Homework

Vlad-Cristian Matei
Departamentul Computer Science
National University of Science and Technology POLITEHNICA
Bucharest, Romania
`vlad_cristian.matei@stud.acs.upb.ro`

1 Dataset

For completing the task, I tried several preprocessing approaches for the dataset. First, I attempted to normalize the text, then formatted the text in the following structure: "[${author}] ${text}". The two approaches were as follows: one involved taking the lines of text exactly as they appeared in the dataset, while the other involved merging consecutive lines from the same author, thereby constructing a direct dialogue with longer lines for each character.

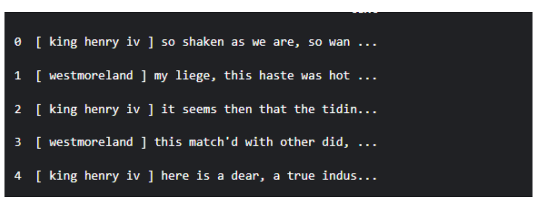


Figure 1: Preprocessing

In the end, I proceeded with the second method, where I created longer, unified lines.

2 Tokenization

For completing the task, I used both proposed tokenizers, the character-level tokenizer and the imported one. Specifically, for the subword tokenizer, I used "gpt2" from the transformers.GPT2Tokenizer library.

```
tokenizer = SubTok()
print(tokenizer.pad_token, tokenizer.pad_token_id)
text = "Anne, has apples."
tokens = tokenizer.tokenize(text)
print("Tokens:", tokens)
ids = tokenizer.convert_tokens_to_ids(tokens)
print("Token IDs:", ids)
reconstructed_tokens = tokenizer.convert_ids_to_tokens(ids)
print("Reconstructed Tokens:", reconstructed_tokens)
reconstructed_text = tokenizer.detokenize(reconstructed_tokens)
print("Reconstructed Text:", reconstructed_text)
reconstructed_text_via_decode = tokenizer.decode(ids)
print("Reconstructed Text via Decode:", reconstructed_text_via_decode)
```

```
tokenizer_config.json:   0%|          | 0.00/26.0 [00:00<?, ?B/s]
vocab.json:   0%|          | 0.00/1.04M [00:00<?, ?B/s]
merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/1.36M [00:00<?, ?B/s]
config.json:   0%|          | 0.00/665 [00:00<?, ?B/s]
<PAD> 50259

Tokens: ['<SOS>', 'Anne', ',', 'Ġhas', 'Ġapples', '.', '<EOS>']

Token IDs: [50257, 43227, 11, 468, 22514, 13, 50258]

Reconstructed Tokens: ['<SOS>', 'Anne', ',', 'Ġhas', 'Ġapples', '.', '<EOS>']

Reconstructed Text: Anne, has apples.

Reconstructed Text via Decode: Anne, has apples.
```

Figure 2: SubWord tokenization

3 Models

I trained multiple models to address the task. However, in the final form, I will present only four of them.

3.1 Small Model - Character Tokenizer

During the experiments, I settled on the following configuration: a batch size of 128, a maximum sequence length of 512, an embedding dimension of 384, 6 attention heads and layers, and a hidden dimension of 128, and Adam Optimizer with a learning rate of 1e-4.

