Learning what to learn: Generating language lessons with BERT

João Pedro Olinto Dossena - r0874700

Supervisor: Prof. dr. Bettina Berendt
Assistant-Supervisor: Ing. Pieter Delobelle
Assessors: Ing. Pieter Delobelle
Prof. dr. G. Marra

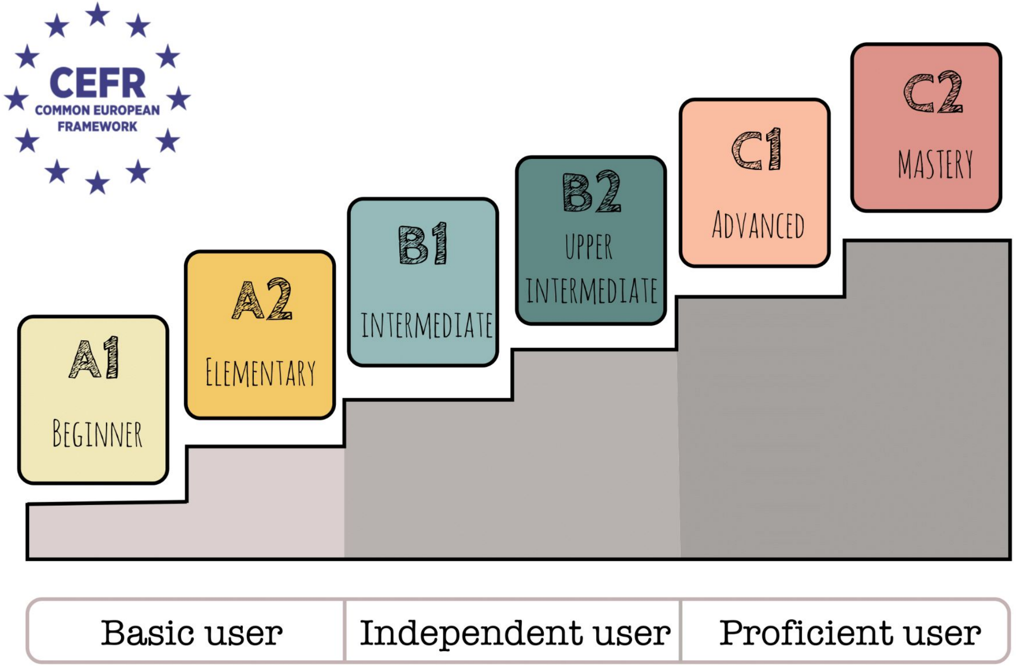
Introduction

- **Context**:
  Duolingo has manually designed skill trees organized by topic (e.g. food, animals), which demands a lot of work by humans

  Transformer language models have shown to be effective in a wide range of NLP tasks

- **Aim**:
  Develop transformer-based method to create skill trees (partially) automatically from existing sentences

Introduction



Introduction

- **Aim**:
  Develop transformer-based method to create skill trees (partially) automatically from existing sentences:

  (1) Develop method to predict text difficulty (in CEFR scale)
  +
  (2) Develop algorithm to construct language skill tree (from difficulty and from topic)

  Case study in Portuguese

Research Questions

The research questions that I aim to answer are:

- **RQ1**: *What impact does a monolingual language model have on classifying the CEFR level of phrases in Portuguese?*

- **RQ2**: *How can a language skill tree be automatically built from CEFR-labeled sentences?*

RQ1: CEFR Prediction

- Work has been done for predicting CEFR level of text in different languages, such as Swedish, Estonian and Portuguese

- Most of these use extracted (syntactical, lexical, morphological, etc) features to predict text difficulty

- A paper introduces the use of Transformer language models for such a task in Portuguese

