

# Fine-tuning with Uniform Information Density-based regularization for Dutch language modelling

Sander van den Bent  
Maria Tepei  
Jelke Bloem

SANDER.VAN.DEN.BENT@STUDENT.UVA.NL  
M.TEPEI@UVA.NL  
J.BLOEM@UVA.NL

*Institute for Logic, Language and Computation, University of Amsterdam*

## Abstract

The uniform information density (UID) hypothesis states that the information within utterances of communication should be evenly distributed for optimal communication. Human speakers have the natural tendency to have an even information density within their communication, subject to the rules of grammar. For large language models (LLMs) potential inductive biases that impact their information density are still an area of investigation. Previous research has indicated that modifying the (pre)training loss function with regularizers based on information-theoretic principles has had a favorable impact on the general perplexity and information density of generated responses of English LLMs. We investigate the effects of fine-tuning a Dutch pre-trained GPT-2 model using these regularizers on the perplexity and information density of generated responses. We observe that UID regularizers at the fine-tuning stage can improve information density metrics, though these effects are less strong than those reported in previous work that pre-trained models with such regularizers from scratch.

## 1. Introduction

With the increasing use of large language models (LLMs), such as generative pretrained transformer (GPT) models, and the rapid adoption of generative artificial intelligence (GenAI), there is growing interest in how these models distribute information in generated language. In particular, it remains an open question to what extent principles proposed for human language production also apply to neural text generation. One such principle is the Uniform Information Density (UID) hypothesis, which states that information within an utterance tends to be distributed evenly in order to optimise communication (Frank and Jaeger 2008, Fenk and Fenk 1980). According to UID, speakers and writers structure language so as to maintain a relatively stable flow of information, avoiding peaks and troughs that may hinder comprehension. Information density is commonly operationalised using *surprisal*, defined as the unexpectedness of a linguistic unit given its context (Hale 2001). For larger sequences of tokens, information density is often quantified using *perplexity*. Previous work suggests that evenly distributed information density reduces processing effort and cognitive load for the listener (Jaeger 2010).

Several linguistic strategies have been identified through which information density can be controlled. These include *syntactic reduction*, where speakers may optionally include or omit elements within an utterance (Jaeger and Levy 2006), lexical variation (Mahowald et al. 2013), and sentence linking within discourse (Torabi Asr and Demberg 2015). Typological differences between languages may also influence information distribution, with subject–verb–object (SVO) languages generally exhibiting more uniform information density (Clark et al. 2023). English, French, and Dutch are examples of such languages.

The UID hypothesis is also relevant for LLMs, particularly decoder-only architectures that generate text by predicting tokens based on preceding context. Firstly, large language models are good estimators of human language processing effort, as language model perplexity has been shown to correlate with human reading times, though not necessarily scaling with model size (Oh et al. 2024).

Secondly, it has been shown that language models pretrained with UID regularizers achieve better test set perplexity compared to equivalent unregularized language models across a range of languages. This is taken as evidence for the truth of the UID hypothesis in human language production (Wei et al. 2021). More specifically, Wei et al. (2021) augment the standard maximum likelihood estimation (MLE) objective with uniform information density-based regularizers during training and observe improved perplexity, as well as increased lexical diversity in generated text in a range of 10 typologically diverse languages.

However, this study does not include Dutch and focuses on pre-training. Information-guided regularization has also been applied for fine-tuning by Sharma et al. (2024), but they do not explore UID-based regularization. Therefore, we address this gap by exploring the effect of UID-based regularization in a fine-tuning setting (using a continued pre-training approach). While the use of regularizers at this stage is likely to be less effective than in the pre-training stage, it is a lot less costly to fine-tune models than to pre-train them from scratch. Therefore, it would be useful to know whether this technique can also yield benefits at the fine-tuning stage if full pre-training from scratch is not feasible.

In our study, we investigate three UID-based regularization techniques: the two applied by Wei et al. (2021), which are variance-based regularization and local consistency regularization, as well as the max regularization approach proposed by Meister et al. (2020a). While Wei et al. (2021) applied them in pretraining of fairseq transformer models with six decoder layers and eight attention heads, we apply them during fine-tuning of a Dutch pretrained GPT-2 model with 12 decoder layers and 12 attention heads. Because GPT-2 is an open-source model, surprisal can be computed directly from next-token probability distributions, making it suitable for testing UID-related hypotheses. As in previous work, we evaluate both model perplexity and the information density of generated text.<sup>1</sup>

The aim of this study is to assess to what extent Wei et al. (2021)’s findings hold in fine-tuning settings and for Dutch, that is, whether adding information density-based regularization to the standard MLE objective leads to more uniformly distributed information in generated responses, while also evaluating changes in perplexity relative to a baseline model.

## 2. Related Work

This section reviews previous work on the Uniform Information Density (UID) hypothesis in relation to Dutch, discusses linguistic properties of Dutch that may affect information density, and outlines information-theoretic regularization methods relevant to neural language models.

### 2.1 Dutch and Uniform Information Density

Dutch is a West Germanic language with both fixed and flexible syntactic properties that may influence information density. Main clauses typically follow a subject–verb–object (SVO) order, while subordinate clauses exhibit subject–object–verb (SOV) order. These word orders are among the most common cross-linguistically and have been associated with more uniform information density distributions (Maurits et al. 2010). Dutch further adheres to the verb-second (V2) constraint in main clauses, allowing different elements to precede the verb without changing propositional meaning. This syntactic flexibility has been argued to support more even information distribution (Clark et al. 2023). Another characteristic of Dutch syntax is scrambling, whereby constituents such as adverbs can appear in different positions within a clause without altering meaning (De Hoop 2003). This positional flexibility may likewise contribute to smoothing information density across an utterance. In addition to syntactic properties, Dutch exhibits compounding, in which multiple lexical items are combined into a single word, with the rightmost element serving as the head (Booij 2019). Compounding has been argued to potentially increase surprisal by concentrating information into

---

1. Code for these experiments is available at: <https://github.com/SandervandenBent/Master-Thesis-Dutch-GPT2-UID-Hypothesis>

fewer items (Gamboa et al. 2024). At the same time, by reducing the number of syntactic units in a sentence, compounding may also contribute to more even information distributions. Empirical evidence for either of these effects remains inconclusive.

With this syntactic flexibility, and a range of other word order alternations, Dutch word order is somewhat more flexible than English (Neeleman and Weerman 1997, Hawkins 2012), which does not have scrambling, for example. Some Dutch word order variation phenomena have been specifically hypothesized to be affected by the uniform information density hypothesis, such as the word order in the verbal cluster (Bloem 2016) and the optionality of the complementizer *om* in *to*-infinitival complement clauses (Bouma 2017). This makes theoretical proposals on the influence of uniform information density on word order more relevant for Dutch than for English.

