An Analysis of Surprisal Uniformity in Machine and Human Translations

Josef Jon Charles University jon@ufal.mff.cuni.cz

Ondřej Bojar Charles University bojar@ufal.mff.cuni.cz

Abstract

This study examines neural machine translation (NMT) and its performance on texts that diverege from typical standards, focusing on how information is organized within sentences.

We analyze surprisal distributions in source texts, human translations, and machine translations across several datasets to determine if NMT systems naturally promote a uniform density of surprisal in their translations, even when the original texts do not adhere to this principle. The findings reveal that NMT tends to align more closely with source texts in terms of surprisal uniformity compared to human translations. We analyzed absolute values of the surprisal uniformity measures as well, expecting that human translations will be less uniform. In contradiction to our initial hypothesis, we did not find comprehensive evidence for this claim, with some results suggesting this might be the case for very diverse texts, like poetry.

1 Introduction

Natural language processing tools based on machine learning, such as machine translation, autocorrect, predictive typing, search, and text generation, have become integral to our daily lives. With the advancement of Large Language Models (LLMs), it's anticipated that interacting with these technologies will become a critical aspect of our

work and societal engagement. However, numerous questions about these technologies persist. In this work, we take a look specifically at neural machine translation (NMT) and at one such question: How well do these tools work on an input that is different from a typical text, not in terminology or domain, but in a way the information content is organized within an utterance? Are there any biases within the algorithms themselves that can be beneficial for ordinary types of texts, but harmful for specific cases that deviate from the usual rules found in mundane text content?

We propose that Neural Machine Translation (NMT) will be more effective with texts adhering to the Uniform Information Density (UID; (Levy and Jaeger, 2006)) hypothesis, meaning that the level of surprisal is consistently spread out throughout the sequence. One of the culprits could be the beam search decoding, which has been shown to adhere to the UID principle (Meister et al., 2020), i.e. even if the input has a diverse distribution of surprisal, the distribution in the translation will be more uniform. The UID-enforcing property of beam search has been shown as the key to its ability to produce high-quality, human-like texts (compared to exact search under the same model), even being dubbed the *beam search blessing* by (Meister et al., 2020).

We hypothesize that while in general, this property is positive, there are use-cases where too much emphasis on the uniformity of suprisals hurts the final translation quality. In this work, we look for such examples by comparing distributions of suprisals over source texts, machine translation and human translation across multiple test sets.

^{© 2024} The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

2 Related work

In this section, we will list the work exploring the presence of the UID principle in human-produced language as well as its presence and links to MT algorithms.

There is an extensive body of psycholinguistic work concerning the relationship between text predictability or surprisal and reading comprehension. The results on whether the effects of surprisal on reading comprehension is linear or super-linear (which would be consistent with the UID hypothesis) are mixed: For example, (Meister et al., 2021; Hoover et al., 2023) found support for super-linear relationship.

One of the most recent and largest studies (Shain et al., 2024) uses a wide array of open-source datasets, new Large Language Models for the surprisal estimation (GPT-3) and novel evaluation methods (deep learning based non-linear regression for analyzing continuous-time systems (Shain and Schuler, 2023). Their findings support a linear relationship between word surprisal and sentence reading times, suggesting that any pressure for UID seen in natural language is not motivated by an easier comprehension.

(Meister et al., 2020) ask why, empirically, beam search produces higher quality outputs than exacted search under the same model. To find the inductive bias embedded in beam search that allows this, they reverse engineer the objective that beam search is a solution for. They found that beam search can be reformulated as an exact search with a uniformity regularizer which enforces UID and that this property is the key to its effectivity. (Wei et al., 2021) employ a similar regularizer in the training of the model, which led to improved translation quality.

3 Methods

This section introduces the measures we use to operationalize the surprisal distribution uniformity concept, closely following the definition by (Meister et al., 2021).

3.1 Uniform information density

Surprisal theory, as outlined by (Hale, 2001), establishes a direct relationship between cognitive effort and the surprisal value of words; in other words, the effort required to comprehend a word is directly proportional to its level of predictability within a given context. To elaborate, for any given

utterance, denoted as u and consisting of elements (e.g. words) u_n , the surprisal of each element can be calculated as $s(u_n) = -\log p(u_n|\mathbf{u}_{\leq n})$, i.e. negative log-probability of the word given the previous context. Therefore, the total cognitive effort needed can be represented as

$$
\mathrm{Effort}(u_n) \propto s(u_n)
$$

Suppose we apply the same approach to a longer sequence of words, such as a sentence. In that case, we arrive to a counter-intuitive conclusion: If the surprisal of the sentence is a sum of surprisals of particular words and this sentence-level surprisal is predictive of processing effort (e.g. reading time), then any way of distributing the information across the utterance is the same in terms of the effort needed for comprehension.

