Deep Learning for Natural Language Processing

Introduction to Module 2

Marco Kuhlmann & Richard Johansson

Module 1: Representation and categorisation

- **Lectures**

  building blocks for representation learning; embeddings

- **Applications**

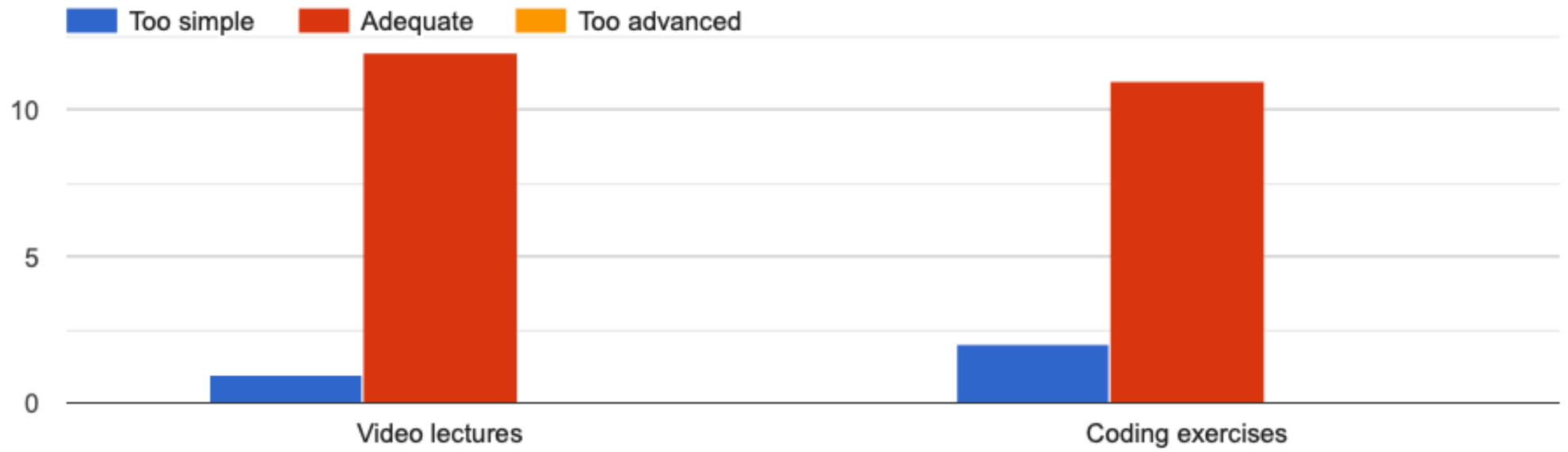
  text categorisation tasks

