Semantic Information: A Difference that Makes a Difference

Anonymous ACL submission

Abstract

In the framework of distributional semantics, we introduce a novel notion and operationalisation of semantic information for natural language. The key idea is as follows: a linguistic sign carries semantic information about a document if it reduces the amount of surprisal for a language processor. We consider two systems, an informed one and an uninformed one, and describe semantic information 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. 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 pre-information through semantic surprisal, the current level of surprisal for the language processor is reduced. The conclusion is that the degree of semantic information results from the degree of semantic prior information.

1 Introduction

Semantics of natural language can be captured through computational methods, as exemplified by Firths famous assertion: 'You shall know a word by the company it keeps'(Firth, 1957). This is the guidung principle that underpins *distributional semantics* (Harris, 1954; Turney and Pantel, 2010; Mikolov et al., 2013), which models linguistic meaning based on co-occurrence patterns in large corpora. Our study is carried out in this theoretical framework, using statistical models to derive semantic effects from linguistic data.

Surprisal is a key concept in psycholinguistics, introduced to model human sentence processing through information-theoretic means (Hale, 2001; Jaeger and Levy, 2007). Surprisal theory posits that processing difficulty is proportional to the unexpectedness of a linguistic unit such as a word in context and to the effort required to process the linguistic unit. At this point it is already important to emphasise that we strictly separate the concepts of semantic surprisal and semantic information because, as we will show, in our model semantic surprisal is the prerequisite for the determination of semantic information (we will abbreviate our concept of semantic information as *SemI* in the following). In the corse of this paper, we employ the *Topic Context Model* (TCM) (Kölbl et al., 2020, 2021; Philipp et al., 2022, 2023a,b) that extends the distributional approach by incorporating topic distributions, providing a framework for computing *semantic surprisal*.

Empirical studies confirm surprisal effects in reading times, eye movements, and neural responses (Boston et al., 2008; Levy, 2008; Demberg and Keller, 2008; Roark et al., 2009; Levy, 2011; Monsalve et al., 2012; Smith and Levy, 2013; Brouwer et al., 2021; Bentum, 2021). Our model aligns with this research by computing semantic surprisal to examine its role in language comprehension.

Shannons information theory (Shannon, 1948) was designed to quantify information transmission and describes an optimal code for information compression. It is a model of the degree of certainty (and uncertainty) within any systems, but Shannon himself explicitly did not want to see his model applied to the semantics of natural language. But Shannons theory is probabilistic, and information in this view is a context-dependent entity. That is, the inherently distributional nature of Shannon information makes it a candidate for semantic modeling, despite the (initial) reservations of its creator.

Several approaches have applied information theory to semantics of language, including formal logic (Carnap et al., 1952), epistemology (Dretske, 1981; Floridi, 2004), and statistical physics (Kolchinsky and Wolpert, 2018). The mentioned models are only weakly empirical and closely related to the *Correspondence Theory of*

Truth. In these models, the philosophically controversial postulate applies that for a proposition to have information, the proposition must be 'true' in a model of the world. In contrast, our model is strongly empirically grounded, as large corpora serve as the basis for the computation of SemI. Furthermore, our model is not model-theoretic and not truth functional.

Influential philosophical perspectives on information and meaning see information represented by a 'difference' prominently expressed by Batesons notion of *a difference that makes a difference* (Bateson, 2000) and Chalmers' distinction between formal and semantic information (Chalmers, 1997). These perspectives highlight the relationship between information, knowledge, and cognitive processing, reinforcing our computational approach to SemI:

Our model quantifies SemI as the reduction of surprisal in an informed language processor (LP). We compare surprisal in informed and uninformed systems, distinguishing between surprisal based on word frequency and semantic surprisal that is derived from contextual topic distributions through TCM which we introduced above.

This framework allows us to test the hypothesis that **SemI reduces surprisal**. From this it follows that SemI will facilitate language comprehension. That is to say, SemI denotes a difference in the level of knowledge of the language processor.

