Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement

Leen Cleuren, Jacques Duchateau, Pol Ghesquière, Hugo Van hamme

Katholieke Universiteit Leuven
Belgium
Leen.Cleuren@ped.kuleuven.be, Jacques.Duchateau@esat.kuleuven.be

Abstract

Within the scope of the SPACE project, the CHildren's Oral REading Corpus (CHOREC) is developed. This database contains recorded, transcribed and annotated read speech (42 GB or 130 hours) of 400 Dutch speaking elementary school children with or without reading difficulties. Analyses of inter- and intra-annotator agreement are carried out in order to investigate the consistency with which reading errors are detected, orthographic and phonetic transcriptions are made, and reading errors and reading strategies are labeled. Percentage agreement scores and kappa values both show that agreement between annotations, and therefore the quality of the annotations, is high. Taken all double or triple annotations (for 10% resp. 30% of the corpus) together, % agreement varies between 86.4% and 98.6%, whereas kappa varies between 0.72 and 0.97 depending on the annotation tier that is being assessed. School type and reading type seem to account for systematic differences in % agreement, but these differences disappear when kappa values are calculated that correct for chance agreement. To conclude, an analysis of the annotation differences with respect to the '*s' label (i.e. a label that is used to annotate undistinguishable spelling behaviour), phoneme labels, reading strategy and error labels is given.

1. Introduction

In Flanders (Belgium), primary school children's school progress is regularly assessed in order to detect early learning difficulties such as reading disabilities. Unfortunately, the traditional paper-and-pencil reading assessment process is a very time-consuming one, taking away valuable time that could much better be invested in the actual guidance and treatment of those children that experience reading difficulties. Moreover, another disadvantage of this way of reading skill assessment is that its evaluation suffers from examiner bias. As a result, an automated, more reliable and objective reading assessment tool is in great demand.

A way to address these disadvantages is by implementing automatic speech recognition in the reading assessment process. Recently, different research projects have made a great effort in implementing speech technology in the assessment (and intervention) of reading difficulties, such as the LISTEN project (Carnegie Mellon University) (Mostow & Aist, 2001), the Foundations to Literacy Reading Tutor project (Colorado University) (Wise et al., 2005), and the TBALL project (UCLA) (Kazemzadeh et al., 2005). In Flanders, the SPACE project (SPeech Algorithms for Clinical and Educational applications) aims to automate the reading assessment process and to make it more objective. Additionally, the project wants to develop interactive, speech technology supported tools that enable a virtual reading tutor to act as a fluent reader model and to provide immediate feedback to the learning child. More information on the SPACE project can be found on its website: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

In spite of the availability of the text being read, the use of automatic speech recognition within the context of reading assessment and instruction is a very challenging task due to reading-related developmental processes: e.g. oral reading of novice readers or readers with reading difficulties can be fraught with oral reading errors. Within the scope of the SPACE project - in order to improve the speech recognizer's ability to accurately detect oral reading errors - we are developing CHOREC (CHildren's Oral REading Corpus), a Dutch database of recorded, transcribed and annotated children's oral readings (Cleuren et al., 2006). CHOREC provides us with sufficient data to statistically characterize children's reading behavior, and to supply the speech recognizer with a model that contains information on the nature and prevalence of likely oral reading errors (cf. Mostow et al., 2002). However, this database is not only of use for the development and testing of speech technology applications. It also offers the unique possibility to researchers in the field of learning disabilities to accurately characterize the phenotypic reading performance pattern of children with and without reading disabilities with respect to reading strategy use and reading error occurrence. As such, this contribution adds to the existing children's speech databases for both speech technology and educational research.

Before this corpus can be used for automated analysis, preferably all recorded reading sessions are manually segmented, transcribed and labeled. The quality of these human annotations relies heavily on various annotator characteristics, such as familiarity with the recorded material, amount of annotation training, motivation etc.; and on various external influences such as time pressure, changes in annotation protocol etc. (Bayerl & Paul, 2007). It is clear then that a quality estimation of these annotations, i.e. an analysis of annotator agreement and consistency in segmentation, transcription, and labeling, is recommended (Cucchiarini, 1996). To do so, different methods to assess the quality of corpus annotations have been proposed in the literature, such as pair wise transcriber percentage agreement and the kappa statistic (e.g. Dilley et al., 2006; Kazemzadeh et al., 2005; Pitt et al., 2005; Yoon et al., 2004).

In this paper, we present an analysis of inter- and intra-annotator agreement in the transcription and labeling of

recorded children's oral readings. More specifically, we investigate the consistency with which orthographic and phonetic transcriptions are made, reading errors are detected, and reading errors and reading strategies are labeled.

2. Method

2.1. Corpus Creation and Annotation

2.1.1. Participants

The speakers/readers in the CHOREC corpus are 400 Dutch speaking elementary school children (6-12 years old) enrolled in a Flemish regular elementary school (n = 274) or in a school for children with specific learning disabilities (n = 126). For all children, parental consent was obtained, as well as relevant information on sex, age, grade, school curriculum, place of birth, place of residence, mother tongue, reading level, reading method used to learn how to read, and presence of reading disabilities.