## 2.2 Surprisal and Information Density

Information density is commonly operationalised using surprisal, defined as the negative logarithm of the conditional probability of a word given its context (Shannon 1948, Hale 2001, Levy 2008). Because a universal probability distribution over a language is not available, surprisal is typically estimated using statistical language models (Wei et al. 2021). For a word  $w_t$  in context  $W_{<t}$ , surprisal is defined as:

$$s(w_t) = -\log p(w_t | W_{<t})$$

Transformer-based language models estimate these probabilities through self-attention mechanisms and a softmax layer over the vocabulary (Vaswani et al. 2017, Brown et al. 2020). GPT-2 follows an autoregressive generation framework, predicting each token conditioned on preceding tokens (Radford et al. 2019). For longer sequences, information density is often evaluated using perplexity, defined as the exponential of the average surprisal:

$$\text{Perplexity}(W) = \exp\left(-\frac{1}{T} \sum_{t=1}^T \log p_{\theta}(w_t | w_{<t})\right)$$

where  $\theta$  are the parameters of the model and  $t$  is a token. Perplexity is a standard evaluation metric for language models and has been widely used in studies investigating information-theoretic regularization (Wei et al. 2021, Brown et al. 2020).

## 2.3 Information Density-Based Regularization

Recent work has proposed editing the standard maximum likelihood estimation (MLE) objective with regularization terms that target information density (Wei et al. 2021, Sharma et al. 2024). These regularizers penalize undesirable surprisal patterns during training and are weighted by a hyperparameter  $\lambda$  that controls the strength of the penalty. The optimal value of  $\lambda$  depends on the dataset and task.

Wei et al. (2021) evaluate two such regularizers, variance-based and local consistency regularization, and discuss their effects on both perplexity and information density. These regularizers are added to the MLE loss and differ in which properties of the surprisal sequence they penalize. They experiment with EuroParl corpus data for five languages (Czech, English, French, German and Spanish) as well as Wikipedia dumps of five more typologically diverse languages (Finnish, Indonesian, Turkish, Swahili and Tagalog). They also perform a data ablation study for English. They train fairseq transformers with six decoder layers and eight attention heads using these regularizers on these languages, finding perplexity improvements (usually in the range of 0-1%) across languages, with larger improvements for the lowest-resourced languages in their dataset and when using smaller training datasets for English. As far as languages related to Dutch go, in their experiment, the variance regularizer shows better performance on English and the local consistency regularizer shows better performance for German.

In addition to the two regularizers from this work, Meister et al. (2020b) proposes a max-based regularizer that targets extreme surprisal values. They assess the three regularizers in the English-French machine translation decoding task and the German-English machine translation decoding task, but do not report results separately per language.

Variance regularization penalizes sequences with high variance in surprisal values:

$$R(S) = \frac{1}{|w|} \sum_{t=1}^{|w|} (s(w_t) - \mu)^2$$

where  $\mu$  is the mean surprisal of the sequence. Local consistency regularization penalizes large surprisal differences between adjacent tokens:

$$R(S) = \frac{1}{|w| - 1} \sum_{t=1}^{|w|-1} (s(w_t) - s(w_{t+1}))^2$$

The max regularizer penalizes sequences based on their highest surprisal value:

$$R(S) = \max_{t=1}^{|S|} s(w_t)$$

We evaluate these regularization strategies during fine-tuning of a Dutch pretrained GPT-2 model, focusing on their effects on perplexity and the information density of generated Dutch text. This will show whether UID regularization might also be beneficial for Dutch language modelling, particularly in low resource settings.

### 3. Methodology

We investigate the effect of information density-based regularization during fine-tuning of a Dutch pretrained GPT-2 language model. We add multiple regularization techniques to the standard training objective and evaluate them with respect to their influence on perplexity and the information density of generated text.

#### 3.1 Datasets

Two Dutch-language datasets are used in this study: OpenSubtitles and Europarl. These datasets differ in domain and register, allowing evaluation of regularization effects across different types of text.

##### 3.1.1 OPENSUBTITLES

The English-to-Dutch OpenSubtitles dataset is part of the OpenSubtitles corpus available via the OPUS platform (Lison and Tiedemann 2016). Subtitles are extracted from audiovisual material and reflect informal and spoken language. The dataset has been used in a variety of NLP tasks, including sequence labeling and conversational modeling (Chapuis et al. 2020, Xu et al. 2018). In this study, version v1 is used, consisting of 1,371,836 Dutch sentences with substantial variation in length and syntactic structure.

##### 3.1.2 EUROPARL

The Europarl corpus consists of proceedings from the European Parliament and is available in multiple languages (Koehn 2005). The English-to-Dutch subset includes both original Dutch speeches and translated material. Version v3 is used, representing formal and highly structured language.

### 3.1.3 DESCRIPTIVE STATISTICS

Table 1 presents descriptive statistics for both datasets, including sentence and token counts, vocabulary size, and sentence length.

Metric	OpenSubtitles	Europarl
Total sentences	1,371,832	1,619,273
Total tokens	7,678,350	36,421,549
Vocabulary size	280,876	420,461
Average sentence length	5.60	22.49
Median sentence length	5.00	20.00

Table 1: Descriptive statistics of the datasets.

## 3.2 Model Architecture

All experiments are conducted using `gpt2-small-dutch`, an open-source Dutch pretrained GPT-2 model available via HuggingFace (de Vries and Nissim 2021). The architecture is identical to the original GPT-2 model, with modifications limited to lexical embeddings. Approximately 13GB of Dutch text was used during pre-training, including Wikipedia articles, books, and news. The model provides token-level probability estimates, enabling direct computation of surprisal. A baseline model is obtained by fine-tuning without additional regularization.

## 3.3 Loss Function and Regularization

Fine-tuning is performed using the standard cross-entropy (negative log-likelihood) loss function:

$$L(\theta) = - \sum_{t=1}^T \log p_{\theta}(w_t | w_{<t}).$$

Information density-based regularizers are added to this objective during fine-tuning. We evaluate three regularization strategies: surprisal variance regularization, local consistency regularization, and a max surprisal regularizer. Each regularizer is weighted by a hyperparameter  $\beta$ .

### 3.3.1 SURPRISAL VARIANCE REGULARIZATION

Surprisal variance regularization penalizes high variance in surprisal values within a sequence:

$$L'(\theta) = L(\theta) + \beta \cdot \frac{1}{T} \sum_{t=1}^T (u(w_t) - \bar{u})^2$$

### 3.3.2 LOCAL CONSISTENCY REGULARIZATION

Local consistency regularization penalizes large differences in surprisal between adjacent tokens:

$$L''(\theta) = L(\theta) + \beta \cdot \frac{1}{T-1} \sum_{t=2}^T (u(w_t) - u(w_{t-1}))^2$$

### 3.3.3 MAX REGULARIZATION

The max regularizer penalizes sequences based on their highest surprisal value:

$$L'''(\theta) = L(\theta) + \beta \cdot \max_t u(w_t)$$

### 3.4 Training Procedure

All regularizers are evaluated using the same training pipeline. Fine-tuning is performed with fixed hyperparameters across conditions, with  $\lambda \in \{0.1, 0.01, 0.001\}$ . Training and testing splits are set to 80–20. To examine the effect of data size, experiments are conducted using 5%, 25%, and 100% of each dataset. Random seeds are fixed and shuffling is disabled to ensure reproducibility. Table 2 provides an overview of the hyperparameter settings.