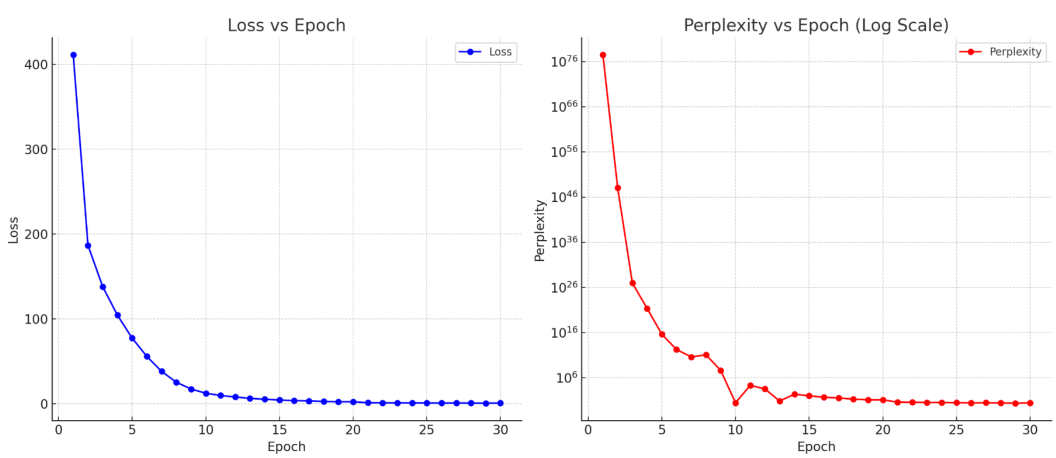


Figure 3: Small Model - Character Tokenizer

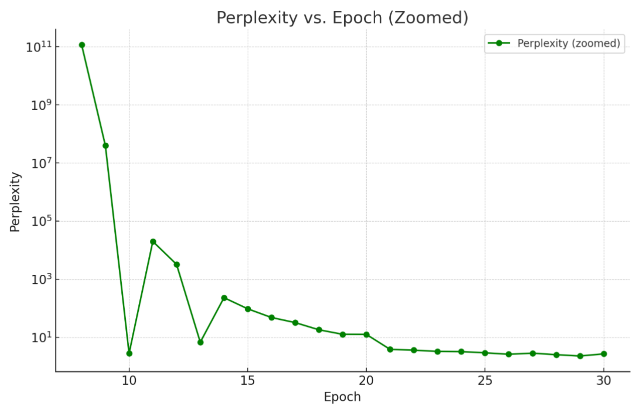


Figure 4: Small Model - Character Tokenizer

The evolution of the losses is shown in the figures below. As can be observed, there is a convergence, with the final loss on the test dataset being 1.3589. Additionally, the last perplexity value calculated in the final epoch is 2.69.

Qualitative Evaluation

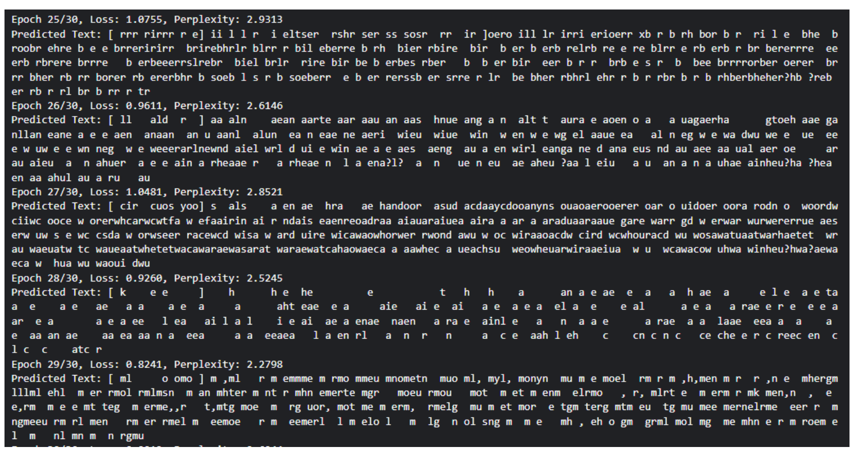


Figure 5: Qualitative Evaluation - Small Model Character Tokenizer

Regarding qualitative evaluation, although the model seems to converge, I don't think it generates well. Even though it starts from the premise of being a model without pretraining, the results don't seem to illustrate anything very coherent. Different word forms are not being formed with each epoch; the model tends to develop a particular style in which it prioritizes certain letters more or less. However, there are instances where it seems to attempt to write different words.

It's not possible to talk about a BLEU score because appropriate words can't be identified, nor are the letters arranged in an intuitive, suitable order.