Recall that surprisal is a cognitive quality, and its reduction is a process that we assume to apply to Bateson's dictum from above (Bateson, 2000)[1]. In our study we exploit English and German corpora, making a novel contribution to the computational analysis of semantic information. According to Shannon, maximum disorder means maximum entropy and maximum uncertainty. The supply of information to a system leads to a reduction in uncertainty. This is the basis of our SemI-model, which manifests itself as a reduction of surprisal.

The structure of this paper is as follows: Section 2 outlines prior research on semantic information and its relationship to information states. In Section 3 we present the methodology for measuring semantic information, introducing the surprisal-based approach and the use of probabilistic models, and in Section 4 we describe the datasets used in the study, along with preprocessing steps. Section 5 details the probability distributions and the workflow for computing semantic information, distinguishing between informed and uninformed language processors. Section 6 reports the results and finally, Section 7 provides a discussion and conclusion, interpreting the findings in relation to semantic information theory and potential applications.

[1]As indicated above and further explained below (Section 2). Bateson's notion of information is not only about reducing uncertainty, but is inseparable from how it affects the system.(Bateson, 2000).

2 Points of departure and relevant work

Inspirations for our study are Chalmers (1997); Tononi (2004); Floridi (2004, 2009). The concept of 'difference' is central to these works. Chalmers (Chalmers, 1997) sketches a model of consciousness in which an information space is a structure with information states and differences between them. For example, if there were only the two information states 0 and 1, we could regard '0' as an uninformed state and '1' as an informed state, and the difference as SemI in the sense of our models. In *Integrated Information Theory* (Tononi, 2004), a transition matrix describes the transition from one state to another information state.

Floridi's non-modeltheoretic approach handles information differences between two distinct systems and distinguish meaningful and meaningless information Floridi (2004, 2009). This difference is termed *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., 2019) on idiom detection, translation-classification and predictive language comprehension, respectively. Here, too, 'differences' are central: these approaches and studies have in common that differences between information states and systems 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

Let us imagine the following situation: we have a language processor (LP) which processes texts word by word. Each new word creates more or less processing effort for the LP. For the sake of

simplicity, we will assume that the processing effort of each word is constant, regardless of where and how often a word appears in a text. This assumption, however, is non-essential and can easily be dropped. An LP with this property can be modeled with a probability distribution P on the set of all words, which we will denote by X.

In such a situation, we can model the processing effort for a word w as the *surprisal* of that word (see Formula 1).

$$S(w) = -\log_2 P(w) \tag{1}$$

Given a document $d = (w_1, w_2, \ldots, w_k)$ consisting of k not necessarily distinct words, we can consider the *average processing effort* of d. Notice that this is coincides with the *cross-entropy* of the inner distribution P of the LP relative to the 'true' distribution T that describes the relative frequency of every given word inside the document d (see Formula 2).

$$\begin{aligned} \bar{S}(d) &= -\frac{1}{k}\sum_{i=1}^{k} \log_2 P(w_i) \\ &= -\sum_{w \in X} T(w) \log_2 P(w) \\ &= H(T, P) \end{aligned} \tag{2}$$

In particular, if $T = P$, we get exactly the *entropy* of the distribution T. Within the context of our model, this coincides with the lowest possible average processing effort an LP can experience: the internal probability distribution P of the LP is taylormade to fit the 'true' distribution T. We say that in this case, the LP has *full information* about d, since in practice, it is impossible to just guess T without knowing it.

Recall that the *Kullback-Leibler-Divergence* (KL) of a pair of distributions T and P is given as the difference of the cross-entropy $H(T, P)$ and the entropy $H(T, T)$. As such, it measures the 'coding inefficiency' of P on a T-distributed set. Within the context of our model, $KL(T, P)$ measures how much *surplus* in processing effort the LP has to exert in order to process d, relative to the optimal value. In particular, if the LP has full information, we get $KL(T, T) = 0$.

