Linguistic Frontiers • 3(1) • 54–59 • 2020
DOI: 10.2478/lf-2020-0007

Linguistic **Frontiers**

Orthographic neighbourhood in Czech language and its use in research of reading impairment

Original Study

Monika Ptáčková, Kateřina Vitásková
Palacký University Olomouc

Received: June 2020; Accepted: June 2020

Abstract: Reading is a complex function that affects our daily living. It requires the cooperation of many structures and functions and is influenced by many different factors. One of the factors influencing the word reading process may be the characteristics of the word, such as its frequency of use in the language, its length, orthographic neighbourhood, and others. These characteristics can affect the speed and accuracy of reading in readers with and without learning disability. The following paper presents some aspects of our current research focused on the influence of the properties of the word on their reading in the Czech language environment. The concept of orthographic neighbourhood will be explained in more detail including some specifics in its calculation and use in the Czech language.

Keywords: reading, orthographic neighbourhood, Damerau-Levenshtein distance, visual word recognition

INTRODUCTION

Reading is a complex process consisting of many important components, such as visual recognition, phonological and orthographic processing, and others (Grainger 2017). There is a variety of different impairments of reading ability, which can be generally divided into two main groups—developmental reading impairments (see e.g. Penolazzi et al. 2009) and acquired reading impairments (see e.g. Karanth 2003). In the present research, the authors focus mainly on the developmental reading disorders, in some classifications also called dyslexia (see e.g. Hulme, Snowling 2016). Dyslexia research involves a lot of interesting aspects including linguistic characteristics of words. It suggests how the reading process is involved in some of these word characteristics in people with or without reading impairment. It can then help us understand reading disorders better in the whole context of language acquisition.

In foreign literature, there is plenty of research on this topic focusing on the effect of word characteristics on visual word recognition and reading in people with dyslexia (see Davies, 2007) and in the typical population (see Grainger 2017). Due to the variance between languages, the results of research studies differ as well[1]. Although in the last ten years many kinds of research in various language environments have been performed, most experiments are still based on the English language and English speakers (Share 2008).

The important word characteristics include for example word length, word frequency (frequency of using the word), and orthographic neighbourhood (ON). The frequency effect generally means that reading of high-frequency words is usually faster and more correct than low-frequency words (Ghyselink et al. 2004). This is noticeable in many languages, such as English (Grainger 2017), Spanish (González-Nosti et al. 2013), etc. Ghyselinck et al. (2004) also mentioned "the cumulative frequency hypothesis", which combines the effect of frequency and age of word acquisition. Many studies focus on other

[1] For example, in Spanish (Gonzáles-Nosti et al. 2014) the effect of orthographic neighbourhood density was shown only in short words, and there was stronger length effect than in English.

variables related to word frequency, such as the age of acquisition or frequency trajectory[2] (Juhasz et al. 2019). Information about word frequency can be found in lists based on the national corpus. As far as the Czech language in concerned, the Czech National Corpus website also includes the frequency list[3].

As for the word length effect, it also represents a broad topic which has been in the centre of attention of many studies. Word length usually affects the speed and accuracy of its recognition. Word length can also change the influence of other properties, for example in a Spanish research study (Gonzales-Nosti et al. 2014), the effect of orthographic neighbourhood was significant only in a group of short words. Schroeter and Schroeder (2014) point out the fact that the word length effect decreases during life with the strongest word length effect being usually in childhood.

The orthographic neighbourhood effect is the focus of the present research and is not frequently mentioned in Czech scientific literature. The orthographic neighbourhood may be explained as the quality of a word, which describes the number of words that can be created from the word by changing a single letter (Grainger et al. 2005). This means for example that the Czech word "klíč" has 3 orthographic neighbours ("kleč", "klín" and "klít"), while the word "hmyz" has no orthographic neighbours because no existing Czech word is generated by changing one letter to another in this letter string. Some authors use the term "orthographic neighbourhood density" (Laxon et al. 2002) or "Coltheart's N" (Yarkoni et al. 2008) which refers to the same concept.

Analogically, phonological neighbourhood refers to how many words can be made by changing one phoneme to another (Clarkson et al. 2017). This is more important in languages with non-transparent orthography such as English or French, while in the Czech language (with transparent orthography) similar to German, in most cases the orthographic and phonological neighbourhood density is the same (Marian et al. 2012).

