An Investigation of Reading Development Through Sensitivity to Sublexical Units

Candidate: Valentina Nicole Pescuma

Supervisor: Davide Crepaldi
Co-supervisor: Maria Ktori

Thesis submitted for the degree of Doctor of Philosophy
in Cognitive Neuroscience

Scuola Internazionale Superiore di Studi Avanzati
Trieste, April 2021

Table of Contents

Chapter I

General Introduction

Accounts of reading acquisition and word recognition

Reading is a uniquely human ability that allows us to perform a diverse set of tasks that are essential to functioning in today's society, such as reading the news, correctly interpreting the instructions on a medicine bottle, signing a contract, or even just for ordering a meal from a restaurant menu. It is estimated that teenagers in their second year of high school can recognise around 80000 English words (Adams, 1990; Castles et al., 2007). From a developmental perspective, learning to read constitutes an important milestone as it provides children with a direct means to knowledge and education, and thus carries a widespread impact on academic performance, self-confidence, and subsequent adult-life chances. It may therefore come as no surprise that reading, and its acquisition, have always been in the limelight of psychology research (for a review, see Castles et al., 2018).

A few major computational models of reading have been proposed throughout the last decades for skilled reading, focusing mostly on reading aloud. However, when it comes to the cognitive processes underlying reading development, there is still no definitive answer as to what mechanisms are at play. The dual-route cascaded (DRC) model (Coltheart et al., 1993, 2001) is a computational model of visual word recognition and reading aloud, which features a lexical route and a non-lexical one for print-to-sound mapping. Specifically, the lexical – or direct – route entails direct access to whole-word orthographic representation, and subsequently to phonology on the one hand and semantic representation on the other. Instead, along the non-lexical – or indirect – route, print information undergoes a letter-to-sound rule procedure before the whole-word (phonological and semantic) level is accessed. This model postulates the existence of local, rather than distributed, word representations in the reading system.

A different perspective is offered by models which posit distributed word representations, such as parallel distributed processing (PDP) models of reading aloud (Seidenberg & McClelland, 1989; Plaut et al., 1996). In particular, Plaut's triangle model (Plaut et al., 1996) proposes networks in which interactive connections are established between phonological, orthographic and semantic representations. Furthermore, weights allocated to units are sensitive to the way in which the network is influenced by the "statistical structure of the environment", reflecting the degree of consistency between mappings, for item representations. Therefore, in this sense, the triangle model can account for how beginners learn to read words (see also Nation, 2009).

An essential first step to successful reading aloud, common to all major reading models, is the establishment of correspondences between print and meaning. This may happen through the mediation of phonology or – as is typical of effortless skilled reading – through direct access to meaning. What is still largely unclear is how these mappings develop, and on which units developing readers rely, in order to become skilled readers. Unfortunately, when one turns to developmental research, there still are no comprehensive models as to how reading acquisition unfolds.

A first essential step towards successful, effortless word reading is the ability to access phonological representations of words through the acquisition of letter-to-sound mappings, that is, *phonological recoding*, the core feature of Share's *self-teaching hypothesis* (Share, 1995). According to this theorization, every successful decoding instance of a new word equips the reader with some new information about the orthography of that word, such that, with even just a few encounters, a word's orthography is successfully acquired. In this sense, the ability to perform phonological recoding, in interaction with the amount of exposure to a given word, would act as a self-teaching mechanism that allows the creation of new orthographic representations. Self-teaching would proceed in a serial, letter-by-letter fashion. Importantly, it would also take place in an item-based manner, meaning that, for instance, higher-frequency words will be recognized faster and depend less on their phonology (Castles et al., 2018; Share, 1995).

While Share's hypothesis is among the most relevant about the development of skilled reading and it crucially recognizes the role of exposure, it does not fully explain how successful exposure happens (Castles et al., 2018). In fact, when one turns to consider that the ultimate goal of reading development to achieve fast and efficient reading of both regular and irregular words – by directly accessing their meaning – the self-learning mechanism does not provide much insight as to how this ability will eventually be attained. In other words, it lacks an explicit explanation of how exposure would allow to reach the so-called *orthographic* stage of word recognition, a proxy of skilled reading, at which words are processed directly as whole entities mapping onto a specific meaning (e.g., Castles et al., 2007; Rastle, 2019).

A hypothesis that has been put forward to explain how lexical representations are refined along development is the *lexical tuning hypothesis* (Castles et al., 1999, 2007), which postulates that the ability to map separate lexical entities onto different print representations – instantiated, for example, by the ability to differentiate words that differ by just a letter (e.g., *cat* vs *mat*) – determines the quality of one's lexical representation of a given word. What follows is that the quality of a reader's lexical representation is not just influenced by print exposure (i.e., reading experience), but also by factors that are inherent to the nature of the orthography, such that the number of potential competitors of a given word's orthography interacts with the amount of print exposure, hence the phrase *lexical tuning*.

In the brief theoretical excursus outlined thus far, what still remains largely underspecified is the core learning mechanism of reading acquisition and, perhaps more importantly, the linguistic units on which such learning capitalizes. An account for the role that sublexical units might play in the successful formation of phonological representation is posited by the *psycholinguistic grain size theory*. Ziegler and Goswami, in their 2005 paper, theorized that what facilitates grapheme-to-phoneme conversion is the shared *grain size* (for example letters, syllables, up to whole words) between the levels of phonology and orthography. A language with inconsistent GPC (grapheme-to-phoneme correspondence), like English, features a phonological system that relies on larger grain size, to which a relatively slow acquisition of sound-to-spelling mappings might be ascribed. Mastering GPCs allows successful phonological recoding. The psycholinguistic grain size theory thus predicts that beginning readers of a language like Italian, which features a highly transparent orthography, would easily rely on the smallest grain size (single phonemes and graphemes) for successful reading acquisition.

While this theory has the merit of bringing sublexical units into the reading acquisition picture, its primary focus is phonology. However, as noted by Castles et al. (2018) in their comprehensive review, reliance on GPC might explain sufficiently well decoding of simple, monomorphemic words, predominantly encountered in the beginning stages of reading (Masterson et al., 2010). Yet, when it comes to more advanced stages of reading development, involving encounters with complex words regularities between orthography and semantics ought to be brought into the picture as well, besides orthography-to-phonology mappings.

Morphemes are sublexical entities that represent “islands of regularity” in the language. They are the smallest units that carry meaning (e.g., a word such as farmer is composed of a stem, farm-, and a suffix, -er) and feature form-meaning consistency. Analyzing the morphological constituents of a novel word thus may allow its interpretation. It is clear how this passage supports and facilitates print-to-meaning mapping, with morphemes representing the shared grain size between orthography and semantics. Therefore, being able to access the morphological structure of words unsurprisingly represents an advantage for the acquisition of reading skills (Rastle, 2019).

A host of studies conducted over the last four decades have provided evidence of the special role that morpho-orthographic processing plays in visual word identification. The hypothesis that morphemes act as chunks whose co-occurrence regularities are captured by the reading system has been outlined in different accounts. Chunking of sublexical units, such as letter clusters forming morphemes (i.e., morpho-orthographic units), is a well-established mechanism underlying orthographic processing (see, for instance, the fine-grained code in the dual-route model described by Grainger & Ziegler, 2011). According to dual-route accounts, this chunking mechanism allows developing readers to establish a print-to-meaning mapping, which in turn enables them to access the phonological representation of an unknown word.

Specifically, as far as reading acquisition is concerned, it has been proposed that, also in Italian, which features a shallow orthography, morphemes could serve as intermediate grain-size units (between single grapheme and whole word processing) aiding lexical processing of complex words, especially in younger and less skilled readers, for whom whole-word processing is more demanding. This is supported by evidence that nonword reading aloud is faster when the nonword structure is morphological (Burani et al., 2002, 2008), whereas no differences are found in reading aloud complex vs simple words by more skilled or adult readers. It is hypothesized that this group of readers would still benefit from morphologically complex (i.e., decomposable) words, where possible, but that their whole-word processing would not be hampered either. This may appear somewhat inconsistent with the above-mentioned psycholinguistic grain-size account, proposed by Ziegler and Goswami (2005), according to which developing readers of orthographically transparent languages are expected to make use of the smallest available grain-size units.

More recently, Grainger and Beyersmann (2017) postulated that morpho-orthographic processing is prompted by the activation of edge-aligned embedded words. Stems are often free-standing words and therefore encountered as words separated by spaces. According to this account, morpho-orthographic segmentation would be initiated by the activation of stems, not as morphemes, but rather as independent whole words, with which developing readers are likely already familiar, too. On the other hand, affixes would be the only entities processed as morpho-orthographic units proper, and they would be activated when present, albeit not stripped (see Taft & Forster, 1975, on prefix-stripping, and successive more general conceptualization, such as Taft & Nguyen-Hoan, 2010). Children would first acquire morpho-semantic relations (e.g., between *farm* and *farmer*; see Beyersmann et al., 2012; Grainger & Beyersmann, 2017), while it is still debated at which stage of reading development sensitivity to morpho-orthographic relations is completed – that is, at what age sensitivity emerges to the relation between a stem and a pseudo-derived word, such as *corner* and *corn* (see a lexical decision study by Quémart et al., 2011, where priming is found for both types of prime-target pairs, as opposed to masked priming studies, such as Beyersmann et al., 2012; Hasenäcker et al., 2016, where no evidence is found for morpho-orthographic segmentation in primary school children).

Relevant work with adults has, on the other hand, established skilled readers' sensitivity to the morphological structure of words, to the point that decomposition is initiated even when pseudo-complex words are presented. The most notable findings in this respect come from a masked priming study by Rastle et al. (2004), where significant priming was found in the morphologically opaque condition (such as the pseudo-complex prime *corner* and the target *CORN*) and was comparable to the effect observed with morphologically transparent pairs (e.g., *cleaner-CLEAN*), while no significant priming effect was detected in the case of mere orthographic overlap, with a non-morphologically decomposable prime (e.g., *brothel-*

BROTH). Such findings show that skilled readers rely on morphological chunking mechanisms so largely that they tend to decompose a word into its constituent morphemes even when they only share a pseudo-morphological relation.

Overall, it is evident that a crucial aspect that must be taken into account for a theory of reading development is the role that the sensitivity to linguistic regularities plays in the formation and refinement of visual word recognition processes.

The role of statistical learning in reading acquisition

So far, the role of regularities has been described as predominantly across different linguistic levels (i.e., as orthography-to-phonology or orthography-to-semantics mappings). For instance, sensitivity to form-to-meaning mappings manifesting as early as second–third grade (see Grainger & Beyersmann, 2017; Castles et al., 2018) is regarded as essential to the refinement of morpho-orthographic processing along reading development. In general, the human cognitive system constantly extracts patterns of regularities from the environment with which it interacts. Through this ability, we learn to expect and predict specific patterns of co-occurrences or sequences of events. The ability to extract salient statistical information from the perceptual input, across domains, is termed *statistical learning* (for reviews, see Armstrong et al., 2017; Aslin, 2017; Christiansen, 2019; Frost et al., 2019; Newport, 2016). As evidenced by several studies, our visual system is particularly reliant on the ability to extract regularities, starting at a very young age; statistical learning has indeed been proposed as a mechanism that even infants use to make sense of their surroundings (Gibson, 1969; Fiser & Aslin, 2001).

Research such as that conducted by Fiser and colleagues has demonstrated the ability to extract co-occurrence regularities from visual patterns to which we are exposed, without any prior instruction or supervision of any kind. Namely, participants were able to tease apart shapes based on how they co-occurred in pairs when displayed to them (Fiser & Aslin, 2001).

In another experiment, participants showed to have formed a visual representation not of individual shapes, but only of the pairs or quadruples in which those single shapes were embedded during the familiarization phase (Fiser & Aslin, 2005). This led the authors to propose statistical learning as a bootstrapping mechanism for complex visual representations.

Sensitivity to orthographic regularities in reading development

Since the crucial role of statistical learning has been generally acknowledged in the visual domain (Fiser & Aslin, 2001, 2005; Kim et al., 2009; Orbán et al., 2008), there is no obvious reason why beginning readers would not exploit this learning mechanism, relying on the extraction of salient statistical information from written language in order to achieve skilled reading.

To explore whether readers extract statistics from frequently co-occurring letter clusters that they encounter, the effects of bigram frequency effects on reaction times have often been investigated, however yielding mixed findings. Already in 1984, Gernsbacher pointed out conflicting evidence from prior studies investigating the effects of bigram frequency on word recognition. Those studies were inconsistent when it came to establishing the significance and the nature – whether inhibitory or facilitatory – of the effects of bigram frequency in visual word processing. The debate is still open, to date. For instance, Schmalz and Mulatti (2017) used Bayes Factor to reanalyze previous literature on bigram frequency effects on reaction times in lexical decision tasks. The results of their analysis were inconsistent between the two English databases used (BLP; Keuleers et al., 2012; ELP; Balota et al., 2007), and overall suggested an absence of bigram frequency effects in lexical decision tasks, while inconsistently supporting the presence of a facilitatory effect in reading aloud.

If studies on the role of the extraction of statistical regularities in visual word processing have proven inconclusive with skilled readers, the picture is also not clear and far from complete when one turns to developmental studies.

Pacton and colleagues (Pacton et al., 2001) proposed that the extraction of patterns in the written language allows even beginning readers to capture orthographic regularities, demonstrating that French-speaking children, as early as in first grade (thus with very little print exposure), presented with nonwords, would judge as word-like those items whose letter co-occurrence and distributional statistics were compatible with those observed in real words (for instance, they would choose *illaro* over *ivvaro*, because *l* is doubled within words in French, while *v* is not), and similar findings are also reported for English (Cassar & Treiman, 1997).

Overall, in spite of general agreement (see Chetail, 2015, for a review) on the fact that orthographic regularities are extracted by readers and that sensitivity to them increases through print exposure, the role that such regularities play in visual word processing, especially in reading development, is still unclear. This is also due to the fact that most studies with beginning readers have so far employed highly refined lab-based experimental paradigms, such as Artificial Grammar Learning (AGL), Serial Recall Task (SRT) (for a review, see Schmalz et al., 2016), and implicit learning tasks with visual stimuli (e.g., Kidd & Arciuli, 2016; von Koss Torkildsen et al., 2019). Furthermore, only indirect connections between statistical learning abilities and reading skills have been established in a few studies (e.g., Arciuli & Simpson, 2012; von Koss Torkildsen et al., 2019).

Therefore, the present work aims at exploring whether and how visual word identification capitalises on statistical regularities extracted from clusters of frequently co-occurring letters, across reading development. In particular, we have used an ecologically valid approach to the study of reading using eye tracking (Chapters II and III of this thesis), as will be described later in this Introduction.

Sensitivity to morpho-orthographic regularities as a form of statistical learning

Lelonkiewicz et al. (2020) have recently provided evidence, through a statistical learning task in which affix-like chunks were embedded in pseudoword strings (in an unknown non-alphabetic script), for the extraction of affix-like chunks, which were frequently co-occurring character clusters in a given position of the string. We furthermore know that morpho-orthographic units (such as *-er* in corner) are spotted even in the absence of a semantic relationship, such that processing of a word like *CORN* is facilitated by the delivery of a masked prime like *corner* (Rastle et al., 2004; similarly for French, Longtin et al., 2003), even though *-er* in *corner* does not signify 'someone who corns'. In this respect, morphemes may very well be defined as clusters of letters that happen to display a consistent form-meaning relationship, such that their orthographic identity can be accessed even when morphological decomposition is not warranted.

The above-described findings suggest that visual information about morphemes is extracted by virtue of the fact that they represent frequently encountered instances of form-to-meaning regularities. What follows is that morphemes can be considered as a special case of frequently co-occurring letter clusters (n-grams, hereafter), and therefore as salient linguistic units.

As mentioned earlier in this Chapter, behavioural findings suggest that children are sensitive to the morphological structure of words quite early on (around Grade 3; e.g., Beyersmann et al., 2012; Hasenäcker et al., 2016). However, while the role of morphemes as reading units (i.e., morpho-orthographic processing) has been recognised in skilled readers, the developmental trajectory of morpho-orthographic processing is still underspecified, as research has yielded mixed evidence in this respect (e.g., Beyersmann et al., 2012; Dawson et al., 2018; Quémart et al., 2011).

While a few studies with French (Casalis et al., 2015; Quémart et al., 2011) and Italian primary school children (Burani et al., 2002, 2008) provide some evidence for sensitivity to nonwords with a morphological structure, other findings disagree with this. More recently, Dawson et al. (2018) investigated the emergence of sensitivity to morphemes within nonwords at different developmental stages, finding that an adult-like pattern, with lower accuracy and slower reaction times for pseudomorphological nonwords (e.g., *earist*), compared to control nonwords (*earilt)* fully emerges at a relatively late stage of development (16-17 years of age), but not in primary school children (7-9 years old) and young adolescents (12-13 years old). Studies in English and German with primary school children using a masked priming paradigm, considered to tap into automatic visual processes, reported priming effects only for morphologically related pairs (e.g., *golden-GOLD*), but not for pseudomorphologically (e.g., *mother-MOTH*) or orthographically related ones, such as *spinach-SPIN* (Beyersmann et al., 2012), or reported no difference in priming between pairs with suffixed word primes (*kleidchen-KLEID*), suffixed nonword primes (*kleidtum-KLEID*), nonsuffixed nonword primes (*kleidekt-KLEID*), compared to unrelated controls (e.g., *träumerei-KLEID*; Hasenäcker et al., 2016). However, in a similar study in French with children from Grades 3, 5 and 7, Quémart et al. (2011) found comparable priming effects from suffixed and pseudosuffixed primes, as opposed to nonsuffixed and orthographic ones, across all grades. These findings show sensitivity to morphemes in words and pseudowords as early as in third grade, providing some evidence for morpho-orthographic processing in young children.

As outlined earlier in this Chapter, in the account of the developmental trajectory of morpho-orthographic processing proposed by Grainger and Beyersmann (2017), beginning readers rely on representations of stems, which are quite familiar from early reading development stages due to their existence as free-standing words in the language. Such reliance on *embedded, edge-aligned stems* would serve as a bootstrapping mechanism for initiating morpho-orthographic segmentation. However, full morpho-orthographic processing (as indexed by the detection of affixes in pseudocomplex words) would only be completed in the final developmental stage, even though it is still debated at which point, along development, such is reached.

In conclusion, evidence is mixed with respect to morpho-orthographic processing in reading development, although it has been suggested to only fully mature at a relatively late developmental stage. Nonetheless, as mentioned, taking a statistical learning perspective, we can fairly assume that also morpho-orthographic clusters are regular units on which we capitalize to become skilled readers. In order to explore this, as described in the following section, I will present a magnetoencephalography (MEG) investigation of the neural underpinnings of morpheme identification, at two different stages of reading development: Grades 5-6 and adulthood.

Research contribution of the present work

This thesis project aims at providing an important contribution to the field by adopting a statistical learning approach to the investigation of the cognitive processes underlying visual word identification and its development, exploring natural reading on the one hand and morpheme identification in pseudowords on the other. The two main techniques used were eye tracking and magnetoencephalography. In the following subsections, I will present the four chapters comprised in this thesis.

Much of our knowledge about reading and its development comes from highly refined experimental paradigms, which hardly reflect the way in which we approach written material in our daily life. We thus aimed at carrying out an ecologically valid investigation of eye movement parameters in natural reading, across development. In parallel, we aimed at filling a gap in developmental reading research, by building an eye tracking tool based on multiline reading. In Chapter II, I present *EyeReadIt*, a database of eye movement measures recorded from a large sample (N=141) of Italian developing readers aged 8-12 (Grades 3-6) and from adult controls (N=33), as they silently read multiline passages from kids' story books for comprehension.

An analysis of developmental changes in eye movement behavior during reading, as well as of well-known linguistic effects (word length and word frequency) on eye movements, yielded comparable results with the existing literature. We thus conclude that *EyeReadIt* is a valid database, and that it represents a solid resource for different analyses, such as the effects of morphological complexity, as outlined in this Chapter. Finally, this database can also be employed to address more advanced research questions, as I will describe in Chapter IV. The database is being finalized and will be made available at this link: https://osf.io/hx2sj/

Chapter III - Algorithms for the automated correction of vertical drift in eye tracking data (with Dr Jon W. Carr)

In Chapter III, I will introduce a methodological study in collaboration with Dr Jon Carr, spurred as a further development of *EyeReadIt*. This chapter is a slightly adapted version of a manuscript accepted for publication, to appear as: Carr, J. W., Pescuma, V. N., Furlan, M., Ktori, M., & Crepaldi, D. (2021). Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods*.

We document ten algorithms – two of which novel – which allow to improve the quality of eye-tracking recordings featuring misaligned fixations, by automatically realigning these to the correct lines of text. Using both data from *EyeReadIt* and computational simulations, we evaluated the performance of several realignment algorithms (two of which were novel to this work) on correcting various phenomena related to so-called "vertical drift". Vertical drift refers to the progressive displacement of fixation registrations on the vertical axis over time and may importantly affect data analysis, especially in multiline reading experiments. We provide guidance for eye-tracking researchers in selecting and applying these correction methods to their data, while also proposing two novel algorithms. In particular, we present one of these (*warp,* based on Dynamic Time Warping) as particularly successful in realigning fixations in children's recordings. All resources have been made freely available and can be found at the following link: https://doi.org/10.17605/OSF.IO/7SRKG

Chapter IV - Eye movements during natural reading reveal sensitivity to orthographic regularities in children

In Chapter IV, I will tackle an aspect of the core of our investigation around the emergence of sublexical regularities in reading development. Since statistical learning has been acknowledged as a mechanism that our visual system exploits, we would expect that the extraction of statistical regularities from the input also occurs in reading. Furthermore, we ask whether and how reading development interacts with the sensitivity to statistical regularities.

As most statistical learning and reading-related questions have been addressed through highly artificial tasks, we intended to complement the literature by using a more ecologically valid approach; to this end, we conducted our analyses on *EyeReadIt* (presented in Chapter II). After all, if statistical cues are extracted during reading and sensitivity changes across reading development, the role that orthographic regularities play should surface even in a less controlled approach.

Here, I will present an investigation of the effects of n-gram frequency metrics on eye movement measures. Specifically, postulating that co-occurring groups of letters within words (n-grams) are used to extract statistics about written language by the reading brain, we examined the effect of different frequency metrics (minimal, average and maximal frequency) of differently-sized n-grams on commonly examined eye tracking measures (first-of-many-fixation duration, gaze duration, total reading time), and we checked whether and how their effect was modulated by reading development.

Our analysis shows that n-gram frequency effects (in particular related to maximum/average frequency metrics) are present even in developing readers, suggesting that sensitivity to sublexical regularities of words in reading is present as soon as it is possible for the developing system to pick it up – in our specific case, as early as in third grade.

Chapter V - Automatic morpheme identification across development: magnetoencephalography (MEG) evidence from Fast Periodic Visual Stimulation

As mentioned in this Introduction, and as will be extensively described in Chapter V, the course of morpho-orthographic processing along reading development is still unclear, and very little research has so far been conducted to examine its neural bases. In this Chapter, I will present an MEG investigation of selective neural responses to morphemes at different stages of reading development, carried out in collaboration with Dr Lisi Beyersmann, Prof Anne Castles and Prof Paul Sowman at Macquarie University, Sydney, Australia. Here, I present the MEG responses of 28 adults and 17 native English-speaking children (Grades 5 and 6) to the presentation of pseudowords containing morphemes, as they underwent MEG recording, using Fast Periodic Visual Stimulation (FPVS) with an oddball design.

Rapid sequences (base stimulation frequency: 6 Hz) of pseudoword combinations of stem/non-stem and suffix/non-suffix components were interleaved with oddball stimuli appearing periodically every fifth item (oddball stimulation frequency: 6 Hz/5 = 1.2 Hz), and were specially designed to examine either stem or suffix detection (e.g., stem+suffix oddballs, such as *softity*, embedded in a sequence of non-stem+suffix base items, such as *terpity*). Successful detection of morphemes would be indexed by a robust peak in the MEG response at the oddball frequency.

Sensor-level analysis was conducted both with a theory-driven approach, by defining a left occipito-temporal region of interest to map onto an area corresponding to the ventral occipito-temporal cortex, and with a data-driven one, using cluster-based permutations. Overall, both in developing and skilled readers, a successful oddball response was found in experimental conditions in which the oddball stimuli were fully decomposable pseudowords – that is, when oddballs were made up of real stems and suffixes (e.g., *softity*).

These results provide evidence for automatic morpheme identification, even at relatively early stages of reading development. Critically, they also suggest that morpheme identification can be modulated by the context in which the morphemes appear. Additional ongoing analyses aim at providing more refined source-level information, in order to help shed light on the neural underpinnings of morpheme identification in visual word processing.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286-304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: past, present and future. https://doi.org/10.1098/rstb.2016.0047

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1-2), e1373. https://doi.org/10.1002/wcs.1373

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459. https://doi.org/10.3758/BF03193014

Beyersmann, E., Castles, A., & Coltheart, M. (2012). Morphological processing during visual word recognition in developing readers: Evidence from masked priming. *Quarterly Journal of Experimental Psychology*, *65*(7), 1306-1326. https://doi.org/10.1080%2F17470218.2012.656661

Burani, C., Marcolini, S., De Luca, M., & Zoccolotti, P. (2008). Morpheme-based reading aloud: Evidence from dyslexic and skilled Italian readers. *Cognition*, *108*(1), 243-262. https://doi.org/10.1016/j.cognition.2007.12.010

Burani, C., Marcolini, S., & Stella, G. (2002). How early does morpholexical reading develop in readers of a shallow orthography?. *Brain and Language*, *81*(1-3), 568-586. https://doi.org/10.1006/brln.2001.2548

Casalis, S., Quémart, P., & Duncan, L. G. (2015). How language affects children's use of derivational morphology in visual word and pseudoword processing: Evidence from a cross-language study. *Frontiers in Psychology*, *6*, 452. https://doi.org/10.3389/fpsyg.2015.00452

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letters in words. *Journal of Educational Psychology*, *89*(4), 631. https://doi.org/10.1037/0022-0663.89.4.631

Castles, A., Davis, C., Cavalot, P., & Forster, K. (2007). Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology*, *97*(3), 165-182. https://doi.org/10.1016/j.jecp.2007.01.006

Castles, A., Davis, C., & Letcher, T. (1999). Neighbourhood effects on masked form priming in developing readers. *Language and Cognitive Processes*, *14*(2), 201–224. https://doi.org/10.1080/016909699386347

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology*, *6*, 645. https://doi.org/10.3389/fpsyg.2015.00645

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468-481. https://doi.org/10.1111/tops.12332

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589. https://doi.org/10.1037/0033-295X.100.4.589

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204. https://doi.org/10.1037/0033-295X.108.1.204

Dawson, N., Rastle, K., & Ricketts, J. (2018). Morphological effects in visual word recognition: Children, adolescents, and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 645. https://doi.org/10.1037/xlm0000485

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499-504. https://doi.org/10.1111%2F1467-9280.00392

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, *134*(4), 521. https://doi.org/10.1037/0096-3445.134.4.521

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128. https://doi.org/10.1037/bul0000210

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256. https://doi.org/10.1037/0096-3445.113.2.256

Gibson, E. J. (1969). *Principles of Perceptual Learning and Development*. New York: Appleton-Century-Crofts.

Grainger, J., & Beyersmann, E. (2017). Edge-aligned embedded word activation initiates morpho-orthographic segmentation. In *Psychology of Learning and Motivation* (Vol. 67, pp. 285-317). Academic Press. https://doi.org/10.1016/bs.plm.2017.03.009

Grainger, J., & Ziegler, J. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology, 2*, 54. https://doi.org/10.3389/fpsyg.2011.00054

Hasenäcker, J., Beyersmann, E., & Schroeder, S. (2016). Masked morphological priming in German-speaking adults and children: Evidence from response time distributions. *Frontiers in Psychology, 7*, 929. https://doi.org/10.3389/fpsyg.2016.00929

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behaviour Research Methods, 44*(1), 287-304. https://doi.org/10.3758/s13428-011-0118-4

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184-193. https://doi.org/10.1111/cdev.12461

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: is it long-term and implicit?. *Neuroscience Letters*, *461*(2), 145-149. https://doi.org/10.1016/j.neulet.2009.06.030

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language, 115*, 104152. https://doi.org/10.1016/j.jml.2020.104152

Longtin, C. M., Segui, J., & Hallé, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *18*(3), 313-334. https://doi.org/10.1080/01690960244000036

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology, 101*(2), 221-242. https://doi.org/10.1348/000712608X371744

Nation, K. (2009). Form–meaning links in the development of visual word recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3665-3674. https://doi.org/10.1098/rstb.2009.0119

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*(3), 447-461. https://doi.org/10.1017/langcog.2016.20

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745-2750. https://doi.org/10.1073/pnas.0708424105

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401. https://doi.org/10.1037/0096-3445.130.3.401

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56. http://dx.doi.org/10.1037/0033-295X.103.1.56

Quémart, P., Casalis, S., & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of*

Experimental Child Psychology, *109*(4), 478-496. https://doi.org/10.1016/j.jecp.2011.02.008

Rastle, K. (2019). The place of morphology in learning to read in English. *Cortex, 116*, 45-54. https://doi.org/10.1016/j.cortex.2018.02.008

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review, 11*(6), 1090-1098. https://doi.org/10.3758/BF03196742

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia*, *67*(2), 147-162. https://doi.org/10.1007/s11881-016-0136-0

Schmalz, X., Beyersmann, E., Cavalli, E., & Marinus, E. (2016). Unpredictability and complexity of print-to-speech correspondences increase reliance on lexical processes: More evidence for the Orthographic Depth Hypothesis. *Journal of Cognitive Psychology*, *28*(6), 658-672. https://doi.org/10.1080/20445911.2016.1182172

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon*, *12*(2), 263-282. https://doi.org/10.1075/ml.17009.sch

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523. http://dx.doi.org/10.1037/0033-295X.96.4.523

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638-647. https://doi.org/10.1016/S0022-5371(75)80051-X

Taft, M., & Nguyen-Hoan, M. (2010). A sticky stick? The locus of morphological representation in the lexicon. *Language and Cognitive Processes*, *25*(2), 277-296. https://doi.org/10.1080/01690960903043261

von Koss Torkildsen, J., Arciuli, J., & Wie, O. B. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences, 69,* 60-68. https://doi.org/10.1016/j.lindif.2018.11.003

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. https://doi.org/10.1037/0033-2909.131.1.3

Ziegler, J. C., Bertrand, D., Lété, B., & Grainger, J. (2014). Orthographic and phonological contributions to reading development: Tracking developmental trajectories using masked priming. *Developmental Psychology*, *50*(4), 1026. https://doi.org/10.1037/a0035187

Chapter II

EyeReadIt: A Developmental Eye-Tracking Corpus of Text Reading in Italian

Reading is a uniquely human ability and, undoubtedly, a fundamental skill to functioning in today’s society. It allows us to perform a diverse set of tasks that can be as essential as understanding the instructions featured on a medicine bottle, and as ordinary as choosing a meal from a restaurant menu. From a developmental perspective, learning to read constitutes an important milestone as it provides children with a direct means to knowledge and education, and as such carries a widespread impact on academic performance, self-esteem, and subsequent adult-life chances (see e.g., Castles et al., 2018). It therefore comes as no surprise that a substantial part of psychological research has been dedicated to the study of reading and its development.

One methodological approach that has been particularly influential in this line of research is the recording of participants’ eye movements as they read. When we read, our eyes make a series of rapid, ballistic movements from one place in the text to another (saccades). These are separated by pauses (fixations) during which text information is collected. Critically, the pattern of our eye-movement behaviour provides an excellent online indication of the cognitive processes underlying reading. For example, the duration of fixations increases when we encounter words that are more difficult to identify (e.g., low-frequency words; Rayner & McConkie, 1976; Rayner & Pollatsek, 1981). Similarly, when we encounter sentences that are syntactically ambiguous, we tend to make regressive saccades that move the eyes backward in the text in order to re-read it (e.g., garden path constructions; Frazier & Rayner, 1982).

While the first eye-movement experiments date back to the end of the 19th century, it was the technological advancements in eye-tracking systems in the mid-1970s that marked what has been the most prolific era of eye-movement research in reading to date (see Rayner, 1998, for a review of early work). Since then, increasingly accurate eye-tracking measures and a variety of experimental techniques (i.e., ranging from simple visual display to innovative eye-contingent display change paradigms) have enabled researchers to investigate reading at different levels of processing (see Liversedge & Findlay, 2000; Rayner, 1998, 2009, for reviews in skilled adult readers, and Blythe & Joseph, 2011, for a review in developing readers). For example, measurements from eye-movement records have been used to indicate the scope of the perceptual span (i.e., how many letters readers process in any one fixation; Henderson et al., 1997; McConkie & Rayner, 1975; for developmental evidence, Häikiö et al., 2009; Rayner, 1986) and to reveal the effects that lexical properties (e.g., length and frequency; Rayner & McConkie, 1976; Rayner & Pollatsek, 1981; see also studies with developing readers, such as: Blythe et al., 2009, 2010; Huestegge et al., 1999; Hyönä & Olson, 1995; Joseph et al., 2009; Rayner, 1986) and contextual constraints (e.g., predictability; Rayner, Binder, Ashby & Pollatsek, 2001; developing readers: Johnson et al., 2018) can exert on word identification. At the same time, eye-movement behaviour has also been used to examine higher order processes involved in the comprehension of written language processing (e.g., through syntactic ambiguities and semantic inconsistencies during sentence and discourse processing; Frazier & Rayner, 1982; Garrod et al., 1994; Rayner, Chace, Slattery, & Ashby, 2006; Spivey & Tanenhaus, 1998; Joseph & Liversedge, 2013). Indeed, over the last 50 years or so, this line of research has grown to the point where sophisticated computational models can account for many of the phenomena associated with eye movement behaviour during reading in both skilled (e.g., the SWIFT model, Engbert et al., 2002; Engbert et al., 2005; the

E-Z Reader model, Reichle et al., 1998; Reichle et al., 2006; Reichle et al., 1999, 2003; the OB1 model, Snell et al., 2018) and developing (e.g., Reichle et al., 2013) readers.