3.1 A flexible LP

Let us now assume that our LP has more than one setting: depending on a piece of information about the text, the LP is capable of anticipating the words it is going to encounter. In other words, it stores more than one probability distribution. Suppose we have $n + 1$ different distributions, denoted by $U, I_1, I_2, \ldots, I_n$, where U is the default distribution (U stands for 'uninformed') and each of the I_i corresponds to a specific *topic* τ_i (I stands for 'informed').

The distribution U is generic in the sense that it does not make any assumptions about the composition of d. In practice, it could be obtained by counting the words in a large and diverse corpus, like the British National Corpus. On the other hand, the distribution I_i makes an assumption about d, namely that its content belongs to the topic τ_i (for the sake of example, let τ_i indicate the topic of biology). It assigns higher probabilities to words that are associated to the topic τ_i (in our example, words like 'animal', 'plant', 'metamorphosis', etc) and hence cause the LP to experience a lower average processing effort if d contains more of these words. In practice, I_i could be obtained by counting the words in a specialised corpus that contains exclusively texts belonging to topic τ_i.

In this setup, we call the τ_i *carriers of semantic information* about d. Note that in general, the τ_i need not be topics. Their precise interpretations depend on the LP and the way it adjusts its processing strategy.

Our aim is to quantify the *semantic information content* that each τ_i carries about d. For that matter, we propose Formula 3.

$$\begin{aligned} SemI(\tau_i) &= -\log_2 \frac{KL(T, I_i)}{KL(T, U)} \\ &= \log_2 KL(T, U) - \log_2 KL(T, I_i) \end{aligned} \tag{3}$$

In words, Formula 3 measures the relative reduction of the surplus in processing effort obtained by supplying the LP with the piece of semantic information τ_i compared to the surplus in processing effort when no additional piece of information is given. Under the assumption that τ_i does reduce processing effort at all, the fraction lies between 0 and 1, which we project to the set of all non-negative reals by taking the negative logarithm.

Note that it is possible to make changes to Formula 3. For example, one could omit the logarithm like in Formula 4 or switch out the KL-divergence for the cross-entropy like in Formula 5, or do both (which yields Formula 4 again).

$$SemI'(\tau_i) = KL(T,U) - KL(T,I_i) \quad (4)$$

$$\begin{aligned} SemI''(\tau_i) &= -\log_2 \frac{H(T,I_i)}{H(T,U)} \\ &= \log_2 H(T,U) - \log_2 H(T,I_i) \end{aligned} \quad (5)$$

Either modification yields different behaviour. For example, the logarithm in Formula 4 looks only at the *absolute* improvement of the surplus in processing effort, while Formula 3 looks at the *relative* improvement. At the same time both Formulae 3 and 4 focus on the surplus in processing effort alone, while Formula 5 accounts for processing effort in its entirety.

However, regardless of the differences, it is easy to see that all three behave essentially the same:

1. if I_i approximates T much better than U does the values are high,

2. if I_i approximates T slightly better than U does, the values are low,

3. if I_i approximates T worse than U does, the values are negative.

In our experiments, we choose Formula 3 because it highlights the relative improvement of the surplus. In fact, one unit of semantic information content computed this way corresponds to a reduction of the surplus processing effort by the factor 2. See Figure 1 for a proof of concept.

3.2 Misinformation and disinformation

As mentioned before, it can happen that I_i approximates T worse than U does. In these cases, $SemI(\tau_i) < 0$ and we call τ_i a *carrier of semantic misinformation*. A semantic carrier of misinformation gives an LP a false sense of the type of text it is going to process, thereby increasing its average processing effort.

In research on the detection of fake news, the terms 'misinformation' and 'disinformation' both describe false information, but the latter carries the connotation of deliberate deception. Since our model does not capture the intent with which a carrier of semantic misinformation was given to the LP, we use the neutral term.

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.

