A Joint Approach for Automatic Analysis of Reading and Writing Errors

Wieke Harmsen, Catia Cucchiarini, Roeland van Hout, Helmer Strik
Centre for Language Studies
Radboud University Nijmegen, The Netherlands
{wieke.harmsen, catia.cucchiarini, roeland.vanhout, helmer.strik}@ru.nl

Abstract

Analyzing the errors that children make on their ways to becoming fluent readers and writers can provide invaluable scientific insights into the processes that underlie literacy acquisition. To this end, we present in this paper an extension of an earlier developed spelling error detection and classification algorithm for Dutch, so that reading errors can also be automatically detected. The strength of this algorithm lies in its ability to detect errors at Phoneme-Corresponding Unit (PCU) level, where a PCU is a sequence of letters corresponding to one phoneme. We validated this algorithm and found good agreement between manual and automatic reading error classifications. We also used the algorithm to analyze words written by second graders and words read by first graders. The most frequent PCU errors were *ei*, *eu*, *g*, *ij* and *ch* for writing, and *v*, *ui*, *ng*, *a* and *g* for reading. This study shows how a joint approach for the automatic analysis of reading and writing errors can be implemented. In future research the value of this algorithm could be tested by analyzing corpora containing initial reading and writing data from the same children.

Keywords: Automatic spelling and reading error detection, Reading and writing instruction, Child read speech corpora, Child written language corpora, Phoneme-grapheme alignment, Dutch

1. Introduction

Reading and writing are both skills that children acquire after long periods of intensive instruction and practice. In this sense, reading and writing are essentially different from speaking and listening, which are skills that children spontaneously acquire in daily interaction. The processes of learning to read and write require children to engage in long and sustained practice, preferably under teachers' guidance. During practice, children inevitably make reading and writing errors that teachers need to correct to make children aware of their gaps in knowledge and to help them improve their reading aloud and writing skills. Analyzing the errors that children make on their ways to becoming fluent readers and writers can provide invaluable scientific insights into the processes that underlie literacy acquisition.

Analyzing the child development of reading and spelling errors is not an easy task. There are three important reasons that make this task challenging. Firstly, there is no corpus available that contains longitudinal reading and writing data from the same children. Secondly, a classification scheme that can capture a large variety of both reading and writing errors has not yet been developed. Thirdly, the task of detecting and classifying reading and writing errors manually is laborious and time-consuming.

Recent developments in the field of language and speech technology have made it possible to overcome these challenges partially. For Dutch, a medium-sized language, there are four corpora available that can be used for either reading error or writing error analysis. These are JASMIN, a small corpus of Dutch and Flemish child speech (Cucchiarini et al., 2006), CHOREC, a corpus of Dutch speech by Flemish elementary school children (Cleuren et al., 2008), BasiScript, a corpus of written texts and dictations produced by children in primary school (Tellings et al., 2018a), and DART, a larger corpus of child read speech by first graders (6-7 years old) (Bai et al., 2022). These corpora make it possible to develop scientifically and pedagogically sound error classification schemes, as well as algorithms that apply these schemes to analyze both reading aloud and writing data automatically.

Using a joint classification scheme opens up opportunities for research on literacy acquisition in which reading and writing development are investigated in combination to gain insight into their differences and their interaction. This type of research would profit enormously from longitudinally collected data collections on reading and writing of course, but such databases are not available for Dutch, although we hope that these might be come about in the near future. In any case, an important prerequisite for comparing reading and writing skills is the development of a joint classification scheme that captures both reading aloud and spelling errors. For developing such a scheme, it is not necessary to have reading and writing data of the same children.

Our aim is to present a joint classification scheme for Dutch reading and spelling errors, together

Second Workshop on Computation and Written Language (CAWL 2024) @LREC-COLING-2024, pages 8–17
May 21, 2024.

with an algorithm that can automatically detect and classify these errors in corpora of read speech and written language. As a first step in this direction, we expanded an existing spelling error detection and classification algorithm (Harmsen et al., 2021b,a) so that it can also detect and classify reading errors. In addition, we applied this expanded algorithm to separate corpora of reading and spelling data from children in Dutch primary school. We describe the type and frequency of reading and spelling errors that we found.

2. Background

2.1. Three essential competences for reading and spelling in Dutch

When children learn to read and write, three competences become relevant. In the first place, phonological awareness (e.g., van Druenen et al. (2019)), which is knowing that words are built from phonemes (i.e., sounds). Phonologically aware learners are able to segment words into phonemes (auditory analysis) and to combine phonemes into words (auditory synthesis).

Written Dutch uses the Latin alphabet, so the second competence involved is knowledge of the alphabetical principle: a grapheme (i.e., single letter) or sequence of graphemes represents a phoneme and vice versa. Borgwaldt et al. (2004) compared the orthographic transparency of five languages that use the Latin alphabet. A transparent orthography is defined as an orthography with both a high feedforward consistency (one-to-one grapheme to phoneme mappings) and high feedback consistency (one-to-one phoneme to grapheme mappings). They conclude that Dutch orthography has as intermediate transparency. In addition, Dutch has a higher feedforward consistency than feedback consistency. This is one reason why reading in Dutch is considered to be less difficult than spelling (Bosman and Van Orden, 1997).