3.2 Big Model - Character Tokenizer

This time, I tried to scale up the model. Thus, the configuration looked as follows: batch size of 64, max sequence length of 512, embedding dimension of 512, attention heads and layers equal to 8, and hidden dimension of 1024. Additionally, I used the Adam optimizer with a learning rate of 1e-4. Everything was trained for 90 epochs, using a GPU P100, with the condition to fit within the 12 hours of resources provided by Kaggle.

Thus, the number of parameters increased from 786,432 to 25,165,824.

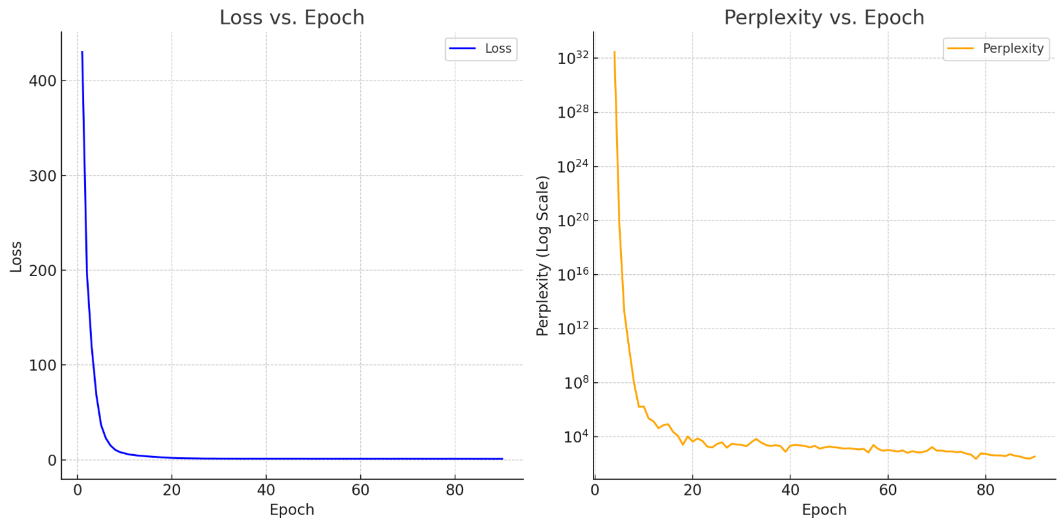


Figure 6: Big Model - Character Tokenizer

As can be seen in the figure above, the model saturates. Perhaps, if I had trained it for a longer period, I would have reached lower scores, as seen in Figure 7, but in a way, with this configuration, these were the results.