More recently, however, eye-movement research in reading appears to be entering a new era. Consistent with the notion that *as reading is a highly ecological skill, so should the methods used to investigate it be* (see Jarodzka & Brand-Gruwel, 2017, for a discussion), this era is marked by a movement away from the conventional design of eye-movement experiments. Experimental materials are no longer restricted to a set of stimuli rigorously selected to test specific hypotheses, nor are presented in highly-controlled, often single-sentence, reading paradigms. Using instead a more naturalistic range of stimuli, this new generation of eye-movement studies adopt multiline reading paradigms during which participants read passages of text, ranging from short passages (e.g., Foster et al., 2018; Spichtig et al., 2017; Kuperman et al., 2018; Luke & Christianson, 2018; Tiffin-Richards & Schroeder, 2018, 2020; Kuperman et al., under review) to entire novels (e.g., Cop et al., 2015; Cop et al., 2017). This experimental approach enables researchers to examine reading as we experience it in our everyday life, considering the influence of a number of variables at different levels of processing (word-level, sentence-level, paragraph-level) and their possible interactions.

An additional advantage of this approach is that it allows the collection of large amounts of eye-tracking data, which, in turn, can become information-rich research resources. In particular, a corpus of eye-tracking data during multiline text reading can be readily used for hypothesis testing that is not limited to the scope of a given study, and can contribute to the development of comprehensive accounts of eye movements and visual information processing during natural reading conditions. Indeed, several eye-tracking corpora of text reading have emerged in the last few years (GECO, Cop et al., 2017; Provo, Luke & Christianson, 2018; ZuCO, Hollenstein et al., 2018; MECO, Kuperman et al., under review).

However, all of the corpora available contain data from skilled readers, creating an obvious need for such an invaluable resource for the purposes of developmental reading research. Here we introduce EyeReadIt, the first developmental eye-tracking corpus of natural multiline reading. The corpus consists of eye-tracking measures collected from 141 children between Grade 3 and Grade 6, and a control group of 33 adults, all Italian readers, as they silently read passages from children's books for comprehension.

Furthermore, we report the results of a series of analyses that examine whether, and to what extent, well-known phenomena documented in the literature on children's eye movements reading are replicated in the context of our ecologically valid reading paradigm. As indicated by Blythe and Joseph's (2011) review, there is a considerable degree of consistency with respect to how the global characteristics of eye movements change as children who are learning to read become skilled adult readers. Furthermore, these changes appear to be consistent across the different languages and education systems that have been thus far examined. Accordingly, our first set of analyses inspects the general developmental pattern of eye-movements during reading. In a second set of analyses, we investigate the benchmark effects of word length (i.e., long words are fixated longer than short words) and word frequency (i.e., high frequency words are fixated longer than low frequency words) on eye-movement measures as a function of reading development (Blythe et al., 2009, 2010; Huestegge et al., 1999; Joseph et al., 2009; Reichle et al., 2013). Finally, we investigate the effects of an additional word-level variable, namely, morphological complexity, on children's eye-movements during multi-line text reading, as a way to demonstrate how *EyeReadIt* can be used to advance our understanding of lexical processing under normal reading conditions.

EyeReadIt is being finalised and will soon be publicly available as a free resource, that researchers will be able to download from the Open Science Framework at https://osf.io/7srkg/

Materials & Methods

Participants

The study was approved by the Ethics Committee of SISSA. Prior to their participation children gave oral consent, while written consent was obtained from their parents. Adult participants gave written consent. Participants' age and gender, as well as other demographic information, are shown in Table 1. Both children and adults were native speakers of Italian and/or received formal education in Italian. They all had normal or corrected-to-normal vision, and no record of reading disability or neurological impairment.

Developing readers

156 children from Grade 3 to 6 (age range: 8-12 years) participated in the study and received a book as a reward. Data collection was part of a hands-on science activity organised between regional schools in Trieste, Italy and SISSA, and took place between June 2017 and May 2018. Data from fifteen children were excluded from the analyses due to technical issues that occurred during data acquisition (two participants), excessive noise artifacts in the eye-movement data (e.g., vertical drift, head movement; 10 participants), or session interruptions (2 participants). This left an effective sample of 141 children to be included in the analyses.

All children completed a reading proficiency and a non-verbal intelligence test. Reading proficiency was assessed with a subtest of the MT Reading Test for Primary School (Cornoldi & Colpo, 1998). In particular, all children were asked to read aloud a short story, recommended for the assessment of reading ability at the beginning of Grade 4 (i.e., *L'indovina che non indovinò*; in English, *The fortune teller who couldn't tell fortune*), while their voice was recorded. Raw scores for reading accuracy and speed (syllables/second and seconds/syllable) were calculated according to the test's scoring guidelines. The raw scores for reading speed were converted into z-scores, normalised on the experimental population.

Children's non-verbal intelligence was assessed with Raven's Coloured Progressive Matrices (Raven, 1949), and the raw scores were also converted into z-scores following an analogous procedure.

Skilled readers

37 young adults, mostly students from the University of Trieste and SISSA, participated in the study in exchange for monetary compensation. Data from four participants were discarded due to excessive noise artifacts in the eye-movement data, leaving a sample of 33 skilled adult readers to be included in the analyses. Adult participants were administered a reading proficiency test for older adolescents/young adults (Cornoldi & Candela, 2015), similar to the MT test used with children, while Raven's (2003) Standard Progressive Matrices were used for the non-verbal intelligence assessment. Z-scores were obtained for both tests.

Table 1 *Participants' demographic characteristics.*

	Developing readers			
	Grade	***N***	***M* age *(SD)***	**Range**
N = 141 F = 73, M = 68	3	37	8.22 (0.42)	8-9
	4	20	9.22 (0.41)	9-10
	5	41	10.05 (0.44)	9-11
	6	43	10.98 (0.34)	10-12
	Skilled adult readers			
	***M* years of education (*SD*)**	***N***	**M age (SD)**	**Range**
F = 21, M = 12	14.42 (2.28)	33	23.39 (3.32)	19-33

Design & Materials

Experimental materials comprised 12 passages of connected multiline text extracted from six children's books (i.e., two passages per book). These were retrieved from various websites online and were slightly adapted in order to appear more relatable to Italian children (i.e., references to foreign proper nouns were substituted with Italian ones; e.g., *Alì Babà* became *Fabio*; *pence* became *euro*). The passages were equally divided into two experimental stimulus sets, with each set containing a passage from each book. This step ensured some variability in the experimental material and accounted for potential idiosyncrasies in the selection of the specific passages.

Each participant was administered one stimulus set of six passages (i.e., Set A or Set B). The order of passage presentation within each set was fixed and was based on the text difficulty as determined by the target readership's age of each book provided by the books' publishers. Passages suitable for the youngest group of children (Grade 3) were displayed first (e.g., *Goldilocks, The Bremen Town Musicians*) and those suitable for older children were displayed last (e.g., *The Call of the Wild, Twenty Thousand Leagues Under the Sea*). This step was taken in order to encourage young children to read as many passages as possible, thus accommodating maximum data collection. The assignment to a stimulus set was performed in a counterbalanced order across participants. Prior to the experimental passages all participants were presented with a practice passage extracted from a well-known story (i.e., *Little Red Riding Hood*). All passages are shown in Appendix A.

Passages are an average of 130.5 words long (range: 109-170) and contain 6 sentences on average (range: 3-10). Sentences are on average 21.75 words long (range: 2-79). Across all passages, there are a total of 1566 word tokens, of which 762 are distinct types. A summary of the characteristics of the lexical tokens available in EyeReadIt is presented in Table 2.

Table 2 Summary of the characteristics of the lexical tokens contained in *EyeReadIt*’s text passages. Word Zipf frequency extracted from SUBTLEX-IT (Crepaldi et al., 2016). The parts of speech considered for content words in this analysis are nouns, verbs, and adjectives.

N word tokens	1566
N word types	762
Unique parts of speech	11
Mean word count per text (range)	130.5 (109–170)
Mean word length (range)	4.66 (1–15)
Mean word Zipf frequency (range)	5.35 (1.19–7.26)
N content word tokens (types)	792 (593)
N complex word tokens (types) - content words	449 (338)
N simple word tokens (types) - content words	343 (256)
N derived word tokens (types) - content words	68 (67)
N inflected word tokens (types) - content words	381 (273)

Apparatus

Eye movements were recorded using an EyeLink 1000 Plus tower mount eye-tracker (SR Research, Canada) at a sampling rate of 1000 Hz. A head-and-chin rest was used to minimise head movements. Viewing was binocular, but eye movements were recorded from the right eye only. Stimuli were delivered on a 27″ monitor with a resolution of 1920x1080 px, at a viewing distance of approximately 63 cm. The screen refresh rate was set at 144 Hz. Text was displayed in black 20-point Courier New font on a light grey background. Each passage appeared as one multi-lined paragraph on the screen, spanning between 10 and 13 lines of text. The lines were double spaced, and each character subtended approximately 0.45 degrees of visual angle or 16 pixels.

Procedure

Participants were tested individually in a quiet and dimly lit room. The experiment began with a nine-point-calibration procedure until calibration error was below an average of 0.5 degrees of visual angle. At the beginning of each trial, a drift correction was carried out to correct for any small movements that may have occurred. For this, a small circular target appeared at the location of the first letter of the first word of the sentence (i.e., left-aligned with a 360-pixel horizontal and a 140-pixel vertical screen offset). A stable fixation on this target was required for the trial to proceed; otherwise, a recalibration procedure was initiated. Following the drift correction, a passage was displayed on the screen. Participants were instructed to read each passage silently and press a button to indicate they had finished reading it and proceed to the next trial. A simple comprehension question was presented after every two trials as an index of task engagement, whereby participants were asked to choose between two possible answers. To reduce head movements, children were instructed to indicate their response by raising their right hand for selecting the right-hand answer as displayed on the screen, and their left hand for selecting the left-hand answer; the experimenter would then press the right or left key according to the children's responses. Adult participants indicated their response with a button press. Mean accuracy was 93% for children and 93% for adults. Experimental trials were preceded by one practice trial consisting of a small 7-lined passage (*Little Red Riding Hood*). The calibration procedure was repeated whenever necessary during the experiment.

Each eye-tracking session with children lasted around 15-20 minutes, in order to fit into the hands-on activity scheduled rotation, and to prevent physical discomfort and loss of attention. For adults, sessions were often even shorter (approximately 10 minutes), thanks to generally smooth calibration and shorter reading times. Participants were free to stop at any point, and data from children were retained even when they did not read all six passages presented in the experimental session.

Data Analysis

Two main sets of analyses were performed. First, we sought to capture the general trends of developmental changes in eye-movement reading behaviour, as reported in the literature (e.g., Blythe et al., 2009, 2011; for an overview, see Blythe & Joseph, 2011). For this set of analyses we calculated the following measures: reading rate, which is the number of words fixated per minute; saccade length, expressed as the number of text characters covered by each saccade; fixation duration; number of fixations per 100 words of text; skipping probability, pertaining to the probability that the target word did not receive a direct fixation during first-pass reading; refixation probability, referring to the probability that a word receives additional fixations following the first fixation, during first-pass reading; and regression probability, that is, the probability of a leftward saccade out of the target word, provided that it was fixated during first-pass reading.

The second set of analyses examined the well-established effects that word length and word frequency have on reading time measures as well as their interaction with development (as expressed in terms of school grade in the current design). The dataset, reflecting natural reading experience, displays a strong correlation (-.79) between word length and word frequency. Due to this, length and frequency have been used as independent predictors in different models.

Additionally, we investigated the effect of morphological complexity on reading times as a function of development, in two sets of analyses. In the first set, morphological complexity is treated as a factorial variable, with the two levels being “complex” (i.e., multimorphemic) vs “simple” (i.e., all base forms). In the second set of analyses, the two levels of morphological complexity are instead “derived” (i.e., complex derived words) and “inflected” (complex inflected words).

To assess the effects of all psycholinguistic variables we calculated the following word-level reading measures extracted from EyeReadIt: first-of-many fixation duration (for short, *FoM*), pertaining to those instances of first fixation duration in which the first fixation was followed by at least another fixation during first pass, was taken to index an early stage of visual processing; gaze duration (*GD*), the summed duration of all fixations performed on a word during first-pass reading, before moving rightward, was selected as a measure of lexical access; total reading time (*TRT*), the summed duration of all fixations performed across all runs, was taken as reflecting later post-lexical processing (e.g., integration processes).

All data analyses were carried out using *R* (version 4.0.4; R Core Team, 2021) and the *lme4* package (version 1.1-26; Bates et al., 2015). In order to satisfy the assumption of normality, all continuous dependent variables were log-transformed, as indicated by a Box-Cox test (from the R *car* package; Fox & Weisberg, 2019), and then back-transformed exponentially for a clearer illustration of the nature of the effects. For the analysis of developmental changes, models with a continuous dependent variable (*DV*) were constructed according to the following basic scheme: *lmer(DV ~ IV + MT_zscore + Raven_zscore + (1|subjectID) + (1|excerptID))*, whereby *DV* is each of the above-described reading measures, and *IV* (independent variable) is Grade (as a factor with levels: 3, 4, 5, 6, adults). Random intercepts for subjects and passages were included as random effects factors. For models with probability measures as *DV*s, the same scheme was adopted; in these cases, however, the function *glmer,* with *family = “binomial”*, from the *lme4* package, was used.

Statistical models of the effects of lexical variables on eye-tracking measures were constructed very similarly, with the only relevant difference being the model predictor, modelled as an interaction term of the lexical *IV* of interest (length, frequency or morphological complexity) and Grade. Random intercepts for subjects, passages and items (word spellings) were included as random effects factors. This resulted in the following model structure: *lmer(DV ~ IV*Grade + MT_zscore + Raven_zscore + (1|subjectID) + (1|excerptID) + (1|spelling)).* Furthermore, in these models, Grade was contrast-coded using backward difference coding, in order for the effect of each grade to be contrasted with that of the previous grade. In all models, Z-scores of the MT reading speed test and of the Raven matrices (non-verbal intelligence) were included as covariates. The significance of the fixed effects was determined with either type II (for models of developmental changes, with Grade as the main predictor) or III (for lexical variables, in which the main predictors were interaction terms) model comparisons, using the Anova function in the *car* package (Fox & Weisberg, 2019). Below we report and discuss the results pertaining to the factors of interest. All results revealed a significant effect of MT reading speed test, with eye-movement reading measures decreasing as the scores on the reading efficiency test increased, whereas the effect of the non-verbal intelligence test was never significant.

Results

Developmental changes in eye-movements during reading

As expected, results revealed that as the school grade increased, so did children's overall reading rate, growing from 90.11 words per minute (wpm) with the 3rd Graders to 167.30 wpm with the 6th graders, and 289.95 wpm with the adults (Fig. 1a; $\chi^2(4, 114349) = 564.18$, p <.001). This increase in proficiency with school grade was also reflected in all the other eye-movement measures examined. First, mean saccade length increased with school grade, increasing from 3.77 characters with the 3rd graders to 5.51 characters with the 6th graders and 7.80 characters with adults (Fig. 1b; $\chi^2(4, 184071) = 354.39$, p <.001). Second, the mean fixation duration decreased with grade, ranging from 241.56 ms with the 3rd graders to

207.66 ms with the 6th graders and 180.07 ms with the adults (Fig. 1c; χ^2(4, 199094) = 157.89, p <.001). Third, the mean number of fixations per 100 words also decreased with grade, ranging from 306.41 with the 3rd graders to 161.60 with the 6th graders and 109.54 with the adults (Fig. 1d; χ^2(4, 199094) = 318.02, p <.001). Fourth, skipping probability increased with grade, ranging from 0.12 with the 3rd graders to 0.19 with the 6th graders and 0.28 with the adults (Fig. 1e, χ^2(4, 114350) = 141.94, p <.001). Fifth, refixation probability decreased with grade, ranging from 0.41 with the 3rd graders to 0.29 with the 6th graders and 0.17 with the adults (Figure 1f, χ^2(4, 91430) = 285.81, p <.001). Finally, regression probability decreased with grade, ranging from 0.35 with the 3rd graders to 0.30 with the 6th graders and 0.21 with the adults (Figure 1g, χ^2(4, 91430) = 74.91, p <.001).

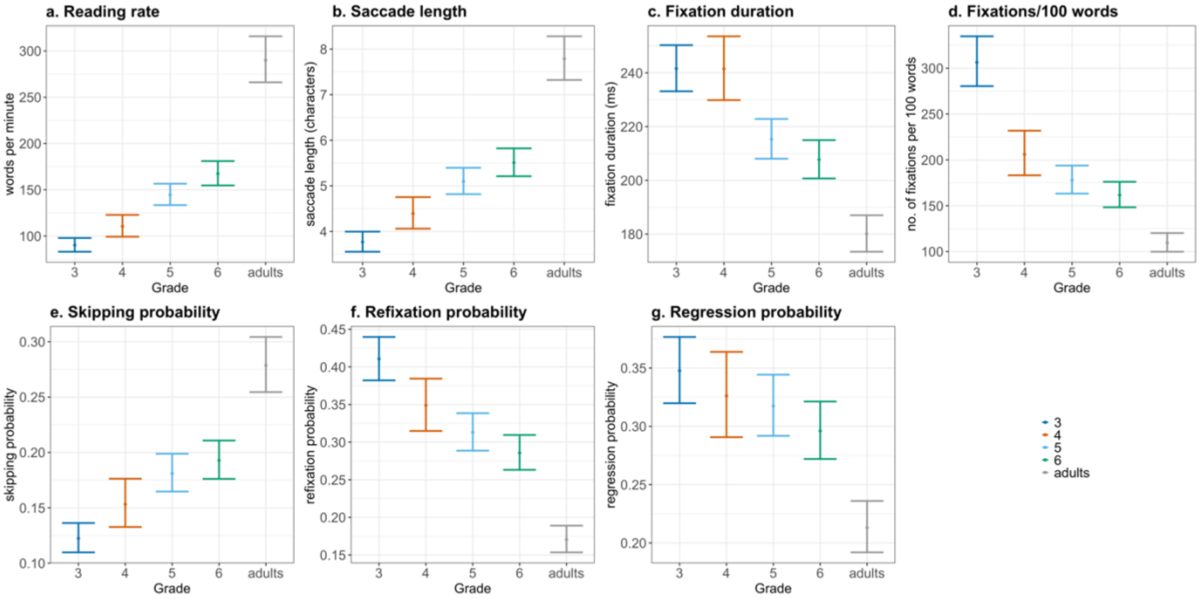


Figure 1 Illustration of the mixed model estimates of the effects of grade on fundamental eye-tracking measures indexing reading behaviour. Colour code for grade.

Benchmark effects of word length and word frequency

Two of the most robust effects in the adult and children eye movement literature are those of word length and word frequency. For this reason, we have chosen to evaluate the validity of *EyeReadIt* by seeking evidence for the influence of these two word-level variables on the identification of words during passage reading.

Word length

The effects, as estimated by our models, are reported in Table 3; for plots, see Figures 2a, 2b, and 2c. With respect to the models in which word length was analysed in interaction with grade, we report a significant main effect of word length on first-of-many fixation duration (p<.001), gaze duration (p<.001) and total reading time (p<.001), with longer durations of all three measures for longer words. A significant main effect of grade is found on all three eye-tracking measures (all p<.001), with progressively shorter durations in older readers. The effect of the interaction between word length and grade is significant for gaze duration and total reading time (all p<.001), with the effect of word length getting progressively smaller as a function of grade for both measures, but not for first-of-many fixation duration (p=.68).

Word frequency

Model estimates are reported in Table 3; for plots, see Figures 2d, 2e, and 2f. With respect to the models in which word frequency was analysed in interaction with grade, we report a significant main effect of word frequency on first-of-many fixation duration, gaze duration and total reading time (all p<.001), with shorter durations for higher-frequency words, and a significant main effect of grade on all three measures as well (all p<.001), with progressively shorter durations in more advanced readers. The effect of the interaction between word frequency and grade is significant for gaze duration and total reading time (all p<.001), with the effect of word frequency getting progressively smaller as a function of grade for both measures, but not for first-of-many fixation duration (p=.80).

Table 3 Linear mixed model results of the effects of length and frequency, in interaction with grade, on eye-tracking measures of interest.

Eye-tracking variable	**Predictor**	**Main effect of predictor**	**Main effect of grade**	**Interaction effect**
First-of-many-fixation duration (ms)	Length	Chisq(1, 27262) = 21.158, p<.001***	Chisq(4, 27262) = 73.515, p<.001***	Chisq(4, 27262) = 2.307, p=.679
Gaze duration (ms)	Length	Chisq(1, 91430) = 782.195, p<.001***	Chisq(4, 91430) = 165.433, p<.001***	Chisq(4, 91430) = 347.913, p<.001***
Total reading time (ms)	Length	Chisq(1, 91430) = 1185.617, p<.001***	Chisq(4, 91430) = 218.251, p<.001***	Chisq(4, 91430) = 859.701, p<.001***
First-of-many-fixation duration (ms)	Frequency	Chisq(1, 27262) = 47.825, p<.001***	Chisq(4, 27262) = 69.391, p<.001***	Chisq(4, 27262) = 1.615, p=.806
Gaze duration (ms)	Frequency	Chisq(1, 89911) = 542.821, p<.001***	Chisq(4, 89911) = 755.711, p<.001***	Chisq(4, 89911) = 320.469, p<.001***
Total reading time (ms)	Frequency	Chisq(1, 89911) = 915.429, p<.001***	Chisq(4, 89911) = 1195.397, p<.001***	Chisq(4, 89911) = 832.745, p<.001***

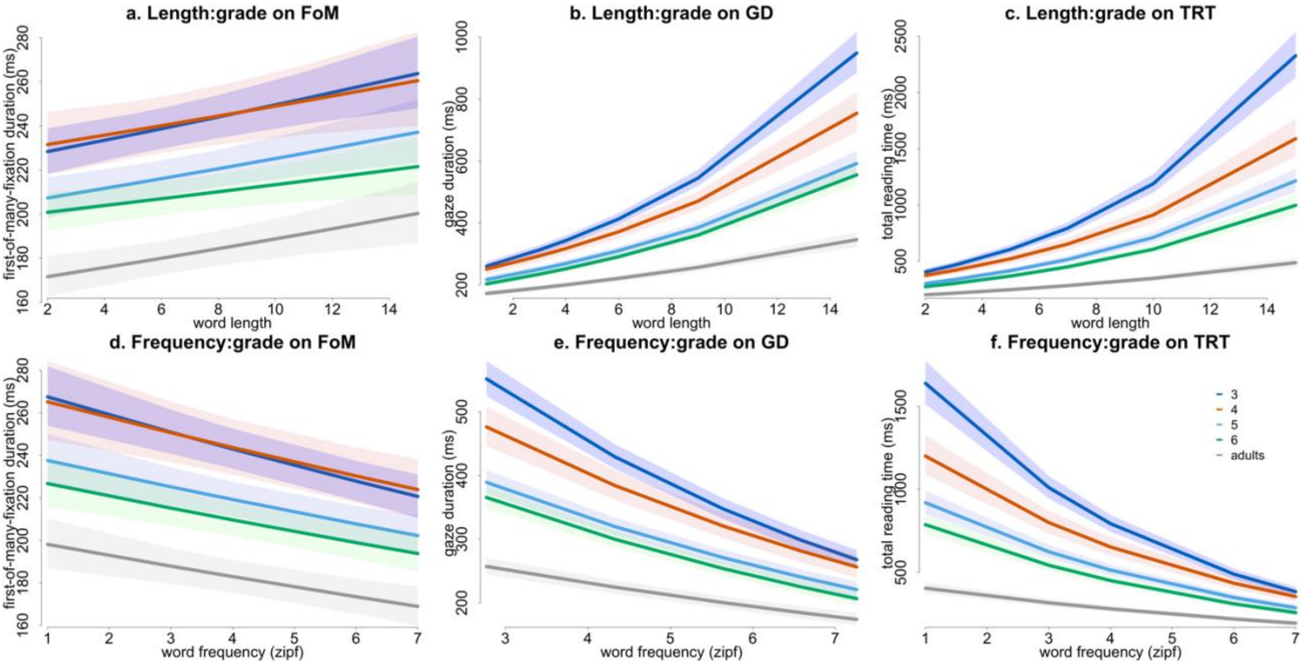


Figure 2 Effects of the interaction between word length and grade, and between word frequency and grade on first-of-many fixation duration (FoM), gaze duration (GD) and total reading time (TRT). Colour code for grade (key in bottom right subplot).

Morphological complexity effects on reading times

In a first set of analyses, morphological complexity is treated as a factorial variable, with the two levels being “complex” (i.e., multimorphemic) vs “simple” (i.e., all base forms). All model estimates are reported in Table 4; for plots, see Figure 3. With respect to the effects of morphological complexity and grade (Figure 3a–c), we report a significant main effect of morphological complexity on gaze duration and total reading time (all p<.001), with longer durations for complex words, but not for first-of-many fixation duration (p=.399). We also report a significant main effect of grade on all three measures considered (all p<.001), with progressively shorter durations in older readers. The interaction between morphological complexity and grade is significant for gaze duration (p=.011) and total reading time (p<.001), with the effect of morphological complexity getting progressively smaller as a function of grade for both measures, but not for first-of-many fixation duration (p=.970). In particular, with respect to gaze duration and total reading time, Grade 6 readers showed a more marked effect of morphological complexity than adult readers (t=2.423 and t=4.439, respectively), while no significant difference emerges between developing readers of any grade. This suggests that young readers are more sensitive than adults to the effects of morphological complexity in reading.

The second set of analyses, in which the two levels of morphological complexity are “derived" and “inflected”, reveals a very similar pattern of results (Figure 3d–f), as reported in Table 6, with inflected words yielding shorter gaze durations and total reading times, this effect being modulated by grade, with younger readers showing greater sensitivity to derived words compared to older ones. As a matter of fact, in this set of analyses, too, gaze duration and total reading time reveal a significant interaction of grade and morphological complexity. Grade 3 and Grade 6 readers display a greater effect than Grade 4 (t=-2.480) and adult readers (t=-2.127), respectively, when it comes to gaze duration. Grade 6 readers display a greater effect than adult readers (t=-4.787), when it comes to total reading time.

Table 4 Linear mixed model results of the effects of morphological complexity variables, in interaction with grade, on eye-tracking measures of interest, considering content words only.

Eye-tracking variable	**Predictor**	**Main effect of predictor**	**Main effect of grade**	**Interaction effect**
First-of-many-fixation duration (ms)	MorphComplex (vs Simple)	Chisq(1, 19972) = 0.594, p=.441	Chisq(4, 19972) = 102.474, p<.001***	Chisq(4, 19972) = 0.532, p=.970
Gaze duration (ms)	MorphComplex (vs Simple)	Chisq(1, 52451) = 17.033, p<.001***	Chisq(4, 52451) = 440.325, p<.001***	Chisq(4, 52451) = 13.005, p=.011*
Total reading time (ms)	MorphComplex (vs Simple)	Chisq(1, 91430) = 36.895, p<.001***	Chisq(4, 91430) = 583.534, p<.001***	Chisq(4, 91430) = 83.827, p<.001***
First-of-many-fixation duration (ms)	MorphDerived (vs Inflected)	Chisq(1, 12076) = 1.034, p=.309	Chisq(4, 12076) = 99.217, p<.001***	Chisq(4, 12076) = 2.431, p=.657
Gaze duration (ms)	MorphDerived (vs Inflected)	Chisq(1, 30010) = 23.736, p<.001***	Chisq(4, 30010) = 412.554, p<.001***	Chisq(4, 30010) = 16.799, p<.001***
Total reading time (ms)	MorphDerived (vs Inflected)	Chisq(1, 38993) = 32.596, p<.001***	Chisq(4, 38993) = 582.044, p<.001***	Chisq(4, 38993) = 124.298, p<.001***

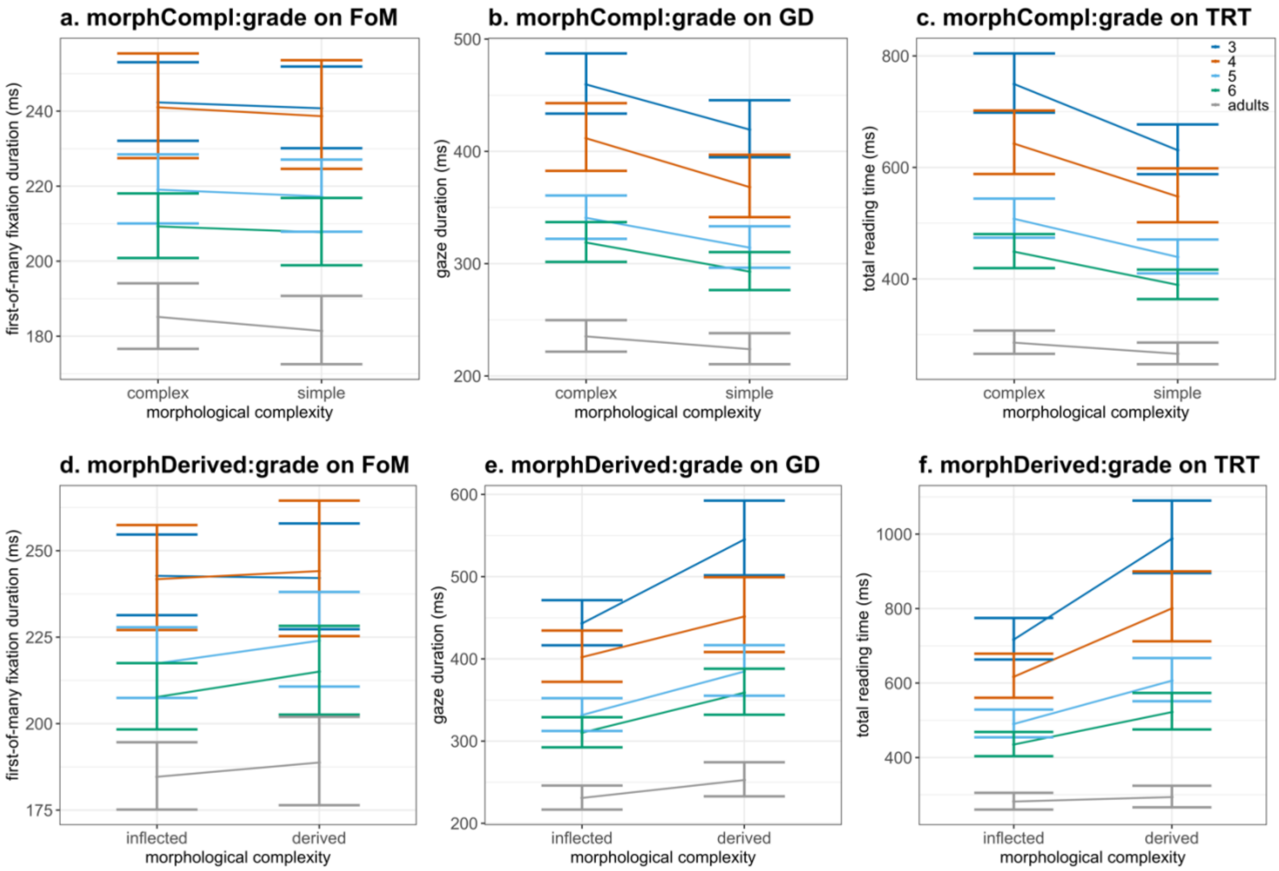


Figure 3 Effects of the interaction between morphological complexity (a, b, c: complex vs simple; d, e, f: derived vs inflected) and grade on first-of-many fixation duration (FoM), gaze duration (GD) and total reading time (TRT). Colour code for grade (key in top-right corner of panel c).

Discussion

The present work introduces *EyeReadIt*, the first developmental database of eye-movement measures in natural reading. While other eye-tracking corpora are based on multiline text reading performed by skilled adult readers, *EyeReadIt* is, to the best of our knowledge, the first publicly available resource focusing on eye-tracking data obtained from a very large sample of developing readers, while using a naturalistic multiline reading task. Our corpus outnumbers the existing ones in terms of participants (141 children and 33 adults); currently, the richest eye-tracking resource is the Provo Corpus (Luke & Christianson, 2018) containing data from 84 native English speaking adult participants. However, it should be noted that we have a much lower number of tokens than other corpora (1566 tokens, vs over 56000 in the Dundee Corpus: Kennedy, 2003; Kennedy et al., 2003; Kennedy & Pynte, 2005), and, more generally, we submitted each participant to much less text than other resources (a maximum of 6 passages per participant; 55 short passages per participant in the Provo Corpus; an entire novel in GECO: Cop et al., 2017). This is due, of course, to the fact that we tested children, and therefore had to design an experiment that would not be too tiring or long, while allowing us to collect a sufficient amount of data.

With EyeReadIt, we aim at providing researchers with the opportunity to both conduct more exploratory analyses and tackle refined research questions, from a developmental and potentially cross-linguistic perspective, without the need to design experiments and collect data themselves. Consistent with the new direction of eye-movement research in reading, the nature of the experimental task used in *EyeReadIt* provides, as an added value, the opportunity to perform eye-tracking analyses on connected text, and to thus investigate phenomena associated with eye movements during natural reading.

The present work ensured that this novel database allowed to highlight some well-consolidated effects on eye-tracking measures, which are typically obtained with more artificial (commonly sentence-reading) paradigms. We aimed at extending the scope of such effects to natural reading with a developmental population. Our eye-tracking data, in keeping with the literature (see Rayner, 1986; Blythe & Joseph, 2011; Reichle et al., 2013), demonstrated that as children's school grade increased (along with their age and reading ability), their patterns of eye movements came to more closely resemble those of adults, such that they made both longer saccades and fewer, shorter fixations, fewer of which occurred in the same word or after regressions.

Furthermore, we showed that both the length of a word and the frequency with which it appears in the written language influence the time developing and skilled readers spend looking at it: long and high-frequency words require longer reading times than short and low-frequency words, respectively (for a review, see e.g. Reichle et al., 2013). Both of these lexical variables influenced our participants' reading times in a comparable way. As far as the durations of first-of-many fixations are concerned, length and frequency effects were present across all reader groups and were not modulated by development. Given that first-of-many fixation durations are considered an early measure of lexical processing, these findings suggest that length and frequency exert an immediate effect on early visual word identification. Interestingly, the influence of length and frequency on gaze durations and total reading times interacted with development. These effects were larger in children than adult readers, and decreased with grade. These findings reflect that the length and frequency have a more substantial impact on children's lexical processing as compared to that of adults (Blythe & Joseph, 2011).