To address this counter-intuitive consequence, the Uniform Information Density theory (UID) suggests a super-linear relationship between the surprisal levels of units and the total effort involved, incorporating the length of the utterance, denoted as N, into its framework (Aylett and Turk, 2004; Fenk and Fenk-Oczlon, 1980; Levy and Jaeger, 2006; Bell et al., 2003; Genzel and Charniak, 2002):

$$
\text{Effort}(\mathbf{u}) \propto \sum_{n=1}^{N} s(u_n)^k + c \cdot N, k > 1
$$

This definition suggests that utterances with a more uniform distribution of surprisal are simpler for human comprehension, indicating a preference for evenly spreading surprisal to effectively communicate a message.

We will demonstrate the intuitive concept of surprisal uniformity on the following two sentences:

- A) More uniform: *"When she got home after a long day at work, she decided to relax by reading her favorite novel and having a cup of tea."*
- B) Less uniform: *"London's annual festival was filled with activities, food stands, windsurfing, and drinks, but the sudden unveiling of a Yetti statue caught everyone's attention."*

Most people would consider the second sentence as more surprising, some of the words feel unexpected. We show the surprisal profiles of both sentences in Figure 1. Indeed, we can see that the

Figure 1: Surprisal behavior for the two examples sentences, measured by GPT-2 model.

surprisal behavior of the second sentence (orange) looks less uniform.

To express the uniformity as a measurable quantity, we experiment with multiple formulas, like Local Variance (LV), Coefficient of Variation (CV) , Global Variance (GV), Gini coefficient, and Super-linear Relationship (SL), and super-linear syntactic log-odds ratio (SLOR, (Kann et al., 2018; Pauls and Klein, 2012)):

• LV(**u**) = $\frac{1}{N-1} \sum_{n=2}^{N} (s(u_n) - s(u_{n-1}))^2$ • $\text{CV}(\mathbf{u}) = \frac{\sigma(\mathbf{u})}{\mu(\mathbf{u})}$

•
$$
GV(\mathbf{u}) = \frac{1}{N} \sum_{n=1}^{N} (s(u_n) - \mu(corpus))^2
$$

•
$$
SL(\mathbf{u}) = \frac{1}{N} \sum_{n=1}^{N} s(u_n)^k \quad (k > 1)
$$

• SLOR(u) =
$$
\frac{1}{N} \sum_{n=1}^{N} s(u_n)^k
$$
 -
\n $s_u(u_n)^k$ $(k > 1)$

Function s denotes surprisal in of a word in context, s_n is a unigram, context-free surprisal.

3.2 Surprisal distribution and translation

We hypothesize that the uniform distribution of surprisal is implicitly enforced by algorithms used for training and decoding in NMT, most prominently by the beam search (see (Meister et al., 2020) for the rationalization). In practical terms, we suggest that source texts characterized by highly uneven surprisal distributions would maintain such distribution upon translation by a human, but translation by MT engine would result in a more uniform distribution. We conducted measurements across multiple datasets, employing the uniformity measures described in the previous section.

4 Results

This section describes the settings and presents the results of measuring difference in surprisal uniformity in human and machine translations.

4.1 Models and Datasets

The first part of our investigation into the uniformity of surprisal across source texts, human translations, and machine translations focuses on the English-French language pair. We utilize a diverse set of corpora: the Books corpus (Zhu et al., 2015) (*books*), Global Voices (Nguyen and Daumé III, 2019) (*global*), Newstest2014 (Bojar et al., 2014) (*wmt*), and a French translation of a poem by Oscar Wilde translated by Jean Guiloineau (*wilde*). For the next set of experiments, involving multiple reference translations in English-Czech direction, we draw upon the dataset provided by (Zouhar and Bojar, 2024; Zouhar et al., 2023), which we refer to as *ORT*.

For comparing surprisals in English to French translations, we turn to BLOOM-1B7 (BigScience Workshop, 2022) for our estimates. For the analysis involving Czech translations with multiple references, surprisal estimates are obtained from MU-NLPC/CzeGPT-2 (Hájek and Horák, 2024) and BUT-FIT/Czech-GPT-2-XL-133k (Fajčík et al., 2024).

Machine translation (MT) systems are also used in our experiments: In the case of English to French translations, Google Translate ($mt1$) and facebook/nllb-200-distilled-600M (Team et al., 2022) as $(mt2)$ serve as our MT systems. For English to Czech tasks, translations are provided by Google Translate $(mt1)$ and one of the top-performing systems from WMT22 (Jon et al., 2022) as mt2. We are aware that using an external MT engine harms the replicability of the experiments. On the other hand, we wanted to analyze if our hypothesis applied to real-world, non-NLP community scenarios, where similar engines are often used.