The impact of the orthographic neighbourhood on reading is examined by means of various types of tasks. For example, as used in the present research, a lexical decision task can be used in which the participant has to decide as soon as possible whether the presented letter string is an existing word or a pseudoword. In this type of experiment, reaction times and error rates are reported. Another approach is the use of technology-based methods, such as data collection from eye-tracking or fMRI during word reading (for more information see for example Schloss 2017). This method provides more comprehensive information, but at the same time the procedure is difficult and sometimes it is not possible to examine sufficient number of participants and collect sufficient data.

There are various findings concerning the orthographic neighbourhood effect on reading in foreign literature—either an inhibitory effect, no effect, or a facilitatory effect on reading and visual word recognition. For example, Weekes et al. (2006) suggests a hypothesis that the ON effect changes during development. A research study by Andrews and Hersch (2010) suggests a conclusion that in individuals with worse spelling skills, there is a higher facilitatory effect of the orthographic neighbourhood, while in individuals with good spelling skills the effect is generally inhibitory. The question is how language affects this variable. Furthermore, properties such as word length, frequency, and orthographic neighbourhood are relatively closely related because short words are generally more frequent and have more orthographic neighbours than long words (largely addressed by Gonzales-Nosti et al. (2014) in their research).

Many recent research studies suggest another approach to the "orthographic neighbourhood density" value, which is The Damerau-Levenshtein distance. As reported by Yarconi et al. (2008), it can be more exact than the traditionally used Coltheart's N. The Damerau-Levenshtein distance (DL) also reports about word similarity compared with other words in the corpus. In this way, the word is compared with all words, not only words of the same length. This value reflects the minimum number of operations needed to change the word to another one. An operation is considered substitution, addition, deletion, or transposition of one letter (two letters in the case of transposition). The resulting value is the sum of the 20 lowest values. In other words, it suggests how many operations are needed to transform the word into the twenty closest words. This means that, for example, in the Czech language the word "soused", which has 0 orthographic neighbours according to the Coltheart's N (no other word arises by changing one letter), has a DL value of 32, which is quite low as in 20 proximate words there is only 1 or 2 operations needed to change to the word "soused".

The present research uses both the Coltheart's N and LD distance, so the influence of both values and their differences can be compared.

EXPERIMENTAL PROCEDURES

The aim of this paper is to introduce a tool which the authors developed for counting the Coltheart's N and DL distance in the Czech language and propose its application in ongoing research. The research aims to answer whether or not the Coltheart's N and Damerau-Levenshtein distance can be used in reading research in the Czech language and what the specifics of its use are in the Czech orthography system.

Foreign literature includes many tools for ON calculation such as N-watch, which is a program that works with any language database (corpus). Using this programme, it is possible to identify the same characteristics as in the

[2] Frequency trajectory expresses how common is the word in childhood and adulthood.

[3] Available at: <https://wiki.korpus.cz/doku.php/seznamy:srovnavaci_seznamy>

case of the Coltheart's N only by insertion of source data (national corpus) and the list of words one is interested in (Davis 2005). Unfortunately, this tool does not work witch Czech diacritics—´ and ˇ, so it is not practicable for the purposes of the present research.

As for English research, there is also a special database that can be used for this purpose[4] which directly generates the number of ONs or other characteristics as it has its English database of source words. Similar tools exist also in some other languages (Marian et al. 2012), but as for our information there is no similar tool in the Czech language. For this reason, the authors have developed their self-designed tool which is reported by Jirásková (2019). This paper presents a revision of this tool, which was performed to make the tool more usable in future research. The following text briefly describes the tool with an emphasis on the changes made.

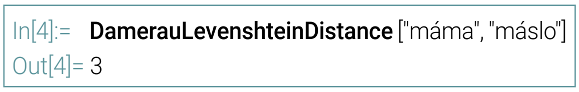
The tool is generally based on VBA for MS Excel. As the source data, the authors used the Czech National Corpus SYN2015, which is a set of about 97 million items. There are about 380,000 word tags. As presented by Jirásková (2019), the tool helped edit the source data and identify any interesting tags for the current research.