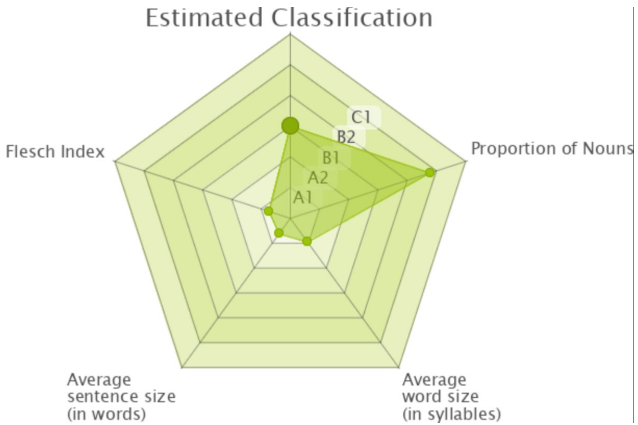
RQ1: CEFR Prediction

Santos et al.:

- Compares transformers to feature-based methods for CEFR level prediction
- Uses 2 transformer models: GPT-2 and RoBERTa
  - GPT-2 was initialized with a model fine-tuned from English to Portuguese
  - RoBERTa was trained with 10M sentences in English and 10M in Portuguese
- The corpus used for training is private
- The Transformer-based models had better metrics than the feature-based models

RQ1: CEFR Prediction

- When testing their best model (GPT-2) through an API, I noticed it didn't seem very accurate, which prompted me to investigate. More on metrics later.
- For example, the simple phrase "*Olá, meu nome é João, sou estudante e moro na Bélgica.*" should be classified as A1



RQ1: CEFR Prediction

- I then fine-tuned a monolingual model (BERTimbau) on a public corpus (COPLE2) for the task of CEFR level prediction for Portuguese

- BERTimbau is a BERT model trained on 17 GB of data in Portuguese

- COPLE2 is a corpus made of CEFR-annotated essays in Portuguese written by second language learners of Portuguese