2.1.2. Reading Material

For every child, a newly developed computerized reading test battery was administered which contains a real word reading test (RWRT), a pseudoword reading test (PWRT), and a story reading test (SRT). Both the RWRT and the PWRT contain three lists of respectively 40 1-syllable, 40 2-syllable and 40 3- or 4-syllable real words or pseudowords. The SRT consists of 9 graded text stories (vocabulary of 538 distinct word forms), ranging from AVI 1 to AVI 9[1] in difficulty, and 103 to 223 words in length. Real words and pseudowords were presented individually; text stories were presented paragraph by paragraph on the computer screen. Each child read minimally one and maximally three real word lists and pseudoword lists, and minimally one and maximally four text stories depending on that particular child's reading level. Recordings were done in 3 sessions of maximally 20 minutes each.

2.1.3. Recording Material and Setup

Recordings were made in any room available at the participating school. People were restricted from entering the room during recordings, but environmental noise such as school bells, children entering the play ground etc. could not be avoided. For the making of the recordings, the ESAT Reading Tutor[2] was developed, a tool - installed on a laptop computer - that enabled us to record each list of words and each story separately. Children's speech was recorded at 22050 Hz by means of 2 microphones: a close-talking microphone and a desk microphone, both connected with the laptop through a preamplifier. In total, 42 GB or 130 hours of speech was recorded.

Children were instructed to try to read the words, pseudowords or paragraphs presented on the screen, as accurately and as fluently as possible. Most children were very motivated and eager to participate in the reading recordings. Some children however, often having severe reading difficulties, needed to be motivated by the promise that they could listen to their own recordings afterwards (cf. Kazemzadeh et al., 2005).

2.1.4. Annotation Procedure

The recordings were segmented, transcribed and labeled manually by means of a customized version of 'Praat' (http://www.Praat.org/). This tool provides the possibility to attach a time-aligned text-grid (containing different tiers) to the speech sound. As such, each tier provides another layer of descriptive information about the speech sound that particular tier is attached to. The 8 annotation tiers used in the CHOREC database include both information on utterances directly resulting from the child's efforts to read what is presented on the computer screen as well as information on background noise and reading task-related and task-unrelated unforeseen utterances made by both the child and the examiner.

Initially, when a recording was loaded in Praat, 2 tiers that contain the original reading task (with original resp. adjusted boundaries), were automatically loaded along with it. Then, two passes were made through each speech file during annotation. Through a first pass (resulting in a 'partial' (p) annotation file containing 6 tiers), the following 4 tiers were annotated and added: (1) orthographic transcription tier, (2) a tier for the broad-phonetic transcription of what is actually read, (3) a tier for the annotation of utterances made by the examiner, and (4) a tier for the annotation of background noise. Through a second pass, only for those words the child hesitates or makes a reading error, labels of oral reading strategies and errors were added to the p-annotation file (resulting in a 'full' (f) annotation file containing 8 tiers). During annotation, the annotator could rely on the audio signal and on visual speech wave representations of the recorded speech. For more information on the exact content of each tier, see Cleuren et al. (2006).

A detailed annotation protocol made sure that annotation was done uniformly by all annotators. The basis for this protocol was obtained by adopting the protocol used in the project CGN (Corpus Gesproken Nederlands, Spoken Dutch Corpus, see http://lands.let.kun.nl/cgn): e.g. use of Praat, protocol on orthographic transcriptions, and protocol on broad phonetic transcriptions (same phone set and symbols). However, at some points, deviations from those protocols were needed. For instance, for CHOREC, some CGN annotation layers do not apply (e.g. syntactic annotation, lexicon link-up), while other annotation layers needed to be added (e.g. text to be read, reading strategy annotation, reading error annotation). Furthermore, the orthographic protocol could be simplified, and it now addresses the topic of disfluencies: since phonetic transcriptions are always available in CHOREC (in CGN only for 10%), a detailed orthographic transcription at the grapheme level of what has been said (during disfluencies) was not needed; as such avoiding a non-unique mapping between the grapheme and phoneme string.

A total of 10 annotators (1 doctoral researcher and 9 students), all native speakers of Dutch and none of them having prior experience in annotating audio files, worked in the project. The doctoral researcher supervised the project and trained each of the other annotators by letting them in-

[1] In the Netherlands and Flanders, the AVI-index is used to distinguish between texts of different technical difficulty level. The index is largely based on a reading index which takes into account word, sentence and text features.

[2] Reading Tutor developed at the Department of Electrical Engineering (ESAT), K.U.Leuven, Belgium.

dependently work through the annotation protocol and by giving them personal feedback after manually correcting their first two days of annotation work. After that, constant supervision was given so that questions regarding the annotation protocol and particular annotation problems could immediately be answered. After the annotations were finished, these were formally checked automatically and manually corrected if necessary.