After preparation of the source data set, the authors used the second part of the tool to calculate the orthographic neighbourhood density. This tool also uses VBA in MS Excel and shows how many orthographic neighbours any word in the list has, and it also shows the list of these neighbours. This time, unlike in the previous research, the authors used the tool for all-length words, by which is it possible to identify any trends in the orthographic neighbourhood and word length.

There are some important changes in the tool that have been made since the last research (Jirásková 2019):

First, the tool Majka (Šmerk 2007) has been integrated, which can be used for a morphological analysis of word tags. The reason was that in the current research, the authors work only with nouns in their basic form. In the previous research, the list of lemmas was used to filter the basic word forms, but the authors were unable to filter only one word type. Integrating Majka in the tool makes it easier to identify and filter any word types (one or more), which can be useful also for other research studies.

Majka was also used as a good source of data, which means that any item not identified by the Majka tool was deleted. Of the total of 282,000 words, 79,472 were not identified by Majka and deleted. These mostly included interjections, typing errors in source texts, etc. The mean frequency of the deleted words was 41.5 (or 0.43 per million words), which means it is a marginal part of the word set. Using these words in the later analysis could be misleading as some of these words can be marked as orthographic neighbours, although they are only wrong entries/non-literary forms of the same word or they do not make sense in the Czech language (for example "mamí" or "bábi").



```
In[4]:=  DamerauLevenshteinDistance ["máma", "máslo"]
Out[4]= 3
```

Figure 1: Example of WolframMathematica using the Damerau-Levenshtein distance count

As for DL distance, the Wolfram Mathematica software[5] was used. In this software, the "Damerau-Levenshtein distance" function is integrated, so the number of operations can be counted for every word pair (an example is shown in Figure 1—the number of minimum operations needed to change the word "máma" to the word "máslo"). Then a simple script in Wolfram Mathematica identifies the value of the sum of 20 lowest values for every word in the corpus (so we get a recalculated value according to Yarconi et al. (2008) often used in similar research studies).

The tools produce two values related to orthographic similarity for every word in our corpus—DL distance and Coltheart's N. In the previous research, a total of 200 words with low and high orthographic similarity were selected, and these two values were merged by sorting the words according to Diagram 1.

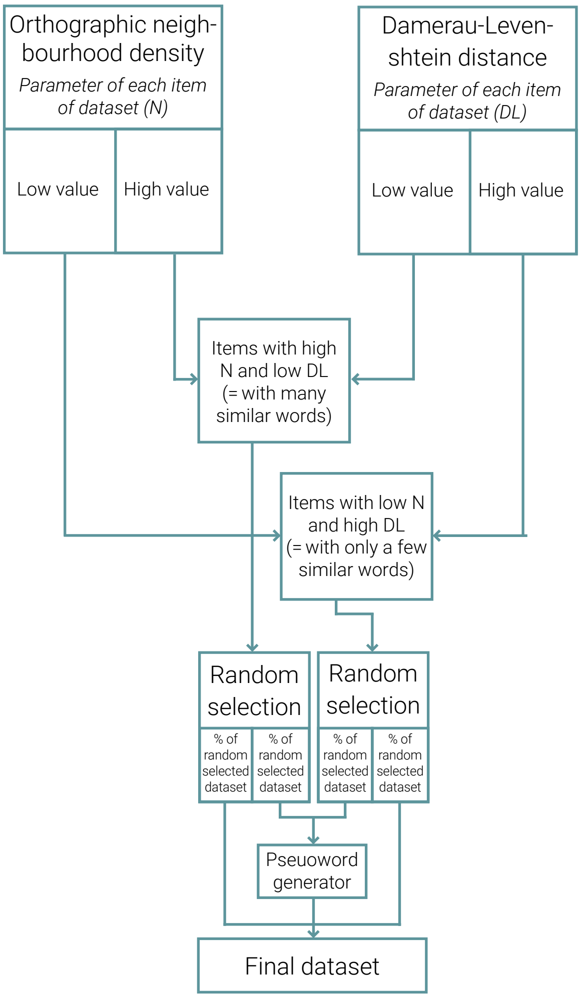

Orthographic neighbourhood density
Parameter of each item of dataset (N)
Low value
High value
Damerau-Levenshtein distance
Parameter of each item of dataset (DL)
Low value
High value
Items with high N and low DL (= with many similar words)
Items with low N and high DL (= with only a few similar words)
Random selection
% of random selected dataset
% of random selected dataset
Random selection
% of random selected dataset
% of random selected dataset
Pseuoword generator
Final dataset

