Exploring a Joint Approach for Analyzing Reading and Writing Errors in Dutch

Wieke Harmsen, Roeland van Hout, Catia Cucchiarini and Helmer Strik

Centre for Language Studies
Radboud University Nijmegen
The Netherlands

14th AWLL international workshop on writing systems and literacy
10-12 November 2023

Introduction

- Reading and writing are important skills in our literate society
- Both are learned skills, that require:
  - Direct instruction [1]
  - Active practice [2]
  - Feedback
- Analyzing reading and writing errors can provide insights into the processes underlying literacy acquisition.

Introduction

This type of research is now possible:

1. Availability of **large corpora** of children's oral reading and writing data in Dutch

Reading corpora
- JASMIN (1st-6th graders, story reading) [3]
- DART (1st graders, word and story reading) [4]
- ASTLA (2nd and 3th graders, word and story reading)

Writing corpora
- BasiScript (dictations and essays) [5]

2. Language and speech technology to analyze these corpora

Research Aim
Developing a joint approach for automatic analysis of reading and writing errors in Dutch

Error Detection

Error Classification

Background

Reading and writing in Dutch

- Dutch is written using **Latin alphabet**
- Alphabetical principle: letters represent speech sounds
- Dutch is a relatively transparent language

Terminology: PCUs [6]

- Reading errors and spelling errors are recognized at **PCU-level**.
- A **Phoneme-Corresponding Unit (PCU)** is
  a sequence of graphemes that corresponds to one phoneme.

Example of PCU segmentation

school

Dutch (IPA)	s	χ	oː	ɬ
English (IPA)	s	k	u	ɬ
PCUs	**s**	**ch**	**oo**	**l**

The analysis pipeline: inputs

Writing

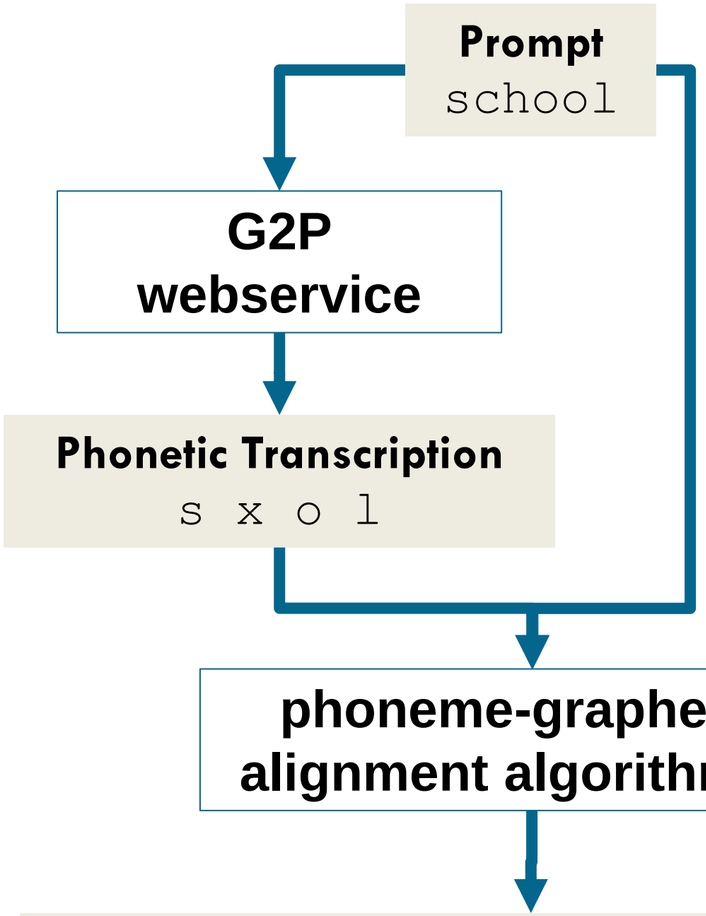
Target	Realized
school	sgool

Reading

Target	Realized
s x o l[1]	s x u l

[1]All phonetic transcriptions are written in the phonetic alphabet of the Spoken Dutch Corpus (version 2).