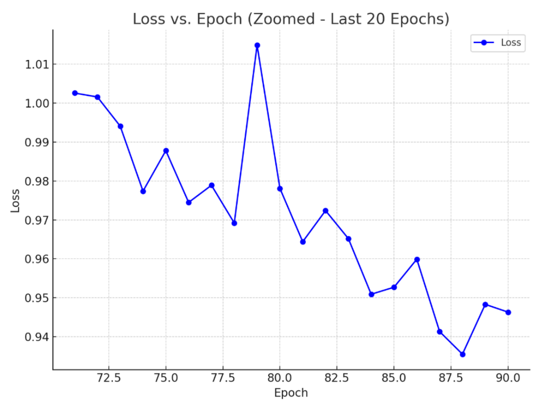


Figure 7: Big Model - Character Tokenizer - Loss

Qualitative Evaluation

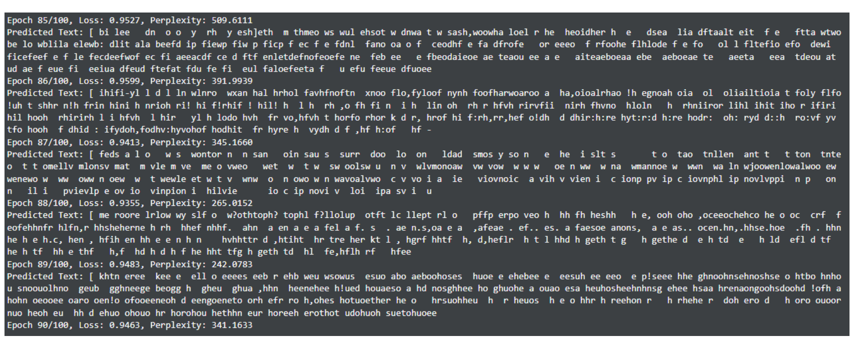

```
Epoch 85/100, Loss: 0.9527, Perplexity: 509.6111
Predicted Text: [ bi lee   dn  o o  y  rh  y esh]eth  m thmeo ws wul ehsot w dnwa t w sash,woowha loel r he  heoidher h  e   dsea  lia dftaalt eit  f e   ftta wtwo
be lo wblila elewb: dlit ala beefd ip fiewp fiw p ficp f ec f e fdnl  fano oa o f  ceodhf e fa dfrofe   or eeeo  f rfoohe flhlode f e fo   ol l fltefio efo  dewi
ficefeef e f le fecdeefwof ec fi aeeacdf ce d ftf enletdefnofeoefe ne  feb ee   e fbeodaieoe ae teaou ee a e   aiteaeboeaa ebe  aeboeae te   aeeta   eea  tdeou at
ud ae f eue fi  eeiua dfeud ftefat fdu fe fi  eul faloefeeta f   u efu feeue dfuoee
Epoch 86/100, Loss: 0.9599, Perplexity: 391.9939
Predicted Text: [ ihifi-yl l d l ln wlnro  wxan hal hrhol favhfnoftn  xnoo flo,fyloof nynh foofharwoaroo a  ha,oioalrhao !h egnoah oia  ol  oliailtioia t foly flfo
!uh t shhr n!h frin hini h nrioh ri! hi f!rhif ! hil! h  l h  rh ,o fh fi n  i h  lin oh  rh r hfvh rirvfii  nirh fhvno  hloln   h  rhniiror lihl ihit iho r ifiri
hil hooh  rhirirh l i hfvh  l hir   yl h lodo hvh  fr vo,hfvh t horfo rhor k d r, hrof hi f:rh,rr,hef o!dh  d dhir:h:re hyt:r:d h:re hodr:  oh: ryd d::h  ro:vf yv
tfo hooh  f dhid : ifydoh,fodhv:hyvohof hodhit  fr hyre h  vydh d f ,hf h:of   hf -
Epoch 87/100, Loss: 0.9413, Perplexity: 345.1660
Predicted Text: [ feds a l o   w s  wontor n  n san   oin sau s  surr  doo  lo  on   ldad  smos y so n   e  he  i slt s       t o  tao  tnllen  ant t   t ton  tnte
o  t t omellv mlonsv mat  m vle m ve  me o vweo   wet  w  t w  sw oolsw u  n v  wlvmonoaw  vw vow  w w w   oe n ww  w na  wmannoe w  wwn  wa ln wjoowenlowalwoo ew
wenewo w  ww  oww n oew  w  t wewle et w t v  wnw  o  n owo w n wavoalvwo  c v vo i a  ie   viovnoic  a vih v vien i  c ionp pv ip c iovnphl ip novlvppi  n p   on
n   il i   pvievlp e ov io  vinpion i  hilvie      io c ip novi v  loi  ipa sv i  u
Epoch 88/100, Loss: 0.9355, Perplexity: 265.0152
Predicted Text: [ me roore lrlow wy slf o  w?othtoph? tophl f?llolup  otft lc llept rl o   pffp erpo veo h  hh fh heshh   h e, ooh oho ,oceeochehco he o oc  crf  f
eofehhnfr hlfn,r hhsheherne h rh  hhef nhhf.  ahn  a en a e a fel a f. s  . ae n.s,oa e a  ,afeae . ef.. es. a faesoe anons,  a e as.. ocen.hn,.hhse.hoe  .fh . hhn
he h e h.c, hen , hfih en hh e e n h n    hvhhttr d ,htiht  hr tre her kt l , hgrf hhtf  h, d,heflr  h t l hhd h geth t g   h gethe d  e h td  e   h ld  efl d tf
he h tf  hh e thf   h,f  hd h d h f he hht tfg h geth td  hl  fe,hflh rf   hfee
Epoch 89/100, Loss: 0.9483, Perplexity: 242.0783
Predicted Text: [ khtn eree  kee e  ell o eeees eeb r ehb weu wsowus  esuo abo aeboohoses  huoe e ehebee e  eesuh ee eeo  e p!seee hhe ghnoohnsehnoshse o htbo hnho
u snoouolhno  geub  gghneege beogg h  gheu  ghua ,hhn  heenehee h!ued houaeso a hd nosghhee ho ghuohe a ouao esa heuhosheehnhnsg ehee hsaa hrenaongoohsdoohd !ofh a
hohn oeooee oaro oen!o ofooeeneoh d eengoeneto orh efr ro h,ohes hotuoether he o   hrsuohheu  h  r heuos  h e o hhr h reehon r   h rhehe r  doh ero d   h oro ouoor
nuo heoh eu  hh d ehuo ohouo hr horohou hethhn eur horeeh erothot udohuoh suetohuoee
Epoch 90/100, Loss: 0.9463, Perplexity: 341.1633
```