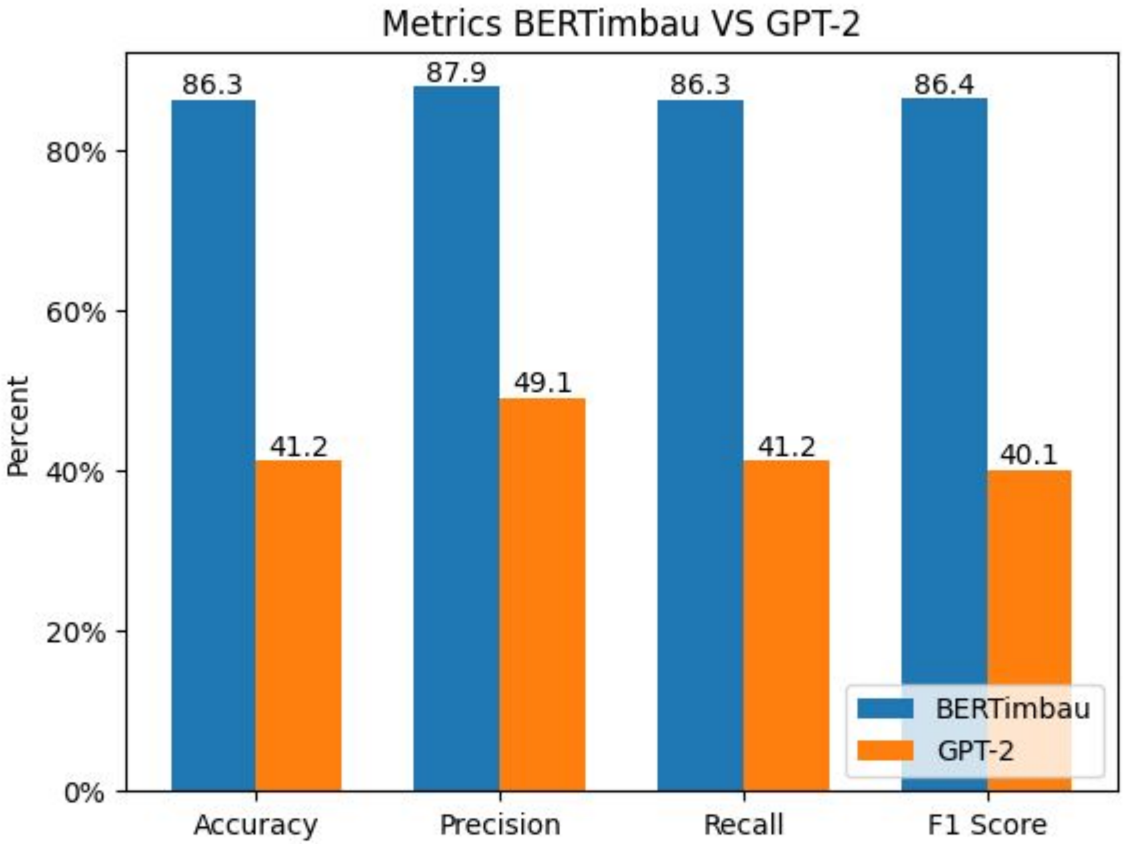
RQ1: CEFR Prediction

- Trained BERTimbau on a training-validation-testing set split of 816-102-102 essays of similar class distribution

- Compared the BERTimbau model to the GPT-2 model of Santos et al., on an unseen subset of the COPLE2 corpus

- Both models made CEFR label predictions to essays on the COPLE2 corpus, and accuracy, precision, recall and F1 score were measured

RQ1


Metrics BERTimbau VS GPT-2
Percent
0%
20%
40%
60%
80%
86.3
41.2
87.9
49.1
86.3
41.2
86.4
40.1
Accuracy
Precision
Recall
F1 Score
BERTimbau
GPT-2

RQ1: CEFR Prediction

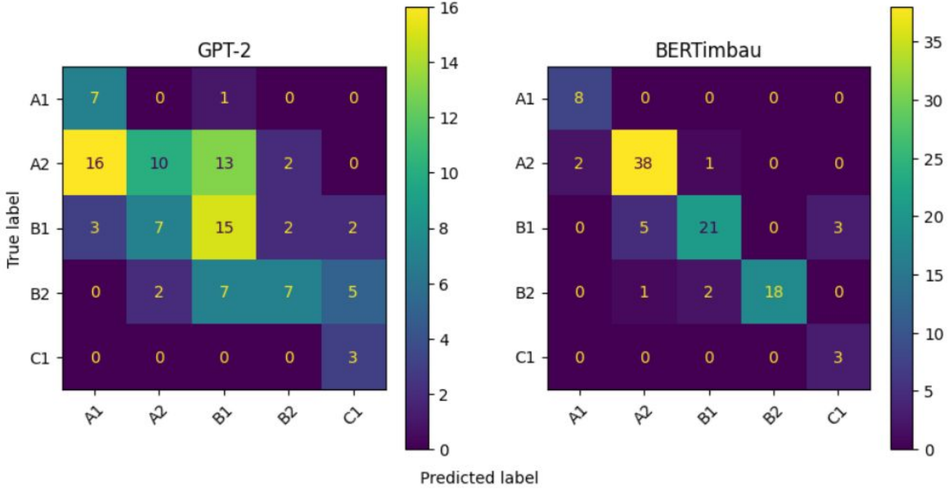


FIGURE 4.1: Confusion matrices for GPT-2's and BERTimbau's predictions on the test set.

RQ1: CEFR Prediction

- The results from GPT-2 are below expected probably because the model has been converted from English to Portuguese on ~1GB of data
- BERTimbau, on the other hand, was trained on 17GB of data in Portuguese

- Another possible aspect that contributes to the performance difference is that BERTimbau was trained and tested on COPLE2; while GPT-2 was trained on another corpus and tested on COPLE2
- So there might be an expected difference if the corpora are vastly different

CEFR Prediction of Tatoeba Sentences

- “**Tatoeba is a collection of sentences and translations.** It's collaborative, open, free and even addictive”

- One can add new sentences in a certain language, and one can translate already existing sentences into another language

CEFR Prediction of Tatoeba Sentences

I then used BERTimbau to classify the CEFR level of over 270k phrases in Portuguese from Tatoeba