5 Probability distributions and Workflow

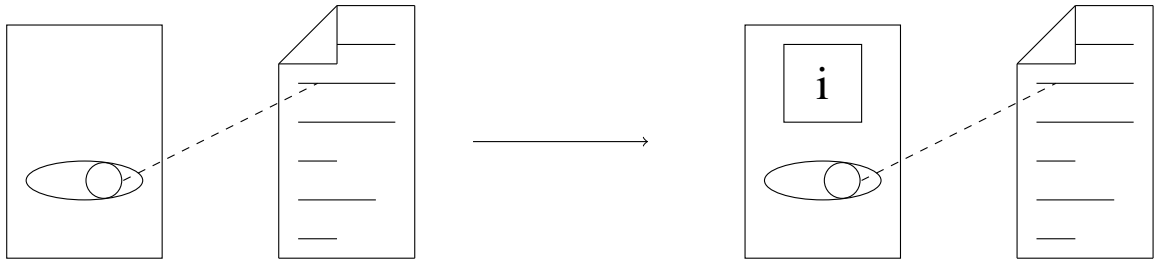
5.1 The distributions

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 6.

[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.



$f_U(w_1) = 0.20$	$f_T(w_1) = 0.25$	$f_I(w_1) = 0.27$	$f_T(w_1) = 0.25$
$f_U(w_2) = 0.10$	$f_T(w_2) = 0.25$	$f_I(w_2) = 0.23$	$f_T(w_2) = 0.25$
$f_U(w_3) = 0.30$	$f_T(w_3) = 0.20$	$f_I(w_3) = 0.21$	$f_T(w_3) = 0.20$
$f_U(w_4) = 0.05$	$f_T(w_4) = 0.15$	$f_I(w_4) = 0.12$	$f_T(w_4) = 0.15$
$f_U(w_5) = 0.20$	$f_T(w_5) = 0.10$	$f_I(w_5) = 0.11$	$f_T(w_5) = 0.10$
$f_U(w_6) = 0.15$	$f_T(w_6) = 0.05$	$f_I(w_6) = 0.06$	$f_T(w_6) = 0.05$

Figure 1: A document consisting of the words w_1 through w_6 whose distribution is given by f_T is processed first by an uninformed LP which estimates a distribution f_U. Then the same document is processed by an LP that holds a piece of information about it. It estimates another distribution f_I which much more closely resembles f_T.

$$P_U(w) = \frac{N + 10^{-17}}{\sum_{w \in \text{training and test corpus}} (N + 10^{-17})} \quad (6)$$

Where N is the number of occurrences of w in the training corpus. In this study, the *Topic Context Model* (TCM) (Kölbl et al., 2020, 2021; Philipp et al., 2022, 2023a,b)[4] utilises the topic detection model *Latent Dirichlet Allocation* (Blei et al., 2003) (LDA). We initialise 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 7.

$$P_I(w|v_d) = \sum_{i=1}^{n} P(w|t_i)P(t_i|d) \quad (7)$$

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 el-

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

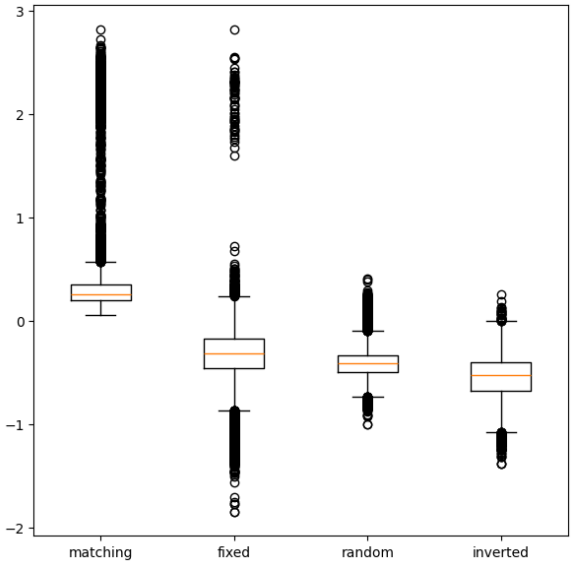


Figure 2: SemI calculated with Formula 3 for the FAZ corpus. The amount of SemI can be read off the y-axis.