- **Assignment**

  word sense disambiguation

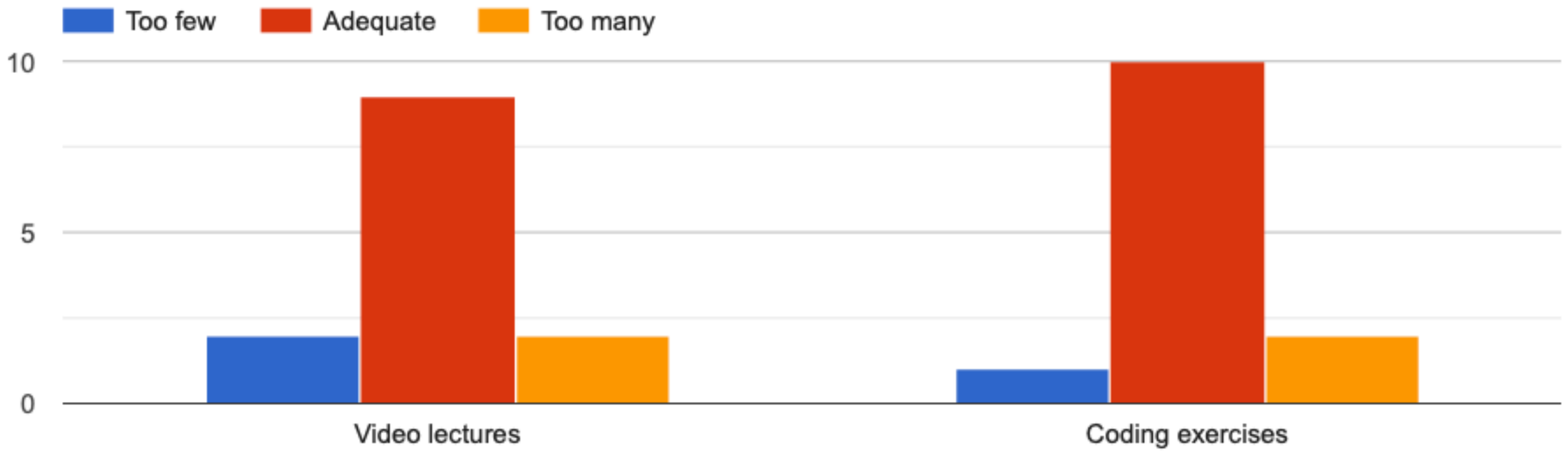
Feedback on Module 1

How challenging are the video lectures and coding exercises?

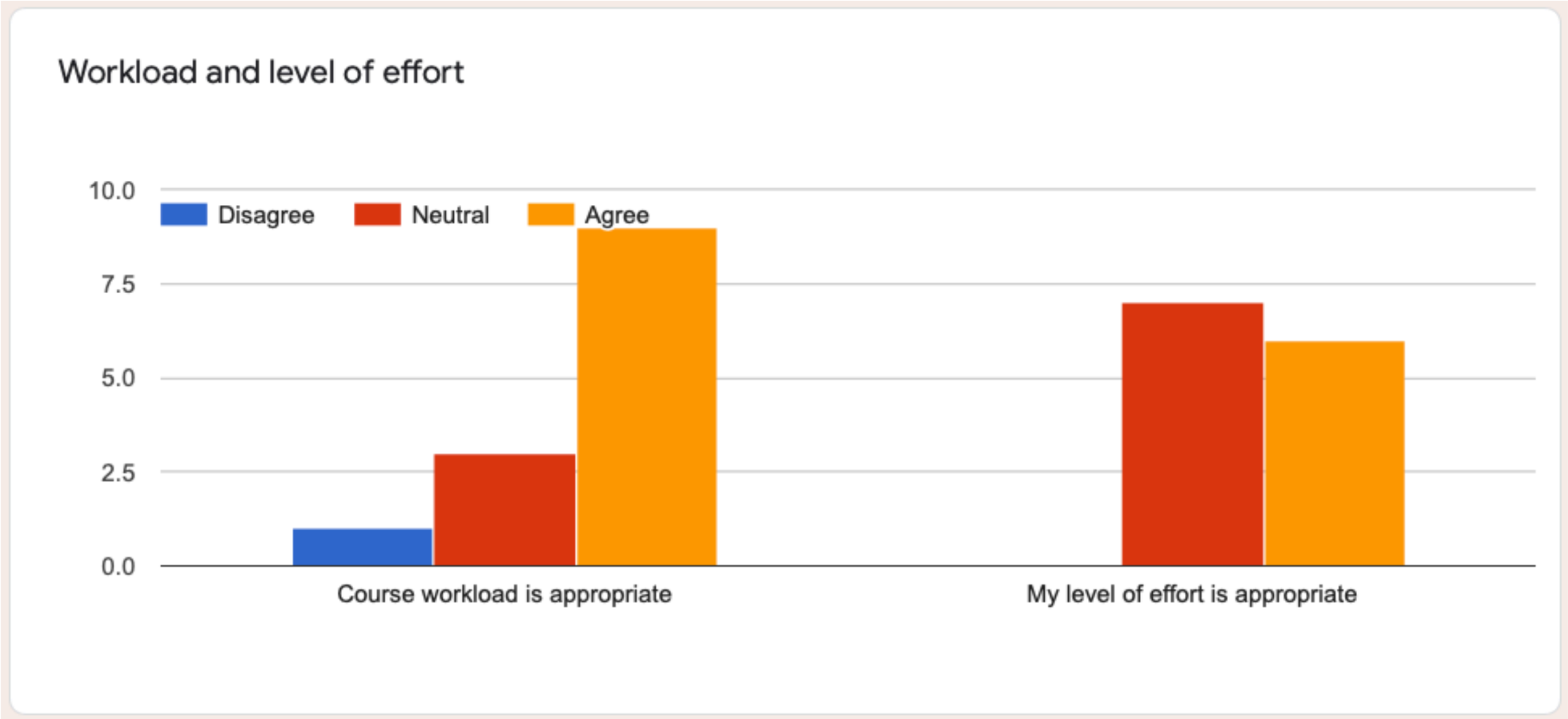


Feedback on Module 1

Are there enough video lectures and coding exercises?