Hyperparameter	Value
Sequence length	32
Batch size	16
Epochs	1
Learning rate	$5 \times 10^{-5}$
Optimizer	AdamW
$\lambda$ values	$\{0.1, 0.01, 0.001\}$
Train fraction	$\{0.05, 0.25, 1\}$
Test fraction	$\{0.05, 0.25, 1\}$
Train split	0.8
Test split	0.2
Training shuffle	False
Test shuffle	False

Table 2: Hyperparameters used during fine-tuning.

### 3.5 Evaluation

Two evaluation approaches are applied: perplexity and analysis of generated text. Our experimental setup includes two conditions: one where the models are trained and evaluated on data from the same source (OpenSubtitles or Europarl) and a cross-validation where the models are trained on the training set from one source, and then evaluated on the evaluation set from the other source. The latter condition can show to what extent this type of regularized tuning can generalize across text types.

#### 3.5.1 PERPLEXITY

Perplexity is computed on the test sets to evaluate language modeling performance. Scores are compared against the baseline model fine-tuned without regularization, with lower perplexity indicating more predictable token distributions.

#### 3.5.2 GENERATED TEXT ANALYSIS

Generated text is evaluated using predefined prompts for each dataset (Table 3). These prompts are phrases from the respective corpora and were chosen either for their everyday conversational style or political tone. For each prompt, 100 samples are generated, resulting in 1,000 samples per dataset.

A maximum output length of 50 tokens is applied. For evaluation metrics, we use average sequence length, lexical diversity (2-, 3-, and 4-grams), output entropy, surprisal variance, local consistency, and max surprisal. Wei et al. (2021) find that their approach generally increases the lexical diversity of generated text, which is a good property for generated text to have.

OpenSubtitles prompts	Europarl prompts
Wie ben jij? <i>Who are you?</i>	Waarover gaat de vergadering vandaag? <i>What is the meeting about today?</i>
Ik geloof je niet. <i>I don't believe you.</i>	Dit is een cruciaal economisch beleid. <i>This is a crucial economic policy.</i>
Gaan we naar huis? <i>Are we going home?</i>	Het Europese parlement moet zich inzetten voor.. <i>The European Parliament must commit to..</i>
Waarom ga je? <i>Why are you going?</i>	Wij roepen de Commissie nu op. <i>We now call on the Commission.</i>
We moeten nu gaan. <i>We have to go now.</i>	Reflecterend op de recente gebeurtenissen.. <i>Reflecting on recent events..</i>
Ik weet niet wat ik moet zeggen. <i>I don't know what to say.</i>	Wat vereist samenwerking tussen de lidstaten? <i>What requires cooperation between member states?</i>
Mis je mij? <i>Do you miss me?</i>	Dit voorstel zal bijdragen aan consensus. <i>This proposal will contribute to consensus.</i>
Denk je dat ze terugkomen? <i>Do you think they will come back?</i>	De verantwoordelijkheid ligt bij de wetenschap. <i>The responsibility lies with science.</i>
Hij zei dat het daar gevaarlijk was. <i>He said it was dangerous there.</i>	Juist daar ligt de complexiteit. <i>That is precisely where the complexity lies.</i>
Dat is niet waar, dat gebeurde niet. <i>That's not true, that didn't happen.</i>	De accordering van het wetsvoorstel. <i>The approval of the bill.</i>

Table 3: Prompts used for text generation.

## 4. Results

We compare the perplexity scores without and with the regularizers. For every dataset proportion, a baseline score is set. Perplexity scores obtained after training with the regularizers are then compared to the baseline score. Subsequently, we analyse and evaluate the generated text.

For reference, Wei et al. (2021) report perplexity improvements of 0.7% (local consistency regularizer) to 1.2% (variance regularizer) for English and 0.4% (variance regularizer) to 0.6% (local consistency regularizer) for German. Improvements at the fine-tuning stage are likely to be smaller than those at the pre-training stage. Therefore, improvements of over 0.2% can be considered notable.

**Training: OpenSubtitles | Testing: OpenSubtitles**

100% of dataset			
Baseline: 57.15			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	57.27 ( $\uparrow 0.21\%$ )	57.16 ( $\uparrow 0.02\%$ )	57.19 ( $\uparrow 0.07\%$ )
UID Local Consistency	57.37 ( $\uparrow 0.38\%$ )	57.17 ( $\uparrow 0.03\%$ )	57.12 ( $\downarrow 0.05\%$ )
UID Max	61.04 ( $\uparrow 6.81\%$ )	57.27 ( $\uparrow 0.21\%$ )	57.23 ( $\uparrow 0.14\%$ )
25% of dataset			
Baseline: 66.61			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	66.58 ( $\downarrow 0.05\%$ )	66.56 ( $\downarrow 0.08\%$ )	66.51 ( $\downarrow 0.15\%$ )
UID Local Consistency	66.51 ( $\downarrow 0.15\%$ )	66.28 ( $\downarrow 0.50\%$ )	66.61 ( $\downarrow 0.00\%$ )
UID Max	66.50 ( $\downarrow 0.17\%$ )	66.57 ( $\downarrow 0.06\%$ )	66.57 ( $\downarrow 0.06\%$ )
5% of dataset			
Baseline: 64.68			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	64.76 ( $\uparrow 0.12\%$ )	64.61 ( $\downarrow 0.11\%$ )	64.61 ( $\downarrow 0.11\%$ )
UID Local Consistency	65.00 ( $\uparrow 0.49\%$ )	64.76 ( $\uparrow 0.12\%$ )	64.62 ( $\downarrow 0.09\%$ )
UID Max	67.13 ( $\uparrow 3.79\%$ )	64.91 ( $\uparrow 0.36\%$ )	64.62 ( $\downarrow 0.09\%$ )

Table 4: Perplexity scores for the OpenSubtitles training and testing set.

**Training: OpenSubtitles | Testing: Europarl**

100% of dataset			
Baseline: 276.03			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	283.84 ( $\uparrow 2.79\%$ )	274.91 ( $\downarrow 0.41\%$ )	279.06 ( $\uparrow 1.09\%$ )
UID Local Consistency	279.47 ( $\uparrow 1.24\%$ )	275.30 ( $\downarrow 0.26\%$ )	276.32 ( $\uparrow 0.11\%$ )
UID Max	295.82 ( $\uparrow 6.92\%$ )	279.31 ( $\uparrow 1.18\%$ )	274.73 ( $\downarrow 0.47\%$ )
25% of dataset			
Baseline: 203.95			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	203.38 ( $\downarrow 0.28\%$ )	204.28 ( $\uparrow 0.16\%$ )	204.06 ( $\uparrow 0.05\%$ )
UID Local Consistency	204.20 ( $\uparrow 0.12\%$ )	204.89 ( $\uparrow 0.46\%$ )	204.64 ( $\uparrow 0.34\%$ )
UID Max	214.25 ( $\uparrow 4.93\%$ )	204.73 ( $\uparrow 0.38\%$ )	204.60 ( $\uparrow 0.32\%$ )
5% of dataset			
Baseline: 156.08			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	154.62 ( $\downarrow 0.94\%$ )	155.56 ( $\downarrow 0.33\%$ )	155.60 ( $\downarrow 0.31\%$ )
UID Local Consistency	154.91 ( $\downarrow 0.75\%$ )	156.09 ( $\uparrow 0.01\%$ )	155.73 ( $\downarrow 0.22\%$ )
UID Max	172.44 ( $\uparrow 9.96\%$ )	156.03 ( $\downarrow 0.03\%$ )	156.00 ( $\downarrow 0.05\%$ )

Table 5: Perplexity scores for the OpenSubtitles training and Europarl testing set.

