

A Multidialectal Geotagged Corpus for Limburgish

Andreas Simons

andreas.simons@uliege.be

University of Liège / Maastricht University



In a Nutshell

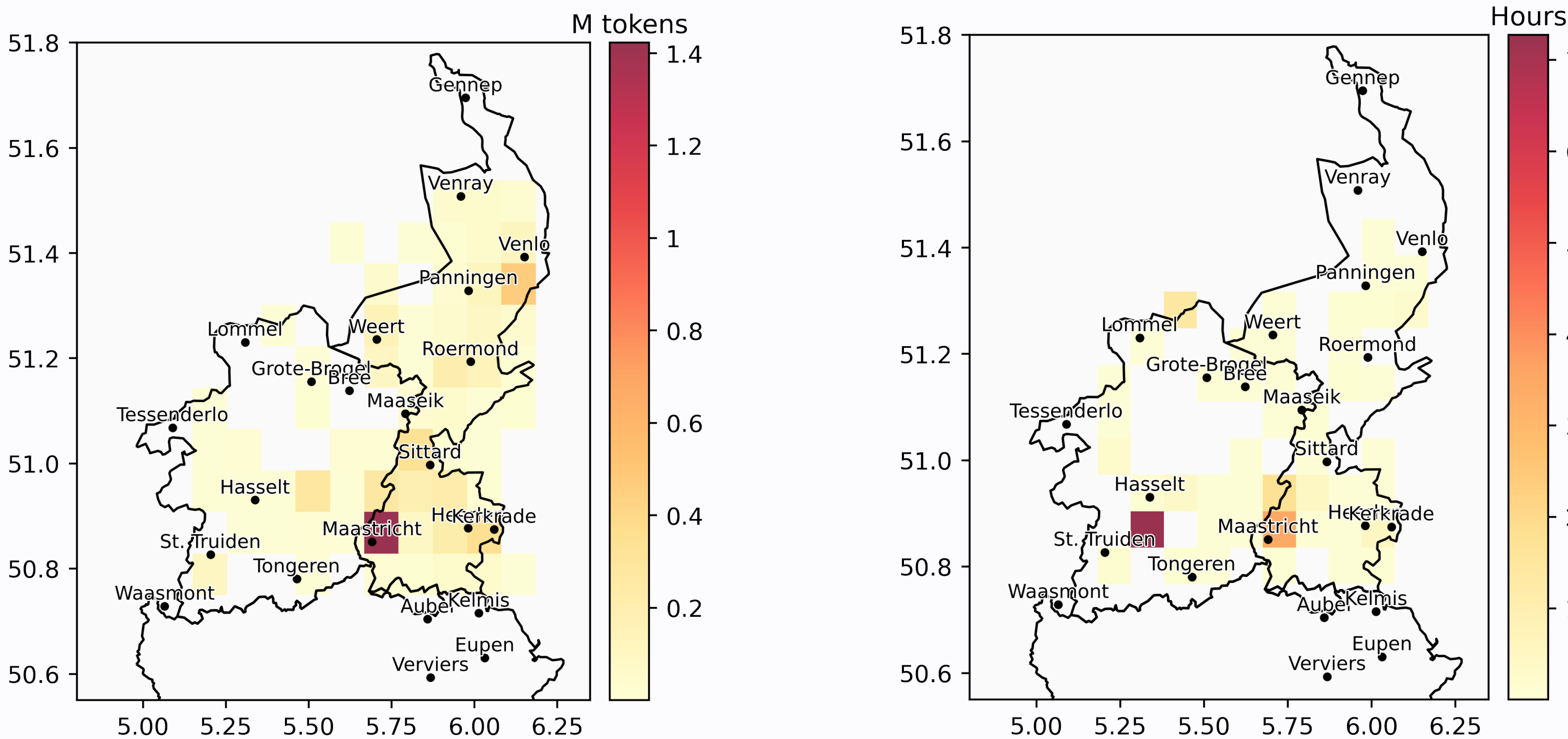
We introduce a very fine-grained, geotagged dialect corpus for written and spoken Limburgish, a non-standardized Germanic language, featuring linguistic variation on all grammatical levels, with extensive intersectional sociolinguistic metadata. This corpus aims to mitigate the lack of available Limburgish corpus material, and contributes to the availability of variational corpora. We also present our challenges and findings in collecting and curating the corpus in an ethical, GDPR-compliant methodology, as well as our efforts in co-creating this corpus with the Limburgish language communities through a crowdsourcing approach.

What is Limburgish?

