Citation: Szewczyk, J., Chiat, S., Marecka, M. & Wodniecka-Chlipalska, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: evidence from a multi-factorial analysis. Cognition, 179, pp. 23-36. doi: 10.1016/j.cognition.2018.06.002

Accepted for publication in *Cognition* (17 May 2018)

Nonword repetition depends on the frequency of sublexical representations at different grain sizes: evidence from a multi-factorial analysis

Jakub Szewczyk[1], Shula Chiat[2], Marta Marecka[1] & Zofia Wodniecka-Chlipalska[1]

[1]Jagiellonian University, Krakow, Poland
[2]City, University of London, London, UK

Abstract

The nonword repetition task (NWR) has been widely used in basic cognitive and clinical research, as well as in clinical assessment, and has been proposed as a clinical marker for Specific Language Impairment (SLI). Yet the mechanisms underlying performance on this task are not clear. This study offers insights into these mechanisms through a comprehensive examination of item-related variables identified in previous research as possibly contributing to NWR scores and through testing the predictive power of each in relation to the others. A unique feature of the study is that all factors are considered simultaneously. Fifty-seven typically developing children were tested with a NWR task containing 150 nonwords differing in length, phonotactic probability, lexical neighbourhood and phonological complexity. The results indicate that phonological processing of novel words draws on sublexical representations at all grain sizes and that these representations are phonological, unstructured and insensitive to morphemehood. We propose a novel index – mean ngram frequency of all phonemes – that best captures the extent to which a nonword draws on sublexical representations. The study demonstrates the primacy of sublexical representations in NWR performance with implications for the nature of the deficit in SLI.

Introduction

The accuracy with which people repeat non-existing, but phonologically possible words, such as *kipser*, is a remarkably good index of their language-related abilities. Performance in this very simple task, known as nonword repetition test (henceforth NWR), has proved an important predictor of novel word learning in both native and foreign language acquisition, as indexed by correlations of NWR scores with vocabulary size (Baddeley, Gathercole, & Papagno, 1998; Farnia & Geva, 2011; Gathercole & Baddeley, 1989; Gathercole, Hitch, Service, & Martin, 1997; Gathercole & Masoura, 2003; Gathercole, Willis, Emslie, & Baddeley, 1992; Jarrold, Thorn, & Stephens, 2009; Service & Kohonen, 1995). Clearly, the NWR task emulates the situation in which learners encounter new lexical items for the first time. In both cases, the learner is presented with new verbal material that needs to be perceived, processed and then repeated. Hence, finding the best predictors of NWR might help to identify key factors entailed in word-learning.

NWR task has also proved effective in differentiating children diagnosed with language disorders such as Specific Language Impairment (SLI) from typically developing children (see Estes, Evans, & Else-Quest, 2007 for a metanalysis). SLI is an umbrella term for a group of language impairments that do not result from any general cognitive deficit (Joanisse & Seidenberg, 1998; 2003; Leonard, 2014). It is characterized by a range of symptoms including problems with grammar (processing of syntactically complex sentences and applying morphological rules; Bortolini, Caselli, & Leonard, 1997), word learning (Morley, Court, Miller, & Garside, 1955; Trauner, Wulfeck, Tallal, & Hesselink, 2000) and phonology (Elliott, Hammer, & Scholl, 1989). NWR has been considered in many studies as a potential diagnostic tool for SLI (see e.g. Estes et al., 2007; Conti-Ramsden & Botting, 2001; Weismer et al., 2000) and it seems that NWR performance is directly related to the core deficit underlying SLI. Therefore, by illuminating mechanisms underlying performance in the NWR task, we should be able to tap into the nature of SLI itself.

The NWR task was initially used as a test of phonological short-term memory (henceforth pSTM) in studies of vocabulary development (Gathercole & Baddeley 1989, 1990a, 1990b). The authors of these studies argued that NWR was a purer measure of pSTM than, for instance, the traditionally used digit span, because it does not involve processing of any lexical information (Gathercole & Baddeley, 1990b). However, very early on the claim that NWR is a test of pSTM was brought into question. It was pointed out that the task might also involve phonological and articulatory skills, as well as lexical knowledge (Bowey, 1996; 2001; Snowling, Chiat, & Hulme, 1991). Even the authors of the early NWR papers conceded that the measure is a complex one and that it measures skills and knowledge beyond pSTM (Gathercole, 1995; Gathercole 2006). This has led to extensive debate about what NWR actually measures, and extensive research investigating the factors that influence NWR performance as evidence of the skills and knowledge involved.

However, the results of most previous studies have been limited by their underlying methodological approach, with few, and often just one, factor being manipulated in a factorial design. These factors are then attributed to a specific type of cognitive representation or process. Some of these factors correspond to mechanisms (phonological STM, articulatory dexterity, perceptual acuity) and some to representations (lexical and sublexical knowledge) hypothesized to underlie NWR. Each will be briefly reviewed below.

Phonological STM capacity

It has been argued that participants who can hold more information in their pSTM can maintain a temporary representation of the nonword long enough to repeat it, resulting in better NWR performance (Gathercole & Baddeley, 1989; 1990a). Participants with poorer pSTM are not able to maintain the nonword in memory and thus have problems with repeating it. The idea that NWR provides an index of pSTM capacity is supported by two classes of findings. First, participants repeat short nonwords more accurately than long nonwords which tax pSTM to a greater degree (e.g. Gathercole & Baddeley, 1989; Stokes, Wong, Fletcher & Leonard, 2006; Weismer et al., 2000). Second, NWR correlates with tasks traditionally associated with phonological memory, namely digit span and the immediate serial recall tasks (Archibald & Gathercole, 2006; Gathercole & Baddeley, 1990b).

Lexical and sublexical knowledge

Length is by no means the only factor to affect NWR performance, indicating that NWR involves more than a purely quantitative capacity. Many researchers have pointed out that long-term memory may contribute to NWR and participants' knowledge of the language might be a key factor determining accuracy of NWR (Bowey, 1996; 2001; Jones, 2011; Jones & Witherstone, 2011; Jones, Tamburelli, Watson, Gobet, & Pine, 2010; Metsala, 1999; Snowling et al., 1991). In this case, learners who are more proficient in a given language should be better at repeating nonwords which resemble real words in this language. This idea is supported by findings showing that NWR accuracy correlates with vocabulary size in L1 (Bowey, 1996; 2001; Gathercole et al., 1992; Gathercole & Baddeley, 1989) and in L2 (Gathercole & Masoura, 2003; Masoura & Gathercole, 1999; Service & Kohonen, 1995). Moreover, participants are more accurate in repeating highly wordlike nonwords (Archibald & Gathercole, 2006; Gathercole, 1995; Munson, Kurtz & Windsor, 2005). However, these findings may reflect different kinds of linguistic knowledge. Zooming in on more specific parameters of nonwords might help to pinpoint the most relevant aspects of this knowledge.

Some authors have proposed that repeating nonwords relies on lexical knowledge and that participants use phonological representations of whole word forms as an aid in temporarily representing the nonword in pSTM (Gathercole, 2006; Roodenrys & Hinton, 2002). They argue that the presentation of a nonword partially activates representations of its lexical neighbors – phonologically similar words that are already known to the learner. This enables more efficient representation of the nonword in pSTM. Phonological neighbors can also be used to repair a decaying memory trace of a particular nonword in pSTM (a process known as lexical redintegration; Brown & Hulme, 1995; Hulme et al., 1991; Hulme et al., 1997; Hulme et al., 1999; Roodenrys & Hinton, 2002; Roodenrys et al., 2002; Schweickert, 1993). The notion that lexical representations support NWR is in line with studies showing that nonwords coming from dense phonological neighborhoods are repeated faster and more accurately than those coming from sparse phonological neighborhoods (Janse & Newman, 2013; Roodenrys & Hinton, 2002; Vitevitch & Luce, 1998; 1999; 2005)[1]. It is also supported by studies showing that nonwords containing real words (e.g. *bathesis*) lead to higher repetition accuracy than nonwords that do not contain embedded words (e.g. *fathesis*; Dollaghan, Biber, & Campbell, 1993; 1995).

Nonword repetition may be supported not only by representations at the lexical level, but also by representations at a sublexical level. Sublexical representations contain phoneme combinations occurring more or less frequently in a given language. The existence of this level of phonological knowledge has been proposed by frameworks based on the Adaptive

---

[1] Note that Vitevitch and colleagues showed that lexical neighborhood density predicts the speed of nonword repetition, but they did not claim that lexical representations are necessarily used during nonword repetition.

Resonance Theory (Grossberg, 1986; Grossberg et al., 1997; Vitevitch & Luce 1998; 1999), the EPAM-VOC and CLASSIC theories (Jones, Gobet, & Pine, 2007; Jones et al., 2014; Jones, 2016), and by the Lexical Restructuring family of theories (Bowey, 2001; Metsala, 1999). The above-mentioned theories (with the exception of Lexical Restructuring) propose that the presentation of a nonword activates representations of "chunks" (or sequences) of phonemes in long-term memory that are present in this nonword. This activation, in turn, facilitates nonword repetition in similar ways to those that have been proposed for lexical representation: more efficient representation of nonwords in phonological STM and/or sublexical redintegration. The facilitative role of sublexical representations in NWR performance is supported by studies showing that nonwords with high phonotactic probability (i.e. containing combinations of phonemes that are typical for the language) are repeated faster (Vitevitch & Luce, 1998; 1999; 2005) or more accurately (Coady & Aslin, 2004; Edwards, Beckman, & Munson, 2004; Gathercole, Frankish, Pickering, & Peaker, 1999; Majerus et al., 2004; Messer, Leseman, Boom & Mayo, 2010; Munson, Edwards & Beckman, 2005; Munson, Kurtz, et al., 2005; Roodenrys and Hinton, 2002; Thorn, Gathercole, & Frankish, 2005; Zamuner, 2009; Zamuner, Gerken, & Hammond, 2004) than nonwords with low phonotactic probability. In all of these studies, phonotactic probability was indexed using phonemic bigram frequency, i.e. the mean frequency of all pairs of adjacent phonemes (or letters) occurring in the nonword.

However, phonemic bigram frequency is not the only possible measure of phonotactic probability. Sublexical representations do not necessarily consist solely of two-phoneme combinations. In fact, EPAM-VOC and CLASSIC theories (Jones et al., 2007; Jones et al., 2014; Jones, 2016) assume that language learners store sublexical representations of varied lengths and that the greater the experience with a particular language, the longer the sequences of phonemes stored. This raises the possibility that a measure of phonotactic probability taking into consideration the frequency of phoneme sequences at many different lengths (phonemic bigrams, trigrams, and so on) might be a better predictor of NWR than simple phonemic bigram frequency. Another possibility is that the most rudimentary units of sublexical representation are syllables or subsyllabic elements such as onsets or codas rather than phonemes. Syllables and subsyllabic elements are typically proposed alongside phonemes as units of speech perception and production in phonology (Côté, 2012; Zec, 2007) and some studies suggest that these units might be indeed more basic and natural than syllables (Anthony et al., 2003; Morais, Cary, Alegria, & Bertelson, 1979; Massaro, 1987; Pierrehumbert & Nair, 1995; Read, Yun-Fei, Hong-Yin, & Bao-Qing, 1986; Treiman, 1983; Ziegler & Goswami, 2005)

Articulatory difficulty

Apart from pSTM and language-specific knowledge, NWR performance may also depend on articulatory difficulty and hence oromotor dexterity, as indicated by two types of evidence. First, NWR performance is correlated with scores on tasks testing oromotor skills in which participants are asked to repeat complex oral movement sequences after a model (Krishnan et al., 2013; Stark & Blackwell, 1997). Second, nonwords containing consonant clusters, structures that are deemed to be phonologically complex and difficult to articulate, are less likely to be repeated correctly in typically developing children (Archibald & Gathercole, 2006; Estes et al., 2007 Gathercole & Baddeley, 1989; 1990) and in children with SLI (Bishop, North, & Donlan, 1996; Briscoe, Bishop, & Norbury, 2001; Leclercq, Maillart, & Majerus, 2013; Munson et al., 2005). Articulatory complexity of consonant clusters stems from the fact that they require a range of complex and rapidly changing movements, rather than the simple oscillatory pattern of closing and opening the jaw entailed in producing a simple Consonant-Vowel (CV) sequence (MacNeilage & Davis, 2000; 2005). Alternatively, the observed difficulty in repeating nonwords containing consonant clusters may simply result from their relative scarcity in the native languages of children tested in these studies (English, French), meaning that the children have had relatively less practice producing such sequences. Polish, the language of the current study, is very rich in consonant clusters (Dobrogowska, 1992) and thus provides a perfect ground for testing if repetition of nonwords containing consonant clusters remains difficult even when consonant clusters are frequent in the participant's native language.