4.2 Results

We studied how some of the uniformity measures change during the translation process, both for human (HT) and machine translation (MT). Surprisal estimates, obtained using models detailed in Section 4.1, are measured on a word level without tokenization, i.e. they consider punctuation as part of adjacent words. Additional results, with including tokenization and excluding punctuation surprisals, are available in Appendix A.1. The initial word's surprisal is excluded due to unreliable first token estimates from GPT-style models, though similar results were observed when included.

dataset	measure	HТ	MT1	MT ₂
	LV^2	0.39	0.58	0.42
books	$_{CV}$	0.42	0.51	0.50
	GV^2	0.46	0.69	0.54
	LV^2	0.43	0.54	0.58
wmt	$_{CV}$	0.49	0.55	0.57
	GV^2	0.46	0.57	0.64
	LV^2	0.69	0.73	0.78
global	$_{CV}$	0.65	0.70	0.63
	GV^2	0.72	0.74	0.80
	LV^2	0.72	0.79	0.83
global_doc	CV	0.68	0.81	0.82
	GV^2	0.76	0.83	0.86
	LV^2	0.16	0.40	0.53
wilde	$_{CV}$	0.07	0.39	0.54
	GV^2	0.16	0.40	0.53

Table 1: Pearsons' r for sentence-level surprisal uniformity of measurements between source and either HT, MT1 or MT2.

For conciseness, we present the results on three measures: local variance squared $(LV²)$, sentencelevel coefficient of variation (CV), and global variance squared $(GV²)$, with global variance calculated as the mean across all surprisals in the text or translation. Detailed findings for other measures are presented in Appendix A.1.

Table 1 presents Pearson's r, comparing sentence-level surprisal uniformity between the source and HT or MT, showcasing that MT aligns more closely with source text surprisal distribution than HT across all datasets and measures. In Appendix A.1, we also show the scatter plots of values of the measures for source sentence and either HT, MT1 or MT2 across datasets.

This result suggests that the source's distribution of surprisal is followed more closely by an MT system than a human translator, at least on a sentence level. We hypothesized that the human translators do not translate the sentences one by one in isolation and might distribute the surprisal variance in larger text units. This is the reason why we also measured the uniformity on a document level in *global doc*, treating each document as a single sequence of tokens for the purposes of surprisal estimation. The results do not support our hypothesis – surprisal distribution uniformity of MT is still better correlated with the source than in HT.

Absolute values of the uniformity measures (Ta-

Figure 2: Comparison on HT, MT1, and MT2 LV^2 scores per dataset (whole datasets, without MT quality filtering).

ble 2a) indicate MT is generally (with some exceptions, depending on the measure and the dataset) as uniform or less uniform than HT. Historgrams of the values support the same conclusion and are shown in the Appendix. This contradicts our initial hypothesis that MT will be more uniform in surprisal. We explored whether translation errors could cause an increase in surprisal diversity in MT – if the MT system translates the input with some obvious mistakes, then these mistakes might be very surprising given the rest of the sentence. We used reference-free COMET (wmt22-cometkiwi-da, (Rei et al., 2022)) scores to assess translation quality. We note that this approach is not without issues – COMET scores have been shown as unreliable on segmentlevel (Moghe et al., 2022).

Figure 3 shows the behavior of the value of LV^2 measure for examples where the MT COMET score is above the threshold (the threshold is on the x-axis). We see that the uniformity behavior is consistent between the HT (green) and the MT (blue and orange), except for the *wilde* dataset, where the unevenness of HT is higher when we only consider the better-scoring sentences. This result might suggest that for very creative texts (such as poetry), MT is more uniform than a human if we disregard wrongly translated utterances.

The plots show another interesting property – the COMET scores are the highest for the most uniform sentences. Since we select the examples based on the COMET scores for the MT in these plots, it can be interpreted as a property of the MT system: it translates the most uniform sentences the best. However, this behavior in the plots is very similar even when the COMET scores are computed for the HT (see Appendix A.1), suggesting a