This was the class distribution:

A1	45.071	16.5%
A2	205.286	75%
B1	12.534	4.6%
B2	8.853	3.2%
C1	1.970	0.7%
Total	273.714	100%

CEFR Prediction of Tatoeba Sentences

PT	EN	CEFR
De vez em quando vamos ao cinema juntos.	We go to the movies together once in a while.	A1
Onde você encontrou o gato deles?	Where did you find their cat?	A2
Tom e Mary dizem que não acham que John precise fazer isso.	Tom and Mary say they don't think John has to do that.	B1

CEFR Prediction of Tatoeba Sentences

PT	EN	CEFR
Mesmo se eu fosse um anão, seria de qualquer forma um gigante.	Even if I were a dwarf, I would still be a giant.	B2
À medida que os estilistas negros continuam a ganhar reconhecimento na indústria global da moda, uma nova onda de cultura africana está a emergir nas ruas de Brooklyn, em Nova Iorque.	As black designers continue to gain recognition in the global fashion industry, a new wave of African culture is surfacing on the streets of Brooklyn, New York.	C1

RQ2: Topic Modeling

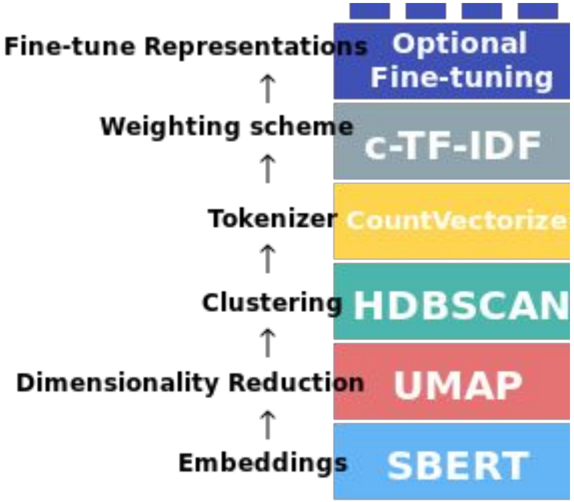
- For this research question, topic modeling was performed with BERTopic
- BERTopic clusters a collection of sentences into semantically similar topics
- Ideally there is a way to gauge the quality of topics extracted by BERTopic

RQ2: Topic Modeling

- How to evaluate **RQ2**?
    - User study? Hard to quantify and tricky to get enough users
    - **Some sort of metric**? Hard to find one
    - **Comparing to existing skill trees**? Different ≠ Better/Worse

