International Journal of Language & Communication Disorders

INT J LANG COMMUN DISORD, SEPTEMBER–OCTOBER 2017,
VOL. 52, NO. 5, 595–611

Research Report

Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds

Peter Howell, Kevin Tang, Outi Tuomainen, Sin Kan Chan, Kirsten Beltran, Avin Mirawdeli and John Harris

Division of Psychology and Language Sciences, University College London, London, UK

(*Received April 2016; accepted October 2016*)

Abstract

Background: Stuttering and word-finding difficulty (WFD) are two types of communication difficulty that occur frequently in children who learn English as an additional language (EAL), as well as those who only speak English. The two disorders require different, specific forms of intervention. Prior research has described the symptoms of each type of difficulty. This paper describes the development of a non-word repetition test (UNWR), applicable across languages, that was validated by comparing groups of children identified by their speech and language symptoms as having either stuttering or WFD.
Aims: To evaluate whether non-word repetition scores using the UNWR test distinguished between children who stutter and those who have a WFD, irrespective of the children's first language.
Methods & Procedures: UNWR was administered to ninety-six 4–5-year-old children attending UK schools (20.83% of whom had EAL). The children's speech samples in English were assessed for symptoms of stuttering and WFD. UNWR scores were calculated.
Outcomes & Results: Regression models were fitted to establish whether language group (English only/EAL) and symptoms of (1) stuttering and (2) WFD predicted UNWR scores. Stuttering symptoms predicted UNWR, whereas WFD did not. These two findings suggest that UNWR scores dissociate stuttering from WFD. There were no differences between monolingual English-speakers and children who had EAL.
Conclusions & Implications: UNWR scores distinguish between stuttering and WFD irrespective of language(s) spoken, allowing future evaluation of a range of languages in clinics or schools.

What this paper adds

What is already known on the subject?

Children in UK schools can have either expressive speech problems or WFD. These can be distinguished based on analyses of connected speech to identify distinct type of speech symptoms. Many children in schools (in the UK and other countries) do not speak the native language of their home country when they enter school. It is widely recognized that they need to be treated equitably. Non-word repetition tests differentiate children with various cognitive difficulties, including fluency difficulties, from those who have no such difficulties. However, these tests cannot be used except for the languages they were designed for.

What this paper adds to existing knowledge

We now know that the speech-based procedure identifies fluency difficulty when children are tested in English even when English is not their native language. Scores on a Universal non-word repetition test (UNWR) provide a sensitive marker of fluency difficulty. UNWR can be used as an alternative to the speech-based procedure for assessing children for speech versus WFD in schools. UNWR is the first test designed to compare performance across children who speak a range of different native languages. UNWR is suitable for use in clinics as well as schools, particularly when EAL speakers are the clients.

Address correspondence to: Peter Howell, Division of Psychology and Language Sciences, University College London, Gower Street, London WC1E 6BT, UK; e-mail: p.howell@ucl.ac.uk

International Journal of Language & Communication Disorders
ISSN 1368-2822 print/ISSN 1460-6984 online
DOI: 10.1111/1460-6984.12305

What are the potential or actual clinical implications of this work?
Children can have expressive speech problems when they enter school. Schools are well-placed to identify these problems. Mirawdeli (2016) showed that schools want something that allows them to identify problems in a systematic way and that does so fairly with all children (e.g., children with EAL). We consider that the procedures we have described (speech and NWR) allow schools to achieve this. The next stop will be to explore the roles clinicians can take in schools to offer improved diagnostic and intervention procedures with these children.

Introduction

Not all disruptions to speech fluency are an indication of stuttering. For instance, children who are not fluent in the test language can experience word-finding difficulty (WFD) which can lead to repetition of whole monosyllable words (WWR) whilst a child tries to retrieve a word (Clark and Clark 1977, Fathman 1980, MacWhinney and Osser 1977). Although WFD impedes the forward flow of speech, it is a vocabulary, not a fluency problem. Therefore, it is desirable to have test formats that can separate fluency difficulty and WFD.

WFD can happen in monolingual children, but it is not exclusive to them as children with English as an additional language (CwEAL) often experience WFD as well. In UK schools there are more than 1 million CwEAL, with in excess of 600 alternative languages spoken. Also, the numbers of infant school CwEAL in England is growing with a reported increase between 1997 and 2013 from 7.8% to 18.1% according to the National Association for Language Development in the Curriculum (NALDIC) (2013).

Children with fluency difficulties show NWR deficits (Anderson and Wagovich 2010, Bakhtiar *et al*. 2007, Hakim and Ratner 2004) whilst there is no evidence that children with WFD show such deficits. This observation suggests that NWR tests could dissociate children with fluency difficulty from those with WFD in samples of children with diverse language backgrounds. However, to date, it has not been possible to test the predictions that NWR should be affected when there is fluency difficulty, but not when there is WFD in diverse language-background samples because NWR tests favour the language for which they were developed (Masoura and Gathercole 1999, Windsor *et al*. 2010). Consequently, if an English NWR test was used in a country where English is the native language, CwEAL would show deficits that could be mistakenly interpreted as indications of fluency difficulty.

A new NWR test (the 'Universal' NWR (UNWR) test) was designed for 20 languages (including English) commonly spoken in UK schools. To date, UNWR has been evaluated for the 20 languages by consulting native users of each language, and has been comprehensively evaluated using computational techniques for the five languages used in the current study.[1] Based on past literature, it was predicted that children with fluency difficulty (indicated by high rates of word fragmentation) would have problems with NWR whatever language they spoke. It was also predicted that NWR performance should not be affected when a child has a high rate of WWR alone because this pattern indicates WFD. This prediction applies both to children with English Only (CwEO) and CwEAL. The predictions about NWR were tested on a sample of children with diverse language backgrounds whose speech had been assessed for fluency difficulty and WFD.

Next, literature is summarized that identified the symptoms to use as indications of fluency difficulty and work that shows that high rates of WWR alone are indications of WFD rather than fluency difficulty. Then work is reviewed that reports that children with fluency difficulty show poor NWR performance but CwEAL do not. Finally, features taken into consideration in the design of UNWR are presented.

Symptoms indicative of fluency difficulty and the role of WWR

According to Brocklehurst (2013) and Wingate (2001), WWR are not symptoms that indicate fluency difficulty. Support for this observation is that the most widely used instrument for assessing fluency, Riley's (2009) Stuttering Severity Instrument (SSI), does not include them but uses symptoms that involve word fragmentation (e.g., prolongation, part-word repetition or word break) instead. Other support for the view that WWR are not indications of fluency difficulty is reviewed in Howell (2010). Subsequently Jiang *et al*. (2012) examined whether or not WWR should be grouped with the fragmentary symptoms of fluency difficulty. People who stuttered heard an incomplete sentence, which they completed whilst lying in an MRI scanner. Each trial was classified according to any symptoms of non-fluency that occurred, and the associated brain activation patterns for different types of non-fluencies were examined. The trials where there had been fragmentary dysfluencies activated different areas of the brain from those where there were other common non-fluencies not associated with fluency difficulty (multi-syllable word repetitions, phrase repetitions and pauses). The brain activation

patterns from trials where there had been a WWR were then automatically assigned to either the fragmentary or the common non-fluency class. The patterns that occurred in WWR were placed with the common non-fluencies, supporting the view that they are not symptoms of fluency difficulty.

The symptom measure used in SSI (% syllables with fragmentary dysfluencies/all syllables, %SS) distinguishes those 4–5-year-old children who do and do not have fluency difficulty (Howell 2013, Mirawdeli and Howell 2016). Identification of fluency difficulty using fragmentary symptoms worked even when samples contained approximately 40% CwEAL (Mirawdeli and Howell 2016). Howell's (2013) procedure does not classify children with high rates of WWR as having fluency difficulty. If WWR had been included as symptoms of fluency difficulty, symptom rate would have been approximately 25% higher (Yairi and Ambrose 2005), more children's symptom counts would have exceeded the threshold, and some children could have been misclassified as having fluency difficulty. The prediction that inclusion of WWR in symptom counts would result in false alarms about fluency difficulty was confirmed by Howell (2013) who showed that models to identify children who stuttered that included WWR along with fragmentary symptoms had lower sensitivity and specificity than models that excluded them.

Howell's (2013) procedure allows the WWR symptoms to be used to indicate WFD. Consistent with this, WWRs are used as fillers when words are not known (Bada 2010, Clark and Clark 1977, MacWhinney and Osser 1977) and high rates of WWR symptoms alone are indicative of WFD (Bergmann *et al.* 2015, Fathman 1980, Fox *et al.* 1996, German 1991, Hilton 2008). Moreover, WWR rate indicates WFD irrespective of a child's language background (Lennon 1990, Rydland and Aukrust 2005). Hence, the proposal to separate off children with fluency difficulty using %SS and then to use rate of WWR in the remainder as an indication of WFD should be applicable to CwEAL as well as CwEO. Although the 40% CwEAL in Mirawdeli and Howell (2016) had high rates of WWR because of WFD, they were not misclassified as having fluency difficulty in comparison with teachers' indications.

NWR performance in children with fluency difficulty and WFD

Non-word repetition (NWR) performance could be used to validate the symptom-based approach in samples where there is diversity in language backgrounds. NWR performance is affected when a child has fluency difficulty (Anderson and Wagovich 2010, Bakhtiar *et al.* 2007, Hakim and Ratner 2004), These studies have reported repetition of two- and three-syllable non-words is poorer in children who stutter relative to fluent control children probably because of impaired phonological processing (Gathercole *et al.* 1994).