Overall, the successful replication of the effects of established predictors in our experiment allows some interesting considerations. First, this pattern of results corroborates the validity of the proposed corpus for the conduction of linguistic analyses, and more generally of a much-needed ecological approach to the study of reading development. Second, our results are theoretically relevant as well, as they suggest that readers as young as third graders have already acquired enough information about word distribution in their lexicon of reference (for similar findings in English see, e.g., Joseph et al., 2013) for it to reflect in their eye-movement behaviour. Furthermore, the fact that length and frequency effects are observed more prominently in younger developing readers, and that the magnitude of the effects gets progressively smaller with reading development, confirms trends from prior research (Blythe

& Joseph, 2011; Joseph et al. 2013), with the added value of using a multiline text reading task, instead of single sentence reading.

We also used *EyeReadIt* to analyse the effects of morphological complexity in interaction with grade on the above-described eye-movement measures. Coherently with what one might predict, in the light of the more fine-tuned nature of the effects of morphological complexity and of the naturalistic paradigm used, our results show neither an effect of morphological complexity nor a significant interaction with grade, at the earliest stage of lexical processing as indexed by first-of-many fixation duration. However, a role of morphological complexity surfaces upon lexical access and post-lexical levels of processing (gaze duration and total reading time), in a grade-modulated fashion, as detailed above.

Evidence for an interaction between morphological complexity and reading experience is in agreement with morphological decomposition theories (see, e.g., Burani et al., 2008), according to which developing readers as young as second and third graders already display morphological awareness and use morphological information in reading (Grainger & Beyersmann, 2017; Hasenäcker et al., 2016; Quémart et al., 2011). Our pattern of results suggest that children exploit the presence of a morphological structure during online processing of written words in naturalistic task, and that they do so from a very young age. In this sense, we can interpret these effects of morphological complexity as a proxy of ongoing sublexical processing (i.e., extraction of information from morphemes), which sits at the interface between whole-word processing and the extraction of letter-level statistics. Furthermore, consistent with the notion that younger and less experienced readers rely more heavily on sublexical processing, effects of morphological complexity are more pronounced in our younger children compared to those of older children and adults that can identify both morphologically complex and simple words as whole-units (Burani et al., 2008). These results further corroborate the potential of a resource like *EyeReadIt*, as it allows to investigate even finer aspects of linguistic processing through eye movements without any artificial constraint.

References

Amenta, S., Marelli, M., & Crepaldi, D. (2015). The fruitless effort of growing a fruitless tree: Early morpho-orthographic and morpho-semantic effects in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1587. https://doi.org/10.1037/xlm0000104

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286-304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behaviour Research Methods, 39*(3), 445-459. https://doi.org/10.3758/BF03193014

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bertram, R. (2011). Eye movements and morphological processing in reading. *The Mental Lexicon, 6*(1), 83-109. https://doi.org/10.1075/ml.6.1.04ber

Beyersmann, E., Castles, A., & Coltheart, M. (2012). Morphological processing during visual word recognition in developing readers: Evidence from masked priming. *Quarterly Journal of Experimental Psychology, 65*(7), 1306-1326. https://doi.org/10.1080%2F17470218.2012.656661

Biederman, G. B. (1966). Supplementary report: The recognition of tachistoscopically presented five-letter words as a function of digram frequency. *Journal of Verbal Learning and Verbal Behaviour, 5*(2), 208-209. https://doi.org/10.1016/S0022-5371(66)80020-8

Blythe, H. I., Häikiö, T., Bertam, R., Liversedge, S. P., & Hyönä, J. (2011). Reading disappearing text: Why do children refixate words?. *Vision Research, 51*(1), 84-92. https://doi.org/10.1016/j.visres.2010.10.003

Blythe, H. I., & Joseph, H. S. S. L. (2011). Children's eye movements during reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *Oxford Library of Psychology. The Oxford Handbook of Eye Movements* (pp. 643–662). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0036

Blythe, H. I., Liversedge, S. P., Joseph, H. S., White, S. J., & Rayner, K. (2009). Visual information capture during fixations in reading for children and adults. *Vision Research, 49*(12), 1583-1591. https://doi.org/10.1016/j.visres.2009.03.015

Burani, C., Marcolini, S., & Stella, G. (2002). How early does morpholexical reading develop in readers of a shallow orthography? *Brain and Language, 81*(1-3), 568-586. https://doi.org/10.1006/brln.2001.2548

Burani, C., Marcolini, S., De Luca, M., & Zoccolotti, P. (2008). Morpheme-based reading aloud: Evidence from dyslexic and skilled Italian readers. *Cognition, 108*(1), 243-262. https://doi.org/10.1016/j.cognition.2007.12.010

Castles, A., Davis, C., & Letcher, T. (1999). Neighbourhood effects on masked form priming in developing readers. *Language and Cognitive Processes, 14*(2), 201–224. https://doi.org/10.1080/016909699386347

Castles, A., Davis, C., Cavalot, P., & Forster, K. (2007). Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology, 97*(3), 165-182. https://doi.org/10.1016/j.jecp.2007.01.006

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology, 6*, 645. https://doi.org/10.3389/fpsyg.2015.00645

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS One, 10*(8), e0134008. https://doi.org/10.1371/journal.pone.0134008

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behaviour Research Methods, 49*(2), 602-615. https://doi.org/10.3758/s13428-016-0734-0

Cornoldi, C., & Candela, M. (2015). *Prove di Lettura e Scrittura MT-16-19*. Edizioni Erickson.

Cornoldi, C., & Colpo, G. (1998). *Prove di Lettura MT per la Scuola Elementare-2 Quarta elementare*. Firenze: Giunti, OS Organizzazioni Speciali.

Crepaldi, D., Amenta, S., Mandera, P., Keuleers, E., & Brysbaert, M. (2016). Frequency estimates from different registers explain different aspects of visual word recognition.

International Meeting of the Psychonomic Society, Granada, Spain, 5–8 May. http://crr.ugent.be/subtlex-it/

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*(5), 621-636. https://doi.org/10.1016/S0042-6989(01)00301-7

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777. https://doi.org/10.1037/0033-295X.112.4.777

Foster, T. E., Ardoin, S. P., & Binder, K. S. (2018). Reliability and validity of eye movement measures of children's reading. *Reading Research Quarterly*, *53*(1), 71-89. https://doi.org/10.1002/rrq.182

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression,* 3rd Edition. Thousand Oaks, CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178-210. https://doi.org/10.1016/0010-0285(82)90008-1

Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, *33*(1), 39-68. https://doi.org/10.1006/jmla.1994.1003

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*(2), 256. https://doi.org/10.1037/0096-3445.113.2.256

Goswami, U., & Ziegler, J. C. (2006). Fluency, phonology and morphology: a response to the commentaries on becoming literate in different languages. *Developmental Science, 9*(5), 451-453. https://doi.org/10.1111/j.1467-7687.2006.00511.x

Grainger, J., & Beyersmann, E. (2017). Edge-aligned embedded word activation initiates morpho-orthographic segmentation. In *Psychology of Learning and Motivation* (Vol. 67, pp. 285-317). Academic Press. https://doi.org/10.1016/bs.plm.2017.03.009

Häikiö, T., Bertram, R., Hyönä, J., & Niemi, P. (2009). Development of the letter identity span in reading: Evidence from the eye movement moving window paradigm. *Journal of Experimental Child Psychology*, *102*(2), 167-181. https://doi.org/10.1016/j.jecp.2008.04.002

Hasenäcker, J., Beyersmann, E., & Schroeder, S. (2016). Masked morphological priming in German-speaking adults and children: Evidence from response time distributions. *Frontiers in Psychology, 7*, 929. https://doi.org/10.3389/fpsyg.2016.00929

Henderson, J. M., McClure, K. K., Pierce, S., & Schrock, G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception & Psychophysics*, *59*(3), 323-346. https://doi.org/10.3758/BF03211901

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, *5*(1), 1-13. https://doi.org/10.1038/sdata.2018.291

Hyönä, J., Bertram, R., & Pollatsek, A. (2004). Are long compound words identified serially via their constituents? Evidence from an eyemovement-contingent display change study. *Memory & Cognition, 32*(4), 523-532. https://doi.org/10.3758/BF03195844

Hyönä, J., & Pollatsek, A. (1998). Reading Finnish compound words: Eye fixations are affected by component morphemes. *Journal of Experimental Psychology: Human Perception and Performance, 24*(6), 1612. https://doi.org/10.1037/0096-1523.24.6.1612

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: towards a model of real-world reading. *Journal of Computer Assisted Learning, 33*(3), 193-201. https://doi.org/10.1111/jcal.12189

Johnson, R. L., Oehrlein, E. C., & Roche, W. L. (2018). Predictability and parafoveal preview effects in the developing reader: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 44*(7), 973. https://doi.org/10.1037/xhp0000506

Joseph, H. S., Liversedge, S. P., Blythe, H. I., White, S. J., & Rayner, K. (2009). Word length and landing position effects during reading in children and adults. *Vision Research, 49*(16), 2078-2086. https://doi.org/10.1016/j.visres.2009.05.015

Joseph, H. S., Nation, K., & Liversedge, S. P. (2013). Using eye movements to investigate word frequency effects in children's sentence reading. *School Psychology Review, 42*(2), 207-222. https://doi.org/10.1080/02796015.2013.12087485

Kennedy, A. L. A. N. (2003). The Dundee corpus [CD-rom]. *Psychology Department, University of Dundee*.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153-168. https://doi.org/10.1016/j.visres.2004.07.037

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behaviour Research Methods, 44*(1), 287-304. https://doi.org/10.3758/s13428-011-0118-4

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184-193. https://doi.org/10.1111/cdev.12461

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology, 63*(9), 1838-1857. https://doi.org/10.1080%2F17470211003602412

Kuperman, V., Matsuki, K., & Van Dyke, J. A. (2018). Contributions of reader-and text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(11), 1687–1713. https://doi.org/10.1037/xlm0000547

Kuperman, V., Siegelman, N., & Schroeder, S. (under review). MECO.

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language, 115*, 104152. https://doi.org/10.1016/j.jml.2020.104152

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*(1), 6-14. https://doi.org/10.1016/S1364-6613(99)01418-7

Longtin, C. M., Segui, J., & Hallé, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes, 18*(3), 313-334. https://doi.org/10.1080/01690960244000036

Luke, S.G. & Christianson, K. (2018). The Provo Corpus: A Large Eye-Tracking Corpus with Predictability Ratings. *Behaviour Research Methods, 50,* 826-833. https://doi.org/10.3758/s13428-017-0908-4

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography-Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology, 68*(8), 1571-1583. https://doi.org/10.1080%2F17470218.2014.959709

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology, 101*(2), 221-242. https://doi.org/10.1348/000712608X371744

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17*(6), 578-586. https://doi.org/10.3758/BF03203972

McConkie, G. W., Zola, D., Grimes, J., Kerr, P. W., Bryant, N. R., & Wolff, P. M. (1991). Children’s eye movements during reading. *Vision and Visual Dyslexia*, *13,* 251-262. https://doi.org/10.1093/oxfordhb/9780199539789.013.0036

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283-312. https://doi.org/10.1006/jmla.1997.2543

Quémart, P., Casalis, S., & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of Experimental Child Psychology, 109*(4), 478-496. https://doi.org/10.1016/j.jecp.2011.02.008

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rastle, K. (2019). The place of morphology in learning to read in English. *Cortex, 116*, 45-54. https://doi.org/10.1016/j.cortex.2018.02.008

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090-1098. https://doi.org/10.3758/BF03196742

Raven, J. C. (1949). *Progressive matrices (1947), sets A, Ab, B: board and book forms.* London: Lewis.

Raven, J. (2003). Raven progressive matrices. In *Handbook of nonverbal assessment* (pp. 223-237). Springer, Boston, MA.

Rayner, K. (1986). Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology*, *41*(2), 211-236. https://doi.org/10.1016/0022-0965(86)90037-8

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457-1506. https://doi.org/10.1080%2F17470210902816461

Rayner, K., Binder, K. S., Ashby, J., & Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research*, *41*(7), 943-954. https://doi.org/10.1016/S0042-6989(00)00310-2

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241-255. https://doi.org/10.1207/s1532799xssr1003_3

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements?. *Vision Research*, *16*(8), 829-837. https://doi.org/10.1016/0042-6989(76)90143-7

Rayner, K., & Pollatsek, A. (1981). Eye movement control during reading: Evidence for direct control. *Quarterly Journal of Experimental Psychology*, *33*(4), 351-373. https://doi.org/10.1080/14640748108400798

Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H. S., White, S. J., & Rayner, K. (2013). Using EZ Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, *33*(2), 110-149. https://doi.org/10.1016/j.dr.2013.03.001

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125. https://doi.org/10.1037/0033-295X.105.1.125

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E–Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*(1), 4-22. https://doi.org/10.1016/j.cogsys.2005.07.002

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ Reader model. *Vision Research, 39*(26), 4403-4411. https://doi.org/10.1016/S0042-6989(99)00152-2

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445. https://doi.org/10.1017/S0140525X03000104

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon, 12*(2), 263-282. https://doi.org/10.1075/ml.17009.sch

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia, 67*(2), 147-162. https://doi.org/10.1007/s11881-016-0136-0

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review, 125*(6), 969. https://doi.org/10.1037/rev0000119

Sperlich, A., Schad, D. J., & Laubrock, J. (2015). When preview information starts to matter: Development of the perceptual span in German beginning readers. *Journal of Cognitive Psychology, 27*(5), 511-530. https://doi.org/10.1080/20445911.2014.993990

Spichtig, A., Pascoe, J., Ferrara, J., & Vorstius, C. (2017). A comparison of eye movement measures across reading efficiency quartile groups in elementary, middle, and high school students in the US. *Journal of Eye Movement Research, 10*(4), 5. https://doi.org/10.16910/jemr.10.4.5

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behaviour, 14*(6), 638-647. https://doi.org/10.1016/S0022-5371(75)80051-X

Taft, M., & Nguyen-Hoan, M. (2010). A sticky stick? The locus of morphological representation in the lexicon. *Language and Cognitive Processes*, *25*(2), 277-296. https://doi.org/10.1080/01690960903043261

Tiffin-Richards, S. P., & Schroeder, S. (2015). Children's and adults' parafoveal processes in German: Phonological and orthographic effects. *Journal of Cognitive Psychology, 27*(5), 531-548. http://dx.doi.org/10.1080/20445911.2014.999076

Tiffin-Richards, S. P., & Schroeder, S. (2018). The development of wrap-up processes in text reading: A study of children's eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(7). http://dx.doi.org/10.1037/xlm0000506

Tiffin-Richards, S. P., & Schroeder, S. (2020). Context facilitation in text reading: A study of children's eye movements. *Journal of Experimental Psychology. Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0000834

Traficante, D., Marelli, M., & Luzzatti, C. (2018). Effects of reading proficiency and of base and whole-word frequency on reading noun-and verb-derived words: an eye-tracking study in Italian primary school children. *Frontiers in Psychology*, *9*, 2335. https://doi.org/10.3389/fpsyg.2018.02335 von Koss Torkildsen, J., Arciuli, J., & Wie, O. B. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences, 69,* 60-68. https://doi.org/10.1016/j.lindif.2018.11.003

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. https://doi.org/10.1037/0033-2909.131.1.3

Chapter III

Algorithms for the Automated Correction of Vertical Drift in Eye-Tracking Data[1]

Reading is a fundamental skill for navigating modern society and, as such, is subject to intense study in the cognitive and language sciences. Among the many tools that researchers use to investigate reading in the laboratory, eye tracking occupies a prominent position. Using this technique, participants' eye movements may be recorded as they read written material, providing a window into the relevant cognitive processes as they unfold. Technological advancements in eye tracking, particularly from the 1970s (see, e.g., Rayner, 1998), have allowed researchers to collect increasingly accurate measures of eye movements during reading tasks, leading to great improvements in the investigation of the cognitive processes underlying reading and reading acquisition.

Many eye-tracking studies involve the reading of single words or sentences. For example, researchers may embed target words into different sentence contexts and manipulate predictability (e.g., Rayner et al., 2001), display isolated words to gain insight into how a reader's eye moves when processing a word (e.g., Vitu et al., 2004), or reveal parts of words in a gaze-contingent fashion to investigate parafoveal processing (e.g., Schotter et al., 2012). Sentence reading experiments have also been essential in revealing the cognitive processes behind different levels of written language processing, from the width of the perceptual span

[1] This chapter is a slightly adapted version of a manuscript accepted for publication, to appear as: Carr, J. W., Pescuma, V. N., Furlan, M., Ktori, M., & Crepaldi, D. (2021). Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01554-0

(e.g., Blythe et al., 2009; Rayner, 1986) to the effects that word length and frequency have on eye movements (e.g., Joseph et al., 2009; Tiffin-Richards & Schroeder, 2015), as well as the effects of syntactic (e.g., Frazier & Rayner, 1982; Pickering & Traxler, 1998) and lexical (e.g., Sereno et al., 2006) ambiguity.

In our everyday experience, however, we often do not encounter sentences in isolation; a good part of our reading experience involves connected text that is distributed over multiple lines. Therefore, experiments based on paragraph reading also provide insight into the reading experience, while allowing us to address levels of processing that are simply not available when one reads a single sentence, such as the role of broader context or the integration of syntactic relations across sentence boundaries (Jarodzka & Brand-Gruwel, 2017). Indeed, studies of multiline reading have become more prevalent in recent years, with researchers using passage reading tasks to investigate, for example, the effect of text- and participant-level characteristics on eye movements (Kuperman et al., 2018) or of contextual facilitation on developing readers' eye movements (Tiffin-Richards & Schroeder, 2020). Several multiline-reading datasets have also been released, including GECO (Cop et al., 2017), MECO (Kuperman et al., under review), and Provo (Luke & Christianson, 2018).

A technical issue that arises from the particular circumstances of multiline reading is so-called "vertical drift," which we define as the progressive displacement of fixation registrations on the vertical axis *over time*. In other words, fixations may be recorded above or below the line of text that the participant was actually reading, and the degree and directionality of this error may fluctuate dynamically with each subsequent fixation, making it nontrivial to eliminate. Fig. 1a depicts a reading trial exhibiting vertical drift phenomena; in this case, fixations—especially those on the left-hand side—are recorded around one line higher than where the reader was actually fixating, but they also tend to slope down to the right such that fixations on the right hand side seem to be better aligned.

Vertical drift can occur quite unpredictably, even following good quality calibration, and it is likely caused by spatial phenomena such as degraded eye tracker calibration at the corners of the screen, or temporal phenomena such as subtle movements in head position or pupil dilation, which can be difficult to control for, even in a laboratory setting. Such sources of measurement error are often exacerbated in the context of multiline reading because, in comparison to single words or sentences, passages of text are distributed over a larger portion of the screen, including areas where general calibration may be worse, and they take longer to read, during which time calibration may begin to degrade. There are also less frequent opportunities to recalibrate the device during passage reading, since this can only be performed between trials or pages and not during the reading of a passage.

Whatever the cause and however it manifests itself, vertical drift will ultimately have a negative impact on the analysis of eye-tracking data because fixations will be mapped to words that were not actually being fixated at a given point in time (as we can see in Fig. 1a). It is therefore incumbent on the researcher to recognize such issues when they occur and to take corrective measures. Specialized software packages, such as EyeLink Data Viewer (SR Research, Toronto, Canada) or EyeDoctor (UMass Eyetracking Lab, Amherst, USA), provide the ability to manually move fixations, either individually or in small batches. However, manual realignment can be very time-consuming and is likely to be error-prone. In particular, the realignment process can be greatly complicated by other sources of noise or idiosyncratic reading behaviors. For example, Fig. 1b depicts a reading trial by a child reader; in this case, not only are the fixations affected by drift issues, but there are also various natural reading behaviors, such as within- and between-line regressions, which add an additional layer of complexity to the task of realignment, not to mention the baseline level of noise and unusual features such as the arching sequence of fixations targeting line 4.

A number of methods have previously been developed to automate post-hoc vertical drift correction. *FixAlign*, an R package developed by Cohen (2013), is currently the most well-established method in the experimental psychology community, although other methods have recently been proposed by Schroeder (2019) and Špakov et al. (2019). In addition, there is a disparate body of work from several subfields of computer science, such as biometrics (Abdulin & Komogortsev, 2015), educational technology (Hyrskykari, 2006), and user-interface design (Beymer & Russell, 2005), in which various ad-hoc algorithms have been reported (see also Carl, 2013; Lima Sanches et al., 2015; Lohmeier, 2015; Martinez-Gomez et al., 2012; Mishra et al., 2012; Nüssli, 2011; Palmer & Sharif, 2016; Sibert et al., 2000; Yamaya et al., 2017).

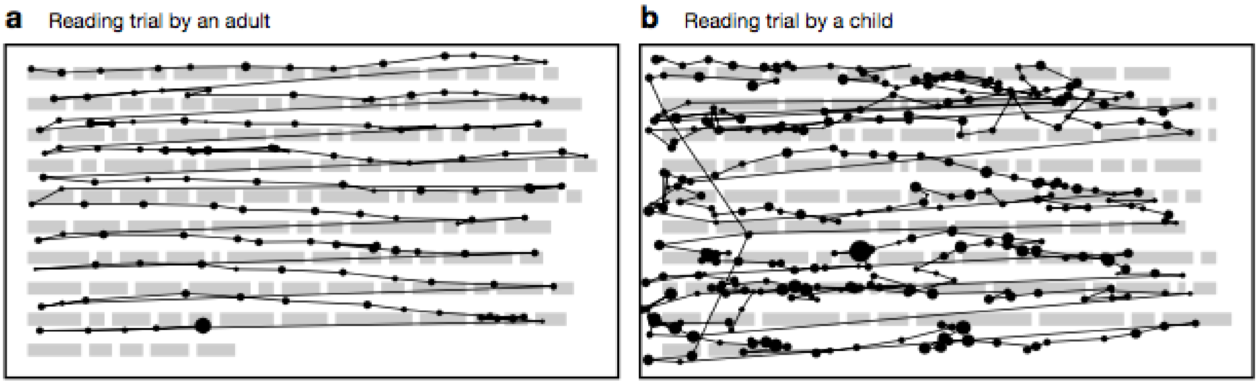


Figure 1 Example reading trials from (**a**) an adult participant and (**b**) a child participant taken from Pescuma et al. (in prep.). Each dot represents a fixation and the size of the dots represents duration. The adult trial exhibits upward shift, especially in the lower left part of the passage. The child trial is extremely noisy and exhibits not just vertical drift issues but also many natural reading phenomena that will pose challenges to the algorithms.

These reported methods can be difficult to evaluate and use because they vary widely in terms of their availability, design choices, implementation languages, usability, level of documentation, expected input data, and the extent to which they rely on project-specific heuristics or particular eye tracker hardware. Furthermore, these methods have largely been developed in isolation from each other, and there has been little attempt to systematically evaluate them, so drift correction software is moving forward blindly without an evidence base to support new directions. In this paper, we attempt to classify the reported methods into ten major approaches, which we formalize as ten simple algorithms that adopt a consistent design model. In other words, we do not attempt to evaluate existing software implementations; rather, we explore the spectrum of drift correction algorithms by isolating and evaluating the core principles on which previous methods have been based. Our goal is to provide a systematic comparison of these algorithms in order to guide researchers' choices about the most suitable methods and to lay a solid foundation for future drift correction software.

To be clear, the algorithms we consider in this paper are restricted to one specific problem. Firstly, we only consider algorithms designed for the ordinary reading of passages of text; other uses of eye tracking, such as visual search and scene perception, can also undergo drift correction, but the methods required are quite different (see, e.g., Vadillo et al., 2015; Zhang & Hornof, 2011, 2014). Similarly, the reading of source code has received some attention, but the affordances and constraints in this domain are quite different from ordinary linguistic reading (see, e.g., Nüssli, 2011; Palmer & Sharif, 2016). Secondly, we only consider the problem of post-hoc correction; vertical drift can also be corrected in real time, but this imposes a more restrictive set of constraints that are better handled by other types of algorithm (see, e.g., Hyrskykari, 2006; Sibert et al., 2000). Thirdly, we only consider fully automated algorithms that do not require human supervision.

The paper proceeds in four main sections. First, we outline the algorithms. Second, we test the algorithms on simulated fixation sequences afflicted with various types of measurement error. Third, we test the algorithms on an eye-tracking dataset (two examples from which are presented in Fig. 1). And finally, we discuss the major properties of the algorithms, provide guidance to researchers about their use, and suggest ways in which they can be improved further. All code and data required to reproduce the analyses reported in this paper, as well as Matlab/Octave, Python, and R implementations of the algorithms, are available from the public data archive associated with this paper: https://doi.org/10.17605/OSF.IO/7SRKG

Algorithms

In this section, we describe ten algorithms for the automated, post-hoc correction of vertical drift. The reader may also wish to refer to Supplementary Item 1 where we present the algorithms in pseudocode alongside other technical details.

Attach

The `attach` algorithm is the simplest of the algorithms considered in this paper. The algorithm simply attaches each fixation to its closest line. While this has the benefit of being extremely simple, it is generally not resilient to the kinds of drift phenomena described above. However, `attach` serves as a useful baseline algorithm, since it essentially corresponds to an eye-tracking analysis in which no correction was performed—a standard analysis of eye-tracking data would simply map fixations to the closest words or other areas of interest. We return to this point later in the paper.

Chain

The `chain` algorithm is based closely on one of the methods implemented in the R package *popEye* (Schroeder, 2019) and can be seen as an extension of `attach`. Fixations are first linked together into "chains"—sequences of consecutive fixations that are within a specified x and y distance of each other. Fixations within a chain are then attached to whichever line is closest to the mean of their y values. This procedure is similar to the slightly more complex methods reported by Hyrskykari (2006) and Mishra et al. (2012), so we consider these to be special cases of `chain`.

The `chain` algorithm generally provides better performance over `attach` by exploiting the sequence's order information. A disadvantage of the method, however, is that it is necessary to specify appropriate thresholds that determine when a new chain begins. If these thresholds are set too low, `chain` becomes equivalent to `attach`; if they are set too high, `chain` will group large numbers of fixations together and force them onto a single inappropriate line. By default, popEye sets the *x* threshold to 20 × the font height and the *y* threshold to 2 × the font height. It is not exactly clear how these defaults were chosen, but we would tentatively suggest that the *x* threshold should be set to approximately one long saccade length (we use 192 px), and the *y* threshold to around half a line height (we use 32 px).

Cluster

The `cluster` algorithm is also based on one of the methods implemented in popEye (Schroeder, 2019). `cluster` applies *k*-means clustering[2] to the *y* values of all fixations in order to group the fixations into *m* clusters, where *m* is the number of lines in the passage. Once each fixation has been assigned to a cluster, clusters are mapped to lines based on the mean *y* values of their constituent fixations: The cluster with the smallest mean *y* value is assigned to line one and so forth.

Unlike `attach` and `chain`, `cluster` does not assign fixations to the closest line in absolute terms; instead, it operates on the principle that fixations with similar *y* values must belong to the same line regardless of how far away that line might be. As such, the algorithm

[2] In both the cluster and split algorithms, we used the KMeans function from the Python library Scikit-learn (Pedregosa et al., 2011). There are many variations of *k*-means clustering, which we have not systematically compared.

generally handles drift issues quite well. However, `cluster` will often not perform well if there is even quite mild overlap between fixations from different lines. In addition, since *k*-means clustering is not guaranteed to converge on the same set of clusters on every run, the `cluster` algorithm is nondeterministic and can be somewhat unpredictable across multiple runs on the same reading trial, which is an important consideration from the point of view of reproducible research output.

Compare

The `compare` algorithm is directly based on the method reported by Lima Sanches et al. (2015) and is very similar to the more complex methods described by Yamaya et al. (2017). The fixation sequence is first segmented into “gaze lines” by identifying the return sweeps—long saccades that move the eye from the end of one line to the start of the next. The algorithm considers any saccade that moves from right to left by more than some threshold value (we use 512 px) to be a return sweep. Gaze lines are then matched to text lines based on a measure of similarity between them. Lima Sanches et al. (2015) considered three measures of similarity and found dynamic time warping (DTW; Sakoe & Chiba, 1978; Vintsyuk, 1968) to be the best method (we discuss DTW in more detail later in this section). Similarly, Yamaya et al. (2017) use the closely related Needleman–Wunsch algorithm (Needleman & Wunsch, 1970).

The gaze lines and text lines are compared in terms of their *x* values under the assumption that the fixations in a gaze line should have a good horizontal alignment with the centers of the words in the corresponding text line. Relying only on the *x* values helps the algorithm overcome vertical drift issues, but it is also problematic because in many standard reading scenarios the lines of text in a passage tend to be horizontally similar to each other; each line tends to contain a similar number of words that are of a similar length, resulting in potential ambiguity about how gaze lines and text lines should be matched up. To alleviate this issue, both Lima Sanches et al. (2015) and Yamaya et al. (2017) only compare the gaze line to a certain number of nearby text lines (we set this parameter to 3, which is effectively the closest line plus one line above and one line below).

Merge

The `merge` algorithm is closely based on the post-hoc correction method described by Špakov et al. (2019). The algorithm begins by creating "progressive sequences"— consecutive fixations that are sufficiently close together. This is similar to `chain`, except that the sequences are strictly progressive (they only move forward), so a regression will initiate a new progressive sequence. The original method uses several parameters to define what constitutes "sufficiently close together," but here we boil this down to a single parameter, the `y_thresh`, which determines how close the *y* values of two consecutive fixations must be to be considered part of the same progressive sequence (we use 32 px).

Once these sequences have been created, they are repeatedly merged into larger and larger sequences until the number of sequences is reduced to m, one for each line of text. On each iteration of the merge process, the algorithm fits a regression line to every possible pair of sequences (with the proviso that the two sequences must contain some minimum number of fixations). If the absolute gradient of the regression line or its error (root-mean-square deviation) is above a threshold (we use 0.1 and 20 respectively), the candidate merger is abandoned. The intuition here is that, if two sequences belong to the same text line, the regression line fit to their combined fixations will have a gradient close to 0 and low regression error. Of the candidate mergers that remain, the pair of sequences with the lowest error are merged and added to the pool of sequences, replacing the original two sequences and reducing their number by one. This process is repeated until no further mergers are possible.

The algorithm then enters the next “phase” of the process, in which the criteria are slightly relaxed, allowing more mergers to occur. These phases could in principle be defined by the user, but we follow the four-phase model reported by Špakov et al. (2019), which effectively builds a set of heuristics into the algorithm. In Phase 1, the first and second sequences must each contain a minimum of three fixations to be considered for merging; in Phase 2, only the second sequence must contain a minimum of three fixations; in Phase 3, there is no minimum number of fixations; and in Phase 4, the gradient and regression error criteria are also entirely removed. Of course, as soon as the number of sequences is reduced to *m* the algorithm exits the merge process, so not all four phases will necessarily be required. Finally, the set of *m* sequences is matched to the set of text lines in positional order: The sequence with the smallest mean *y* value is mapped to line one and so forth.

A similar sounding method is reported by Beymer and Russell (2005) whose technique is based on “growing” a gaze line by incrementally adding fixations until this results in a poor fit to a regression line, at which point a new gaze line is begun. However, the description of the method lacked sufficient detail for us to consider it further.

Regress

The `regress` algorithm, which is closely based on Cohen’s (2013) R package *FixAlign*, treats the fixations as a cloud of unordered points and fits *m* regression lines to this cloud. These regression lines are parameterized by a slope, vertical offset, and standard deviation, and the best parameters are obtained by minimizing[3] an objective function that determines the overall fit of the lines through the fixations. The algorithm has six free parameters which are used to specify the lower and upper bounds of the slope, offset, and

[3] In both the regress and stretch algorithms, we used the minimize function from the Python library SciPy (Virtanen et al., 2020). We have not systematically compared the choice of optimizer settings.

standard deviation. Here we directly adopt FixAlign's defaults: [−0.1, 0.1], [−50, 50], and [1, 20] respectively. Once the *m* best-fitting regression lines are obtained, regress assigns each fixation to the highest-likelihood regression line, which itself is associated with a text line.

regress tracks FixAlign very closely, except that we did not implement the "run rule," an option that is switched on by default in FixAlign. This option maps ambiguous fixations to the same line as the surrounding fixations, if the surrounding fixations were classified unambiguously (Cohen, 2013, p. 680). Cohen's run rule is a more general method that could in principle be applied to the output of any algorithm, so in the interest of isolating the core concept of FixAlign and comparing all algorithms on a level playing field, we did not to implement the option here.

regress has some conceptual similarities with merge but differs in several important respects. Notably, regress takes a top-down approach, where merge is more bottom-up, and the regression lines that regress fits to the fixations cannot take independent values—it is assumed that all fixations are sloping in the same direction, with the same vertical offset, and with the same amount of within-line variance. In addition, unlike merge, regress does not utilize the order information; instead, like cluster, it views the fixations as a collection of unordered points.

Segment

The segment algorithm is a slight simplification of the method described by Abdulin and Komogortsev (2015). The fixation sequence is first segmented into *m* disjoint subsequences based on the *m*−1 most extreme backward saccades along the *x*-axis (i.e., the saccades that are most likely to be return sweeps). These subsequences are then mapped to the lines of text chronologically, under the assumption that the lines of text will be read in order. Abdulin and Komogortsev (2015) do not state precisely how they identify the return sweeps, but it seems they potentially allow for more than *m* subsequences to be identified, in which case, rereadings of a previous line, based on a threshold level of similarity, are discarded. The version of the algorithm considered here does not discard any fixations and instead always identifies exactly *m* subsequences.

The advantage of this general approach, as emphasized by Abdulin and Komogortsev (2015), is that the *y* values of the fixations are completely ignored, rendering any vertical drift entirely invisible to the algorithm. However, the approach does not allow for the possibility that the lines of text might be read out of order or that a line of text might be read more than once, which is not uncommon in normal reading behavior. Therefore, the great strength of `segment`—its identification of *m* consecutive subsequences, permitting a chronological, as opposed to positional, mapping—is also its great weakness: If a large regression is mistakenly identified as a return sweep, this will lead to a catastrophic off-by-one error in subsequent line assignments.

