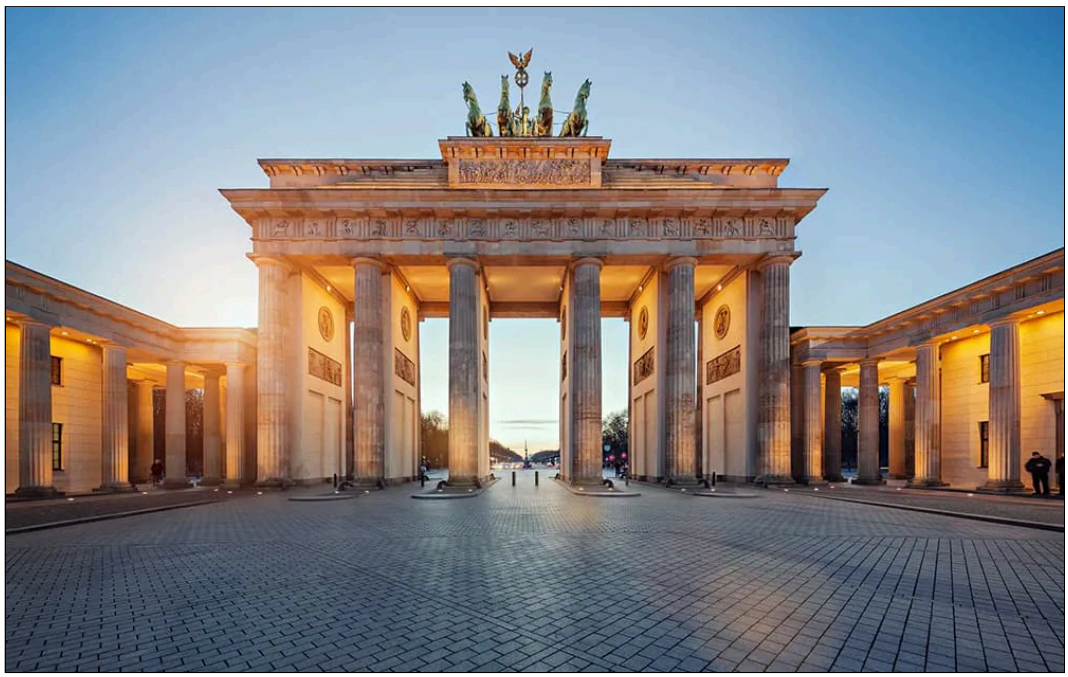
The High Carbon Footprint of German Auto-Translation Models

Originally published at https://www.unite.ai/the-high-carbon-footprint-of-german-auto-translation-models/



New research into the carbon footprint created by machine learning translation models indicates that German may be the most carbon-intensive popular language to train, though it is not entirely clear why. The new report is intended to open up additional avenues of research into more carbon-efficient AI training methods, in the context of growing awareness of the extent to which machine learning systems consume electricity.

The preprint paper is titled *Curb Your Carbon Emissions: Benchmarking Carbon Emissions in Machine Translation*, and comes from researchers at India's Manipal Institute of Technology.

The authors tested training times and calculated carbon emission values for a variety of possible inter-language translation models, and found 'a notable disparity' between time taken to translate the three most carbon-intensive language pairings, and the three most carbon-economical models.

Language Pair	CO2 Emissions (in g)	BLEU Score
En-De	64.33 ± 0.273	29.5 ± 0.61
En-Fr	47.32 ± 6.67	46.2 ± 0.14
De-En	52.65 ± 0.182	33.1 ± 0.51
De-Fr	36.41 ± 1.744	31.6 ± 0.0
Fr-De	42.13 ± 5.283	25.5 ± 0.0
Fr-En	34.69 ± 0.331	43.3 ± 0.61

Language Pair	CO2 Emissions (in g)	BLEU Score
En-De	9.3777 ± 0.2260	33.61 ± 0.324
En-Fr	9.7988 ± 0.0587	53.19 ± 0.000
De-En	9.4135 ± 0.1003	34.55 ± 0.324
De-Fr	9.7924 ± 0.0116	35.26 ± 0.021
Fr-De	9.9836 ± 0.1061	30.13 ± 0.354
Fr-En	9.5342 ± 0.1058	46.96 ± 0.000

An average of carbon emissions released over 10 epochs of training. On the left, results using ConvSeq (see below), on the right, Transformers. Source: https://ar

The paper found that the most 'ecological' language pairings to train are English>French, French>English and, paradoxically, German to English, while German features in all the highest-consuming pairs: French>German, English>German and German>French.