Feedback on Module 1



Module 2: Structured prediction tasks

- **Lectures**

  neural architectures that produce structured outputs

- **Applications**

  named entity recognition, relation extraction

- **Assignment**

  dependency parsing

Search and learning

predicted output

input

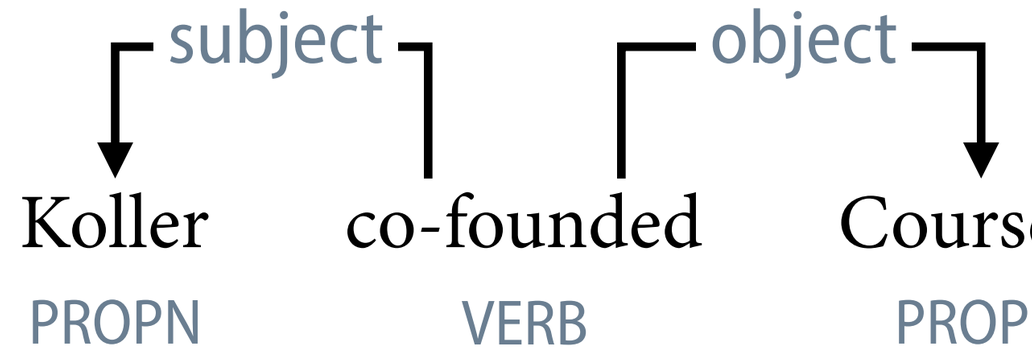
$$\hat{y} = \underset{y}{\operatorname{argmax}}\ \operatorname{score}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

candidate output

model parameters

Eisenstein (2019), § 1.2.2

Linguistic representations

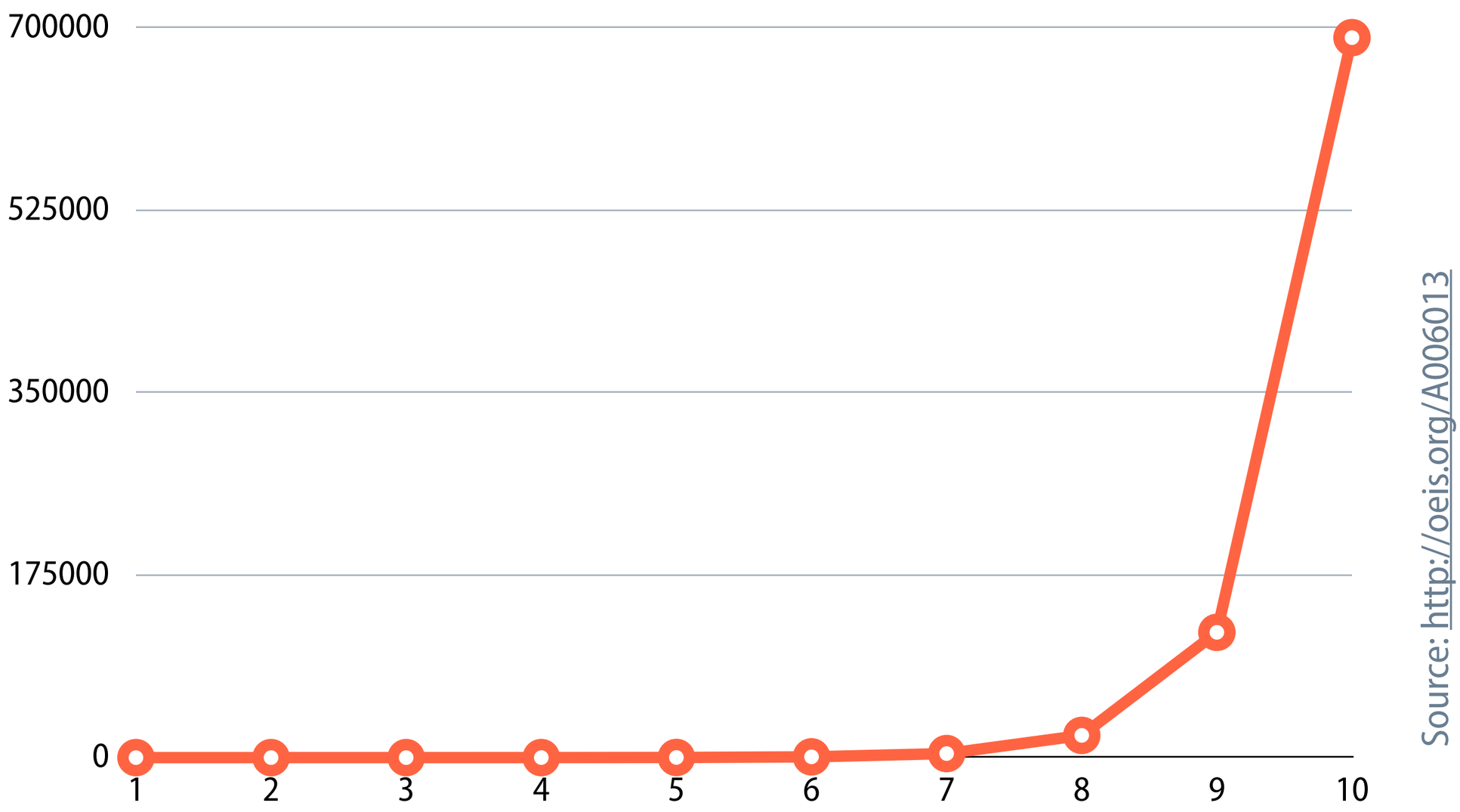


```
dbr:Coursera dbo:foundedBy dbr:Daphne_Koller
```