Split

As far as we know, the `split` algorithm takes an approach that is distinct from any thing previously reported, although it is conceptually similar to `segment`. Like `segment`, the `split` algorithm works on the principle of splitting the fixation sequence into subsequences by first identifying the return sweeps. However, `split` is not restricted to finding exactly *m*−1 return sweeps; instead, it identifies the most likely set of return sweeps, however many that turns out to be. There are various ways of approaching this classification problem, but here we use *k*-means clustering to partition the set of saccades into exactly two clusters. Since return sweeps are usually highly divergent from normal saccades (i.e., a return sweep is usually represented by a large negative change on the *x*-axis), one of the two clusters will invariably contain the return sweeps, which can then be used to split the fixation sequence into subsequences. However, since this is not guaranteed to produce *m*−1 return sweeps (and therefore *m* subsequences), an order-based mapping is not possible, so `split` must use absolute position: Subsequences are mapped to the closest text lines in absolute terms. `split` has the advantage of generally finding all true return sweeps, and even if it identifies some false positives, the resulting subsequences can still be mapped to the appropriate lines by absolute position. However, this also means the algorithm is less resilient to vertical drift issues.

Stretch

The `stretch` algorithm is loosely based on the method proposed by Lohmeier (2015) and shares some similarities with Martinez-Gomez et al. (2012) and Nüssli (2011). Lohmeier's (2015) original method was designed for the reading of source code and therefore takes advantage of the fact that code has very irregular line lengths and indentation levels. The method works by finding an *x*-offset, *y*-offset, and scaling factor that, once applied to the fixations, minimizes alignment error between the fixations and lines of text.

The framework we adopt herein never adjusts the *x* values, and we also assume that an ordinary passage of text is being read, so line length is substantially more constant than during code reading and therefore less informative. Therefore, we simplified the original method by dispensing with all dependencies on the *x* values. Instead, `stretch` finds a *y*-offset, o∗, and a vertical scaling factor, s∗, that minimizes the sum absolute difference between the corrected fixation positions (fy s + o) and the corrected fixation positions once attached to their closest lines. The equations presented in Lohmeier (2015, pp. 37–38) therefore simplify to:

$$o^{\star}, s^{\star} = \operatorname*{arg\,min}_{o,s} \sum_{f \in F} \left|(f_y s + o) - \mathrm{attach}(f_y s + o)\right|, \tag{1}$$

where attach(·) returns the *y*-axis position of the nearest line of text. In other words, the algorithm seeks a transformation of the fixations that results in minimal change following the application of `attach`.

To constrain the minimization problem, the user must specify appropriate lower and upper bounds for the offset and scaling factor, resulting in four free parameters. Here we adopt offset bounds of [−50, 50], following the `regress` algorithm, and scaling factor bounds of [0.9, 1.1]. Effectively, this means the algorithm can move the set of fixations up or down by up to 50 pixels and stretch their positions on the vertical axis by between 90% and 110%. While approaching the problem from a different angle, `stretch` is computationally similar to `regress`, except that it emphasizes systematic offset issues rather than systematic slope issues.

Warp

The final algorithm we consider, `warp`, is novel to this paper, although it is mostly a wrapper around a preexisting algorithm—dynamic time warping (DTW; Sakoe & Chiba, 1978; Vintsyuk, 1968). DTW was used by the `compare` algorithm to provide a measure of dissimilarity between a gaze line and a text line. To our knowledge, however, there have been no previous reports of DTW being used directly to align fixations to text lines. This is somewhat surprising because DTW is the natural computational choice for tackling drift and alignment problems. The closest previously-described method is Carl (2013), who uses a basket of reading-related measures to place a cost on different paths through a lattice of fixation-to-character mappings and selects the path with minimal cost. This is quite complex, however, and we consider it to be a special case of `warp`, which is a direct application of the standard DTW algorithm to eye-tracking data.

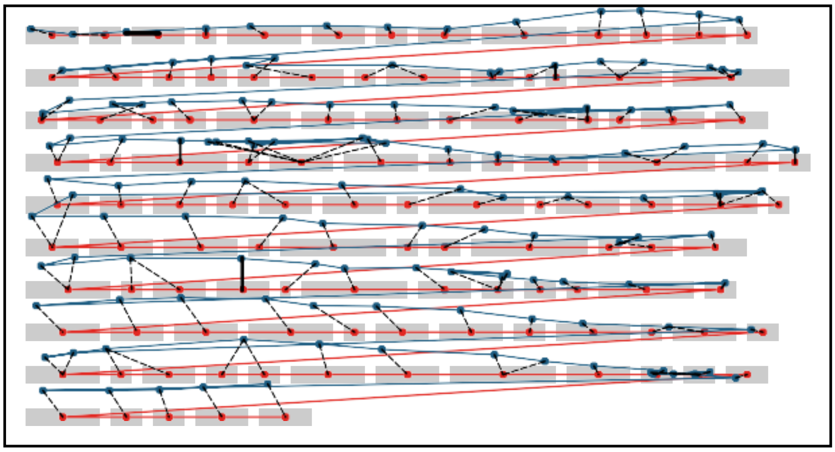
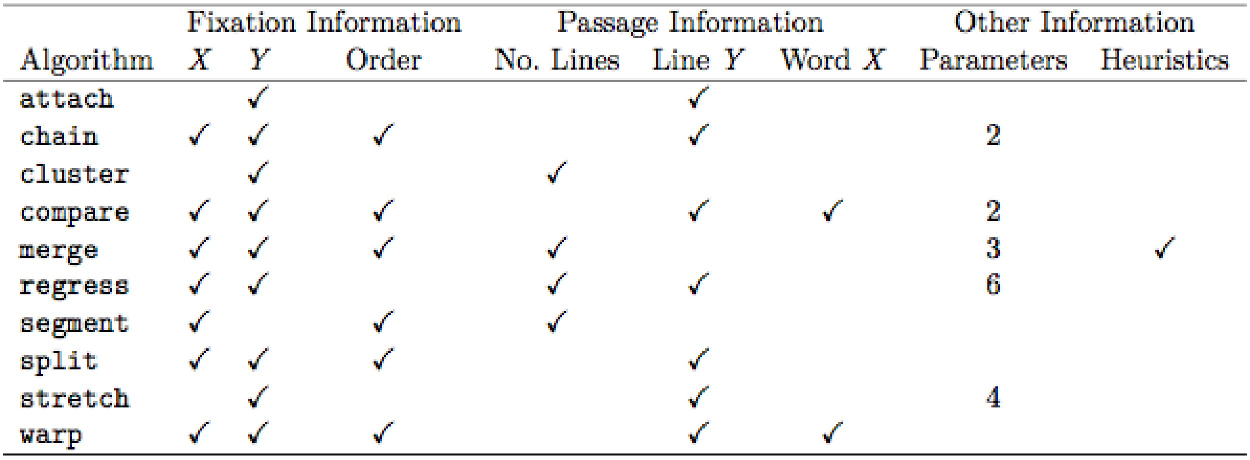


Figure 2 Illustration of the `warp` algorithm. The veridical fixation sequence is represented in blue, and the expected fixation sequence (the sequence of word centers) is represented in red. The dashed black lines show the DTW warping path—the optimal way to align the two sequences, such that the sum of the Euclidean distances between matched points (i.e., the sum of the dashed lines) is minimized.

DTW is typically useful when you have two sequences, not necessarily of the same length, and you want to (a) calculate how similar they are (as is the case in the `compare` algorithm) or (b) align the two sequences by mapping each element in one sequence to a corresponding element in the other. For example, DTW may be used to calculate the similarity between a signature, which can be expressed as a sequence of *xy*-coordinates over time, and a reference signature (e.g., Lei & Govindaraju, 2005; Riesen et al., 2018). Importantly, the two sequences do not need to be perfectly matched in terms of overall magnitude or patterns of acceleration and deceleration for a good alignment to be found. In the case of signature verification, for example, it does not matter if the candidate signature has the same size as the reference or that it was drawn at the same speed, what matters is that there is a good match in the overall shape and that the strokes were drawn in the same order. DTW finds many other applications in, for example, genomics (Aach & Church, 2001), medicine (Caiani et al., 1998), and robotics (Vakanski et al., 2014).

In order to use DTW to realign the fixation sequence to the text, we first need to specify an expected fixation sequence. Since we expect the reader to traverse the passage from left to right and from top to bottom, we can use the series of word centers as the expected sequence, under the assumption that readers will target the centers of words (O'Regan et al., 1984). Given the expected and veridical sequences as inputs, the DTW algorithm finds the optimal way to nonlinearly warp the sequences on the time dimension such that the overall Euclidean distance between matched points across the two sequences is minimized, while maintaining a monotonically increasing mapping.[4] In the "warping path" that results from this process, every fixation is mapped to one or more words and every word is mapped to one or more fixations (see Fig. 2 for an example). It is then simply a case of assigning each fixation to whichever line its mapped word(s) belong(s) to. In the unlikely event that the mapped words belong to different lines, the majority line wins or an arbitrary choice is made in the case of ties.

Table 1 Information utilized by the algorithms



	Fixation Information			Passage Information			Other Information	
Algorithm	X	Y	Order	No. Lines	Line Y	Word X	Parameters	Heuristics
attach		✓			✓			
chain	✓	✓	✓		✓		2	
cluster		✓		✓				
compare	✓	✓	✓		✓	✓	2	
merge	✓	✓	✓	✓			3	✓
regress	✓	✓		✓	✓		6	
segment	✓		✓	✓				
split	✓	✓	✓		✓			
stretch		✓			✓		4	
warp	✓	✓	✓		✓	✓		

[4] Specifically, the first fixation must be mapped to (at least) the first word; the last fixation must be mapped to (at least) the last word; every other fixation must be mapped to at least one word; and, if fixation *i* is mapped to word *j*, then fixation *i*+1 must be mapped to word(s) ≥ *j*. And vice versa for the mapping from words to fixations.

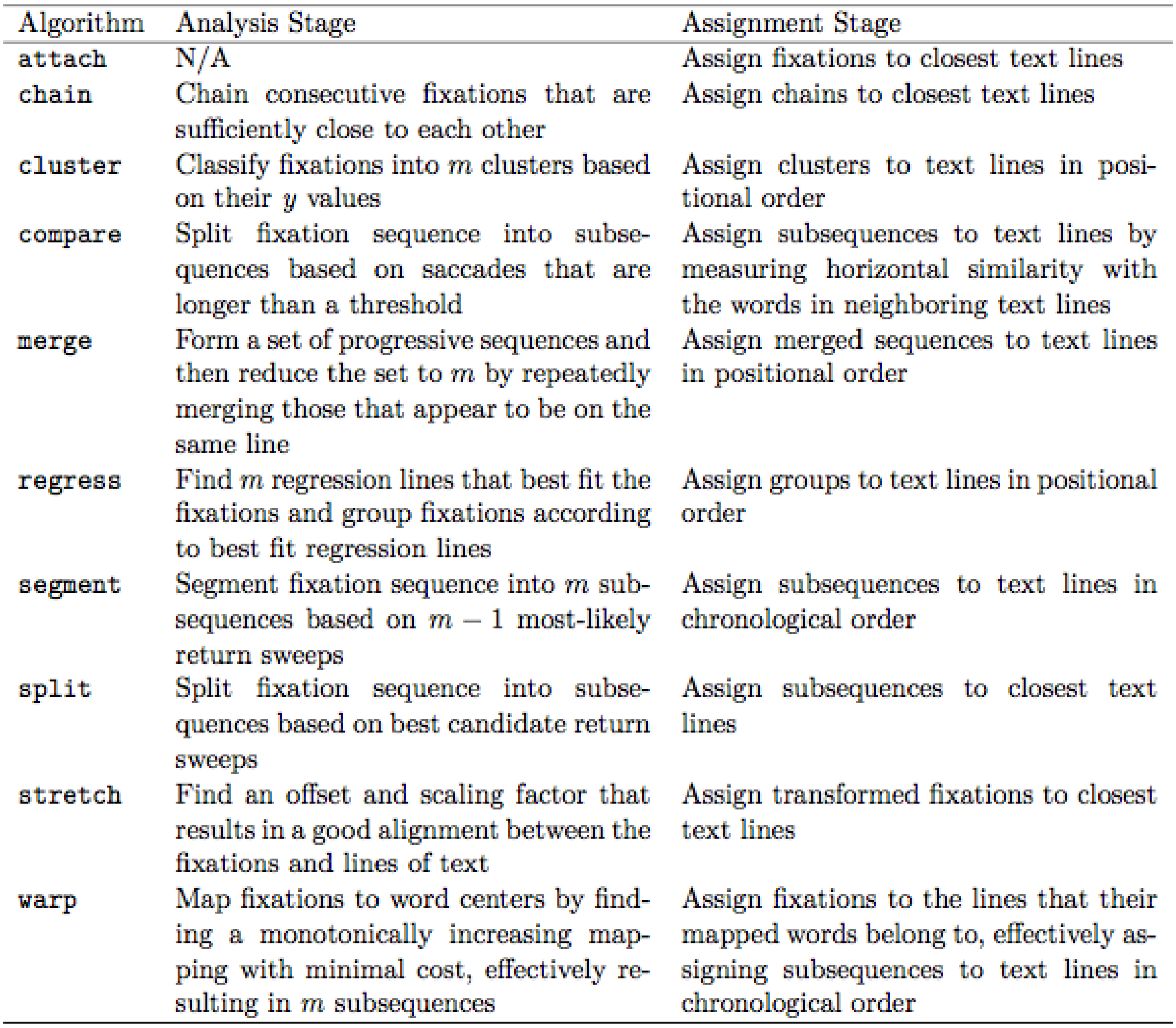
If the final fixation on line *i* were mapped to the first word on line *i*+1, this would result in a large increase in the overall cost of the mapping, so line changes act as major clues about the best alignment. The upshot of this is that `warp` effectively segments the fixation sequence into exactly m subsequences, which are mapped to the lines of text in chronological order. In this sense, `warp` behaves very much like `segment`. However, the additional benefit of `warp` is that it can simultaneously consider different possibilities about which saccades are the return sweeps, selecting only those that result in the best fit to the passage at a global level. Nevertheless, `warp` is ultimately limited by the veracity of the expected fixation sequence, which encodes one particular way of reading the passage—line by line from start to end. If the reader deviates from this assumption (e.g., by rereading or skipping lines), `warp` can fail to correctly assign fixations to lines.

Summary

In this section we have described ten algorithms for aligning a fixation sequence to a multiline text, each of which takes a fundamentally different approach. A summary of the information utilized by the algorithms is provided in Table 1; each algorithm uses at least one piece of information about the fixations and at least one piece of information about the passage, and some also rely on additional parameters set by the user or built-in heuristics.

Broadly speaking, the algorithms proceed in three stages, analysis, assignment, and update the one exception being `attach` which has no analysis stage. In the analysis stage, the fixations are analyzed, transformed, or classified in some sense. The rationale behind this process varies by algorithm, but in general the algorithms can be categorized into those that classify the fixations into *m* groups (i.e., one group per text line; `cluster`, `merge`, `regress`, `segment`, and `warp`) and those that do not (`attach`, `chain`, `compare`, `split`, and `stretch`).

Table 2 Summary of the analysis and assignment stages of each algorithm



Algorithm	Analysis Stage	Assignment Stage
`attach`	N/A	Assign fixations to closest text lines
`chain`	Chain consecutive fixations that are sufficiently close to each other	Assign chains to closest text lines
`cluster`	Classify fixations into m clusters based on their y values	Assign clusters to text lines in positional order
`compare`	Split fixation sequence into subsequences based on saccades that are longer than a threshold	Assign subsequences to text lines by measuring horizontal similarity with the words in neighboring text lines
`merge`	Form a set of progressive sequences and then reduce the set to m by repeatedly merging those that appear to be on the same line	Assign merged sequences to text lines in positional order
`regress`	Find m regression lines that best fit the fixations and group fixations according to best fit regression lines	Assign groups to text lines in positional order
`segment`	Segment fixation sequence into m subsequences based on $m - 1$ most-likely return sweeps	Assign subsequences to text lines in chronological order
`split`	Split fixation sequence into subsequences based on best candidate return sweeps	Assign subsequences to closest text lines
`stretch`	Find an offset and scaling factor that results in a good alignment between the fixations and lines of text	Assign transformed fixations to closest text lines
`warp`	Map fixations to word centers by finding a monotonically increasing mapping with minimal cost, effectively resulting in m subsequences	Assign fixations to the lines that their mapped words belong to, effectively assigning subsequences to text lines in chronological order

In the assignment stage, the fixations are assigned to text lines. If the analysis stage does *not* produce *m* groups, then assignment must be based on absolute position (or similarity in the case of compare, although it still uses absolute position to select neighboring lines to compare to). If the analysis stage *does* produce *m* groups, then they can be assigned to text lines based on order; this generally allows for better handling of vertical drift because absolute position is ignored. In the case of `cluster`, `merge`, and `regress`, which produce unordered groups at the analysis stage, groups are matched to text lines based on the order in which they are positioned vertically (i.e., mean *y* value). In the case of `segment` and `warp`, the groups are assigned to text lines in chronological order, which is only possible because these two algorithms produce subsequences that inherit the order of the original fixation sequence. An overview of the analysis and assignment methods is provided in Table 2 for quick reference.

Finally, in the update stage, the original fixation sequence is modified to reflect the line assignments identified in the previous stage. In the versions of the algorithms reported in this paper, we always use the same update approach: The *y* values of the fixations are adjusted to the *y* values of the assigned lines, while the *x* values and the order of the fixations are always left untouched. In principle, however, there are other ways of performing the update stage (e.g., retaining the original *y*-axis variance or discarding ambiguous fixations).

Performance on Simulated Data

We now test the ability of each algorithm to correctly recover the intended lines from simulated fixation sequences. These fixation sequences are simulated with particular characteristics, allowing us to understand how the algorithms respond to specific, isolated phenomena.

Method

In each simulation, we generate a passage of "Lorem ipsum" dummy text consisting of between 8 and 12 lines with up to 80 characters per line and 64 px of line spacing. We then generate a fixation sequence consisting of one fixation for every word in the passage: The *x* value of a fixation (f_x) is set randomly within the word; the *y* value of a fixation (f_y) is calculated according to:

$$f_y = \mathcal{N}(l_y, d_{\text{noise}}) + f_x d_{\text{slope}} + l_y d_{\text{shift}}, \qquad (2)$$

where l_y is the vertical center point of the intended line—the *y* value that the reader is targeting. This models three types of distortion: noise, slope, and shift. Additionally, we simulate two types of regression that are characteristic of normal reading behavior but which can nevertheless disrupt algorithmic correction. Together, these five phenomena are illustrated in Fig. 3 and described below.

Noise Distortion

The noise distortion parameter, d_{noise}, controls the standard deviation of the normally distributed noise around the intended line and represents imperfect targeting by the reader and/or measurement error. In our exploration of this parameter, we use values of $\mathrm{d}_{\text{noise}} = 0$, representing no noise, through $d_{\text{noise}} = 40$, representing extreme noise. The noise parameter is also a proxy for line spacing (raising the noise level effectively corresponds to tightening the line spacing), so this parameter also provides an indication of how the algorithms will perform under different degrees of line spacing.

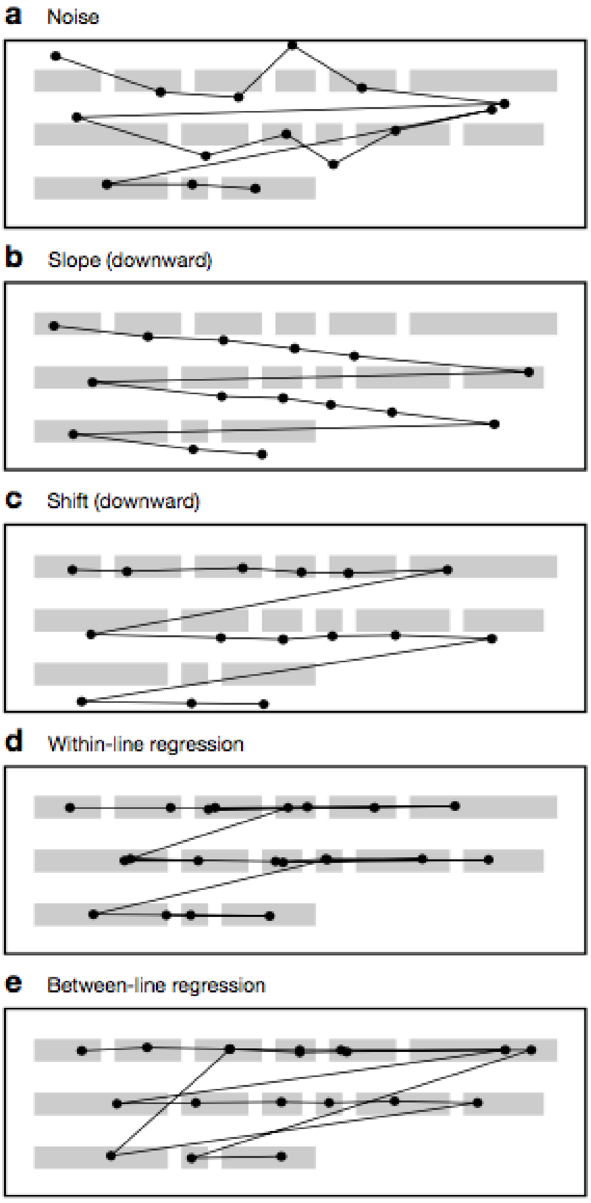


Figure 3 Example simulated fixation sequences under five phenomena considered in this paper. The algorithms must overcome these phenomena in order to correctly infer the intended line of each fixation.

Slope Distortion

The slope distortion parameter, d_{slope}, controls the extent to which fixations progressively move downward as the reader moves from left to right across the passage; fixations on the left edge of the passage will be correctly located, but for every one pixel the reader moves to the right, the fixations will drift downward by d_{slope} pixels. Unlike noise, this is solely attributable to measurement error. In our exploration of this parameter, we use values of $d_{\text{slope}} = -0.1$, representing extreme upward slope, through $d_{\text{slope}} = 0.1$, representing extreme downward slope.

Shift Distortion

The shift distortion parameter, d_{shift}, controls the extent to which fixations progressively move downward as the reader moves from one line to the next; fixations on the first line will be correctly located, but for every one pixel of intentional downward movement, the fixations will drift downward by a further d_{shift} pixels. Like slope, this represents systematic measurement error. Our exploration of this parameter uses values of $d_{\text{shift}} = -0.2$, representing extreme upward shift, through $d_{\text{shift}} = 0.2$, representing extreme downward shift.

Within-Line Regression

As mentioned above, we also consider the effects of two types of regression. The first of these is within-line regressions, which is where the reader momentarily jumps back to a previous point in the current line. The extent to which the reader performs within-line regressions is formalized by a probability. If this probability is set to 1, the reader will perform a regressed fixation after every normal fixation, doubling the number of fixations on the line; if the parameter is set to 0, the reader will never perform a regression within the line. The x position of the regressed fixation is located randomly between the start of the line and the current fixation with longer regressions being linearly less probable than shorter regressions. The y value of the regressed fixation follows Equation 2.

Between-Line Regression

The second type of regression, between-line regressions, is where the reader rereads text from a previous line. Between-line regressions are expressed in terms of the probability that the reader will go back to a previous line at some point during reading of the current line. Once the regression is completed, the reader returns to the point in the passage before the regression occurred. If the parameter is set to 1, the reader will reread part of a previous line for every line they read; if it is set to 0, the reader will never perform a regression to a previous line. When a between-line regression occurs, the previous line is determined randomly, with more recent lines linearly more probable than less recent lines; the section of the previous line is determined randomly by two uniformly distributed *x* values. The *y* values of the regressed fixations follow Equation 2.

Results

For each phenomenon, we ran 100 simulations for each of 50 gradations in the parameter space, and each of these 5000 simulated reading scenarios was corrected by all ten algorithms. Accuracy is measured as the percentage of fixations that were correctly mapped back to the target line. Before describing the results, there are three important things to note. Firstly, the extreme values we have chosen for each phenomenon are arbitrary, so the algorithms should only be compared within, and not across, phenomena.[5] Secondly, we have not modeled the interactions between phenomena because it is inherently difficult to explore the effects of five dimensions on accuracy and it is not clear how the dimensions should be weighted a priori. Thirdly, for the algorithms that have free parameters (`chain`, `compare`, `merge`, `regress`, and `stretch`), we use the default parameter settings defined in the previous section. We have not systematically manipulated the parameter settings because (a) this would result in an explosion in the number of algorithm/parameter combinations that we

[5] For example, regress appears to have worse performance on shift compared to slope; however, if we had simulated a narrower range of shift values, the results might have led us to the opposite conclusion.

must consider, (b) manipulating a parameter to deal with one phenomenon can have unexpected consequences for other phenomena,[6] and (c), in a sense, these algorithms ought to incur a penalty for not being parameter-free.

Performance on Noise

Results for the noise distortion parameter are shown in Fig. 4a. Under zero noise, all algorithms perform at 100% accuracy, but six of the algorithms are adversely affected by noise when it reaches a sufficiently high level of around 10: Of these, `chain` performs best, closely followed by `attach`, then `cluster`, `regress`, and `stretch`, and finally `merge`. Of the remaining algorithms, `compare` and `split` are highly resilient to noise, while `segment` and `warp` are entirely invariant.

Performance on Slope

In terms of slope distortion (Fig. 4b), when the parameter is set to zero, all algorithms perform perfectly, but as the slope becomes more extreme (in either the upward or the downward direction), five of the algorithms experience a sustained loss in accuracy. Of these, `cluster` and `stretch` generally perform best and, initially at least, `attach` performs worst; `chain` and `split` initially perform better than `attach`, but are eventually outperformed. Interestingly, although `regress` is mostly resilient to slope, it has two weak spots around the values of -0.03 and 0.03. When the slope takes one of these values, `regress` struggles to disambiguate between (a) zero offset combined with the appropriate slope and (b) a large offset combined with slope in the opposite direction; if it selects the wrong option, fixations on one half of the passage will be misaligned, causing a substantial drop in accuracy.

[6] For example, if the standard deviation bounds of regress are widened, it may be possible to improve performance on noise, but the algorithm will be less capable of dealing with slope.

This reveals a hidden weakness of the `regress` algorithm, and we will see an example of it later. Of the remaining algorithms, `compare` and `merge` are highly resilient to slope, while `segment` and `warp` are invariant.

Performance on Shift

In terms of shift (Fig. 4c), when the parameter is zero, all algorithms perform perfectly, but as it becomes more extreme, five of the algorithms—`attach`, `chain`, `regress`, `split`, and `stretch`—drop in accuracy. In fact, `attach`, `chain`, and `split` produce identical results in the case of shift because they are all fundamentally reliant on absolute position. Somewhat surprisingly, `stretch` does not perform especially well on shift. This is because `stretch` can only handle up to one full line of shift; any more than this and the bounds have to be relaxed, but this results in an objective function with multiple maxima which is difficult to optimize. The `compare` algorithm is mostly resilient to shift, while the remaining four algorithms—`cluster`, `merge`, `segment`, and `warp`—are invariant.

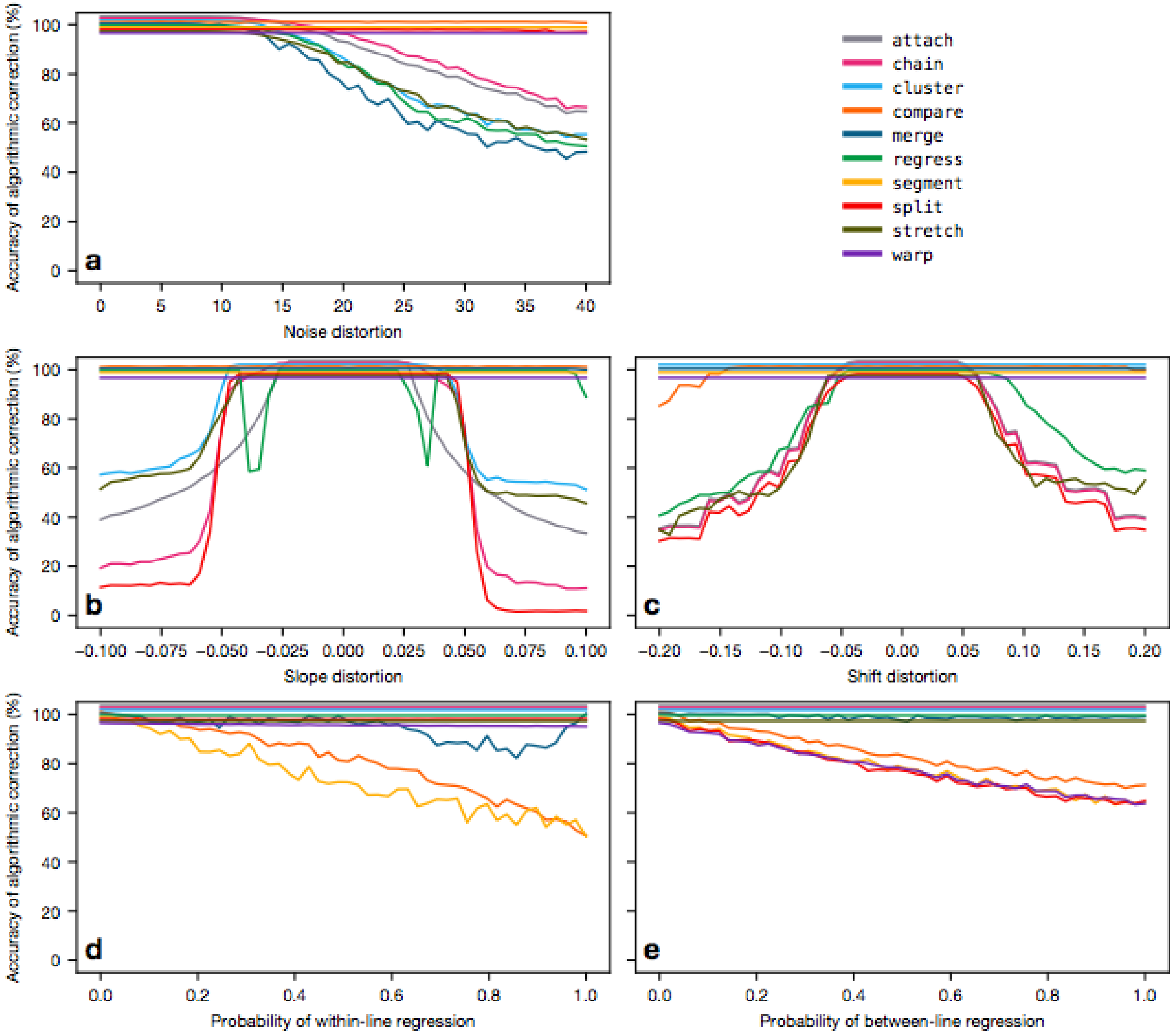


Figure 4 Mean accuracy of the ten algorithms in response to the five eye-tracking phenomena. For example, some algorithms (`attach`, `chain`, `cluster`, `merge`, `regress`, and `stretch`) are adversely affected as the noise level is increased, while the other algorithms are either resilient to noise (`compare` and `split`) or entirely invariant to noise (`segment` and `warp`). The plotted lines have been vertically staggered to aid visualization.

Performance on Within-Line Regression

Results for within-line regressions are shown in Fig. 4d. When there are no within-line regressions, all algorithms perform at 100%, but three of the algorithms drop off as the probability of within-line regression is increased. Of these, `compare` and `segment` track each other quite closely because they rely on identifying the return sweeps; `merge` is generally quite resilient, except when the parameter is around 0.7–0.9 because these values cause a large number of progressive sequences to be generated which cannot then be merged very freely, so the merge process tends to get trapped in local minima (i.e., bad mergers that happen early on cannot later be reverted). Of the remaining algorithms, `split`[7] and `warp` are highly resilient, while `attach`, `chain`, `cluster`, `regress`, and `stretch` are invariant.

Performance on Between-Line Regression

In terms of between-line regressions (Fig. 4e), four algorithms are negatively impacted by increases in this parameter. Of these, compare and `split` can in principle find more than *m* gaze lines, but they have difficulties identifying when a between-line regression occurs, while `segment` and `warp` are limited to identifying exactly *m* gaze lines in strictly sequential order, so they fundamentally cannot handle between-line regressions. Of the remaining algorithms, `merge` is resilient to between-line regressions, while `attach`, `chain`, `cluster`, `regress`, and `stretch` are entirely invariant.

Summary

In this section, we have simulated five eye-tracking phenomena that are particularly

[7] Unlike compare and segment, even if split misidentifies a regression as a return sweep, it will still be able to map the resulting gaze line to the appropriate text line because it assigns based on position rather than order.

relevant to understanding the performance characteristics of the algorithms. Fig. 5 summarizes how accurately the algorithms perform on each phenomenon. No single algorithm is invariant—or even resilient—to all phenomena, although `merge` and `warp` come quite close: `merge` is only weak on noise, while `warp` is only weak on between-line regressions. In general, there tends to be a tradeoff between how well an algorithm can handle distortion and how well it can handle regressions; the ability to deal with one tends to come at the cost of the other. Nevertheless, in real world scenarios, performance will very much depend on the degree and relative prevalence of the phenomena. Furthermore, there are likely to be other important forms of measurement error and reading behavior that we have neglected to consider here, and those that we have considered are likely to interact in complex, unpredictable ways. It is therefore important to test the algorithms against natural eye-tracking data to get a more holistic understanding of their performance.

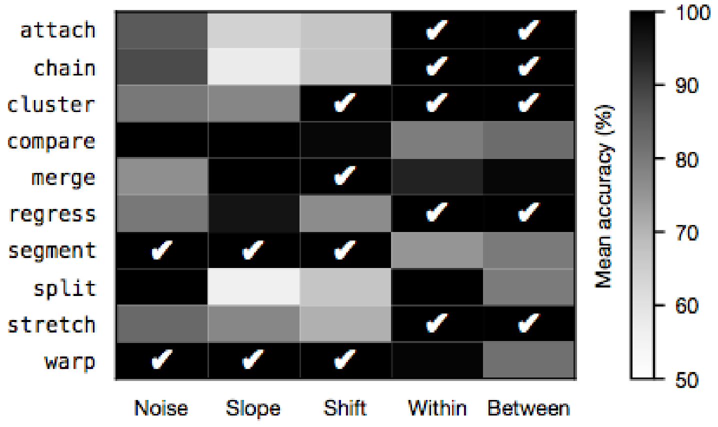


Figure 5 Mean accuracy of the algorithms for each of the eye-tracking phenomena. Darker cells indicate phenomena that an algorithm performs well on. A checkmark indicates that the algorithm is entirely invariant to the phenomenon in question, scoring 100% in all 5000 simulations.