## 4.1 OpenSubtitles

Table 4 shows the results using the OpenSubtitles dataset as training and testing set. Taking the full dataset for training and testing showed no improvements with the exception of applying local consistency with a lambda value of 0.001, though that improvement being slight with a lowering of test set perplexity by 0.05%. The 25% proportion of the dataset showed the most improvements with almost all regularizers and lambda values showing a decrease in test set perplexity, with the only exception being local consistency with lambda value 0.001. The largest improvement is seen using the local consistency regularizer at the 25% proportion of the dataset using a lambda value of 0.01. Another notable result comes from applying the max regularizer with lambda value 0.1, showing that a relatively harsh penalty with strong lambda value resulted in an improvement on 25% of the dataset whereas applying this same method on the other dataset proportions of 100% and 5% resulted in the largest worsening in perplexity scores (increases of 6.81% and 3.79%, respectively).

For the cross-validation results, see Table 5. After training on OpenSubtitles and testing on Europarl, generally higher perplexity scores were seen on both baseline and regularization. This indicates that fine-tuning on the OpenSubtitles dataset provides no generalizability. Despite the overall higher perplexity scores, improvements were still observed after applying the regularization methods. Compared to using the OpenSubtitles testing set, the 100% and 5% proportions saw a higher frequency of improvements along with higher percentual decreases of perplexity scores. Notably, using the 25% proportion caused only improvements in using the OpenSubtitles testing set but close to no improvements in the Europarl testing set, indicating a lack of generalizing and overfitting of the training data, despite the generalization attempt of the regularization methods.

## 4.2 Europarl

Table 6 shows the results of the perplexity scores compared to the baselines of the different dataset proportion and lambda values. The regularization methods showed little to no improvements on the perplexity metric, often showing even worsening scores. Only in the 5% subset of the dataset and applying the surprisal variance regularizer with lambda value 0.001 a slight improvement of 0.03% was seen. Similar to the OpenSubtitles training and testing set, a notable result comes from applying the max regularizer with lambda value 0.1, resulting in a relatively strong worsening of perplexity scores in the 5% subset of Europarl. With this methodology, in the 100% dataset no change in perplexity was seen, in the 5% dataset proportion it led to a worsening of 12.09%.

The cross-validation results are shown in Table 7. Applying the OpenSubtitles testing set resulted in much higher general perplexity scores compared to using the Europarl testing set. However, with the application of the OpenSubtitles testing set, in the 25% training subset the application of all the regularizers resulted in improvements of perplexity scores with lambda values 0.01 and 0.001 which were not seen in the Europarl testing set. For the 100% and 5% proportions, for neither testing set any improvements were seen, with the only exception being the surprisal variance regularizer with lambda value 0.001 on the Europarl testing set.

**Training: Europarl | Testing: Europarl**

100% of dataset			
Baseline: 24.88			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	25.05 (↑0.68%)	24.89 (↑0.04%)	24.88 (↓0.00%)
UID Local Consistency	25.29 (↑1.65%)	24.88 (↓0.00%)	24.88 (↓0.00%)
UID Max	24.88 (↓0.00%)	24.88 (↓0.00%)	24.88 (↓0.00%)
25% of dataset			
Baseline: 29.05			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	29.06 (↑0.03%)	29.06 (↑0.03%)	29.07 (↑0.07%)
UID Local Consistency	29.05 (↓0.00%)	29.06 (↑0.03%)	29.05 (↓0.00%)
UID Max	29.06 (↑0.03%)	29.05 (↓0.00%)	29.05 (↓0.00%)
5% of dataset			
Baseline: 36.30			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	36.50 (↓0.55%)	36.30 (↓0.00%)	36.29 (↓0.03%)
UID Local Consistency	36.87 (↑1.57%)	36.34 (↑0.11%)	36.33 (↑0.08%)
UID Max	40.69 (↑12.09%)	36.41 (↑0.30%)	36.35 (↑0.14%)

Table 6: Perplexity scores for the Europarl training and testing set.

**Training: Europarl | Testing: OpenSubtitles**

100% of dataset			
Baseline: 399.03			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	401.48 (↑0.61%)	400.30 (↑0.32%)	400.37 (↑0.34%)
UID Local Consistency	406.88 (↑1.95%)	401.81 (↑0.70%)	400.39 (↑0.34%)
UID Max	468.21 (↑15.95%)	404.11 (↑1.27%)	399.92 (↑0.22%)
25% of dataset			
Baseline: 339.92			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	342.13 (↑0.65%)	339.43 (↓0.15%)	339.21 (↓0.21%)
UID Local Consistency	340.63 (↑0.21%)	339.06 (↓0.25%)	339.31 (↓0.18%)
UID Max	375.97 (↑10.07%)	338.84 (↓0.32%)	339.90 (↓0.01%)
5% of dataset			
Baseline: 228.30			
	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$
UID Variance	233.82 (↑2.39%)	228.97 (↑0.29%)	229.12 (↑0.36%)
UID Local Consistency	234.28 (↑2.59%)	230.50 (↑0.96%)	228.96 (↑0.29%)
UID Max	248.88 (↑8.63%)	231.99 (↑1.60%)	228.50 (↑0.09%)

Table 7: Perplexity scores for the Europarl training and OpenSubtitles testing set.

### 4.3 Analysis of generated text

Based on our previous results in sections 4.1 and 4.2, we select those dataset proportions and lambda values that generated the highest improvements in test set perplexity for our second evaluation. Table

8 shows a summary overview of the highest achieved results per dataset, along with corresponding lambda values and dataset proportions. The selected results are from the overviews in which training and testing set were the same. In cases where no improvement was observed, we chose the parameters for which the biggest improvements in perplexity were seen on the other testing set. The 2, 3- and 4-gram scores indicate the n-gram *ratios*: the proportion of the observed n-gram divided by the total number of n-grams in a generated sequence. Entropy scores are based on average entropy in prediction throughout the sequence.

Dataset	Regularizer	Proportion	$\lambda$	Improvement
OpenSubtitles	UID Variance	0.25	0.001	0.15%
	UID Loc. cons.	0.25	0.01	0.50%
	UID Max	0.25	0.1	0.17%
Europarl	UID Variance	0.25	0.001	0.21%
	UID Loc. cons.	0.25	0.01	0.25%
	UID Max	0.25	0.01	0.32%

Table 8: Overview of highest improvements in test set perplexity scores

Table 9 shows the results of the analysis of the generated text for both datasets. Contrary to Wei et al. (2021), we mostly observe decreases in lexical diversity despite the perplexity gains. We do see an entropy decrease on the Europarl dataset, and with the Variance regularizer on the OpenSubtitles dataset.