Another measure of phonological complexity that might reflect articulatory difficulty and thereby affect NWR is adherence to sonority-sequencing rules. Sonority is a feature of speech sounds, usually defined as the degree of opening of the mouth during the production of the sound (Jespersen, 1904; Keating, 1988) or the amplitude of the speech sound relative to other sounds produced in the same conditions (Parker, 2008; Selkirk, 1984; Steriade, 1982). Sonority-sequencing theories assume that a syllable is less phonologically complex if it begins with a less sonorous sound (e.g. an obstruent) and sonority rises steadily with each phoneme until it reaches its peak at a vowel (Dziubalska-Kołaczyk, 2002; Selkirk, 1984). For example, an initial consonant cluster /sp/, which consists of two obstruents (both low in sonority) violates sonority sequencing rules and thus is more complex than an initial cluster /pl/, which consists of an obstruent (low sonority) and a sonorant (high sonority). Sequences conforming to sonority rules tend to be easier to articulate, since such sequences typically require fewer articulatory gestures. The least articulatorily complex sequences are ones that conform to a simple pattern of closing and opening the jaw with minimal additional movement (MacNeilage & Davis, 2000; 2005). Measures of articulatory complexity based on the sonority hierarchy (e.g. the number of sonority violations) have not, to the best of our knowledge, been used in research on NWR. The only exception is a study by Nimmo and Roodenrys (2002) who carried out a post-hoc analysis to determine whether phoneme sequences that conformed to the sonority hierarchy were more likely to be retained. However, the results of this study were inconclusive. Therefore, explicitly manipulating the sonority measure in an NWR study might provide information about the involvement of articulatory skills in the repetition of nonwords.

Perceptual difficulty

Phonologically complex nonwords may not only be more difficult to articulate but also more challenging to perceive. Simple *consonant+vowel* sequences tend to be more perceptually salient because vowels and consonants are acoustically distinct, which facilitates perception (Ohala & Kawasaki-Fukumori, 1997). Furthermore, vowels strengthen the acoustic cues to consonants, while two consonants occurring in sequence can mask each other's acoustic cues, especially if the sequence violates sonority-based rules and contains two obstruents (Wright, 2004; Henke, Keisse, & Wright, 2012). So, if NWR performance depends on the phonological complexity of items (indexed by number of consonants or adherence to the sonority hierarchy), this could be because the task taps articulatory dexterity and/or perceptual acuity.

Limitations of previous approaches

Clearly, repetition of a nonword entails perception of relevant details in input, temporary storage, and articulatory planning and production. Summarizing research to date, NWR tests have been proposed to tap into many different cognitive representations and mechanisms (phonological short term memory, lexical representations, sublexical representations as well articulatory dexterity, with little mention of perceptual acuity), and empirical evidence has been advanced in support of these. However, the number and diversity of proposed factors and the cognitive processes associated with these poses problems for identifying sources of NWR performance. Although previous studies have pointed to the importance of these factors in NWR performance, they have largely ignored the fact that all the item parameters influencing NWR performance are likely to be correlated with one another, and therefore in part explain common variance. For example, lexical neighborhood correlates with phonotactic probability (measured by phonemic bigram frequency), which leaves open the question which (one or both) truly helps nonword repetition (Messer et al., 2015; Metsala & Chisholm, 2010; Storkel, Armbrüster & Hogan, 2006; Vitevitch & Luce, 1998; 1999; 2005). Similarly, the presence of consonant clusters is correlated with phonotactic probability (consonant clusters are rare in English) and with subjective ratings of wordlikeness, at least in English-like nonwords (Coady & Evans, 2008; Gathercole, Willis, Emslie, & Baddeley, 1991). Likewise, length is highly correlated with neighborhood density (the shorter the word or nonword, the more neighbors it has, as exemplified in Table 2). When sequences of nonwords are repeated, the effect of neighborhood eclipses that of length (Jalberta, Neath, Bireta, & Surprenant, 2011) and it remains an open issue whether the same holds for nonword repetition.

The intercorrelation of the proposed factors highlights the first problem with previous approaches to investigating the processes of nonword repetition. Given the multitude of correlated factors, it is necessary to consider many item parameters in parallel in order to identify those that most directly explain the difficulty of repeating the nonword. However, previous studies have investigated only a few, and in most cases just one selected variable, precluding identification of the most direct predictors of NWR. Moreover, these studies often lacked sufficient statistical power to explore a larger number of item-related variables, due to the limited number of items (typically fewer than 30). Dichotomization of the selected (continuous) variable(s) further limits statistical power. In addition, the items used have often been highly specific, with length often limited to one syllable, and/or limited to one syllable structure (Gathercole et al., 1999; Messer et al., 2010; Messer et al., 2015; Roodenrys & Hinton, 2002; Thorn & Frankish, 2005; Vitevitch & Luce, 1998; 1999; Zamuner, Gerken, & Hammond, 2004; Zamuner, 2009).

A second shortcoming of previous studies is that some have evaluated a theoretical claim using indices that do not provide an adequate test of that claim. For example, support for theories positing that nonword repetition relies on sublexical representations hinged on the fact that phonemic bigram frequency predicts NWR accuracy. However, phonemic bigram frequency is a limited index of sublexical representation, providing no evidence regarding sublexical chunks greater than two phonemes. This limits evidence of the role of sublexical knowledge and implications for what is involved in NWR.

Current study

To go beyond previous studies and throw more light on key processes involved in NWR, we took a radically different approach to investigating factors influencing NWR performance. Making no a priori theoretical commitments, we set out to test the contribution of a wide array of item parameters using a much larger pool of items than those typically utilised. We targeted most of the predictors previously investigated, as well as including several new predictors that are theoretically justified (see Methods for details). The items were crafted in such a way that they spanned the entire space of possible values of the targeted item parameters.

In addition, our study targeted Polish rather than English. As Vitevitch, Chan, and Goldstein (2014) point out, testing English-based theories in languages other than English is crucial for advancing theories and evidence, and research on typologically different languages may be particularly informative. Polish provides an excellent source of evidence regarding the relative contribution of parameters such as syllable complexity (presence of consonant clusters) and phonotactic probability, because in contrast to English, consonants clusters are very frequent in Polish words. For example, English allows for 46 double consonant clusters and 11 triple consonants clusters word-initially (Trnka, 1966), while Polish allows for as many as 160 initial double clusters, around 100 initial triple clusters, as well as 20 quadruple clusters (Dobrogowska, 1992). This relates to the fact that Polish is a heavily consonantal language with 31 consonants, but only 6 oral vowels and 2 nasal diphthongs (in contrast, English has 24 consonants, but 12 vowels and 8 oral diphthongs; Roach, 2004; Gussman, 2007; Jassem, 2003). Moreover, using Polish made it particularly important to check if the effect of sublexical knowledge on NWR is moderated by the knowledge of morphology, since Polish is a morphologically rich, inflectional language and Polish nonwords are likely to contain morphemes. For instance, Polish nouns are inflectionally marked for number and case (there are seven cases in Polish) and their declension depends on gender, with six gender classes in the language (Nagórko, 2007). There is also a rich declension system of adjectives and complex system of verb conjugation, with eleven conjugation classes, that produce a wide range of inflectional morphemes.

Our approach will help uncover the most important factors influencing nonword repetition, and thereby elucidate factors shaping the processing and learning of novel words. It will also have implications for the nature of deficits in children with SLI whose performance on NWR tasks is typically compromised.

Methods

Participants

Seventy-five children were recruited from four kindergartens in Kraków, Poland. All were monolingual speakers of Polish. Parents of all children signed informed consent, and filled in a short questionnaire about parental education level, history of children's language and hearing problems, and ear infections. Ten children were excluded because they did not complete all tests (see below); six children because parents reported hearing deficits, or serious ear infections; one child because he was receiving speech therapy; and one child because of low scores on the receptive vocabulary test (< 2 SD). The remaining 57 children were selected for analysis (26 female, mean age: 5;5, range: 4;5 - 6;10). To ensure that children had no hearing deficits, audiometric screening was conducted using a modified Hughson-Westlake procedure (Carhart and Jerger, 1959). All participants had hearing thresholds at or below 25 db HL for frequencies in range 1-4kHz.

Materials

Standardised tests

Children took part in two standardised tests: a test of receptive vocabulary (OTS-R; Haman & Fronczyk, 2012), and a test of nonverbal IQ (Columbia Maturity Scale; Burgemeister, Blum & Lorge, 1972; Polish adaptation: Ciechanowicz, 1990). Both tests were administered according to the instructions in the test manuals. Descriptive statistics for raw scores (used in the analyses) and normalised scores on the tests are given in Table 1.

	range	mean	SD
OTS-R (raw, max 88)	30 - 84	63.1	12
OTS-R (sten)	2 - 9	5.8	1.9
Columbia (raw, max 67)	24 - 47	36.7	5.3
Columbia (sten)	2.7 - 7.9	5.9	1.3

Table 1. Descriptive statistics for raw and normalized results of the test of receptive vocabulary (OTS-R) and of nonverbal IQ (Columbia Maturity Scale). Sten scores have a mean of 5.5 and SD of 2.

Construction of the nonword repetition test

The NWR test was constructed based on parameters of Polish, derived from the National Corpus of Polish using the balanced subcorpus of about 250 million words (Przepiórkowski et al., 2012). Only words with a frequency above 0.1 per million were used for computing these corpus statistics. The following sections describe how the parameters were calculated (1-3), how the nonwords were generated (4), and procedures for the final selection and presentation of nonwords (5-7):

1. Phonologisation of the corpus

All words occurring in the corpus were automatically converted to a phonological form, taking advantage of the nearly perfect orthographic transparency of Polish, with each letter corresponding to one phoneme of Polish (Grzegorczykowa et al., 1998). In addition, a special character ('#') was added to the onset and offset of each word (see below).

2. Segmentation of words into chunks of varying grain-size

All words were automatically segmented into chunks of three different grain sizes: phonemes, syllables, and subsyllabic elements. Subsyllabic elements were obtained by splitting each syllable into onset, nucleus and coda. For example, segmentation of the word #klarnet# (English: clarinet) using the three methods would yield the following results: for phonemes: #-k-l-a-r-n-e-t-#; for syllables: #-klar-net-#; for subsyllabic elements: #-kl-a-r-n-e-t-#. # symbols denoting the beginning and end of the nonword were necessary to capture positional frequency of syllables, subsyllabic elements and phonemes.