2.2. Analysis of Inter- and Intra-Annotator Agreement

2.2.1. Re-annotations

To be able to assess inter-annotator agreement with respect to the detection of reading errors, and orthographic and phonetic transcriptions, p-annotation files were independently provided by 3 different annotators (A1, A2, and A3) for 30% of the corpus. For 10% of the corpus, f-annotation files were annotated twice by one annotator with a period of at least one month (up to one year) between the two annotations. For 50% of the double f-annotations, the reading strategy and reading error tier were stripped from the f01-file which was then re-annotated for these tiers (resulting in an f01b-file). For the other 50% of double f-annotations, p02-files were used as the basis annotation to add the reading strategy and reading error tier to (resulting in an f02-file). The audio files chosen for triple p-annotations (p01, p02, p03) and double f-annotations (f01, f01b/f02) were selected in such a way that they form a representative sample of all reading tasks used in the making of CHOREC (RWRT, PWRT, SRT), and both school types (regular schools, special schools).The resulting database consists of individual p- and f-annotation files for each list of (pseudo)words or each story read by the child, and for each version of re-annotation.

2.2.2. Tiers under Analysis

For the analysis of inter-annotator agreement, we focus on the orthographic transcription tier and the phonetic transcription tier; for the analysis of intra-annotator agreement, the reading strategy tier and the reading error tier are the tiers of interest. Figure 1 illustrates these 4 tiers by giving an example of an annotated erroneously read sentence.

In the *orthography tier*, a '*' denotes an erroneous attempt to read the word whereas a correct attempt is orthographically written down. If the erroneous attempt results in another existing Dutch word (e.g. 'als' means 'if'), this word is added between brackets. A '*s' visualizes the fact that the annotator hears 'something' but can not write it out because she is not able to distinguish the different reading attempts and to register how these attempts were exactly pronounced. Whenever the last attempt for a word is a '*' or '*s', this word is assessed as being read incorrectly (even if the child first read the word correctly but made an error or hesitated after that); when the last attempt is correctly read (and therefore is fully written down orthographically), this word is assessed as being read correctly (even if the child made one or more preceding errors or hesitations before this last correct attempt).

Second, the *phonetics tier* provides us with a broad phonetic transcription of what was actually read by the child. Before any manual annotation starts, this tier automatically provides us with a concatenation of all correct phonetic transcriptions for a particular word. If all words in a segment are directly read correctly, these canonical transcriptions remain; else, all words in that segment are transcribed phonetically. '*s' is just copied from the orthography tier.

Third, the strategy used by the child to read a particular word is annotated in the *reading strategy tier* (e.g. 'f' = wrong within the first trial, '*sg' = correct after some nearly inaudible reading attempts, 'g' = correct within the first trial, 'agg' = repetition of an initial part of the word; 'O' = omission of a word). Fourth, for those words for which the last attempt is erroneous, the (combination of) reading errors made are labeled in the *reading error tier*. An error classification system containing 40 reading error categories (each category represented by a letter or a number) lies at the basis of this tier. In case of a correct last attempt, a '0' label is attributed to the word in order to make the total amount of entities in this tier correspond to the amount of words the child had to read.

Thus, for each (erroneous or correct) attempt of the child to read a particular word, a different string (made of letters and/or tags) is annotated in the orthography and phonetics tier, such that the total amount of entities (strings) in the two tiers correspond. In the reading strategy tier, one string of letters represents all reading attempts for a particular word; whereas in the reading error tier, a string of numbers (joined by a separator symbol) represents the errors made in the last reading attempt of a word. Therefore, the total amount of entities in these two tiers corresponds to the amount of words the child had to read.

2.2.3. Agreement Metrics

The two most popular measures to express annotator agreement are used in the present study: (1) percentage agreement, and (2) the kappa statistic.

Percentage Agreement. Pair wise inter-annotator agreement scores (expressed as % agreement) were calculated for each of the three pairs (A1-A2, A2-A3, A1-A3) of annotators; and this with respect to both orthographic transcription and phonetic transcription agreement. Intra-annotator agreement scores (expressed as % agreement) are calculated with respect to reading error and strategy labeling agreement. Here, the pair wise transcriber agreement results represent only a single comparison pair (Yoon et al., 2004). Percentage agreement is obtained by dividing the total number of pair wise agreements between annotators for each word, by the total number of possible pair wise agreements (sum of total number of disagreements and total number of agreements), and then multiplying this by 100% (Cucchiarini, 1996).

Cohen's Kappa. When one wants to correct for chance agreement, the unweighted kappa statistic (which varies between 0 and 1) is commonly used to evaluate annotator agreement at the label level (for the exact formula see Cohen, 1960). A kappa statistic of 0.6 or higher indicates a substantial agreement; a kappa of 0.8 or higher indicates an almost perfect agreement (Landis & Koch, 1977).