dataset						\mathbf{m} $\mu(s)$ $\rho(s)$ $\mu(ht)$ $\rho(ht)$ $\mu(mt1)$ $\rho(mt1)$ $\mu(mt2)$ $\rho(mt2)$			
books	LV^2 CV GV^2		31.9 29.6 0.72 0.17 20.2 25.1	30.2 0.71 21.3	30.6 0.18 31.0	28.0 0.73 21.1	29.0 0.18 30.0	29.3 0.73 20.8	23.4 0.16 20.4
wmt	LV^2 CV GV^2		28.1 16.6 0.76 0.16 18.1 13.0	22.5 0.80 15.4	13.2 0.16 9.2	22.5 0.81 15.8	13.3 0.17 9.9	23.5 0.80 15.8	13.8 0.16 9.5
global	LV^2 CV GV^2		33.0 32.5 0.73 0.21 22.5 24.6	29.3 0.75 21.7	32.1 0.21 27.2	30.2 0.79 21.8	34.9 0.22 27.8	29.8 0.78 20.9	30.2 0.21 22.8
global_doc CV	LV^2 GV^2	26.6 15.2	5.9 0.94 0.07 3.3	22.1 0.98 13.1	5.4 0.08 3.2	23.4 1.00 13.9	5.5 0.09 3.2	24.3 1.17 16.5	17.1 0.49 9.7
wilde	LV^2 CV GV^2	16.0	29.7 12.1 0.67 0.08 5.7	26.0 0.63 15.1	10.0 0.09 5.2	26.7 0.69 14.6	11.5 0.10 5.5	25.3 0.72 13.4	10.5 0.11 5.4

(a) Uniformity measures for source, MT and HT across datasets for whole datasets, including examples of poor MT quality.

different underlying cause, for example, COMET bias to score more uniform sentences higher. Such biases are a base for further investigation since they do not allow us to directly automatically compare translation quality between diverse and nondiverse texts (i.e. if one system's translations are more uniform, they could be unfairly scored better than less uniform translations). This property diminishes the validity of our approach to filtering and in future work, we will focus on better ways of selecting high-quality translations for evaluation. For GV^2 score, the behavior is similar, however, we see a different trend for CV (Appendix A.1).

Based on our inspection of the translations, we have chosen COMET thresholds for which the translations seem acceptable, without serious translation errors. These thresholds are dataset specific, since COMET scores are domain dependent.¹. The results are presented in Table 2b. Again, the only notable difference to the whole dataset is on the *wilde* test set, where HT is less uniform (in LV^2) considering only examples with high-quality MT.

4.3 Multiple references

In this analysis, we explore how translation processes, both human (ref) and machine (mt), affect the uniformity of surprisal distributions, utilizing the ORT dataset to compare multiple high-quality human translations against machine translations. Surprisal estimates were generated using the MU-NLPC/CzeGPT-2 model, with parallel ex-

dataset								\mathbf{m} $\mu(s)$ $\rho(s)$ $\mu(ht)$ $\rho(ht)$ $\mu(mt1)$ $\rho(mt1)$ $\mu(mt2)$ $\rho(mt2)$	
books	LV^2 CV GV^2		30.4 27.7 0.72 0.17 19.6 24.4	30.1 0.71 21.8	32.6 0.18 34.3	27.1 0.73 21.1	26.0 0.18 29.6	28.9 0.73 21.0	24.0 0.16 21.3
wmt	LV^2 CV GV^2		25.6 14.4 0.78 0.16 16.8 10.5	20.7 0.82 14.7	12.3 0.17 8.5	20.3 0.84 14.8	11.1 0.16 8.8	21.1 0.82 14.8	12.1 0.16 8.6
global	LV^2 CV GV^2	32.9	32.6 0.73 0.20 22.2 24.3	29.2 0.76 21.5	32.2 0.21 27.1	29.9 0.79 21.5	34.8 0.21 27.4	29.6 0.78 20.8	30.2 0.20 22.4
global_doc	LV^2 CV GV^2	26.4 15.3	7.1 0.96 0.08 3.8	21.8 1.00 13.3	6.3 0.09 3.8	23.2 1.03 14.0	6.7 0.11 3.9	26.5 1.10 17.4	18.6 0.27 12.3
wilde	LV^2 CV GV^2	25.4 0.69 14.0	10.7 0.08 3.9	24.5 0.64 14.7	8.9 0.09 4.7	21.7 0.69 12.7	9.1 0.10 4.0	21.0 0.72 11.7	7.0 0.10 2.8

(b) Uniformity measures for source, MT and HT across datasets, with manually set COMET thresholds for each dataset to only select examples with high quality MT.

metric	src	ref1	ref ₂	ref3	ref4	mt	mt2
$\mu(s^{0.25})$	1.72	1.49	1.49	1.53	1.48	1.51	1.52
$\rho(s^{0.25})$	0.08	0.10	0.09	0.10	0.10	0.10	0.10
$\mu(s)$	10.21	7.07	6.91	7.73	6.94	7.13	7.59
$\rho(s)$	1.62	1.87	1.57	1.97	1.85	1.70	1.96
$\mu(s^3)$	2797	2902	1939	3289	2862	2130	3290
$\rho(s^3)$	1497	3455	2546	3615	3487	2763	3663
$\mu(qini)$	0.33	0.45	0.42	0.44	0.45	0.42	0.44
$\rho(qini)$	0.05	0.07	0.06	0.07	0.07	0.07	0.07
$\mu(CV)$	0.62	0.94	0.85	0.90	0.95	0.84	0.92
$\rho(CV)$	0.11	0.23	0.19	0.22	0.24	0.19	0.23
$\mu(LV^2)$	70.6	83.5	62.7	92.8	83.0	65.1	93.2
$\rho(LV^2)$	30.4	83.6	65.3	84.6	85.3	65.5	82.4
$\mu(GV^2)$	42.0	53.7	40.3	58.2	53.2	41.1	58.2
$\rho(GV^2)$	18.7	47.6	36.1	47.9	48.2	35.4	47.2