Compound Interest

The findings suggest that lexical diversity 'is directly proportional to training time to achieve an adequate level of performance', and note that the German language has the highest lexical diversity score among the three tested languages as estimated by its Type-Token Ratio (TTR) – a measurement of vocabulary size based on text length.

The increased demands of processing German in translation models are not reflected in the source data that was used for the experiment. In fact, the German language tokens generated from the source data has fewer (299445) derived tokens than English (320108), and far fewer than French (335917).

Language	TTR	Vocabulary	# of Tokens
English	0.0306	9802	320108
German	0.0584	17488	299445
French	0.0336	11287	335917

The challenge, from a Natural Language Processing (NLP) standpoint, is to decompose compound German words into constituent words. NLP systems often have to accomplish this for German without any of the pre-'split' surrounding grammar or contextual clues that can be found in languages with lower TTR scores, such as English. The process is called *compound splitting* or *decompounding*.

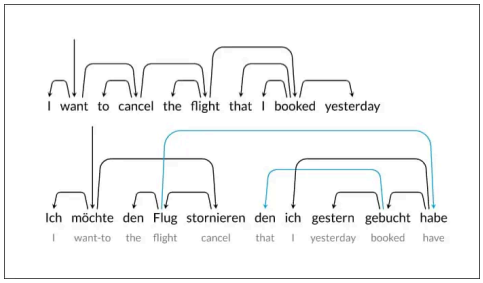
The German language has some of the longest individual words in the world, though in 2013 it lost official recognition of its 65-character former record-breaker, which is long enough to require its own line in this article:

> *Rindfleischetikettierungsueberwachungsaufgabenuebertragungsgesetz*

The word refers to a law delegating beef label monitoring, but fell out of existence due to a change in European regulations that year, conceding the place to other popular stalwarts, such as 'widow of a Danube steamboat company captain' (49 characters):

> *Donaudampfschifffahrtsgesellschaftskapitaenswitwe*

In general, German's syntactic structure requires a departure from the word-order assumptions underpinning NLP practices in many western languages, with the popular (Berlin-based) spaCY NLP framework adopting its own native language in 2016.



Projective mappings in an English and German phrase demonstrate the complex interrelationships between lexical elements in the German language. Source: https://explosion.ai/blog/german-model

Data and Testing

For source data, the researchers used the Multi30k dataset, containing 30,000 samples across the French, German and English languages.

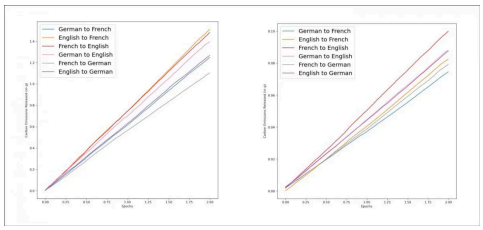
The first of the two models used by the researchers was Facebook AI's 2017 Convolutional Sequence to Sequence (ConvSeq), a neural network that contains convolutional layers but which lacks recurrent units, and instead uses filters to derive features from text. This allows all operations to take place in a computationally efficient parallel manner.

The second approach used Google's influential Transformers architecture, also from 2017. Transformers uses linear layers, attention mechanisms and normalization routines. Admittedly, the original released model has come under criticism for carbon inefficiency, with claims of subsequent improvements contested.

The experiments were carried out on Google Colab, uniformly on a Tesla K80 GPU. The languages were compared using a BLEU (Bilingual Evaluation Understudy) score metric, and the CodeCarbon Machine Learning Emissions Calculator. The data was trained over 10 epochs.

Findings

The researchers found that it was the extended duration of training for German-related language pairs that tipped the balance into higher carbon-consumption. Though some other language pairs, such as English>French and French>English had even higher carbon consumption, they trained more quickly and resolved more easily, with these spurts of consumption characterized by the researchers as 'relatively insignificant' in relation to the consumption by language pairings that include German.



Analysis of the language pairs be encoder/decoder carbon emissions.

The researchers conclude:

> *'Our findings provide clear indication that some language pairs are more carbon intense to train than others, a trend that carries over different architectures as well.'*

They continue:

> *'However, there remain unanswered questions regarding why there are such stark differences in training models for a particular language pair over another, and whether different architectures might be more suited for these carbon-intense language pairs, and why this would be the case if true.'*

The paper emphasizes that the reasons for disparity of carbon consumption across training models are not entirely clear. They anticipate developing this line of study with non Latin-based languages.

1.20pm GMT+2 – Text error corrected.