Also, in the UK, whilst NWR performance should be sensitive to fluency difficulties, it should not be affected when there is WFD in CwEAL: They often have to produce phoneme sequences in English words that they are not familiar with and this makes similar demands to NWR (Bialystok *et al.* 2003). The same should apply to CwEO who have WFD if the problem is due to poor vocabulary (Ellis Weismer *et al.* 2000). Consistent with this position, no studies reporting NWR deficits in children with WFD were returned in literature searches. In summary, NWR performance is: (1) an indication of fluency difficulty; (2) not impaired if children only have WFD; and (3) at least as good for fluent CwEAL as it is for fluent CwEO.

Design requirements of UNWR

Masoura and Gathercole (1999) reported that Greek children who learned English at school performed better on a Greek than an English NWR test. Also, Windsor *et al.* (2010) showed that children whose first language was Spanish were more accurate than CwEO on a Spanish NWR test, but the opposite was true when an English NWR test was used. Therefore, NWR tasks designed for one language cannot be applied to other languages and a task that applies to more than one language is required when NWR ability is assessed in heterogeneous language samples. To provide such a test, a common core of syllabic phonotactic constraints was identified that apply to the languages spoken by the children who were tested. Non-words generated according to these constraints are phonologically well-formed for all these languages and are appropriate for testing children who speak any of the targeted languages.

UNWR tests non-words that are between two and five syllables in length. Tests that are used in schools need to be short (to minimize disruptions to children's learning time). To keep UNWR brief, the number of non-words at each syllable length was set at seven. A child had to get all seven items correct to progress to the next syllable length ($p < .05$, by Sign test). UNWR has to be straightforward to administer so that teachers can conduct it rather than employing outside professionals. Equipment requirements should be minimal, again for reasons of efficiency. Consequently, Conti-Ramsden *et al.*'s (2001) procedure was adopted, where the experimenter spoke the material. These considerations led to differences between the UNWR and Gathercole *et al.*'s (1994) Child test of NWR (CNRep) which is widely used to test English-speaking children. CNRep has ten non-words per syllable length, delivers all syllable

lengths to a child irrespective of performance on shorter non-words and works from recordings.

UK schools insist that testing is conducted in English as they do not have the staff to deal with all the languages they encounter (NALDIC, 2013) and teaching English is a focal educational goal.[2] Together these considerations make UNWR a brief, easily administered test that can be used to assess all children who speak one of the 20 languages.

The UNWR non-words were checked to ensure they were not words in any of the other languages (by native respondents for all 20 languages and computationally using lexicons for the five languages used in this study). Non-words were also checked for word-likeness for the five targeted languages. Non-words that obey the phonotactic constraints of an individual's native language are high in word-likeness and this influences accuracy of repetition (Dollaghan *et al*. 1995, Gathercole 1995, Munson *et al*. 2005). Luce's (1986) neighbourhood density measure was used to determine how many phonologically similar words there are to a target non-word.

Current study

Short samples of speech were used to determine participants' degree of fluency difficulty and degree of WFD (%SS, and %WWR instances out of all syllables, respectively). Multiple regression analyses were conducted that tested whether: (1) %SS and language group (CwEO versus CwEAL) were predictors of UNWR; (2) %WWR and language group were predictors of UNWR; and (3) whether UNWR scores varied across languages. The corresponding predictions based on Howell (2013) were: (1) %SS should predict UNWR; (2) %WWR should not predict UNWR; (3) UNWR should not favour any of the languages and, therefore, it should be immune to differences in English language ability of the CwEAL. Therefore, language group was not expected to be a significant predictor in any regression. The children were also tested on CNRep for comparison with UNWR

Method

Participants

All children from reception classes in five mainstream primary schools were tested (three in the London borough of Merton and two in Ipswich. Ipswich had a population of 133,400 in 2014, of which 82.9% were White British. The average weekly pay for men in Ipswich was £456.[3] This was lower than that of England overall (£513). Approximately a quarter of the Ipswich population (26.6%) lived in the most deprived conditions in England.[4] The population of Merton was 199,700 in 2015 and 75.0% of the population were white British. The median gross weekly pay in Merton was £535.50, which was fourteenth out of 32 London boroughs.[5] Based on these statistics, both regions are in mid to low socio-economic areas.

None of the children had neurological deficits. Language spoken was reported by the schools. Thirteen children were excluded because they spoke a language not in the set of 20 UNWR languages and 18 more were excluded because a lexicon was not yet available for that language for checking word status and word-likeness properties computationally. Five children were excluded because their %SS was more than three standard deviations from the mean. The 17% of children outside that limit were considered outliers where random factors led to variation in %SS (e.g., a child had an emotional issue at home or at school or was starting with an illness). The remaining 96 children spoke English (79.17%), Urdu (7.29%), Polish (5.21%), Portuguese (7.29%) or Romanian (1.04%). There were 49 males (41 of whom were CwEO) and 47 females (35 of whom were CwEO). The mean age of the CwEO group was 4.49 years (SD = 0.50 years) and of the CwEAL was 4.60 years (SD = 0.50 years). This difference in age between the two groups was not statistically significant using Wilcoxon rank-sum test with continuity correction ($p = .3730$). Ethics approval was granted by UCL's IRB (4374/001).

UNWR stimuli

The UNWR stimuli varied in the number of syllables within a non-word (syllable length) and the number of consonants within a syllable. Syllable-internal complexity was varied according to two main parameters that have the same settings in all 20 target languages (English included): (1) the number of consonants in the syllable onset (one versus two) and (2) whether or not a syllable was closed by a coda consonant. Two lower thresholds of syllabic complexity were fixed: (1) every syllable contained a vowel (some of the languages lack syllabic consonants), and (2) every onset contained at least one consonant (some of the languages lack onsetless syllables). Two upper thresholds were also fixed: (1) onsets contained at most two consonants (some of the languages lack onset clusters larger than this), and (2) codas contained at most one consonant (some of the languages lack complex coda clusters).

This combination of parametric settings yielded the following set of syllable template: CV (consonant + vowel) versus CCV (simple versus complex onset); CV versus CVC (open versus closed syllable). These templates were strung together to form polysyllabic nonwords. Consonant clusters have two sources: (1) they occur as complex onsets (CCV), or (2) they straddle the boundary between a closed syllable and a following onset

(VC-CV). For each syllable template, all possible phone sequences were created using *Python* (Python Software Foundation, Beaverton, OR, USA) with the following additional constraints:

- Twelve consonants ([p, t, k, b, d, g, f, s, m, n, r, l]) were selected. According to the UCLA Phonological Segment Inventory Database UPSID (Maddieson 1984), these account for 57.57% of consonant occurrences in the target languages and typically developing 5–6-year-olds, produce 98.05% correctly (Raitano *et al.* 2004).
- Long monophthongs and diphthongs were not permitted as UNWR vowels since they are absent in some of the languages (e.g., Polish). Five short vowels were selected: [i, e, a, u, o] that account for 41.46% of the vowel occurrences in the target languages (Maddieson 1984).
- Word stress was not indicated in the basic set of UNWR stimuli. Although all 20 target languages have stress, they differ considerably with respect to where this falls within words. Moreover there are also major differences in terms of the impact stress has on vowel quality and vowel length. UNWR abstracts away from these differences in two ways: (1) by focusing on consonant phonotactics and (2) by allowing stress, vowel length and vowel quality to vary according to the first language of the speaker producing the non-word stimuli.
- In view of the third point above, another set of otherwise identical non-word stimuli was generated where all the vowels were reduced to schwa. Both the unreduced and the reduced sets of non-words were examined in the lexical checks described below, as it is possible for an unreduced non-word to be a real word in the reduced form.
- The first consonant of an onset cluster was always an obstruent ([p, t, k, b, d, g, f]) and the second always a liquid ([r, l]), except that neither [t] nor [d] were followed by [l] ([tl, dl] onsets are barred in all 20 languages). Only these initial clusters are permitted by all 20 languages (Harris 1994).
- Word-final consonants could only be [p, t, k, m, n]. All cross-syllable consonant clusters (i.e., C.C in [. . . VC.CV . . .]) were restricted to a coda nasal [m, n] followed by an onset plosive of the same place of articulation (e.g., [mb, mp, nt, nd]. These constraints were implemented to reflect the heavy restrictions that some of the target languages impose on what consonants can appear in a coda (Harris 1994).
- [s] was only used as a singleton and not in initial clusters. Not all of the languages allow [s] plus consonant to occur word-initially. Also, there is evidence that [s] in initial clusters is not integrated into the following onset and that medial [s]C clusters are always syllabified with the [s] in the coda rather than the following onset (Harris 1994).

Non-word candidates were selected for each syllable length as follows. First, all permitted syllable combinations which could serve as templates (e.g., [CV.CV] for two-syllable combinations) were created for each syllable length, and 100 of these templates per syllable length were selected at random. Second, consonant and vowel phones were selected randomly, apart from the above constraints, and entered into the template for each syllable length. Third, individual syllables were combined. The stimulus set featuring vowel reduction was generated by converting all vowels to schwa. The phonetic transcriptions in table 1 show the difference between two versions of UNWR: a language-independent ('universal') form in which stress, vowel length and reduction are absent and an English implementation in which these properties are present.