Table 3: Mean values and standard deviations of sentenecelevel uniformity measures for source, two machine translations and the three human reference sets. The texts are not tokenized for the surprisal estimation, thus the estimates for punctuation are often summed up with the adjacent words in the calculations of the uniformity metrics. Across most of the metrics, *ref2* and *mt* are the most uniform translations.

metric	src	ref1	ref ₂	ref3	ref4	mt	mt2
$\mu(s^{0.25}$	1.72	1.45	1.47	1.51	1.45	1.48	1.48
$\rho(s^{0.25})$	0.08	0.10	0.09	0.11	0.09	0.09	0.09
$\mu(s)$	10.34	6.50	6.57	7.52	6.48	6.55	6.85
$\rho(s)$	1.59	1.77	1.62	2.19	1.75	1.34	1.41
$\mu(s^3)$	2738	2422	1802	3445	2436	1565	2340
$\rho(s^3)$	1388	3958	3336	4176	4102	2479	2953
$\mu(qini)$	0.32	0.45	0.43	0.45	0.45	0.41	0.44
$\rho(qini)$	0.06	0.08	0.07	0.08	0.07	0.07	0.08
$\mu(CV)$	0.60	0.94	0.86	0.93	0.93	0.83	0.90
$\rho(CV)$	0.13	0.25	0.20	0.23	0.24	0.21	0.24
$\mu(LV^2)$	68.0	71.8	60.6	96.6	73.2	52.9	73.0
$\rho(LV^2)$	32.1	99.7	88.8	98.6	105.8	58.7	70.9
$\mu(GV^2)$	40.4	47.7	38.7	61.8	47.6	34.2	46.7
$\rho(GV^2)$	18.3	51.5	44.3	54.0	53.8	31.8	39.2

Table 4: Mean values and standard deviations of sentenecelevel uniformity measures for source, two machine translations and the three human reference sets, using only examples where COMET score is above 0.88 for both m t and m t 2.

¹The thresholds are: wmt: 0.88 , books: 0.81 , global: 0.7, wilde: 0.72, global doc: 0.65

Figure 3: Relationship between COMET scores of the MT and the LV^2 measure. As a proxy of translation quality, we use COMET score threshold to filter out low-quality translations.

periments conducted using an alternative language model (BUT-FIT/Czech-GPT-2-XL-133k) for comparison, detailed in Appendix A.1.3.

Figure 4 presents Pearson's r for three uniformity measures (LV^2 , CV , GV^2), revealing the degree of correlation between the source and translations.

Table 3 summarizes the mean values and standard deviations of sentence-level uniformity measures, showing variations across source, human, and machine translations. We see that mt usually scores as the most uniform, while $m\text{t}$ is among the least uniform translations, showing large variance among different MT systems.

Again, we seek to filter out translation errors in MT which could increase the diversity of surprisals by producing expressions unrelated to the rest of the sentences. This approach allows for a focused analysis on translations that accurately convey the source text's meaning without significant errors, which could otherwise distort the surprisal distribution. Upon inspection of the translations, we set the COMET threshold to 0.88 for both mt and mt2 simultaneously. Figure 5 shows the LV^2 measure for both the unfiltered and filtered datasets. Similar bar charts for CV and GV^2 scores can be found in Appendix A.1.3. We see that $m \pm 1$ usually scores as the most uniform, while $m \pm 2$ is among the least uniform translations, showing large variance among different MT systems. We see that while mt2 is the most diverse translation for the whole dataset, after filtering out the lower quality translations, ref3 becomes the most diverse in terms of LV^2 score. This result again suggests that at least a part of the surprisal diversity of MT is caused by translation errors.

5 Future work

The results of our study are inconclusive and we plan to obtain more diverse test sets, where the MT adherence to generating uniform texts could be harmful. Suppose we find such texts, where traditional, encoder-decoder MT models using beam search decoding struggle. In that case, we will evaluate also large language models, where the decoding algorithm is usually based on sampling.

We speculate that the constraints on surprisal distribution imposed by beam search might be

Figure 4: Correlation coefficients (Pearson's r) of three sentence-level measures of uniformity across the source texts, four human references and the machine translation.