Step 1: Phoneme-grapheme alignment




```
target_phon      s x  o  l
target_graph     s ch oo l
```

Rule-based algorithm, contains:

All possible PCUs to write a certain phoneme:

```
phoneme   : PCUs
"a"       : ["aa", "a", "ä", "á"]
"A"       : ["a", "e", "ä", "ah"]
"b"       : ["bb", "b"]
"d"       : ["dd", "d", "t"]
…
```

Step 2: ADAGT & ADAPT

WRITING

ADAGT [7]

- Algorithm for Dynamic Alignment of ***Grapheme*** Transcriptions

```
target_graph  school
real_graph    s-gool
```

- Vowels are aligned with vowels, consonants with consonants
- Equal distance between graphemes

READING

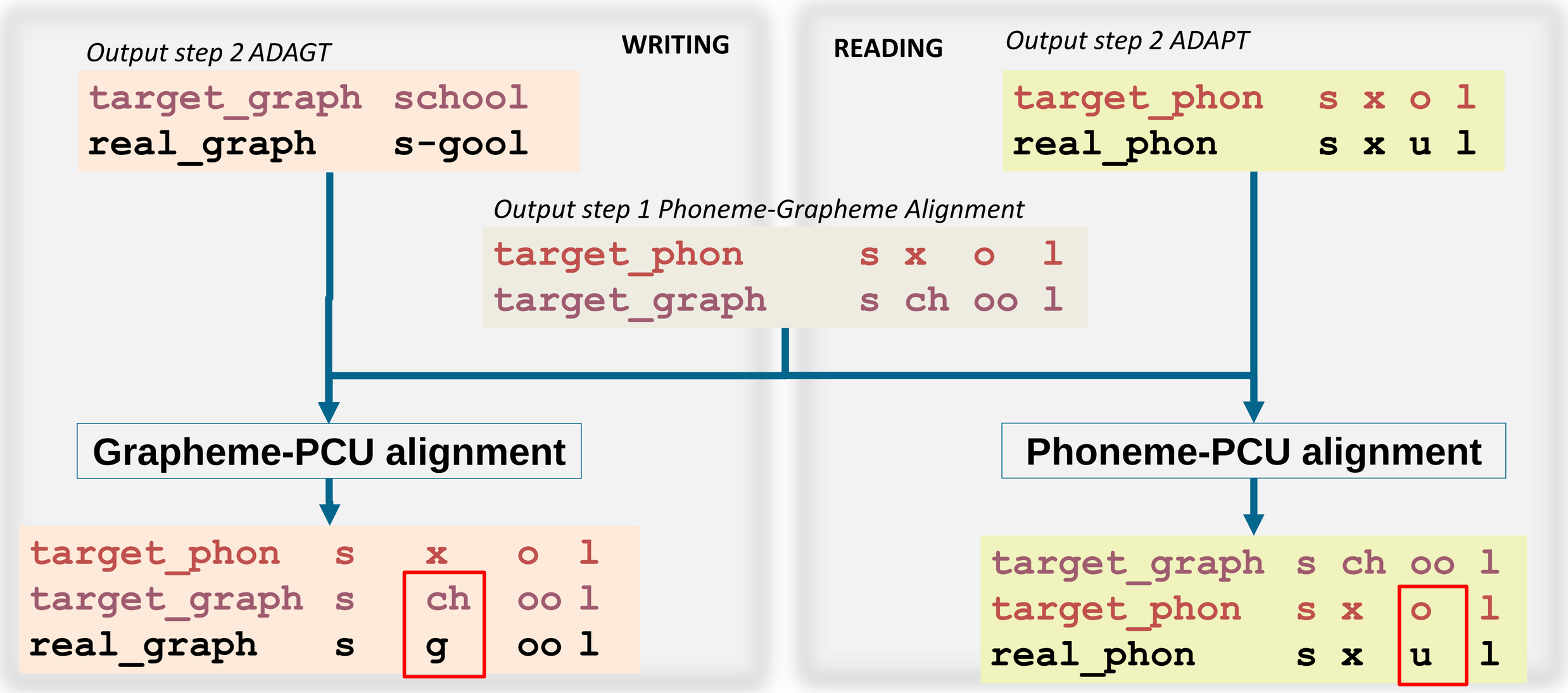
ADAPT [8]