OpenSubtitles					
Model	Seq. Len	2-gram	3-gram	4-gram	Entropy
Baseline	49.85	0.2504	0.5603	0.7642	8.1525
UID Variance	49.86	0.2480	0.5584	0.7656	8.1256
UID Loc. cons.	49.91	0.2356	0.5051	0.6900	8.4489
UID Max	49.87	0.1815	0.4244	0.6262	8.4538

  

Europarl					
Model	Seq. Len	2-gram	3-gram	4-gram	Entropy
Baseline	49.81	0.3609	0.6453	0.7923	3.4902
UID Variance	49.85	0.3546	0.6376	0.7884	3.4834
UID Loc. cons.	49.94	0.3326	0.6047	0.7625	3.4094
UID Max	49.84	0.3518	0.6334	0.7863	3.4587

Table 9: Lexical diversity results of generated text.

For both baseline and the regularizers for both datasets the sequence lengths approximate 50; this suggests that without setting the generating limit to 50 the model would indefinitely generate new text. This indicates a lack of prediction of end-of-sequence tokens. These predictions are contrary to the general structure of the OpenSubtitles dataset which contains relatively short sentences all having an end-of-sequence token. The Europarl dataset also contains an end-of-sequence token in every line of the dataset.

For the Europarl dataset, all three regularizers show slight decreases in entropy compared to baseline, indicating that the model is more confident in its output predictions. The surprisal variance regularizer showed the smallest changes, having an improved 4-gram ratio but a worsening of the 2- and 3-gram ratio and entropy. The local consistency regularizer showed the biggest differences, indicating that penalizing with a modest lambda value of 0.01 a lack of a smooth distribution of

token surprisal levels more strongly affects the generated text with a decrease in entropy (which is good) but also a decrease in lexical diversity (which is bad).

OpenSubtitles			
Model	Surp. var. (avg.)	Loc. cons. (avg.)	Max surp. (avg.)
Baseline	4.9609	9.2163	13.5754
UID Variance	4.9201	9.2189	13.4497
UID Loc. cons.	5.0982	9.3646	13.7136
UID Max	5.2179	10.2833	13.4588

  

Europarl			
Model	Surp. var. (avg.)	Loc. cons. (avg.)	Max surp. (avg.)
Baseline	7.5709	13.6029	16.7345
UID Variance	7.3126	13.3730	16.5591
UID Loc. cons.	6.6190	12.4014	16.5230
UID Max	7.4573	13.6608	16.6505

Table 10: Average surprisal variance, local consistency and max surprisal of the generated text.

Table 10 shows the average surprisal variance, local consistency and max surprisal of the generated text sentences after applying the regularizers compared to baseline scores. For OpenSubtitles, the surprisal variance regularizer delivered the best results, only showing a slight reduction in local consistency. The local consistency regularizer showed only worsening of scores and the max regularizer only an improvement in the average max surprisal. For Europarl, the surprisal variance and local consistency regularizers showed improvements in all metrics, with the local consistency regularizer showing the greatest improvements, particularly in surprisal variance. The max regularizer showed improvements in surprisal variance and max surprisal, but a worsening of local consistency.

#### 4.4 Examples of generated text

In Appendix A, we show various examples of text outputs generated by these models, with and without regularization. All examples are from the 0.25 training dataset proportion, which showed the largest difference in our metrics. For the regularized examples, the local consistency regularizer with lambda value 0.01 was used, which also showed the largest improvements.

For the OpenSubtitles-based models, we observe that models generally generate incoherent text, though some word orders are correct. This is due to the generally limited capabilities of GPT-2 Small compared to larger and more modern generative decoder models, especially in spoken registers. The following contrasting non-regularized and regularized samples from section A.1.2 illustrate this.

Prompt: *Ik geloof je niet.*

Non-regularized sample:

- (1) *ik wil niet dat jij dat deed. ", je gelooft me niet. en de ander had geen keus. en ik ga je laten zien. kom binnen en zeg, ik ben een paar minuten naar binnen, ok? ", wat ga je het*

Regularized sample:

- (2) *ik wil niet dat jij dat deed. ", je gelooft me niet. je begrijpt er niks van. je weet wel van waar je een beetje van het voor me hebt gedaan? en wat denk je daarvan, ok? het was een droom. het*

In these examples, we cannot easily observe the improvements from the local-consistency regularizer that our quantitative metrics indicate. In these particular samples, the non-regularized model (1) moves from the second-person ‘je’ focus to a more interpersonal context, while for the regularized sample (2) the focus stays on the second-person ‘je’. But there is no clear trend of increased topic stability across all the samples in Appendix A.1.

Compared to the Opensubtitles results, the generated output from fine-tuning on Europarl delivers more coherent results. For the prompt “*Wie ben jij?*” (Appendix A.2.1), the non-regularized outputs exhibit noticeable syntactic instability and repetition (for example, “*luisteren en luisteren*”, “*duidelijk duidelijk*”), as well as abrupt topic changes within sentences (Example 3). In contrast, the regularized outputs display more coherent sentence structures with more consistent patterns and more clear progressions, particularly in the use of formal parliamentary language use (Example 4).

Prompt: *Wie ben jij?*

Non-regularized sample:

- (3) *Ik wil niet luisteren en luisteren naar wat ik heb gezegd. Ik wil het debat niet opnieuw verstoren. Ik wil mijn woorden duidelijk duidelijk maken dat de Commissie niet akkoord zal gaan met de resolutie, en dat zij ook van mening is dat die het Parlement niet*

Regularized sample:

- (4) *Ik wil niet luisteren en ik wil niet van de Raad horen of het verslag van collega Cohn-Bendit een beetje duidelijk is geweest op deze twee onderwerpen. Het lijkt wel alsof zij niets aan het spel stond, en geen enkel effect heb.*

The other Europarl examples in Appendix A.2 show similar trends. The non-regularized outputs tend to mix unrelated topics and contain grammatical irregularities or redundant elements (such as repeated words and incomplete outputs). The regularized outputs, however, show improved local consistency, with smoother transitions between clauses. Although repetition is still present, it appears in a more controlled form. The differences are most pronounced for the third prompt (Appendix A.2.3), where the non-regularized model produces unstable outputs, including sequences with excessive punctuation and incoherent parts. In contrast, the regularized model generates more stable and well-formed sentences, maintaining a consistent structure and avoiding random outputs. The text exhibits clearer syntactic continuity and more predictable transitions between clauses.

Overall, these examples suggest that information density-based regularization has a stronger qualitative impact with the Europarl dataset. Further research would be needed to determine what factors cause this difference (e.g. similarity to the pre-training data, lexical diversity, frequency of vocabulary or register/style factors). The regularized model produces outputs that are more syntactically stable, locally coherent, and consistent, supporting the claim that UID-based regularization promotes smoother information flow. At the same time, some repetition remains, indicating that improvements in local consistency may still come at the cost of reduced variability.