Figure 5: Difference of LV^2 scores between all (orange) and high-quality (green) MT translations.

compensating for the models' inherent lack of global planning. In an ideal scenario, a model might introduce a word with high surprisal intentionally, planning to balance this with lower surprisal words in subsequent segments. However, the current model designs, focusing on next-token prediction, might not accurately forecast these future steps, i.e. the model could produce high surprisal word with a "plan" to get the surprisal "back" in the future timesteps, but, due to next-token-only objective, the future steps are miscalculated. The beam search will not produce such word, due to the adherence to the local uniformity, so the modelling flaw stays hidden.

If this hypothesis turns out to be true, our focus will shift to improve the global planning capabilities of the models, e.g. by employing an alternative training objective.

6 Conclusions

Overall, we do not have reliable proof that MT produces texts that are more uniform in surprisal distribution than humans yet. Either our hypothesis is false, or our measurement methodology is flawed. One possible reason could be that the LMs we used to estimate the surprisals are trained on human text, not on MT outputs so it overestimates surprisal of some phenomena in MT. We plan further experiments to improve our methodology and extend the analysis to more datasets.

While our study was not able to reliably prove our initial hypothesis that MT systems make the distribution of surprisal more uniform in their translations than a human translator, we have gained some insights from the experiments we carried out. Firstly, NMT systems demonstrate a tendency to produce translations that exhibit surprisal uniformity closely aligned with source texts, more so than a human translator. Secondly, the absolute values of uniformity measures are similar between HT and MT as well, however, it depends on the MT system used. Some of the systems produce more uniform translations than humans.

Notably, in more varied datasets, such as those containing literary works or poetry, human translations showed greater diversity compared to MT outputs, according to some of the measures.

Some of the findings indicate that part of the variance in surprisal distribution observed in MT may stem from translation inaccuracies. By scoring the translations using quality estimation metrics and filtering out low-scoring examples, in MT surprisal uniformity on one of the datasets increases, while HT uniformity stays the same.

7 Acknowledgements

This work was supported by the Grant Agency of Charles University in Prague (GAUK 244523) and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

References

- Aylett, Matthew and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56. PMID: 15298329.
- Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024, 01.
- BigScience Workshop. 2022. BLOOM (revision 4ab0472).
- Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.
- Fajčík, Martin, Martin Dočekal, Jan Doležal, Karel Beneš, and Michal Hradiš. 2024. Benczechmark: Machine language understanding benchmark for czech language. *arXiv preprint arXiv:insertarxiv-number-here*, March.
- Fenk, August and Gertraud Fenk-Oczlon. 1980. Konstanz im kurzzeitgedächtnis - konstanz im sprachlichen informationsfluß? *Zeitschrift fur experi- ¨ mentelle und angewandte Psychologie*, 27:400–414, 01.
- Genzel, Dmitriy and Eugene Charniak. 2002. Entropy rate constancy in text. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.
- Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Hoover, Jacob Louis, Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. 2023. The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing. *Open Mind*, 7:350–391, 07.
- Hájek, Adam and Aleš Horák. 2024. Czegpt-2 training new model for czech generative text processing evaluated with the summarization task. *IEEE Access*, 12:34570–34581.
- Jon, Josef, Martin Popel, and Ondřej Bojar. 2022. CUNI-bergamot submission at WMT22 general translation task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, ` Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 280– 289, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.
- Kann, Katharina, Sascha Rothe, and Katia Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In Korhonen, Anna and Ivan Titov, editors, *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium, October. Association for Computational Linguistics.
- Levy, Roger and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. volume 19, pages 849–856, 01.
- Meister, Clara, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online, November. Association for Computational Linguistics.
- Meister, Clara, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Moghe, Nikita, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2022. Extrinsic evaluation of machine translation metrics.
- Nguyen, Khanh and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In Wang, Lu, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China, November. Association for Computational Linguistics.
- Pauls, Adam and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In Li, Haizhou, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 959–968, Jeju Island, Korea, July. Association for Computational Linguistics.

- Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, Jose G. C. de Souza, Taisiya Glushkova, Duarte ´ Alves, Luisa Coheur, Alon Lavie, and Andre F. T. ´ Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costajussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, Andre Martins, Makoto Morishita, Christof Monz, ´ Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.
- Shain, Cory and William Schuler. 2023. A deep learning approach to analyzing continuous-time systems.
- Shain, Cory, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.
- Team, NLLB, Marta R. Costa-jussà, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.
- Wei, Jason, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5191–5202, Online, August. Association for Computational Linguistics.
- Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies

and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December.

- Zouhar, Vilém and Ondřej Bojar. 2024. Quality and quantity of machine translation references for automated metrics.
- Zouhar, Vilém, Věra Kloudová, Martin Popel, and Ondřej Bojar. 2023. Evaluating optimal reference translations.