- Algorithm for Dynamic Alignment of ***Phonetic*** Transcriptions

```
target_phon   s x o l
real_phon     s x u l
```

- Vowels are aligned with vowels, consonants with consonants
- Distance between phonemes determined by articulatory feature vectors

Step 3: Error Detection



Step 4: Error Annotation

Annotate each incorrect PCU-phoneme mapping with:

- Which PCU (spelling) or phoneme (reading) was written in the realized transcription

WRITING

target_graph	s	ch	oo	l
target_phon	s	x	o	l
real_graph	s	g	oo	l

“ch”-/x/ -> “g”
Interpretation
“ch” is written incorrectly as “g”. Both “g” and “ch” can be pronounced as /x/.

READING

target_graph	s	ch	oo	l
target_phon	s	x	o	l
real_phon	s	x	u	l

“oo”-/o/ -> /u/
Interpretation
“oo” is read incorrectly as /u/, it should be read as /o/.

Application of Algorithm

Data selection

BasiScript	Dictation words by 2nd graders
DART	Read words by 1st graders

Sound pure words

Words consisting only of primary PCU-phoneme mappings

Primary PCU-phoneme mappings

How initial readers are taught to pronounce certain PCUs at the beginning of primary school

Vowels

PCU	Phoneme
aa	/a/
a	/A/
ee	/e/
e	/E/
ei	/EI/
ij	/EI/
ie	/i/
i	/I/
oo	/o/
eu	/EU/
u	/U/
ui	/UI/
o	/O/
oe	/u/
au	/AU/
ou	/AU/
uu	/y/
u	/U/

Consonants

PCU	Phoneme
b	/b/
d	/d/
f	/f/
h	/h/
j	/j/
k	/k/
l	/l/
m	/m/
n	/n/
ng	/N/
p	/p/
r	/r/
s	/s/
t	/t/
v	/v/
w	/w/
g	/x/
ch	/x/
z	/z/