For each percentage agreement score and kappa value,

Expected	Els zoekt haar schoen onder het bed. [Els looks for her shoe under the bed.]						
Observed	Als *(says 'something')* zoekt haar sch…schoen onder bed. [Als *(says 'something')* looks for her sh…shoe under bed.]						
	Els	zoekt	haar	schoen	onder	het	bed.
Orthography	*(als)	*s zoekt	haar	* schoen	onder		bed
Phonetics	Als	*s zukt	har	sx sxun	Ond@r		bEt
Strategy	f	*sg	g	agg	g	O	g
Error	e/4						

Figure 1: Example of an annotated sentence.

a 95% confidence interval (CI) is calculated in order te analyze whether differences in agreement are statistically significant.

3. Results

3.1. Description of Agreement between Annotators

Table 1 gives an overview of all % agreement scores and kappa values that were calculated to assess annotator agreement for the different annotation levels.

3.1.1. Detection of Reading Errors

The first analyses concerned the annotation of a word assessed as being read correctly (last attempt is the orthographic transcription of the word under assessment) versus a word assessed as being read incorrectly (last attempt is a '*' or a '*s'). Two annotators were said to agree if they both annotated the last attempt as a '*' or '*s', or both annotated the last attempt orthographically. For these analyses, added real words between brackets (cf. 2.2.2.) were not taken into account.

The overall inter-annotator agreement for the detection of reading errors was 95.96% with a 95% confidence interval (CI) of 95.87-96.04 (κ = 0.796; 95% CI 0.792-0.800). When looking at agreement between annotators for the 2 different school types, the % agreement score for regular schools was higher than for special schools: 96.32% (96.22-96.42) versus 95.21% (95.04-95.37); whereas the kappa value for regular schools was lower than for special schools: 0.779 (0.773-0.785) versus 0.816 (0.809-0.822) (both differences were statistically significant). When comparing agreement scores for the different reading tasks, there were clear and statistically significant differences with respect to the % agreement scores: 95.20% (RWRT; 95.00-95.39) versus 90.59% (PWRT; 90.32-90.86) versus 98.37% (SRT; 98.30-98.44); as well as significant differences with respect to the corresponding kappa values: κ = 0.735 (RWRT; 0.725-0.746) versus κ = 0.776 (PWRT; 0.770-0.783) versus κ = 0.794 (SRT; 0.785-0.804).

3.1.2. Orthographic Transcriptions

The second group of analyses concerned the exact orthographic transcription. Two annotators were said to agree if both annotated a word exactly alike, i.e. with the same number of '*', and the same number of orthographic transcriptions of the word under assessment; and all that in the same order. For these analyses, '*s' was not taken into account but words between brackets were. Kappa values at the orthographic label level were not calculated because of the fact that there were exactly 2716 different labels in the orthography tier. According to Cicchetti (1972), we would need a total of 2 x 2716^2 observations to obtain meaningful kappa values; a condition that was not fulfilled at all (number of observations = 259804).

The overall inter-annotator agreement for the making of orthographic transcriptions was 90.79% (95% CI 90.66-90.91). Transcriptions for regular schools showed significantly better agreement between annotators then for special schools: 92.13% (91.99-92.27) versus 87.93% (87.68-88.18). The agreement score for the annotation of SRTs (95.56%; 95.44-95.68) was significantly higher than for RWRTs (88.92%; 88.63-89.21), which was again significantly higher than for PWRTs (80.50%; 80.13-80.86).

3.1.3. Phonetic Transcriptions

A third cluster of analyses was carried out to assess agreement between annotators with respect to the broad phonetic transcription of the word under assessment. Two annotators were said to agree if their phonetic transcription of a particular word were exactly alike. For these analyses, only those words were assessed for which the canonical phonetic transcription, often containing several allowed pronunciations for that word, was reduced to only one transcription option (otherwise, it was not known which transcription was the 'chosen' one). % agreement scores were calculated at the word level, whereas kappa values were calculated at the phoneme label level.

The overall inter-annotator agreement for the making of phonetic transcriptions was 86.37% (95% CI 86.20-86.54) with a corresponding kappa value of 0.930 (95% CI 0.929-0.930). Audio files from regular schools were annotated with significantly higher agreement than files coming from special schools: percentages agreement were 88.51% (88.31-88.71) versus 82.18% (81.85-82.51); the corresponding kappa values were 0.937 (0.936-0.937) versus 0.917 (0.915-0.918). The agreement percentage for phonetic transcriptions of SRTs (94.34%; 94.19-94.48) was much higher than those for RWRTs (78.87%; 78.35-79.38) and PWRTs (68.45%; 67.94-68.95); a tendency that was also shown by the kappa values for the different reading tasks (κ = 0.964; 0.963-0.965 vs. 0.907; 0.906-0.909 vs.