Diagram 1: Words selection process

[4] like the English Lexicon Project available at: <https://elexicon.wustl.edu/query13/query13.html>

[5] Available at <https://www.wolfram.com/mathematica/>

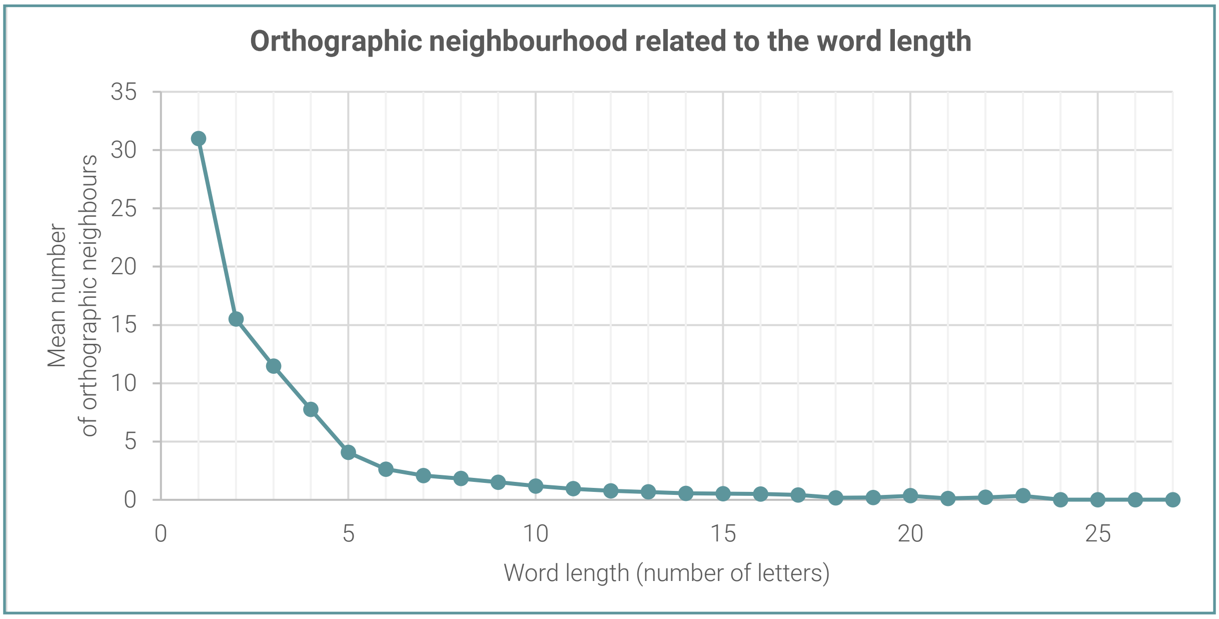


Diagram 2: Orthographic neighbourhood related to word length (low N value = low similarity of the compared word)

This means that only those words were used that had both values corresponding with the criteria of low or high orthographic similarity.

The present research focuses on a comparison of the influence of these two values in more detail. This time the authors used words that are either appropriate in DL and N or different in DL and N (such as the mentioned word "soused"). This provides an opportunity to better discuss the results concerning the explanatory value of the variables in the Czech language.

RESULTS

In the preparatory part of the present research, the authors produced and used self-designed tools for counting the Coltheart's N and Damerau-Levenshtein distance as an index of orthographic similarity in the Czech language.

Similarly to Spanish (Gonzales-Nosti et al. 2014) and some other languages, there is a tendency in the Czech language according to which short words have many more similar orthographic words than longer words. This is suggested by Diagram 2 and Diagram 3. Generally, N and DL stand for word similarity. N describes the similarity of any word pair in only two alternatives (similar or different words), whereas DL expresses this property in a natural-number value. The tendency of N and DL related to word length should be analogous, but DL provides a finer accuracy of word-pair similarity.