3. Computation of statistical structure (ngrams) of Polish words

For each grain size (phonemes, subsyllabic elements, syllables), we computed frequencies for all possible substrings. For example, the word #dom# divided into phonemic ngrams contributed to frequency counts of the following ngrams: #d, #do, #dom, do, dom, dom#, om, om#, m#. Each word contributed to frequency counts of sequences proportionally to its frequency in the corpus. For example, because the word "dom" occurred in the corpus 34537 times, this value was added to the frequency counts of all above phonemic ngrams. This procedure was repeated for all words in the corpus. From now on, for brevity we will refer to the sequences of any size as ngrams (phonemic ngrams, syllabic ngrams, etc.), while sequences with a defined length will be referred to as bigrams, trigrams, and so on.

4. Generation and selection of maximally varied nonwords

A pool of 3500 2-4-syllable nonwords was generated by putting together randomly selected phonemic or sub-syllabic bi- and trigrams. In order to generate maximally varied nonwords, for some nonwords the generation algorithm was biased to draw more frequent combinations of ngrams, while for other nonwords it was biased to pick less frequent combinations of ngrams. Some nonwords were generated by putting together phonemes, others by putting together subsyllabic elements. In this way, nonwords varied with respect to their proximity to the source language. All generated nonwords were phonologically legal in Polish. Legality was defined as the requirement that all subsyllabic bigrams had to occur in the corpus of subsyllabic chunks.

For all nonwords, three ngram parameters were computed, representing the averaged log-frequencies of all phonemic, subsyllabic, and syllabic bigrams in a nonword. A subset of 180 nonwords was then selected, with 60 at each length (2-, 3- and 4-syllable). Within each length, the nonwords were selected to be representative of the entire pool of n-syllabic nonwords, preserving the variability in the pool of 3500 created to contain maximally varied nonwords.

5. Recording and final selection of nonwords

The 180 nonwords represented in orthographic form were read with a standard Polish accent by a professional actress and digitally recorded to a computer audio file. The recording took place in a sound-attenuated chamber using a Rode M3 condenser microphone. Each nonword was spliced and placed in a separate file. To ensure that the stimuli did not differ in volume, all nonword files were equated for overall RMS amplitude using the Audacity program.

All stimuli were later screened by the experimenter for sound artifacts, and for pronunciation inaccuracies. This led to the selection of a final set of 150 nonwords (51 2-syllable, 51 3-syllable, and 48 4-syllable).

6. Assessment of wordlikeness

For all 150 nonwords we obtained wordlikeness ratings from 30 adult native speakers of Polish. Rating was carried out in two phases. First, the raters passively listened to all nonwords presented in a randomised order via headphones in order to familiarise them with the variability of the nonwords. The raters were then asked to listen to the nonwords again and assess each on wordlikeness on a scale from 1 to 5 (1 - unlike a Polish word; 5 - this would make a perfect Polish word). The mean ratings for each nonword were then used as one of the predictors of nonword repetition accuracy.

7. Creation of the final experimental procedure

The 150 nonwords were divided into 3 lists of 50 nonwords. The nonwords in each list were matched with respect to length in syllables, and phonemic, sub-syllabic and syllabic average log-frequencies of all neighbouring pairs of chunks. Each child was tested with two of the three lists of the NWR task.

The nonword repetition test was embedded in a game in which children learnt an alien language. Each list of nonwords was wrapped into a PowerPoint presentation. The presentation pictured an alien in a flying saucer who does funny things after every five nonwords. The nonwords were presented in the order of expected difficulty, such that shorter nonwords occurred earlier, and within a given length in syllables, nonwords with higher chunk frequencies preceded those with lower chunk frequencies. This was to avoid poorer performing children being discouraged by difficult items at the beginning of the test. After 25 nonwords there was a break in the presentation.

Procedure

Children were tested individually in a quiet room in their kindergarten, in two sessions held on separate days. In the first session children received the Columbia Mental Maturity test, the audiometric test, and one nonword repetition list. In the second session, they received the receptive vocabulary test and a second nonword repetition list. All children's repetitions were recorded on a digital voice recorder. After each session children were given a sticker with an alien.

The lists were balanced across children, so that in total, all lists were used an equal number of times. Six children were tested in one session only: they took part in all accompanying tests but repeated only one list of nonwords (data from these children is included in the analyses). After the exclusion of participants (see Participants section), across children, the lists were used 41, 38 and 31 times, for the first, second and third lists respectively[2].

Scoring

Nonword repetitions of each child were independently assessed by two trained judges. The judges were instructed to transcribe the utterances in orthographic form with maximum faithfulness to the original, and to score whether the repetition was accurate or not. We chose whole-item scoring, because clinically it is considered more appropriate and sensitive in discriminating between typically developing children and children with SLI (Roy & Chiat, 2004; see comparison of scoring methods in Boerma et al., 2015). Children’s repetitions were scored correct if they included all and only the target phonemes in the correct order; addition/deletion/substitution counted as errors, but not substitution of a phoneme with a substandard allophonic variant. The two transcriptions were then reviewed by a third judge. The task of the judge was to review all instances where the two transcriptions of a nonword repetition did not match, and to adjudicate which one was correct. As a final step, the judge determined whether the child had developmental articulation problems with some sounds (e.g. substituting /r/ with [l] or [j]), and amended the score for those repetitions which were correct apart from developmental articulation problems. In the same way, the third reviewer also eliminated errors resulting from minor regional variations in accent. In the final analyses, only the binary scoring correct / incorrect was used.

Predictors and statistical analyses

Our approach in this study was fully exploratory, making no assumptions concerning the influence of item-related factors that might influence nonword repetition. That is why we set out to test as many plausible item-related predictors of nonword repetition as possible. The item-related variables included:

- Three indices of length and/or information load:
    - number of vowels (syllables) - the most frequently used index of nonwords' length
    - number of consonants
    - duration of the target nonword (as presented to the child) in milliseconds
- Two measures of lexical neighbourhood:

---

[2] While in traditional analysis such imbalance in the number of children repeating the three sets of items could be a problem, this is not an issue for mixed-effects models which are robust to unbalanced designs.

  - PLD20 - the average Levenshtein distance (the minimal number of substitutions, insertions, or deletions to be made in order to edit one string of any length into another) to the 20 closest phonological neighbours
  - Coltheart's N - the number of phonological neighbours at Levenshtein distance equal to 1
- Six indices of phonotactic probability at various levels of granularity:
  - mean log frequency of phonemic bigrams, i.e. sequences of 2 adjacent phonemes (corresponding to the most widely used index of phonotactic probability)
  - mean log frequency of subsyllabic bigrams, i.e. sequences of 2 adjacent subsyllabic elements (e.g. onset and vowel)
  - mean log frequency of syllabic bigrams, i.e. sequences of 2 adjacent syllables
  - mean log frequency of all phonemic ngrams - the mean was obtained by summing all log-frequencies of all ngrams (ngrams not occurring in the corpus were counted as zero), and dividing it by the number of ngrams occurring in the nonword (equal to n(n-1)/2, where n is the length of the nonword)
  - mean log frequency of all sub-syllabic ngrams
  - mean log frequency of all syllabic ngrams
- Three indices of phonological complexity:
  - number of consonant clusters
  - maximum length of a constituent consonant cluster
  - number of sonority violations (Shariatmadari, 2006). Since there is no widely accepted hierarchy of sonority (Selkirk, 1984; Ohala 1992), we used one that is quite broad, and in our intuition works well for Polish: voiceless obstruents < voiced obstruents < nasals < liquids and glides. Any consonant pair within a syllable onset that had decreasing sonority, or any consonant pair within a coda that had increasing sonority counted as a violation of sonority

Although not central to our analysis, we also tested a few participant-related predictors typically controlled in NWR studies. These included:

- age
- raw score on the receptive vocabulary test
- raw score on the Columbia Mental Maturity Scale
- sex
- parents' education level.

A note on collinearity

	n cons	n vowels	phon bigram freq	ssyl bigram freq	syl bigram freq	phon ngram freq	ssyl ngram freq	syl ngram freq	n coltheart	PLD 20	max cluster len	n cluster	sonority	word like ness
n vowels	0,52													
phon bigram freq	-0,32	-0,04												
ssyl bigram freq	-0,58	0,00	0,77											
syl bigram freq	-0,36	-0,36	0,32	0,37										
phon ngram freq	-0,75	-0,64	0,62	0,65	0,49									
ssyl ngram freq	-0,79	-0,58	0,57	0,71	0,45	0,97								
syl ngram freq	-0,76	-0,49	0,49	0,65	0,44	0,87	0,90							
n coltheart	-0,54	-0,43	0,30	0,43	0,30	0,66	0,69	0,61						
PLD20	0,86	0,70	-0,39	-0,48	-0,38	-0,83	-0,82	-0,76	-0,51					
max cluster len	0,71	0,12	-0,31	-0,70	-0,32	-0,50	-0,61	-0,60	-0,43	0,52				
n cluster	0,88	0,19	-0,40	-0,73	-0,30	-0,63	-0,71	-0,72	-0,42	0,69	0,80			
sonority	0,41	0,08	-0,42	-0,60	-0,29	-0,42	-0,44	-0,41	-0,19	0,39	0,45	0,44		
wordlikeness	-0,41	-0,47	-0,06	0,05	0,06	0,38	0,38	0,40	0,18	-0,55	-0,23	-0,33	-0,06	
recording len	0,88	0,69	-0,32	-0,45	-0,41	-0,78	-0,78	-0,74	-0,54	0,86	0,57	0,70	0,32	-0,44

Table 2. Correlation matrix for all tested predictors.

As can be seen in Table 2, there is considerable collinearity between the item-related variables. Given that most of the variables stem from some aspect of lexical phonology, this is to be expected. Phonotactic probability and indices of lexical neighbourhood are correlated, because words similar to many other words necessarily include frequent chunks of phonemes. Likewise, a correlation between lexical neighbourhood and length is to be expected, because shorter words tend to come from dense neighbourhoods (the space of possible short words is small, and thus it is densely packed; see section '*Limitations of previous approaches*' in the Introduction). Thus, the high intercorrelation between the different theoretically motivated variables is an intrinsic property of this set of variables. This problem has to be dealt with in some way, because each of the closely correlated variables warrants qualitatively different theoretical conclusions. We assumed that, although the variables to a large extent explain common variance, identifying those that best explain the NWR variance would enable us to infer which representations and/or mechanisms are critical for NWR. Entering such correlated variables into a regression leads to the situation in which a variable's estimate reflects its contribution while taking into account the contribution of all other correlated variables. As a result, the predicted shared variance is distributed across all correlated variables. Moreover, building models that are based on (too) many variables also straightforwardly leads to overfitting. To circumvent these problems and to distill a minimal set of variables that maximise model fit we used a backwards stepwise regression strategy. An alternative approach to dealing with multicollinearity, for example using principal component analysis, would mask possible differences between variables that could be informative about nonword processing.