Computational checks that UNWR non-words are not words in any of the five languages and evaluation of word-likeness of UNWR and CNRep stimuli

Phonemic lexicons

Each UNWR non-word was checked to ensure that the candidate string was not a word in any of the five languages for the full and reduced forms. Subtitle-based lexicons (SUBTLEX) that employ TV/film subtitles to identify words in the languages and their frequencies were employed (New *et al.* 2007). SUBTLEX predict performance variables like reaction times in lexical decision tasks well. SUBTLEX-UK (van Heuven *et al.* 2014), SUBTLEX-PL (Mandera *et al.* 2014) and SUBTLEX-PT (Soares *et al.* 2014) were used for British English, Polish ad European Portuguese respectively.

A Romanian lexicon (SUBTLEX-RO) was custom built for this study using a similar procedure to that applied to Brazilian Portuguese (Tang 2012) except that de-duplication was not done for SUBTLEX-RO. De-duplication removes multiple translations of the same film/TV episode and avoids potential inflation of word frequency of words that occur in the duplicated material. However, there is no evidence that this is the case.

Urdu is phonologically almost indistinguishable from Hindi (Masica 1991) (there are differences with respect to loanwords in each language). Hindi was used as a proxy for Urdu as there was no material for building a lexicon for Urdu. Hindi Wikipedia (2014)

Table 1. Orthographic and phonetic transcriptions of two-, three-, four- and five-syllable UNWR non-words. Two- and three-syllable non-words are given in (a) with orthographic forms to the left and phonetic forms to the right. Four- and five-syllable non-words are given in a similar way in (b)

(a)

Two-syllable non-word		Three-syllable non-word	
'Universal'	English	'Universal'	English
Frofrat	ˈfɹəw.fɹat	flifoflip	ˈflɪ.fəw.flɪp
blimpruk	ˈblɪm.pɹʌk	laklemban	ˈlɑː.klɛm.bɑːn
dromprok	ˈdɹɔm.pɹɔk	bletondri	ˈblɛ.tɔn.dɹɪj
blomplin	ˈblɔm.plɪn	gokimap	ˈgəw.kɪ.map
fekrip	ˈfɛ.kɹɪp	flontrendrut	ˈflɔn.tɹən.dɹʊt
glamblen	ˈglam.blɛn	grundrimpop	ˈgɹʌn.dɹɪm.pɔp
dimblit	ˈdɪm.blɪt	prempilut	ˈpɹɛm.pɪ.lʊt
brafre	ˈbɹɑː.fɹɛj	gimbompop	ˈgɪm.bɔm.pɔp
flamplon	ˈflam.plɔn	tridrumblap	ˈtɹɪ.dɹʌm.blap
brentrik	ˈbɹɛn.tɹɪk	flotrambe	ˈfləw.tɹəm.bə

(b)

Four-syllable non-word		Five-syllable non-word	
'Universal'	English	'Universal'	English
mimblaplifla	mɪm.blɑː.ˈplɪ.flɑː	britragokoga	bɹɪ.tɹɑː.gə.ˈkəw.gɑː
tetombogret	tɛ.ˈtɔm.bə.gɹɛt	suntofrafepre	sʌn.tə.ˈfɹɑː.fə.pɹə
kundrebrembam	kən.dɹə.ˈbɹɛm.bam	blegrenugrantra	blə.gɹɛ.nə.ˈgɹan.tɹɑː
gideplasot	gɪ.dɛ.ˈplɑː.sɔt	pugantumpoluk	pə.gan.ˈtəm.pə.lək
pruligloma	pɹə.lɪ.ˈgləw.mə	trablenduntimbla	tɹɑː.blɛn.ˈdʌn.tɪm.blɑː
bofodoplup	bə.fə.ˈdəw.pləp	plendetendumbat	plɛn.də.tən.ˈdəm.bat
rimbefripep	ɹɪm.bə.ˈfɹɪ.pɛp	sembumpiklempet	sɛm.bəm.pɪ.ˈklɛm.pɛt
glubumbipa	glə.bʌm.ˈbɪ.pɑː	klodrinigandin	kləw.dɹɪ.nɪ.ˈgan.dɪn
gliklolimpi	glɪ.klə.ˈlɪm.pɪj	kumbripomplibet	kəm.bɹɪ.ˈpɔm.plɪ.bɛt
godrampimbum	gəw.dɹam.ˈpɪm.bʌm	grodrantutripe	gɹəw.dɹan.tə.ˈtɹɪ.pə

Note: Transcriptions in the 'Universal' columns have language-independent IPA phonetic values. The corresponding transcriptions in the 'English' columns follow the format of CUBE (Lindsey and Szigetvári 2014), to reflect contemporary Southern British English.

articles were used to create the Urdu/Hindi lexicon. Encyclopaedia articles may reduce validity as they overestimate the frequency of words that are used in formal settings and underestimate the frequency of words used in conversations.

Pre-processing (pruning and phonetic transcription)

Table 2, column 3, shows that the number of tokens and types ranged across the lexicons from 21.6 million words (Urdu/Hindi) to 237.1 million words (SUBTLEX-RO). The number of word types depends on the size of corpora (more word types tend to be found in large corpora), which would affect word-likeness. Rare words were removed to offset any bias caused by differences in tokens and types across the five languages by excluding words that occurred in fewer than three contexts.[6] For the subtitle corpora, each subtitle text was a context, while for the Urdu/Hindi Wikipedia corpus, each article was a context.

Phonemic lexicons were created next. Pronunciation sources to convert from orthographic to phonemic form were available for British English, European Portuguese and Romanian (sources are in column five of table 2). Any words that were not in the corresponding dictionaries were converted using a Grapheme to Phoneme converter that used the corresponding pronunciation resource as training data (Jiampojamarn *et al*. 2010). Polish was converted by a customized rule-based method since the orthographic system of Polish is phonemically regular. Urdu/Hindi was transcribed using the eSpeak (2013) text-to-speech toolkit that converts orthographic forms to IPA forms using rules and dictionary conversions. The pronunciation sources varied in the amount of detail given. However, conversion into phonemic form normalized any such differences across the languages. Column 6 of table 2 gives the source of phonemic analysis for each language.

Word-likeness

Luce's (1986) phoneme edit metric was used to determine the number of phonologically similar words there are to a target. A target non-word has a word as a neighbour if that word results when a single operation at a phonemic level is made. The operations are insertion, deletion or substitution of a phoneme, and these are

Table 2. Summary of sources and statistics for the five target languages spoken by children in the report

Lexicon	Source of lexicon	Token size (million words)	Type size (Pruned)	Source of pronunciation	Source of phonemic analysis
British English	SUBTLEX-UK (Full) (van Heuven *et al.* 2014)	201.7	332,987	Lindsey and Szigetvári (2014)	Lindsey (2012)
Polish	SUBTLEX-PL (Mandera *et al.* 2014)	146	987,911	Custom built	Gussmann (2007)
European Portuguese	SUBTLEX-PT (Soares *et al.* 2014)	78	132,710	Veiga *et al.* (2013)	Mateus and d'Andrade (2000)
Romanian	SUBTLEX-RO (custom built)	237.1	495,302	Schlippe *et al.* (2010)	Chitoran (2002)
Urdu/Hindi	Hindi Wikipedia Dump (Wikipedia 2014)	21.6	513,018	eSpeak (2013)	Ohala (1999)

Note: Source of lexicon and size in terms of token and type (the latter after pruning as described in the text), pronunciation sources and phonemic conversion sources are listed in columns 2–6.

made at all positions in the target phone string. The measure indicates the number of words differing by a single phoneme from the target non-word.

The word-likeness estimate of the target non-word was the total number of word neighbours found after all operations were applied (the more neighbours, the higher the word-likeness). Each non-word was checked against all the words in the lexicons of the five languages using Levenshtein distance (Levenshtein 1966).

Using full non-word target transcriptions to compute word-likeness scores weights vowel and consonant phone changes equally. Vowels vary more across languages than do consonants and this can impact on the number of neighbours the same non-word has in different languages. Vowels were reduced to schwa (the reduced set) and word-likeness scores were recalculated This normalized comparisons across languages that have different numbers of vowels, avoided problems that arise because vowels have different qualities across the languages and gave vowels less weight than consonants when determining word neighbours of non-word strings. Word-likeness scores were computed for the two non-word repetition sets (UNWR versus CNRep), two types of vowel information (full versus. reduced), four syllable lengths (length ranged from two to five syllables in UNWR and CNRep) and the five selected languages.

Procedure

Children were tested individually in a quiet room in a 15-min session, which was audiotaped on a Zoom H4n recorder using an X/Y stereo microphone. A speech sample was elicited using probe questions and Riley's (2009) picture material. The mean length of the samples was 333.21 syllables. Finally, the two NWR tests were performed. The order of presentation of UNWR and CNRep was counterbalanced across children. Children were told prior to the non-word tests that they would hear made-up words that they should repeat. For UNWR, the non-words were spoken by a female experimenter using English stress patterns (table 1). Two- and three-syllable items had strong–weak and strong–weak–weak patterns respectively. The stress pattern for four- and five-syllable English words varied (their stress patterns are indicated in table 1). The stress patterns for CNRep stimuli are given in table 3 (Gathercole *et al.* 1994). Each NWR test began with the two-syllable stimuli, and syllable lengths were increased successively. Three stimuli per syllable-length were used as practice material. The seven test stimuli for a particular syllable length were randomized. Each child was allowed as much time as was necessary to respond after each non-word was presented. The accuracy of the child's production was evaluated as each word was tested. When evaluating a child's repetition of a UNWR stimulus, only consonants were scored; any differences in stress, vowel length or quality were ignored. All stimuli at a given syllable length were tested even when a child made an error, but a child only progressed to the next syllable length when all seven non-word stimuli at the present syllable length were correct. The total score for UNWR and CNRep was the number of non-words that were correct across all syllable lengths that were attempted.