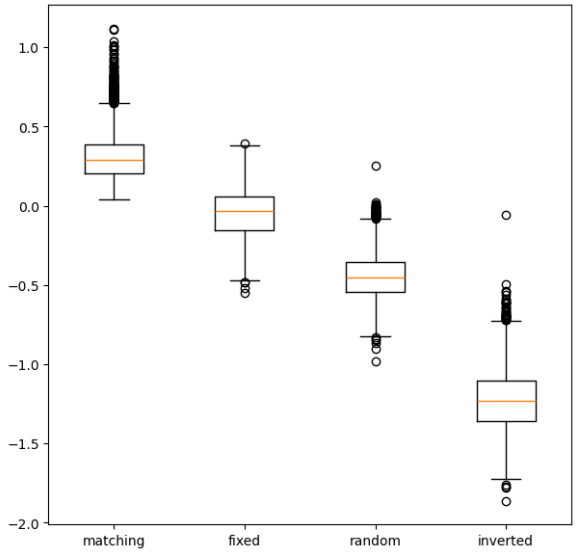


Figure 3: SemI calculated with Formula 3 for the Heise corpus. The amount of SemI can be read off the y-axis.

ement 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 3 and 4.

6 Results

First, we wanted to know 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 2, 3 and 4 illustrate that the 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

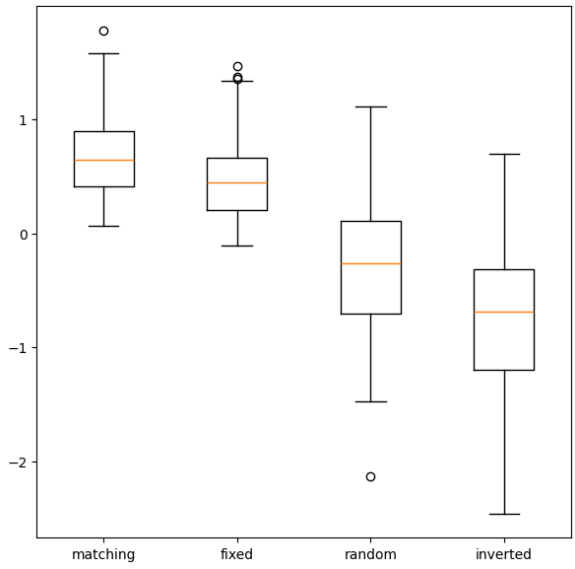


Figure 4: SemI calculated with Formula 3 for the fairy tale corpus. The amount of SemI can be read off the y-axis.

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

7 Discussion and conclusion

Within the test-setting of our study on the empirical application of our model we see our hypothesis confirmed that **SemI reduces surprisal**.

Across all corpora, we observe that the matching topic vectors carry the largest amount of SemI, followed by the fixed, then random, and then inverted vectors. This shows that the matching vectors give the most accurate SemI about its document. Fixed vectors being next in line can be explained by the fact that some documents are similar to one another w.r.t. their topics. Hence, if a document is similar to the one the fixed vector is taken from, it will carry higher SemI. 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 surprisal from the relative frequency of words in the training corpus 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

Gregory Bateson. 2000. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago press.

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.

Marisa F. Boston, Shravan Vasishth, Richard L. Lewis, and Hiroko Drenhaus. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Harm Brouwer, Francesca Delogu, Noortje J Venhuizen, and Matthew W Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

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.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. In *Proceedings of the 11th International Conference on Cognitive Modeling (ICCM)*, pages 213–218. Routledge.

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.

JR Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.

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.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2):146–162.

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.

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

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. 2011. Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119.

Irene F. Monsalve, Roger Levy, and Edward Gibson. 2012. Frequency and surprisal in predictive human sentence processing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 103–111. Association for Computational Linguistics.

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)*.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Murat Kzlkaya. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 324–333. Association for Computational Linguistics.

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.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Giulio Tononi. 2004. An information integration theory of consciousness. *BMC Neuroscience*, 5:42.

Peter D. Turney and Patrick Pantel. 2010. *Distributional Semantics*. The Handbook of Computational Linguistics and Natural Language Processing, Oxford.

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