## 5. Discussion

Fine-tuning the Dutch GPT-2 model with information density-based regularizers yields small but consistent reductions in test set perplexity, primarily on smaller subsets of the data. This aligns with earlier findings that regularization effects are more pronounced in data-scarce settings (Wei et al. 2021). On larger datasets, improvements are rare and performance often worsens, suggesting diminishing returns when sufficient training data is available.

Across our experiments, the surprisal variance regularizer produces the most frequent perplexity improvements for Dutch, though not always the largest ones. This is the regularizer from Meister et al. (2020b) that was not included in the evaluation of Wei et al. (2021). In contrast, the local con-

sistency and max regularizers achieve stronger improvements in specific configurations, particularly on the 25% dataset subsets. This difference likely reflects their focus on more targeted sequence properties—local smoothness and extreme surprisal values—whereas surprisal variance penalizes overall sequences. Moderate regularization weights ( $\lambda = 0.01$ ) are most consistently effective, indicating that mild penalties outperform both overly strong and overly weak regularization. A more systematic hyperparameter search could further explain these findings.

Cross-dataset evaluation results in substantially higher perplexity scores, reflecting structural and lexical differences between OpenSubtitles and Europarl. Improvements under cross-testing are limited and inconsistent, preventing conclusions about generalizability.

Analysis of generated text shows that lexical diversity and information density capture distinct aspects of quality of generation. For OpenSubtitles, local consistency and max regularization increase n-gram repetition and increase entropy, indicating worse lexical diversity. In contrast, surprisal variance regularization most strongly improves information density metrics, reducing surprisal variance and max surprisal without substantially affecting lexical diversity. For Europarl, changes are smaller overall, with local consistency regularization yielding the largest information density metric gain but also the largest loss in lexical diversity. These results highlight the importance of evaluating multiple dimensions of generative quality.

Overall, we find that the potential of UID-based regularization, observed during the pre-training stage in previous work, also extends to fine-tuning to some extent. We observe improvements in test set perplexity but we do not observe improvements in lexical diversity that were noted in previous work.

Sharma et al. (2024) show decoder language modelling performance improvements by applying regularizers during fine-tuning, but they do not use UID-based regularizers. Our results do show that UID-based regularizers can be beneficially applied in fine-tuning as well as in pre-training, which has not been tested previously, as far as we are aware.

In addition, by performing this experiment on Dutch we show that UID-based regularization also has potential for Dutch. Lastly, as the effectiveness of UID-based regularization has been taken as evidence for the Uniform Information Density hypothesis in theories of language processing, we also consider our findings to be evidence for the UID hypothesis in Dutch. UID in Dutch has been hypothesized to affect word order choice in the verbal cluster (Bloem 2016), as well as diachronic changes in verb cluster word order (Olthof et al. 2017).

## 5.1 Limitations

One limitation of our study is that the regularization techniques are only applied during fine-tuning, possibly making them less effective than when applied from the start in pre-training. Previous research applied the regularizers during pre-training (Wei et al. 2021). Given their consistent improvements across languages, it might be expected that these regularizers would also benefit Dutch language model pre-training, but we have not been able to validate this at the GPT-2 scale.

Another limitation is that, because earlier results for English were observed at the pre-training stage, we cannot directly compare our Dutch results at the fine-tuning stage to those results. Such a comparison would make it possible to see which language benefits more from UID regularization in language models. As outlined in section 2.1, Dutch has more options for word order variation, providing speakers with more opportunities to manage information density, so it might benefit more than English. It would be interesting to explore how language modelling for different languages is affected by UID regularization more broadly. Such a study would have to consider languages with more fixed word order (e.g. English) and more free word order (e.g. Russian), as well as the amount of morphological marking a language has. This factor interacts with word order typology, with free word order languages typically having more morphological marking.

A further limitation is that our evaluation of generated text was limited to automated quantitative metrics. A human judgement experiment would be required to assess whether human readers

consider text from UID-regularized language models to be more natural or easier to read than text produced by models with the standard MLE language modelling objective.

There are also some general limitations of this line of work that have not been reflected on much in the literature. Firstly, surprisal is a relative quantity, so while we can try to make it more uniform, it is not clear whether there are any guarantees that models do not settle on a consistently very high or consistently very low information density. A comparison to information density in human-written text could be used to show whether this happens, but further research is needed to see whether this is preventable, if it happens. Secondly, information-theoretic surprisal is a limited operationalization of the concept of informativeness in general. Aspects such as the semantic informativeness or the novelty of the information to the reader are not necessarily covered by this operationalization. Other approaches would be needed to optimize language modelling for these other types of informativeness.

## 6. Conclusion

This study examined the effect of information density-based regularization during fine-tuning of a Dutch pretrained GPT-2 model. Results show that adding regularizers can yield small improvements in uniform information density, particularly when training data is limited. This adds to previous results where such regularizers have been shown to be effective in pre-training for a variety of other languages, with smaller fairseq models.

Among the tested methods, surprisal variance regularization most consistently improves information density metrics across both datasets, reducing surprisal variance and extreme surprisal values. Local consistency regularization is especially effective for the more structured Europarl data, improving both local predictability and information flow. The max regularizer shows limited impact on information density but increases lexical diversity for OpenSubtitles.

Overall, these findings suggest that UID-based regularization during fine-tuning can modestly steer language models toward more evenly distributed information without necessarily increasing lexical diversity. While effects are limited in scale and sensitive to data size and regularization strength, the results indicate that information-theoretic regularizers offer a viable mechanism for influencing the information structure of generated text and imitating human language processing patterns.

## References

- Bloem, Jelke (2016), Testing the processing hypothesis of word order variation using a probabilistic language model, *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pp. 174–185.
- Booij, Geert (2019), *The morphology of Dutch*, Oxford University Press.
- Bouma, Gosse (2017), Om-omission, in Wieling, Martijn, Martin Kroon, Gertjan van Noord, and Gosse Bouma, editors, *From Semantics to Dialectometry: Festschrift in honor of John Nerbonne*, Tributes, College Publications, pp. 65–73.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020), Language models are few-shot learners, *Advances in neural information processing systems* **33**, pp. 1877–1901.
- Chapuis, Emile, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel (2020), Hierarchical pre-training for sequence labelling in spoken dialog, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, pp. 2636–2648.