Analysis and reliability estimates

The spontaneous speech samples were analyzed to obtain %SS (Riley 2009). For %WWR, all the repetitions on the monosyllabic word were counted as a single non-fluent event, which is how Riley (2009) treated stuttering symptoms to obtain SS. The total syllable count was adjusted by deducting the number of WWR repeats from the total syllable count in the %SS analysis in which each repeated word was counted. %WWR was the percentage of WWR non-fluent events out of all syllables.

Table 3. Orthographic and phonetic transcriptions of two-, three-, four- and five-syllable CNRep non-words. Two- and three-syllable non-words are given in (a) with orthographic forms to the left and phonetic forms to the right. Four- and five-syllable non-words are given in a similar way in (b)

(a)

Two-syllable non-word		Three-syllable non-word	
Orthographic	Phonetic	Orthographic	Phonetic
Ballop	ˈba.ləp	bannifer	ˈba.nɪ.fə
Bannow	ˈba.nəw	barrazon	ˈba.ɹə.zən
Diller	ˈdɪ.lə	brasterer	ˈbɹa.stə.ɹə
Glistow	ˈglɪ.stəw	commerine	ˈkɔ.mə.ɹɪn
hampent	ˈham.pənt	doppelate	ˈdɔ.pə.lɛjt
Pennel	ˈpɛ.nəl	frescovent	ˈfɹɛ.skə.vənt
Prindle	ˈpɹɪn.dəl	glistering	ˈglɪ.stə.ɹɪŋ
Rubid	ˈɹʉw.bɪd	skiticult	ˈskɪ.tɪ.kʌlt
sladding	ˈsla.dɪŋ	thickery	ˈθɪ.kə.ɹɪj
Tafflest	ˈta.fləst	trumpetine	ˈtɹʌm.pɪ.tɪjn

(b)

Four-syllable non-word		Five-syllable non-word	
Orthographic	Phonetic	Orthographic	Phonetic
blonterstaping	ˈblɔn.tə.ˈstɛj.pɪŋ	altupatory	al.ʧʉw.ˈpɛj.tə.ɹɪj
commeecitate	kə.ˈmɪj.sə.tɛjt	confrantually	kən.ˈfɹan.ʧʉw.lɪj
contramponist	kən.ˈtɹam.pə.nɪst	defermication	də.ˈfɜː.mɪ.ˈkɛj.ʃən
empliforvent	ɛm.plɪ.ˈfoː.vənt	detratapillic	də.tɹa.ta.ˈpɪ.lɪk
fenneriser	fɛ.nə.ˈɹɑj.zə	pristoractional	pɹɪ.stə.ˈɹak.ʃə.nəl
loddenapish	lɔ.də.ˈnɛj.pɪʃ	reutterpation	ɹɪj.ˈʌ.tə.ˈpɛj.ʃən
pennerriful	pə.ˈnɛ.ɹɪ.fəl	sepretennial	sɛ.pɹə.ˈtɛ.nɪj.əl
perplisteronk	pə.plɪ.stə.ˈɹɔŋk	underbrantuand	ʌn.də.ˈbɹan.tʉw.ənd
stopograttic	ˈstɔ.pə.ˈgɹa.tɪk	versatrationist	ˈvəː.sə.ˈtɹɛj.ʃə.nɪst
woogalamic	wɵ.gə.ˈla.mɪk	voltularity	vɔl.ʧə.ˈla.ɹɪ.tɪj

Note: Transcriptions in the 'orthographic' columns are those given for these materials by Gathercole *et al.* (1994). The corresponding transcriptions in the 'phonetic' columns follow the format of CUBE (Lindsey and Szigetvári 2014), to reflect contemporary Southern British English (stresses in these columns correspond to those given by Gathercole *et al.* 1994).

The speech recordings were reanalyzed for %SS and %WWR so that intra- and inter-judge reliability could be calculated. For intra-judge reliability the original judge reassessed 10 randomly selected samples. Agreement about non-fluent events for the 10 samples varied between 87 and 95% for %SS and between 90 and 95% for %WWR. All κ coefficients were above .75, 'excellent' according to Fleiss (1971). For inter-judge reliability, a second judge estimated %SS and %WWR on ten randomly selected children's speech samples. Agreement, calculated as above, across the judges for the ten samples varied between 81 and 90% for %SS and between 85 and 92% for %WWR. All κ coefficients were again above .75.

Ten UNWR and ten CNRep tests were reassessed by the same judge and by a second judge. Inter- and intra-judge reliability for UNWR and CNRep consonant accuracy was assessed as percentage agreement across the two overall correct scores. κ coefficients indicated excellent agreement in all cases (Fleiss 1971).

Results

Evaluation of whether UNWR is appropriate for equitable testing across languages and comparison with CNRep

The average number of neighbours is an index of word familiarity. Row one of figure 1 shows the data where there was no vowel reduction: None of the UNWR non-words had any word neighbours for any of the five lexicons; Row two shows that CNRep non-words had neighbours for two-syllable non-words across all lexicons, and for three-syllable non-words for the British English lexicon and there were few neighbours for any lexicon for the four- and five-syllable non-words. Comparison across languages showed that the CNRep stimuli had more neighbours for the English lexicon than for the non-English lexicons. The third and fourth rows in figure 1 show the data when the vowels were reduced. Again, CNRep non-words had more neighbours than UNWR non-words and only the British English, Polish and Romanian two-syllable UNWR non-words had neighbours. Overall, CNRep had more neighbours

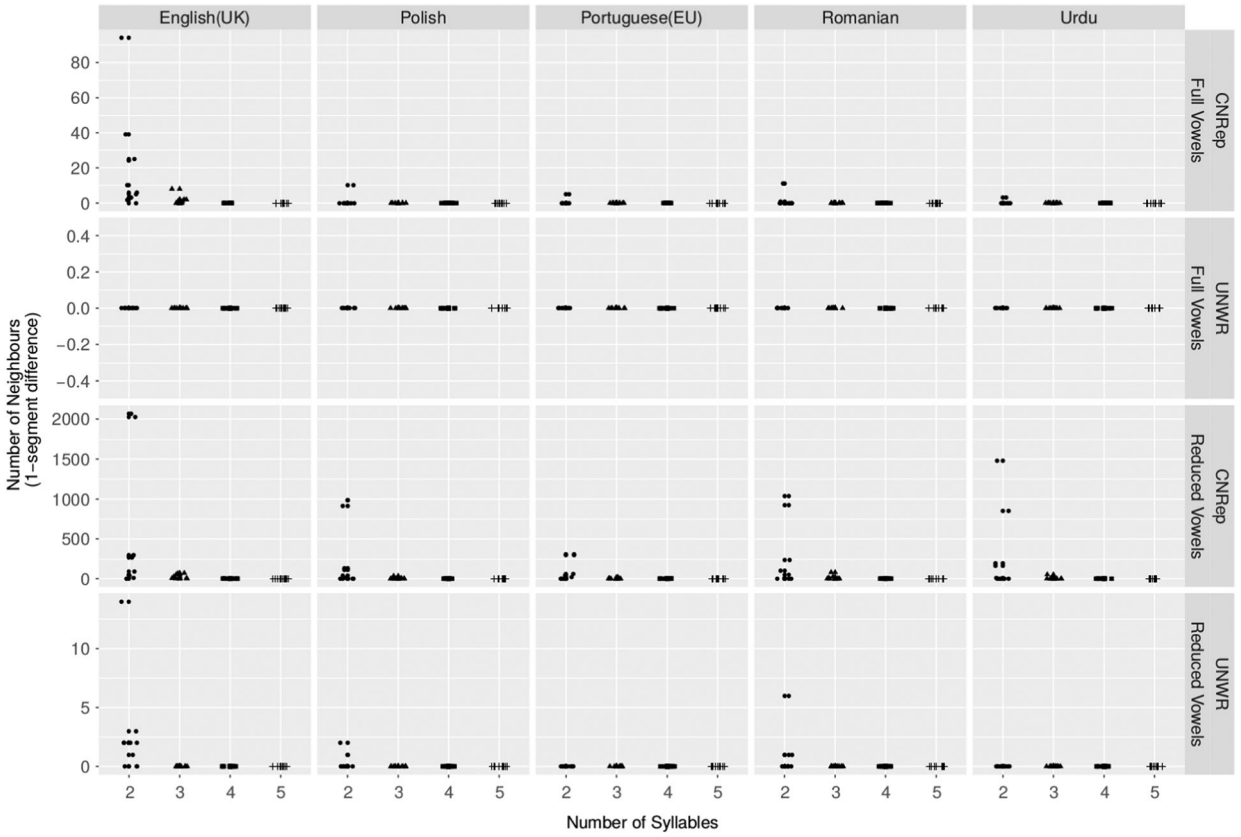

Figure 1. Neighbourhood estimates with phoneme edit distance–repetition set (2) × vowel reduction (2) × lexicon (5) × syllable size (4)

than UNWR, particularly when non-words had few syllables. Statistical models evaluated this pattern of variation in the number of neighbours to establish whether UNWR is appropriate to assess the children who participated in the study.