Finally, the child has to acquire an explicit morphological awareness, since Dutch morphological principles are part of the writing system. This may result in the phenomenon that words are pronounced differently than one would expect based on their written form and the set of learned phoneme-grapheme mappings. For example, the verb *hij vindt* (he finds) consists of two morphemes: *vind* and *t*, since it is constructed by taking the root *vind* and adding the third person singular suffix *t*. In this verb, the *dt* is pronounced as a single /t/[1], which means that the pronunciations of *vind* and *vindt* are exactly the same: /v I n t/ (due to final devoicing, the *d* in the root *vind* is pronounced as /t/).

2.2. Existing classification schemes

So far, each study researching reading or spelling errors in Dutch used its own classification scheme. Most classification schemes manually labeled each misspelled word with a label describing the reading or spelling error (e.g. Kleijnen (1997); Cleuren et al. (2008); Tellings et al. (2018b); Limonard et al. (2020)). A disadvantage of this approach is that it is not clear which part of the word is written incorrectly, and which letters are substituted, deleted or inserted in comparison with the target word.

Another type of classification scheme that was mainly used in research on automatic pronunciation assessment in alphabetic languages, defined reading errors as phonemes that are inserted, deleted or substituted in comparison with the phonetic transcription of the target word (e.g., Zhang et al. (2021); Lin and Wang (2022); Gelin et al. (2023)). These studies were able to return the phoneme that was read incorrectly, but not the letters in the target word that represented this phoneme. That means that important diagnostic information was missing.

2.3. Phoneme-Corresponding Units

In Dutch, a single phoneme can be represented by one or a sequence of two or even more graphemes. In line with Laarmann-Quante (2016), we refer to these grapheme representations as Phoneme-Corresponding Units (PCUs). For example, the Dutch word *maan* (moon) consists of three phonemes /m a n/, and thus three PCUs: *m*, *aa* and *n*. The Dutch word *bureau* (desk) consists of four phonemes /b y r o/, and thus four PCUs: *b*, *u*, *r* and *eau*. Unfortunately the number of phonemes and the PCUs a word consists of are not always equal. Sometimes, a word can have more PCUs than phonemes. This is possible because some graphemes in Dutch are not pronounced. An example is the diminutive name of cupboard *kastje* (*kast* (noun) + *je* (diminutive suffix), little cupboard) with phonetic transcription /k A s j @/ and PCU segmentation *k*, *a*, *s*, *t*, *j*, *e*. In this example, the PCU *t* is not pronounced.

2.4. Primary PCU-phoneme mappings and sound pure words

A distinction can be made between primary and secondary PCU-phoneme mappings. The primary PCU-mappings mark a fixed, unique link between a PCU and a phoneme. They are the PCU-phoneme mappings that are taught initially to beginning readers and writers in first grade of primary school. The primary PCU-phoneme mappings are

[1]All phonetic transcriptions in this paper are written between slashes and in the computer phonetic alphabet CGN2 (Gillis, 2001).

VOWELS		CONSONANTS	
Phoneme	**PCU**	**Phoneme**	**PCU**
/a/	aa	/b/	b
/o/	oo	/d/	d
/y/	uu	/f/	f
/e/	ee	/h/	h
/i/	ie	/j/	j
/A/	a	/k/	k
/O/	o	/l/	l
/U/	u	/m/	m
/E/	e	/n/	n
/I/	i	/N/	ng
/EI/	ei	/p/	p
/EI/	ij	/r/	r
/EU/	eu	/s/	s
/UI/	ui	/t/	t
/u/	oe	/v/	v
/AU/	au	/w/	w
/AU/	ou	/x/	g
/@/	e (schwa)	/x/	ch
		/z/	z

Table 1: Dutch primary PCU-phoneme mappings.

presented in Table 1. This table contains for each phoneme just one way to write it, except for the phonemes /EI/ and /OU/, which can be written in two ways. Words containing only primary PCU-phoneme mappings are called *sound pure* words (*klankzuiver* in Dutch).
Not all words can be written correctly using only primary PCU-phoneme mappings. For example, the Dutch language also contains loanwords, words containing a schwa, and words whose spelling depends on morphology. In these cases, secondary PCU-phoneme mappings are used to write these words correctly. Examples of secondary PCU-phoneme mappings are: *e*-/@/ in *bestaan* (to exist) and *eau*-/o/ in *bureau* (desk).