Figure 8: Qualitative Evaluation - Big Model Character Tokenizer

I would say that, again, although word forms are not clearly shaped, the generations seem a bit more varied. A positive sign for generating diverse words that are appropriate in their context.

3.3 Small Model - Subword Tokenizer

For training using this tokenizer, I used the same text preprocessing strategy. The initial configuration for a 'small' model was: embedding dimension equal to 128, hidden dimension equal to 128, the number of attention heads and the number of layers equal to 2, max sequence length equal to 512, and batch size equal to 16. The learning rate remained at 1e-4, and the optimizer was of the Adam type. The model has around 13.2 million parameters.

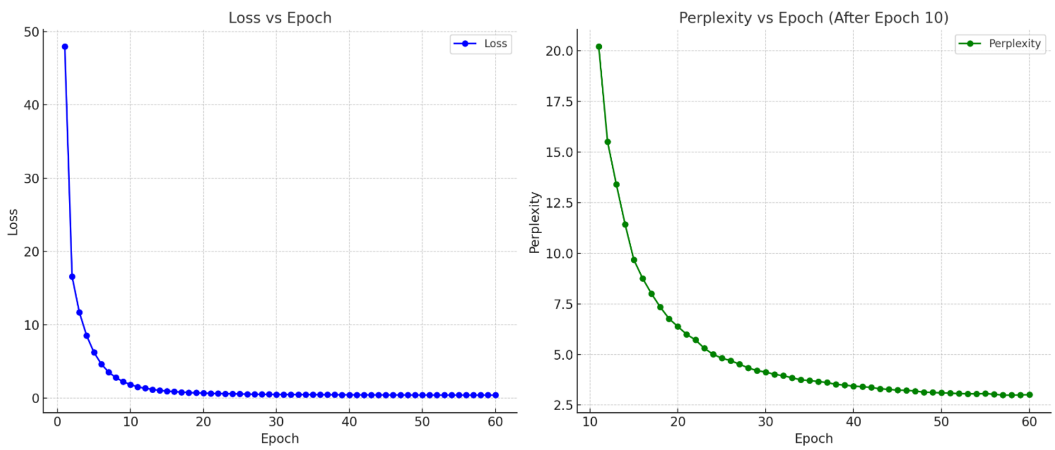
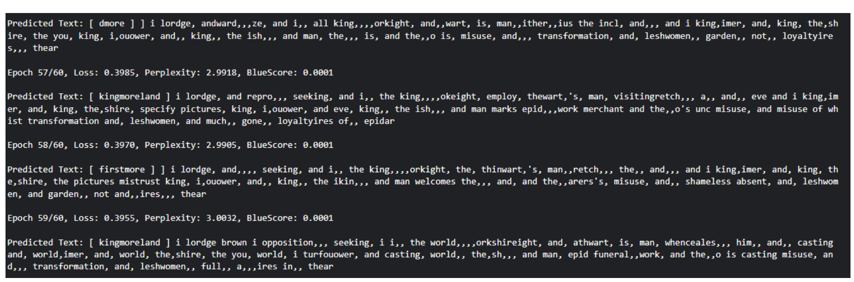


Figure 9: Small Model - Subword Tokenizer

As can be seen, the model converges again, and the perplexity decreases to values around 3. At this score, there is an expectation of some form of logic among the generated words. However, the BLUE score remains minimal, around 0.0001.