Performance on Natural Data

In this section, we test the algorithms against an eye-tracking dataset that has been manually corrected by human experts. Unlike the simulations, there is no ground truth, and we cannot isolate particular phenomena; however, the benefit of this approach is that the phenomena are combined in a realistic way, allowing us to estimate how well the algorithms are likely to perform in real-world scenarios.

Method

We tested the algorithms on an eye-tracking dataset collected by Pescuma et al. (in prep.), which includes reading data for both adults and children, allowing us to test the algorithms on two distinct populations. Our general approach is illustrated in Fig. 6 and discussed over the following sections.

The Dataset

Pescuma et al. (in prep.) collected eye-tracking data for 12 passages from Italian children's stories (e.g., a passage from *Goldilocks*). The passages were around 130 words in length, spanning 10–13 lines and were presented in 20-point Courier New (each character occupying around 0.45 degrees of visual angle). Either of two sets, each comprised of six passages, was administered for silent reading to a large sample of children aged 8–12 (N = 141) and a smaller sample of adult controls (N = 33) for a total of 877 reading trials. Eye movements were recorded using a tower mounted EyeLink 1000 Plus eye tracker (SR Research, Toronto, Canada) for which a typical accuracy of 0.25–0.50 degrees is reported by the manufacturer. Recording was monocular (right eye) with a 1000 Hz sampling rate.

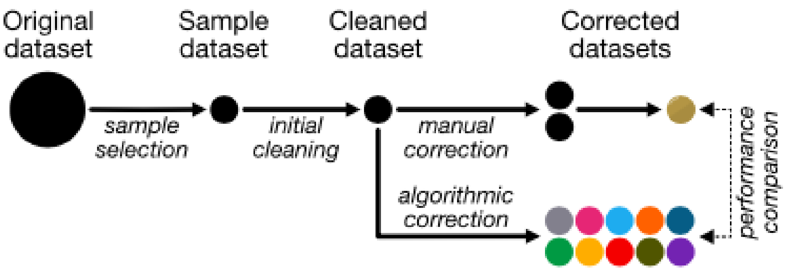


Figure 6 Pipeline for testing the algorithms on a natural dataset. The original dataset was first reduced to a smaller sample, which then underwent some initial cleaning steps. This cleaned dataset was then corrected by the ten algorithms and two human correctors, whose corrections were merged to form the gold standard. Performance is measured by how closely the algorithmic corrections match the gold standard correction.

Selection of the Sample for Manual Correction

Since it was impractical to manually correct all 877 trials (to do so would require months of work), we selected a sample for manual correction. For each of the 12 passages, we selected two reading trials by adult participants and two reading trials by child participants, for a total of 48 trials (5.5% of the full dataset). The reading trials were selected pseudorandomly such that no single participant was represented more than once. Additionally, we manually checked and adjusted the sample to ensure it contained an equal balance of easy and challenging cases, as well as examples of all the various eye-tracking phenomena discussed previously.

Initial Data Cleaning

We performed two initial cleaning steps in order to isolate the core problem of line assignment from two extraneous issues. Firstly, any fixation that was located more than 100 px from any character in the passage was removed (i.e., out-of-bounds fixations that occur in the margins or off-screen). This is because the algorithms are not designed to detect and discard these fixations, and such cases can hinder their ability to match fixations to the appropriate lines. Secondly, prior to reading a passage and on its completion, a reader's fixations will typically jump around the text unpredictably; again, since the algorithms are not designed to automatically discard such fixations, we manually removed any such cases from the starts and ends of the fixation sequences, allowing the algorithms to concentrate on the core problem of assigning fixations to lines.

Manual Correction Procedure

The cleaned sample dataset was corrected independently by two human correctors (JWC and VNP). To perform the correction, each corrector studied plots of the participants' fixation sequences and recorded, fixation by fixation, which line each one belonged to, guided by fixation position, saccade trajectories, textual cues, and fixation duration, as well as general knowledge of eye tracking and reading behavior. Unlike the algorithms, the human correctors also had the option to discard fixations as they saw fit. This is because there were cases where it was clear a fixation should be discarded—for instance, due to spatial misplacement or ultra-short duration—and it would have therefore felt disingenuous to assign these cases to a line anyway.

Across the 48 reading trials, the correctors initially disagreed on 299 of 10,245 fixations (2.9%). Of these 299 disagreements, only 15 related to which line a fixation was assigned to; on inspection, all 15 cases turned out to be human error on the part of one corrector or the other. The other 284 disagreements related to whether or not a fixation should be discarded; following discussion of these cases, the correctors reached consensus about how these fixations should be treated. This resulted in a single manual correction, which we consider to be the gold standard against which the algorithms can be evaluated. In this gold standard correction, a total of 255 fixations were discarded across all 48 trials (2.5%; 5.3 fixations per trial).

It is interesting to note that, although the two correctors had slightly different intuitions about when it was appropriate to discard a fixation, they essentially had perfect agreement about which line a fixation ought to be assigned to if it was retained. This suggests that the correction of vertical drift is actually quite objective—there is usually an unequivocally correct solution to any given trial, even if that solution may be difficult and time-consuming to obtain.

Results

We analyze the performance characteristics of the algorithms in four ways. Firstly, we look at how the algorithms fare against the gold standard manual correction; secondly, we look at what proportion of trials are likely to be usable following drift correction; thirdly, we look at how the algorithms perform in comparison to using no drift correction at all; and finally, we look at how the algorithms relate to each other, regardless of their accuracy.

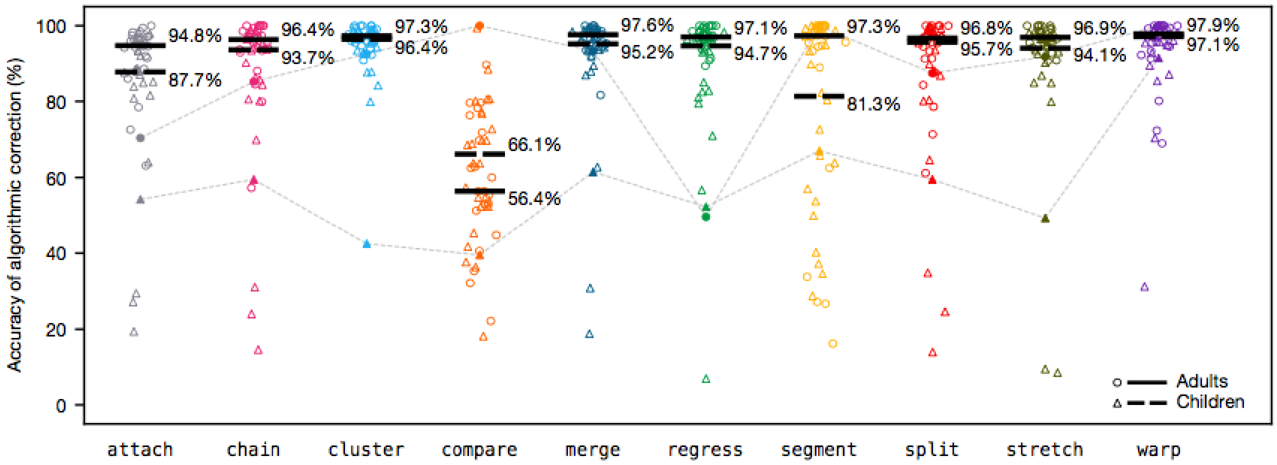


Figure 7 Accuracy of the algorithms on adult reading trials (circles) and child reading trials (triangles). The *y*-axis measures the percentage of fixations assigned to the correct line, as defined by the gold standard manual correction. The filled points, linked together by dashed lines, correspond to the two example trials illustrated in Figs. 8 and 9. The black bars show median accuracy for the adults (solid bars) and children (broken bars).

Accuracy against the Gold Standard

As with the simulations, accuracy is measured as the percentage of fixations that the algorithm mapped to the correct line; the ground truth is defined by the gold standard manual correction. In cases where the correctors chose to discard a fixation, the algorithm is automatically wrong, which amounts to a constant baseline level of error that all algorithms suffer from equally. Fig. 7 plots accuracy on the 48 sample trials by algorithm. The most striking result is `compare` with overall median accuracy of 61.2%, substantially worse than all other algorithms.[8] This contrasts with our simulations, which indicated that `compare` should at least be relatively strong on distortion. The reason for this discrepancy is that the simulated fixation sequences were generated directly from the lines of text with one fixation per word, so the artificial gaze lines that `compare` identified tended to be very horizontally similar to the artificial text lines. In the natural dataset, however, this is not the case; when the data contains a lot of natural noise and regressions, gaze lines cannot be reliably matched to text lines based on similarity, even if the set of candidate text lines is narrowed down to the three closest neighbors. Given that `compare` exhibited such poor performance, we consider it to be an algorithmic deadend and do not discuss it any further.

[8] This concurs with the 60% accuracy reported by Lima Sanches et al. (2015, p. 1231). Yamaya et al. (2017, p. 104), who describe a slightly more complex method with better sweep detection, report accuracy of 87%, which is still quite low compared to the other results obtained in this paper.

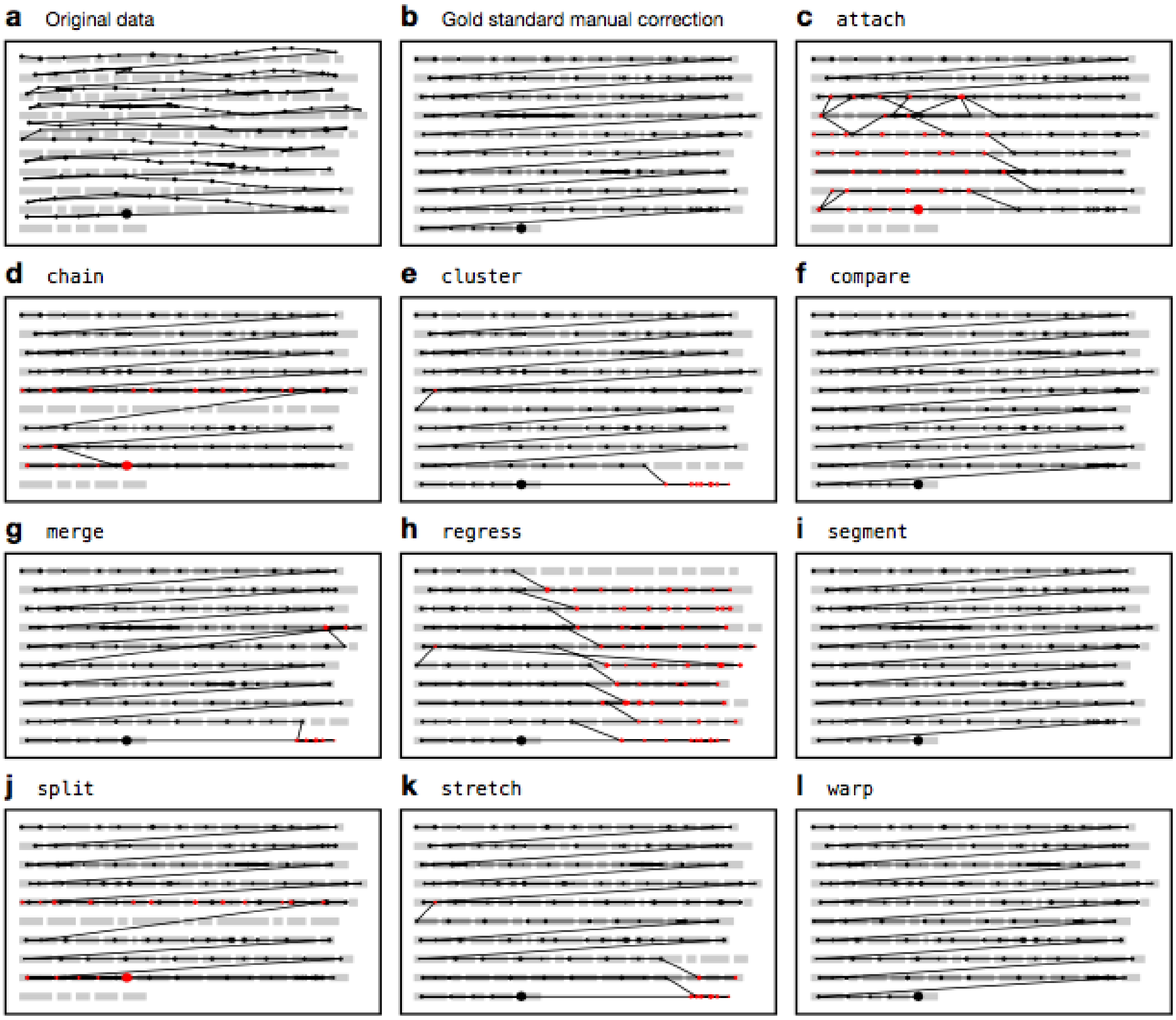


Figure 8 Original data and corrections of an adult trial. Fixations in red have been assigned to the wrong line. The algorithmic corrections correspond to the filled circles in Figs. 7 and 11.

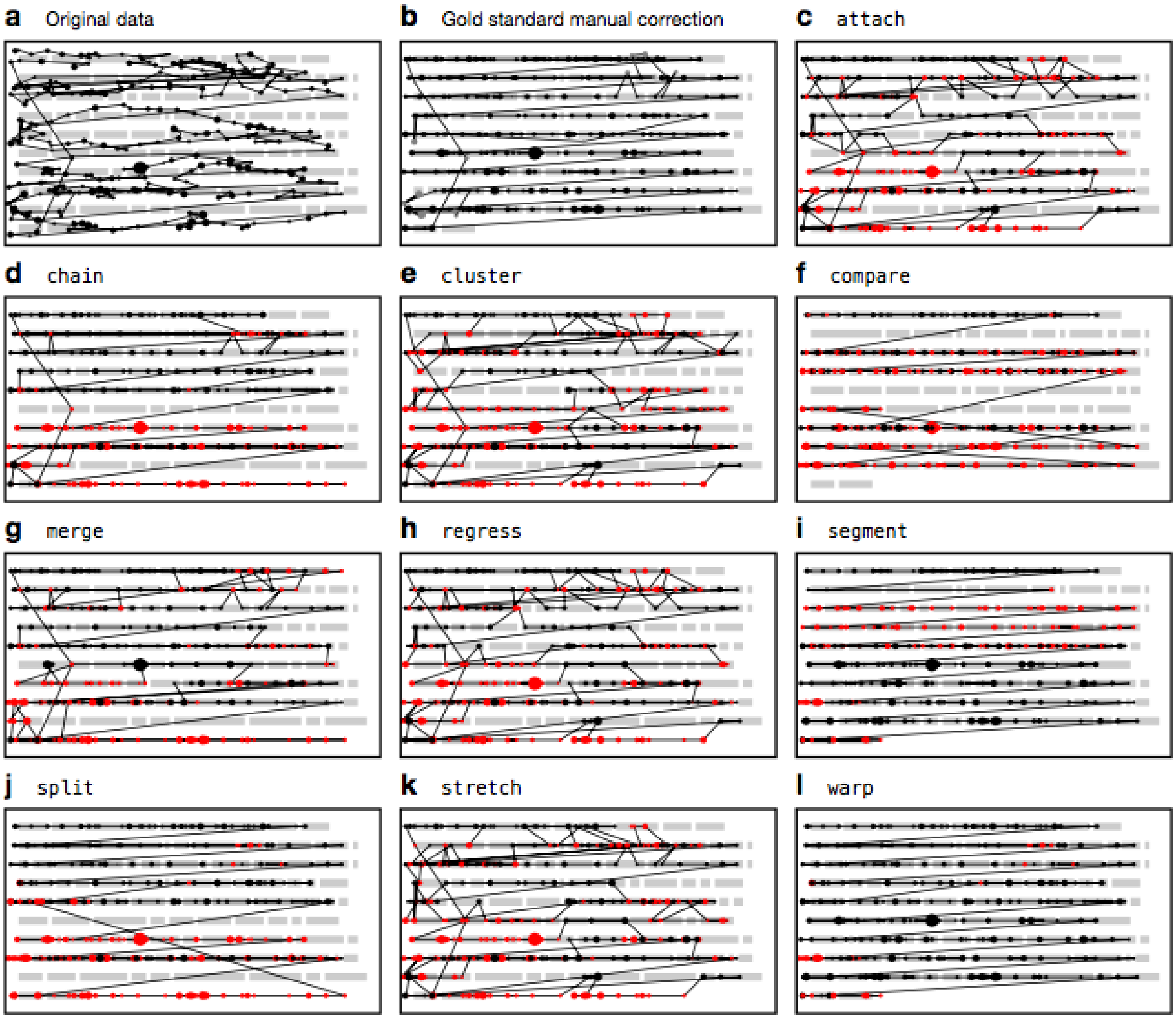


Figure 9 Original data and corrections of a child trial. Fixations in red have been assigned to the wrong line. Fixations that were discarded in the gold standard manual correction are shown in gray. The algorithmic corrections correspond to the filled triangles in Figs. 7 and 11.

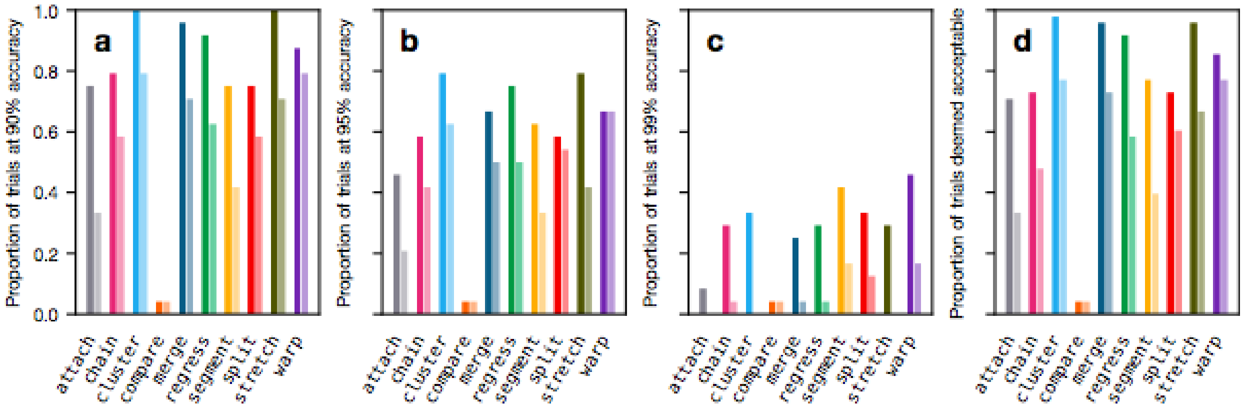


Figure 10 Proportion of trials that surpassed (**a**) 90%, (**b**) 95%, and (**c**) 99% accuracy, and (**d**) the proportion of corrections deemed acceptable by two human raters. The dark and light bars represent the adult and child datasets respectively.

Of the remaining algorithms, median accuracy is typically around 95%, the worst performer being `attach` at 92% and the best performer being `warp` at 97.3%. Accuracy on child trials tends to be lower and more variable than accuracy on adult trials; however, the difference in medians was usually quite small. The major exception to this was `segment` for which median accuracy on adult trials was 97.3%, while median accuracy on child trials was 81.3%, making `segment` one of the best algorithms in terms of adult data but one of the worst in terms of child data. This may be because children tend to perform more regressions (e.g., Blythe & Joseph, 2011) and have more disfluent return sweeps (e.g., Parker et al., 2019), both of which create obstacles for the `segment` algorithm.

Median performance alone conceals the fact that accuracy is often highly variable and long tailed. In the best case scenario, an algorithm will produce a perfect correction that is identical to the gold standard—all algorithms (even `compare`) scored 100% in at least one trial. In the worst case scenario, an algorithm will perform as low as 10–30% accuracy. In addition, the algorithms often differ markedly on particular trials. We have highlighted this in

Fig. 7 by singling out two trials, one by an adult and one by a child, which are represented by the filled data points that are linked together with dashed lines. Algorithmic corrections of the adult trial (filled circles) are depicted in Fig. 8. In this particular case, compare, segment, and warp were able to correctly recover the intended line of every fixation. However, the trial presented problems for some of the other algorithms; in particular, attach failed to handle the upward shift in the lower left quadrant of the passage, and regress misinterpreted the situation as a case of upward slope, resulting in fixations on the right hand side of the passage being forced down by one line, a potential weakness highlighted by our simulations.

Fig. 9 depicts the algorithmic corrections of the child reading trial, which are represented by the filled triangles in Fig. 7. Performance on this trial is much worse due to the large amount of noise. cluster, for example, has struggled to correctly classify the fixations due to the large amount of overlap between fixations intended for adjacent lines, and segment has identified one particularly long within-line regression as a return sweep, resulting in some misalignment in the middle of the passage. Only warp was able to recover the intended lines for the majority of fixations, and the few errors it did make appear to be cases where the correctors chose to discard some fixations. Overall, these two example trials highlight that, although the algorithms have a similar level of performance on average, performance on a particular trial can be quite divergent depending on its particular characteristics. Illustrated corrections of all 48 trials by each algorithm can be found in Supplementary Item 2.

Proportion of Corrections Likely to be Usable

Fig. 10 reports the proportion of corrections that surpassed an accuracy level of 90%, 95%, and 99% by algorithm. If you are willing to accept relatively low accuracy at the trial level (e.g., 90%, Fig. 10a), then cluster, merge, and stretch will provide the best performance—a large proportion of corrections will meet this criterion. In comparison, if you have more stringent accuracy requirements at the trial level (e.g., 99%, Fig. 10c), then `segment`, `split`, and `warp` are likely to provide better performance. Of course, the cost of a more stringent accuracy criterion is that fewer corrections will be usable overall, and it is not possible to know which trials have low accuracy in the absence of manual correction data. This highlights the fact that it is currently not possible to confidently achieve a high level of accuracy in a high proportion of trials, so researchers may still need to invest a significant amount of time if a high level of accuracy is demanded.

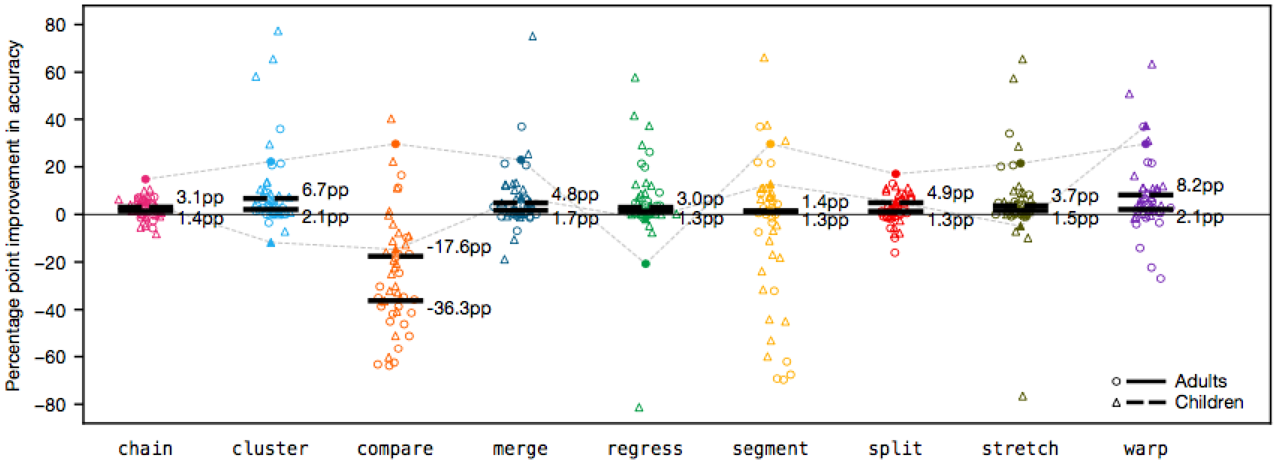


Figure 11 Improvement in accuracy in comparison to performing a standard eye-tracking analysis with no drift correction. The *y*-axis measures the percentage point increase (or decrease) in accuracy beyond the baseline accuracy of the `attach` algorithm. The filled points, linked together by dashed lines, correspond to the two example trials illustrated in Figs. 8 and 9. The black bars show median improvement for the adults (solid bars) and children (broken bars).

To estimate how usable the corrections are likely to be to a typical researcher, we performed a more subjective analysis of their quality. All 480 algorithmic corrections were presented blind and in random order to two raters (JWC and VNP), who independently classified every correction as either "acceptable" or "needs more work." The raters did not discuss in advance what criteria they would use to make these judgments, but agreement was nevertheless very high at 94%. In addition to the overall number of errors, the raters weighed up other factors, such as how the errors were distributed over the passage, how challenging the input data seemed to be, and what effect the errors might have for downstream analyses. The results, which are shown in Fig. 10d, suggest that `cluster`, `merge`, and `stretch` are likely to produce very satisfactory results on adult data.

Improvement over no Drift Correction

Another way to measure performance is in terms of how much of an improvement an algorithm provides in comparison to applying no drift collection at all. To estimate this, we first need to define a baseline level of accuracy. As mentioned previously, the `attach` algorithm essentially corresponds to a standard eye-tracking analysis; it is equivalent to drawing maximal, nonoverlapping bounding boxes around the words in a passage and then mapping fixations to whichever bounding box they fall into (as would be the case in a standard analysis of eye-tracking data using the widely adopted Area-of-Interest paradigm). Therefore, we can estimate the potential improvement that a given algorithm offers by comparing its accuracy to the accuracy of the `attach` algorithm.

The results of this analysis are plotted in Fig. 11. The *y*-axis shows the percentage point increase (or decrease) in accuracy that results from applying vertical drift correction. The zero line represents the baseline of no drift correction (equivalent to `attach`). As before, the datapoints themselves tell us a lot more than the medians. The `chain` and `split` algorithms tend to be quite conservative, while the others tend to have more extreme effects. In the best case, `cluster` resulted in a 77 percentage point increase in accuracy in comparison to leaving the data uncorrected (i.e., `attach = 19%, cluster = 96%`); while in the worst case, `regress` resulted in an 81 percentage point drop in accuracy, badly corrupting the original input data (i.e., `attach` = 88%, `regress` = 7%).

These results highlight that, although in most cases the application of vertical drift correction can improve data quality, the process is not without risk. Furthermore, there is potentially more to gain from applying drift correction to child data, since the baseline level of accuracy tends to be lower to begin with; for example, `warp` offered a modest 2.1 percentage point increase in accuracy on adult data but an 8.2 percentage point increase on child data.

Relationships between Algorithms

As noted above, some algorithms tend to produce very similar output where others produce quite different output. This raises the issue of how the algorithms relate to each other regardless of their performance characteristics on real or simulated data. To investigate this, for each pair of algorithms, we measured the DTW distance between the corresponding algorithmic corrections of each of the 48 sample trials and took the median distance as an estimate of how dissimilar those two algorithms are.[9] We then analyzed the pairwise distances in two ways.

Firstly, we used agglomerative hierarchical clustering to produce a dendrogram (see Fig. 12a), which yields an approximate taxonomy of the algorithms based on their similarity. The root node represents all algorithms, which initially fork into two major groups. The "sequential algorithms," `segment` and `warp`, both operate on the principle of identifying the return sweeps and mapping the resulting subsequences to the lines of text in sequential order; in other words, their analysis stages can only produce groups consisting of fixations that were arranged consecutively in the original fixation sequence. This means they tend to produce similar outcomes—they both, for example, force fixations onto inappropriate lines in order to preserve sequentiality. Of the "positional algorithms," `split` is the first to branch off, perhaps

[9] Since compare was highly divergent from all other algorithms due to its poor performance, it is not included in this analysis.

because—like the sequential algorithms—it leans heavily on the return sweeps, and among the remaining algorithms, there is a clear dichotomy between those that assign based on relative position (`cluster`, `merge`, `regress`, and `stretch`) and those that assign based on absolute position (`attach` and `chain`).

Secondly, we used multidimensional scaling to locate the algorithms in some latent "algorithm space." Fig. 12b shows the output of this analysis projected into two hypothetical dimensions: Algorithms that are close together in this space tend to produce similar results, while algorithms that are far apart tend to produce dissimilar results. The two dimensions of the space appear to roughly correspond to the algorithms' analysis strategy (*x*-axis) and assignment strategy (*y*-axis). `split`, for example, shares its analysis strategy with `segment` (it groups fixations based on return sweeps), but it has an assignment strategy that is more similar to `chain` (it assigns based on absolute position). We also see that `regress` and `stretch` tend to produce very similar output and are therefore likely to be somewhat interchangeable. Interestingly, the human correctors—represented by the gold standard manual correction—are located in a relatively unexplored part of algorithm space: Their analysis strategy appears to be more similar to `chain` or `merge` (finding local linear clusters), while their assignment strategy seems to be more global and sequential, like `warp`. Anecdotally, this aligns with our experience of performing the manual corrections, and this observation is suggestive of fertile ground for the future development of correction algorithms.

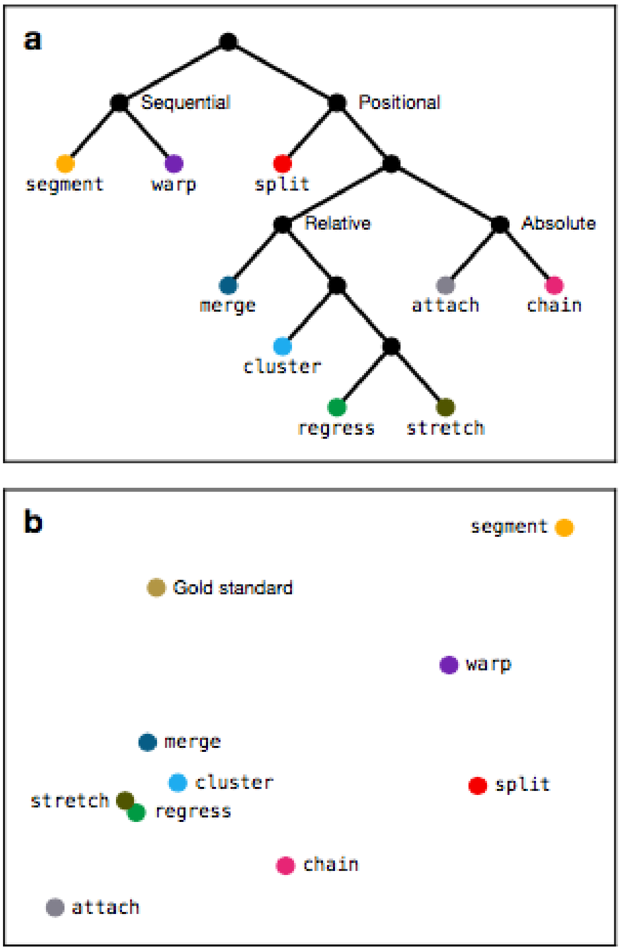


Figure 12 (**a**) Hierarchical clustering analysis of the algorithmic outputs, providing an approximate taxonomy of the algorithms. (**b**) Multidimensional scaling analysis of the algorithmic outputs; the distance between two algorithms corresponds to how dissimilar their corrections tend to be, so the space as a whole approximates how the algorithms relate to each other on two hypothetical dimensions.

Summary

In this section, we tested the ten algorithms on a real eye-tracking dataset. Although `warp` was marginally the most performant algorithm across the majority of measures, our results indicate that the best algorithm will largely depend on the particular characteristics of a given trial, as well as the general characteristics of the dataset being corrected. All 48 reading trials could be improved by at least one of the algorithms; the difficulty for the researcher, of course, is in knowing which algorithm to apply to a given trial in the absence of a gold standard.

Discussion

We have identified ten core approaches to the methodological problem of correcting vertical drift in eye-tracking data. We instantiated each of these approaches as a simple algorithm that can be evaluated in a consistent and transparent way. Our first analysis using simulated data allowed us to identify which phenomena the algorithms are invariant to and to quantify how the algorithms respond to increasing levels of those phenomena. Our second analysis validated the algorithms on a real eye-tracking dataset and allowed us to strengthen our qualitative intuitions about their similarities and differences. In the remainder of the paper, we sum up what we learned about the algorithms, provide some practical guidance for researchers in the field, and conclude with some thoughts about how vertical drift correction can be improved going forward.

Major Properties of the Algorithms

The algorithms can be placed into three major categories. The **sequential algorithms**, `segment` and `warp`, hinge on their ability to correctly identify the return sweeps. If successful, these algorithms have excellent performance because any vertical drift in the data essentially becomes invisible. However, if the text is read nonlinearly, the premise on which the sequential algorithms are based breaks down. Therefore, one should apply these algorithms with great caution when the data are rich in regressions, either within-line (which `warp` tends to handle well) or between-line. The risk is particularly high with `segment`, which has good median performance on adult data, but can also lead to catastrophic errors if large regressions are mistakenly interpreted as return sweeps.

The **relative-positional algorithms**, `cluster`, `merge`, `regress`, and `stretch`, are mostly dependent on their ability to correctly classify the fixations into *m* groups, each using a slightly different technique to do so. So long as the identified groups are sound, then the use of relative position to assign the fixations to lines is generally very resistant to vertical drift.

The **absolute-positional algorithms**, `attach`, `chain`, and `split`, generally tend to be the worst at dealing with vertical drift because they assign based on absolute position; this feature makes them generally weaker than other algorithms at dealing with noise, slope, and shift. However, a benefit of these three algorithms is that they tend to be quite conservative and do not make dramatic changes to the data, which makes them a reasonable choice for researchers who would prefer a more minimalist data transformation or whose data are not overly affected by distortion issues.