Data analysis

The data were analysed using generalised linear mixed-effects regression, with logit link function, using the lme4 package in R (Bates, Maechler, Bolker & Walker, 2015). Because of the number of predictors, it was impossible to test the maximal model, including all fixed effects and full random effects structure. Therefore, we first fitted an anti-conservative model containing all fixed effects, with random effects limited to the by-participant and by-item random intercepts. Then, we reduced this model via backwards stepwise selection procedure: we removed predictors one by one, until exclusion of any predictor would result in a decrease in the Akaike Information Criterion by value smaller than 2 (a decrease in AIC reflects improvement in a model's goodness of fit, while penalizing excessively complex models). Then we found the best random effects structure (following the recommendations by Bates, Kliegl, Vasishth, & Baayen, 2015) using stepwise forward regression for this model. After extending the random effects structure, the fixed effects were established anew using backward regression. The final model did not require adjusting the random effects (removing or adding random slopes did not improve the model fit). All predictor variables were centered before conducting the analyses (however, they were decentered for visualization in Figure 1). Thus, the intercept of the model corresponds to the overall mean log-odds of accurately repeating a nonword.

Results

On average, children correctly repeated 63.7% of the nonwords (range: 16%-85%, SD: 15.0%). On average, items were correctly repeated by 64.5% of the children (range: 6.7% - 97.3%, SD: 20.4%). This shows that the demands of the items were appropriate to the children's level of phonological development, with no floor or ceiling effects. At the same time, items varied widely in their overall difficulty, affording variance necessary for estimating predictors of repetition accuracy.

The best model contained fixed effects for four item-related factors (number of consonants, mean log frequency of all phonemic ngrams, sonority violation index as measured by violations of the sonority scale, wordlikeness), and one participant-related factor (raw score on the receptive vocabulary test). It also included by-subject and by-item random intercepts, and a by-subject slope for the mean log frequency of all phonemic ngrams. The fixed effects of the model are given in Table 3. The partial effects of each of the predictors are shown in Figure 1.

Fixed effect	**Estimate**	**SE**	**z**	**p**
Intercept	0.72	0.11	6.5	< .0001
N consonants	-0.21	0.07	-3.0	< .01
Sonority violations	-0.38	0.16	-2.4	<.05
Phonemic ngram frequency	0.33	0.12	2.7	<.01
Wordlikeness	0.27	0.12	2.4	<.05
Receptive vocabulary	0.03	0.01	5.1	< .0001

Table 3. LMM estimates (on the logit scale) of fixed effects in the final model.

Random effect	**SD**	**Correlation parameters**
by-Item		
Intercept	0.74	
by-Participant		
Intercept	0.66	
Phonemic ngram Frequency	0.25	-0.86

Table 4. LMM estimates of random effects in the final model. The correlation parameter reflects correlation between by-participant intercept and slope for Phonemic ngram frequency.

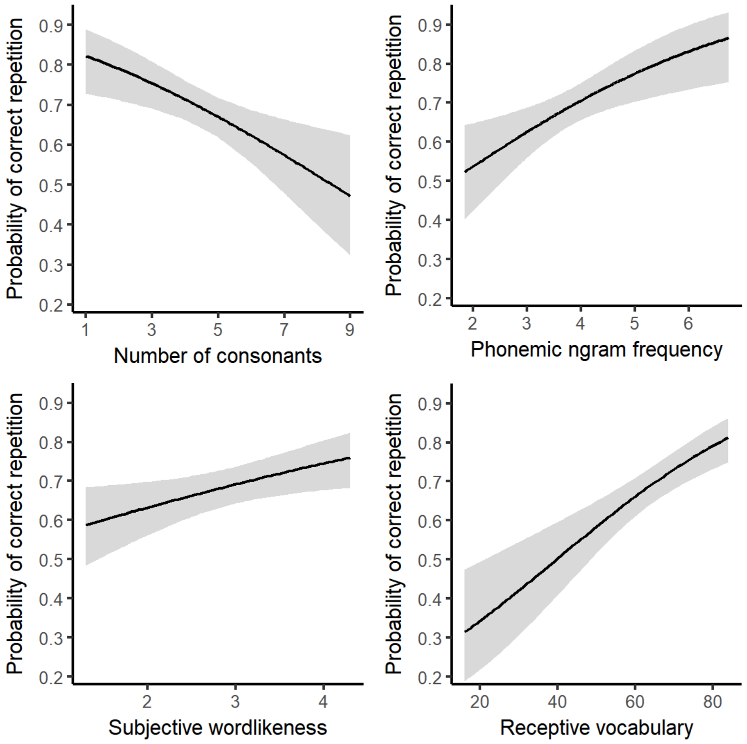

Figure 1. Partial effects in the model predicting likelihood of correct repetition of nonwords. Shaded areas reflect 95% confidence bands. Note that the values on the Y-axis are transformed to the probability scale.

The two most influential item-related predictors were phonemic ngram frequency and number of consonants. The nonword with the most frequent ngrams had 1.6 higher log-odds of correct repetition, than the nonword with the least frequent ngrams. The log-odds of repeating a nonword with 9 consonants were lower by 1.7, relative to repeating a nonword with 1 consonant. The two remaining item-related fixed effects were weaker: the full effect of Wordlikeness increased log-odds of correct repetition by 1.1, while increasing the number of sonority violations by 2 decreased log-odds of correct repetition by 0.8. Importantly, these four factors overshadowed all other variables tapping into phonotactic probability and lexical neighbourhood.

Out of all the participant-related predictors, receptive vocabulary score best explained variance in nonword repetition accuracy, overshadowing other predictors, such as age, intelligence or parents' educational level.

As we can see in Table 4, children varied significantly with respect to their sensitivity to phonemic ngram frequency of nonwords. This sensitivity strongly and negatively correlated with the overall by-subject mean, indicating that the better a child was at repeating nonwords, the less susceptible he or she was to ngram frequency. This presumably reflects a simple ceiling effect: in the extreme cases when a child repeated almost all nonwords correctly, nonwords' ngram frequency by necessity no longer played any role.

Having established that receptive vocabulary was the strongest participant-related predictor, while consonant number and ngram frequency were the most important item-related predictors, we tested whether there were any significant interactions between the participant-related predictor and the item-related predictors. The likelihood ratio comparisons showed that there were not: neither a model that included the consonant number by receptive vocabulary interaction term, nor a model that included the ngram frequency by receptive vocabulary interaction term was better than the original model not including these interactions.

Likewise, we checked that the contribution of the significant predictors was linear. We built a series of alternative models in which we added quadratic versions of the significant predictors. None of these models was better than the original model.

Fixed effect variance explained by the predictors of the best model

There is no widely agreed method of estimating the goodness-of-fit of (generalized) mixed effects model (see Bolker et al., 2017). To estimate the individual contribution of each effect identified above to predicting nonword repetition accuracy, we used the method described by Nakagawa and Schielzeth (2013) and Johnson (2014) as implemented in sem.model.fits in the piecewiseSEM package in R (version 1.2.1, Lefcheck, 2016). This method does not have the many theoretical problems associated with other definitions of R^2 for mixed effect models (e.g. negative R^2 values). It extracts marginal and conditional R^2 components of the mixed effects model. Here we focus on marginal variance that indexes variance explained by fixed effects only. For model comparison we used constant random effects structure that was the same as in the best model described above. We tested marginal variance explained when each individual effect was entered into the model alone. This analysis showed that receptive vocabulary explains 0.037 of total marginal variance. For item-related effects, the most predictive were phoneme ngram frequency and number of consonants, explaining .105 and .096 of marginal variance, respectively. When combined, these two effects explained .118 variance. Subjective wordlikeness and number of sonority violations explained much less variance when entered individually (.035 and 0.036, respectively). In total, all item-related variables explained .125, while the full model, including receptive vocabulary size, explained .162 of total marginal variance.

Further explorations of the ngram frequency factor

Since phonemic ngram frequency was the most important predictor of all variables representing phonotactic probability and lexical neighbourhood density, we took a closer look at this factor. We addressed two questions: first, whether phonemic ngrams of all lengths contributed to the predictive value of this variable, and second, if all constituent ngrams contributed to nonword repetition accuracy, or perhaps only those that corresponded to morphemes existing in Polish.

We addressed the first question by fitting a series of mixed effects models. Each model consisted of the same effects (fixed and random) as the original best model but with the phonemic ngram frequency fixed effect replaced by a variant of the effect based on ngrams of a fixed length: bigrams, trigrams, up to 8+-grams (ngrams of length 8 and more). We then compared this series of models with two baseline models: the original best model (containing the full phonemic ngram frequency effect), and a model without any ngram effect included.

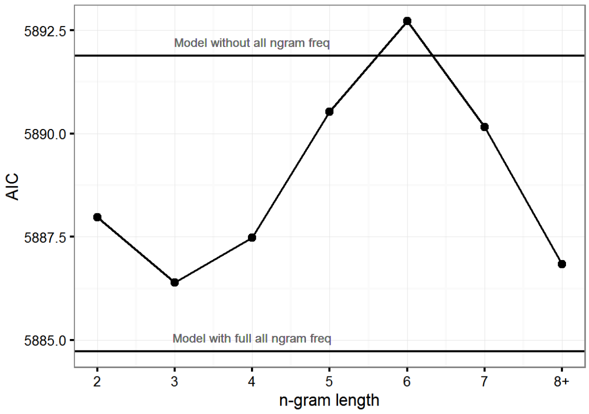

Figure 2. AIC of the model as a function of length of ngrams based on which ngram frequency was computed. The lower the AIC value, the better fit of the model. Since there were very few 8-, 9- and 10-grams, they were collapsed under the 8+ category.

Figure 2 shows the results of these analyses. As can be seen, with the exception of one ngram length (discussed below), models with phonemic ngrams of all fixed lengths improved accuracy of the model, relative to a model without any ngram effect included. This confirms that the sublexical representations helping in repetition of nonwords are not restricted to only short or only long ngrams. Only 6-grams (and correlated 5- and 7-grams) seem not to be particularly useful. Our working assumption is that this result is due to sampling error and is an idiosyncratic property of our set of nonwords and participants, but this will have to be verified in future research. Our analysis showed in addition that no model based on phonemic ngrams of fixed length was a better fit than the model containing ngrams of all lengths (i.e. the original model). This confirms that ngrams of all lengths cooperate in supporting nonword repetition.

The second question was whether phonemic ngrams contribute to nonword repetition regardless of their morpheme status in Polish. Because Polish is a morphologically rich language (see Introduction) and because the procedure for nonword generation based on the distribution of ngrams in Polish, many of our nonwords containing easily identifiable lexical units, as well as inflectional and derivational affixes. Therefore, nonword fragments (ngrams) whose form is consolidated in long-term memory and associated with some conceptual representation may be entirely responsible for the explanatory effect of phonemic ngrams. To test whether morpheme status mattered, we extracted all 3006 unique phonemic ngrams occurring in our nonwords that had non-zero frequency. Next, we hand-marked whether each of them corresponded to a morpheme or not. We followed an inclusive strategy, classifying as morphemes inflectional, derivational and lexical morphemes. Moreover, we marked as morphemes ngrams corresponding to inflected stems (root + affix), as well as ngrams corresponding to sole affixes, and sole word roots. We reasoned that if the predictive power of ngram frequency is not due to morpheme status, then even under this inclusive strategy, the effect of non-morphemic ngrams should still remain significant.

This procedure resulted in 18% of the 3006 unique non-zero frequency ngrams being marked as morphemic. After classifying the ngrams, for each nonword we computed separate frequency scores for morphemic and non-morphemic ngrams (their sum equal to the original ngram frequency score). Next, we modified the original model best predicting nonword repetition by substituting the ngram frequency fixed effect (and corresponding random slopes) with these two new scores. The resulting model is shown in table 5.