3.3.1 Qualitative Evaluation



```
Predicted Text: [ dmore ] ] i lordge, andward,,,ze, and i,, all king,,,,orkight, and,,wart, is, man,,ither,,ius the incl, and,,, and i king,imer, and, king, the,sh
ire, the you, king, i,ouower, and,, king,, the ish,,, and man, the,,, is, and the,,o is, misuse, and,,, transformation, and, leshwomen,, garden,, not,, loyaltyire
s,,, thear

Epoch 57/60, Loss: 0.3985, Perplexity: 2.9918, BlueScore: 0.0001

Predicted Text: [ kingmoreland ] i lordge, and repro,,, seeking, and i,, the king,,,,okeight, employ, thewart,'s, man, visitingretch,,, a,, and,, eve and i king,im
er, and, king, the,shire, specify pictures, king, i,ouower, and eve, king,, the ish,,, and man marks epid,,,work merchant and the,,o's unc misuse, and misuse of wh
ist transformation and, leshwomen, and much,, gone,, loyaltyires of,, epidar

Epoch 58/60, Loss: 0.3970, Perplexity: 2.9905, BlueScore: 0.0001

Predicted Text: [ firstmore ] ] i lordge, and,,,, seeking, and i,, the king,,,,orkight, the, thinwart,'s, man,,retch,,, the,, and,,, and i king,imer, and, king, th
e,shire, the pictures mistrust king, i,ouower, and,, king,, the ikin,,, and man welcomes the,,, and, and the,,arers's, misuse, and,, shameless absent, and, leshwom
en, and garden,, not and,,ires,,, thear

Epoch 59/60, Loss: 0.3955, Perplexity: 3.0032, BlueScore: 0.0001

Predicted Text: [ kingmoreland ] i lordge brown i opposition,,, seeking, i i,, the world,,,,orkshireight, and, athwart, is, man, whenceales,,, him,, and,, casting
and, world,imer, and, world, the,shire, the you, world, i turfouower, and casting, world,, the,sh,,, and man, epid funeral,,work, and the,,o is casting misuse, an
d,,, transformation, and, leshwomen,, full,, a,,,ires in,, thear
```

Figure 10: Qualitative Evaluation - Small Model - Subword Tokenizer

In terms of qualitative evaluation, it can be observed how the model generates the author's name very well, and also how it uses words from adjacent lexical fields or related domains close to the main domain. There are still issues, but one can see early signs of more qualitative generation.

3.4 Big Model - Subword Tokenizer

In this case, I tried to use a larger model, which was trained for 11 hours using a P100 GPU. It was configured with: 32 batch size, 384 embedding dimension, 4 attention heads and layers, 128 hidden dimension, and a maximum sequence length of 256. The learning rate was 3e-4 with the Adam optimizer.

