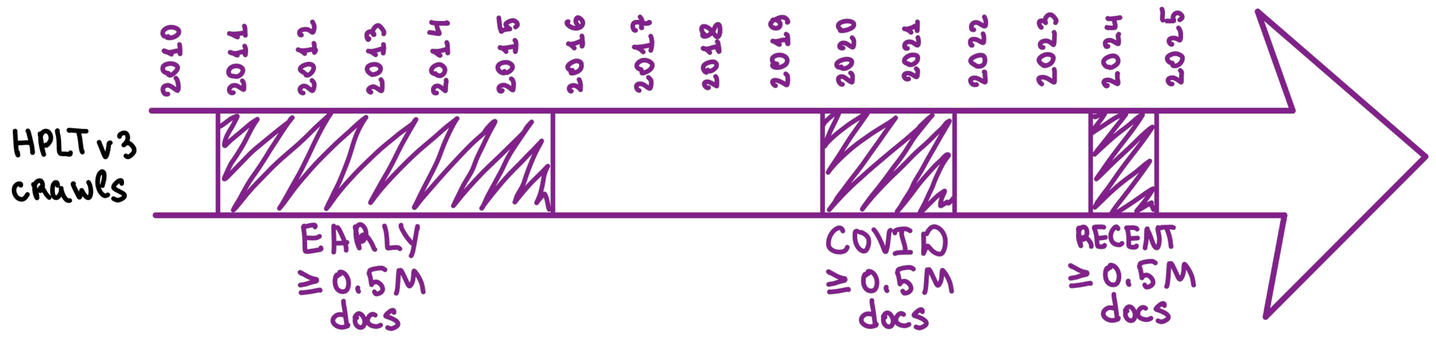
DHPLT: large-scale multilingual diachronic corpora and word representations for semantic change modelling

Mariia Fedorova, Andrey Kutuzov, Khonzoda Umarova

University of Oslo, Cornell University

- **No diachronic corpus for your language?**
- **DHPLT comes to help!**
  - **41 languages**, 12 language families
  - **3 standard time periods**:
    1. 2011-2015 (“Early”)
    2. 2020-2021 (“Covid”)
    3. 2024 (“Recent”)
  - 0.5-1 million documents per time period
  - **~60 billion words** total
  - sourced from web-crawled **HPLTv3** (https://hplt-project.org/datasets/v3.0)



...comes with pre-computed semantic representations for immediate use

DHPLT languages

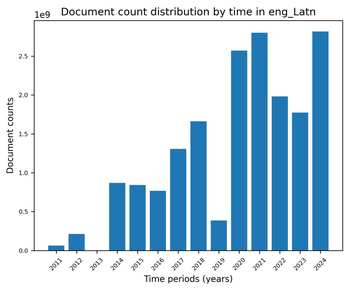
Language selection criteria:

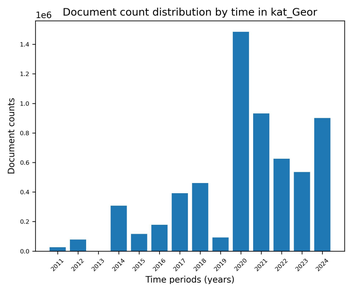
- HPLTv3 must have at least 0.5M documents in each time period
- Language must have a corresponding T5 model pre-trained on HPLTv3 (https://hf.co/collections/HPLT/hplt-30-t5-models)