- Clark, Thomas Hikaru, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy (2023), A cross-linguistic pressure for uniform information density in word order, *Transactions of the Association for Computational Linguistics* **11**, pp. 1048–1065, MIT Press.
- De Hoop, Helen (2003), Scrambling in Dutch: Optionality and optimality, *Word order and scrambling* pp. 201–216, Wiley Online Library.
- de Vries, Wietse and Malvina Nissim (2021), As good as new. How to successfully recycle English GPT-2 to make models for other languages, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 836–846.
- Fenk, August and Gertraud Fenk (1980), Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß?, *Zeitschrift für experimentelle und angewandte Psychologie* **27** (3), pp. 400–414.
- Frank, Austin F and T Florain Jaeger (2008), Speaking rationally: Uniform information density as an optimal strategy for language production, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 30.
- Gamboa, John CB, Leigh B Fernandez, and Shanley EM Allen (2024), Investigating the uniform information density hypothesis with complex nominal compounds, *Applied Psycholinguistics* pp. 1–46, Cambridge University Press.
- Hale, John (2001), A probabilistic Earley parser as a psycholinguistic model, *Second meeting of the North American Chapter of the Association for Computational Linguistics*.
- Hawkins, John A. (2012), The drift of English toward invariable word order from a typological and Germanic perspective, *The Oxford Handbook of the History of English*, Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199922765.013.0053>.
- Jaeger, T and Roger Levy (2006), Speakers optimize information density through syntactic reduction, *Advances in neural information processing systems*.
- Jaeger, T Florian (2010), Redundancy and reduction: Speakers manage syntactic information density, *Cognitive psychology* **61** (1), pp. 23–62, Elsevier.
- Koehn, Philipp (2005), Europarl: A parallel corpus for statistical machine translation, *Proceedings of Machine Translation Summit X: Papers*, pp. 79–86.
- Levy, Roger (2008), Expectation-based syntactic comprehension, *Cognition* **106** (3), pp. 1126–1177, Elsevier.
- Lison, Pierre and Jörg Tiedemann (2016), Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles, European Language Resources Association.
- Mahowald, Kyle, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson (2013), Info/information theory: Speakers choose shorter words in predictive contexts, *Cognition* **126** (2), pp. 313–318, Elsevier.
- Maurits, Luke, Dan Navarro, and Amy Perfors (2010), Why are some word orders more common than others? A uniform information density account, *Advances in neural information processing systems*.

- Meister, Clara, Elizabeth Salesky, and Ryan Cotterell (2020a), Generalized entropy regularization or: There’s nothing special about label smoothing, in Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 6870–6886.
- Meister, Clara, Ryan Cotterell, and Tim Vieira (2020b), If beam search is the answer, what was the question?, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2173–2185.
- Neeleman, Ad and Fred Weerman (1997), L1 and L2 word order acquisition, *Language Acquisition* **6** (2), pp. 125–170, Taylor & Francis.
- Oh, Byung-Doh, Shisen Yue, and William Schuler (2024), Frequency explains the inverse correlation of large language models’ size, training data amount, and surprisal’s fit to reading times, in Graham, Yvette and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, St. Julian’s, Malta, pp. 2644–2663. <https://aclanthology.org/2024.eacl-long.162/>.
- Olthof, Marieke, Maud Westendorp, Jelke Bloem, and Fred Weerman (2017), Synchronic variation and diachronic change in Dutch two-verb clusters, *Tijdschrift voor Nederlandse Taal-en Letterkunde* **1** (1), pp. 34–60.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019), Language models are unsupervised multitask learners, *OpenAI blog* **1** (8), pp. 9.
- Shannon, Claude E (1948), A mathematical theory of communication, Vol. 27, Nokia Bell Labs, pp. 379–423.
- Sharma, Mandar, Nikhil Muralidhar, Shengzhe Xu, Raquib Bin Yousuf, and Naren Ramakrishnan (2024), Information guided regularization for fine-tuning language models, *First Conference on Language Modeling*.
- Torabi Asr, Fatemeh and Vera Demberg (2015), Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission, in Purver, Matthew, Mehrnoosh Sadrzadeh, and Matthew Stone, editors, *Proceedings of the 11th International Conference on Computational Semantics*, Association for Computational Linguistics, London, UK, pp. 118–128.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need, *Advances in neural information processing systems*.
- Wei, Jason, Clara Isabel Meister, and Ryan Cotterell (2021), A cognitive regularizer for language modeling, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1, Association for Computational Linguistics, pp. 5191–5202.
- Xu, Xinnuo, Ondřej Dušek, Ioannis Konstas, and Verena Rieser (2018), Better conversations by modeling, filtering, and optimizing for coherence and diversity, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3981–3991.

## Appendix A. Examples of generated text

This appendix contains some examples of model-generated text, so that readers who read Dutch can get an impression of the differences beyond the quantitative results presented in the paper. Generated texts are shown for equivalent models without and with UID regularization, trained on the two different text types, for three simple prompts. Per prompt, 3 samples of generated output are shown. All examples are from the 0.25 training dataset proportion, which showed the largest difference in our metrics. For the regularized examples, the local consistency regularizer with lambda value 0.01 was used, which also showed the largest improvements.

### A.1 OpenSubtitles

#### A.1.1 PROMPT 1

Prompt: *Wie ben jij?*

Non-regularized samples:

1. *. ik was een vrouw. maar dat niet. ik dacht, ga gewoon... hier weg, anders is ik dan dat. -ik ben niet echt verliefd, ik zou wel liever je hebben uitgenodigd moeten hebben, de volgende keer dan mijn vader. en*
2. *ik ben hier vanavond om te bedanken voor wat me betreft... voor deze mooie jurk. en de jas van de hele wereld gekocht had. 't Is misschien de ultieme eer. voor je om te bedanken dat ik dit heb gekregen, voor de mooie*
3. *. ik ben niet het man. dat is 't. 't is 't niet. en hij is 'n boer. en als er iets in 't water kookt?. 't Is je schuld. en je houdt het op. '*

Regularized samples:

1. *ik was een vrouw. 't Was mijn bedoeling. 'n Ritz, maar daar kan je het over krijgen. 'n Necromd, is je niet? en wanneer ben je in hemelsnaam de volgende?. 't is niet*
2. *'n Hetero?. 'n Een hetero. 'n Dat is gewoon een klootzak!. 'n Wat?. 'n In Wat?!. wat is er dan?. 'n Wat ben jij? een klootzak, 'n*
3. *. ik ben niet het man. dat is 't. 't is 't niet. 't is 'n boer. 't is 'n klootzak. 'n stomme boer, dat is je. jij is de klootzak! 'n klootzak*

#### A.1.2 PROMPT 2

Prompt: *Ik geloof je niet.*

Non-regularized samples:

1. *niet zoals deze. het is een leugen. in het begin. het is een leugen, niet zoals die. het is een droom. je droom. het is een droom, het is een droom. dit is geen droom. het is een droom.*
2. *ik wil niet dat jij dat deed. ", je gelooft me niet. en de ander had geen keus. en ik ga je laten zien. kom binnen en zeg, ik ben een paar minuten naar binnen, ok? ", wat ga je het*
3. *ik voel me niet schuldig. en ben je gewoon een sukkel? het spijt me echt. ik wil jullie allemaal bedanken voor deze boot, maar dit is de enige wat ik voel. "t lijkt alsof je het moet doen. ik was niet echt voorbereid*

Regularized samples:

1. *niet zoals deze. het is een leugen. in het begin. het is een leugen, niet zoals die. niet waar, niet waar. ik ben het niet. ik weet het niet. ik weet dat het geen leugen is, weet je wat het*
2. *ik wil niet dat jij dat deed. ”, je gelooft me niet. je begrijpt er niks van. je weet wel van waar je een beetje van het voor me hebt gedaan? en wat denk je daarvan, ok? het was een droom. het*
3. *ik voel me niet schuldig. en ben niet erg goed. ik ben zo vreselijk. 't Is de meest fantastische kamer dat ik ooit had. 't zou zo makkelijk geweest zijn... maar ik heb 'm wel gezien. ik was niet echt goed*

