DOUBLING THE AMOUNT OF TRAINING DATA: DOES IT HELP?
A New Training Corpus for Slovene and Its Impact on Automatic UD Annotation

Univerza v Ljubljani Fakulteta za računalništvo in informatiko
Institut "Jožef Stefan" Ljubljana, Slovenija
Univerza v Ljubljani Filozofska fakulteta

Luka Terčon **Nikola Ljubešić** **Kaja Dobrovoljc**

Expanding the training corpus for Slovene:

The ssj500k training corpus

Approx. 500,000 tokens

→

The SUK training corpus

Approx. 1,000,000 tokens

Two experiments:

- Doubling the amount of training data
- Adding an inflectional lexicon