2.5. Dutch PCU-based spelling error detection and classification

In an earlier study (Harmsen et al., 2021a,b), an algorithm was presented that could detect spelling errors by aligning the realized spelling (including spelling errors) with the target spelling using the phonetic transcription of the target spelling. In this algorithm, a spelling error was defined as an inserted, deleted or substituted PCU.
We illustrate the strength of this algorithm with an example. Where a Levenshtein-based alignment algorithm would most likely detect two errors in the misspelling *lag* of the word *lach* (laugh), namely a substitution of *c* with *g* and a deletion of *h*, the newly presented algorithm that also uses the phonetic transcription of the target (i.e., /l A x/) could recognize that both the *ch* and *g* correspond to the same phoneme /x/ and align them with each other. In this way, the detected spelling error is defined as a substitution of the PCU *ch* pronounced as /x/) with *g*, which is more informative.

2.6. The current research

Given that our aim is to develop a joint classification scheme for reading aloud and spelling, we address the following more specific research questions:

1. To what extent can we accommodate and expand an automatic spelling error detection and classification algorithm (Harmsen et al., 2021a,b) in such a way that it is able to detect sound pure reading errors on the basis of phonetic transcriptions of audio recordings?

2. We address three subquestions to make a comparison between reading and spelling:

   (a) Which sound pure PCUs are most frequently written incorrectly by initial writers?

   (b) Which sound pure PCUs are most frequently read incorrectly by initial readers?

   (c) How do these patterns compare?

3. Method

3.1. Data sets

The reading data set consists of phonetic transcriptions of audio that were collected within the project *Dutch Automatic Reading Tutor (DART)* (Bai et al., 2020). In this project, grade 1 pupils (6-7 years old) practiced reading for six weeks (twice a week for 10 minutes) with a system that provided feedback on their reading aloud. Before and after these practice weeks, each pupil had to take three pretests and three posttests. Each test consisted of a list of 24 words that the pupils had to read in one go while their speech was recorded.
For the current study, we selected phonetic transcriptions of 28 audio recordings from the DART pretest and posttest dataset. An annotator made the phonetic transcriptions of the audio. In total, the reading data consisted of phonetic transcriptions of 672 words.
The written data set used in this study consists of 2352 dictations from the BasiScript corpus (Tellings et. al., 2015) that were written by grade 2 pupils (7-8 years old). Each pupil wrote the same dictation, consisting of 25 words. The dictations were originally handwritten by the pupils, digitized (typed) and stored in the BasiScript corpus. In total, the writing data consisted of 58,800 words.

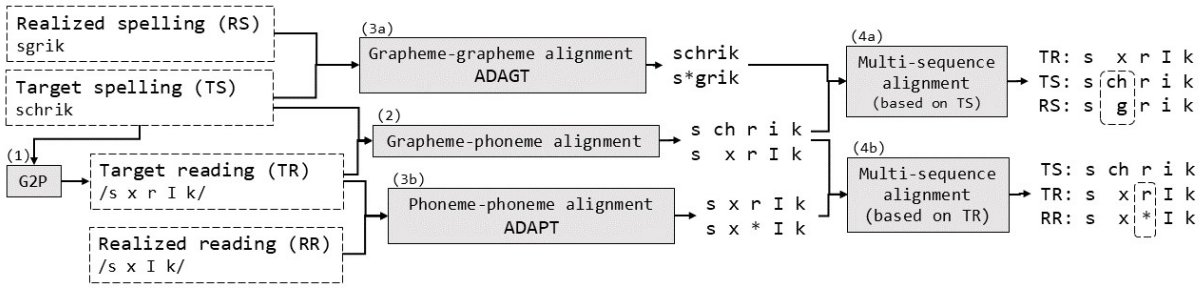

Figure 1: An example of application of the spelling and reading error detection algorithm on the word *schrik* (fright), spelled incorrectly as *sgrik* and read incorrectly as /s x I k/.

3.2. Spelling and reading error detection

To automatically detect reading and spelling errors at PCU-level in the realized readings and spellings, we extended an earlier presented algorithm developed for automatic spelling error detection (Harmsen et al., 2021a). This spelling error detection algorithm consists of four parts. To be able to detect reading errors at PCU-level, we had to create separate variants of parts 3 and 4 of the algorithm. Figure 1 visualizes the four parts in the analysis using an example.

First, the phonetic transcription of the target spelling was obtained automatically using a Dutch grapheme-to-phoneme converter (G2P) webservice (Ten Bosch, 2019). This is the target reading. Secondly, the target reading and target spelling are aligned. This is done using a dictionary that defines all possible ways a phoneme can be written in Dutch. For example, the phoneme /x/ can be written as *g* and as *ch*. Starting from the third step, the analysis procedure is different for the spelling and reading error detection. With respect to reading error detection, the order of the phonemes of the target reading and realized reading is first reversed, and subsequently aligned using the *Algorithm for Dynamic Alignment of Phonetic Transcriptions (ADAPT)* (Elffers et al., 2013). In this algorithm, articulatory features are incorporated to define the distance between two phonemes (Cucchiarini, 1993, 1996). The resulting alignment is again reversed, so that the phonemes are in the correct order again. The readings are reversed, because we want the target reading to match with the final attempt of the speaker to read a word. For spelling error detection, step three consists of aligning the graphemes of the target spelling and realized spelling with each other using the *Algorithm for Dynamic Alignment of Graphemic Transcriptions (ADAGT)* (Bai et al., 2021; Harmsen et al., 2021a). This is an adaptation of ADAPT, made suitable for grapheme alignment. This algorithm aligns vowels only with vowels and consonants only with consonants. In the fourth step, multi-sequence alignment is performed. The output alignments of step 2 and 3 are combined based on their overlapping transcription: the target spelling for spelling error detection and the target reading for reading error detection. In this way, the PCU-segmentation of the realized spelling is deduced. In this final spelling alignment, spelling errors can be detected as PCUs that are spelled incorrectly. In the final alignment output of the reading error detection pipeline, reading errors can be detected as PCUs that are read incorrectly.