- Germanic language with very dense vowel system and lexical pitch accent
- (e.g. Maastricht dialect has 24 vowels)
- No spoken or written standard: High degree of orthographic and phonological variation
- Hundreds of dialects considered equal with individual codifications
- Local dictionaries, historical spellings, overarching standards (Veldeke)
- Rapidly growing need for resources and tools
→ Standardization very undesirable!

Example: variation of "school" in 5 dialects with respective codified spelling:
sjaol (Maastricht), **sjeuwel** (Hasselt), **schoul** (St-Truiden), **schoeël** (Venlo), **sjoëël** (Kerkrade)

Geographic distribution of corpus (text & speech)



Overview of the Limburgish Corpus

	Text	Audio
corpus size	6.2M tokens	176 h
files	13200	1656
tokens (h) per file	472 ± 2530 tokens/file	0.11 ± 0.23 h/file
dialects	≈ 340	≈ 80
time period	1850–2025	1990–2025
gender ratio (F-M)	17-83	NA
genre representativeness	carnival lyrics 53%, stories 21%, poetry 12%, articles 4%, historical documents 3%...	music 53%, podcasts 21%, interviews 9%, stories 8%, theater plays 2%...

- Metadata:** author(s), gender, year of birth, year of publication, genre, dialect, specific spelling used, copyright license ...
- Availability:** www.limburgscorpus.com. For non-scientific use momentarily limited due to GDPR and copyright compliance.

Future work

- Elaborate digitization & corpus collection campaign carried out by Maastricht University & Hoes veur 't Limburgs
- Digitization of existing paper archives (Veldeke associations, local history chapters, private collections)
 - Digitization of public broadcaster archives, speech crowdsourcing

Corpus collection

Crowdsourcing

Volunteers can upload any Limburgish texts/audio on our website. With **informed consent**, they can freely decide which metadata to contribute, which copyright license to provide, and how their materials can be used.