- Solution: similarity metric compared to existing skill trees

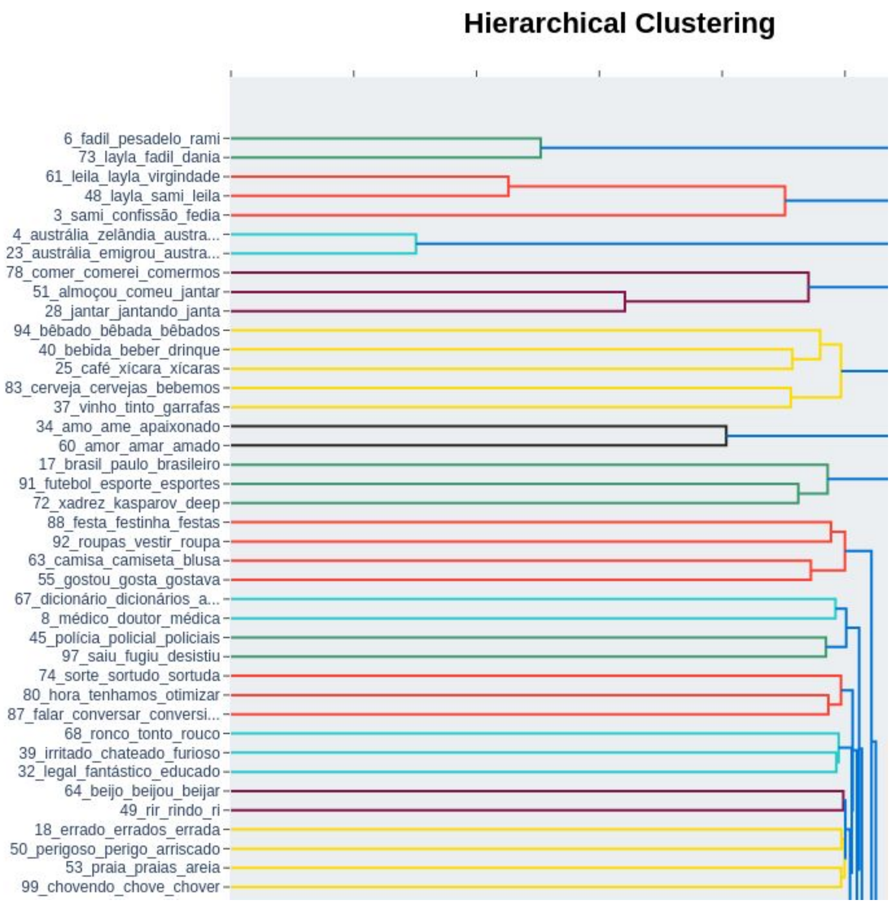
RQ2: Topic Modeling



- Using the default parameters, BERTopic clustered 273k sentences into over 5k topics

RQ2: Topic Modeling

- (Part of) hierarchical clustering of the top 100 topics out of 5137
- Many topics with the same color are semantically similar
- Suggests merging topics might be beneficial

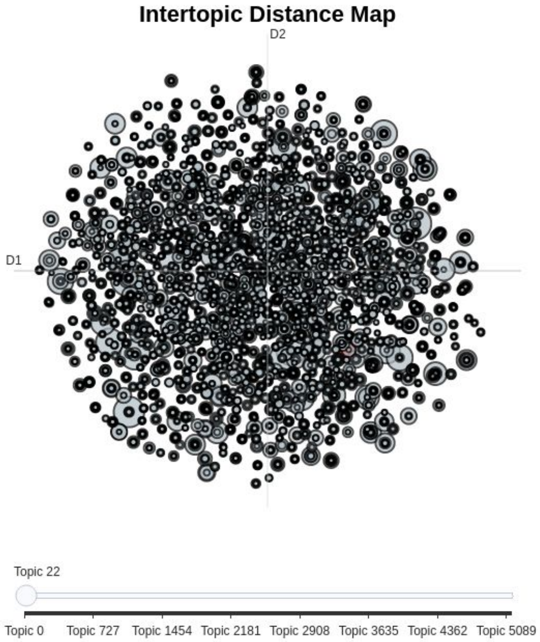


RQ2: Topic Modeling

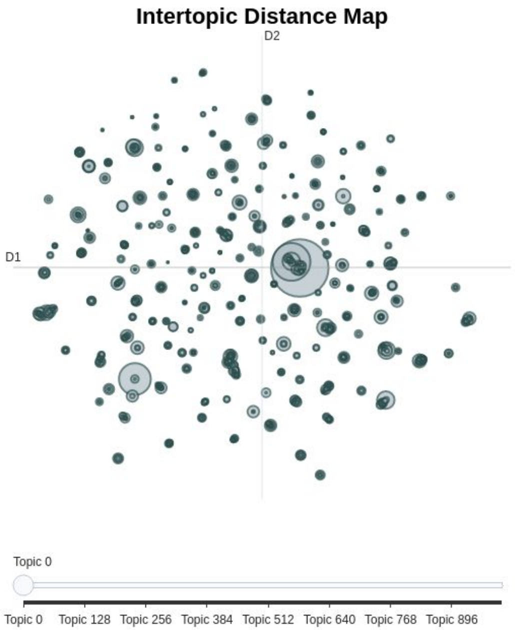
- BERTopic offers the option of merging most similar topics until you get a prespecified N amount

