# Swedish Language Complexity Generator for Adaptive Language Learning

LT2326 H24 Machine learning for statistical NLP: Advanced
Final Project
Tianyi Geng

## 1.Introduction

My project intends to generate Swedish sentences of different complexity for adaptive language learning. The project idea originates from my personal project of building a language game for learning Swedish. In my game, the player will meet different NPCs that speak Swedish of different adjustable difficulty levels, which, in this case, are called ”complexity levels”. The complexity level ranges from 1 to 5, as simple to the most difficult. For example, if the complexity is set to level 1, the NPC says *Jag spelar piano* (I play piano), while a level-5 NPC will say something like *Jag förlorar mig i de harmoniska tonerna från pianot, där varje ackord reflekterar min passion för musik* (I lose myself in the harmonious tones of the piano, where each chord reflects my passion for music).

As one of the state-of-the-art large language model, ChatGPT is able to complete this task with sufficient competence:

[Below are generated Swedish sentences of 1-5 complexity level by ChatGPT]
**Level 1 (Simple)**:
*Jag gillar att spela datorspel.*
(I like to play video games.)

**Level 2 (Basic)**:
*Min vän och jag brukar spela datorspel tillsammans efter skolan.*
(My friend and I usually play video games together after school.)

**Level 3 (Intermediate)**:
*När jag var yngre, brukade jag tillbringa flera timmar med att utforska olika spel på datorn.*
(When I was younger, I used to spend several hours exploring different games on the computer.)

**Level 4 (Advanced)**:
*Trots att många tror att datorspel bara är underhållning, tycker jag att de kan hjälpa till att utveckla viktiga färdigheter som problemlösning och samarbete.*
(Although many people think that video games are just entertainment, I believe they can help develop important skills such as problem-solving and teamwork.)

**Level 5 (Complex)**:
*Även om vissa forskningsstudier har pekat på potentiella negativa effekter av överdrivet spelande, har andra rapporter visat att strategiska och interaktiva datorspel kan förbättra kognitiva funktioner och främja sociala interaktioner i digitala miljöer.*
(Even though some research studies have pointed out potential negative effects of excessive gaming, other reports have shown that strategic and interactive video games can enhance cognitive functions and promote social interactions in digital environments.)

Through talking with ChatGPT, it seems to me that the underlying criteria for such level generation is the Common European Framework of Reference for Languages (CEFR), where low-level sentences are likely constructed with relatively common words and shorter length, while high-level sentences are likely using verbs of past tenses, more advanced words, and more complex grammar structures.

Following the lately released paper ”From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation” (Ali et al. 2024) where a CEFR-Aligned Language Model was built for this task, my projects focused on recreating similar results by fine-tuning smaller scaled open-source LLMs.

## 2.Background

A series concepts are relevant in defining ”language complexity” and many of them are especially influential in the field of second language acquisition (SLA). Flesch (1948) refers to ”text difficulty based on sentence length and syllable count” for the concept of ”readability”. While sentence length being one of the most obvious elements contributing to the difficulty of the texts, other elements in language use such as ”syntactic complexity”, ”lexical complexity”, ”grammatical accuracy”, and ” coherence in writing”, introduced in the CAF (”complexity”, ”accuracy” and ”fluency”) model proposed by Housen et al. (2009) exhibit different layers in analyzing language difficulty. In 2001, the Common European Framework of Reference for Languages (CEFR) was developed by the Council of Europe in 2001 and it provides a comprehensive system for standardizing language proficiency with 6 levels from A1 to C2. The CEFR has predominant influence globally and thereafter has become the guideline for many European languages tests and a thorough framework for language evaluation and research.

Despite being widely-used officially and well-structured, CEFR level descriptions (e.g. for reading, C1 level refers to ”I can understand specialized articles and longer technical instructions, even when they do not relate to my field”) are fundamentally qualitative rather than quantitative, thus resulting in the lack of precision. Furthermore, the data used for training any CEFR related LLM are human-labeled so the model can perform efficiently well in assessing CEFR levels through learning the patterns but is not necessarily being able to quantify the concrete differences of each level.

Ali et al. (2024) have trained a CALM (CEFR-Aligned Language Model) which is built by fine-tuning open-source models like ”LLama2-7B” and ”Mistral-7B” and is claimed to be able to outperform GPT-4 in generating language of proficiency control. Since my project is about Swedish generation, I will use other pertained open-source models that focus on Swedish data and explore how to fine-tune the models for this task.