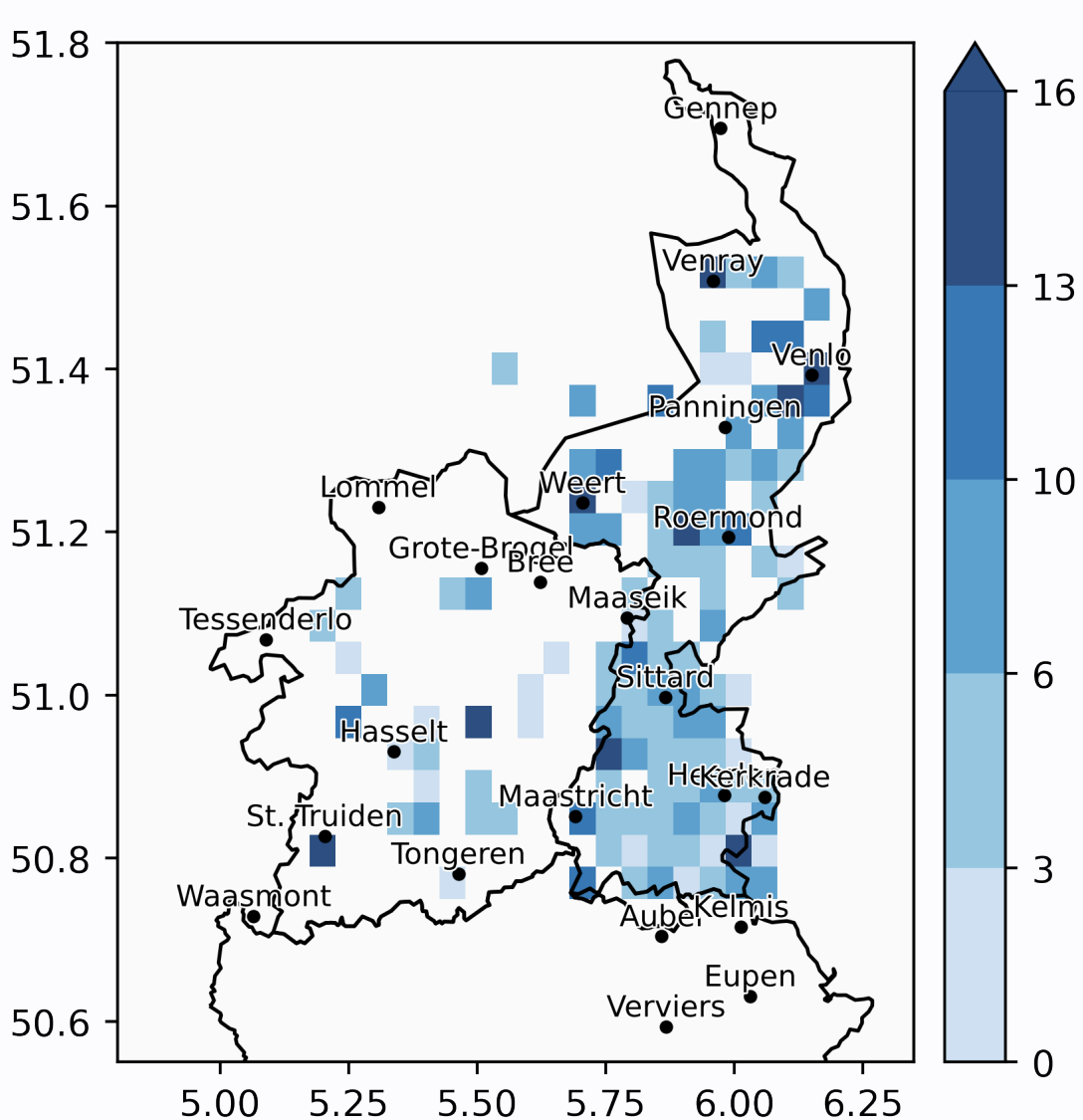
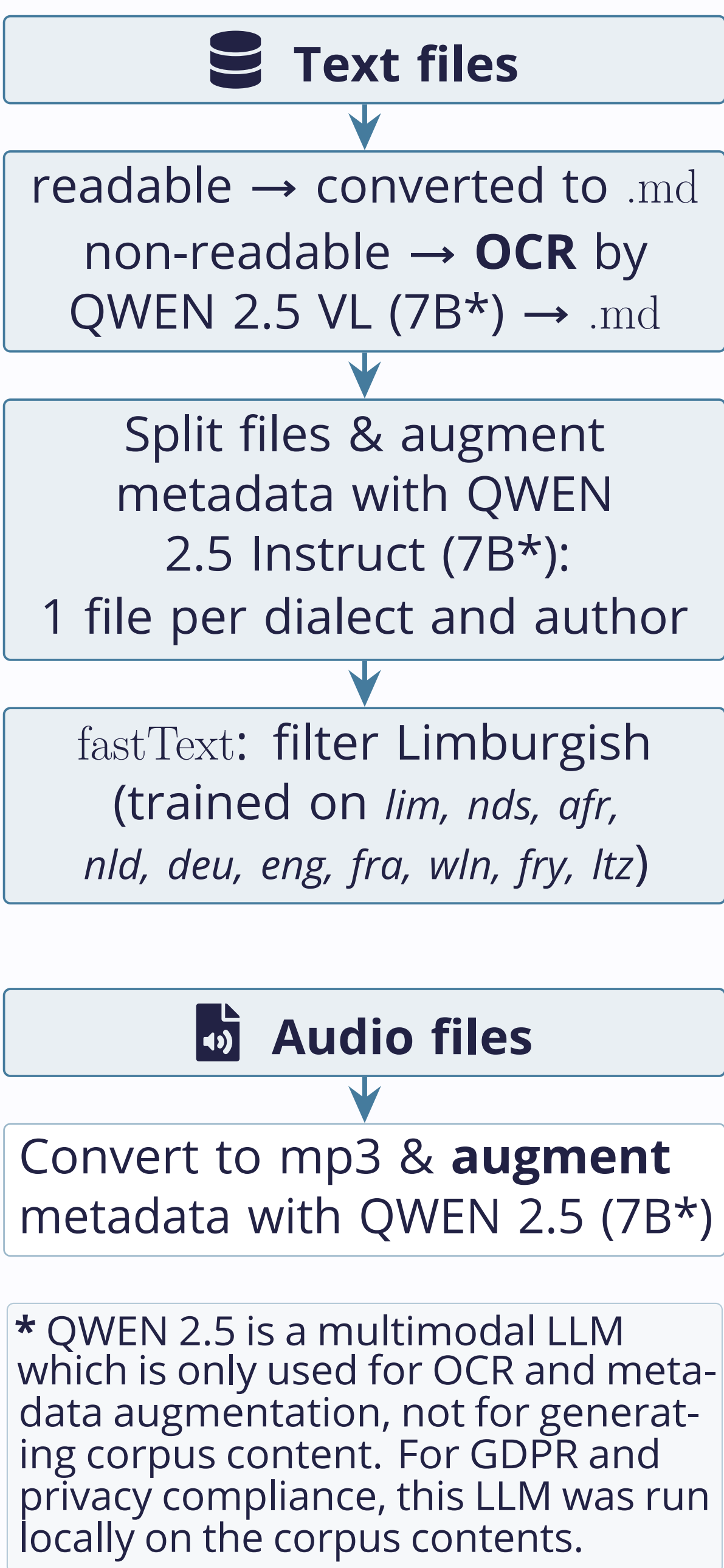
Scraping

Limburgish text and audio are scraped from manually selected websites and YouTube channels.