In groups of 3—10-letter words, the authors found words with both a low and a high level of orthographic similarity, which suggests that this index can be used as a variable in an experiment with individuals with and without a specific reading impairment.

There is a specific feature related to the Czech language system that the authors had to deal with; this is mentioned in the Discussion part.

DISCUSSION

During the design of the tool for N and DL counting, there were some ambiguities that had to be resolved. One of the specifics of the Czech language in counting orthographic neighbourhood density is the digraph "ch", which is generally considered one letter, but N counts it as two letters "c" and "h". Moreover, during optical reading "ch" is perceived as two letters, and for example in eye movement measurement it can cause some differences. For the purposes of the present research, no words containing "ch" were used for stimulations, so there was no scope for misinterpretation as a result of this. Nevertheless, in N and DL distance calculation "ch" stands for 2 letters, which means for example that the word "chata" is not an orthographic neighbour of the word "pata" although there is only one change from *ch* to *p* in the first position.

Another important thing is the inflection of Czech words. Again, in the stimulation material, the authors used always only one word form (nominative singular), but in the tools for ON and DL counting the whole corpus is inserted (including more than one word form of some words). This means that the word forms can be considered "neighbours"[6]. In fact, when the reader does not know the instruction that the stimuli are only basic word forms, for example the word "emoce" can be confused with "emocí" in the same way as "kůň" and "kůl", which have a completely different meaning. Another question is semantic processing, which may cause the word "emoce", whose orthographic neighbours are only its different forms

[6] For example the word "emoce" has two orthographic neighbours "emocí" and "emoci" which are in fact only different forms of the same word (dative singular and genitive plural).