		All Data	School Type		Task Type		
			Regular Schools	*Special Schools*	*RWRT*	*PWRT*	*SRT*
Reading Error Detection							
	(1)	95.96%	96.32%	95.21%	95.20%	90.59%	98.37%
	(2)	0.796	0.779	0.816	0.735	0.776	0.794
Orthographic Transcriptions							
	(1)	90.79%	92.13%	87.93%	88.92%	80.5%	95.56%
Phonetic Transcriptions							
	(1)	86.37%	88.51%	82.18%	78.87%	68.45%	94.34%
	(2)	0.930	0.937	0.917	0.907	0.888	0.964
Reading Strategy Labeling							
f01-f01b	(1)	98.64%	98.72%	98.45%	98.35%	96.75%	99.26%
	(2)	0.966	0.961	0.971	0.960	0.966	0.956
f01-f02	(1)	91.50%	93.09%	88.38%	91.25%	77.79%	95.96%
	(2)*	0.779	0.802	0.744	0.774	0.733	0.711
Reading Error Labeling							
f01-f01b	(1)	97.77%	98.01%	97.22%	97.55%	92.55%	99.32%
	(2)	0.911	0.899	0.921	0.896	0.575	0.933
f01-f02	(1)	94.14%	95.39%	91.71%	94.57%	80.88%	98.24%
	(2)	0.717	0.722	0.706	0.709	0.660	0.848

* For the calculation of these kappa values, strings containing a '*s' strategy label were excluded.

Table 1: % agreement scores (1) and kappa values (2) for the different analyses.

0.888; 0.886-0.889). All differences were statistically significant.

3.1.4. Reading Strategy Labeling

The fourth series of analyses concerned the intra-annotator agreement with respect to the labeling of reading strategies. Two annotations were said to agree if exactly the same labels in the same order were attributed to the word under assessment. Results for the f01-f01b and f01-f02 comparison are discussed separately. % agreement scores were calculated at the word level, whereas kappa values were calculated at the reading strategy label level.

f01-f01b comparison. The overall intra-annotator agreement for the labeling of reading strategies was 98.64% (95% CI 98.41-98.83) whereas the kappa value was 0.966 (95% CI 0.961-0.970). When looking at agreement between annotations for the 2 different school types, % agreement scores were almost equal: 98.72% (regular schools; 98.45-98.94) versus 98.45% (special schools; 97.99-98.81); the kappa values were 0.961 (regular schools; 0.955-0.968) versus 0.971 (special schools; 0.964-0.978). However, the % agreement score for the annotation of SRTs (99.26%; 99.04-99.43) was significantly higher than for RWRTs (98.35%; 97.71-98.81), which was again significantly higher than for PWRTs (96.75%; 95.91-97.42). On the contrary, kappa values for the three different reading tasks did not show this tendency: κ = 0.956 (0.945-0.966) versus κ = 0.960 (0.949-0.971) versus κ = 0.966 (0.958-0.973) (the differences were not statistically significant).

f01-f02 comparison. The overall intra-annotator agreement for the labeling of reading strategies was 91.50% with a 95% confidence interval of 90.95-92.02 (κ = 0.779; 0.767-0.791). Reading strategy labeling for regular schools showed significantly better agreement between annotations then for special schools: 93.09% (92.47-93.66) versus 88.38% (87.28-89.39), κ = 0.802 (0.786-0.817) versus κ = 0.744 (0.725-0.763). The agreement score for the annotation of SRTs (95.96%; 95.45-96.42) was significantly higher than for RWRTs (91.25%; 89.96-92.39), which was again significantly higher than for PWRTs (77.79%; 75.93-79.54). However, when comparing the corresponding kappa values, it seems that agreement is lower for SRTs (κ = 0.711; 0.686-0.736) than for RWRTs (κ = 0.774; 0.748-0.800), and that there are no significant differences in agreement between annotations for SRTs, RWRTs and PWRTs (κ = 0.733; 0.713-0.752). It must be noted here that, because of automatic alignment difficulties, annotation pairs containing a '*s' strategy label were not used in the kappa value calculations. In these cases, '*s' could be substituted for an enormous amount of different strategy label combinations, as such making it impossible to obtain interpretable aligned strings.

3.1.5. Reading Error Labeling

The last series of analyses were carried out to assess the intra-annotator agreement with respect to the labeling of reading errors. Two annotations were said to agree if exactly the same labels (not necessarily in the same order) were attributed to the word under assessment. Again, results for the f01-f01b and f01-f02 comparison are discussed separately. % agreement scores were calculated at the word level, whereas kappa values were calculated at the reading error label level.

f01-f01b comparison. The overall intra-annotator agreement for the labeling of reading errors was 97.77% (95% CI 97.49-98.02); the corresponding kappa was 0.911 (95% CI 0.903-0.919). Audio files from regular schools were annotated with comparable agreement as files coming from special schools: 98.01% (97.68-98.29) versus 97.22% (96.63-97.71), κ = 0.899 (0.887-0.911) versus κ = 0.921 (0.909-

0.933). The % agreement for phonetic transcriptions of SRTs (99.32%; 99.11-99.48) was significantly higher than that for RWRTs (97.55%; 96.80-98.13) and PWRTs (92.55%; 91.35-93.59); the corresponding kappa's were 0.933 (SRT; 0.919-0.948) versus 0.896 (RWRT; 0.875-0.917) versus 0.575 (PWRT; 0.557-0.593).