- Reducing to N = 1000 topics (guessing)

RQ2: Topic Modeling

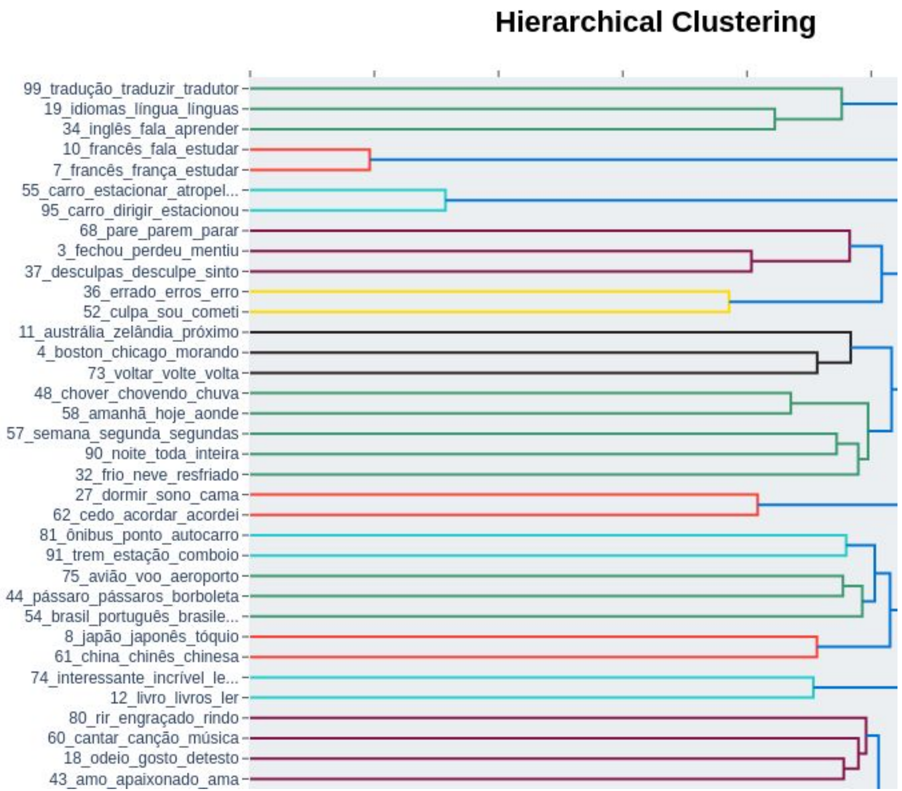


5137 topics
VS
1000 topics



RQ2: Topic Modeling

- (Part of) hierarchical clustering of the top 100 topics out of 1000
- Many topics with the same color are STILL semantically similar
- Suggests merging topics STILL might be beneficial
- Will reduce it to N = 250 topics



RQ2: Topic Modeling

- Duolingo, Babbel and Memrise are 3 of the most popular language apps
- Get topics from the apps, remove grammatical topics
- Compare topics extracted by BERTopic to semantical topics from apps

RQ2: Topic Modeling

- BERTopic’s *find_topics* method that allows one to find the N most similar extracted topics to a certain search term

- The method also returns a semantic similarity percentage

- Pass language learning apps’ semantical topics as search terms into the method

RQ2: Topic Modeling

Memrise semantical topic	Most similar topic	Similarity	Topic frequency
1 - Activities	217_especialidade_depois_livre_passatempo	51.89%	72
2 - Basics	184_formulário_página_site_blog	41.8%	150
3 - Education	24_escola_professor_aula_professora	68.88%	1467
4 - Food	5_jantar_comer_bolo_pão	73.31%	3384
5 - Health	202_saúde_mental_meditando_psicólogo	67.86%	98
6 - Introductions	184_formulário_página_site_blog	46.22%	150
7 - Miscellaneous	184_formulário_página_site_blog	44.34%	150
8 - Opinions	51_futuro_ideia_plano_pensar	56.38%	857
9 - Relationships	26_amor_amo_coração_beijo	47.03%	1395
10 - Shopping	127_supermercado_loja_compras_shopping	74.58%	312
11 - Social Life	135_vida_universo_alma_mundo	49.9%	293
12 - Society	30_nós_temos_juntos_vamos	43.44%	1323
13 - Sports	53_tênis_futebol_jogar_beisebol	68.4%	839
14 - Travel	112_férias_viajar_viagem_país	71.99%	385
15 - Work	52_trabalho_emprego_trabalhar_escritório	67.08%	852

