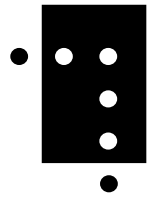
NB Uttale: A Norwegian Pronunciation Lexicon with Dialect Variation

Marie Iversdatter Røsok & Ingerid Løyning Dale

The Language Bank

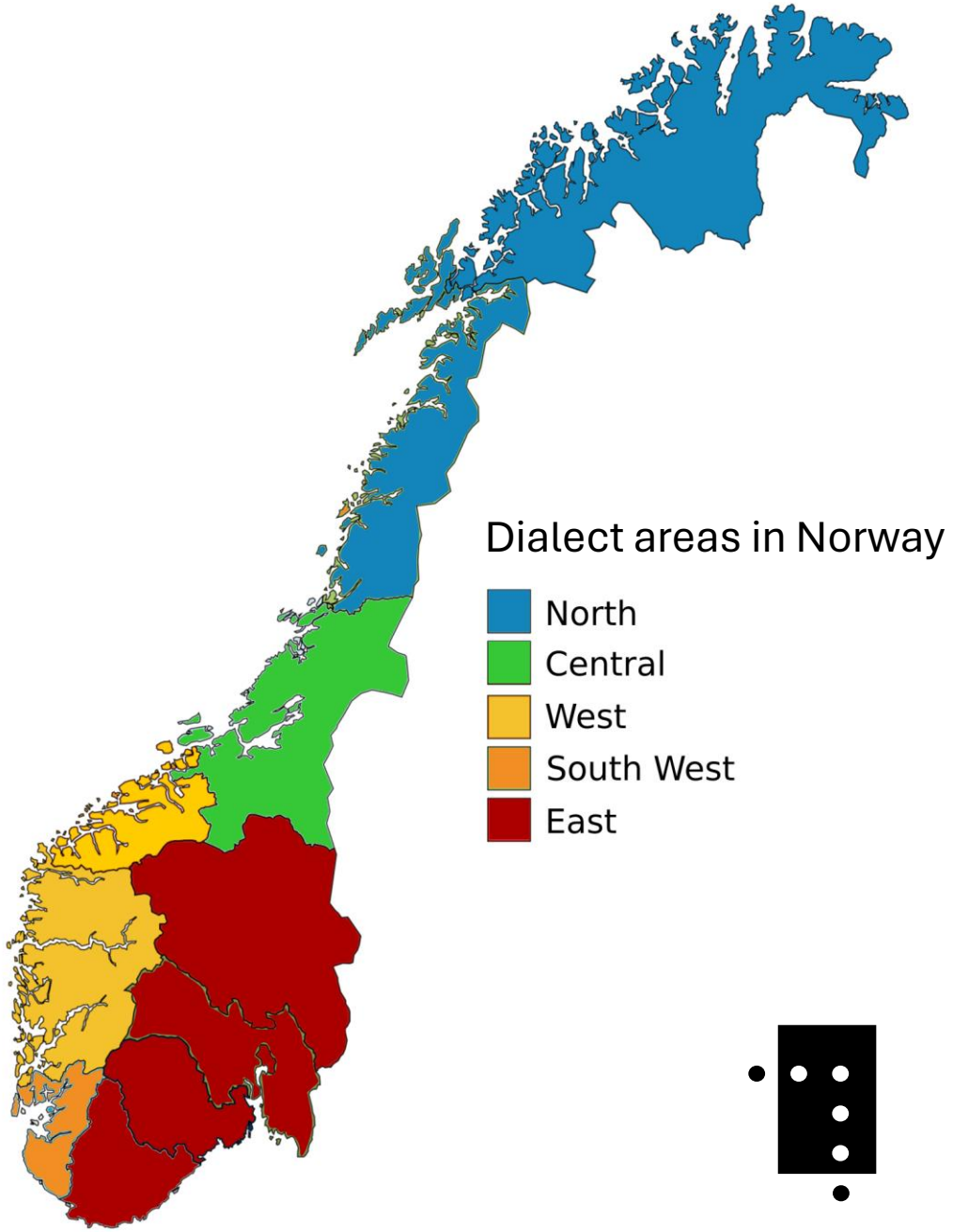
The National Library of Norway

LREC-Coling 2024



NB Uttale

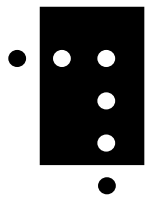
- Nearly 800 000 word entries for Norwegian Bokmål
- Phonemic transcriptions for five dialect areas
  - East (e)
  - West (w)
  - Southwest (sw)
  - Central (t)
  - North (n)
- Close-to-spoken and close-to-written style variants



```
# Replace AX0 N AX0 adjective ending with AH0 N AX0 S
# S P EH2 N NX0 AX0 --> S P EH2 N AH0 N AX0 S
dialect_anes_adjective = {
    'name': 'dialect_anes_adjective',
    'areas': [
        'n_spoken',
        'sw_spoken'],
    'rules': [
        {
            'pattern': r'\b(AX0 N|NX0) AX0( S)?$',
            'replacement': r'AH0 N AX0 S',
            'constraints': [
                {
                    "field": 'pos',
                    "pattern": r'JJ|VB',
                    "is_regex": True
                },
                {
                    "field": 'wordform',
                    "pattern": r'endes?$',
                    "is_regex": True
                },
            ],
        },
    ]
}
```

Transformation rules

- Based on NST pronunciation lexicon (*Nordisk språkteknologi*)
- Phonemic differences between dialects
- Nofabet transcription standard
- Regular expressions
- Constraints



Error corrections

- Syllabic retroflex nasal

  bakeren (the baker), /ˈbɑː.kæ.**ɳ̩**/ → /ˈbɑː.kæ**ɳ̩**/

Dialect updates

- *Non-retroflexes (west, southwest):*

  forsker (scientist), /fɔ**ʂ**kər/ → /fɔ**rs**kər/

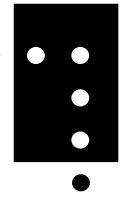
- *Apocope (central, north):*

  elvene (the rivers), /ɛlʊ**ənə**/ → /ɛlʊ**ɑn**/

- Hard *d* (west):