We found that the phonetic transcriptions provided by the G2P webservice were not always consistent. To overcome this problem, we made the following decisions. Words with a grapheme transcription ending in *-en* (e.g., *kijken* (to watch)) get a phonetic transcription ending in /@/ (schwa). In addition, we specified that words ending in the graphemes *-auw* or *-ouw* get the phonetic transcription /OU/ and not /OU w/. Finally, we decided to make no distinction between /x/ (unvoiced) and /G/ (voiced), since the G2P output is not consequent in making this decision, and Dutch speakers do not consistently pronounce them as either voiced or unvoiced either.

3.3. Annotation

Phoneme-PCU transparency The result of step two of the error detection algorithm was the alignment of target graphemes and phonemes, which resulted in a sequence of PCU-phoneme mappings the word consists of. In this step, we labeled each PCU-phoneme mapping as either sound pure (in case it occurred in Table 1), or not sound pure. When a target word only contained primary PCU-phoneme mappings, the word was annotated as sound pure. This annotation layer was called *SoundPure*.

Multiple attempts A frequently occurring phenomenon in initial readers is that they need multiple attempts to read the target word. In the aligned target graphemes, target phonemes and realized phonemes, this is visible as one or more

insertions at the beginning of the alignment. Using a search function, we automatically annotated multiple attempts at the word level. The results were stored in the annotation layer *MultipleAttempts_Automatic*.

Reading error type The next step is to label each aligned target phoneme and realized phoneme with an error type: insertion (a phoneme was added), deletion (a target phoneme was not read), substitution (a target phoneme was substituted). This is done by comparing each target phoneme and realized phoneme pair one-by-one. After that, from these alignments on phoneme level, a classification at word level is computed: correct, insertion (only one phoneme was inserted in the complete word), deletion (only one phoneme was deleted in the complete word), substitution (only one phoneme was substituted in the complete word), multi (there are multiple phonemes inserted, deleted and/or substituted) and delWord (the complete word is not read). These classifications were saved in the categorical annotation layer *ErrorType_Automatic*.

3.4. Analysis 1: Validation

The spelling error detection algorithm was validated in an earlier study (Harmsen et al., 2021a). To validate the performance of the automatic reading error detection algorithm, one annotator manually annotated all readings from the selected DART data in two layers. The first layer contains for each target reading a boolean value which indicates whether there are multiple attempts or not (*MultipleAttempts_Manual*). The second layer, *ErrorType_Manual*, contains a categorical value for each reading. This value describes the reading error type using the following values: correct, insertion, deletion, substitution, multi and delWord. To validate the performance of the reading error detection and classification algorithm, we computed Matthew's Correlation Coefficient (MCC) between the automatic and manual annotation layers of *MultipleAttempts* and *ErrorType*. We used the MCC metric since our dataset is unbalanced, as it has more correct than incorrect readings. The MCC is proven to be more trustworthy than Cohen's Kappa on imbalanced datasets (Chicco et al., 2021).

3.5. Analysis 2: Application

From both the reading data set (phonetic transcriptions of 672 read words) and the spelling data set (digitized writings of 58,800 dictation words), we automatically selected the sound pure target words, using the annotation layer *SoundPure*. The sound pure target readings and spellings are listed in Table 2.

Corpus	Sound pure target words
BasiScript (N=9)	fee huis keus ligt lip monteur rijst schrik steil
DART (N=51)	bal blauw boomstam buik deuk dof flits fop gat geit hout jaap jong juicht keelpijn klets koen kous lach lift lijn lus markt meetlat melk mug muis muur nicht proost reis saus schoen schraal schrift schrik schroef schroot schuur specht spierkracht sportpark sterk stoep strik toch vang vorst vuur warmst zwart

Table 2: The sound pure target words from the DART reading tests and BasiScript dictations.

The spelling and reading error detection pipeline yields for each realized spelling an alignment of the target graphemes, target phonemes and realized phonemes, and for each realized reading, an alignment of target graphemes, target phonemes and realized phonemes. From these two multi-sequence alignments, we extract a list of target PCUs that are spelled at least one time incorrectly, and a list of target PCUs that are read at least one time incorrectly. For each PCU in each list, we compute the following measures:

Number of different targets The number of different target words in which the target PCU occurs. If this number is small, the target words themselves are printed.