Language	ISO Code	Family
Albanian	als_Latn	Indo-European
Arabic	arb_Arab	Afro-Asiatic
Bosnian	bos_Latn	Indo-European
Bulgarian	bul_Cyrl	Indo-European
Catalan	cat_Latn	Indo-European
Czech	ces_Latn	Indo-European
Chinese	cmn_Hans	Sino-Tibetan
Danish	dan_Latn	Indo-European
German	deu_Latn	Indo-European
Estonian	ekk_Latn	Uralic
Greek	ell_Grek	Indo-European
English	eng_Latn	Indo-European
Finnish	fin_Latn	Uralic
French	fra_Latn	Indo-European
Hebrew	heb_Hebr	Afro-Asiatic
Croatian	hrv_Latn	Indo-European
Hungarian	hun_Latn	Uralic
Armenian	hye_Armn	Indo-European
Indonesian	ind_Latn	Austronesian
Italian	ita_Latn	Indo-European
Japanese	jpn_Jpan	Japanese
Georgian	kat_Geor	Kartvelian
Korean	kor_Hang	Korean
Lithuanian	lit_Latn	Indo-European
Latvian	lvs_Latn	Indo-European
Macedonian	mkd_Cyrl	Indo-European
Dutch	nld_Latn	Indo-European
Norwegian	nob_Latn	Indo-European
Polish	pol_Latn	Indo-European
Portuguese	por_Latn	Indo-European
Romanian	ron_Latn	Indo-European
Russian	rus_Cyrl	Indo-European
Slovak	slk_Latn	Indo-European
Slovenian	slv_Latn	Indo-European
Spanish	spa_Latn	Indo-European
Swedish	swe_Latn	Indo-European
Tamil	tam_Taml	Dravidian
Thai	tha_Thai	Tai-Kadai
Turkish	tur_Latn	Altaic
Ukrainian	ukr_Cyrl	Indo-European
Vietnamese	vie_Latn	Austro-Asiatic

Time period separation

- We don’t know the **date of creation** for any random web page
- ...but we know its **web crawling timestamp**
  - i.e., when a crawler downloaded this page
- **DHPLT** uses crawling timestamps to to bin documents by time periods
- **Upper boundaries** for creation dates:
  - *If a text was crawled in 2015, it can’t be created later than 2015*
- Different from traditional diachronic corpora!
  - Period 3 can contain documents created in periods 1 and 2 and 3 (or earlier)
  - Period 2 can contain documents created in periods 1 and 2 (or earlier)
  - Period 1 can contain documents created in period 1 (or earlier)
- DHPLT documents preserve precise crawling timestamps:
  - come up with your own time splits, if necessary





“Refined” DHPLT: semantic representations

- About 19,000 potentially interesting target words per language
- For them, semantic representations are pre-computed on DHPLT diachronic corpora

Representation types:

- **Token embeddings** for 1000 random occurrences by:
  - HPLT v3 T5 models
  - XLM-R model
  - HPLT v3 GPT-BERT models
- Top 15 **lexical substitutes** for 100 random occurrences by
  - XLM-R model
  - HPLT v3 GPT-BERT models
- **Static word type embeddings**
  - Word2vec SGNS
- Frequency counts

- Ready to use in your experiments
- Do not spend compute again
- Want different target words?
  - Re-run our code and produce your own representations:

https://github.com/ltgoslo/scdisc_hplt

Is it actually useful for change modeling?

1: 2011-2015	2: 2020-2021	3: 2024-
multiplayer	chatbots	generative
NPCs	IoT	AI’s
RPG	robotics	GenAI
animations	RPA	ChatGPT
FPS	intelligence	LLMs

1. Evolution of the English word “AI”: 5 nearest neighbours according to DHPLT static word embeddings

1: 2011-2015	2: 2020-2021	3: 2024-
BETA	AI	generativa
PS	artificial	artificial
AI	algoritmos	AI
jugabilidad	learning	inteligencia
artificial	inteligencia	ChatGPT

2. Same for Spanish: DHPLT static word embeddings

Period pairs	‘ai’	‘remote’	‘legislative’	‘jurisdiction’
1 to 2	0.5533	0.4586	0.4117	0.4495
1 to 3	**0.5646**	0.4619	0.4141	0.4497
2 to 3	0.48	0.4548	0.4191	0.4351

3. Semantic change discovery for English with DHPLT T5 embeddings: the word “AI” has much higher change score compared to the control words.

Download DHPLT for your language now!

https://data.hplt-project.org/three/diachronic/