f01-f02 comparison. The overall intra-annotator agreement for the labeling of reading errors was 94.14% with a 95% CI of 93.67-94.57 (κ = 0.717; 95% CI 0.702-0.731). Reading error labeling for regular schools showed significantly better agreement between annotations then for special schools with resepect to % agreement: 95.39% (94.87-95.86) versus 91.71% (90.76-92.57). However, kappa values did not differ significantly: κ = 0.722 (0.701-0.742) versus κ = 0.706 (0.686-0.726). The % agreement score for the annotation of SRTs (98.24%; 97.89-98.53) was significantly higher than for RWRTs (94.57%; 93.51-95.46), which was again significantly higher than for PWRTs (80.88%; 79.12-82.53). The kappa value for SRTs (κ = 0.848; 0.791-0.906) seems to be significantly higher than the value for RWRTs (κ = 0.709; 0.673-0.744) and PWRTs (κ = 0.660; 0.639-0.681). However, there are no significant differences between the kappa values for RWRTs and PWRTs.

3.2. Interpretation of Differences in Agreement

3.2.1. All Data

Taking all double and triple annotations together, kappa values range from 0.717 to 0.966 and % agreement scores range from 86.37 to 98.64%. This indicates that agreement between annotators, and therefore the quality of annotations, is high.

More specifically, when comparing the different annotation tiers, it can be seen that the inter-annotator kappa value becomes larger when the annotation is more detailed: kappa values are always smaller for the detection of reading errors than for phonetic transcriptions. A comparison of the intra-annotator kappa values shows us that, in case of both the f01-f01b and f01-f02 comparison, the labeling of reading strategies is done with higher agreement than the labeling of reading errors; except for the f01-f02 kappa value obtained for the SRTs where an unexpected inverse relation was found. When we oppose the kappa values for the f01-f01b and f01-f02 comparison for both reading strategy and reading error labeling, we see that agreement is almost always higher for the f01-f01b comparison. This is easily explained by the influence of differences in annotations at the phonetic transcription level for the f01-f02 comparison. However, we did not find an explanation yet for the inverse relationship in kappa values with respect to reading error labeling in PWRTs.

Percentage agreement diminishes when the annotation becomes more detailed: scores are always higher for the detection of reading errors than for orthographic transcriptions, and % agreement scores for the latter are always higher than for phonetic transcriptions. This is exactly the opposite as seen for the kappa values, and can be explained by the fact that, when calculating % agreement scores, there is no correction for agreement by chance. A comparison of the intra-annotator % agreement scores shows us that, in case of the f01-f01b comparison, the labeling of reading strategies is done with higher agreement than the labeling of reading errors; however, in case of the f01-f02 comparison, the picture is exactly the opposite. When we oppose the % agreement scores for the f01-f01b and f01-f02 comparison for both reading strategy and reading error labeling, we see again that agreement is always higher for the f01-f01b comparison. This is again explained by the influence of differences in annotations at the phonetic transcription level for the f01-f02 comparison.

3.2.2. Influence of School Type

When comparing kappa's for regular schools versus special schools, values range from 0.706 to 0.971; % agreement scores range from 82.18 to 98.72%. This indicates that, for both school types seperately, agreement is high.

Further, we did not detect systematic significant differences in inter- and intra-annotator kappa values: the picture was different for the different annotation tiers. More precisely, annotations for regular schools agreed worse than for special schools with respect to the detection of reading errors, but better than for special schools with respect to the phonetic transcriptions. With respect to reading strategy and reading error labeling, there were no significant differences between the two school types. Although we did find a significant difference for the reading strategy f01-f02 comparison, this difference could be explained by the influence of differences in annotations at the phonetic transcription level.

With respect to the % agreement scores, it misleadingly seems that the amount of inter- and intra-annotator agreement depends on school type: annotations for regular schools are always better than those for special schools. But in fact, these differences go hand in hand with differences in the amount of errors made by these 2 groups. Children enrolled in a special school (19.3% errors) make more reading errors than those enrolled in a regular school (12.2% errors) (which is not counterintuitive). The more errors are made, the more opportunities there are for annotators to possibly disagree.

3.2.3. Influence of Reading Task Type

Kappa values for the different reading tasks range from 0.575 to 0.966, whereas % agreement scores range from 68.45 to 99.32%. This indicates that, for the different reading tasks seperately, agreement between annotators is substantial to almost perfect.

Again, with respect to the kappa values, the picture is not the same for all annotation tiers. Regarding the detection of reading errors, agreement for SRTs is higher than for PWRTs, and agreement for PWRTs is higher than for RWRTs. With respect to the phonetic transcriptions, the agreement for SRTs is again the highest, but now agreement for PWRTs is higher than for RWRTs. Reading strategy labeling agreement is the same for all three reading tasks; there only seems to be a significant difference in the f01-f02 comparison that shows better agreement for RWRTs than for SRTs. However, reading error labeling agreement seems to be different for the different reading tasks: SRTs are annotated with higher agreement than RWRTs, and annotations for RWRTs agree better than for PWRTs (except

in the f01-f02 comparison, there agreement is the same for RWRTs and PWRTs).