A Translation

A.1 English to French dataset

A.1.1 Correlations

Scatter plots 6, 7, 8, 9, 10, 11 show how well the $1v^2$ and GV^2 measures correlate between source sentence and HT, MT1 and MT2 sentences on *books, wmt and wilde datasets.*

A.1.2 Absolute values

Figures 12, 21, 22, 13, 14, 15, 16, 17 and 18 show histograms of the senetence-level values of $1v^2$, CV and GV^2 across *books*, wmt and wilde datasets. Figures 19 and 20 show the $1v^2$ histograms on *wilde* and *wmt* after applying filtering based on COMET threshold of the MT.

Figures 24 and 25 show values of CV and GV^2 depending on the COMET threshold for MT translations.

Figure 23 shows the relationship between the COMET score threshold on the *human* translation and the mean lv2 score of the source sentence and all the translations. We see that in *books, global and wmt*, both the source sentence and the translations are more uniform for high COMET scores. This might suggest a preference of COMET score for uniform surprisal (for example, rooted in training data).

A.1.3 Multiple reference dataset

We experimented also with alternative preprocessing and LMs for estimating the wordlevel surprisals in the Czech translations of the ORT dataset. The estimates in the main text are computed on untokenized input, i.e. surprisals of punctuation adjacent to a word are summed with that word's surprisal. In Tables 5 and 6, we present the results on tokenized (i.e. punctuation surprisals are considered separately) texts and texts with punctuation removed altogether. We also used an alternative language model (BUT-FIT/Czech-GPT-2-XL-133k) to estimate the surprises. See Tables 7, 8, 9 for untok-

Figure 6: Scatter plots of $1v^2$ on *books* dataset between source, and either HT, MT1 or MT2.

Figure 7: Scatter plots of $1v2$ on *wmt* dataset between source, and either HT, MT1 or MT2.

Figure 8: Scatter plots of $1v2$ on *wilde* dataset between source, and either HT, MT1 or MT2.

Figure 9: Scatter plots of GV^2 on *books* dataset between source, and either HT, MT1 or MT2.

Figure 10: Scatter plots of GV^2 on *wmt* dataset between source, and either HT, MT1 or MT2.

Figure 11: Scatter plots of GV^2 on *wilde* dataset between source, and either HT, MT1 or MT2.

Figure 12: Histogram of lv2 on *books* dataset for source, HT, MT1 and MT2.

Figure 13: Histogram of GV^2 on *books* dataset for source, HT, MT1 and MT2.

Figure 14: Histogram of GV^2 on *wmt* dataset for source, HT, MT1 and MT2.

Figure 15: Histogram of GV^2 on *wilde* datasets for source, HT, MT1 and MT2.

Figure 16: Histogram of CV on *books* dataset for source, HT, MT1 and MT2.

Figure 17: Histogram of CV on *wmt* dataset for source, HT, MT1 and MT2.

Figure 18: Histogram of CV on *wilde* datasets for source, HT, MT1 and MT2.

Figure 19: Histograms of lv2 on *wmt* dataset for source, HT, MT1 and MT2, after applying COMET threshold, only keeping examples where MT1 COMET score is above 0.88

Figure 20: Histograms of lv2 on *wilde* dataset for source, HT, MT1 and MT2, after applying COMET threshold, only keeping examples where MT1 COMET score is above 0.72

Figure 21: Histogram of 1v2 on *wmt* dataset for source, HT, MT1 and MT2.

enized, tokenized and punctuation-free results, respectivelly.

Figures 26 and 27 show mean values of CV and GV^2 scores on unfiltered, whole dataset (orange) and dataset containing only examples where COMET score for both mt and mt2 is above 0.88.

Table 5: Mean values and standard deviations of sentenecelevel uniformity measures for source, two machine translations and the three human reference sets. The texts are tokenized for the surprisal estimation, thus the estimates for punctuation are considered separately in the uniformity measures' calculations.

Table 6: Mean values and standard deviations of sentenecelevel uniformity measures for source, two machine translations and the three human reference sets. The surprisal estimates for punctuation are discarded for the uniformity measures' calculation.

) 2814 1156 1214 1264 1097 1249 1360

) 1316 948 923 947 867 897 1097

) 72.4 42.2 43.5 44.4 41.4 42.8 48.3

) 29.5 25.7 26.2 25.1 24.1 23.7 30.7

) 41.8 28.2 28.6 29.5 27.5 29.0 31.4

) 17.2 16.4 16.4 16.5 15.9 16.5 19.5

glob) 28.1 28.1 28.6 29.6 27.4 29.1 31.6

glob) 16.0 16.0 16.4 16.8 15.3 16.9 20.0

 $\mu(GV^2)$

 $\rho(GV^2)$

Figure 22: Histogram of lv2 on *wilde* dataset for source, HT, MT1 and MT2.

