

University of Groningen

## The Little Data That Could

Edman, Lukas

DOI:  
[10.33612/diss.907471451](https://doi.org/10.33612/diss.907471451)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2024

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*  
Edman, L. (2024). *The Little Data That Could: Making the Most of Low-Resource Natural Language Processing*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.907471451>

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Propositions accompanying the thesis

# The Little Data That Could

## Making the Most of Low-Resource Natural Language Processing

by Lukas Edman

- Neural networks are excellent for modeling problems, except when there is little data.
- It is best to keep the input short for models learning from scratch on little data.
- Translating without parallel data is possible, at least for languages that aren't that different from each other.
- Character-level models should be the state-of-the-art, if only they weren't so slow.
- Character-level models can be faster, as long as they behave a bit more like subword-level models.
- Curriculum learning doesn't work because language models aren't as stupid as us.
- Working in research is essentially working to make your previous work outdated.
- Making lots of small talk with your supervisors is a great way to distract them from the fact that you got nothing done since the last meeting.
- Every mention of the word "train" in this dissertation is a cheeky reference to the title.
- I will happily graduate having completed a dissertation in linguistics without ever taking a linguistics course.