Number of realized writings/readings (N) How often the target words that contain the selected target PCU are spelled/read in the complete dataset.

Absolute incorrect count How often the target PCU was spelled/read incorrectly.

Relative incorrect percentage How often the target PCU was spelled/read incorrectly with respect to how often this target PCU had to be spelled/read in total.

Aligned realized PCUs/phonemes A list of incorrect realized PCUs (in case of spelling errors) or phonemes (in case of reading errors) of the target PCU. The list is sorted from most to least frequently occurring realization

4. Results

4.1. Validation of the reading error algorithm

We computed the agreement between *MultipleAttempts_Manual* and *MultipleAttempts_Automatic*

Corpus	Total	Correct	Incorrect
BasiScript	21168	15323 (72%)	5845 (28%)
DART	321	207 (64%)	114 (36%)

Table 3: The number of analyzed realized writings (BasiScript) and readings (DART) of sound pure target words, together with their classification as either correctly or incorrectly read/spelled.

and found MCC = 0.87. In addition, we computed the agreement between *ErrorType_Manual* and *ErrorType_Automatic* and we found MCC = 0.92. So, for both *MultipleAttempts* and *ErrorType*, we found a high agreement between the manually and automatically obtained values.

4.2. Application

4.2.1. Description of the data

Table 3 presents the number of times a sound pure target word is read or written in the two data sets. The sound pure target words from BasiScript are written 21168 times by 2352 different writers. 28% of these written words contain at least one error. The sound pure words from DART are read 321 times by 28 different readers. From these read words, 36% contains at least one reading error.

4.2.2. Frequently made errors

Table 4 and Table 5 present respectively all incorrectly spelled target PCUs and all incorrectly read target PCUs, together with measures computed from the multi-sequence alignments. The results in Table 4 and 5 are ordered in descending absolute incorrect count, and in the right column per row in descending percentage.

In Table 4, we can observe that the primary PCU *ei* that represents the phoneme /EI/ is the PCU that is most frequently written incorrectly, i.e. in more than 80% of the times that this phoneme had to be written. In almost half of the times (49.91%) it was misspelled as *ij* and in 24.0% as *e*. Four other PCUs with a relative error percentage higher than 10% are *eu* (substituted most often with *u*), *g* (substituted with *ch*), *ij* (substituted with *ei*) and *ch* (deleted). In addition, we see that the * appears relatively high in the left column of the table (representing an insertion of a PCU), since it occurred 439 times in the selected data. The PCU *e* is most often (48.3%) inserted, followed by the PCU *u*.

Table 5 presents all incorrectly read target PCUs. The target PCUs that are relatively most frequently read incorrectly are *v* (33.33% of 9 readings), *ui* (23.53% of 17 readings), *ng* (22.22% of 9 readings), *a* (18.06% of 72 readings) and *g* (15.79% of 9 readings). These PCUs have the highest relative incorrect percentage.

5. Discussion

In this paper, we have presented an extension of the automatic spelling error detection and classification algorithm (Harmsen et al., 2021a,b) that is capable of detecting reading errors in phonetic transcriptions of audio recordings, in such a way that they are comparable with spelling errors. For reading error detection, the inputs to the algorithm are the target spelling and a realized (incorrect) phonetic transcription. For spelling error detection the inputs are the target spelling and the realized (incorrect) spelling. The output of the joint algorithm consists both in the reading and writing condition of a PCU-segmentation of the target spelling. In addition, each target PCU from the PCU-segmentation is aligned with its target phoneme and has a marking indicating whether it was read or spelled correctly in the realized reading or spelling. In case of a reading error, the substituted or inserted phoneme is returned. In the case of a spelling error, the substituted or inserted PCU is returned.

To answer research question 1, the reading error detection algorithm was evaluated by comparing automatically detected reading errors with manually annotated reading errors in a selection of phonetic transcriptions of read words from the DART corpus. We found a high level of agreement between the automatic and the manual scores.

Next, we applied the reading and spelling error detection algorithm to analyze reading and spelling errors in a selection of sound pure words in two separate corpora, one containing read words by first graders and one containing written words by second graders. To answer research question 2a, we analyzed the realized writings. We found that the PCUs that are more often written incorrectly are *ei*, *eu*, *g*, *ij* and *ch*. We observed that *ij* and *ei* are often exchanged, most probably because they sound the same (as /EI/). In addition, these target PCUs were presented in the target words *steil* (steep) and *rijst* (rice), in which substitution of the *ei* or *ij* results in the words *stijl* (style) and *reist* ((he) travels), which are both existing Dutch words. Earlier studies have proven that these two aspects make spelling more difficult (Bosman and de Groot, 1996; van Assche et al., 2014), which explains the fact that children make this specific error in writing this word. The same two explanations seem to hold for the finding that *g* in the word *ligt* ((he) lays) was substituted by *ch* in almost one fourth of the cases it had to be written. *g* and *ch* correspond to the same phoneme (i.e. /x/) and *licht* (light) is also an existing word in Dutch.