Effect	Estimate	SE	z	p
Intercept	0.79	0.11	7.1	<.0001
N consonants	-0.21	0.07	-3.0	<.01
Articulatory difficulty	-0.37	0.16	-2.3	<.05
Ngram frequency non-morphemic	0.38	0.13	2.9	<.01
Ngram frequency morphemic	0.34	0.18	1.9	.057
Wordlikeness	0.26	0.11	2.3	<.05
Receptive vocabulary	0.04	0.01	5.0	<.0001

Table 5. Fixed effects in the LMM model with phonemic ngram frequency split into morphemic and non-morphemic ngrams.

First and foremost, this analysis shows that the effect of non-morphemic ngram frequency remained as a significant contributor to the model. Moreover, the two predictors – morphemic and non-morphemic ngrams – have very similar estimates of effect size (though the effect of morphemic ngram frequency turns out to be non-significant, presumably due to the relatively low number of morphemic ngrams). Thus, it seems that ngrams contribute to nonword repetition performance independently of whether they are or are not associated with some conceptual representation or grammatical function.

Considering statistical power

Testing many predictors in parallel inevitably brings the question whether there is enough statistical power in the dataset to detect all theoretically interesting effects.

When considering statistical power, one usually establishes the size of the effect and its variability from previous studies. Using these parameters, one determines the number of participants/items that are required to detect potential effects. In our case, this approach was impossible, because no prior study considered all our predictors jointly. Most of the predictors that we considered are highly correlated and this leads to a reduction in their effect sizes when the correlated predictors are entered into the model simultaneously. In other words, the sizes of the effects estimated alone are much larger than when these effects are considered with other correlated predictors.

An alternative approach to power analysis is testing whether our best model would be capable of detecting an additional effect besides the ones identified, if the effect had size that would still be theoretically interesting to consider. One effect which was not present in the best model, but whose statistical significance would certainly bear on the interpretation of our findings, is lexical neighborhood size. We carried out Monte Carlo simulations using the powerSim function from the simr package in R (Green & MacLeod, 2016) to check what effect size would make neighbourhood density detectable with a decent power (when the model already contains all other predictors of the best model). Simulations revealed that the PLD20 effect size of .45 would be detectable with over 80% power. For our sample of children, this effect size means that increasing PLD20 by half of its interquartile-distance (by 0.8) would result in an increase of average repetition accuracy from 65% to 72% (with other predictors centered at their mean and held constant). Given that such a small effect would still be detectable by our model with a high probability, we conclude that we have sufficient power to reveal all relevant effects. Our failure to observe significant predictors besides the ones discussed most likely resulted from the fact that none of these could explain any further variance.

Discussion

In this study, we explored the mechanisms and representations underlying performance in the NWR task by testing the relationship between the scores on this task and potentially relevant item- and participant-related factors. Many of these factors have been considered in previous research and literature, but usually only individually. In this study, we explored them simultaneously, and making no a priori assumptions, in a large set of nonwords crafted in such a way as to span parameter space as broadly as possible.

We found four critical item-level predictors: number of consonants, phonemic ngram frequency, number of sonority violations, and wordlikeness, as well as one participant-level predictor, receptive vocabulary. We will now discuss, in the order of their importance to the model, how each of these predictors might contribute to our understanding of the processes behind nonword repetition and, more broadly, behind phonological processing of novel words.

Phonemic ngram frequency

Our data show that one of the most important item-level predictors of accurate nonword repetition is the mean log frequency of phonemic ngrams. This index reflects the corpus frequency of any and all phonemic sequences contained in a particular nonword (phonemic bigrams, trigrams, etc.). It is the only corpus-based statistic that remained in the final model. The primacy of this factor over all other corpus-derived factors is novel and interesting for two reasons. First, it shows the superiority of ngram frequency with respect to lexical variables, suggesting the contribution of sublexical representations to novel word processing. Second, it demonstrates that other indices of phonotactic probability at different grain sizes were less capable of explaining NWR variance, which suggests that sublexical representations consist of chunks of all sizes. Below we will discuss each of these observations in more detail.

Lexical vs. sublexical processing

In previous studies, bigram frequency and neighborhood density were sometimes pitted against each other as predictors of NWR to clarify whether novel word processing draws on lexical or sublexical representations. These studies showed mixed results: some found that lexical neighborhood better explained NWR performance when phonotactic probability was controlled (Janse & Newman, 2013; Roodenrys & Hinton, 2002), while others found that both factors explained unique variance in NWR scores (Bailey & Hahn, 2001; Thorn & Frankish, 2005). Strong support for the role of sublexical representations in nonword repetition came from a study showing that listeners exposed to an artificial language with particular phonotactic patterns found it easier to repeat nonwords conforming to these patterns (Majerus et al., 2004). An even more complex picture was painted by studies looking at predictors of novel word acquisition. These have shown that phonotactic probability and neighborhood density might be connected with different aspects of novel word learning, with lexical neighborhood facilitating only later aspects of a word’s consolidation (Storkel & Lee, 2011; Storkel, Armbruster, & Hogan, 2006).

The present study was particularly suited to adjudicate between lexical and sublexical predictors. First, it included multiple predictors of both types of representation, giving them the best chance to reveal their contribution. Second, it employed a much larger sample of nonwords than typically used, which enabled us to manipulate the corpus-derived predictors in their full range. In contrast, many previous studies used nonwords with restricted variance in sublexical or lexical predictors which might well have distorted their results in favor of one or other type of representation (c.f. Thorn & Frankish, 2005). Because our new index of phonotactic probability (i.e. phonemic ngram frequency) turned out to be one of the best predictors of the NWR scores (and better than any of the lexical neighborhood measures we tested), our results strongly support the notion that sublexical representations have a critical role in nonword repetition, over and above lexical representations.

Grain size of sublexical representations

Among all indices of sublexical patterns in Polish, phonemic ngram frequency (in contrast to bigram frequency) turned out to be the best predictor. This result suggests that processing of a nonword is sensitive not only to phonemic bigram frequency, which is generally used as the index of ‘phonotactic probability’ (see e.g. Coady & Aslin, 2004; Edwards, Beckman, & Munson, 2004; Gathercole, Frankish, Pickering, & Peaker, 1999; Majerus et al., 2004; Messer, Leseman, Boom & Mayo, 2010; Munson, Edwards, & Beckman, 2005; Munson, Kurtz, et al., 2005; Roodenrys & Hinton, 2002; Thorn, Gathercole, & Frankish, 2005; Zamuner, 2009; Zamuner, Gerken, & Hammond, 2004); rather, it is influenced by frequency of chunks at all grain sizes. Our analyses revealed that of all individual grain-sizes of phonemic ngrams, trigrams (chunks of three phonemes) explained the most variance, but the best results are achieved when chunks of all grain sizes are considered jointly (as defined by the index of phonemic ngram frequency). Furthermore, the fact that frequency of phonemic ngrams was superior to indices relying on syllabic structure (frequencies of syllabic and sub-syllabic ngrams) implies that the phonological chunks stored in long-term memory are organized around phonemes. The fact that syllabic and sub-syllabic ngram frequency did not show up in the model suggests that syllables and sub-syllabic elements do not play any part in speech processing based on sublexical representations. Conversely, sublexical representations comprise any frequently occurring combination of phonemes and do not depend on syllabic boundaries. Similarly, there is no difference in explanatory power between the frequency of phonemic ngrams containing existing morphemes (content or functional) and those that do not contain any morphemes. This contradicts previous claims that nonwords containing morphemes are easier to repeat (Dollaghan et al., 1993; 1995; Archibald & Gathercole, 2006). The present study shows that this may be a corollary of the fact that morphemes constitute frequent phonemic ngrams and thus have well entrenched phonological representations.

Perception, maintenance in STM or production?

Our finding that NWR is related to sublexical representations at different grain sizes is in line with several theoretical and computational models of speech perception positing the existence of chunk-level representations. First, it supports the adaptive resonance theory, according to which speech recognition is based on the recognition of speech chunks of different lengths, where all known chunks of phonemes compete to represent the perceived phonological input in working memory (Grossberg, 1986; Grossberg, Boardman, & Cohen, 1997; Vitevitch & Luce, 1999). Our findings are also in line with computational models of speech perception by Jones and colleagues: EPAM-VOC and CLASSIC (Jones et al., 2007; Jones et al., 2014; Jones, 2016). These models demonstrate that with greater exposure to a given language, learners learn increasingly larger chunks of phonemes that repeatedly occur in the input. Thanks to this, when confronted with a novel word (from the learner perspective: a nonword), they can encode it in short-term memory using relatively few longer chunks instead of many short chunks or separate phonemes. The longer the chunks, the less learners are limited by their short-term memory capacity, because maintaining a few longer chunks strains STM capacity to a lesser extent than maintaining a greater number of shorter chunks.

The above two theories imply that sublexical representations support nonword perception and its maintenance in pSTM. However, the predictive power of phonemic ngram frequency can be also explained from the perspective of speech production: participants found it easier to repeat nonwords with high ngram frequency because they have practice with producing frequent combinations of phonemes. In other words, just as exposure to frequent combinations of phonemes is likely to generate perceptual patterns facilitating speech recognition (sublexical representations consisting of chunks of phonemes), the frequent production of certain combinations of phonemes is likely to foster generation of articulatory patterns (motor programs) that facilitate speech production. In fact, it is likely that both the perceptual and articulatory stages of nonword repetition contribute to performance in the task (Jones & Witherstone, 2011) and that both perceptual and articulatory representations are sublexical in nature.

Vocabulary

Our study did not aim to test a comprehensive set of participant-related variables (in contrast to item-related variables) and we included only the variables most frequently investigated in research on NWR. Findings on the effects of these participant-related variables mesh with our findings on item-related variables in interesting and informative ways.

The only significant participant-level predictor of NWR performance was receptive vocabulary. This measure eclipsed all other participant-level predictors: age, IQ, sex and parents' education level. While the last three predictors served merely as control variables, age is known to significantly predict children’s accuracy in nonword repetition (Gathercole, Willis, Baddeley, & Emslie, 1994). In the current dataset, age would also be a significant predictor if it were not for the presence of receptive vocabulary in the model. This implies that, at least in 4-8-year-old children, age explains variance in nonword repetition accuracy because it is a proxy for vocabulary size – or any other construct that is tapped by receptive vocabulary.

This raises the question of what makes receptive vocabulary such a good predictor of NWR. On the one hand, the lack of a lexical neighborhood effect in our model (see above) suggests that lexical representations do not contribute to accuracy of nonword repetition. On the other hand, vocabulary size, which directly measures the richness of lexical representations, does have a significant effect. This apparent paradox can be resolved by assuming that vocabulary size, as well as estimating the repertoire of lexical representations, is a proxy for the repertoire and the entrenchment of sublexical representations. In typically developing children, lexical representations (vocabulary) grow hand in hand with sublexical representations. This effect has most recently been demonstrated by Jones and Rowland (2017), who performed a simulation of vocabulary acquisition, NWR and novel word learning in children using the CLASSIC computational model. They found that feeding more diverse input into the model, i.e. input containing more varied types of words, led to a greater increase in vocabulary and novel word learning performance than merely increasing the amount of input, because it resulted in a wider repertoire of sublexical representations. Having more sublexical representations in turn improved the efficiency of encoding novel words. All this suggests that in our linear mixed effects models, a single underlying construct – sublexical representations – is indexed by two predictors, one participant-level (receptive vocabulary) and one item-level (ngram frequency).