Application of Algorithm

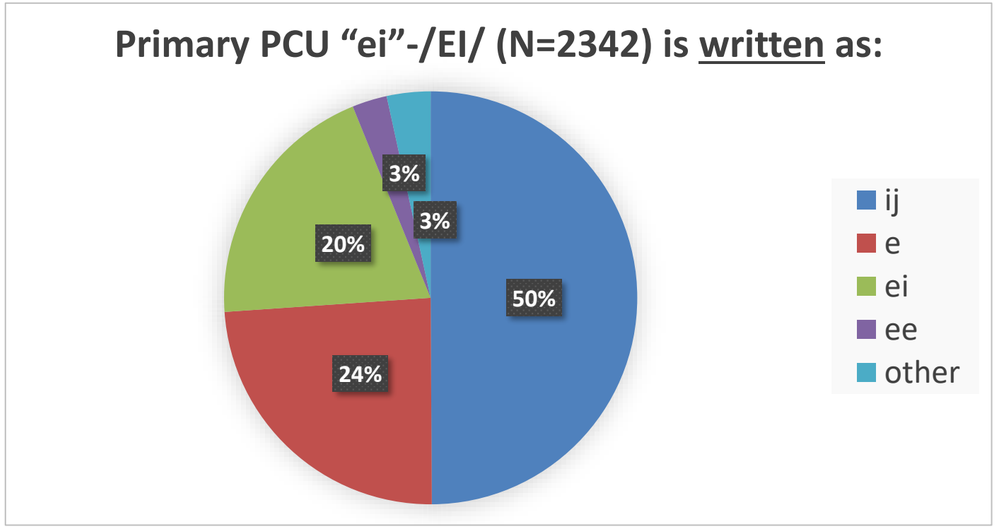
Data selection

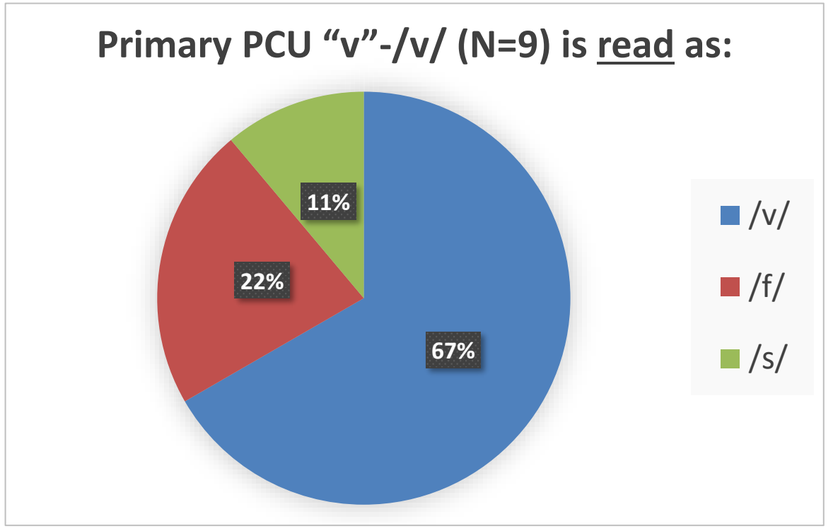
- Sound pure prompts

Corpus	Word types	Word tokens	Tokens Correct	Tokens Incorrect
BasiScript (writing)	9	21168	15323 (72%)	5845 (28%)
DART (reading)	51	323	207 (64%)	114 (36%)

Relatively most frequent incorrect PCUs:

- Spelling: ’ei’, ’eu’, ’g’, ’ij’, ’ch’
- Reading: ’v’, ’ui’, ’ng’, ’a’ , ’g’





Discussion and Conclusion

We presented a first approach for a method to automatically detect reading and spelling errors at PCU-level.

Advantages

- Reading and writing errors are comparable at target grapheme level and target phoneme level.
- Bridge between phonetic and grapheme representations in both reading and writing

Limitations

- Pronunciation variation not taken into account
- Only applicable on sound pure words, otherwise number of target PCU-phoneme mappings will explode and results will be hard to interpret.
- Multiple attempts (typical for reading) not taken into account

Future Directions

- Extent classification scheme
  - Combine current categories in a supercategory
  - Take into account more complex rules, marked by morphology, etymology, semantics
- Use ASR technology to automatically obtain phonetic transcriptions

Radboud University

Questions?

wieke.harmsen@ru.nl

Radboud University Nijmegen, The Netherlands

This work is part of ASTLA project with project no. 06.20.TW.009, which is (partly) financed by the Dutch Research Council (NWO).

References

[1] E. M. H. Assink, “Verkennen kinderen spontaan orthografische regels?,” in *Tijdschrift voor Taalbeheersing,* vol. 8, pp. 106-118, 1986.

[2] K. A. H. Cordewener, L. Verhoeven and A. M. T. Bosman, “Improving Spelling Performance and Spelling Consciousness,” in *The Journal of Experimental Education*, vol. 84, no .1, pp. 48-74, 2016.

[3] Cucchiarini, C., Van hamme, H., Herwijnen, O., & Smits, F. (2006). JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*. European Language Resources Association (ELRA).

[4] Y. Bai, F. Hubers, C. Cucchiarini, R. van Hout, and H. Strik. 2022. The Effects of Implicit and Explicit Feedback in an ASR-based Reading Tutor for Dutch First-graders. In *Proc. Interspeech 2022*, pages 4476–4480.

[5] A. Tellings, N. Oostdijk, I. Monster, F. Grootjen, and A. van den Bosch, “BasiScript: a corpus of contemporary Dutch texts written by primary school children,” in *International Journal of Corpus Linguistics*, vol. 23, no. 4, pp. 494–508, 2018.

[6] R. Laarmann-Quante, “Automating multi-level annotations of orthographic properties of German words and children’s spelling errors,” in *Proceedings of the 2nd language teaching, learning and technology workshop, 2016. LTLT 2016*. pp. 14-22, 2016.

[7] Harmsen, W. N., Cucchiarini, C., & Strik, H. (2021). Automatic Detection and Annotation of Spelling Errors and Orthographic Properties in the Dutch BasiScript Corpus. *Computational Linguistics in the Netherlands Journal*, *11*, 281–306. Retrieved from https://www.clinjournal.org/clinj/article/view/140