## 3.Methodology

For models, I have tried the LLama3-7B model but ended up using the ”gpt-sw3” models provided by AI Sweden on HuggingFace due to the unstable performance of the former one in generating Swedish. The gpt-sw3-356m model and gpt-sw3-1.3B models are used for the suitable sizes to run on the server GPUs. The prompts are written in Swedish instead of bilingual or English to ensure the quality of the generated Swedish texts.

I have prepared my own datasets that are labeled with 1-5 complexity levels. I also used LoRA fine-tuning tools due the capacity limits of the MLTgpu.

## 4.Data

My training datasets are made of around 200 instances of ”input sentence (text)”, ”input sentence complexity level (number from 1 to 5)”, ”target output sentence complexity level (number from 1 to 5)” and ”target output sentence (text)”. The texts cover different topics from ”psychology and mental health” to ”media and entertainment”. The data with reversed input/output are also included for bidirectional generation training.

Example instances are as follows:

Input Sentence : Jag gillar musik.
Input Sentence Complexity Level: 1
Target Output Sentence Complexity Level: 5
Target Output Sentence: Många av mina vänner är musiker, och vi brukar ha jamkvällar där vi spelar olika instrument tillsammans.

Input Sentence : Många av mina vänner är musiker, och vi brukar ha jamkvällar där vi spelar olika instrument tillsammans.
Input Sentence Complexity Level: 5
Target Output Sentence Complexity Level: 1
Target Output Sentence: Jag gillar musik.

## 5.Results

To test the result, I used ”Jag gillar att läsa böcker” as input sentence of complexity level ”1” and asked the model to generate a corresponding sentence of complexity level ”5”. Notably, the gpt-sw3-1.3B model showed worse performances than than gpt-sw3-356m model. The results also varied largely across different prompts. For instance, the model didn’t work very well with abstract or complicated prompts and tended to produce output that was either irrelevant or was a repetition of the input sentence (Table 1).

| Model | Prompt | Generation before Fine-tuning | Generation after Fine-tuning |
|---|---|---|---|
| gpt-sw3-356m | För varje exempelmening och dess svårighetsnivå, skapa en ny mening på svenska som har en målsvårighetsnivå. Använd nedanstående format: Exempelmening:{Input Sentence}, Svårighetsnivå för exempelmening:{Input Sentence Complexity}, Målsvårighetsnivå: {Target Sentence Complexity}, Generera en mening som matchar målsvårighetsnivån: | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): 1.Jag läser böcker för att jag tycker att det är roligt.2.Jag tycker att böcker är bra<br>för att de är roliga att läsa.3.Böcker är bra eftersom de är bra att läsa för att det finns många olika typer<br>av böcker.4.Böckerna är bra därför att de har många | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): 10. Exempel på exempelmeningar är:Jag tycker att det är roligt att spela fotboll,Jag tycker om att träna,Jag<br>är glad över att jag är bra på att läsa,Jag gillar böcker,Jag vill spela fotboll.<br>Tips för att skapa en meningsfull mening:<br>1. Skapa en |
| gpt-sw3-356m | Original: {input_text}<br>Nuvarande nivå: {input_level}<br>Målnivå: {target_level}<br>Svar: | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): Jag vill ha en bok som ger en inblick i mitt liv. | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): Jag gillar att läsa böcker |

| | | | |
|---|---|---|---|
| gpt-sw3-1.3B | För varje exempelmening och dess svårighetsnivå, skapa en ny mening på svenska som har en målsvårighetsnivå.<br>Använd nedanstående format:<br>Exempelmening:{Input Sentence},<br>Svårighetsnivå för exempelmening:{Input Sentence Complexity},<br>Målsvårighetsnivå: {Target Sentence Complexity},<br>Generera en mening som matchar målsvårighetsnivån: | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): Skriv först en målsvårighetsnivå och en matchande mening<br>i följande ordning:<br>(1) Exempelmening,<br>(2) Svårighetsnivå,<br>(3) Målsvår | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): Använd nedanstående format:<br>Exempelmening:Jag tycker om att läsa böcker,<br>Målsvårighetsnivå:1,<br>Svårighetsnivå för exempelmening:2,<br>Generera en mening som matchar |
| gpt-sw3-1.3B | Original: {input_text}<br>Nuvarande nivå: {input_level}<br>Målnivå: {target_level}<br>Svar: | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): 2<br>Antal böcker: 16<br>Senaste bok: Förlorarnas son av Paulo Coelho<br><br>Jag har läst:<br>Jag gillar att läsa böcker<br>Nuvarande nivå: 5<br>Målnivå: | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): 3 |
| gpt-sw3-1.3B | ### Uppgift: Skriv om texten till en annan svårighetsgrad.<br><br>Texten: {input_text}<br>Nuvarande svårighetsgrad: {input_level}<br>Önskad svårighetsgrad: {target_level}<br>Ny text: | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): Jag gillar att läsa böcker.<br>Nuvarande svårighetsgrad: 1<br>Önskad svårighetsgrad: 5<br><br>Öka | Input (Level 1): Jag gillar att läsa böcker<br>Generated (Level 5): Jag gillar att läsa böcker<br><br>Texten: Jag gillar att läsa böcker<br>Ny svårighetsgrad: 1<br><br>Texten: |