General Guidance for Researchers

The analyses presented in this paper clearly indicate that each algorithm performs best on a different set of factors, some of which we have not considered here in detail. For example, if the line spacing is quite tight, the eye-tracking data is more likely to be negatively impacted by distortion, making a sequential algorithm a better choice; conversely, if lines are spread far apart, a relative-positional algorithm may be more appropriate. Overall, different sets of data will require different correction algorithms, and a qualitative inspection of the data will be required to detect the relative severity of general noise, drift issues, and regression phenomena. To help in this process, one option might be to hand-correct a sample of the trials in order to assess which algorithm performs best on those specific cases and then apply this algorithm to the entire dataset. However, there might very well be too much trial-by-trial (or participant-by-participant) variability to use a single algorithm across an entire dataset. In this case, it may be preferable to create subsets of data exhibiting comparable patterns of eye-tracking phenomena and deal with those subsets with different algorithms. Another idea would be to run several algorithms over the dataset and manually inspect cases where there is disagreement.

Recording children's eye movements poses extra challenges relative to adults', particularly due to the difficulty that younger participants often experience sitting still for relatively long periods of time (Blythe & Joseph, 2011), which can lead to a loss of calibration. Therefore, especially in the case of multiline reading, developing readers' eye movements are generally characterized by more noise, as well as by greater slope and shift, than adults'. This would suggest resorting to algorithms like `segment` and `warp`, which are entirely invariant to noise, slope, and shift (Fig. 5). However, children tend to generally make more regressions than adults (e.g., Blythe & Joseph, 2011; Reichle et al., 2003), which is exactly the phenomenon that affects `segment` and `warp` the most. The general tradeoff between the ability to handle distortion and the ability to handle regressions is at play here, and only an attentive, qualitative check of the data will tell the researcher which way to go. If there does not appear to be too much of an issue with between-line regressions, then `warp` is probably the best choice; otherwise, `cluster` or `merge` might be a better option.

Regarding the practicalities of the algorithm application pipeline, we suggest performing a few cleaning steps before drift correction, in order to isolate the line assignment problem from other issues that the algorithms considered here were not designed to deal with, and which would otherwise impair their performance. For example, researchers may first want to discard any fixations that lie beyond the text area and merge or eliminate extremely short fixations. Only after these basic cleaning steps have been performed, can algorithmic correction be safely applied.

Another important aspect to consider is the presence of free parameters. The `chain`, `merge`, `regress`, and `stretch` algorithms take additional input parameters that must be set appropriately by the user. In practice, Špakov et al.'s (2019) and Cohen's (2013) suggested defaults for the `merge` and `regress` algorithms seemed to work well on our test dataset, but in the case of `chain`, it was somewhat unclear how to set the *x* and *y* thresholds appropriately, so experimentation might be required to produce the best results. We also found that `stretch` was very sensitive to its parameter settings and that the upper and lower bounds must be tightly constrained around likely values for it to produce sensible results. An advantage of all other algorithms is that they are parameter free, making drift correction easier to perform, document, and justify.

It is also worth considering the complexity of the algorithms. Some, such as `chain` and `segment`, are very simple and intuitive, while others, such as `merge` and `warp`, are quite complex. Although complexity is not an important consideration from a performance perspective (in general, we should prefer whichever algorithm works best), it is worth considering how complexity might impinge on real-world use. For example, users may be less inclined to use an algorithm if they cannot intuitively understand how it will manipulate their data, so algorithms should, where possible, be designed in a way that researchers find easy to understand and easy to convey to their readership. In that regard, we hope that this paper will give researchers more confidence in the algorithms, which we have validated and benchmarked.

Finally, it is worth noting that most of the algorithms have linear time complexity and can process a reading trial in fractions of a second, so runtime does not warrant any special consideration. The one exception to this is `merge` which scales quadratically with the number of fixations; in our testing, for example, it took 100 ms for a trial consisting of around 100 fixations but up to 31 s for a trial of around 500 fixations.

Improvements on the Algorithms

We would not wish to claim that the algorithms, as presented here, are the only approaches one may take nor that they are the ultimate form of each core method; all can be improved in one way or another. Furthermore, there are likely to be ways of combining the outputs of multiple algorithms to increase confidence in particular solutions. The goal of this paper, however, was to evaluate the algorithms in their more abstract, idealized forms in order to make general recommendations and to provide a solid foundation for the future development of vertical drift correction software. Nevertheless, here we briefly note some of the most obvious ways in which the algorithms could be improved.

Chain

The main weakness of the `chain` algorithm is its reliance on threshold parameters that must be set by the user, but this situation could be improved if the parameters defaulted to sensible values based on reliable heuristics. For example, it may be the case that the parameters can be reliably estimated from the line and character spacing (as appears to be the case in Schroeder's (2019) implementation in popEye) or other known properties of the passage, language, or reader. Secondly, our simulations showed that `chain` does not respond well to slope distortion, performing worse than `attach` under extreme values. This can be alleviated by a *y* threshold that grows as the reader progresses over the line, as is the case in Hyrskykari's (2006) sticky lines algorithm.

Cluster

The biggest weakness of `cluster` was its ability to deal with general noise (or, equivalently, tight line spacing). One potential way to improve this would be to utilize the *x* values of the fixations. Unfortunately, it is not simply a case of performing a two-dimensional *k*-means clustering on the *xy* values because this leads to situations where clusters are identified that span multiple lines because they have similar *x* values. However, it might still be possible to utilize the *x*-axis information, perhaps by weighting the two axes differently in some way. `Cluster` analysis is a very broad topic in data science, and there are likely to be many other candidate algorithms, beyond simple *k*-means clustering, that will be worth investigating.

Merge

The core principle of `merge` is to start with small groups of fixations and gradually build them up into gaze lines, guided by their fit to regression lines. The most extreme version of this algorithm would start with every fixation in an individual group, and the algorithm would consider every sequence in which mergers could be performed (i.e., the entire binary search tree). This would allow the algorithm to explore cases where it is first necessary to make a bad merger in order to make a great merger later on (i.e., it would avoid becoming stuck in local maxima). Such an algorithm would be intractable, however, due to a combinatorial explosion in the number of possible merge sequences. To avoid this, `merge` uses an initial `chain`-like strategy to seed the merge process with a reduced set of groups, and it then explores just one possible path through the search tree, selecting only the most promising merger at each step. One way to improve the algorithm, then, would be to use more advanced tree traversal techniques, such as beam search in which several of the most promising mergers are fully explored on each iteration. This would come at the cost of making an already slow algorithm even slower, but it would probably result in better solutions and might also allow for the removal of the thresholds and heuristics.

Regress

The main weakness of the `regress` algorithm is that the *m* regression lines it fits to the data cannot take independent slope or offset values, limiting its ability to handle complex cases, especially those involving shift. Thus, one obvious way to advance the algorithm would be to allow for such independent values. However, even the simplest case of having a single slope parameter, a single standard deviation parameter, and one offset parameter per line of text would result in an objective function with m + 2 parameters, which may become difficult or impossible to minimize, especially as the number of lines increases. Another avenue for improving `regress` would be to try some form of nonlinear regression. In Fig. 9a, for example, we see a case where a gaze line forms a nonlinear arc, which a linear regression line cannot fully capture (cf. Fig. 9h).

Segment

The performance of the `segment` algorithm hinges on its ability to identify the true return sweeps; when it works, it tends to work very well, but when it fails, it does so catastrophically. One way to improve the `segment` algorithm would therefore be to encode additional heuristics about how to distinguish true return sweeps from normal regressions. For example, a return sweep is not just an extreme movement to the left but also a movement downward by a relatively predictable amount (one line space), ultimately landing near the left edge of the passage. Introducing such heuristics would not be without caveats, however; in the case of downward slope, for example, return sweeps can appear quite flat (see, e.g., the final sweep in Fig. 8a) and would therefore go unnoticed under this change.

Split

Like `segment`, `split` could also benefit from better sweep detection, as well as better detection of between-line regressions. There are likely to be many ways of approaching this classification problem, but one simple option would be to use both dimensions in the saccade clustering—the return sweeps would then be the cluster of saccades that have large negative change on the *x*-axis, as well as a small positive change on the *y*-axis. More generally, it might be possible to combine the `split` and `segment` algorithms, since they are quite closely related computationally. For example, the set of saccades that most resemble return sweeps could first be identified, and then the $m - 1$ most extreme of these could be treated as major segmentation points, allowing for a sequential assignment, while the remainder could be treated as minor segmentation points, allowing for the identification of between-line regressions.

Stretch

Our analyses showed that `stretch` behaves very similarly to `regress`. This is because they are essentially two variants on the same basic idea: Detect the magnitude of the underlying slope (`regress`) or shift (`stretch`) and then reverse it. However, this does not work so well if the underlying calibration error is fluctuating in time or space. One way to improve the method, then, would be to search a more complex transformation space by including rotations and shears, for example, or by applying separate transformations to each quadrant of the text. Additionally, as Equation 1 makes clear, `stretch` essentially has the `attach` algorithm embedded within it, but in principle it should be possible to substitute this with any of the positional algorithms. For example, a `stretch-chain` algorithm would find a transformation of the fixations that results in minimal change when you apply the `chain` algorithm.

Warp

The primary weakness of `warp` is that the expected fixation sequence cannot encode unpredictable reading behavior that might be present in the veridical sequence, and there is no feasible way such unpredictability could be encoded. Instead, improving the `warp` algorithm is likely to involve relaxing DTW's requirement that matches between sequences increase monotonically, allowing the algorithm to find a mix of global and local sequence alignments. In this respect, the so-called "glocal" alignment algorithms could prove useful (Brudno et al., 2003), as well as many other sequence alignment algorithms that ought to be systematically investigated for the present purposes (e.g., Keogh & Pazzani, 2001; Tomasi et al., 2004; Tormene et al., 2009; Uchida, 2005). One simpler option—which could also be applied to `segment`—would be to use `attach` as a fallback method in cases where a fixation's revised *y*-axis coordinate is substantially different from its original *y*-axis coordinate. This would deal with cases where the strict sequentiality requirement forces fixations on to lines that are very far from their original positions.

Improvements on the Benchmarking

Aside from improving the algorithms themselves, it would also be useful to produce a much larger, heterogeneous benchmarking dataset, with data contributed from many different laboratories. This would offer more generalizable results and would help us understand how the algorithms respond to specific factors. For example, one useful feature of the dataset we used in this paper is that it includes data from both adults and children on the same passages of text, allowing us to compare how the algorithms respond to these two distinct populations. However, there are many other factors that will ultimately determine how the algorithms behave, such as the layout of the text, the complexity of the reading material, and the peculiarities of the eye tracker hardware. In addition, there are likely to be important linguistic factors at play: For example, our current set of results may not generalize well to logographic writing systems, such as Chinese, right-to-left scripts, such as Hebrew, orthographically opaque languages, such as English, or agglutinating languages, such as Turkish, where fixation patterns might differ in ways that the algorithms are sensitive to. However, the main difficulty we foresee in creating such a heterogeneous dataset—aside from producing the required manual corrections—would be ensuring it is representative of the kinds of experiments that researchers most typically run, while also diverse enough to capture all relevant factors.

Conclusion

Our intentions with this paper were twofold. Firstly, we wanted to systematically evaluate the various vertical drift correction algorithms that have been reported in the literature in order to provide guidance to researchers about how they work, when they should be used, and what their limitations are. In this respect, our most important observation was that there is no one killer app; different datasets—and even different trials within a dataset—will require different solutions, so researchers should select their correction method carefully. We hope that the guidance we have provided herein will be helpful in this regard.

Secondly, we wanted to lay a solid foundation for future work on post-hoc vertical drift correction by delimiting the core algorithms, providing constraints that future work can operate inside, and offering new perspectives on how drift correction techniques can be improved going forward. In this respect, we have provided basic implementations of the ten algorithms in multiple languages, which can be used as a starting point for building new versions or, indeed, as a comparison group against which entirely new algorithms can be compared. Several of the algorithms are already implemented in the Python package *Eyekit* (https://jwcarr.github.io/eyekit/) and the R package *popEye* (https://github.com/sascha2schroeder/popEye), which provide higher level tools for processing and analyzing reading data more generally. In time, we hope that the algorithms might also be implemented in other software packages.

Finally, we have introduced two novel methods in this paper that are distinct from those that have previously been presented. The `warp` algorithm, in particular, showed great promise and is likely to be especially useful to researchers working on reading development in children. We also hope that connecting the literature on vertical drift to sequence alignment techniques might also open new avenues for future algorithm development.

Open Practices Statement

All code and data required to reproduce the analyses reported in this paper, as well as Matlab/Octave, Python, and R implementations of the algorithms, are available from the OSF archive associated with this paper: https://doi.org/10.17605/OSF.IO/7SRKG

References

Aach, J., & Church, G. M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics, 17*(6), 495–508. https://doi.org/10.1093/bioinformatics/17.6.495

Abdulin, E. R., & Komogortsev, O. V. (2015). Person verification via eye movement-driven text reading model. *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems*. IEEE. https://doi.org/10.1109/BTAS.2015.7358786

Beymer, D., & Russell, D. M. (2005). WebGazeAnalyzer: A system for capturing and analyzing web reading behavior using eye gaze. *CHI '05 extended abstracts on Human Factors in Computing Systems* (pp. 1913–1916). Association for Computing Machinery. https://doi.org/10.1145/1056808.1057055

Blythe, H. I., & Joseph, H. S. S. L. (2011). Children's eye movements during reading. In S. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 643–662). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0036

Blythe, H. I., Liversedge, S. P., Joseph, H. S. S. L., White, S. J., & Rayner, K. (2009). Visual information capture during fixations in reading for children and adults. *Vision Research, 49*(12), 1583–1591. https://doi.org/10.1016/j.visres.2009.03.015

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., & Batzoglou, S. (2003). Glocal alignment: Finding rearrangements during alignment. *Bioinformatics, 19*(supplement 1), i54–i62. https://doi.org/10.1093/bioinformatics/btg1005

Caiani, E. G., Porta, A., Baselli, G., Turiel, M., Muzzupappa, S., Pieruzzi, F., Crema, C., Malliani, A., & Cerutti, S. (1998). Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume. *Computers in Cardiology 1998*, 73–76. https://doi.org/10.1109/CIC.1998.731723

Carl, M. (2013). Dynamic programming for re-mapping noisy fixations in translation tasks. *Journal of Eye Movement Research*, *6*(2), Article 5. https://doi.org/10.16910/jemr.6.2.5

Cohen, A. L. (2013). Software for the automatic correction of recorded eye fixation locations in reading experiments. *Behavior Research Methods*, *45*(3), 679–683. https://doi.org/10.3758/s13428-012-0280-3

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615. https://doi.org/10.3758/s13428-016-0734-0

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Hyrskykari, A. (2006). Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. *Computers in Human Behavior*, *22*(4), 657–671. https://doi.org/10.1016/j.chb.2005.12.013

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: Towards a model of real-world reading. *Journal of Computer Assisted Learning*, *33*(3), 193–201. https://doi.org/10.1111/jcal.12189

Joseph, H. S. S. L., Liversedge, S. P., Blythe, H. I., White, S. J., & Rayner, K. (2009). Word length and landing position effects during reading in children and adults. *Vision Research*, *49*(16), 2078–2086. https://doi.org/10.1016/j.visres.2009.05.015

Keogh, E. J., & Pazzani, M. J. (2001). Derivative dynamic time warping. In V. Kumar & R. Grossman (Eds.), *Proceedings of the 2001 SIAM International Conference on Data Mining* (pp. 1–11). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972719.1

Kuperman, V., Matsuki, K., & Van Dyke, J. A. (2018). Contributions of readerand text-level characteristics to eye-movement patterns during passage reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1687–1713. https://doi.org/10.1037/xlm0000547

Kuperman, V., Siegelman, N., Schroeder, S., ..., & Usal, K. (under review). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-Movement Corpus (MECO).

Lei, H., & Govindaraju, V. (2005). A comparative study on the consistency of features in on-line signature verification. *Pattern Recognition Letters*, *26*(15), 2483–2489. https://doi.org/10.1016/j.patrec.2005.05.005

Lima Sanches, C., Kise, K., & Augereau, O. (2015). Eye gaze and text line matching for reading analysis. *Adjunct proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 1227–1233). Association for Computing Machinery. https://doi.org/10.1145/2800835.2807936

Lohmeier, S. (2015). *Experimental evaluation and modelling of the comprehension of indirect anaphors in a programming language* (Master's thesis). Technische Universität Berlin. http://www.monochromata.de/master_thesis/ma1.3.pdf

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods, 50*(2), 826–833. https://doi.org/10.3758/s13428-017-0908-4

Martinez-Gomez, P., Chen, C., Hara, T., Kano, Y., & Aizawa, A. (2012). Image registration for text-gaze alignment. *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (pp. 257–260). Association for Computing Machinery. https://doi.org/10.1145/2166966.2167012

Mishra, A., Carl, M., & Bhattacharyya, P. (2012). A heuristic-based approach for systematic error correction of gaze data for reading. *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing* (pp. 71–80).

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nüssli, M.-A. (2011). *Dual eye-tracking methods for the study of remote collaborative problem solving (Doctoral dissertation).* École Polytechnique Fédérale de Lausanne. https://doi.org/10.5075/epfl-thesis-5232

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillère, B. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance, 10*(2), 250–257. https://doi.org/10.1037/0096-1523.10.2.250

Palmer, C., & Sharif, B. (2016). Towards automating fixation correction for source code. Proceedings of the 9th biennial ACM *Symposium on Eye Tracking Research & Applications* (pp. 65–68). Association for Computing Machinery. https://doi.org/10.1145/2857491.2857544

Parker, A. J., Slattery, T. J., & Kirkby, J. A. (2019). Return-sweep saccades during reading in adults and children. *Vision Research, 155*, 35–43. https://doi.org/10.1016/j.visres.2018.12.007

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pescuma, V. N., Crepaldi, D., & Ktori, M. (in prep.). EyeReadIt: A developmental eye-tracking corpus of text reading in Italian. https://doi.org/10.17605/OSF.IO/HX2SJ

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 940–961. https://doi.org/10.1037//0278-7393.24.4.940

Rayner, K. (1986). Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology, 41*(2), 211–236. https://doi.org/10.1016/0022-0965(86)90037-8

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124* (3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Binder, K. S., Ashby, J., & Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research*, *41*(7), 943–954. https://doi.org/10.1016/S0042-6989(00)00310-2

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445–476. https://doi.org/10.1017/S0140525X03000104

Riesen, K., Hanne, T., & Schmidt, R. (2018). Sketch-based user authentication with a novel string edit distance model. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48*(3), 460–472. https://doi.org/10.1109/TSMC.2016.2601074

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49. https://doi.org/10.1109/tassp.1978.1163055

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics, 74* (1), 5–35. https://doi.org/10.3758/s13414-011-0219-2

Schroeder, S. (2019). popEye – An R package to analyse eye movement data from reading experiments. https://github.com/sascha2schroeder/popEye

Sereno, S. C., O’Donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate-bias effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 335–350. https://doi.org/10.1037/0096-1523.32.2.335

Sibert, J. L., Gokturk, M., & Lavine, R. A. (2000). The reading assistant: Eye gaze triggered auditory prompting for reading remediation. *Proceedings of the 13th annual ACM Symposium on User Interface Software and Technology* (pp. 101–107). Association for Computing Machinery. https://doi.org/10.1145/354401.354418

Špakov, O., Istance, H., Hyrskykari, A., Siirtola, H., & Räihä, K.-J. (2019). Improving the performance of eye trackers with limited spatial accuracy and low sampling rates for reading analysis by heuristic fixation-to-word mapping. *Behavior Research Methods, 51*(6), 2661–2687. https://doi.org/10.3758/s13428-018-1120-x

Tiffin-Richards, S. P., & Schroeder, S. (2015). Children's and adults' parafoveal processes in German: Phonological and orthographic effects. *Journal of Cognitive Psychology, 27*(5), 531–548. https://doi.org/10.1080/20445911.2014.999076

Tiffin-Richards, S. P., & Schroeder, S. (2020). Context facilitation in text reading: A study of children's eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(9), 1701–1713. https://doi.org/10.1037/xlm0000834

Tomasi, G., van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*(5), 231–241. https://doi.org/10.1002/cem.859

Tormene, P., Giorgino, T., Quaglini, S., & Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine, 45*(1), 11–34. https://doi.org/10.1016/j.artmed.2008.11.007

Uchida, S. (2005). A survey of elastic matching techniques for handwritten character recognition. *IEICE Transactions on Information and Systems, E88-D*(8), 1781–1790. https://doi.org/10.1093/ietisy/e88-d.8.1781

Vadillo, M. A., Street, C. N. H., Beesley, T., & Shanks, D. R. (2015). A simple algorithm for the offline recalibration of eye-tracking data through best-fitting linear transformation. *Behavior Research Methods, 47* (4), 1365–1376. https://doi.org/10.3758/s13428-014-0544-1

Vakanski, A., Janabi-Sharifi, F., & Mantegh, I. (2014). Robotic learning of manipulation tasks from visual perception using a Kinect sensor. *International Journal of Machine Learning and Computing, 4*(2), 163–169. https://doi.org/10.7763/IJMLC.2014.V4.406

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics, 4* (1), 52–57. https://doi.org/10.1007/BF01074755

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vitu, F., Kapoula, Z., Lancelin, D., & Lavigne, F. (2004). Eye movements in reading isolated words: Evidence for strong biases towards the center of the screen. *Vision Research, 44*(3), 321–338. https://doi.org/10.1016/j.visres.2003.06.002

Yamaya, A., Topić, G., & Aizawa, A. (2017). Vertical error correction using classification of transitions between sequential reading segments. *Journal of Information Processing, 25*, 100–106. https://doi.org/10.2197/ipsjjip.25.100

Zhang, Y., & Hornof, A. J. (2011). Mode-of-disparities error correction of eye-tracking data. *Behavior Research Methods, 43* (3), 834–842. https://doi.org/10.3758/s13428-011-0073-0

Zhang, Y., & Hornof, A. J. (2014). Easy post-hoc spatial recalibration of eye tracking data. *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 95–98). Association for Computing Machinery. https://doi.org/10.1145/2578153.2578166

Chapter IV

Eye Movements During Natural Reading Reveal Sensitivity to Orthographic Regularities in Children

In alphabetic languages, learning to map letters (graphemes) onto sounds (phonemes) provides young children with the initial tools to translate the printed form of words (orthography) into their spoken equivalent (phonology), through which information about meaning can be accessed. Phonological decoding, however, is not sufficient for fluent reading. In order to become skilled readers, children are required to develop a sophisticated orthographic system that would enable them to recognise printed words rapidly and efficiently, and to map orthography directly to meaning (Castles & Nation, 2006, 2008). This process, known as orthographic learning, has been the focus of extensive research and, over the last three decades, we have learned a great deal about the changes that children's orthographic knowledge undergoes as a result of their experiences with printed words (for a review, see Castles et al., 2018). Arguably, however, one aspect of orthographic learning about which we know less is the how. How do children acquire orthographic information about words? Or more specifically, what might be the underpinning learning mechanism(s) that enables them to do so?

The present study examines the hypothesis that the acquisition of orthographic knowledge can be supported by a general learning mechanism that enables children to exploit the statistical regularities present in the orthographic information to which they are exposed. The underlying premise of this proposal is that such a mechanism is not specific to the processing of linguistic material. It refers, instead, to a fundamental domain-general mechanism that underpins the capacity of the human cognitive system to implicitly use statistical properties of the input such as the frequency, distributional variability, and probability with which events co-occur, in order to discover patterned regularities in the environment (for reviews, see Armstrong et al., 2017; Aslin, 2017; Christiansen, 2019; Frost et al., 2019; Newport, 2016).

One domain for which statistical learning has been particularly well documented is visual processing. Specifically, evidence from visual statistical learning studies suggests that observers compute co-occurrence probabilities for visual objects, both when these objects are presented sequentially or embedded within complex scenes (e.g., Fiser & Aslin, 2001; 2002; 2005; Kim, Seitz, Feenstra, & Shams, 2009; Orbán, Fiser, Aslin, & Lengyel, 2008). Hence, by adopting the perspective that written words are instances of visual stimuli, orthographic learning can be easily conceptualised as a form of visual statistical learning: in the same way we (infants, children, adults) can rapidly and automatically extract patterns of statistical regularities from a flow of abstract visual information, we can detect regularities in the distribution of letters and letter sequences in words from exposure to print. Such orthographies regularities could include, for example, the frequency of letter co-occurrences in written words (e.g., in English the letters *S* and *A* co-occur more frequently in words than the letters *J* and *A*, the letter *R* is more often doubled than the letter *D;* Chetail, 2015). Alternatively, orthographic regularities can be estimated on the basis of letter transitional probabilities (e.g., in English the probability that the letter *G* is followed by the letter *O* is roughly three times higher than the probability that *H* is followed by *O*; Chetail, 2015).

A limited experience appears to be sufficient for young readers to capture some orthographic regularities of their written language. Pacton and colleagues (Pacton, Perruchet, Fayol, & Cleeremans, 2001), for example, provided a demonstration of such knowledge in French-speaking children (Grades 1- 5). They showed that children would consider items like *illaro* to be more word-like than items like *ivvaro*, consistent with the fact that *l* is frequently doubled in the French language, whereas *v* is almost never doubled. Similarly, children would consider items like *bukkox* to be more word-like than items like *bukoxx*, consistent with the fact that even though neither the *k* nor *x* is doubled in the language, double consonants never appear at the end of French words (for similar findings in English, see Cassar & Treiman, 1997). Sensitivities to the frequency and position legality of double consonants in words were shown to increase with grade level, but, interestingly, they were already present in the youngest group of participants after having received only a few months of print exposure.

However, while children appear to rapidly pick up information about the patterns with which letters co-occur in words, the role of this information for the development of a skilled orthographic word-recognition system remains grossly underspecified. One main reason for this, is that, with the exception of the aforementioned work, developmental research has examined the relationship between statistical learning and learning to read by measuring the two abilities separately and seeking to provide a correlational link between them (Arciuli & Simpson, 2012; von Koss et al., 2019; West et al., 2018). However, as informative as this approach might be in highlighting an association between the two capacities, that for statistical learning and that for reading proficiency, it cannot speak to the direct influence of statistical learning on the development of reading processes.

Furthemore, the empirical evidence available from studies conducted with adults does not provide a coherent account about the impact of orthographic regularities acquired via statistical learning on word reading, thus limiting our understanding of its contribution to skilled reading. This is clearly illustrated in Chetail's (2015) review of the relevant literature. First, the data yield a mixture of results: while orthographic regularities are sometimes reported to exert a facilitatory effect on word recognition processes (e.g., Conrad et al., 2009; Lima & Inhoff, 1985; Massaro et al., 1981), other times are shown to have a detrimental effect (e.g., Hand et al., 2012; Rice & Robinson, 1975; Westbury & Buchanan, 2002) or no effect at all (e.g., Andrews, 1992; Johnston, 1978; Keuleers et al., 2012). Moreover, this mixed pattern of evidence cannot be ascribed to methodological differences because contradictions are present both across and within experimental paradigms (e.g., lexical decision; Chetail et al., 2014; Conrad et al., 2009; Keuleers et al., 2012; Schmalz & Mulatti, 2017). Finally, there is a clear lack of systematicity with regard to the orthographic regularities that have been thus far examined. Even though, as already indicated, there are several ways to express orthographic regularities, most of the time these are expressed in terms of the frequency with which clusters of letters (i.e., n-grams) occur in the written language. However, while in principle readers can exploit regularities of multiple grain sizes (e.g., bigrams, trigrams, qudrigrams; Vinckier, Dehaene, Jobert, Dubus, Sigman & Cohen, 2007), the investigation of bigram frequency effects appears to have dominated this line of research. Perhaps this is unsurprising, given the fact that bigrams have been ascribed a special status in several theories of visual word processing (e.g., Grainger & Van Heuven, 2003; Grainger & Ziegler, 2011; Whitney, 2001). Nevertheless, as a consequence of this specific focus our understanding concerning the effects that orthographic regularities of multiple grain sizes might exert on visual word recognition remains limited.

More recently, a different approach to examining the role of orthographic regularities acquired via statistical learning was adopted by two learning studies using artificial script. In a study by Chetail (2017) skilled readers were exposed to a stream of unfamiliar character strings, with several bigrams occurring very frequently. Following this exposure, participants demonstrated considerable sensitivity to high-frequency bigrams, that is, they were more likely to judge a previously unseen string as “word-like”, if it contained a high-frequency bigram, and if that bigram appeared in the same position as in the strings seen in exposure. Furthermore, participants detected letters coming from a high frequency bigram more rapidly than letters coming from a low frequency bigram. Using a similar design, Lelonkiewicz et al. (2020) extended these findings by demonstrating that skilled readers also capture information about the frequency and position of larger clusters of co-occurring characters (3- and 4-character long). Although these experiments involved pseudo-linguistic materials clear parallels between these findings and those reported in studies with children can be drawn: even after a short exposure readers can encode pure orthographic regularities (i.e., with no reference to higher order linguistic information) present in the script. However, exposure to an artificial script, albeit the control that it offers for experimentation, is distinctly different from exposure to a real script that encompasses the richness and the many nuances of a written language. Furthermore, while learning studies with adults can provide a proxy of how learning proceeds during the course of development, whether the learning mechanisms used by skilled adult readers are comparable to those employed by developing readers is not currently known.

The main aim of the present work was to provide an investigation into the role of orthographic regularities in reading development. To achieve this, we examined sensitivities to clusters of co-occurring letters (i.e., bigrams, trigrams, quadrigrams) across a wide range of developing readers (Grades 3-6) and a control group of skilled readers. If exposure to print leads readers to capture orthographic regularities via a statistical learning mechanism, then children should become more sensitive to the frequency with which letter sequences occur in the language as they accumulate experiences with printed words (as a function of school grade). Alternatively, early sensitivities to letter co-occurrence might diminish as children build up a set of well-consolidated visual lexical memories (an orthographic lexicon) during the course of reading development. We also sought to adopt a more ecological experimental approach than what has been previously used in the literature. To this aim, we investigated children's sensitivity to n-grams in their eye movements, as they silently read multi-line passages of text for understanding. made use of a large database of eye movement during natural reading in children, which we developed in the lab (Pescuma et al., in prep.; Chapter II in this thesis). The database was built on Italian material, and is called. It is based on a group of 141 developing readers, spanning from Grade 3 to Grade 6, and a control sample of 33 skilled, adult readers. These participants were asked to silently read for understanding several multi-lined stories while their eye movements were recorded.

Methods

We conducted our analyses using a database of eye movement during natural reading in children, which we developed in the lab (*EyeReadIt*; Pescuma et al., in prep.). The database is described more fully in Chapter II of this thesis. *EyeReadIt* is based on Italian, and contains data from a group of 141 developing readers spanning from Grade 3 to Grade 6 (73 girls, 68 boys; N is 37, 20, 41 and 43 for 3rd, 4th, 5th and 6th graders, respectively). For comparison, *EyeReadIt* also includes a control sample of 33 skilled, adult readers (21 female, 12 male; age range 19-33). These participants were asked to silently read passages of text for understanding while their eye movements were recorded.

Eye-tracking metrics (dependent variables)

We extracted from *EyeReadIt* all the data pertaining to words, which yielded 202657 individual fixations (174608 for children and 28049 for adults), on 1566 individual word tokens and 762 word types. The lexical-orthographic features of the word types are illustrated in Table 1. Based on these data, we extracted the following word-level eye-tracking measures for each trial. First-of-many fixation duration (*FoM*) refers to the duration of those instances of first fixation in which further fixations on the same word followed during first pass. This index was taken to track an early stage of lexical processing. Gaze duration (*GD*) is the summed duration of all fixations performed on a word during first-pass reading. This metric is typically considered as the most direct measure of lexical access (Inhoff, 1984; Just & Carpenter, 1980). Total reading time (*TRT*) is the summed duration of all fixations performed across all runs, and is typically taken to also reflect later, post-lexical processing (e.g., meaning integration at the sentence or passage level). The three final datasets for the analyses contained 28481, 116484 and 93253 datapoints, for *FoM*, *GD* and *TRT* respectively. The distribution and correlations of the three eye-tracking metrics are illustrated in Figure 1.

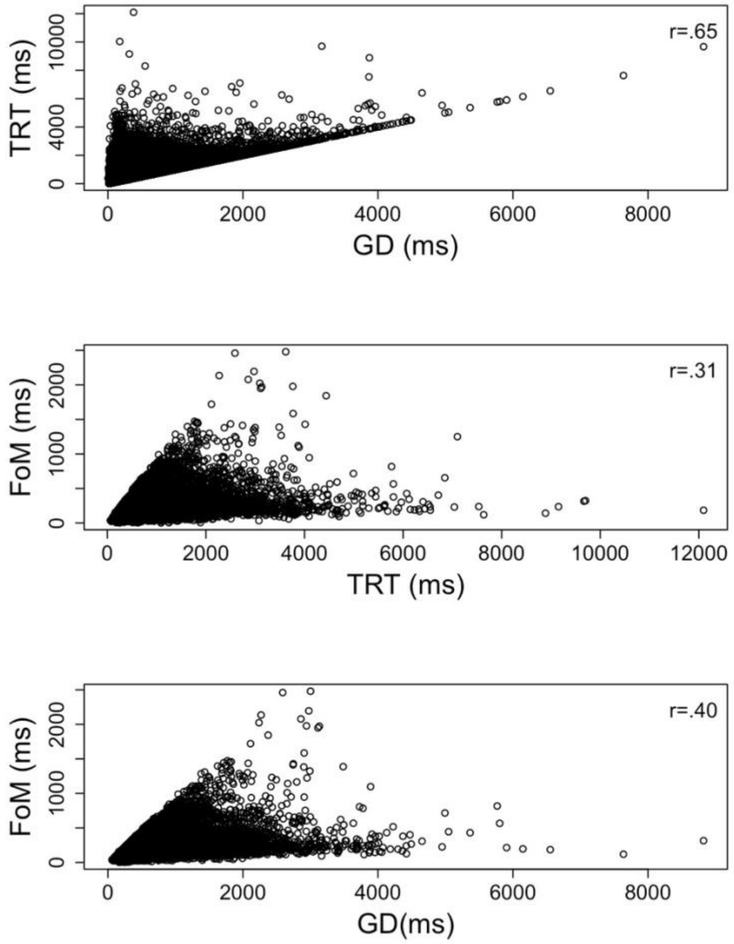


Figure 1 Scatter plots illustrating the distribution and pairwise correlations of all three eye-tracking measures (gaze duration, *GD*, first-of-many fixation duration, *FoM*, and total reading time, *TRT*, all durations expressed in milliseconds, ms) that are treated as dependent variables in the analyses. The correlation between *GD* and *TRT* is fairly high (.65), showing that most words were not fixated anymore after first pass reading. It also suggests that these two variables mostly track similar processes, at least in the present data.