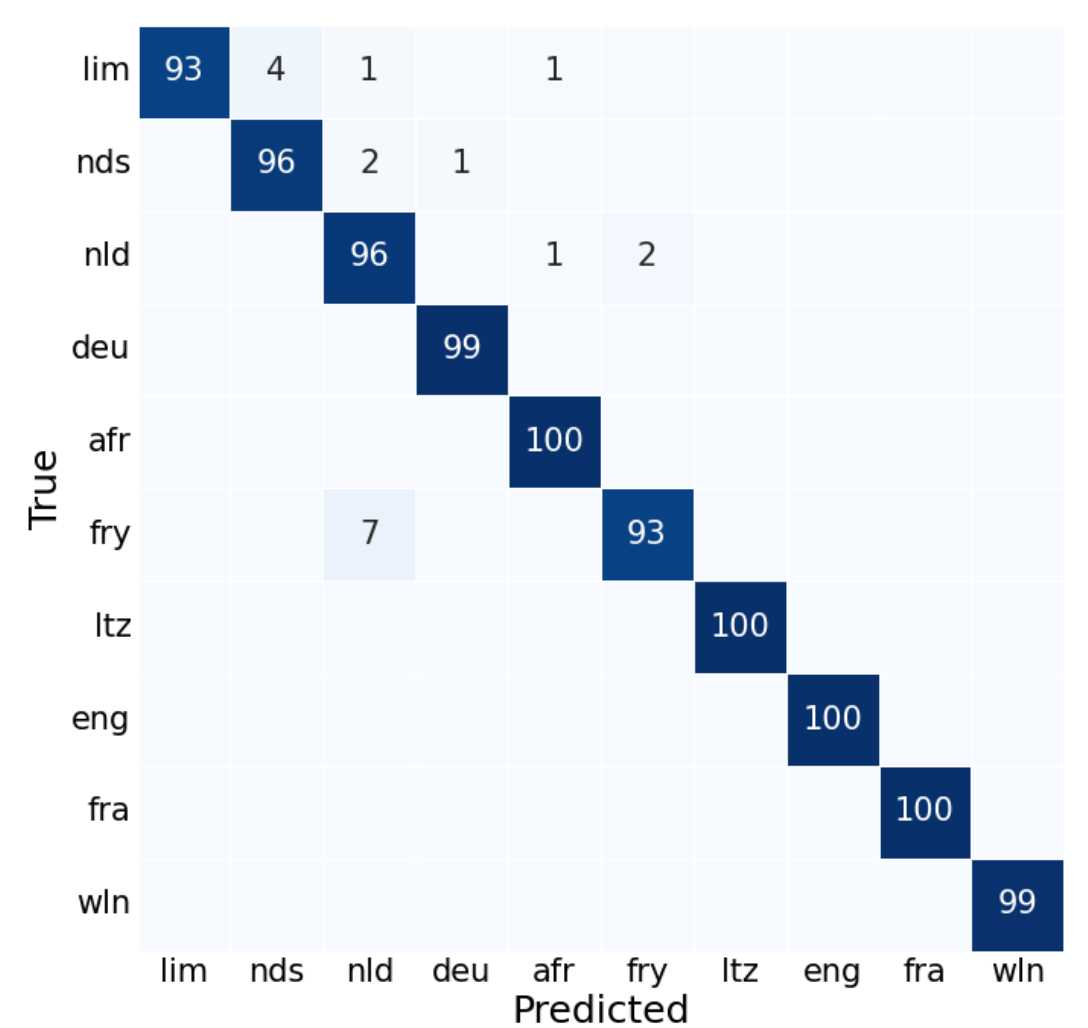
Community Outreach

On-site corpus collection events and calls for contributions through local media, dialect associations, local history chapters, open science events.

Corpus processing



Extra non-Latin graphemes per dialect



Confusion matrix for fastText

Challenges

GDPR/Copyright

- Dialects are intrinsically personal data
- Crowdsourcing/scraping: unsolicited personal data
- Copyright: restricts public access

Metadata

- "Dialect" is ill-defined: give authors maximal flexibility
- Limburgish names vs. official names

Vulnerability

- Elderly volunteers: digital barriers, vulnerable collections
- Distrust towards researchers

Variation

- Total orthographic variation in text, large phonological variation in speech. Not tagged or annotated (yet)