TABLE 4.12: Most similar topics to Memrise's semantical topics

RQ2

Duolingo semantical topic	Most similar topic	Similarity	Topic frequency
1 - Use basic phrases	89_dicionário_esperanto_língua_dicionários	47.39%	493
2 - Describe what's around you	2_quem_você_vocês_pergunta	56.6%	5695
3 - Use polite phrases	17_inglês_língua_idiomas_línguas	45.5%	1960
4 - Greet people	6_feliz_obrigado_felizes_muito	56.65%	3055
5 - Describe your food	5_jantar_comer_bolo_pão	63.44%	3384
6 - Talk about animals	29_cachorro_cavalo_animais_cavalos	76.67%	1332
7 - Use tu	118_funcionar_funciona_como_mostrar	37.69%	348
8 - Use a gente	30_nós_temos_juntos_vamos	61.4%	1323
9 - Describe things	138_significa_palavra_pronuncia_pronunciar	59.58%	282
10 - Express possession	28_dinheiro_rico_caro_pagar	34.74%	1342
11 - Describe clothing	58_camisa_casaco_gravata_roupas	66.25%	783
12 - Order food	5_jantar_comer_bolo_pão	59.15%	3384
13 - Describe colors	110_azul_cor_verde_azuis	63.99%	389
14 - Count up to twenty	201_trinta_dólares_30_quarenta	57.96%	102
15 - Talk about body parts	58_camisa_casaco_gravata_roupas	42.34%	783
16 - Talk about your family	14_família_irmã_pai_pais	76.67%	2210
17 - Describe your home	33_casa_ir_em_assombrada	69.96%	1214
18 - Name common objects	89_dicionário_esperanto_língua_dicionários	34.43%	493
19 - Mention where something is	138_significa_palavra_pronuncia_pronunciar	50.18%	282
20 - Describe people and things	138_significa_palavra_pronuncia_pronunciar	43.72%	282
21 - Ask where people are going	2_quem_você_vocês_pergunta	57.47%	5695
22 - Talk about things around you	217_especialidade_depois_livre_passatempo	63.35%	72
23 - Talk about your job	52_trabalho_emprego_trabalhar_escritório	81.2%	852
24 - Express opinions	51_futuro_ideia_plano_pensar	57.98%	857
25 - Mention dates	107_ano_outono_primavera_mês	49.0%	396
26 - Communicate quantities	32_telefone_carta_celular_senha	37.49%	1226
27 - Describe people	138_significa_palavra_pronuncia_pronunciar	43.82%	282
28 - Talk about abstract things	82_matemática_física_células_química	48.99%	522
29 - Describe where you are	2_quem_você_vocês_pergunta	57.3%	5695
30 - Talk about people	2_quem_você_vocês_pergunta	51.02%	5695
31 - Talk about memories	104_lembro_lembra_memória_esquecer	72.04%	409
32 - Count up to a million	168_pessoas_morreram_milhares_havia	56.24%	188
33 - Describe sizes	125_alto_tamanho_alta_altura	54.33%	320
34 - Say what you need	51_futuro_ideia_plano_pensar	57.71%	857
35 - Express quantity	119_explicar_explicação_exemplo_entendo	37.22%	340