Table 1 Summary of the lexical and sublexical features of the stimuli, as obtained from EyeReadIt. All frequency measures are expressed in Zipf units and were obtained using SUBTLEX-IT (Crepaldi et al., 2016).

Number of word tokens	1566
Number of word types	762
Unique parts of speech	11
Mean word count per text (range)	130.5 (109–170)
Mean word length (range)	4.66 (1–15)
Mean word frequency (range)	5.35 (1.19–7.26)
Mean average bigram frequency (range)	6.31 (3.46–6.82)
Mean average trigram frequency (range)	5.42 (2.88–6.44)
Mean average quadrigram frequency (range)	4.60 (0–5.89)
Mean maximal bigram frequency (range)	6.57 (3.46–6.83)
Mean maximal trigram frequency (range)	5.78 (2.88–6.44)
Mean maximal quadrigram frequency (range)	4.97 (0–5.96)
Mean minimal bigram frequency (range)	5.96 (3.04–6.82)
Mean minimal trigram frequency (range)	5.02 (2.06–6.44)
Mean minimal quadrigram frequency (range)	4.20 (0–5.89)

N-gram frequency metrics (independent variables)

For each individual word for which we had an eye-tracking metric to model, we considered three different sizes of letter n-grams – bigrams, trigrams and quadrigrams. For each n-gram size, we computed minimal, maximal and average frequency. So, for example, the word *house* would be analysed as the sequence of bigrams *ho*, *ou*, *us* and *se*. The log Zipf frequency of these bigrams (calculated on SUBTLEX-UK, Van Heuven et al., 2014) is 3.66, 3.74, 3.67, and 3.85, respectively. Therefore, the minimal bigram frequency for the word *house* would be 3.66, its maximal bigram frequency would be 3.85, and its average bigram frequency would be (3.66+3.74+3.67+3.85)/4=3.73. The same procedure would be applied to the trigrams in the word (i.e., *hou*, *ous* and *use*), and its quadrigrams (i.e., *hous*, *ouse*). We limited our analysis to words that were at least one letter longer than the relevant n-grams (i.e., three-letter words for bigrams, four-letter words for trigrams and five-letter words for quadrigrams), so as to ensure that at least two n-grams were considered for each word. For simplicity, we did not consider open bigrams (e.g., Grainger & Whitney, 2004). N-gram frequency was computed using Zipf units (Van Heuven et al., 2014), extracted from SUBTLEX-IT (Crepaldi et al., 2016).

Because n-gram statistics likely correlate with other established predictors at the word level, we also considered word frequency and word length, so as to ensure that whatever n-gram effect we might find is not a spurious effect out of these correlations (see below). Word frequency was also computed based on SUBTLEX-IT, and was modelled in Zipf units (Van Heuven et al., 2014).

Statistical analysis

Because the n-gram frequency metrics described above were correlated, and also correlated with word length and word frequency (see Results), we first subjected all these variables to a Principal Component Analysis (PCA) with a Varimax rotation (e.g., Grice, 2001; Morucci et al., 2018; Rencher, 1992). To this aim, we used the R function principal from the R package *psych* (version 2.0.12; Revelle, 2020). This procedure rotates the variable space so that each dimension (Principal Component) is uncorrelated with the others, and loads maximally different onto the original variables set, so as to be psychologically interpretable (see below).

We used these newly obtained Principal Components to model the eye-tracking metrics with linear mixed-effects models. The dependent variables were all log-transformed to better approximate a normal distribution. The models included as fixed effects six Principal Components (see below), grade, reading speed and non-verbal intelligence, in addition to the interaction between grade and each of the six Principal Components, to track the evolutionary pattern of sensitivity to letter co-occurrence statistics. The models also included a random intercept for items and participants. Significance was assessed via the comparison between hierarchical models (using the function Anova from the package *car* in R; Fox & Weisberg, 2019) using Type III sum-of-squares and chi-square Wald tests. The nature of the effects was illustrated through the computation of model-based estimates (via the function effects from the package *effects* in R; Fox, 2003; Fox & Weisberg, 2019) and the significance of the model parameters Beta (via the function summary in R). Following Baayen and Milin (2010), we ensured that that the effects were not driven by unduly influential outliers, by fitting additional models from which datapoints with residuals higher than 2.5 SD in absolute value were excluded. No difference between the main model and these “sanity check” models emerged, in terms of patterns or significance level of the effects.

Results

Correlational structure of the n-gram variables and PCA

Figure 2 illustrates the correlational structure of the n-gram variables, word length and word frequency.

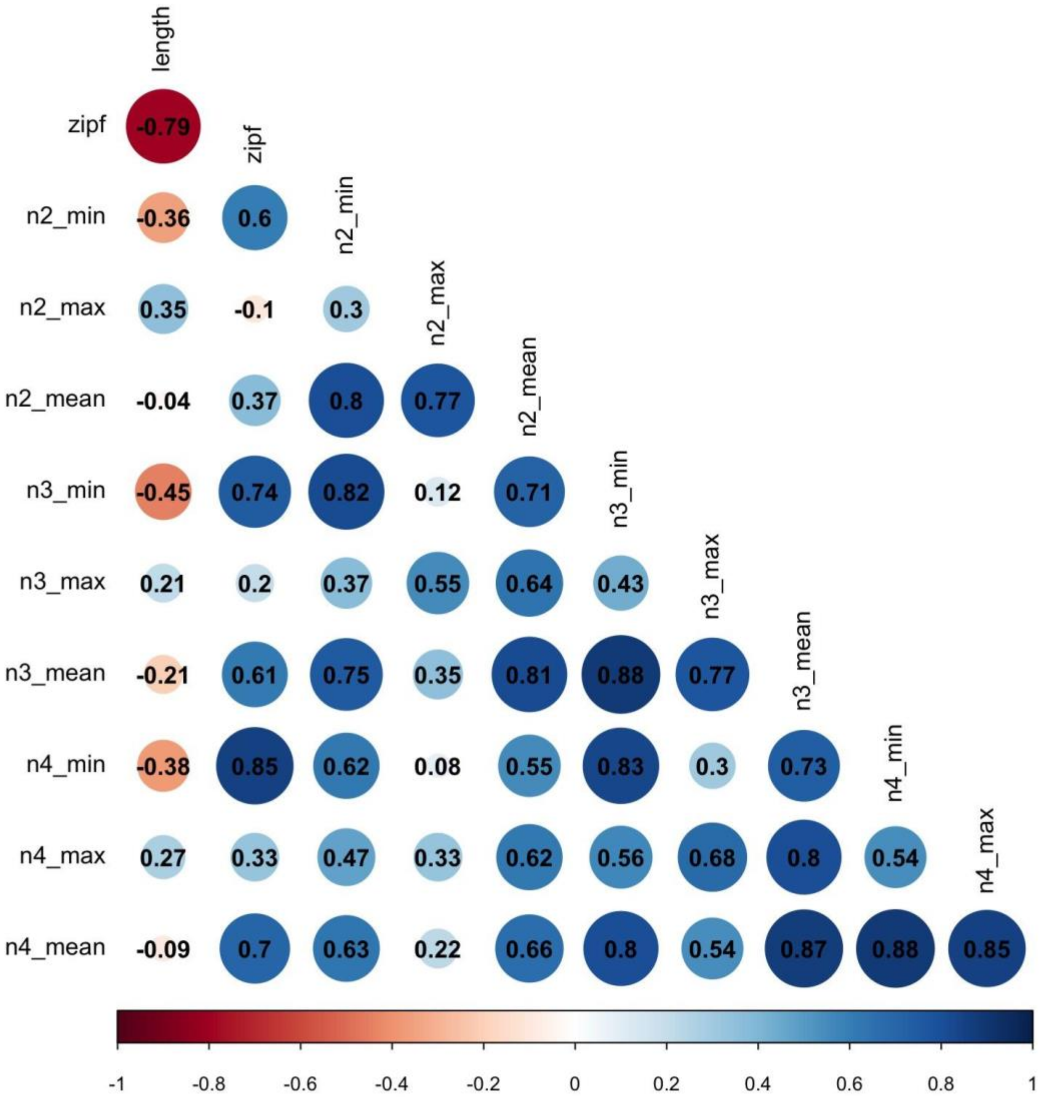


Figure 2 Correlational matrix between the n-gram variables, word length and word frequency. The size of the circles is proportional to the strength of the correlation; color codes for the direction of the correlation (blue is positive, red is negative).

Unsurprisingly, there are several strong correlations; we designed our study to be as naturalistic as possible and thus we did not try to artificially shrink correlations that would tend to naturally occur in an unselected sample. Clearly, this correlational structure makes it impossible to use all the predictors together in a single regression model (e.g., the condition number K is 353, while safe predictor independence is typically indicated by K<30; Baayen, 2008). As mentioned in the Methods, we performed a Principal Component Analysis with a Varimax rotation in order to obtain a set of variables that would be usable in a regression model, and also to understand the structure that underlies these correlations (e.g., Grice, 2001; Morucci, Bottini and Crepaldi, 2018; Rencher, 1992). As illustrated in Figure 3, six Principal Components (PC) accounted for a substantial amount of variance, and were therefore taken into consideration. Importantly, the Varimax rotation ensured that their correlations were quite small (range = –.24-10).

The interpretation of the six PCs is illustrated in Figure 4. The first component, which accounts for .23 of the variance, correlates most strongly with word frequency, and minimal and average quadrigram frequency. Overall, it clearly represents better the statistics of larger units than of smaller ones. Also, because it loads high on whole-word frequency, its eventual effect cannot be taken as a clear sign of sensitivity to sub-word, n-grams statistics. PC2, which accounts for .21 of the overall variance, loads heavily on minimal bigram frequency and, to a lesser extent, to average bigram and trigram frequency, and to minimal trigram frequency; therefore, it seems to track the minimal and average frequency of smaller n-grams. PC3, which explains .15 of the variance, loads heavily on maximal quadrigram frequency, and also tracks average quadrigram frequency (and, to a somewhat lesser extent, maximal and average trigram frequency). This variable nicely represents maximal and average frequency of larger n-grams, particularly of quadrigrams. PC4 (.14 of the overall variance) clearly tracks maximal and average bigram frequency; it seems to be the smaller n-grams counterpart of PC3, particularly tuned to bigrams. PC5, which amounts to .12 of the variance, loads particularly on maximal and average trigram frequency. PC6, instead, correlates very heavily with word length (and, to some extent, to word frequency); this variable accounts for .11 of the overall variance.

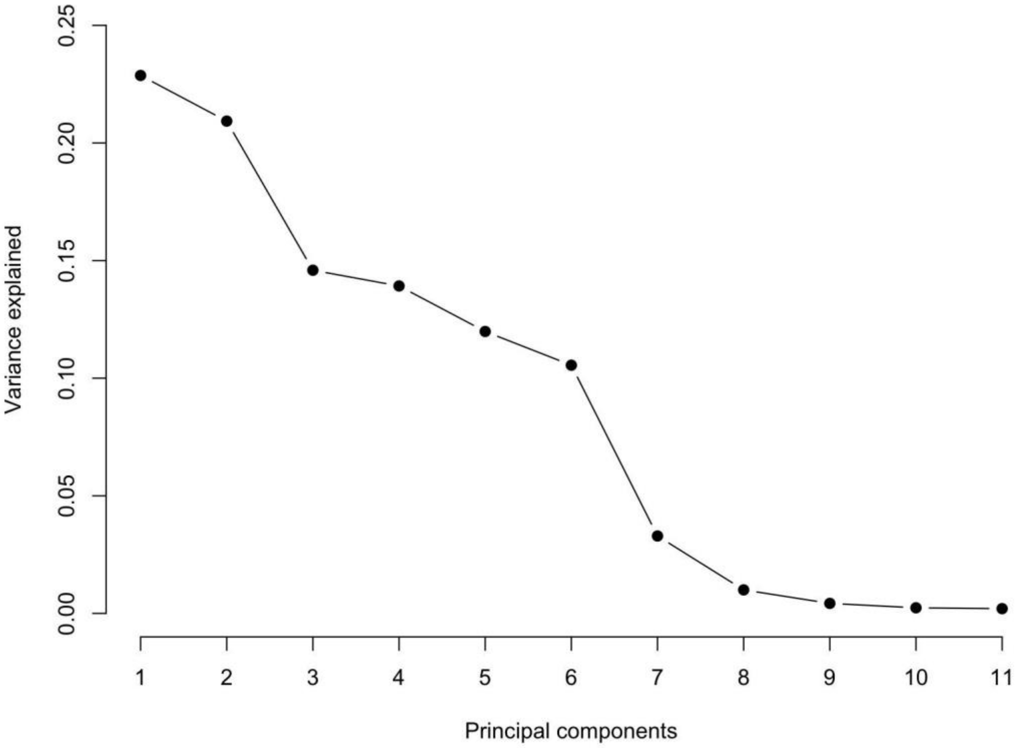


Figure 3 Proportion of explained variance by the 11 Principal Components that emerged from the Varimax-rotated Principal Component Analysis.

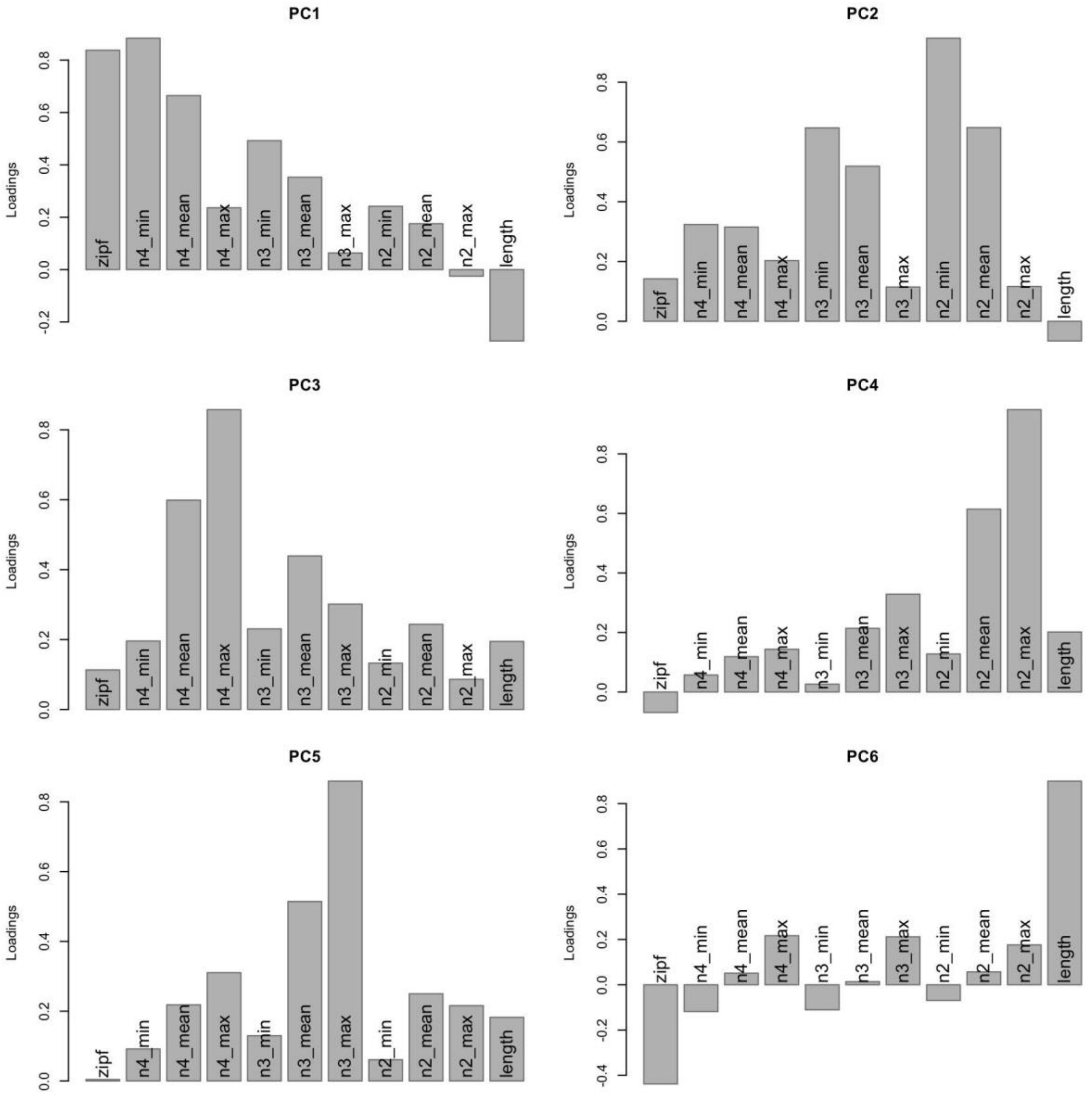


Figure 4 Illustration of the mapping between the six Principal Components that explain most of the variance, and the 11 original variables subjected to PCA. The Y axis reports loadings, which are interpretable as the correlation between the Principal Components and the original variables.

Some important insight emerges here. First, PC2, PC3, PC4 and PC5 show little correlation with word frequency and word length, and therefore can be considered as genuine sub-word n-gram variables whose effect would suggest that children are indeed sensitive to the statistics of such units during natural reading. Second, it turned out very difficult to isolate the effect of average n-gram frequency; in fact, this variable always goes together with either minimal (PC2) or maximal n-gram frequency (PC3, PC4 and PC5). So, despite this is the n-gram variable that was mostly investigated, it turns out to be the least distinct in this unselected sample of language (which might perhaps contribute to explain why data on average bigram frequency are largely inconsistent; e.g., Chetail, 2015). Third, it seems difficult to disentangle n-grams of different size for what concerns minimal/average frequency; PC2 hinges on both bigrams and trigrams, while quadrigrams are captured by the same PC as word frequency, and therefore cannot be assessed independently. However, it is possible to disentangle n-grams of different size with respect to maximal/average frequency; PC3, PC4 and PC5 load particularly on quadrigrams, bigrams and trigrams, respectively. Overall, the structure of the six most relevant PCs allows us to distinguish quite well the role of smaller n-grams (PC2 and PC4) and larger n-grams (PC3 and PC5). Also nicely (and by no means obvious), the role of larger n-grams seems to be identifiable independently of word-level variables, as PC3 and PC5 loads only mildly to word frequency and word length.

Correlational structure of the n-gram variables and PCA

Equipped with the new variables that emerged from the Varimax PCA, we can now move on to the analysis of the eye-tracking data. To make the interpretation of the PCs easier, we renamed them as follows: *PC1-wordFreq*, *PC2-minAv-smallNgrams*, *PC3-maxAv-4grams*, *PC4-maxAv-2grams*, *PC5-maxAv-3grams*, and *PC6-wordLength*. We will start from the analysis of gaze duration, which can be thought of as the most direct eye-tracking metric for visual word identification. We will then check more specifically what happens at the earlier (*FoM*) and later (*TRT*) stages of this process.

Gaze duration was modulated significantly by all Principal Components, except *PC2-minAv-smallNgrams* (see Table 2). The effects are illustrated in Figure 5.

Table 2 Overall significance of the gaze duration effects. *df*, numbers of degrees of freedom. Significant effects appear in bold.

	Chi-square	***df***	***p***
PC1-wordFreq	**237.19**	**1**	**<.001**
PC2-minAv-smallNgrams	1.54	1	.21
PC3-maxAv-4grams	**6.45**	**1**	**.01**
PC4-maxAv-2grams	**24.44**	**1**	**<.001**
PC5-maxAv-3grams	**15.11**	**1**	**<.001**
PC6-wordLength	**304.79**	**1**	**<.001**
Grade	**460.50**	**4**	**<.001**
MT	**130.87**	**1**	**<.001**
Raven	2.59	1	.11
pc1_wordFreq:Grade	**83.56**	**4**	**<.001**
pc2_minAv_smallNgrams:Grade	3.05	4	.55
pc3_maxAv_4grams:Grade	.56	4	.97
pc4_maxAv_2grams:Grade	7.91	4	.09
pc5_maxAv_3grams:Grade	**11.71**	**4**	**.02**
pc6_wordLength:Grade	**50.26**	**4**	**<.001**

PC1-wordFreq and *PC6-wordLength* interact significantly with Grade, mirroring the results reported above for these variables; this constitutes further validation of the results of the PCA. Also, given that all PCs have a similar range, the slopes reported in Figure 5 reflect the relevant effect size; it is clear, therefore, that *PC1-wordFreq* and *PC6-wordLength* have a much larger effect on gaze duration than the other PCs. This suggests that any effect of n-gram frequency is quite smaller than the effects of word frequency and word length, perhaps unsurprisingly.

Nevertheless, *PC3-maxAv-4grams*, *PC4-maxAv-2grams* and *PC5-maxAv-3grams* were statistically significant, showing that n-gram frequency effects can be observed independently of word length and word frequency. The former two effects did not interact with Grade, and show that children (and adults) fixate longer on words whose maximal/average n-gram frequency is higher. PC3 tracks larger n-grams, while PC4 is more tuned to smaller n-grams; yet, their effects do not seem to differ. *PC5-maxAv-3grams*, instead, does show a significant interaction with Grade; again, readers gaze longer on words with higher n-gram frequency, but the effect seems to shrink with age (Figure 5, panel (e)). More specifically, 3rd and 5th graders differ in a statistically significant way from adults, and 4th and 6th graders show a solid trend in the same direction (Table 3, panel (b)). Also, the effect of *PC5-maxAv-3grams* is significant and strong in all groups of children, while it is not significant in adults (although it does show a trend; Table 3, panel (c)).

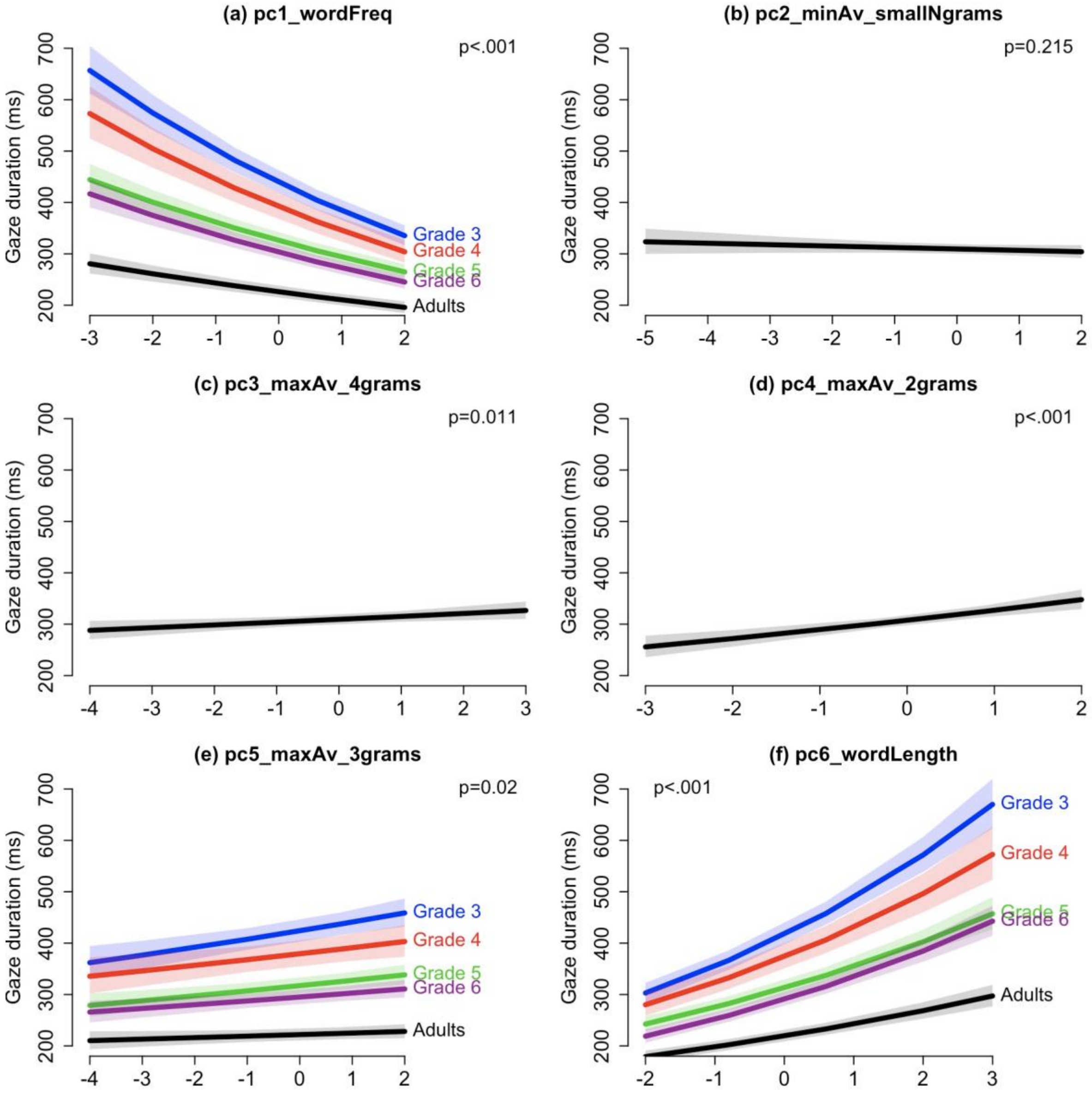


Figure 5 Effects of the six PCs on gaze duration. Colour panels are those where an interaction with Grade was significant; different colors refer to different grades as reported in the legend. The effect of PC2-minAv-smallNgrams, reported in panel (b), was not significant. The *p*-value of the effect is reported in each panel.

Table 3 Model parameters for gaze duration referring to the interaction between Grade and pc5_maxAv_3grams. *SE*, standard error; *df*, numbers of degrees of freedom. Panel (a) refers to a model where Grade was coded with a backward difference scheme, using 3rd grade as the reference level; therefore, the first parameter codes for the difference between 4th and 3rd grade, the second parameter codes for the difference between 5th and 4th grade, and so on. Here the emphasis is on the developmental pathway; however, since the change might be smooth and relatively small between consecutive grades (i.e., the time scale of the change might be larger than one grade), this approach might not be sensitive enough to capture significant differences. Panel (b), on the contrary, refers to a model where Grade was dummy-coded with the adult sample as a reference level; therefore, each parameter illustrates the comparison between each grade and the adult participants. In Panel (c), we illustrate the parameter for the main effect of pc5_maxAv_3grams when we change the reference level in the dummy-coded model to each of the five levels, therefore effectively tracking the presence of a significant pc5_maxAv_3grams in each individual grade (and in the adult sample).

(a)

	Beta	*SE*	*t*	*df*	*p*
pc5_maxAv_3grams:Grade1 (3rd grade vs. 4th grade)	–.008	.010	–.88	61670	.38
pc5_maxAv_3grams:Grade2 (4th grade vs. 5th grade)	.001	.009	.17	61630	.87
pc5_maxAv_3grams:Grade3 (5th grade vs. 6th grade)	–.006	.007	–.87	61600	.39
pc5_maxAv_3grams:Grade4 (6th grade vs. adults)	–.013	.007	–1.74	61610	.08

(b)

	Beta	*SE*	*t*	*df*	*p*
pc5_maxAv_3grams:Grade1 **(3rd grade vs. adults)**	**.026**	**.008**	**3.20**	**61850**	**.001**
pc5_maxAv_3grams:Grade1 (4th grade vs. adults)	.017	.009	1.76	61640	.08
pc5_maxAv_3grams:Grade1 **(5th grade vs. adults)**	**.019**	**.007**	**2.51**	**61600**	**.01**
pc5_maxAv_3grams:Grade1 (6th grade vs. adults)	.012	.007	1.74	61610	.08

(c)

	Beta	*SE*	*t*	*df*	*p*
pc5_maxAv_3grams **(3rd grade)**	**.039**	**.009**	**4.41**	**1640**	**<.001**
pc5_maxAv_3grams **(4th grade)**	**.031**	**.010**	**2.93**	**2999**	**.003**
pc5_maxAv_3grams **(5th grade)**	**.032**	**.008**	**3.85**	**1257**	**<.001**
pc5_maxAv_3grams **(6th grade)**	**.026**	**.008**	**3.19**	**1144**	**.001**
pc5_maxAv_3grams (adults)	.013	.008	1.61	1312	.11

Overall, there seem to be clear effects of n-gram frequency, which, albeit smaller in size, are clearly distinguishable and independent from word-level effects. Also, maximal n-gram frequency seems to play a more important role than minimal n-gram frequency[1], and it correlates positively with gaze duration: higher maximal frequency leads to longer fixations. Evidence is more mixed for what concerns developmental change: sensitivity to *PC5-maxAv-3grams* does shrink with growing grades, while this is not the case for *PC3-maxAv-4grams* and *PC4-maxAv-2grams*.

FoM fixation duration was significantly modulated by *PC1_wordFreq*, *PC5_maxAv_3grams*, and *PC6_wordLength*, as reported in Table 4. None of these effects significantly interacted with Grade.

[1] Minimal/average quadrigram frequency is embedded within PC1, which does have a very strong effect on gaze duration. However, because PC1 also tracks word frequency, we cautiously consider it to reflect word-level processing.

Table 4 Overall significance of the first-of-many fixation duration effects. *df,* number of degrees of freedom. Significant effects appear in bold.

	Chi-square	***df***	***p***
PC1-wordFreq	**26.92**	**1**	**<.001**
PC2-minAv-smallNgrams	1.33	1	.25
PC3-maxAv-4grams	1.41	1	.23
PC4-maxAv-2grams	1.64	1	.20
PC5-maxAv-3grams	**4.48**	**1**	**.03**
PC6-wordLength	**4.54**	**1**	**.03**
Grade	**120.92**	**4**	**<.001**
MT	**33.92**	**1**	**<.001**
Raven	0.19	1	.67
pc1_wordFreq:Grade	2.54	4	.64
pc2_minAv_smallNgrams:Grade	3.21	4	.52
pc3_maxAv_4grams:Grade	3.98	4	.41
pc4_maxAv_2grams:Grade	5.68	4	.22
pc5_maxAv_3grams:Grade	4.21	4	.38
pc6_wordLength:Grade	3.84	4	.43

As illustrated in Figure 6, the duration of FoM fixations shrinks with *PC1_wordFreq* and grows with *PC6_wordLength*, in line with gaze duration. Also in line with gaze duration, FoM fixation grows with *PC5_maxAv_3grams*. In terms of effect size (represented by slope in the figure), word frequency appears to have a much stronger effect, whereas word length and maximal/average trigram frequency have a similar effect.

This pattern of results suggests that maximal/average trigram frequency modulates the fixation pattern at an early processing stage. Furthermore, quite interestingly, the information that is relevant at this early processing stage does not appear to change with age.

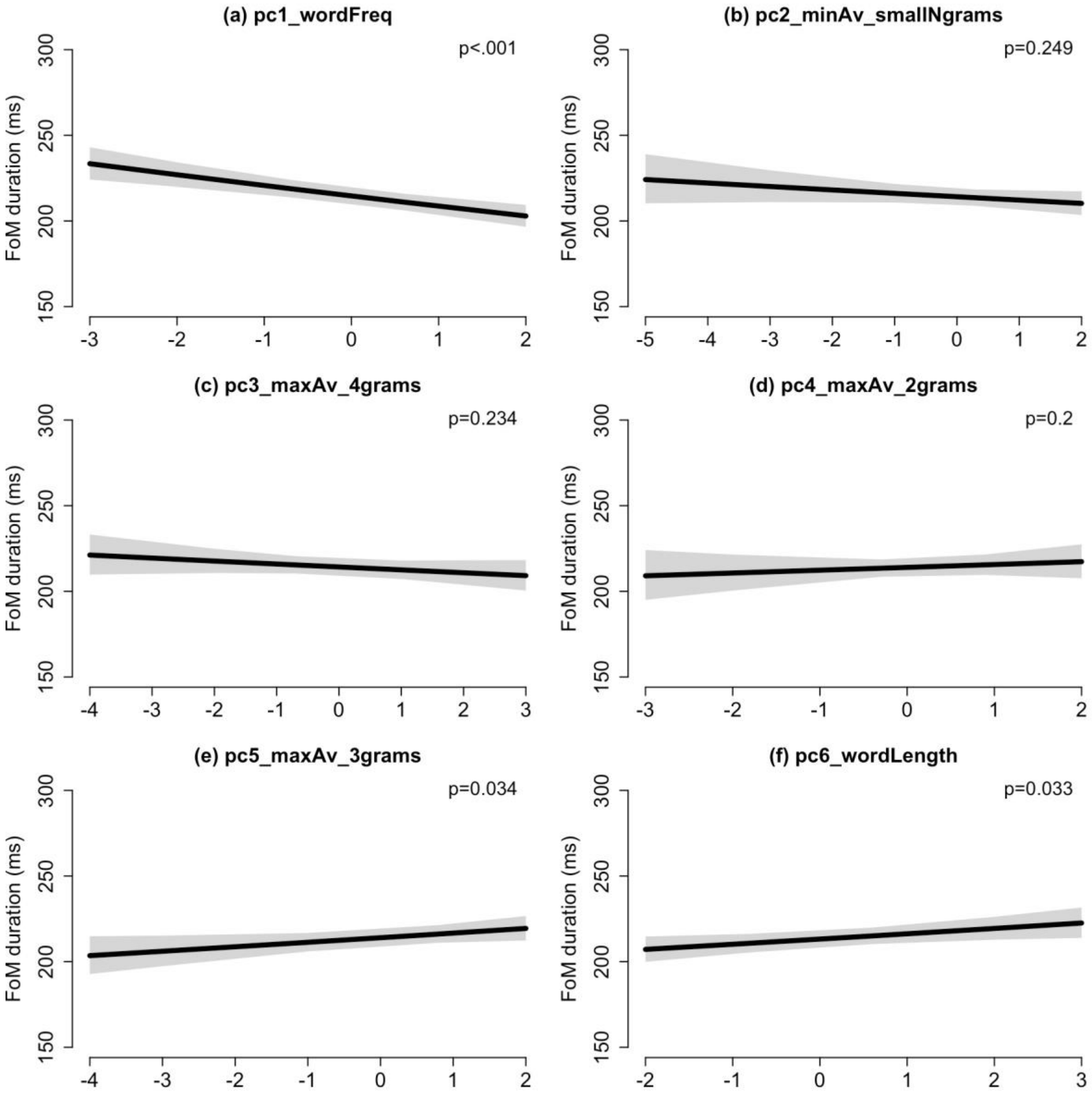


Figure 6 Effects of the six PCs on first-of-many fixation duration. No PC interacted significantly with Grade, therefore only their main effects are illustrated. The p-value of the effect is reported in each panel.