Ambiguity leads to combinatorial explosion

I	want	to	live	in	peace
PRON	VERB	PART	VERB	ADP	NOUN
NOUN	NOUN	ADP	ADJ	ADV	VERB
		ADV	ADV	ADJ	
				NOUN	

Number of projective dependency trees



Algorithmic approaches

- **Exhaustive search**

  Cast structured prediction as a combinatorial optimisation problem over the set of target representations.

  Viterbi algorithm, Eisner algorithm

- **Greedy search**

  Cast structured prediction as a sequence of classification problems: at each point in time, predict one of several options.

  window-based part-of-speech tagging, arc-standard algorithm

Google Research Blog

The latest news from Research at Google

Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source

Thursday, May 12, 2016

Posted by Slav Petrov, Senior Staff Research Scientist

At Google, we spend a lot of time thinking about how computer systems can read and understand human language in order to process it in intelligent ways. Today, we are excited to share the fruits of our research with the broader community by releasing SyntaxNet, an open-source neural network framework implemented in TensorFlow that provides a foundation for Natural Language Understanding (NLU) systems. Our release includes all the code needed to train new SyntaxNet models on your own data, as well as *Parsey McParseface*, an English parser that we have trained for you and that you can use to analyze English text.

Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing more than 150 treebanks in 90 languages. If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Discussion on UD
  - UD-related events
- Query UD treebanks online:
  - SETS treebank search maintained by the University of Turku
  - PML Tree Query maintained by the Charles University in Prague
  - Kontext maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
  - INESS maintained by the University of Bergen
- Download UD treebanks

If you want to receive news about Universal Dependencies, you can subscribe to the UD mailing list. If you want to discuss individual annotation questions, use the Github issue tracker.

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Afrikaans	1	49K		IE, Germanic
Akkadian	1	1K		Afro-Asiatic, Semitic
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	1	52K		IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque

Module 3: Generation tasks, research outlook

- **Lectures**

  neural architectures for text generation

- **Applications**

  machine translation, summarisation

- **Assignment**

  your own project

Project

- You send us a one-page project sketch until 18 May, 8:00 AM.
- We give feedback on your sketch by 20 May.
- You pitch your project during a Zoom session 25–26 May.
- You send us your project report, deadline 19 June.

  think: short conference paper