The frequent misspelling of the PCU *eu* occurring in the words *monteur* (mechanic) and *keus* (choice) might have another explanation. The PCU *eu* (with primary phoneme /EU/) is a vowel

Target PCU	Target words	N	Incorrect Absolute (#)	Incorrect Relative (%)	Realized PCUs (%)
ei	steil	2342	1875	80.06	ij (49.91), e (24.0), *ei (19.94)*, ee (2.73), 11 others
eu	monteur, keus	4663	1401	30.05	eu (69.95), u (17.39), e (4.61) uu (3.0), ui (2.06), 17 others
g	ligt	2345	616	26.27	g (73.73), ch (24.48), 8 others
ij	rijst	2331	590	25.31	ij (74.69), ei (22.35), ie (1.29), ui (0.3), 18 others
*	9 different words	-	439	-	e (48.3), u (14.29), j (6.58), i (5.9), t (3.85), n (3.17), 19 others
t	rijst, steil, monteur, ligt	9362	314	3.35	t (96.65), * (1.52), d (1.04), dt (0.34) tt (0.12), h (0.1), 10 others
ch	schrik	2332	250	10.72	ch (89.28), * (7.59), g (2.02), 7 oth.
r	rijst, monteur, schrik	7007	225	3.21	r (96.79), * (2.88), l (0.1), 8 others
f	fee	2323	181	7.79	f (92.21), v (7.06), t (0.13), * (0.13), 9 others
s	schrik, steil, keus, huis, rijst	11640	144	1.24	s (98.76), * (0.69), z (0.44), 8 others
l	steil, lip, ligt	7029	81	1.15	l (98.85), * (0.53), j (0.17) b (0.11), s (0.1), 9 others
i	schrik, lip, ligt	7018	69	0.98	i (99.02), ie (0.3), * (0.2) ee (0.19), e (0.13), 5 others
ui	huis	2316	63	2.72	ui (97.28), i (2.07), 9 others
p	lip	2341	59	2.52	p (97.48), b (1.28), 7 others
ee	fee	2323	57	2.45	ee (97.55), e (1.68), 6 others
k	schrik, keus	4651	33	0.71	k (99.29), * (0.34), h (0.19), 6 oth.
o	monteur	2344	30	1.28	o (98.72), oo (0.51) * (0.3) a (0.26), 2 others
m	monteur	2344	8	0.34	m (99.66), 3 others
h	huis	2316	3	0.13	h (99.87), 2 others

Table 4: Results of the BasiScript spelling error analysis. For each target PCU that is written incorrectly at least one time, we present the measures described in Section 3.5. An asterisk in the first column represents an insertion of a PCU. An asterisk in the last column represents a deletion of the target PCU.

and is written using two graphemes: *e* and *u*. These graphemes figure in as many as nine other PCUs: *e*, *ee*, *ei*, *u*, *uu*, *ui*, *eu*, *ou*, *au* and *oe*, which might be confusing for initial writers. This explanation is supported by the fact that the PCUs that initial writers write instead of the *eu* are *u*, *e*, *uu* and *ui*, which all contain an *e* or *u*.

With respect to the analyzed readings (research question 2b), we observed that the PCUs *v*, *ui*, *ng*, *a* and *g* are relatively most frequently read incorrectly. For most of these cases, the absolute number of errors is rather small. The errors for *a* are probably caused by the following complex but systematic vowel distinction in Dutch. In Dutch there is a phonological distinction between tense ('long') and lax ('short') vowels, as in the case of *oo* and *o*. In Dutch, letter doubling is used to distinguish between tense and lax vowel phonemes in monosyllables. The lax vowel *o* is written with one letter *o*, but the tense vowel *oo* is written with a single letter in open syllables (*bo-men (*=trees)) and with double letters *oo* in closed syllables (*boom (= tree*)) . This distinction can of course be confusing for beginning learners.

To answer research question 2c, we investigated the overlap between spelling and reading errors. This overlap seems to be limited. We found that *ch* is often deleted, both in spelling and reading, and that a *g* is often substituted by *ch* and /g/ respectively. However, there seem to be no clear reasons that can explain these specific errors. The limited overlap between spelling and reading errors could be explained by the fact that Dutch has a higher feedforward consistency than feedback consistency (Bosman and Van Orden, 1997). However, since several variables were not controlled for (i.e., the target words that had to be read and spelled, the participants, and the size of the datasets), we are not able to make a strong claim on this aspect. However, the joint spelling and reading error analysis method we presented in this study enables further research in this direction.