Interestingly, however, these two predictors did not interact in our models. The absence of such interaction indicates that the effect of phonemic ngram frequency on NWR scores was the same regardless of receptive vocabulary. In other words, all children benefited equally from high-ngram nonwords and were equally set back by low-ngram nonwords, with no ceiling or floor effects. Or, taking a different perspective, children with greater receptive vocabulary were better at repeating nonwords of any ngram frequency (i.e. nonwords with both frequent and infrequent ngrams). This suggests that children in our study knew all the ngrams to some extent, but differed in the strength of entrenchment of the ngrams in phonological long-term memory. If children with the smallest vocabularies did not have the representations of the least frequent ngrams, we would obtain an interaction between receptive vocabulary and ngram frequency. For nonwords with the rarest ngrams, the effect of receptive vocabulary would be flatter, because very few children would have the representations of these ngrams. Similarly, for children with the smallest vocabularies, the effect of ngram frequency would be flatter, because these children would know fewer ngram representations which they could use to process nonwords. Because this was not the case, we conclude that sublexical representations (phonemic ngrams) are not acquired in an all-or-nothing fashion (an ngram is either stored or not), but rather that the development of sublexical representations is a gradual phenomenon, where all ngrams are represented to some extent, but ngram representations vary in their strength. At the same time, it should be noted that in this study we tested children in the age range 4;5 – 6;10. Had we tested younger children, it is likely that we would have obtained an interaction of ngram frequency and vocabulary, as demonstrated by studies testing younger children (Coady & Aslin, 2004; Zamuner, 2009)[3].

---

[3] Note that the absence of vocabulary by mean ngram frequency interaction cannot be attributed to restricted range of vocabulary size values. Our participants demonstrated a full range of vocabulary sizes, with the mean and variance typical for the population (see Table 1).

Number of consonants

Apart from phonemic ngram frequency and receptive vocabulary, the third most important predictor of NWR performance was the number of consonants. Length-like measures were traditionally considered a proxy for information load and were taken to imply the involvement of phonological STM capacity in NWR: since longer nonwords strain pSTM capacity to a greater degree, they are more difficult to repeat (Gathercole & Baddeley, 1990). This interpretation is consistent with previous NWR studies, which have found that participants are better at repeating shorter than longer nonwords, defined by number of syllables (Gathercole et al., 1991; Stokes et al., 2006; Weismer et al., 2000, see also Gathercole, 2006 for a review). In the current study, we tested many length-like indices, including number of consonants and duration of the target nonword in milliseconds, as well as number of syllables (equal to the number of vowels). Out of these indices only number of consonants remained as a significant predictor.

This raises the question why the number of consonants, but not the number of vowels or phonemes[4], was most predictive of NWR accuracy. We believe that this result might be due to idiosyncratic characteristics of Polish – the native language of the participants and the language on which the nonwords were modelled. Polish is a highly consonantal language, with 31 consonants and only 6 vowels (Gussmann, 2007; Jassem, 2003). Vowels therefore carry a very low information load (see Fenk-Oczlon, 2001) and serve primarily as the background for the consonants. It is safe to assume that differentiating consonants constitutes the bulk of phonological perceptual or articulatory difficulty in Polish, and this may account for number of consonants emerging as the best predictor of NWR out of all measures of length. Thus, this interpretation does not attribute the effect of consonant number to the maintenance of nonwords in STM, but rather to perceptual and motor planning phases of nonword repetition. If perceiving and producing each consonant comes with a risk of misperceiving or misarticulating it, the higher the number of consonants, the higher is the chance that the nonword will be repeated incorrectly. This does not preclude that STM capacity also affects NWR performance, but it shows that it is impossible to unequivocally attribute the effect of number of consonants to perception, maintenance or articulatory phases of NWR.

Another consideration related to consonants is that phonological complexity, defined as the number or length of consonant clusters, did not explain any NWR variance on top of consonant number and phonemic ngram frequency. Traditionally, consonants occurring in clusters are considered more phonologically complex, and claimed to be both difficult to produce (MacNeilage & Davis, 2000; 2005), and to perceive (Wright, 2004). A number of NWR studies have proposed that nonwords containing complex consonant clusters are more difficult to repeat, particularly in children with language impairments (Archibald & Gathercole, 2006; Estes et al., 2007; Bishop et al., 1996; Gathercole & Baddeley, 1989; 1990a). However, they tested this effect in languages where consonant clusters are relatively

---

[4] We did not explicitly test the number of phonemes in our models, this factor being redundant since our models already included the number of consonants and the number of vowels. However, the fact that the number of vowels did not contribute as a significant predictor indicates that the number of consonants is a better predictor than the number of vowels. We also confirmed this conclusion in an alternative model where the number of vowels was substituted with the number of phonemes.

infrequent, raising the possibility that clusters' infrequency made them difficult to repeat[5]. Polish offers a good testing ground because consonant clusters are very frequent in Polish. Our results show that when phonemic ngram frequency and the number of consonants are held constant, consonant clusters play no role in determining repetition accuracy. It suggests that in those languages in which consonant clusters are less frequent, nonwords containing clusters may be more difficult to repeat than nonwords not containing consonant clusters precisely because these clusters are rare. This does not detract from the possibility that consonant clusters may be difficult to learn, as suggested by studies showing that they are acquired relatively late across the languages (Demel, 1987; McLeod, van Doorn & Reed, 2001; Tamburelli, Sanoudaki, Jones, & Sowinska, 2015). However, our Polish data suggest that, once mastered, their difficulty may boil down to their frequency in the language.

The substantial contribution of number of consonants bears on a controversy around the word length effect. This effect refers to the observation that longer verbal stimuli (in particular, stimuli that take longer to pronounce) are more difficult to retain in phonological short-term memory than shorter ones. This effect constituted evidence for time-based decay, a defining property of the phonological loop component in Baddeley's theory of working memory (Baddeley, 1992). However, this idea has recently been challenged: it has been shown that at least for some stimulus sets, nonword length is confounded with neighborhood size (short nonwords tend to have many lexical neighbors), and when the latter is controlled, the effect of length disappears (Jalberta et al., 2011). The present study suggests that the conclusion by Jalberta et al. may be premature. Models including number of consonants are superior to models including neighborhood size, when phonemic ngram frequency is also controlled. On the other hand, it is number of consonants that matters and not the actual length of the nonwords in milliseconds (as some researchers have previously proposed, e.g. Lipinski & Gupta, 2005, or as assumed in Baddeley's theory of working memory, Baddeley, 1992), which might be an argument against the time-based nature of decay in phonological short-term memory.

To sum up, our results align with previous research demonstrating the contribution of length-like indices in determining the difficulty of NWR. However, this finding cannot be unambiguously interpreted as evidence that phonological STM capacity affects the accuracy of NWR, since the number of consonants affects the number of opportunities to make a mistake in perception or articulation as well as increasing recall load. Finally, it should be borne in mind that the predictive power of number of consonants in our final model can be explained by the properties of Polish phonology and may not extend to more vowel-heavy languages.

Sonority

Another item-related factor that predicted variance in the NWR task was the number of sonority violations within a particular nonword. This index did not explain much additional marginal variance on top of the other item-related predictors. This effect suggests that while novel word repetition is based primarily on the application of known, language-specific sublexical representations (indexed by ngram frequency), there is also a more universal factor of phonological complexity that influences the mechanism. This factor might be related to

---

[5] Alternatively, the finding that consonant clusters predict nonword repetition could also be explained by the fact that those studies, in contrast to our study, did not include the index of consonant number. In either case, this suggests that the sole fact that consonants occur in a cluster does make a nonword more difficult to repeat.

articulatory difficulty of the sequence and/or its perceptual salience. Sequences of consonants that violate sonority sequencing generalizations are difficult to produce, because they involve additional articulatory gestures which interrupt the jaw cycle (MacNeilage & Davis, 2000; 2005). Moreover, sonority-violating sequences might be less perceptually salient. For example, if an obstruent is followed by another obstruent, the two sounds mask each other's acoustic cues (Wright, 2004; Henke, Keisse, & Wright, 2012).

Our finding that number of sonority violations predicted NWR accuracy over and above effects of phonemic ngram frequency suggests that some speech sequences are simply more difficult to perceive and articulate, regardless of how frequently they occur in a particular language. This suggests that perceptual and articulatory difficulty do affect the processing and repetition of nonwords. This is consistent with previous research showing that NWR performance correlates with oromotor dexterity (Krishnan et al., 2013) and that NWR appears to reflect both perceptual and articulatory processes (Jones & Witherstone, 2011).

An alternative explanation for the role of sonority in nonword repetition is related to the entrenchment of sublexical representations. Chunks of phonemes that do not conform to the sonority hierarchy may be more difficult to store in learners' phonological long-term memory (Nimmo & Roodenrys, 2002; Lee and Goldrick, 2008). As a result, these sequences might be less accessible during NWR and hamper performance.

Wordlikeness ratings

The last factor that turned out to be significant in our model is wordlikeness. Unlike other indices reflecting congruence of the nonword to lexical phonology in the language, this index was based on human judgment, rather than being derived from the corpus. It is not clear what criteria drive raters in their wordlikeness judgments, but there is a chance that they take into account criteria that we failed to capture using the corpus-based indices. However, wordlikeness ratings explained only a little of the total marginal variance on top of other predictors. This reassures us that the previously identified predictors adequately explain why some nonwords are more difficult to repeat than other nonwords, with human ratings adding little to the predictions. It is possible that the additional variance explained by the wordlikeness factor pertains to subtle differences in the prosody of the recorded nonwords, a factor that was not controlled in our study, but to which the participants of the wordlikeness judgment study could be sensitive.

Understanding Specific Language Impairments

As we indicated in the introduction, nonword repetition is considered one of the most sensitive markers of SLI in children. Determining factors that do versus do not affect accuracy of NWR in typically developing children may have implications for understanding the nature of the deficit in children with SLI who attain poor scores in NWR. Previous studies demonstrated that NWR tests employing items highly resembling the participants' native language are better than tests employing less language-like items in differentiating children with SLI from typically developing children (Archibald & Gathercole, 2006; Estes et al., 2007). These studies showed that both groups of children take advantage of the nonwords' similarity to their language and consequently repeat them more accurately than less language-like items, but children with SLI benefit less from the nonwords' resemblance to actual words, performing more similarly when repeating high- and low-language-like nonwords, compared to typically developing children.

Our study revealed that language-likeness of a nonword reflects the degree to which the nonword draws on participants' sublexical representations: items with high ngram frequencies were easier to repeat because our participants knew the chunks of phonemes building up these nonwords. All in all, this suggests that children with SLI might have less developed sublexical representations, implying a deficit in extracting and learning sublexical patterns from input. This hypothesis is in line with recent findings suggesting that children with SLI may be slower, relative to TD children, to exploit the statistical structure of input to facilitate extracting and memorizing frequently occurring chunks of phonemes (Evans, Saffran, & Robe-Torres, 2009; Haebig, Saffran, & Ellis Weismer, 2017). Such a deficit could be construed in terms of previous suggestions that the primary deficit in SLI (or at least in the subpopulation who attain low scores on NWR) is phonological in nature and all remaining deficits (e.g. problem with acquisition of grammar in English) are a corollary of the phonological deficit (Chiat, 2001; Joanisse & Seidenberg, 1998).