### A.1.3 PROMPT 3

Prompt: *Gaan we naar huis?*

Non-regularized samples:

1. *..... we hoeven wel. je bent gewoon te groot. niet zo slim. nee, niet zo slim. niet zo slim. geen idee wat het beste is, ik was vergeten. ik wilde je vragen om geld terug te vragen. voor me uit*
2. *. ik ga er niet heen. ik moet je een gunst bewijzen. ik moet mijn auto lenen. dat is je gelukt, je weet nog wel... maar je weet wel. dat ze gaan daar niet weg. ik denk, toch ben je thuis.*
3. *. -nee, wacht. lk bel niet. -geen probleem. ga maar. ga niet daarheen. -wat dan? jij blijft hier even. de rest is nog even hier. en neem daar wat eten, kom me een fles water.*

Regularized samples:

1. *- we hoeven wel. waar ga je naar binnen? niet zo erg. nee, niet zo erg. nee, ik ga geen auto lenen en we gaan terug. - de auto nemen we. en het geld niet, de caravan. ik zal*
2. *'t Gaat niet om een hond. 't Kan niet. ik had geen enkele vraag stellen. 't Kan niet. 't Komt omdat ik 'm niet herinner dat ze de bus niet hebben. 't heeft toch wel geduurd? dat*
3. *nee. ln bed, ik ben net bezig. hé, ga maar wat van die? 't is 't nu even? de eerste keer... ik weet 't niet. oké. dat maakt me een beetje nerveus om*

## A.2 Europarl

### A.2.1 PROMPT 1

Prompt: *Wie ben jij?*

Non-regularized samples:

1. *zegt dat het zo zou moeten zijn, of niet. In dit Huis zegt dit en wel bij een andere. Het Parlement heeft zich in het ontwerp voor een begrotingsprocedure geplaatst om deze zaak aan te houden.. De Commissie wil hier een oplossing voor zoeken*
2. *Ik wil niet luisteren en luisteren naar wat ik heb gezegd. Ik wil het debat niet opnieuw verstoren. Ik wil mijn woorden duidelijk duidelijk maken dat de Commissie niet akkoord zal gaan met de resolutie, en dat zij ook van mening is dat die het Parlement niet*
3. *? zegt het: hoe ver gaan?... ik heb het niet over de Raad van ministers. Ik zal de situatie op de voet volgen. ”Wat is er precies in dit verslag terechtgekomen, die betrekking heeft op de noodzaak om deze besluiten te nemen?*

Regularized samples:

1. *zegt dat het zo zou moeten zijn, en niet u, maar de heer Barón Crespo, die de hele dag aanwezig is, ook niet als eerste die deze kwestie aan de orde heeft gesteld. Het is een feit dat wij in de Raad*
2. *Ik wil niet luisteren en ik wil niet van de Raad horen of het verslag van collega Cohn-Bendit een beetje duidelijk is geweest op deze twee onderwerpen. Het lijkt wel alsof zij niets aan het spel stond, en geen enkel effect heb.*
3. *Het zou misschien erger zijn om dit antwoord te herhalen, maar het zou ook een goed idee zijn om de Commissie dit te onthouden. Het Europees Parlement is er, zoals de heer Cohn-Bendit zojuist heeft gezegd, namelijk van overtuigd dat*

#### A.2.2 PROMPT 2

Prompt: *Ik geloof je niet.*

Non-regularized samples:

1. *Het is geen recht! ik ben zelf ook niet een voorstander van een verbod op het gebruik van alcohol. "Mijnheer de Voorzitter, als we het om een bepaalde wijn niet moeten proberen, dan zou ik nog een motie van orde indienen in de vorm*
2. *Wij doen het met respect voor die mensen die de mensenrechten respecteren en op grond van hun lot in hun eigen land blijven blijven.Het Europees Parlement heeft ons, dat de mensenrechten in de Europese Unie altijd gerespecteerd, al te duidelijk uitgesproken tegen haar verzoek.*
3. *Mevrouw de Voorzitter, ik wilde graag zien dat u dit onderwerp van gesprek niet opnieuw noemt. Mevrouw de commissaris, ik kan u echter niet beletten met u te spreken over het geval van de Koerdische vluchtelingen en de situatie in Albanië. Ik zal u daarom niet*

Regularized samples:

1. *Het is geen recht! Het is uw zaak! Het is het feit dat het geld wel terug zou moeten vorderen. Het is waar dat de kosten voor de ontwikkelingshulp niet toereikend zijn, ook al hebben we geen gemeenschappelijk landbouwbeleid in stand gehouden*
2. *Wij doen het met respect voor die mensen die de mensenrechten hebben geschonden, als ik weet dat in hun eigen land een rechtbank in het Verenigd heeft vastgehouden. Het slachtoffer is de doodstraf, en het slachtoffer is de doodstraf. Het land van de doodstraf. Het*
3. *Mevrouw de Voorzitter, ik wilde graag zien dat u dit onderwerp van gesprek met deze commissaris wilt voeren, zowel bij het Vragenuur voor vragen over de begroting als bij deze vergadering. Het zal misschien wel een tijdverdrijf zijn om uw toespraak van vanavond tot*

#### A.2.3 PROMPT 3

Prompt: *Gaan we naar huis?*

Non-regularized samples:

1. *???????? te stappen naar het bureau voor wederopbouw????????????????????????????????*
2. *Ik denk dat dat zo is. Van huis? Het is moeilijk om aan te trekken wat er in Turkije gebeurt en wie van de Turkse regering houdt.? En wat in Turkije gebeurt in Turkije? In het verslag van de secretaris-generaal van de Verenigde*

3. *We zijn daar altijd op de aangewezen plaats. Ik wil u vragen of u de commissaris heeft verteld wat er voor ogen is, aangezien u in het debat een aantal vragen stelt die niet in de orde zijn: wanneer denkt u over de toekomst van de informatie*

Regularized samples:

1. *Het Verenigd Koninkrijk of naar mijn eigen land, of naar Finland? Wat zouden dan deze twee landen doen? Het is niet duidelijk welk regime dat heeft gevoerd. Het Verenigd Koninkrijk gaat er ook op aan. Het Verenigd Koninkrijk heeft er zijn eigen burgers te vertegenwoordigen*
2. *Ik denk dat dat zo is. ' . '?... De ministers in de nationale parlementen hebben nu duidelijk over de Top van Tampere in Lissabon besloten om nu al het mogelijke te doen om een besluit ter plekke tot een besluit van het*
3. *We zijn daar altijd op de aangewezen plaats. Ik wil u vragen of u de commissaris heeft verteld wat er voor ogen is, aangezien u in het Verenigd Koninkrijk, waar wij het niet in grote lijnen mee hebben, geen andere bevoegdheden. Het is te verwachten*