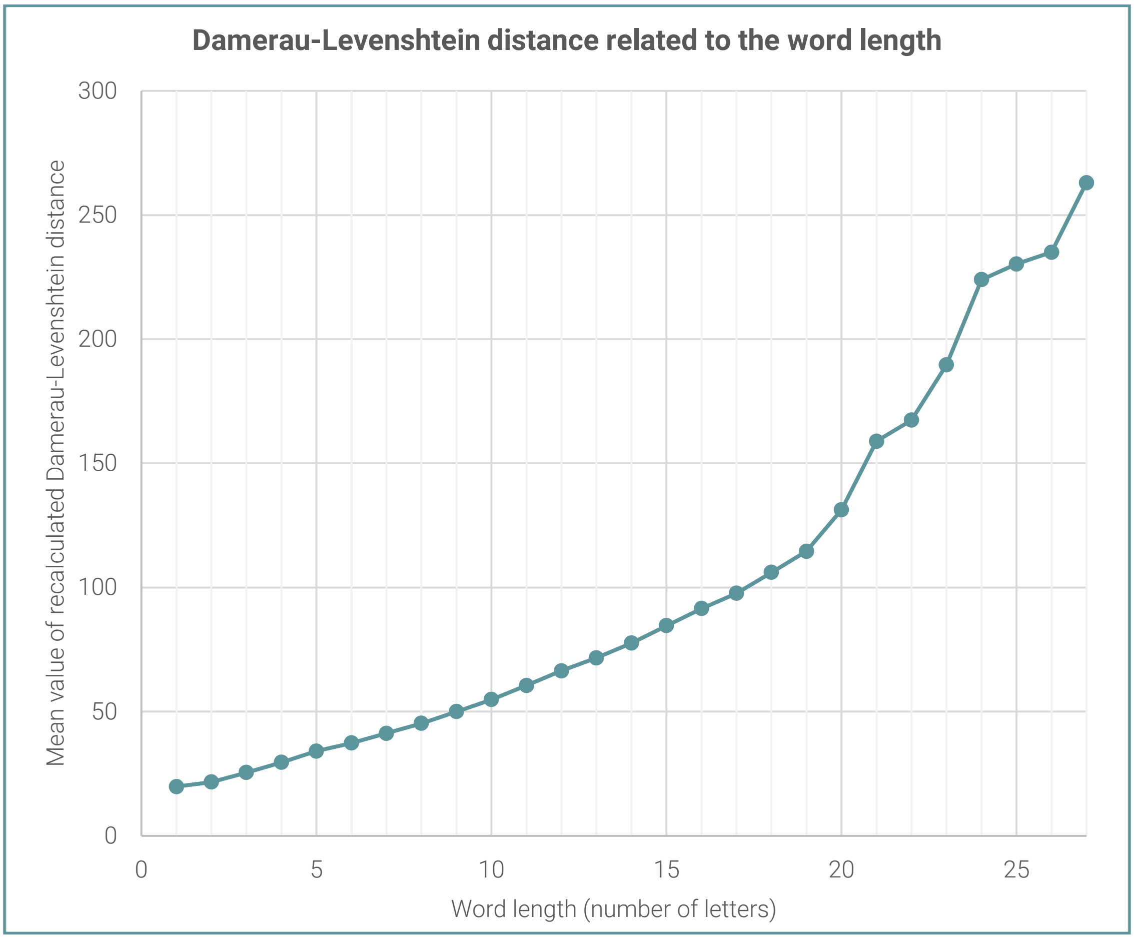


Diagram 3: Damerau-Levenshtein distance related to word length (high DL value = low similarity of the compared word)

(emocí and emoci), to be processed differently than the word “dotaz”, which also has 2 orthographic neighbours, but these are semantically different words (doraz, potaz).

The introduced tools are now being used for creating an experiment based on a lexical decision task. The experiment will use a sample of words sorted by length, frequency, and orthographic similarity. The same number of pseudowords will be produced. The effects of word properties on adults without reading impairment will be examined. The experiment will then be extended to a group of adults with a specific reading impairment and perhaps some others. As far as future research is concerned, other methods may be added, such as eye movement monitoring. It is also possible to add some other word-characteristics-related variables, such as the age of acquisition or the existence of high-frequency orthographic neighbours, which are also frequently used in foreign research studies. Implementation of this is nowadays in the centre of our attention and after realization of the experiment, data relating to influence of orthographic similarity to word reading in Czech will be presented in some other paper.

ACKNOWLEDGMENTS

The paper is based on the partial results of a specific research study IGA_PdF_2020_036 “Research of verbal and nonverbal communication, voice, speech in the context of modern speech—language assessment and therapy” (Principal researcher: Assoc. Prof. Kateřina Vitásková, Ph.D.) conducted at the Faculty of Education, Palacký University Olomouc. There is no presumption of conflict of interest in this study.

REFERENCES

Andrews, S., Hersch, J., 2010. Lexical Precision in Skilled Readers: Individual Differences in Masked Neighbor Priming. *Journal of Experimental Psychology*, 139(2), 299—318, available at: <https://psycnet.apa.org/record/2010-08363-005>.

Clarkson, L., Roodenrys, S., Miller, L. M. et al., 2017. The phonological neighbourhood effect on short-term memory for order. *Memory (Hove, England)*, 25(3), 391–402, available at: <https://www.tandfonline.com/doi/abs/10.1080/09658211.2016.1179330>.

Davies, R., Cuetos, F. Glez-Seijas, R. M., 2007. Reading development and dyslexia in a transparent orthography: a survey of Spanish children. *Annals of Dyslexia*, 57(2), 179-198, avaiable at <https://link.springer.com/article/10.1007/s11881-007-0010-1>.

Davis, C. J., 2005. N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior*

Research Methods, 37(1), 65–70, available at: <https://link.springer.com/article/10.3758/BF03206399>.

Ghyselinck, M., Lewis, M. B., Brysbaert, M., 2004. Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, 115(1), 43–67, available at: <https://www.sciencedirect.com/science/article/pii/S0001691803001070?casa_token=-ArqcT_1yl-cAAAAA:-y3eWbRpkc8laz8k3XLb-87C8A1VHEeiY-xeRXDL_isqbnXAUfItOupU9aJE38OvXLx3B2Kng-Q>.

González-Nosti, M., Barbón, A., Rodríguez-Fereiro J. et al., 2014. Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. *Behavior Research Methods*. 46(2), 517–525, available at: <https://link.springer.com/article/10.3758/s13428-013-0383-5>.