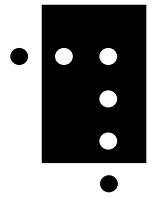
  kveld (night), /kʊɛl/ → /kʊɛl**d**/

rs, rn, rl, rd, rt:
/ʂ ɳ ɭ ɖ ʈ/ → /rs rn rl rd rt/



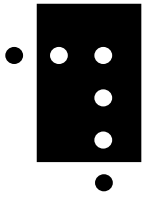
Evaluation

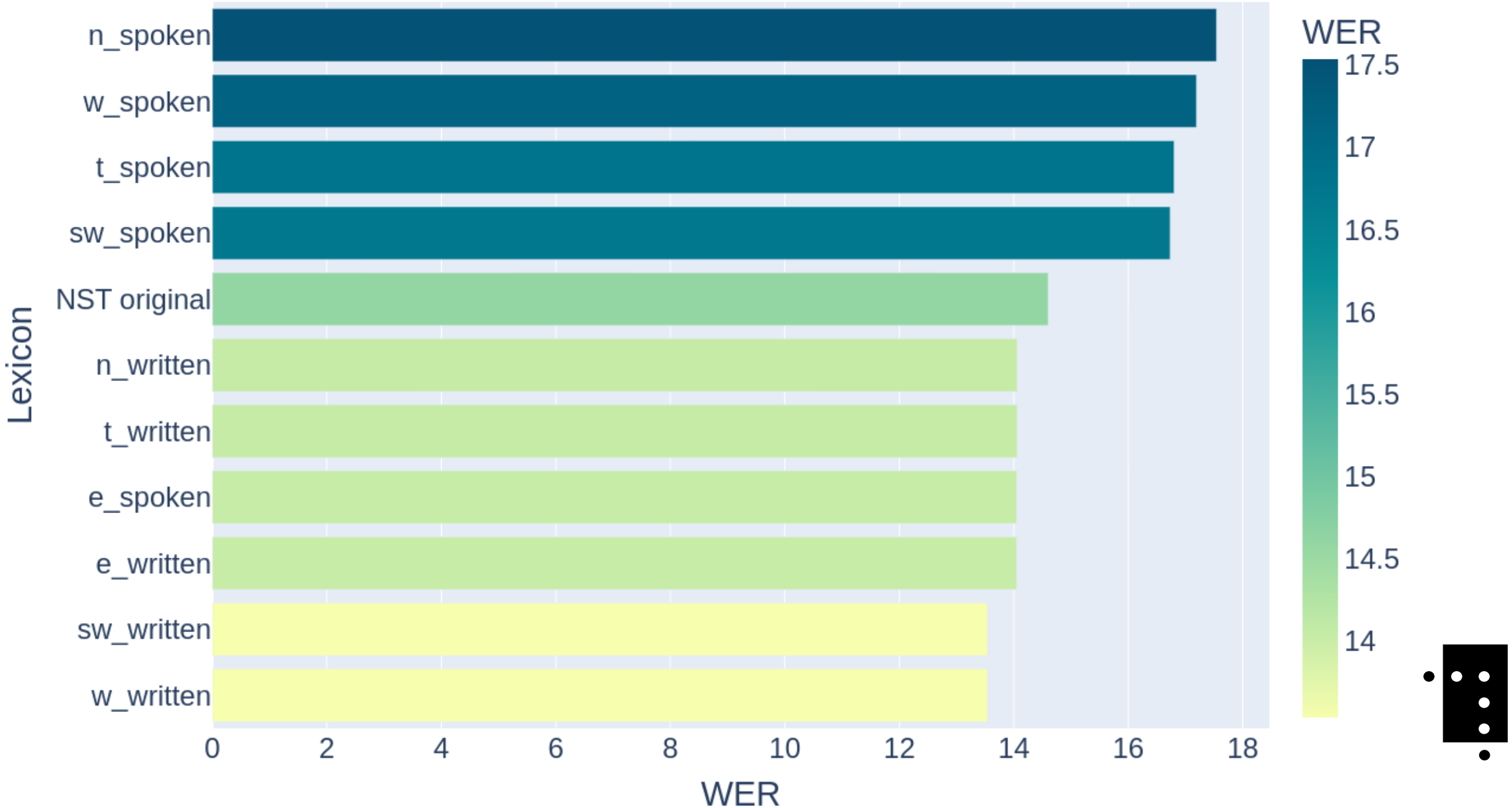
- Consistency
- Correctness
- Statistical predictability as quantitative measure
- Grapheme-to-phoneme models