Analysis of these data is complicated because 88.1% of the data points from figure 1 had zero counts (figure 2). Consequently, a Hurdle model (Mullahy 1986) was used. There are two components to hurdle models: the hurdle component itself which models zero counts versus non-zero counts. A binomial distribution was used to model this component (Zeileis *et al*. 2008); the truncated count component which models positive counts. A negative binomial distribution was used to model the truncated component (Zeileis *et al*. 2008).

In the analyses, the predictee was the neighbour counts and the predictors were as follows for both components:

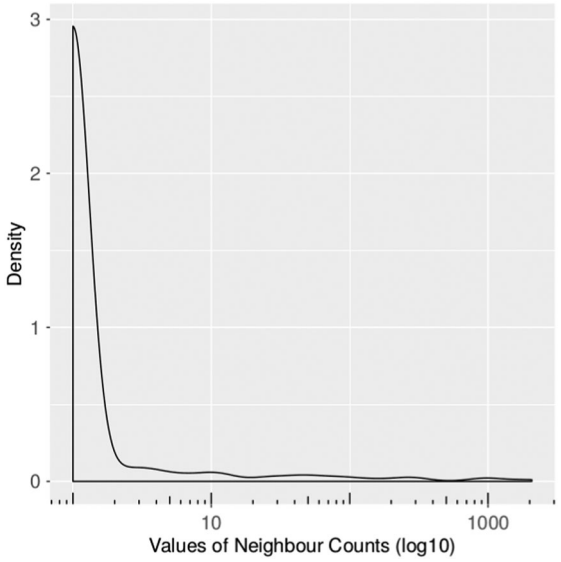

Figure 2. Density plot of neighbour counts with phoneme edit distance

- *Repetition Set* (UNWR versus CNRep). This was dummy-coded with CNRep as the base as this is the standard NWR test used in the UK.
- *Vowel Reduction* (Reduced versus Full). This was simple-coded with Full as the base as these are the stimuli the children heard in the UNWR test.
- *Language* (British English, Polish, European Portuguese, Romanian or Urdu-Hindi). Language was deviation-coded with British English as the

base. This resulted in each of the other languages being contrasted with British English.
- *Syllable Size* (2–5) was modelled as a categorical variable to avoid assuming that effects of syllable-length increase are interval-spaced. It was Helmert-coded, with each subsequent level contrasted with the mean of the previous levels, e.g., three-syllable results were contrasted with two-syllable results, four syllables were contrasted with the mean of two- and three-syllable results etc.

A saturated model with fully crossed predictors (main effects and all interaction terms) was fitted first as a baseline against which improvement in subsequent models could be ascertained. This did not converge. Consequently, the model was simplified by basing the order in which factors were excluded on the hierarchy from most complex (the biggest interaction term) to least complex (single factor terms). The marginality principle was adhered to, meaning that models containing an interaction term were not allowed unless the respective main effects and all lower-order interactions with this factor were significant.

This model converged and provided separate regression terms for the count and hurdle components. Starting with the Converged model, a series of nested model comparisons was performed using a χ^2 likelihood ratio test (lrtest) with $\alpha = .05$ (Zeileis and Hothorn 2002). Retention or omission of factors was based on whether or not model fit changed significantly: A term was retained or omitted depending on whether fit reduced or improved significantly respectively. Again the principle of marginality was adhered to. A series of nested model comparisons was then performed on superset and subset model pairs. A superset model contained more regression terms than a subset model, and all the terms in a subset model had to be present in its superset model. The best-path algorithm was used for model construction where going backward (exclusion) determined whether there were multiple subset models that resulted in p-values that exceeded the α-level in their nested model comparisons with the superset model. This indicated which subset model had the strongest evidence (the highest p-value). When going forward (inclusion), if there were multiple superset models that resulted in p-values below the α-level in their nested model comparisons with the subset model, the superset model with the strongest evidence (the lowest p-value) was selected. The direction of comparisons was first backward (exclusion) then forward (inclusion), and this pattern was repeated until no further terms were included or excluded. The order in which terms were excluded went from the most complex (the largest interaction terms) to the least complex (single terms). The reverse was the case when terms were considered for inclusion. The process was performed for the count and hurdle components.

This procedure provides the optimal combination of the selected factors in the final prediction model (summarized in table 4). There were main effects of Repetition Set, Vowel Reduction and Syllable Size for the count component, and of Repetition Set, Vowel Reduction, Language and Syllable Size for the hurdle component. There were no interaction terms for either component. All the predictors (*Repetition Set*, *Vowel Reduction*, *Language and Syllable Size*) discussed in the descriptive analysis contributed significantly to the prediction of word-likeness except that the Syllable Size contrast [5 versus {2, 3, 4}] was not significant in the count component. This suggests that the syllable size effect decreased with length (an effect only occurred up to four syllables). Language was present in the hurdle component showing that non-words were more word-like in English than the non-English languages. Language was absent in the count component, showing that non-words that have at least one neighbour were not significantly more word-like in English than non-English languages. UNWR had lower word-likeness (fewer neighbours) than did CNRep overall and for each of the five languages with either no difference (Full vowels), or small differences (Reduced vowels), across the languages. The virtual absence of a language effect on word-likeness with UNWR reduced and full stimuli allows comparison of performance across children who speak one or more of the target languages for which it has been comprehensively evaluated.

Speech measures and language group as predictors of UNWR or CNRep scores

For the children's performance data, mean %SS was 3.39 (SD = 4.25) for CwEO and 1.90 (SD = 1.37) for CwEAL. Mean %WWR was 3.33 (SD = 2.53) for CwEO and 2.65 (SD = 2.88) for CwEAL. CwEO tended to show higher rates of both types of difficulty than did CwEAL. The mean %SS and %WWR were not significantly different between the EO and EAL groups by Wilcoxon tests ($p = .1593$ and $p = .1998$ respectively).

Statistical analyses employed one of the two speech measures (%SS or %WWR) and language group (EO versus EAL) to predict either UNWR or CNRep scores giving four models in total. The modelling procedure was equivalent in all cases. First a saturated model with fully crossed predictors (single factors and all interaction terms) was fitted. A series of nested model comparisons was then made using a χ^2 likelihood ratio test ($\alpha = .05$). This showed which factors (main effects and any of their interaction terms) could be left out without model fit being affected. Only those factors left after non-significant

Table 4. Optimal Hurdle model: neighbourhood estimates with phoneme edit distance

(a)

Count model coefficients

Fixed effects	Estimate	SE	Z	p (>|z|)
(Intercept)	−5.1941	60.4219	−0.0860	0.9310
Repetition Set [UNWR versus CNRep (base)]	−5.5401	0.5533	−10.0120	$2 \times 10^{-16***}$
Syllable Size [3 versus 2 (base)]	−2.7292	0.4174	−6.5380	$6.22 \times 10^{-11***}$
Syllable Size [4 versus {2, 3} (base)]	−2.9790	0.5210	−5.7170	$1.08 \times 10^{-8***}$
Syllable Size [5 versus {2, 3, 4} (base)]	−7.2979	120.8404	−0.0600	0.9520
Vowel Reduction [Reduced versus Full (base)]	3.1787	0.4503	7.0600	$1.67 \times 10^{-12***}$
log (theta)	−1.2469	0.3106	−4.0150	$5.95 \times 10^{-5***}$

(b)

Zero hurdle model coefficients

Fixed effects	Estimate	SE	Z	p (>|z|)
(Intercept)	−5.1904	0.4856	−10.68	$2 \times 10^{-16***}$
Repetition Set [UNWR versus CNRep (base)]	−3.9096	0.4998	−7.82	$5.18 \times 10^{-15***}$
Language [Polish versus English (Full) (base)]	−2.7796	0.5313	−5.23	$1.68 \times 10^{-7***}$
Language [Portuguese versus English (Full) (base)]	−3.6368	0.6065	−6.00	$2.02 \times 10^{-9***}$
Language [Romanian versus English (Full) (base)]	−2.1064	0.4823	−4.37	$1.26 \times 10^{-5***}$
Language [Urdu versus English (Full) (base)]	−4.0528	0.6449	−6.28	$3.29 \times 10^{-10***}$
Syllable Size [3 versus 2 (base)]	−2.3137	0.3997	−5.79	$7.07 \times 10^{-9***}$
Syllable Size [4 versus {2, 3} (base)]	−2.5636	0.4126	−6.21	$5.17 \times 10^{-10***}$
Syllable Size [5 versus {2, 3, 4} (base)]	−2.2139	0.5396	−4.10	$4.08 \times 10^{-5***}$
Vowel Reduction [Reduced versus Full (base)]	3.0567	0.4294	7.12	$1.09 \times 10^{-12***}$

Note: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $+p < .1$

factors were pruned were significant predictors of the respective NWR scores.

All four continuous variables (UNWR scores, CNRep scores, %SS and %WWR) were logarithmically transformed (base 10, with Laplace smoothing), and then converted into z-scores. The categorical language variable was coded with contrast values of 1 and –1 for EAL and EO respectively.

%SS and language group were assessed as predictors of UNWR scores in the first model. The model only included %SS as a main effect. The model is formulated as 'UNWR scores ~ %SS' where '~' = predicted by. The fit statistics are given in the first section of table 5 (model 1). Children who had high %SS tended to show lower UNWR scores.

The second model used %WWR and language group to predict UNWR scores. None of the factors were significant as main effects nor interactions. The model is formulated as 'UNWR scores ~ Intercept' and statistics are summarized in the second section of table 5.