Figure 23: Relationship between COMET scores of the HT and the $1v^2$ measure.

Figure 24: Relationship between COMET scores of the MT and the CV measure. As a proxy of translation quality, we use COMET score threshold to filter out low-quality translations.

Figure 25: Relationship between COMET scores of the MT and the GV^2 measure. As a proxy of translation quality, we use COMET score threshold to filter out low-quality translations.

metric	src	ref1	ref ₂	ref3	ref4	mt	mt2
$\mu(s^{0.25})$	1.72	1.35	1.37	1.39	1.33	1.39	1.39
$\rho(s^{0.25})$	0.08	0.10	0.09	0.10	0.09	0.11	0.10
$\mu(s)$	10.21	4.70	4.96	5.25	4.55	5.24	5.30
$\rho(s)$	1.62	1.01	1.02	1.09	0.94	1.26	1.23
$\mu(s^3)$	2797	417	483	565	397	540	615
$\rho(s^3)$	1497	355	383	425	322	460	553
$\mu(gini)$	0.33	0.42	0.42	0.42	0.43	0.42	0.43
$\rho(qini)$	0.05	0.06	0.06	0.06	0.06	0.06	0.06
$\mu(CV)$	0.62	0.80	0.79	0.79	0.82	0.77	0.80
$\rho(CV)$	0.11	0.15	0.15	0.15	0.14	0.14	0.15
$\mu(LV^2)$	70.6	26.6	29.9	32.0	26.5	30.2	34.7
$\rho(LV^2)$	30.4	15.4	16.8	16.3	14.7	16.5	20.4
$\mu(GV^2)$	42.0	14.8	16.3	17.9	14.5	17.2	19.0
$\rho(GV^2)$	18.7	7.9	8.7	9.2	7.5	9.8	11.5
$\mu(GV^2 _glob)$	14.8	14.8	16.3	18.1	14.5	17.3	19.2
$\rho(GV^2 _glob)$	7.6	7.6	8.7	9.5	7.0	10.1	11.9

Table 7: Mean values and standard deviations of sentenecelevel uniformity measures for source, two machine translations and the three human reference sets. The texts are not tokenized for the surprisal estimation, thus the estimates for punctuation are often summed up with the adjacent words in the calculations of the uniformity metrics. The surprisals are calculated by BUT-FIT/Czech-GPT-2-XL-133k model.

Table 8: Mean values and standard deviations of sentencelevel uniformity measures for source, two machine translations and the three human reference sets. The surprisals are calculated by BUT-FIT/Czech-GPT-2-XL-133k model. The texts are tokenized for the surprisal estimation, thus the estimates for punctuation are considered separately in the uniformity measures' calculations.

Figure 26: Difference of CV scores between all (orange) and high-quality (green) MT translations.

metric	src	ref1	ref ₂	ref3	ref4	mt	mt2
$\mu(s^{0.25})$	1.73	1.33	1.35	1.36	1.31	1.37	1.37
$\rho(s^{0.25})$	0.07	0.10	0.09	0.10	0.09	0.11	0.10
$\mu(s)$	10.36	4.52	4.77	4.91	4.38	5.07	5.05
$\rho(s)$	1.48	0.96	0.97	1.01	0.90	1.22	1.17
$\mu(s^3)$	2814	380	433	461	359	488	537
$\rho(s^3)$	1316	342	359	385	304	416	512
$\mu(qini)$	0.32	0.43	0.43	0.42	0.44	0.42	0.43
$\rho(gini)$	0.05	0.06	0.07	0.06	0.06	0.06	0.06
$\mu(CV)$	0.60	0.81	0.80	0.79	0.83	0.78	0.81
$\rho(CV)$	0.10	0.15	0.15	0.15	0.14	0.14	0.16
$\mu(LV^2)$	72.4	25.1	27.9	28.0	25.1	28.2	32.1
$\rho(LV^2)$	29.5	14.6	16.2	14.8	13.9	15.1	20.3
$\mu(GV^2)$	41.8	14.0	15.3	15.7	13.8	16.2	17.8
$\rho(GV^2)$	17.2	7.6	8.2	8.4	7.0	8.9	11.2
$\mu(GV^2 _glob)$	14.0	14.0	15.3	15.8	13.8	16.4	18.0
$\rho(GV^2 _glob)$	7.3	7.3	8.2	8.5	6.6	9.3	11.5

Table 9: Mean values and standard deviations of sentencelevel uniformity measures for source, two machine translations and the three human reference sets. The surprisals are calculated by BUT-FIT/Czech-GPT-2-XL-133k model. The surprisal estimates for punctuation are discarded for the uniformity measures' calculation.

Figure 27: Difference of GV^2 scores between all (orange) and high-quality (green) MT translations.