RQ2

Babbel semantical topic	Most similar topic	Similarity	Topic frequency
1 - First Words and Sentences	101_começar_primeira_primeiro_começo	48.33%	422
2 - Food and Drinks	5_jantar_comer_bolo_pão	62.98%	3384
3 - Animals	29_cachorro_cavalo_animais_cavalos	72.97%	1332
4 - Body	128_dançar_exercício_exercícios_dança	41.14%	311
5 - Society	30_nós_temos_juntos_vamos	43.44%	1323
6 - Sports	53_tênis_futebol_jogar_beisebol	68.4%	839
7 - Communication	32_telefone_carta_celular_senha	52.27%	1226
8 - Digital World	139_facebook_internet_google_twitter	53.4%	281
9 - Clothes	58_camisa_casaco_gravata_roupas	75.17%	783
10 - Vacations	112_férias_viajar_viagem_país	72.25%	385
11 - Feelings and Attitudes	26_amor_amo_coração_beijo	45.35%	1395
12 - Relationships	26_amor_amo_coração_beijo	47.03%	1395
13 - Life	135_vida_universo_alma_mundo	67.94%	293
14 - Festivals and Parties	61_festa_natal_aniversário_presente	62.78%	750
15 - Transportation and Travel	43_ônibus_trem_estação_pegar	59.99%	928
16 - Free Time	178_barato_livre_graça_grátis	60.72%	171
17 - Culture	219_recursos_naturais_países_minerais	46.09%	72
18 - Basic Properties	151_pedra_ouro_ferro_anel	31.46%	254
19 - Academic Fields	24_escola_professor_aula_professora	55.69%	1467
20 - Media	93_televisão_rádio_tv_assistir	53.66%	465
21 - Deparments and Services	207_reduzir_preços_despesas_empresa	45.25%	93
22 - Work	52_trabalho_emprego_trabalhar_escritório	67.08%	852
23 - Home	33_casa_ir_em_assombrada	75.52%	1214
24 - Education	24_escola_professor_aula_professora	68.88%	1467
25 - Landscapes	181_mapa_triângulo_ângulos_quadrado	50.27%	159
26 - Plants	120_jardim_batatas_fazenda_quintal	59.46%	339
27 - Environment	45_chuva_chover_guarda_chovendo	38.41%	920
28 - City	163_rua_estrada_atravessar_atravessando	49.17%	209
29 - Rockstars and Fans	225_concerto_show_circo_sucesso	56.41%	53
30 - Wine, Food and Gastronomy	47_café_leite_chá_xícara	47.43%	886
31 - Lifestyle	135_vida_universo_alma_mundo	42.26%	293
32 - Portuguese for Everyday Life	49_brasil_espanhol_português_espanha	69.56%	877
33 - Portuguese for Your Vacation	49_brasil_espanhol_português_espanha	75.15%	877
34 - Portuguese for Carnival	49_brasil_espanhol_português_espanha	57.97%	877
35 - Communication at Work	52_trabalho_emprego_trabalhar_escritório	57.77%	852

RQ2: Topic Modeling

	Avg. Similarity				
Amount of topics	Duolingo	Babbel	Memrise	Total	Avg. Top Topic size
5137	**68.45%**	**65.97%**	**66.04%**	**67.29%**	62.57
1000	63.83%	61.53%	61.76%	62.78%	242.46
250	56.34%	56.62%	58.20%	56.70%	**1079.36**

TABLE 4.10: Average similarity metrics of most similar BERTopic topic to each app topic

Goal Summary

- Since making skill trees is cumbersome, we wanted to streamline this process in two ways

- Classifying text by difficulty

- Classifying text by topic

Conclusion: RQ1

- Classifying text by difficulty can be used for Automatic Essay Scoring for language institutes

- Having too many students makes manually assessing every essay unfeasible

- Good results using monolingual Transformer models extensively trained in the target language

Conclusion: RQ2

- Classifying text by topic helps streamline the process of mining sentences of a particular topic of interest
- Using both topic and difficulty can be useful for course creators and for independent students to find language learning material

Future Work

- Further investigation on disparity between the BERTimbau and GPT-2 models
- Study the difference between models trained on essays versus sentences

- Examine topic modeling with pre-specified topic labels
- User study evaluating language learning progress of students with material curated using this method

Thank you! Questions?