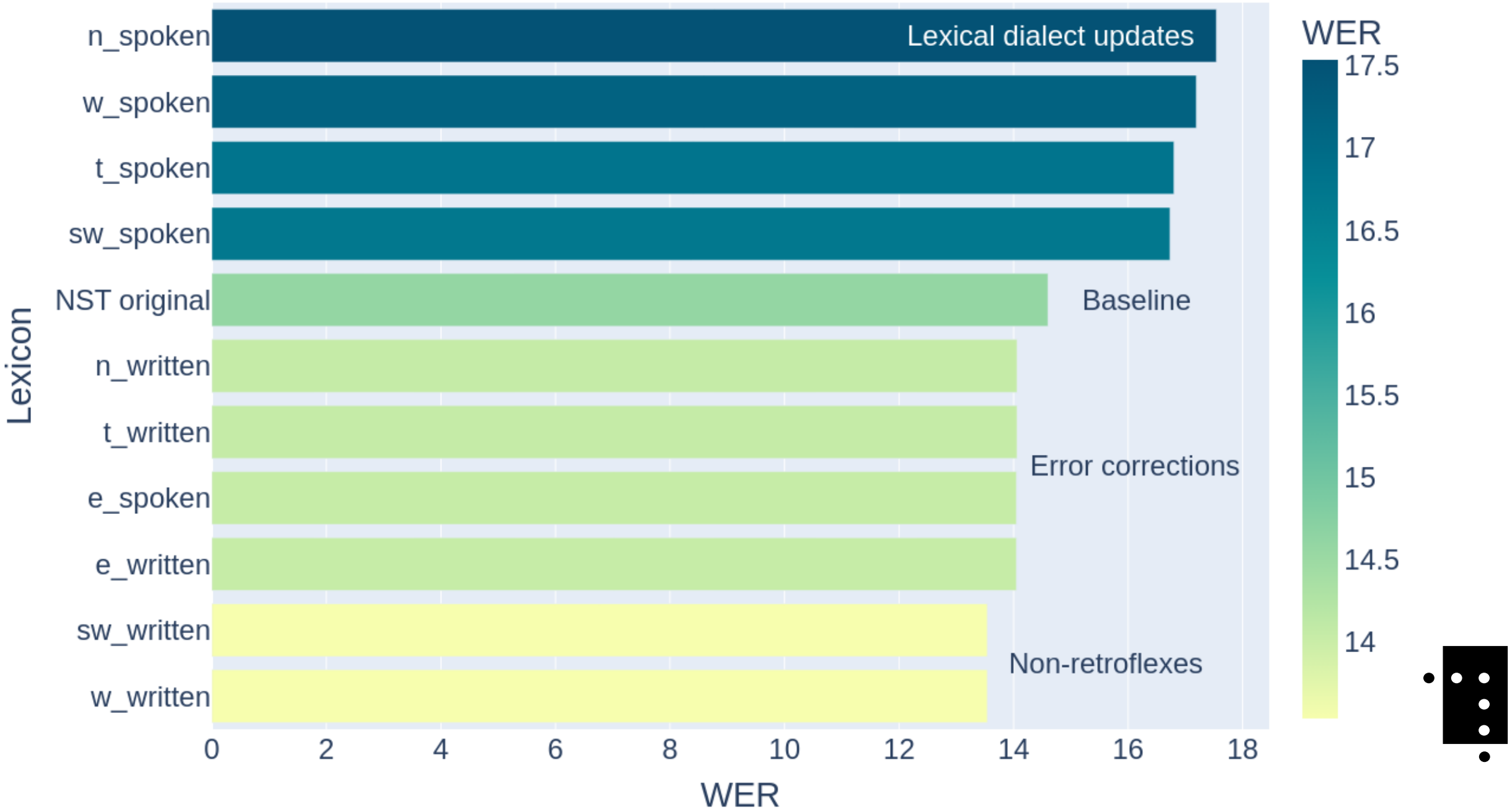


Grapheme-to-phoneme models (G2P)

- 1 per dialect + pronunciation style
- Phonetisaurus + Open FST
  1. Aligned grapheme and phoneme sequences (aligned corpus)
  2. 8-gram models
  3. Finite state transducer models




n_spoken
w_spoken
t_spoken
sw_spoken
NST original
n_written
t_written
e_spoken
e_written
sw_written
w_written
Lexicon
0
2
4
6
8
10
12
14
16
18
WER
WER
17.5
17
16.5
16
15.5
15
14.5
14


n_spoken
w_spoken
t_spoken
sw_spoken
NST original
n_written
t_written
e_spoken
e_written
sw_written
w_written
Lexical dialect updates
Baseline
Error corrections
Non-retroflexes
Lexicon
0
2
4
6
8
10
12
14
16
18
WER
WER
17.5
17
16.5
16
15.5
15
14.5
14

Thank you!

sprakbanken@nb.no

This work has in part been supported by the IKTPLUSS grant for the SCRIBE project financed by the Research Council of Norway (KSP2IPD).