Grainger, J., 2017. Orthographic Processing: a „Mid-Level" Vision of Reading. *Quarterly journal of experimental psychology*, 71(2), 1–72, available at: <https://www.researchgate.net/publication/315802243_Orthographic_Processing_a_Mid-Level_Vision_of_Reading>.

Grainger, J., Muneaux, M., Farioli, F. et al., 2005. Effects of phonological and orthographic neighbourhood density interact in visual word recognition. *Quarterly Journal of Experimental Psychology: Section A,* 58(6), 981–998, available at: <https://www.tandfonline.com/doi/abs/10.1080/02724980443000386>

Hulme, C., Snowling, M. J., 2016. Reading disorders and dyslexia. *Current opinion in pediatrics*, 28(6), 731–735, available at: <https://doi.org/10.1097/MOP.0000000000000411>.

Jirásková, M., 2019. *Dyslexie ve speciálněpedagogickém a neurovědeckém pojetí se zaměřením na fonologické zpracování*, Diploma thesis, Univerzita Palackého v Olomouci, Olomouc, The Czech Republic.

Juhasz, B. J., Yap, M. J., Raoul, A., Kaye, M., 2019. A further examination of word frequency and age-of-acquisition effects in English lexical decision task performance: The role of frequency trajectory. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 45(1), 82–96, available at: <https://psycnet.apa.org/record/2018-17903-001>.

Karanth, P., 2004. *Cross-linguistic study of acquired reading disorders*. Springer US, available at: <https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,uid&db=edseul&AN=edseul.3000072674760&lang=cs&site=eds-live>.

Křen, M., Cvrček, V., Čapka, T. et al., 2015. *SYN2015: reprezentativní korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha, available at: <http://www.korpus.cz>.

Laxon, V., Gallagher A., Masterson, J. 2002. The effects of familiarity, orthographic neighbourhood density, letter-length and graphemic complexity on children's reading accuracy. *British journal of psychology*. 93(2), 269–287, available at: <https://onlinelibrary.wiley.com/doi/abs/10.1348/000712602162580?casa_token=TA24rf7Qj7YAAAAA:mz47UkE-jlWV4nEkr-pr5w2jzW4FyBmGp-xrHXKDgVMc1s8pA9IyzqN-40x5ArCabFPvDr05qO2sH0Lc4>.

Marian, V., Bartolotti, J., Chabal, S. et al., 2012. Clearpond: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*. 7(8), 1–11, available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3423352/>.

Penolazzi, B., Spironelli C., Angrilli, A., 2009. *Children and Language: Development, Impairment and Training*. Hauppauge, NY: Nova Science Publishers, pp. 113–145. ISBN 9781606923955.

Share, D. L., 2008. On the anglocentricities of current reading research and practice: the perils of overreliance on an „outlier" orthography. *Psychological Bulletin,* 134(4), 584–615, available at: <https://pdfs.semanticscholar.org/c063/7a40368398bdaaa93304e97d90b2c233c93a.pdf>.

Schloss, B. J., 2017. *Predicting Concept Evolution during Naturalistic Reading with Simultaneous Eye-tracking and fMRI*, Master thesis, The Pennsylvania State University.

Schroeter, P., Schroeder, S., 2014. Differences in visual word recognition between L1 and L2 speakers the impact of length, frequency, and orthographic neighborhood size in German children. *Studies in second language acquisition.* 40(2), 319–339, available at: <https://www.researchgate.net/publication/320455116_DIFFERENCES_IN_VISUAL_WORD_RECOGNITION_BETWEEN_L1_AND_L2_SPEAKERS>.

Šmerk, P., 2007. Fast Morphological Analysis of Czech. In Sojka P., Horák A. (Eds.), *Proceedings of Third Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. Brno: Masaryk University, pp. 13–16. ISBN 978-80-210-5048-8.

Weekes, B., Castles, A., Davies, R., 2006. Effects of Consistency and Age of Acquisition on Reading and Spelling among Developing Readers. *Reading.* 19(2), 133–169, available at: <https://link.springer.com/article/10.1007%2Fs11145-005-2032-6>.

Yarkoni, T., Balota, D., Yap, M., 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin* 15(5), 971-979, available at: <https://link.springer.com/article/10.3758/PBR.15.5.971>.

FIGURES

Return to front page ↑