Target PCU	# Diff. targets	N	Incorrect Absolute (#)	Incorrect Relative (%)	Realized phonemes (%)
*	28	-	208	-	s (12.02), t (10.58), x (7.69), @ (7.21) r (6.73), p (6.25), k (4.81), , 25 others
t	24	175	17	9.71	t (90.29), * (4.0), s (2.29), l (1.71), f (0.57) k (0.57), p (0.57)
r	18	139	14	10.07	r (89.93), * (5.76), l (2.16), x (0.72), d (0.72), w (0.72)
a	11	72	13	18.06	A (81.9), a (12.5), O (2.8), E (1.4), e (1.4)
l	12	80	10	12.5	l (87.5), r (3.75), * (2.5), p (1.25), d (1.25), k (1.25), m (1.25), h (1.25)
ch	13	86	8	9.3	x (90.7), * (4.65), k (1.16), r (1.16), N (1.16), h (1.16)
s	24	152	7	4.61	s (95.39), t (0.66), p (0.66), S (0.66), b (0.66), z (0.66), * (0.66), v (0.66)
k	13	86	6	6.98	k (93.0), * (3.5), r (1.2), p (1.2), t (1.2)
i	6	43	4	9.3	I (90.7), i (4.65), y (2.33), u (2.33)
ui	3	17	4	23.53	UI (76.47), U (11.76), EU (5.88), I (5.88)
o	6	34	4	11.76	O (88.24), o (5.88), A (2.94), UI (2.94)
g	3	19	3	15.79	x (84.21), g (10.53), S (5.26)
v	3	9	3	33.33	v (66.67), f (22.22), s (11.11)
m	8	63	3	4.76	m (95.24), b (1.59), h (1.59), p (1.59)
p	9	63	3	4.76	p (95.24), r (1.59), * (1.59), b (1.59)
ei	2	18	2	11.11	EI (88.89), UI (5.56), i (5.56)
ng	2	9	2	22.22	N (77.78), x (11.11), n (11.11)
f	6	44	2	4.55	f (95.45), * (4.55)
n	5	21	2	9.52	n (90.48), I (4.76), * (4.76)
ee	2	14	1	7.14	e (92.86), @ (7.14)
uu	3	16	1	6.25	y (93.75), U (6.25)
oo	3	16	1	6.25	o (93.75), AU (6.25)
ou	2	5	1	20.0	AU (80.0), UI (20.0)
e	4	25	1	4.0	E (96.0), @ (4.0)
b	4	24	1	4.17	b (95.83), d (4.17)

Table 5: Results of the DART reading error analysis. For each target PCU that is read incorrectly at least one time, we present the measures described in Section 3.5. An asterisk in the first column represents an insertion of a PCU. An asterisk in the last column represents a deletion of the target PCU.

6. Future Directions

In the introduction to this paper we made clear that this is only the beginning of a research endeavour that will certainly require more suitable data and optimized algorithms. The present study has indeed some limitations that were imposed by the complexity of the task and the scarcity of available data. For a first attempt, we decided to constrain ourselves to analyzing the so-called sound pure words. In a following step, we would like to extend this approach to other, more complex words, but then it is clear that we are likely to get an explosion of phoneme-grapheme mappings that will require a more complex classification scheme. This classification scheme should capture some characteristics that are specific for the speech of initial readers, like multiple attempts to read a word or insertion of vowels (Harmsen et al., 2023). In addition, to gain insight into specific individual difficulties, we would like to study reading and spelling data of one and the same child, preferably longitudinal data, in which children read and write the same words. So one important task for future research could be the collection of reading and spelling data of the same children. Another future direction is to use Automatic Speech Recognition (ASR) to automatically obtain phonetic transcriptions of child read speech. Currently, ASR models for automatic word correctness assessment are available for Dutch (e.g., Molenaar et al. (2023); Harmsen et al. (2023)), but a well performing and evaluated ASR model for phoneme recognition in child read speech has not yet been published. Such a model could be inspired by research by Gelin et al. (2023), who have recently published about developing such models for automatic phoneme recognition in French.

7. Ethical statement

The present research and its results may have a major societal impact as they contribute to significantly increasing the reliability and validity of reading and writing assessment and ultimately paving the way to improving and personalizing learning-to-read-and-write trajectories.

8. Acknowledgements

We would like to thank Stéphanie Kremer for making phonetic transcriptions of the child audio recordings. This publication is part of the ASTLA project with project number 406.20.TW.009, which is (partly) financed by the Dutch Research Council (NWO).

9. Bibliographical References

Y. Bai, F. Hubers, C. Cucchiarini, and H. Strik. 2021. An ASR-based reading tutor for practicing reading skills in the first grade: Improving performance through threshold adjustment. In *Proc. IberSPEECH 2021*, pages 11–15.

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.

S.R. Borgwaldt, F. Hellwig, and A.M.B. De Groot. 2004. Word-initial entropy in five languages: Letter to sound and sound to letter. *Written Language and Literacy*, 7:165–184.

A. M. T. Bosman and A. de Groot. 1996. Phonologic mediation is fundamental to reading: Evidence from beginning readers. *The Quarterly Journal of Experimental Psychology Section A*, 49(3):715–744.

A. M. T. Bosman and G. C. Van Orden. 1997. *Why Spelling is More Difficult than Reading*, pages 173–194. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

D. Chicco, M.J. Warrens, and G. Jurman. 2021. The Matthews Correlation Coefficient (MCC) is more informative than Cohen's Kappa and Brier Score in binary classification assessment. *IEEE Access*, 9:78368–78381.