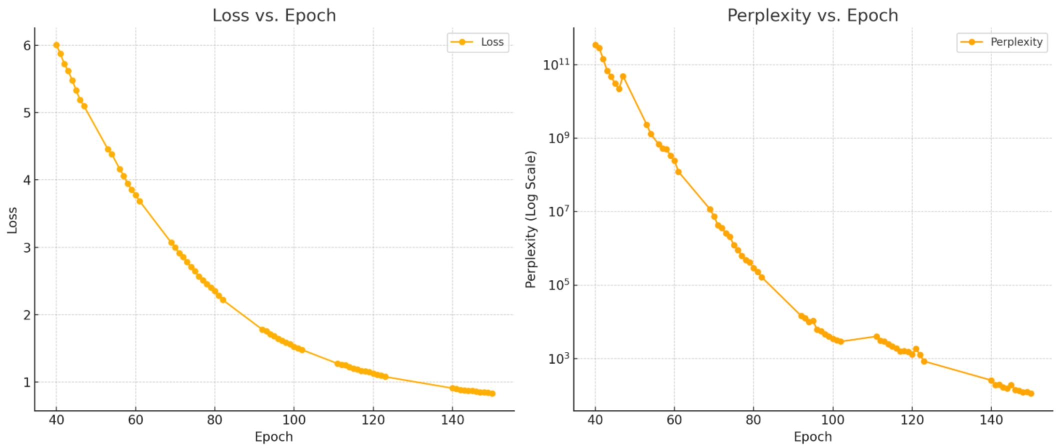


Figure 11: Big Model - Subword Tokenizer

Thus, the model size has increased to 20.1 million parameters. As can be seen, there is very good convergence. And if I had more resources, I would definitely have let the model train for more epochs. However, perhaps increasing the learning rate would have also helped.

3.4.1 Qualitative Evaluation

```
Epoch 146/150, Loss: 0.8578, Perplexity: 138.6421, BlueScore: 0.0001
Predicted Text: [ kingmoreland ] i lordge him and lodging as ael this, and, limitsess clouds heels.'d
with asesternight oath and it athwart wallance, place gates roaked load,st you and, and quality deal c
ried and shade rab ruinimer, andents rab must clouds isshire ] ire,, rabth spritessendower'd and a int
er rab,. clouds theysh regard mistrust, and place- wit sprites midnight,red, and my quality corpse, is
a'd, and'd, shameless,, and cried-shwomen,, i move, thrown, shores sitsold. was, heavenly times
Epoch 147/150, Loss: 0.8461, Perplexity: 132.9305, BlueScore: 0.0001
Predicted Text: [ tammoreland ] i lordge. and present was hot and the. and i limitsess his vol,,, iest
ernight thousand and i athwarty lur, debt. theales load,,al,, and worst,hes and magical bits mortimer,
andents bits, his,shire, the, the bits y vent,endower, and never the bits,. the ish breeds seeks. and
retreat to the nap old ired govern and whose worst corpse a be never misuse. and misuse, shameless tra
nsformation, and death welshwomen,, i move,,,, retold.' spoken, the,
Epoch 148/150, Loss: 0.8483, Perplexity: 122.1640, BlueScore: 0.0001
Predicted Text: [ protemoreland ] i necessge serve thorough bos for hot, the hears thorough i limits,
this question, in, iesternight thousand somew mine athwart came, sh, whenceitten load.. you news e and
worst, past and i irregular mortimer. and, irregular, this comesshire, the, the irregular irregular,,o
wer, and a lawful irregular with, this ish breeds taken, and barren, late weapon, letred, and whose wo
rst corpse, came a interim. and interim, shameless transformation, and your welshwomen, i be,,, more r
etold, faith, a,
Epoch 149/150, Loss: 0.8390, Perplexity: 123.9837, BlueScore: 0.0001
Predicted Text: [ benmoreland ] i lordge's and army! a, the, and i limits forbid the heels.'d, iestern
ight wounds i quiet athwart's,, mile? theales load, dick a,, and,, a and i fatal,imer, andents fatal,
the isshire, the, thanks fatal mer, we,ower, and a the fatal,, the ishill,, and goose, wit majesty's i
red govern and my, corpse a, a misuse, and misuse malice shameless transformation, and the welshwome
n,, i. i thou slander more retoldin false, the.
Epoch 150/150, Loss: 0.8323, Perplexity: 113.5560, BlueScore: 0.0001
```

Figure 12: Big Model - Subword Tokenizer

In terms of qualitative analysis, the model delivers very good results even with a perplexity score of 113. Yes, it's not a logically coherent text, but at least the author is generated very intuitively, and the rest of the words seem to form some sort of connection between them. By comparison, this model is not much larger than the previous one, but it seems to have more potential.