% agreement scores for SRTs are always better than those for RWRTs which are again better than those for PWRTs. Again, these differences can be explained by the differences in the amount of errors made by these 3 groups. For all children it seems that most errors are made in the pseudoword reading test (37.4% errors), and the least in the story reading test (5.5% errors); in the real word reading test children made 14.3% errors. This, again, comes up to our expectations. Pseudoword reading causes more reading errors because for those words, children can not use a direct or lexical route and need to rely more on phonological processes. Story reading causes less reading errors because in that case, extra contextual information defines syntactical and semantical restrictions for a particular word. Again, the more errors are made, the more possible disagreements.

3.3. Confusions

3.3.1. Phoneme Labeling

Since we were interested in which phonemes were more likely to get confused by the different annotators, we set up a confusion matrix and calculated agreement scores (taking all data together) for the different phonemes used for the making of the phonetic transcriptions. Figure 2 indicates for which phonemes the percentage agreement score was less than 90% (agreement scores for all other phonemes ranged from 90.3% to 96.2%). For these phonemes, the matrix reveals that the most common confusions happened between /i/ and /I/, /o/ and /O/, /@/ and /E/, /y/ and /Y/, /2/ and /y/, /x/ and /G/, /n/ and /N/, /S/ and /s/, /g/ and /G/, and /z/ and /Z/. These confusions are largely attributable to the fact that the different annotators came from different Flemish regions. Because of this, their dialects differ and so do their judgments of vowel length and type (explaining confusions between the long and short 'i', 'o', 'u', and between /2/ and /y/), and voicing (explaining confusions between /x/ and /G/). Confusions between /@/ and /E/, and /n/ and /N/ were not unexpected because of the fact that those phonemes show large acoustical resemblance. The /g/-/G/ and /z/-/Z/ confusions were probably just the result of erroneous typing (and do not occur very frequently); whereas the /S/-/s/ confusion is caused by the alignment program that aligns these two phonemes in words like e.g. /mE+[sj/S]@/ (where there is the need to choose between /sj/ and /S/). Additionnally, vowels seem to be deleted/inserted often; this is especially seen for the /@/, the /I/, the /O/, the /E/, the /i/, the /O/ and the /Y/.

3.3.2. Reading Strategy Labeling

With respect to reading strategy labeling, we were interested in which reading strategy labels were more likely to get confused; so we set up a confusion matrix for the f01-f01b comparison. Figure 3 pictures the % agreement scores for the different strategy labels. Agreement ranged from 32.4% up to 99.8%. Most confusions happened between 'lsf' (i.e. wrong letter spelling) and 'af' (i.e. wrong start of a word), 'ssf' (i.e. wrong syllable spelling) and 'alf' (i.e. wrong end of a word), 'af' and 'f' (i.e. wrong decoding within one trial), 'ssg' (i.e. correct syllable spelling) and

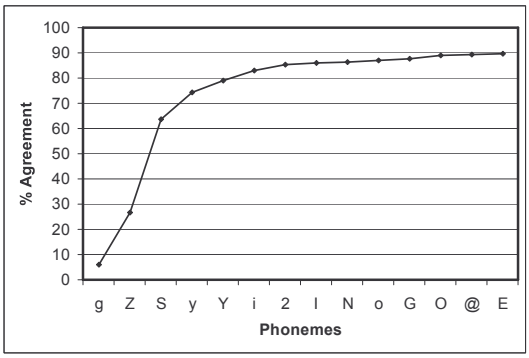

Figure 2: Percentage agreement per phoneme.

'ssf', and 'lsg' (i.e. correct letter spelling) and 'lsf'. The low score for '(W)' (i.e. change of word order) is due to its low occurrence; the % agreement score for 'O' (i.e. word omission) did not reach the 100% because of deletions/insertions of that label. Other labels had % agreement scores ranging from 92.9 to 99.8%. For more details about the meaning of these labels, see Cleuren et al. (2006).

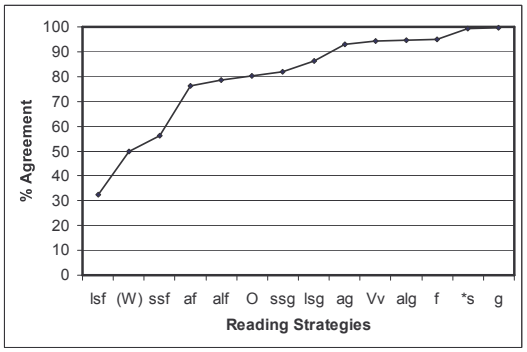

Figure 3: Percentage agreement per strategy label.