Conclusions

The current study is, to the best of our knowledge, the most comprehensive investigation of item-related predictors of NWR to date, and makes a number of novel contributions to the field. First, we propose a new index of sublexical support a nonword receives – average phonemic ngram frequency – which emerged as a much better indicator of nonword difficulty than other existing indices of phonotactic probability and lexical familiarity. We propose that this index should be used to assess phonotactic probability instead of the traditional bigram frequency. Second, by showing the primacy of average phonemic ngram frequency in explaining NWR performance, we provide compelling evidence that sublexical representations are crucial for accurate nonword repetition. This stands in contrast to previous studies which argued for the primacy of lexical neighborhood over phonotactic probability. In fact, we have demonstrated that phonemic ngram frequency is the most effective index of phonotactic probability and that it can explain apparent effects of lexical factors. Third, we show that sublexical representations consist of chunks of phonemes of all grain sizes, that they are phonological, unstructured and insensitive to morphemehood. These findings provide crucial information about the basic representational units used in the repetition (and thus presumably processing) of novel words. Taken together, these results are important not only for the interpretation of NWR performance, but also for broader understanding of how novel words are perceived, maintained in STM and, consequently, learned. Moreover, they have implications for understanding the key deficit in children with SLI who perform poorly on NWR, suggesting that these children have difficulty in acquiring sublexical representations. Such a difficulty will necessarily have repercussions for learning words and other aspects of language that are heavily reliant on precise phonology (such as morphosyntax).

Acknowledgements

This work was initiated as a part the COST Action ISO804. We thank Magdalena Smoczyńska for consulting material preparation, Ewa Haman for sharing details of the vocabulary test, and Marta Białecka-Pikul for providing support in accessing pre-schools and kindergartens to recruit participants. The construction of the nonword repetition test as well as data collection were supported by a subsidy from the Foundation for Polish Science to Z.W. During the work on the manuscript, J.S. and Z.W. were partially supported by the grant

SONATA BIS awarded by the National Science Center Poland awarded to Z.W. (Sonata BIS grant 015/18/E/HS6/00428).

Bibliography

Anthony, J. L., Lonigan, C. J., Driscoll, K., Phillips, B. M., & Burgess, S. R. (2003). Phonological sensitivity: A quasi- parallel progression of word structure units and cognitive operations. *Reading Research Quarterly*, *38*(4), 470–487. http://doi.org/10.1598/RRQ.38.4.3

Archibald, L. M. D., & Gathercole, S. E. (2006). Nonword repetition: a comparison of tests. *Journal of Speech, Language, and Hearing Research*, *49*(5), 970–983. http://doi.org/10.1044/1092-4388(2006/070)

Baddeley A. 1992. Working memory. *Science.* 255:556–59

Baddeley, A. D., Gathercole, S. E., & Papagno, C. C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591. http://doi.org/10.1006/jmla.2000.2756

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword Repetition as a Behavioural Marker for Inherited Language Impairment: Evidence From a Twin Study. *Journal of Child Psychology and Psychiatry*, *37*(4), 391–403. http://doi.org/10.1111/j.1469-7610.1996.tb01420.x

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children Learning Dutch as a Second Language. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1747–1760. http://doi.org/10.1044/2015_JSLHR-L-15-0058

Bolker, B. (2017). Dealing with quasi-models in R. *Compare*, *1*, 5-452305.

Bortolini, U., Caselli, M. C., & Leonard, L. B. (1997). Grammatical Deficits in Italian-Speaking Children With Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *40*(4), 809–820. http://doi.org/10.1044/jslhr.4004.809

Bowey, J. A. (1996). On the Association between Phonological Memory and Receptive Vocabulary in Five-Year-Olds. *Journal of Experimental Child Psychology*, *63*(1), 44–78. http://doi.org/10.1006/jecp.1996.0042

Bowey, J. A. (2001). Nonword Repetition and Young Children's Receptive Vocabulary: A Longitudinal Study. *Applied Psycholinguistics*, *22*(03), 441–469. http://doi.org/null

Briscoe, J., Bishop, D. V. M., & Norbury, C. F. (2001). Phonological processing, language, and literacy: a comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(3), 329–340.

Brown, G. D. A., & Hulme, C. (1995). Modeling Item Length Effects in Memory Span: No Rehearsal Needed? *Journal of Memory and Language*, *34*(5), 594–621. http://doi.org/10.1006/jmla.1995.1027

Burgermeister, B., Blum, L. H., & Lorge, I. (1972) *Columbia Mental Maturity Scale.* (3rd ed.) New York: Harcourt Brace Jovanovich.

Chiat, S. (2001). Mapping theories of developmental language impairment: Premises, predictions and evidence. *Language and Cognitive Processes*, *16*(2-3), 113–142. http://doi.org/10.1080/01690960042000012

Ciechanowicz, A. (1990). *Skala Dojrzałości Umysłowej Columbia.* Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, *89*(3), 183–213. http://doi.org/10.1016/j.jecp.2004.07.004

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of nonword repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *43*(1), 1–40. http://doi.org/10.1080/13682820601116485

Conti-Ramsden, G., & Botting, N. (2001). Psycholinguistic markers for specific language impairment (SLI). *The Journal of Child Psychology and Psychiatry*, *42*(6), 741–748.

Côté, M. (2012). The role of the syllable in the organisation and realization of sound systems. In A.C., Cohn, C. Fougeron, & M. K. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 232-242). Oxford University Press.

Demel, G. (1987). *Elementy logopedii.* Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Dobrogowska, K. (1992). Word initial and word final consonant clusters in Polish popular science text and in artistic prose. *Studia Phonetica Posnaniensia, 3,* 47-121.

Dollaghan, C., Biber, M., & Campbell, T. (1993). Constituent Syllable Effects in a Nonsense-Word Repetition Task. *Journal of Speech, Language, and Hearing Research, 36*(5), 1051–1054. http://doi.org/10.1044/jshr.3605.1051

Dollaghan, C. A., Biber, M. E., & Campbell, T. F. (1995). Lexical influences on nonword repetition. *Applied Psycholinguistics*, *16*(2), 211–222. http://doi.org/10.1017/S0142716400007098

Dziubalska-Kołaczyk, K. (2002). *Beats-and-Binding Phonology*. Frankfurt/Main: Peter Lang.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The Interaction Between Vocabulary Size and Phonotactic Probability Effects on Children's Production Accuracy and Fluency in Nonword Repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421–436. http://doi.org/10.1044/1092-4388(2004/034)

Elliott, L. L., Hammer, M. A., & Scholl, M. E. (1989). Fine-Grained Auditory Discrimination in Normal Children and Children with Language-Learning Problems. *Journal of Speech, Language, and Hearing Research*, *32*(1), 112–119. http://doi.org/10.1044/jshr.3201.112

Estes, K. G., Evans, J. L., & 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. http://doi.org/10.1044/1092-4388(2007/015)

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical Learning in Children With Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 321–335. http://doi.org/10.1044/1092-4388(2009/07-0189)

Farnia, F., & Geva, E. (2011). Cognitive correlates of vocabulary growth in English language learners. *Applied Psycholinguistics*, *32*(04), 711–738. http://doi.org/10.1017/S0142716411000038

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In J. Bybee and P. Hooper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 431-448). Amsterdam: John Benjamins Publishing Company.

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*(1), 83–94.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513–543. http://doi.org/10.1017/S0142716406060383

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*(2), 200–213. http://doi.org/10.1016/0749-596X(89)90044-2

Gathercole, S. E., & Baddeley, A. D. (1990a). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*(3), 336–360. http://doi.org/10.1016/0749-596X(90)90004-J

Gathercole, S. E., & Baddeley, A. D. (1990b). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, *81*(4), 439–454. http://doi.org/10.1111/j.2044-8295.1990.tb02371.x

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 84–95. http://doi.org/10.1037/0278-7393.25.1.84

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children., *33*(6), 966–979. http://doi.org/10.1037/0012-1649.33.6.966

Gathercole, S. E., & Masoura, E. V. (2003). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory, 13*(3-4), 422–429. http://doi.org/10.1080/09658210344000323

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The Children's Test of Nonword Repetition: a test of phonological working memory. *Memory*, *2*(2), 103–127. http://doi.org/10.1080/09658219408258940

Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children’s repetition of nonwords. *Applied Psycholinguistics*, *12*(03), 349–367. http://doi.org/10.1017/S0142716400009267

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology*, *28*(5), 887–898. http://doi.org/10.1037/0012-1649.28.5.887

Green, P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution, 7(4)*, 493–498. https://doi.org/10.1111/2041-210X.12504

Grossberg, S. (1986). The adaptive self-organisation of serial order in behaviour: Speech, language, and motor control. In E.C. Schwab & H.C. Nusbaum (Eds.), *Pattern recognition by humans and machines, Vol.1: Speech perception* (pp. 197–294). New York: Academic Press.

Grossberg, S., Boardman, I., & Cohen, M. (1993). Neural Dynamics of Variable Rate Speech Categorization. *Journal of Experimental Psychology 23*(2) 481–503.

Grzegorczykowa, R., Laskowski, R., & Wróbel, H. (1998). *Gramatyka współczesnego języka polskiego. Morfologia (The modern Polish grammar Morphology)*. Warszawa: PWN.

Gussmann, E. (2007). *The Phonology of Polish*. OUP Oxford.

Haebig, E., Saffran, J. R., & Ellis Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, *47*, 487. http://doi.org/10.1111/jcpp.12734

Haman, E., & Fronczyk, K. (2012). *Obrazkowy Test Słownikowy - Rozumienie (OTSR).* Gdańsk, Poland: Pracownia Testów Psychologicznych i Pedagogicznych.

Henke, E., Kaisse, E. M., & Wright, R. (2012). Is the Sonority Sequencing Principle an epiphenomenon? In S. Parker (Ed.), *The Sonority Controversy* (pp. 65–100). Berlin, Boston: DE GRUYTER. http://doi.org/10.1515/9783110261523.65

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*(6), 685–701. http://doi.org/10.1016/0749-596X(91)90032-F

Hulme, C., Newton, P., Cowan, N., & Stuart, G. (1999). Think before you speak: Pauses, memory search, and trace redintegration processes in verbal memory span. *Journal of Experimental Child Psychology,* 25(2), 447–463.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, M., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1217–1232.

Jalberta, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 338-353.

Janse, E., & Newman, R. S. (2013). Identifying Nonwords: Effects of Lexical Neighborhoods, Phonotactic Probability, and Listener Characteristics. *Language and Speech*, *56*(4), 421–441. http://doi.org/10.1177/0023830912447914

Jarrold, C., Thorn, A. A. S. C., & Stephens, E. (2009). The relationships among verbal short-term memory, phonological awareness, and new word learning: evidence from typical development and Down syndrome. *Journal of Experimental Child Psychology*, *102*(2), 196–218. http://doi.org/10.1016/j.jecp.2008.07.001

Jassem, W. (2003). Polish. *Journal of the International Phonetic Association*, *33*(01), 103–107. http://doi.org/10.1017/S0025100303001191

Jespersen, O. (1904). Lehrbuch der Phonetik. Leipzig: Druck und Verlag von B. G. Teubner.

Joanisse, M. F., & Seidenberg, M. S. (1998). Specific language impairment: a deficit in grammar or processing? *Trends in Cognitive Sciences*, *2*(7), 240–247. http://doi.org/10.1016/S1364-6613(98)01186-3

Joanisse, M. F., & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language*, *86*(1), 40–56. http://doi.org/10.1016/S0093-934X(02)00533-3

Jones, G. (2011). A computational simulation of children's performance across three nonword repetition tests. *Cognitive Systems Research*, *12*(2), 113–121. http://doi.org/10.1016/j.cogsys.2010.07.008

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition*, *153*, 79–88. https://doi.org/10.1016/j.cognition.2016.04.017

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: the case of nonword repetition. *Developmental Science, 17*(2), 298–310. http://doi.org/10.1111/desc.12111

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: a computational model of the learning of new words. *Developmental Science*, *10*(6), 853–873. http://doi.org/10.1111/j.1467-7687.2007.00638.x

Jones, G., Tamburelli, M., Watson, S. E., Gobet, F., & Pine, J. M. (2010). Lexicality and Frequency in Specific Language Impairment: Accuracy and Error Data from Two Nonword Repetition Tests. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1642–1655. http://doi.org/10.1044/1092-4388(2010/09-0222)

Jones, G., & Witherstone, H. L. (2011). Lexical and sublexical knowledge influences the encoding, storage, and articulation of nonwords. *Memory & Cognition, 39*(4), 588–599. http://doi.org/10.3758/s13421-010-0045-0

Keating, Patricia A. 1988. The phonology-phonetics interface. In F. J. Newmeyer, *Linguistics: The Cambridge survey, Vol. I: Linguistic theory: Foundations*. Cambridge:

Cambridge University Press, pp. 281-302.