The next two models looked at either %SS (model 3) or %WWR (model 4) along with language group as predictors of CNRep scores. None of the factors was significant as main effects nor interactions in either model. Both models are formulated as 'CNRep scores ~ Intercept'. The statistics for these models are labelled model 3 and model 4 in table 5. The fit statistics for models 3 and 4 are identical because they are null models predicting the same thing—CNRep scores.

Language group and %WWR were not significant predictors in any of the models. However, %SS predicted UNWR scores ($p = .0044$) but not CNRep scores. Thus, a relationship between fluency difficulty (indicated by %SS) and NWR performance was only found for the UNWR test.

Given the findings from these four models, another regression was performed with both %SS and %WWR included ('UNWR scores ~ %SS + %WWR') to see if such a model predicted UNWR performance better than %SS. alone A nested model comparison showed that the model fit was significantly improved when %SS was included ($\chi^2(1) = 6.9822, p = .0082$) but not when WWR was included ($\chi^2(1) = 0.17376$, $p = .6664$). Therefore, WWR was dropped and the resultant model is identical to model 1 in table 5. The hypothesis that NWR scores can discriminate between fluency difficulty and WFD may apply to CNRep for the EO children that the test was designed for. To assess this, a further model (model 5) was fitted that only used data from the EO group. The model examined whether %SS predicted CNRep scores ('CNRep scores ~ %SS'). %SS was not a significant predictor of CNRep scores and it did not significantly improve the model fit ($\chi^2(1) = 1.3523$,

Table 5. Statistics for the four multiple regression models fitted to the data. Models 1 and 2 predicted UNWR scores, models 3 and 4 predicted CNRep scores. The first model in each pair used %SS, the second used %WWR as predictors. Language group was included in all models

(a) Model 1: Prediction of UNWR score using %SS and language group

Fixed effects	Estimate	SE	t	p (>|t|)
(Intercept)	1.7130×10^{-16}	0.0983	*0.000*	1.000
%SS	−0.0288	0.9877	−2.916	0.0044**

(b) Model 2: Prediction of UNWR score using %WWR and language group

Fixed effects	Estimate	SE	t	p (>|t|)
(Intercept)	2.0400×10^{-16}	0.1021	0	1.000

(c) Models 3 and 4: Prediction of CNRep score using %SS or %WWR and language group

Fixed effects	Estimate	SE	t	p (>|t|)
(Intercept)	3.399×10^{-17}	0.1021	0	1.000

(d) Model 5: Prediction of CNRep score using %SS with the English monolingual group

Fixed effects	Estimate	SE	t	p (>|t|)
(Intercept)	−0.00601	0.1205	−0.05	0.9600

Note: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $+p < .1$

$p = .2449$, $p > .1$), and was therefore dropped. Hence, CNRep scores were not sensitive to fluency difficulty.

Discussion

The proportion of CwEAL was high and a wide variety of languages were spoken in the five schools. The CwEO had a higher propensity to both fluency difficulty and WFD than the CwEAL. This may be because English is more developed in CwEO, putting them at a stage where fluency difficulties and WFD may have started, whereas CwEAL may have later onsets of both types of difficulty.

UNWR was shown to be appropriate for testing children who spoke the five languages as word-likeness was virtually equivalent across the languages. The relationship between speech symptom scores and UNWR and CNRep scores were evaluated in the main four statistical models. The models looked at whether %SS and %WWR, along with language status, predicted either UNWR or CNRep scores. UNWR scores were predicted by %SS, but not by %WWR. The relationship between %SS and UNWR scores confirmed hypothesis one, that UNWR provides a measure of fluency difficulty. The lack of a relationship between %WWR and UNWR scores confirmed hypothesis two, that UNWR scores were not affected when there was WFD. An ancillary analysis that included language and both %SS and %WWR as predictors showed that only %SS predicted UNWR scores, once more confirming that children with fluency difficulty, but not those with WFD, scored worse on this test. The lack of a main effect of language group in all of these regression analyses confirmed hypothesis three that UNWR is appropriate for use across the selected languages.

In the corresponding CNRep analyses, neither %SS nor %WWR were significant predictors of NWR scores (language was not significant, as with UNWR). Further analysis on CwEO alone showed that CNRep scores were not predicted by %SS. The hypothesis that NWR scores can discriminate between fluency difficulty and WFD only seems to hold with UNWR (all languages) and not for CNRep even when, as further analysis showed, CwEO were selected (for whom the test was appropriate). A possible explanation of the differences between the NWR tests is that CNRep is likely to reflect lexical influences on task performance due to the higher neighbourhood density of the non-words whereas UNWR seems to rely more on phonemic processing.

The findings that UNWR was predicted by %SS, but not %WWR are consistent with the claim that these speech symptoms separate fluency difficulty from WFD (Howell 2013). Thus, Howell's (2013) speech-symptom procedure was validated by the results from the UNWR test.

Patterns of non-fluencies in children with fluency difficulty and in CwEAL

One explanation of the relationship between SS and UNWR is that they both reflect phonological planning whereas WWR do not (WWR are an articulatory response used to hold the floor whilst generating or reformulating the following speech output). According to MacWhinney and Osser (1977), 75% of WWR in 5-year-olds occur on function words. Their explanation

was that function-word repetition allows pre-planning of upcoming utterances as opposed to them being prone to errors during production, since function words tend to be phonologically simple. Howell (2010) adapted this idea to account for different symptom patterns within phonological word (PW) units (Selkirk 1984). PW units have a content word (C) as the obligatory nucleus and function words (F) as optional affixes and suffixes. Howell assumed that the content-word in a PW has properties that can lead to its plan not being ready when prior words have been produced. For example, in the PW *He spilt it* (FCF) *spilt* is complex and its plan may not be ready after *he* has been uttered. The speaker can either; (1) repeat the prior function word *he* to buy planning time which would lead to a WWR (MacWhinney and Osser 1977); or (2) continue the utterance and rely on the remainder of the plan being completed whilst the part which is available is produced. If the plan is completed, the speech will be fluent. Otherwise fragmentary non-fluencies on the content word would occur. In this account, WWR and the fragmentary non-fluencies are different ways of tackling the same problem, i.e., that the plan for the content word has not been generated in time. Two factors jointly determine whether non-fluencies will arise in connected speech in this explanation: planning difficulty associated with the content word and the articulation rate on the stretch of function words that precede the content word. If the speech prior to the content word is produced rapidly, there is pressure on when the content-word plan needs to be ready. From this perspective, WWRs are regarded as a way of effectively reducing articulation rate that influences planning indirectly whereas SS indicate that speech has been attempted based on an incomplete plan. Many young children use function-word repetition in preference to fragmentary non-fluencies and this is not a barrier to fluency whereas children who produce fragmentary non-fluencies can experience long-term fluency difficulty (Howell 2010). Therefore, function-word repetition may allow the children to avoid fluency failure on content words.

Function-word repetition could also allow children to recover fluent speech control when there is WFD, as happens when, for instance, CwEAL do not know a word. Advancing to the content word is not an option as the word is absent so fragmentary non-fluencies cannot arise. WWR would allow time for these children to employ alternative ways out of such blocks (Howell 2010). For instance, a word from their native language could be used (code switching) or they could circumlocute using an alternative English word. To summarize: (1) SSs arise during planning and are an indication of fluency difficulty (Howell 2010, Mirawdeli and Howell 2016); and (2) WWRs operate at the articulation stage, which adjusts articulation rate to allow more time for planning when there is fluency difficulty (Howell 2010) and to allow a child with WFD time to determine an alternative formulation of the utterance (Bergmann *et al.* 2015, Fathman 1980, Fox *et al.* 1996, Hilton 2008).

Why should %SS predict UNWR scores whereas %WWR does not?

It follows that if fluency difficulty results from planning problems (Howell 2010), a relationship between %SS and UNWR would occur if UNWR also assays planning problems. Gathercole *et al.* (1994) argued that NWR is related to phonological rehearsal (which sets up plans). Evidence in support of this is that NWR scores correlate with auditory digit span in typically developing participants (Gathercole *et al.* 1994, Gray 2003, Gupta 2003) and with performance on memory tests in neuropsychological patients (Romani 1992). Rehearsal would be impaired in children with fluency difficulty because a plan can only be rehearsed once it has been generated and these individuals have planning impairments. Thus, the planning problem gives rise to SS and UNWR deficits in people with fluency difficulty and would offer one explanation as to why %SS predicts UNWR.

NWR performance also requires non-words to be segmented and perceived accurately (Gathercole *et al.* 1994). Wolk *et al.* (1993) proposed that phonological analysis is a problem for children with fluency difficulty and this could lead to fragmentary disfluencies if these representations are involved in generation of speech plans for output. UNWR would be affected if the proposal that phonological analysis problems lead to NWR deficits (Gathercole *et al.* 1994) applies to individuals with fluency difficulty. Thus, if people with fluency difficulty have problems in segmentation and perceptual analysis, this would affect NWR ability and provide another route by which %SS predicts UNWR.

The account of how WWR arise, which was extended here to CwEAL, argued that this type of disruption to speech does not arise from planning problems per se. Rather WWR solve the planning problem in an indirect way by slowing prior speech rate. The lower speech rate allows more planning time that avoids fragmentary non-fluencies. WWR would not impact on UNWR performance, which is a result of phonological planning, rather than articulation. Therefore, WWR should be independent of UNWR and would not be affected across language groups who have different levels of WFD (as was found).

