifically designed to mislead the LP. Hence, they always come last.

It can also be observed that among the fixed, random, and inverted topics, the SemI values are sometimes even negative. That means that the Kullback-Leibler divergence is *higher* in the informed system than in the uniformed system, meaning that syntactic surprisal does a better job setting the LP's expectations than semantic surprisal, if the underlying semantics are faulty. This goes to show that SemI can be utilised as a measure of model evaluation since the results show that our information model TCM works. It discloses systematic differences between the groups 'matching', 'fixed', 'random' and 'inverted' since the changes in surprisal are, as shown, not due to chance.

As we already pointed out however, the connection between surprisal and semantics is not straightforward. The reduction of surprisal can only give an indirect indication of semantics: for text comprehension, a high degree of SemI ensures a low processing effort, that is, the LP has to process not as much new information. From this we conclude that the LP has got some prior semantic information about the text, and that this is why SemI increases the certainty in language processing and language comprehension: although, in this study, we restricted ourselves to computing the SemI values of given informing (or disinforming) tokens, the results indicate this method's potential for applications to knowledge extraction. Among a set of tokens, the one with the highest semantic information may reveal useful knowledge about the underlying text. This could be used, for example, for measuring the quality of a set of extracted keywords: a set of keywords is of good quality if it *prepares* the LP for the text, resulting in lower processing effort.

Limitations

(i) Theoretical: our concept of semantic information captures the meaning of natural language only indirectly. Also, it derives information from purely frequency-based contexts and does not make use of knowledge of the world a human language processor typically has and leverages.

(ii) Methodological: due to memory limits, we had to base our determination of SemI on relatively small corpora which might restrict the empirical validity and the analytical significance of our findings.

(iii) Empirical: in the frame of this pilot study, we were unable to empirically test our predictions regarding the reduction in processing effort with human test subjects.

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.

Rudolf Carnap, Yehoshua Bar-Hillel, et al. 1952. An outline of a theory of semantic information.

David J Chalmers. 1997. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

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.

Fred Dretske. 1981. *Knowledge and the Flow of Information*. MIT Press.

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.

Luciano Floridi. 2004. Outline of a theory of strongly semantic information. *Minds and machines*, 14:197–221.

Luciano Floridi. 2009. Philosophical conceptions of information. In *Formal theories of information: From Shannon to semantic information theory and general concepts of information*, pages 13–53. Springer.

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.

Micha Heilbron. 2022. *Getting ahead: Prediction as a window into language, and language as a window into the predictive brain*. Ph.D. thesis, [Sl]:[Sn].

Artemy Kolchinsky and David H Wolpert. 2018. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface focus*, 8(6):20180041.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

Paula Vaz Lobo and David Martins De Matos. 2010. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *LREC*, volume 10, pages 1472–1475.

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. 2023a. Perplexed by idioms? In *Knowledge Graphs: Semantics, Machine Learning, and Languages*, pages 70–76. IOS 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.

J. Nathanael Philipp, Michael Richter, Erik Daas, and Max Kölbl. 2023b. Are idioms surprising? *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, California. Association for Computational Linguistics.

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

Noortje J Venhuizen, Matthew W Crocker, and Harm Brouwer. 2019a. Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, 56(3):229–255.

Noortje J Venhuizen, Matthew W Crocker, and Harm Brouwer. 2019b. Semantic entropy in language comprehension. *Entropy*, 21(12):1159.