Krishnan, S., Alcock, K. J., Mercure, E., Leech, R., Barker, E., Karmiloff-Smith, A., & Dick, F. (2013). Articulating Novel Words: Children's Oromotor Skills Predict Nonword Repetition Abilities. *Journal of Speech, Language, and Hearing Research, 56*(6), 1800–14. http://doi.org/10.1044/1092-4388(2013/12-0206)

Leclercq, A. L., Maillart, C., & Majerus, S. (2013). Nonword repetition problems in children with specific language impairment: A deficit in accessing long-term linguistic representations?. *Topics in Language Disorders, 33*(3), 238-254.

Lee, Y., & Goldrick, M. (2008). The emergence of sub-syllabic representations. *Journal of Memory and Language*, *59*(2), 155–168. http://doi.org/10.1016/j.jml.2008.03.002

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods in Ecology and Evolution, 7*(5), 573-579.

Leonard, L. B. (2014). Children with specific language impairment. Cambridge, Massachusetts: The MIT Press.

Lipinski, J., & Gupta, P. (2005). Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration. *Journal of Memory and Language*, *52*(2), 171–192. http://doi.org/10.1016/j.jml.2004.10.004

MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, *288,* 527–531.

MacNeilage, P. F., & Davis, B. D. (2005). The Frame/Content theory of evolution of speech. *Interaction Studies*, *6*(2), 173–199.

Majerus, S., Van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, *51*(2), 297–306. http://doi.org/10.1016/j.jml.2004.05.002

Masoura, E. V., & Gathercole, S. E. (1999). Phonological Short-term Memory and Foreign Language Learning. *International Journal of Psychology*, *34*(5-6), 383–388. http://doi.org/10.1080/002075999399738

Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 254-283). New York: Cambridge University Press.

McLeod, S., van Doorn, J., & Reed, V. A. (2001). Consonant Cluster Development in Two-Year-Olds: General Trends and Individual Difference. *Journal of Speech, Language, and Hearing Research*, *44*(5), 1144–1171. http://doi.org/10.1044/1092-4388(2001/090)

Messer, M. H., Leseman, P. P. M., Boom, J., & Mayo, A. Y. (2010). Journal of Experimental Child Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology*, *105*(4), 306–323. http://doi.org/10.1016/j.jecp.2009.12.006

Messer, M. H., Verhagen, J., Boom, J., Mayo, A. Y., & Leseman, P. P. M. (2015). Growth of verbal short-term memory of nonwords varying in phonotactic probability: A longitudinal study with monolingual and bilingual children. *Journal of Memory and Language*, *84*(C), 24–36. http://doi.org/10.1016/j.jml.2015.05.001

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology*, *91*(1), 3–19. http://doi.org/10.1037/0022-0663.91.1.3

Metsala, J. L., & Chisholm, G. M. (2010). The influence of lexical status and neighborhood density on children's nonword repetition. *Applied Psycholinguistics*, *31*(3), 489–506. http://doi.org/10.1017/S0142716410000081

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, *7*(4), 323–331.

http://doi.org/10.1016/0010-0277(79)90020-9

Morley, M., Court, D., Miller, H., & Garside, R. F. (1955). Delayed Speech and Developmental Aphasia. *British Medical Journal*, *2*(4937), 463–467.

Munson, B., Edwards, J., & 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.

Munson, B., Kurtz, B. A., & Windsor, J. (2005). The Influence of Vocabulary Size, Phonotactic Probability, and Wordlikeness on Nonword Repetitions of Children With and Without Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1033–1047. http://doi.org/10.1044/1092-4388(2005/072)

Nagórko, A. (2007). *Zarys gramatyki polskiej (Outlining Polish grammar)*. Warszawa: PWN.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Nimmo, L. M., & Roodenrys, S. (2002). Syllable frequency effects on phonological short-term memory tasks. *Applied Psycholinguistics*, *23*(04), 643–659. http://doi.org/10.1017/S0142716402004071

Ohala, J. (1992). Alternatives to the Sonority Hierarchy for Explaining Segmental Sequential Constraints. In *Papers from the Parasession on the Syllable* (pp. 319–338). Chicago: Chicago Linguistic Society.

Ohala, J., & Kawasaki-Fukumori, H. (1997). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In S. Eliasson & E. H. A. K. Jahr (Eds.), *Studies and Monographs* (Vol. 100, pp. 343–365). Berlin: Mouton de Gruyter.

Parker, S. (2008). Sound level protrusions as physical correlates of sonority. *Journal of Phonetics*, *36*(1), 55–90. http://doi.org/10.1016/j.wocn.2007.09.003

Pierrehumbert, J., & Nair, R. (1995). Word games and syllable structure. *Language and Speech*, *38*(1), 77–114. http://doi.org/10.1177/002383099503800104

Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B. (2012). *Narodowy korpus języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Read, C., Yun-Fei, Z., Hong-Yin, N., & Bao-Qing, D. (1986). The ability to manipulate speech sounds depends on knowing alphabetic writing. *Cognition*, *24*(1-2), 31–44. http://doi.org/10.1016/0010-0277(86)90003-X

Roach, P. (2004). British English: Received Pronunciation. *Journal of the International Phonetic Association*, *34*(2), 239–245. http://doi.org/10.1017/s0025100304001768

Roodenrys, S., & Hinton, M. (2002). Sublexical or lexical effects on serial recall of nonwords? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 29–33. http://doi.org/10.1037//0278-7393.28.1.29

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(6), 1019–1034. http://doi.org/10.1037//0278-7393.28.6.1019

Roy, P., & Chiat, S. (2004). A Prosodically Controlled Word and Nonword Repetition Task for 2- to 4-Year-Olds: Evidence From Typically Developing Children. *Journal of Speech, Language, and Hearing Research*, *47*(1), 223–234. http://doi.org/10.1044/1092-4388(2004/019)

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, *21*(2), 168–175. http://doi.org/10.3758/BF03202729

Selkirk, E. O. E., Aronoff, M., & Oehrle, R. T. (1984). On the major class features and syllable theory. In M. Aronoff & R. T. Oehrle (Eds.), *Language Sound Structure. Studies in Phonology Presented to Morris Halle by His Teacher and Students* (pp. 107–136). Cambridge: The MIT Press.

Service, E., & Kohonen, V. (1995). Is the Relation Between Phonological Memory and Foreign Language Learning Accounted for by Vocabulary Acquisition? *Applied Psycholinguistics*, *16*(02), 155–172. http://doi.org/10.1017/S0142716400007062

Shariatmadari, D. (2006). Sounds difficult? Why phonological theory needs' ease of articulation'. *School of Oriental and African Studies Working Papers in Linguistics, 14,* 207-226.

Snowling, M., Chiat, S., & Hulme, C. (1991). Words, nonwords, and phonological processes: Some comments on Gathercole, Willis, Emslie, and Baddeley. *Applied Psycholinguistics*, *12*(03), 369–373. http://doi.org/10.1017/S0142716400009279

Stark, R. E., & Blackwell, P. B. (1997). Oral volitional movements in children with language impairments. *Child Neuropsychology*, *3*(2), 81–97. http://doi.org/10.1080/09297049708401370

Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Leonard, L. B. (2006). Nonword Repetition and Sentence Repetition as Clinical Markers of Specific Language Impairment: The Case of Cantonese. *Journal of Speech, Language, and Hearing Research*, *42*(2), 219–236. http://doi.org/10.1044/1092-4388(2006/019)

Steriade, Donca. 1982. Greek prosodies and the nature of syllabification. M.I.T. doctoral dissertation.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating Phonotactic Probability and Neighborhood Density in Adult Word Learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. http://doi.org/10.1044/1092-4388(2006/085)

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. http://doi.org/10.1080/01690961003787609

Tamburelli, M., Sanoudaki, E., Jones, G., & Sowinska, M. (2015). Acceleration in the bilingual acquisition of phonological structure: Evidence from Polish–English bilingual children. *Bilingualism: Language and Cognition*, *18*(04), 713–725. http://doi.org/10.1017/S1366728914000716

Thorn, A. S. C., & Frankish, C. R. (2005). Long-Term Knowledge Effects on Serial Recall of Nonwords Are Not Exclusively Lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 729–735. http://doi.org/10.1037/0278-7393.31.4.729

Thorn, A. S. C., Gathercole, S. E., & Frankish, C. R. (2005). Redintegration and the benefits of long-term knowledge in verbal short-term memory: An evaluation of Schweickert's (1993) multinomial processing tree model. *Cognitive Psychology*, *50*(2), 133–158. http://doi.org/10.1016/j.cogpsych.2004.07.001

Trauner, D., Wulfeck, B., Tallal, P., & Hesselink, J. (2000). Neurological and MRI profiles of children with developmental language impairment. *Developmental Medicine & Child Neurology*, *42*(7), 470–475. http://doi.org/10.1111/j.1469-8749.2000.tb00350.x

Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, *15*(1-3), 49–74. http://doi.org/10.1016/0010-0277(83)90033-1

Trnka, B. (1966). A phonological analysis of present-day standard English (No. 17). Hokuou Pub. Co.

Vitevitch, M. S., Chan, K. Y., & Goldstein, R. (2014). Using English as a 'model language' to understand language processing. In N. Miller & A. Lowit (Eds.), Motor Speech Disorders: A Cross-Language Perspective. Bristol: Multilingual Matters.

Vitevitch, M. S., & Luce, P. A. (1998). When Words Compete: Levels of Processing in

Perception of Spoken Words. *Psychological Science*, *9*(4), 325–329. http://doi.org/10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408. http://doi.org/10.1006/jmla.1998.2618

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, *52*(2), 193–204. http://doi.org/10.1016/j.jml.2004.10.003

Weismer, S. E., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword Repetition Performance in School-Age Children With and Without Language Impairment. *Journal of Speech, Language, and Hearing Research*, *43*(4), 865–878. http://doi.org/10.1044/jslhr.4304.865

Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically Based Phonology* (pp. 34–57). Cambridge: Cambridge University Press.

Zamuner, T. S. (2009). Phonotactic Probabilities at the Onset of Language Development: Speech Production and Word Position. *Journal of Speech, Language, and Hearing Research*, *52*(1), 49–60. http://doi.org/10.1044/1092-4388(2008/07-0138)

Zamuner, T. S., Gerken, L., & Hammond, M. (2004). Phonotactic probabilities in young children's speech production. *Journal of Child Language*, *31*(3), 515–536.

Zec, D. (2007). The syllable. In P. de Lacy (Ed.), *The Cambridge handbook of phonology*, (pp. 161-194). Cambridge: Cambridge University Press.

Ziegler, J. C., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, *131*(1), 3–29. http://doi.org/10.1037/0033-2909.131.1.3