L. Cleuren, J. Duchateau, P. Ghesquière, and H. Van Hamme. 2008. Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

C. Cucchiarini. 1993. *Phonetic transcription: A methodological and empirical study*. Ph.D. thesis, Nijmegen, The Netherlands.

C. Cucchiarini. 1996. Assessing transcription agreement: Methodological aspects. *Clinical Linguistics and Phonetics*, 10:131–155.

C. Cucchiarini, H. van Hamme, O. van Herwijnen, and F. Smits. 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)*, Genoa, Italy. European Language Resources Association (ELRA).

L. Gelin, M. Daniel, T. Pellegrini, and J. Pinquier. 2023. Comparing phoneme recognition systems on the detection and diagnosis of reading mistakes for young children's oral reading evaluation. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 6–10.

W. Harmsen, C. Cucchiarini, and H. Strik. 2021a. Automatic detection and annotation of spelling errors and orthographic properties in the Dutch BasiScript corpus. *Computational Linguistics in the Netherlands Journal*, 11:281–306.

W. Harmsen, C. Cucchiarini, and H. Strik. 2021b. Automatic quantitative analysis of spelling errors in texts written by sixth graders. In *EDULEARN21 Proceedings*, 13th International Conference on Education and New Learning Technologies, pages 8937–8945. IATED.

W. Harmsen, F. Hubers, R. van Hout, C. Cucchiarini, and H. Strik. 2023. Measuring word correctness in young initial readers: Comparing assessments from teachers, phoneticians, and ASR models. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 11–15.

J. Keuning and L. Verhoeven. 2008. Spelling development throughout the elementary grades: The Dutch case. *Learning and Individual Differences*, 18(4):459–470.

M.H.L. Kleijnen. 1997. *Strategieën van zwakke lezers en spellers in het voorgezet onderwijs*. Ph.D. thesis, Vrije Universiteit Amsterdam, Lisse.

R. Laarmann-Quante. 2016. Automating multilevel annotations of orthographic properties of German words and children's spelling errors. In *Proc. Language Teaching, Learning and Technology (LTLT 2016)*, pages 14–22.

K. Landerl and P. Reitsma. 2005. Phonological and morphological consistency in the acquisition of vowel duration spelling in Dutch and German. *Journal of Experimental Child Psychology*, 92(4):322–344.

S. Limonard, C. Cucchiarini, R.W.N.M. van Hout, and H. Strik. 2020. Analyzing read aloud speech by primary school pupils: Insights for research and development. In *Proc. Interspeech 2020*, pages 3710–3714.

B. Lin and L. Wang. 2022. Phoneme mispronunciation detection by jointly learning to align. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6822–6826.

B. Molenaar, C. Tejedor-Garcia, C. Cucchiarini, and H. Strik. 2023. Automatic assessment of oral reading accuracy for reading diagnostics. In *Proc. Interspeech 2023*, pages 5232–5236.

A. Nunn. 1998. Dutch orthography: A systematic investigation of the spelling of Dutch words. Holland Academic Graphics, Den Haag.

A. Tellings, N. Oostdijk, I. Monster, F. Grootjen, and A. van den Bosch. 2018a. BasiScript: : A corpus of contemporary Dutch texts written by primary school children. *International Journal of Corpus Linguistics*, 23(4):494–508.

A. Tellings, N. Oostdijk, I. Monster, F. Grootjen, and A. van den Bosch. 2018b. Spelling errors of 24 cohorts of children across primary school 2012-2015: A BasiScript corpus study. *Computational Linguistics in the Netherlands Journal*, 8:83–98.

E. van Assche, W. Duyck, and R.J. Hartsuiker. 2014. Phonological recoding in error detection: A cross-sectional study in beginning readers of Dutch. *PLOS ONE*, 8(12).

M. van Druenen, M. Gijsel, F. Scheltinga, and L. Verhoeven. 2019. *Leesproblemen en dyslexie in het basisonderwijs: Handreiking voor aankomende leerkrachten*, 3rd edition. Masterplan Dyslexie Expertisecentrum Nederlands, 's-Hertogenbosch.

Z. Zhang, Y. Wang, and J. Yang. 2021. Text-conditioned transformer for automatic pronunciation error detection. *Speech Communication*, 130:55–63.

10. Language Resource References

Bai et al. 2020. *Dutch Automatic Reading Tutor (DART) Corpus*. Radboud University.

Elffers et al. 2013. *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*. Radboud University. [link].

Gillis. 2001. *Protocol for Broad Phonetic Transcriptions*. Corpus Gesproken Nederlands (CGN) Project. [link].

Tellings et. al. 2015. *BasiScript-corpus*. Radboud University. Dutch Language Institute, 1.0. [link].

Ten Bosch. 2019. *Grapheme to Phoneme Converter*. Centre for Language and Speech Technology, 0.3.4. [link].