3.3.3. Did you hear that? The case of '*s'

As described before, a '*s' in the orthographic transcription tier visualizes the fact that the annotator hears 'something' but can not write it out because she is not able to distinguish the different reading attempts (which would otherwise be annotated as a '*' for each attempt), and to register how these attempts were exactly pronounced. Of course, the better an annotator listens to the child's reading, the less '*s' is used and the more difficult the annotation task becomes. As such it is clear that annotator characteristics and external influences come into play here.

A '*s' in the orthography tier is simply copied to the reading strategy tier; whereas every '*' receives a separate reading strategy label in that tier. Therefore, to have a better picture of the importance of these '*s' disagreements, a confusion matrix for the reading strategy f01-f02 comparison was set up. This revealed that 15.3% of all disagreements were caused by a '*s' disagreement, i.e. disagreements where annotator 1 was able to annotate what was said by the child, while the other annotator decided she was not able to do that and put a '*s' instead.

3.3.4. Reading Error Labeling

Again, a confusion matrix was set up to investigate % agreement scores for each reading error label. For the 10 most frequently annotated reading error labels (which account for 92.3% of all reading error label annotations), agreement ranged from 44.1 up to 99.9%; Figure 4 pictures the % agreement scores for these 10 reading error labels (for more details about the error labels, see Cleuren et al., 2006). No systematic confusions were detected; there only seemed to be consistent disagreement with respect to the annotation of short-long vowel substitutions.

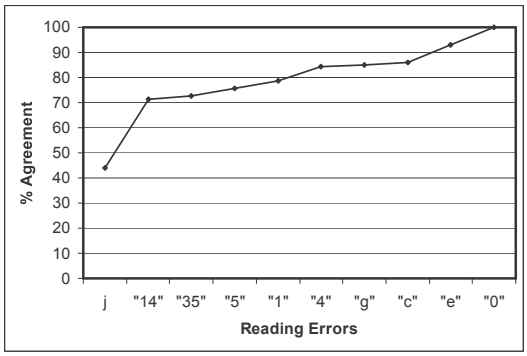

Figure 4: Percentage agreement per error label.

4. Conclusion

For CHOREC, a Dutch database of children's reading, we have analyzed inter- and intra-annotator agreement with respect to the detection of reading errors, orthographic and phonetic transcriptions, and reading strategy and reading error labeling. Percentage agreement scores and kappa values both show that agreement between annotations, and therefore the quality of the annotations, is high. Taken all double or triple annotations together, % agreement varies between 86.4% and 98.6%, whereas kappa varies between 0.72 and 0.97 depending on the annotation tier that is being assessed. School and reading type seem to account for systematic differences in % agreement, but these differences disappear when kappa values are calculated that correct for chance agreement.

When having a closer look at which phonemes were more likely to get confused during phonetic transcriptions, we saw that, for the phonemes with the lowest inter-annotator agreement scores, confusions could be explained by acoustical resemblance between phonemes and dialect differences between annotators. Additionally, vowels seem to be easily inserted or deleted. With respect to reading strategy and reading error label disagreement, systematic confusions were detected for the former but not for the latter. Furthermore, '*s' disagreements seemed to account for 15.3% of all disagreements in the f01-f02 strategy labeling comparison.

5. Acknowledgment

The research in this paper was supported by the IWT project SPACE (sbo/040102): SPeech Algorithms for Clinical and Educational applications, home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

References

Bayerl, P. S., & Paul, K. I. (2007). Identifying sources of disagreement: Generalizability theory in manual annotation studies. *Computational Linguistics*, *33*(1), 3–8.

Cicchetti, D. V. (1972). A new measure of agreement between rank ordered variables. In *Proceedings of the 80th Annual Convention of the American Psychological Association* (pp. 17–18).

Cleuren, L., Duchateau, J., Sips, A., Ghesquière, P., & Van hamme, H. (2006). Developing an automatic assessment tool for children's oral reading. In *Proceedings of Interspeech* (pp. 817–820). Pittsburgh, USA.

Cohen, J. (1960). A cofficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clinical Linguistics and Phonetics*, *10*(2), 131–155.

Dilley, L., Breen, M., Bolivar, M., Kraemer, J., & Gibson, E. (2006). A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). In *Proceedings of Interspeech* (pp. 317–320). Pittsburgh, USA.

Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., et al. (2005). TBALL data collection: The making of a young children's speech corpus. In *Proceedings of Interspeech* (pp. 1581–1584). Lisbon, Portugal.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of project LISTEN. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 169–234).

Mostow, J., Beck, J., Winter, S., Wang, S., & Tobin, B. (2002). Predicting oral reading miscues. In *Proceedings of the Seventh International Conference on Spoken Language Processing* (pp. 1221–1224). Denver, CO.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*, 89–95.

Wise, B., Cole, R., van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., et al. (2005). Learning to read with a virtual tutor: Foundations to literacy. In C. Kinzer & L. Verhoeven (Eds.), *Interactive literacy education: Facilitating literacy environments through technology.*

Yoon, T.-J., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of Interspeech* (pp. 2729–2732). Jeju Island, Korea.