Practical applications of UNWR

UNWR could provide an alternative method from that of Howell (2013) for identifying children with fluency

difficulty. Separating children into those with WFD who are otherwise fluent and those with fluency difficulty raises interesting possibilities for interventions. Group-based procedures could be developed for tackling WFD whereas SLTs could continue with the individually tailored interventions they currently give to children with fluency difficulties. The present results imply that interventions that tackle WFD should have an impact on WWR (i.e., they should reduce as the child learns more vocabulary) even though speech was not treated directly.

Limitations and future work

There are several ways in which UNWR could be improved. First, there are issues regarding test difficulty. CNRep was easier than UNWR and this could lead to ceiling effects. Conversely, UNWR could lead to floor effects. The balance that was struck was that performance on the two-syllable set in UNWR should not be at floor across all the 4–5-year-old children, but should be more difficult than CNRep. The advantage of UNWR being harder than CNRep is that it should be possible to test children out to older ages. At present, no adjustment for level of difficulty is anticipated, but results continue to be scrutinized for floor effects. Second, a scoring scheme was adopted where each repetition attempt was scored as correct or incorrect. Although such schemes yield large effect sizes (Estes *et al*. 2007), future revisions of the test could use a phoneme-by-phoneme scoring method and obtain the percentage of phonemes correctly repeated per non-word. Finally, the UNWR test should be norm-referenced using both clinical and typical populations to enhance its screening utility.

It could also be argued that %SS and %WWR are likely to be low in fluent children, and the limited range of scores could restrict their usefulness in regression analyses. However, this does not arise since Mirawdeli and Howell (2016) reported individual case classification of 879 children using this procedure, the majority of whom were fluent, and found good correspondence with teachers' judgments about the children.

Material was delivered using English vowels and stress patterns. This seems appropriate in the UK setting but a speaker of the native language or dialect should be used in other speech communities. A test is planned to see whether delivering UNWR in different native languages affects performance since the prosodic complexity of non-words has been shown to influence repetition accuracy, independent of the string complexity (Harris *et al*. 2007).

UNWR does not apply to every language spoken. Specific languages not covered are those with contrastive tone or with syllabic structure that is radically different from that of the languages in the present study. Separate tests for these groups may need to be developed but this would not be entirely satisfactory as comparison across language tests would not be possible. A database is being built up of children who speak languages outside those in the UNWR set. Their performance on UNWR will be examined to see whether or not non-included languages show patterns like those seen in included languages.

The CwEAL varied in their language balance. Although NWR tests have some immunity from language experience (Ellis Weismer *et al*. 2000), any potential influences need to be checked. Future work could also take a conventional measure of WFD and compare this with %WWR results.

Non-word material was constructed using phonotactic properties that are shared across the languages that were targeted for testing. This raises the question of what phonological material might have been prepared that is not shared across all the languages (or only shared by some languages). By leaving this material out, the test may be less valid for those languages that allow longer consonant strings than for those that only have the strings included in UNWR. On the other hand, the UNWR approach to stimulus construction appears valid since 'universal' phonotactics identify common underlying articulatory processes (Kawasaki-Fukumori 1992).

Conclusions

Approximately 7% of UK children experience fluency difficulties (Bercow 2008). Fluency difficulties limit an individual's psychosocial development and employment opportunities if not resolved early (Craig *et al*. 2009). A speech-based screen at school entry is desirable so as to minimize the delay between symptom onset and clinic referral if required (Howell 2013). The major challenge is how to identify fluency difficulties in CwEAL; the problem is that CwEAL can have WFD in English that lead to symptoms like WWR that can be misinterpreted as signs of fluency difficulty. This study addressed these issues and showed that: (1) %SS is appropriate to identify a range of fluency difficulties because it predicted UNWR scores; (2) %WWR did not predict UNWR; (3) both symptom and UNWR scores can be used to identify fluency difficulty in samples with diverse languages. The UNWR test could potentially be used as part of a screening procedure to identify fluency difficulties in linguistically diverse populations at school entry.

Acknowledgements

Parts of the research reported were supported by the Dominic Barker Trust. ***Declaration of interest:*** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Notes

1. The 20 languages are English, Polish, Romanian, European Portuguese, Bulgarian, Serbo-Croat-Bosnian, Czech, Dutch, French, German, Hungarian, Slovene, Swedish, Danish, Norwegian, Russian, Latvian, Ukrainian, Urdu-Hindi and Bengali. Children who spoke English, Urdu-Hindi, Polish, Portuguese or Romanian were represented in the sample in this report.
2. The CNRep transcriptions given by Gathercole *et al*. (1994) include North of England vowels that speakers of other varieties of English would pronounce differently. UNWR is explicitly designed to allow for such cross-dialect and cross-language differences in vowel quality.
3. See https://www.ipswich.gov.uk/sites/default/files/State%20of%20Ipswich%20AMReport%20(v1%201)%202014.pdf/.
4. See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6871/1871208.pdf2011/.
5. See http://www.merton.gov.uk/jsna_summary_document_2015_final.pdf/
6. The authors thank Dr Emmanuel Keuleers for suggesting this filter.

References

ANDERSON, J. D. and WAGOVICH, S. A., 2010, Relationships among linguistic processing speed, phonological working memory, and attention in children who stutter. *Journal of Fluency Disorders*, **35**, 216–234.

BROCKLEHURST, P. H., 2013, Stuttering prevalence, incidence and recovery rates depend on how we define it: comment on Yairi and Ambrose's article 'Epidemiology of stuttering: 21st century advances'. *Journal of Fluency Disorders*, **38**, 290–293.

BADA, E., 2010, Repetitions as vocalized fillers and self-repairs in English and French interlanguages. *Journal of Pragmatics*, **42(6)**, 1680–1688. doi:10.1016/j.pragma.2009.10.008

BAKHTIAR, M., ALI, D. A. A. and SADEGH, S. P. M., 2007, Non-word repetition ability of children who do and do not stutter and covert repair hypothesis. *Indian Journal of Medical Sciences*, **61**, 462–470.

BERGMANN, C., SPRENGER, S. A. and SCHMID, M. S., 2015, The impact of language co-activation on L1 and L2 speech fluency. *Acta Psychologica*, **16**, 125–135. doi:10.1016/j.actpsy.2015.07.015

BERCOW, J., 2008, *The Bercow Report: A Review of Services for Children and Young People (0–19) with Speech, Language and Communication Needs* (available at: http://webarchive.nationalarchives.gov.uk/20130401151715/https://www.education.gov.uk/publications/eOrderingDownload/Bercow-Summary.pdf).

BIALYSTOK, E., MAJUMDER, S. and MARTIN, M. M., 2003, Developing phonological awareness: is there a bilingual advantage? *Applied Psycholinguistics*, **24(1)**, 27–44. doi:10.1017/S014271640300002X

CHITORAN, I., 2002, *The Phonology of Romanian: A Constraint-Based Approach* (Berlin: Walter de Gruyter).

CLARK, H. H. and CLARK, E., 1977, *Psychology and Language. An Introduction to Psycholinguistics* (New York, NY: Harcourt).

CONTI-RAMSDEN, G., BOTTING, N. and FARAGHER, B., 2001, Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry*, **42(6)**, 741–748. doi:10.1017/S0021963001007600

CRAIG, A., BLUMGART, E. and TRAN, Y. T., 2009, The impact of stuttering on the quality of life in adults who stutter. *Journal of Fluency Disorders*, **34(2)**, 61–71. doi:10.1016/j.jfludis.2009.05.002.

DOLLAGHAN, C. A., BIBER, M. E. and CAMPBELL, T. F., 1995, Lexical influences on nonword repetition. *Applied Psycholinguistics*, **16(2)**, 211–222. doi:10.1017/S0142716400007098

ELLIS WEISMER, S., TOMBLIN, B., ZHANG, X., BUCKWALTER, P., CHYNOWETH, J. and JONES, M., 2000, Nonword repetition performance in school-aged children with and without language impairment. *Journal of Speech, Language and Hearing Research*, **43(4)**, 865–878 (available at: http://www.csee.ogi.edu/~gormanky/papers/NRT/ellis-weismer_etal_2000.pdf).

ESPEAK, 2013, *eSpeak Text to Speech, Version 1.47.11* (available at: http://espeak.sourceforge.net/).

ESTES, K. G., EVANS, J. L. and ELSE-QUEST, N. M., 2007, Differences in the nonword repetition performance of children with and without specific language impairment: a meta-analysis. *Journal of Speech, Language, and Hearing Research*, **50(1)**, 177–195. doi:10.1044/1092-4388(2007/015\051

FATHMAN, A., 1980, Repetition and correction as an indication of speech planning and execution processes among second language learners. In H. W. Dechert and M. Raupach (eds), *Towards a Cross-Linguistic Assessment of Speech Production* (Frankfurt am Main: Peter Lang), pp. 77–86.

FLEISS, J., 1971, Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76(5)**, 378–382. doi:10.1037/h0031619

FOX, B. A., HAYASHI, M. and JASPERSON, R., 1996, Resources and repair: a cross-linguistic study of syntax and repair. *Studies in Interactional Sociolinguistics*, **13**, 185–237.

GATHERCOLE, S. E., 1995, Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory and Cognition*, **23(1)**, 83–94. doi:10.3758/BF03210559

GATHERCOLE, S. E., WILLIS, C. S., BADDELEY, A. D. and EMSLIE, H., 1994, The children's test of nonword repetition: a test of phonological working memory. *Memory*, **2(2)**, 103–127. doi:10.1080/09658219408258940

GERMAN, D. J., 1991, *Test of Word Finding in Discourse* (Austin, TX: PRO-ED).