Table 1

## 6.Analysis

Both of the gpt-sw3-356m model and gpt-sw3-1.3B model have managed to generate Swedish texts but demonstrated significant limitations in handling the task of complexity transformation. While the overall before- and after-generation performance isn't good, fine-tuning does not seem to improve the performance; on the contrary, the model seems more "confused" after fine-tuning.

The poor performance after fine-tuning could be attributed to:

- The quality of the dataset: There have been only about 200 instances in the dataset, which could be too little for such a difficult task, making the fine-tuning trivial. There is also a lack of coherent complexity progression in the dataset which could result in unstable output.
- Uncertainty in fine-tuning process: The LoRA parameters and relevant fine-tuning configurations might not have been set correctly.

The poor performance before fine-tuning could be attributed to:

- Prompt strategy: It is difficult to give explicit and concise instructions in prompts. The model struggled to generate the desired output when the prompt is too long.
- The task being too unclear and complicated: Most of the time the model couldn't understand language complexity and the desired outcome very well and deviated from the task after having generated the target sentence. Perhaps splitting the task into several sub-tasks will produce more stable and controllable results.

## 7.Discussion

The task of controlled complexity generation has shown several notable challenges:

- The ambiguity in the definition of language complexity itself
- The relationship between sentence length and complexity is not always linear or straightforward, and short-long sentences transformation is a tough task for the model
- The need to maintain the basic semantic meaning of the sentences while increasing/decreasing complexity

For future work, I would improve in these ways:

- Narrow down to one focus of language complexity (e.g. to "Lexical complexity" only)
- Follow the prompt strategies outlined by Ali et al. (2024) and align the complexity levels with CEFR levels with explicit level definitions.
- Break down the task into smaller, controllable sub-tasks (e.g. "Generate a level-2 sentence with the given level-1 sentence", "Generate a level-1 sentence with the given level-2 sentence")
- Develop a larger and more refined dataset with clear progression of complexity levels (ideally with a detailed guideline categorizing vocabulary into different CEFR levels)
- Experiment with other models and fine-tuning methods

## 8.Conclusion

In conclusion, despite my current implementation of the Swedish complexity generation using gpt-sw3-356m and gpt-sw3-1.3B models demonstrated limited outcomes, this project still has a lot of potential in reaching useful results that can contribute to the usage of small-scale language models on specific tasks (e.g. intelligent computer-assisted language learning, or "ICALL"). The challenges I have encountered in the prompting and fine-tuning current open-source models suggested there are a lot of work to do in the task clarification, dataset making and technical implementation. My future focus will be further narrowing down the task and prioritizing the refined, simple-task-specific generation of usable results.

## Reference


Language Complexity:

Mohammadi, H., & Khasteh, S. H. (2022). Text Difficulty Classification by Combining Machine Learning and Language Features. International Journal of Information Technology and Computer Science, 14(1), 1-13. https://doi.org/10.5815/ijitcs.2022.01.01

Vajjala, S., & Meurers, D. (2016). Characterizing Text Difficulty with Word Frequencies. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 84-94. Association for Computational Linguistics

Flesch, R. (1948). *A New Readability Yardstick*. *Journal of Applied Psychology*, 32(3), 221–233. doi:10.1037/h0057532.

Skehan, P. (2009). Modelling second language performance: Integrating Complexity, Accuracy, Fluency, and Lexis. Applied Linguistics, 30(4), 510–532. doi:10.1093/applin/amp047.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Malik, A., Mayhew, S., Piech, C., & Bicknell, K. (2024). From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation.

Model:

AI-Sweden-Models/gpt-sw3-356m
https://huggingface.co/AI-Sweden-Models/gpt-sw3-356m/tree/main

AI-Sweden-Models/gpt-sw3-1.3B
https://huggingface.co/AI-Sweden-Models/gpt-sw3-1.3b

AI-Sweden-Models/Llama-3-8B
https://huggingface.co/AI-Sweden-Models/Llama-3-8B