GRAY, S., 2003, Diagnostic accuracy and test–retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders*, **36(2)**, 129–151. doi:10.1016/S0021-9924(03\05100003-0

GUPTA, P., 2003, Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *Quarterly Journal of Experimental Psychology: Section A*, **56(7)**, 1213–1236. doi:10.1080/02724980343000071

GUSSMANN, E., 2007, *The Phonology of Polish* (Oxford: Oxford University Press).

HAKIM, H. B. and RATNER, N. B., 2004, Nonword repetition abilities of children who stutter: an exploratory study. *Journal of Fluency Disorders*, **29(3)**, 179–199. doi:10.1016/j.jfludis.2004.06.001

HARRIS, J., 1994, *English Sound Structure* (Oxford: Blackwell).

HARRIS, J., GALLON, N. and VAN DER LELY, H., 2007, Prosodic complexity and processing complexity: evidence from language impairment. *Revista da Associação Brasileira de Lingüística*, **6**, 1–19.

HILTON, H., 2008, The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, **36**, 153–166.

HOWELL, P., 2010, Language processing in fluency disorders. In J. Guendouzi, F. Loncke and M. Williams (eds), *The Handbook on Psycholinguistics and Cognitive Processes: Perspectives*

on Communication Disorders (London: Taylor & Francis), pp. 437–464.

HOWELL, P., 2013, Screening school-aged children for risk of stuttering. *Journal of Fluency Disorders*, **38(2)**, 102–123. doi:10.1016/j.jfludis.2012.09.002

JIAMPOJAMARN, S., CHERRY, C. and KONDRAK, G., 2010, Integrating joint n-gram features into a discriminative training framework. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010), USA*, pp. 697–700 (available at: http://delivery.acm.org/10.1145/1860000/1858102/p697-jiampojamarn.pdf?ip=116.48.57.17&id=1858102&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F343-7&CFID=732357301&CFTOKEN=27597819&__acm__=1448200680_922e618c5a8e67bd08d8b844289c7ba7).

JIANG, J., LU, C., PENG, D., ZHU, C. and HOWELL, P., 2012, Classification of types of stuttering symptoms based on brain activity. *PloS ONE*, **7(6)**, e39747. doi:10.1371/journal.pone.0039747

KAWASAKI-FUKUMORI, H., 1992, An acoustic basis for universal phonotactic constraints. *Language and Speech*, **35(1–2)**, 73–86. doi:10.1177/002383099203500207

LENNON, P., 1990, Investigating fluency in EFL. *Language Learning*, **40(3)**, 387–417.

LEVENSHTEIN, V. I., 1966, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10(8)**, 707–710.

LINDSEY, G. A., 2012, *The British English Vowel System* [online], 8 March (available at: http://englishspeechservices.com/blog/british-vowels/).

LINDSEY, G. A. and SZIGETVÁRI, P., 2014, *Current British English Searchable Transcriptions* (online] (available at: http://seas3.elte.hu/cube/\051.

LUCE, P. A., 1986, Neighborhoods of words in the mental lexicon. Doctoral dissertation, Indiana University (available at: http://files.eric.ed.gov/fulltext/ED353610.pdf).

MACWHINNEY, B. and OSSER, H., 1977, Verbal planning functions in children's speech. *Child Development*, **48(3)**, 978–985. doi:10.2307/1128349

MADDIESON, I., 1984, *Patterns of Sounds* (Cambridge: Cambridge University Press).

MANDERA, P., KEULEERS, E., WODNIECKA, Z. and BRYSBAERT, M., 2014, Subtlex-PL: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, **47(2)**, 471–483. doi:10.3758/s13428-014-0489-4

MASICA, C. P., 1991, *The Indo-Aryan Languages* (Cambridge Language Surveys) (Cambridge: Cambridge University Press).

MASOURA, E. V. and GATHERCOLE, S. E., 1999, Phonological short-term memory and foreign language learning. *International Journal of Psychology*, **34(5–6)**, 383–388. doi:10.1080/002075999399738

MATEUS, M. H. and D'ANDRADE, E., 2000, *The Phonology of Portuguese* (Oxford: Oxford University Press).

MIRAWDELI, M., 2016, *Assessing speech fluency problems in typically developing children aged 4 to 5 years* (Unpublished doctoral thesis) (London: University College London).

MIRAWDELI, A. and HOWELL, P., 2016, Is it necessary to assess fluent symptoms, duration of dysfluent events and physical concomitants when identifying children who are at risk of speech difficulties? *Clinical Linguistics and Phonetics*. http://dx.doi.org/10.1080/02699206.2016.1179345.

MULLAHY, J., 1986, Specification and testing of some modified count data models. *Journal of Econometrics*, **33(3)**, 341–365. doi:10.1016/0304-4076(86\05190002-3

MUNSON, B., EDWARDS, J. and BECKMAN, M. E., 2005, Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Journal of Speech, Language, and Hearing Research*, **48(1)**, 61–78 (available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.8737&rep=rep1&type=pdf).

NATIONAL ASSOCIATION FOR LANGUAGE DEVELOPMENT IN THE CURRICULUM (NALDIC), 2013, *EAL Statistics* (online) (available at: http://www.naldic.org.uk/research-and-information/eal-statistics/eal-pupils/) (accessed on 3 April 2016).

NEW, B, BRYSBAERT, M., VERONIS, J. and PALLIER, C., 2007, The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, **28(4)**, 661–677. doi:10.1017/S014271640707035X

OHALA, M., 1999, Hindi. In International Phonetic Association (ed.), *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet* (Cambridge: Cambridge University Press), pp. 100–103.

RAITANO, N. A., PENNINGTON, B. F., TUNICK, B. F., BOADA, R. and SHRIBERG, L. D., 2004, Pre-literacy skills of subgroups of children with speech sound disorders. *Journal of Child Psychology and Psychiatry*, **45**, 821–835.

RILEY, G., 2009, *The Stuttering Severity Instrument for Adults and Children (SSI-4)*, 4th ed. (Austin, TX: PRO-ED).

ROMANI, C., 1992, Are there distinct input and output buffers? Evidence from an aphasic patient with an impaired output buffer. *Language and Cognitive Processes*, **7(2)**, 131–162. doi:10.1080/01690969208409382

RYDLAND, V. and AUKRUST, V. G., 2005, Lexical repetition in second language learners' peer play interaction. *Language Learning*, **55**, 229–274.

SCHLIPPE, T., OCHS, S. and SCHULTZ, T., 2010, Wiktionary as a source for automatic pronunciation extraction. Paper presented at the 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, September 2010.

SELKIRK, E., 1984, *Phonology and Syntax: The Relation between Sound and Structure* (Cambridge, MA: MIT Press).

SOARES, A. P., MACHADO, J., COSTA, A., IRIARTE, Á., SIMÕES, A., DE ALMEIDA, J. J., ... PEREA, M., 2014, On the advantages of word-frequency and contextual diversity measures extracted from subtitles: the case of Portuguese. *Quarterly Journal of Experimental Psychology*, **68(4)**, 680–696 (available at: http://doi.org/10.1080/17470218.2014.964271).

TANG, K., 2012, A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, **24**, 208–214 (available at: http://s3.amazonaws.com/academia.edu.documents/30975281/Tang-UCLWPL.pdf?AWSAccessKeyId=AKIAJ-56TQJRTWSMTNPEA&Expires=1448462554&Signature=5oYbWLKmwUuuuNmVh1M%2FDNU9k-MY%3D&response-content-disposition=inline%3B%20-filename%3DA_61_Million_Word_Corpus_of_Brazilian_Po.pdf).

VAN HEUVEN, W. J. B., MANDERA, P., KEULEERS, E. and BRYSBAERT, M., 2014, Subtlex-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, **67(6)**, 1176–1190. doi:10.1080/17470218.2013.850521

VEIGA, A., CANDEIAS, S. and PERDIGÃO, F., 2013, Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment.

Journal of the Brazilian Computer Society, **19(2)**, 127–134. doi:10.1007/s13173-012-0088-0

WIKIPEDIA, 2014, *Dumps of hiWiki—Wikipedia, The Free Encyclopedia* (online) (available at: http://dumps.wikimedia.org/hiwiki/latest/).

WINDSOR, J., KOHNERT, K., LOBITZ, K. F. and PHAM, G. T., 2010, Cross-language nonword repetition by bilingual and monolingual children. *American Journal of Speech–Language Pathology*, **19(4)**, 298–310. doi:10.1044/1058-0360(2010/09-0064)

WINGATE, M., 2001, SLD is not stuttering. *Journal of Speech, Language, and Hearing Research*, **44**, 381–383.

WOLK, L., EDWARDS, M. L. and CONTURE, E. G., 1993, Coexistence of stuttering and disordered phonology in young children. *Journal of Speech and Hearing Research*, **36(5)**, 906–917. doi:10.1044/jshr.3605.906

YAIRI, E. and AMBROSE, N. G., 2005, *Early Childhood Stuttering*. Austin, TX: PRO-ED.

ZEILEIS, A. and HOTHORN, T., 2002, Diagnostic checking in regression relationships. *R News*, **2(3)**, 7–10 (available at: http://CRAN.R-project.org/doc/Rnews/).

ZEILEIS, A., KLEIBER, C. and JACKMAN S., 2008, Regression models for count data in R. *Journal of Statistical Software*, **27(8)**, 1–25 (available at: http://epub.wu.ac.at/1168/1/document.pdf).