Total reading time is significantly modulated by most of the PCs considered here, all in interaction with Grade, as reported in Table 5. Besides *PC1-wordFreq* and *PC6-wordLength*, a significant effect also emerged for *PC3-maxAv-4grams*, *PC4-maxAv-2grams* and *PC5-maxAv-3grams*.

Table 5 Overall significance of the total reading time effects. Significance values as obtained via *car*::Anova(model). *df* = number of degrees of freedom. Significant effects appear in bold.

	Chi-square	***df***	***p***
PC1-wordFreq	**538.33**	**1**	**<.001**
PC2-minAv-smallNgrams	2.13	1	.14
PC3-maxAv-4grams	**27.23**	**1**	**<.001**
PC4-maxAv-2grams	**23.83**	**1**	**<.001**
PC5-maxAv-3grams	**31.87**	**1**	**<.001**
PC6-wordLength	**531.19**	**1**	**<.001**
Grade	**574.00**	**4**	**<.001**
MT	**223.67**	**1**	**<.001**
Raven	1.24	1	.26
pc1_wordFreq:Grade	**270.48**	**4**	**<.001**
pc2_minAv_smallNgrams:Grade	5.55	4	.23
pc3_maxAv_4grams:Grade	**12.52**	**4**	**.009**
pc4_maxAv_2grams:Grade	**18.48**	**4**	**.001**
pc5_maxAv_3grams:Grade	**46.66**	**4**	**<.001**
pc6_wordLength:Grade	**160.36**	**4**	**<.001**

As illustrated in Figure 7, higher n-gram frequency yields longer reading times, more markedly so in younger readers. The only component whose effect does not reach statistical significance is *PC2_minAv_smallNgrams*, in line with the results of the analyses of gaze and FoM fixation durations. As in the gaze duration analysis, the effects of *PC1-wordFreq* and *PC6-wordLength* are larger in size than the n-gram effects; nevertheless, again similarly to gaze duration, *PC3-maxAv-4grams*, *PC4-maxAv-2grams*, and *PC5-maxAv-3grams* are statistically significant, showing that n-gram frequency effects can be observed independently of word length and word frequency. All effects also significantly interact with Grade, and show that words whose maximal/average n-gram frequency is higher yield longer fixations, and that the effect shrinks with age. The effects do not seem to differ greatly across n-gram sizes; particularly, the effects of *PC3-maxAv-4grams*, which tracks quadrigrams, *PC4-maxAv-2grams*, tuned to bigrams, and *PC5-maxAv-3grams*, tracking trigrams, appear to have a very similar slope (Figure 7, panels (c), (d) and (e)). For all three n-gram metrics, we found a significant difference in their effect on total reading time between 6th graders and adults; as far as *PC5-maxAv-3grams* is concerned, there is also a significant difference in its effect between 3rd and 4th graders, and a trend is observed between 5th and 6th graders (Table 6, panel (a)). A significant difference in the effect of *PC3-maxAv-4grams* is observed between 4th graders and adults, 5th graders and adults, and 6th graders and adults, while for *PC4-maxAv-2grams* we found a significant difference between 3rd graders and adults, and 6th graders and adults, besides a trend between 4th graders and adults. A significant difference of the *PC5-maxAv-3grams* effect is observed between developing readers of each grade and adults (Table 6, panel (b)). Finally, the effects of all three PCs are significant and strong in all groups of children and in adults (Table 3, panel (c)).

Table 6 Model parameters for total reading time, referring to the interaction between Grade and pc3_maxAv_4grams, Grade and pc4_maxAv_2grams, and Grade and pc5_maxAv_3grams. *SE* = standard error; *df* = number of degrees of freedom. See Table 3 for a description of the contrast structure illustrated in each panel, adopted here as well. Significant effects appear in bold.

(a)

	Beta	***SE***	***t***	***df***	***p***
pc3_maxAv_4grams:Grade1 (3rd grade vs. 4th grade)	.008	.010	0.86	61672	.39
pc3_maxAv_4grams:Grade2 (4th grade vs. 5th grade)	-.004	.009	-0.43	61632	.67
pc3_maxAv_4grams:Grade3 (5th grade vs. 6th grade)	.009	.007	1.35	61617	.18
pc3_maxAv_4grams:Grade4 **(6th grade vs. adults)**	**-.025**	**.007**	**-3.56**	**61629**	**<.001**
pc4_maxAv_2grams:Grade1 (3rd grade vs. 4th grade)	–.026	.016	–1.63	61672	.10
pc4_maxAv_2grams:Grade2 (4th grade vs. 5th grade)	-.009	.015	-0.62	61639	.54
pc4_maxAv_2grams:Grade3 (5th grade vs. 6th grade)	.006	.011	0.52	61624	.60
pc4_maxAv_2grams:Grade4 **(6th grade vs. adults)**	**–.024**	**.011**	**–2.07**	**61614**	**.04**
pc5_maxAv_3grams:Grade1 **(3rd grade vs. 4th grade)**	**–.026**	**.010**	**–2.69**	**61667**	**.007**
pc5_maxAv_3grams:Grade2 (4th grade vs. 5th grade)	.004	.009	0.40	61636	.69
pc5_maxAv_3grams:Grade3 (5th grade vs. 6th grade)	–.012	.007	–1.80	61612	.07
pc5_maxAv_3grams:Grade4 **(6th grade vs. adults)**	**–.016**	**.007**	**–2.40**	**61617**	**.02**

(b)

	Beta	***SE***	***t***	***df***	***p***
pc3_maxAv_4grams:Grade1 (3rd grade vs. adults)	.011	.008	1.48	61857	.14
pc3_maxAv_4grams:Grade2 **(4th grade vs. adults)**	**.020**	**.009**	**2.10**	**61660**	**.04**
pc3_maxAv_4grams:Grade3 **(5th grade vs. adults)**	**.016**	**.007**	**2.17**	**61694**	**.03**
pc3_maxAv_4grams:Grade4 **(6th grade vs. adults)**	**.025**	**.007**	**3.56**	**61629**	**<.001**
pc4_maxAv_2grams:Grade1 **(3rd grade vs. adults)**	**.053**	**.012**	**4.28**	**61824**	**<.001**
pc4_maxAv_2grams:Grade2 (4th grade vs. adults)	.027	.015	1.77	61662	.08
pc4_maxAv_2grams:Grade3 (5th grade vs. adults)	.018	.012	1.53	61677	.13
pc4_maxAv_2grams:Grade4 **(6th grade vs. adults)**	**.024**	**.011**	**2.07**	**61614**	**.04**
pc5_maxAv_3grams:Grade1 **(3rd grade vs. adults)**	**.050**	**.008**	**6.59**	**61825**	**<.001**
pc5_maxAv_3grams:Grade2 **(4th grade vs. adults)**	**.025**	**.009**	**2.70**	**61647**	**.007**
pc5_maxAv_3grams:Grade3 **(5th grade vs. adults)**	**.028**	**.007**	**4.04**	**61664**	**<.001**
pc5_maxAv_3grams:Grade4 **(6th grade vs. adults)**	**.016**	**.007**	**2.40**	**61617**	**.02**

(c)

	Beta	*SE*	*t*	*df*	*p*
pc3_maxAv_4grams **(3rd grade)**	**.035**	**.009**	**3.94**	**1426**	**<.001**
pc3_maxAv_4grams **(4th grade)**	**.043**	**.010**	**4.18**	**2592**	**<.001**
pc3_maxAv_4grams **(5th grade)**	**.039**	**.008**	**4.64**	**1179**	**<.001**
pc3_maxAv_4grams **(6th grade)**	**.049**	**.008**	**5.85**	**1081**	**<.001**
pc3_maxAv_4grams **(adults)**	**.024**	**.009**	**2.76**	**1242**	**.006**
pc4_maxAv_2grams **(3rd grade)**	**.092**	**.015**	**6.08**	**1273**	**<.001**
pc4_maxAv_2grams **(4th grade)**	**.067**	**.018**	**3.79**	**2300**	**<.001**
pc4_maxAv_2grams **(5th grade)**	**.057**	**.015**	**3.94**	**1082**	**<.001**
pc4_maxAv_2grams **(6th grade)**	**.063**	**.014**	**4.42**	**1006**	**<.001**
pc4_maxAv_2grams **(adults)**	**.039**	**.015**	**2.67**	**1149**	**.008**
pc5_maxAv_3grams **(3rd grade)**	**.070**	**.009**	**7.76**	**1490**	**<.001**
pc5_maxAv_3grams **(4th grade)**	**.043**	**.010**	**4.24**	**2597**	**<.001**
pc5_maxAv_3grams **(5th grade)**	**.047**	**.008**	**5.62**	**1171**	**<.001**
pc5_maxAv_3grams **(6th grade)**	**.035**	**.008**	**4.27**	**1077**	**<.001**
pc5_maxAv_3grams **(adults)**	**.019**	**.008**	**2.21**	**1218**	**.03**

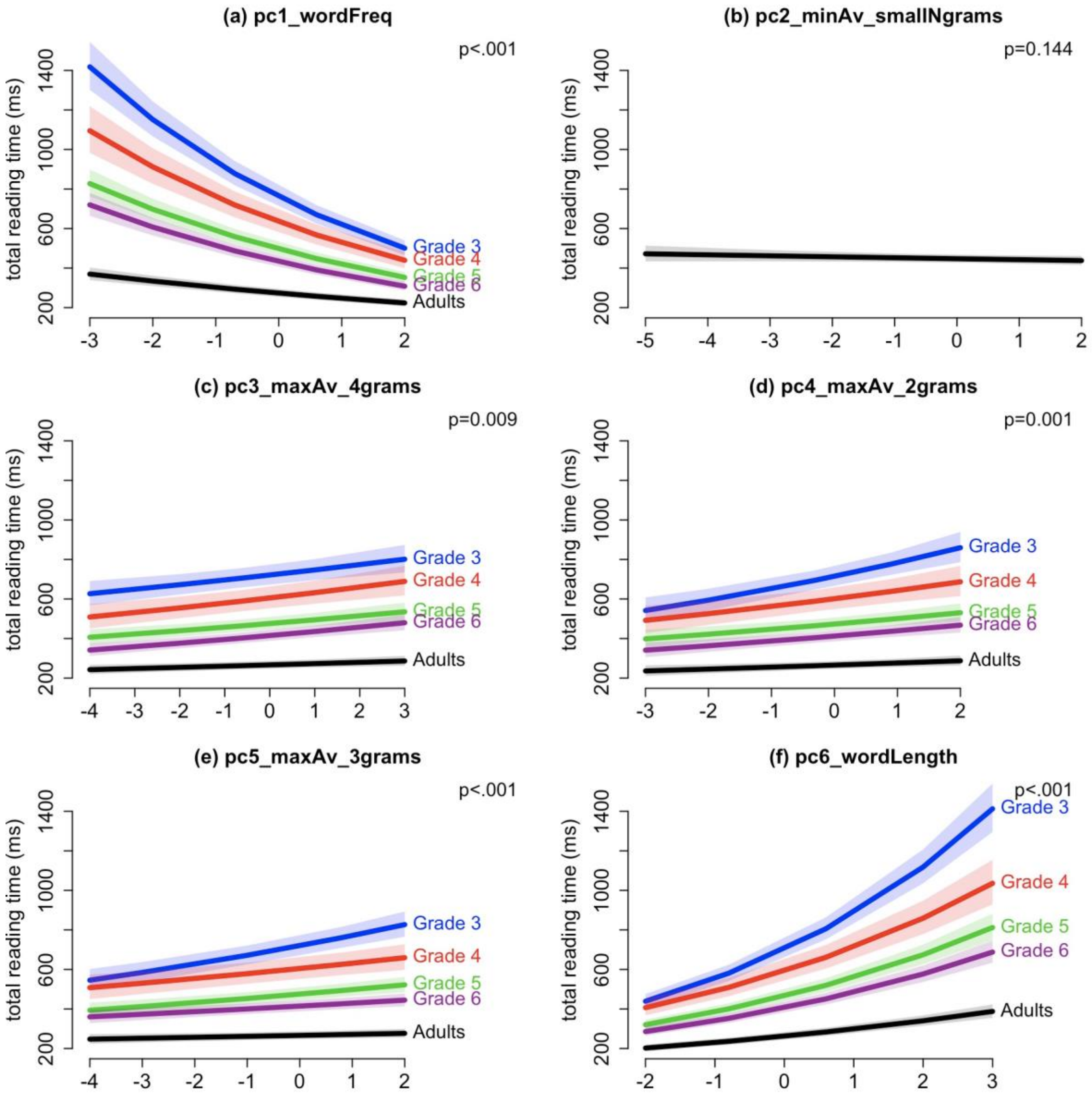


Figure 7 Effects of the six PCs on total reading time. Color panels are those where an interaction with Grade was significant; different colors refer to different grades as reported in the legend. The effect of *PC2-minAv-smallNgrams*, reported in panel (b), was not significant. The *p*-value of the effect is reported in each panel.

Overall, it seems that effects of n-gram frequency are also found consistently later in processing, and that they are distinguishable and independent from word-level effects. Again, maximal n-gram frequency seems to play a more important role than minimal n-gram frequency, and it correlates positively with total reading time: higher maximal frequency leads to longer looking times. Furthermore, sensitivity to all three significant n-gram components (*PC3-maxAv-4grams*, *PC4-maxAv-2grams* and *PC5-maxAv-3grams*) appears to shrink as grade increases.

Discussion

The theoretical question that we addressed with this work was centered on (i) the relationship between statistical learning and reading, more generally, and (ii) the emergence of sensitivity to the statistical properties of written language along reading development, more specifically. While the relevance of regularities between different processing levels (e.g., orthography and phonology, Ziegler & Goswami, 2005; or orthography and meaning,) has been widely explored, it is still not clear whether and how the extraction of visual regularities within the orthographic level benefits reading (see, e.g., Lelonkiewicz et al., 2020; Schmalz & Mulatti, 2017). In spite of general agreement (see Chetail, 2015, for a review) on the fact that orthographic regularities (i.e., clusters of frequently co-occurring letters) are extracted by readers and that sensitivity to them increases through print exposure, the role that such regularities play in visual word processing, and in reading development in particular, is still unclear. Indeed, evidence remains mixed as to whether the effects of bigram frequency, the most commonly analysed metric of orthographic regularity, are facilitatory or inhibitory on word processing times (e.g., Owsowitz, 1963, in Biederman, 1966; Gernsbacher, 1984; Schmalz & Mulatti, 2017). Here, for the first time we addressed these issues : (i) in natural reading (eye movements); (ii) with an unselected sample of realistic text (children stories); (iii) with a large-scale characterization of regularities in letter statistics (through n-gram frequency, on a wide sample of n-grams of different size and a wide sample of words); and (iv) with a large cohort of children participants (N=141, from grade 3 to grade 6) and a control group of adults (N=33).

Overall, our data yielded seven fundamental findings: (i) first and foremost, n-gram frequency affects looking times independently of word-level variables such as frequency and length; (ii) this is generally true already in the youngest cohort of participants, in third grade; (iii) with growing age/grade, the effect of n-gram frequency shrinks; (iv) these developmental effects occur in late, but not early measures of lexical processing (i.e., in gaze duration and total looking time, but not in first-of-many fixation duration); (v) n-gram size does not seem to influence the strength of the n-gram effects, although trigrams seem to yield particularly solid results; (vi) maximal and/or average n-gram frequency across a word seem to work better than minimal n-gram frequency; (vii) contrary to word frequency, higher n-gram frequency implies longer looking times.

The core result of the present paper is that children show sensitivity to the statistics with which letters and graphemes go together in words, independently of other well established predictors of eye movement during reading, like word frequency and word length (and reading skills, of course). This fundamental result sits well with earlier evidence showing that children as young as 6 years old prefer novel words whose spelling conforms to the general characteristics of words in their language (Cassar & Treiman, 1997; Pacton et al., 2001). Crucially, we extend this evidence to a large-scale operationalization of these characteristics in terms of n-gram frequency and, just as important, to the natural reading of real words. This suggests that reading and learning to read do indeed benefit from the coding of regularities with which letters (and/or graphemes) form words—there is a link between reading and learning to read on the one hand, and statistical learning on the other.

How does this sit with the partial evidence that comes from other studies with children (Cassar & Treiman, 1997; Pacton et al., 2001), and from studies on sensitivity to n-gram frequency (particularly bigram) in adults (Chetail, 2015; Schmalz et al., 2017; Schmalz & Mulatti, 2017)? There are methodological considerations that might explain inconsistencies on this front. Several children studies used a correlational approach, reasoning that if statistical learning plays a role in learning to read, children who are good at statistical learning should also become more proficient readers (Arciuli & Simpson, 2012; Kidd & Arciuli, 2016; von Koss Torkildsen et al., 2019). Results were mixed (West et al., 2018), which might be due to the relatively poor reliability of some statistical learning tasks (e.g., Siegelman, Bogaerts and Frost, 2017), or perhaps to the fact that statistical learning at the service of reading and visual word identification captures a specific type of information – regularities in letter co-occurrence – that does not necessarily correlate with the kind of information one learns in the typical statistical learning tasks (e.g., the transition probability between non-linguistic visual objects in Visual Statistical Learning; see Siegelman et al., 2017). If children (and adults) are independently sensitive to how letters go together in real linguistic materials vs. ad-hoc, lab-specific, non-meaningful visual systems, it comes as no surprise that the correlation between performance in classic statistical learning tasks and reading skills is shaky, even if a connection between statistical learning and reading becomes apparent when the relationship is studied entirely within the domain of reading itself, i.e., with real words and with sensitivity to the statistics of co-occurrence between real letters.

The finding that statistical learning might contribute to visual word identification and reading is very relevant for general theories of literacy acquisition, in particular those that focus specifically on the construction of an orthographic lexicon that allows a quicker and more efficient word identification process. For example, the lexical tuning hypothesis suggests that orthographic reading is attained through a progressive sharpening of lexical representations, so that stronger lexical competitors (that is, more similar words) get distinguished progressively better along reading development (Castles et al., 2007). The present data offer a potential account of *how* this happens, that is, via sensitivity to letter co-occurrence statistics, which may identify potential units in the visual input. These candidates for higher-level representations will then be vetted for their functional role in the phonological system (e.g., multi-letter graphemes, like *ou*, or *isl*) or in the semantic system (e.g., morphemes, like *ing*, or *ness*, and words); or perhaps, they might more simply remain statistically cohesive (e.g., frequent) chunks from a merely orthographic point of view, with no particular correspondence with the rest of the language system (e.g., *str*). Of course, more experience would lead to a more precise estimation of the statistics of letter chunks, including words. This will in turn lead to the construction of more and more solid, and more and more precise lexical representations—this account instantiates one possible form of lexical tuning. This account based on statistical regularities in letter co-occurrence does not give any particularly privileged status to words; in fact, any frequent enough letter chunks might develop its own representation. Nicely, this matches with a growing body of evidence, particularly in the adults literature, suggesting that the lexical system also includes non-lexical representations, like affixes (e.g., Taft & Kougious, 2004), morphemes more in general (e.g., Crepaldi et al., 2010; Xu & Taft, 2014), or even letter chunks, like the present account would suggest (e.g., Rastle et al., 2004; Grainger and Ziegler, 2011).

Interestingly, the n-gram effects seem strong and solid already in the youngest cohort of participants in the present study (third graders). This aspect of the results is again in line with the only other investigation of children's sensitivity to letter regularities, although with nonce words and in an artificial, lab-based task. In fact, Pacton et al. (2001) reported that even first graders, whose experience with written language is quite limited, preferred novel words that conformed better with French orthography (e.g., contained higher-frequency doublet consonants). The letter statistics that we considered here are more sophisticated, not much in their quality (we also refer to the frequency of letter chunks), but surely in their quantity (we considered a wide distribution of n-grams included in a realistic text, rather than a few, carefully selected stimuli); and yet, we also find sensitivity to these statistics as early in literacy acquisition as we were able to test.

Where does this sensitivity come from, in such young and quite inexperienced readers? Of course, one possibility is that the visual system is very quick in picking up statistical regularities in letter co-occurrence, so that even a very limited experience is enough to generate statistical learning effects. This seems quite plausible in the case of Pacton et al. (2001), where letter regularities were operationalised as a handful of simple frequency contrasts (e.g., high-frequency consonant doublets vs. low-frequency consonant doublets) or general rules (e.g., vowels never duplicate in French). The present data would attest for a much more powerful mechanism, which does not only capture individual chunks, or orthographic rules in a language, but keeps track of an entire frequency distribution among a large amount of n-grams and words, and does so quite effectively from a relatively early stage (third grade, in this case) with surprising accuracy. One important note of caution is in order, however. Italian has a very transparent mapping between orthography and phonology, and therefore the statistics of co-occurrence between letters (or graphemes) largely overlap to the statistics of co-occurrence between their corresponding phonemes. Thus, we cannot exclude that at least part of the current pattern of sensitivity to statistical information was based on knowledge that was accumulated in the phonological domain, in which children have much more expertise (see, e.g., Ehri, 2005). There are surely elements that would speak against a strong role of phonological regularities here. For example, reading aloud was not required in the present paradigm; although there is evidence that phonology is computed even when words are read silently (see, e.g., Alario et al., 2007; for a review, see Clifton, 2015), the emphasis on understanding and the lack of open articulation should have reduced the contribution of phonological information. Also, early-stage readers are known to make heavy use of phonological recoding (e.g., Share, 1995), but this is mostly a strategic, explicit cognitive mechanism (e.g., Castles et al., 2018), while sensitivity to statistical regularities in the environment is typically implicit and independent of strategic control (e.g., Chetail, 2015; Frost et al., 2019). Nonetheless, further data from a language where orthographic and phonological regularities are decoupled thanks to more opaque grapheme-to-phoneme mappings would certainly be a welcome addition in this respect, and would attest more clearly for the genuine orthographic nature of these effects.

Pacton et al. (2001) also found evidence for a developmental pattern whereby sensitivity to statistics increased with age, at least in some of their results. We also find some developmental trends in the n-gram frequency effects, but they seem to go in the opposite direction: the effects shrink with growing age/grade, instead of strengthening. This might seem surprising, because more experience with a written language should intuitively bring better knowledge of its statistics, and therefore an increased sensitivity to this factor. However, we highlighted above the fact that our data (and Pacton et al.'s) seem to suggest a very early sensitivity to letter co-occurrence statistics (perhaps helped, in the present study at least, by having the same statistics in the phonological domain). Therefore, the developmental changes that we see here might not be related to a better knowledge of the letter statistics, but to other dynamics that characterise reading development. For example, it is known that readers make heavier use of whole-word processing as they become more proficient (Ehri, 2005; Nation, 2009; Perfetti & Hart, 2002); this might reduce the importance of letter processing, therefore also reducing the importance of their statistics of co-occurrence.

Some caution is required here, however, by the fact that also the effects of whole-word frequency (*PC1*) and word length (*PC6*) seem to shrink with age; that is, this developmental pattern is not specific for n-gram frequency. This might suggest that there are other factors in our data driving the statistical reduction of the effects across development, factors that would not be specifically tied to sensitivity to n-grams and their frequency (for example, a mere reduction in the variance of the dependent variables with growing grade/age/proficiency, which might artificially reduce the amount of explained variance in higher grades for *all* variables). A potential explanation is provided by the *linguistic-proficiency hypothesis* outlined by Reichle et al. (2013), according to which, as a reader becomes more skilled, his/her increased proficiency will also result in increasingly adult-like eye-movement patterns. This possibility warrants further consideration.

It is interesting to note that we were only able to see significant developmental trends on relatively late eye-tracking metrics; *FoM* fixation durations were modulated by n-gram frequency (maximal/average trigram frequency, more precisely), but the impact of this variable does not seem to change during the course of reading acquisition. This is another novel insight that stems from these data, which was enabled by the use of eye tracking. Methodologically, this result nicely confirms that the information collected during *FoM* fixations, and the processing thereof, does not entirely overlap with what is tracked by later measures, like gaze duration. From a more theoretical point of view, this suggests that the early visual uptake of information is fully matured by the time developing readers are in third grade, and does not really change substantially as reading proficiency improves. This resonates with much previous work on the maturation of the reading/visual system (e.g., Blythe et al., 2009), and suggests that the very early stages of word/letter processing during natural reading are mostly visual in nature, and perhaps not specific to reading (or at least not entirely specific, to the point that they do not change with changes in reading development).

With this work, we discovered three fundamental properties of sensitivity to n-grams in reading and its development. First, we seem to collect information about n-grams of different sizes, without one specific dominant “spatial resolution”, or “grain size”, to make a more explicit connection with a notable theory of learning to read (Ziegler & Goswami, 2005). Trigrams seem to be more important at the early stages of processing (their maximal/average frequency is the only significant predictor of *FoM* fixation duration), but they are quite on par with bigrams and quadrigrams in later, more fully lexical metrics (gaze duration in particular). Second, maximal/average frequency affects eye movements more strongly than minimal frequency. Finally, words with higher n-gram frequency are fixated for *longer* periods of time, rather than yielding shorter fixations, as it is the case for word frequency. Collectively, these properties allow us to characterize n-gram processing –or, more generally, the way we use statistical information on letter co-occurrence during reading– with much more precision. First, the process seems to be unselective as to what kind of units it would use—it takes statistical information whatever it comes from, and is able to consider different n-gram sizes at the same time. Trigrams seem to be particularly useful early on (i.e., during first fixations); with a bit of speculation, this might be related to the fact that chunks of three letters seem to be a potentially useful stepping stone from individual letters to words, and this might be particularly relevant when information uptake must be maximised, i.e., when readers perform a quick first fixation on a new word, in the view of having further fixations more centrally within that word.

Perhaps more difficult to understand is why maximal, rather than minimal, n-gram frequency within a word emerges as a particularly salient cue. Information is inversely proportional to frequency; if you know that a word contains the highly frequent trigram *peg*, you are left with many alternatives, but if you know that a word contains a rare trigram like *ynx*, you can be pretty sure that the word is *sphynx*. So, this result might sound particularly counter-intuitive; readers should focus more on statistics that promise to deliver more information. One possible account is related to the fact that we measured eye movements here, and therefore the system is not only involved in letter processing, word identification and understanding, but also in deciding where to fixate next, for how long, and what the best possible landing position might be. So, even if maximally frequent n-grams provide a lesser cue to word identity, they might be good processing locations, and therefore the eye movement control system might be specifically tuned to find them. Of course, a more detailed study of how information is distributed within words would be necessary here (see, for example, Alhama et al., 2019, or Underwood et al., 1990), but there are also cognitive reasons that might suggest such a fixation strategy; for example, it might be convenient to have higher-frequency material in the fovea, so that more resources are free to compute parafoveal information (e.g., Kennedy & Pynte, 2005; Veldre & Andrews, 2018).

This account for why maximal frequency is more important than minimal frequency might also help explaining the somewhat surprising direction of the effect—words with higher maximal n-gram frequency attract longer, rather than shorter fixations. If readers are particularly interested in maximal n-gram frequency because this allows them to collect more information from their fixations, then it makes sense that they stay longer on words whose maximal n-gram frequency is higher—there will likely be more information to collect. It is important to note, however, that the literature includes trigram effects that go in either direction, facilitatory (e.g., Hand et al., 2012) or inhibitory (e.g., Lima and Inhoff, 1985). In her review of letter statistic effects, Chetail (2015) does note this inconsistency, and proposes an account to reconcile this contrasting evidence. The account is based on different stages of processing; n-grams would have a facilitatory effect during the early steps of visual word recognition, e.g., orthographic processing, and inhibitory effects on later stages, where lexical competition kicks in. Such an account might also apply to our data, of course, but it would be difficult to reconcile with the fact that we did not find different results between early eye-tracking metrics, such as *FoM* fixation duration, and later ones, like gaze duration.

For what concerns general models of reading and learning to read (e.g., the lexical tuning hypothesis by Castles et al., 2007; the DRC model by Coltheart et al., 1993, 2001; the dual-route approach by Grainger & Ziegler, 2011; the EZ reader model by Reichle, 2011; Reichle et al. 1998; the OB1 reading model by Snell et al., 2018; the psycholinguistic grain size theory by Ziegler & Goswami, 2005), they generally make little explicit connection with statistical learning, and therefore it is difficult to see how the present data would confirm, challenge or qualify them further. Also, this work is only a first step towards a more specific, mechanistic account of how sensitivity to letter statistics connects to reading and reading acquisition; thus, it is probably too early to compare this experimental evidence to the very detailed dynamics that are implemented in some of those models. Most generally, however, these data are obviously compliant with the emphasis on statistical regularities that is typical of connectionist models of reading (e.g., Plaut et al., 1996; Seidenberg & McClelland, 1989; Harm & Seidenberg, 1999, 2004), which, interestingly, was also the theoretical framework in which Pacton et al. (2001) interpreted their results. In the last decade, several other models were proposed, which do not explicitly put themselves in the line of early connectionism, but still emphasise the important role played by statistics in the correspondence between different linguistic domains (form and meaning in particular; e.g., Baayen et al., 2011; Marelli and Baroni, 2015; Marelli et al., 2017). Frost (2012) took up the issue that most models of reading and most experimental evidence in this domain is based on a tiny subset of languages in the world (e.g., English, Dutch, German, Italian, Hebrew), and argues that a more general universal model of reading must be sought for. Critically for the present work, he indicates sensitivity to statistical regularities within the reading system as a potentially unifying principle that might support the construction of such a linguistically universal model. In addition to explicit, mechanistic models of reading and visual word identification, there are also novel experimental effects that emerged in recent years and that are entirely built on the idea that the brain is sensitive to statistical regularities in the connection between form and meaning (e.g., Orthography-to-Semantic Consistency, OSC; Marelli et al., 2015; Marelli and Amenta, 2018; Amenta et al., 2020).

From this perspective, the novelty of the present work is that it focuses entirely on orthography, while the work described above focused specifically on ties *between different* *domains* (e.g., phonology and orthography, form and meaning). The data illustrated here (and in some other recent work; e.g., Chetail, 2017; Lelonkiewicz et al., 2020) suggest that not only statistical cues are relevant to establish ties between units of different nature (e.g., phonemes and graphemes), but even the construction of the relevant representations within one given domain (here, orthography) might be dominated by sensitivity to statistical regularities. This yields even more generality to the statistical learning approach to reading and learning to read.

A final note is in order to specify that, although some models of visual word identification have explicitly proposed that n-grams (bigrams, in particular) are critical processing units in the system (e.g., Dehaene et al., 2005; Whitney, 2001), we would not take a strong stance in this respect. The main reason is that n-gram frequency was assumed in this work only as one possible operationalisation of letter co-occurrence statistics—one among many, which seemed particularly convenient in the context of reading development and visual word identification (for example, because this variable has attracted some attention in this domain). It was outside the scope of the paper to specifically contrast this metric with other potential statistical cues, like, e.g., transitional probability, which has generated comparatively more research in the statistical learning domain (e.g., Bogaerts et al., 2016; Fiser & Aslin, 2001, 2005; Kirkham et al., 2002; Saffran et al., 1996). So, we take these results as surely compatible with accounts of reading and literacy acquisition that stipulate a specific role for n-gram representations. However, we do not think that these data are particularly constraining to those models; in fact, the fundamental theoretical message of the present work, we believe, is that sensitivity to letter statistics is an important player in reading acquisition. We started to draft the specific characteristics of the cognitive mechanisms that might lie behind this general tenet, but this is only the beginning of, we hope, a much longer journey.

References

Alario, F. X., De Cara, B., & Ziegler, J. C. (2007). Automatic activation of phonology in silent reading is parallel: Evidence from beginning and skilled readers. *Journal of Experimental Child Psychology*, *97*(3), 205-219. https://doi.org/10.1016/j.jecp.2007.02.001

Alhama, R. G., Siegelman, N., Frost, R., & Armstrong, B. C. (2019). The role of information in visual word recognition: A perceptually-constrained connectionist account. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)* (pp. 83-89). Austin, TX: Cognitive Science Society.

Amenta, S., Crepaldi, D., & Marelli, M. (2020). Consistency measures individuate dissociating semantic modulations in priming paradigms: A new look on semantics in the processing of (complex) words. *Quarterly Journal of Experimental Psychology, 73*(10), 1546-1563. https://doi.org/10.1177/1747021820927663

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 234. https://doi.org/10.1037/0278-7393.18.2.234

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286-304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: past, present and future. https://doi.org/10.1098/rstb.2016.0047

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(1-2), e1373. https://doi.org/10.1002/wcs.1373

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12-28. https://doi.org/10.21500/20112084.807

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438. https://doi.org/10.1037/a0023851

Biederman, G. B. (1966). Supplementary report: The recognition of tachistoscopically presented five-letter words as a function of digram frequency. *Journal of Verbal Learning and Verbal Behaviour, 5*(2), 208-209. https://doi.org/10.1016/S0022-5371(66)80020-8

Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic Bulletin & Review*, *23*(4), 1250-1256. https://doi.org/10.3758/s13423-015-0996-z

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letters in words. *Journal of Educational Psychology*, *89*(4), 631. https://doi.org/10.1037/0022-0663.89.4.631

Castles, A., Davis, C., Cavalot, P., & Forster, K. (2007). Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology*, *97*(3), 165-182. https://doi.org/10.1016/j.jecp.2007.01.006

Castles, A., & Nation, K. (2006). How does orthographic learning happen. *From Inkmarks to Ideas: Current Issues in Lexical Processing*, *151*. https://doi.org/10.4324/9780203841211

Castles, A., & Nation, K. (2008). Learning to be a good orthographic reader. *Journal of Research in Reading*, *31*(1), 1-7. https://doi.org/10.1111/j.1467-9817.2007.00367.x

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Chetail, F., Balota, D., Treiman, R., & Content, A. (2014). What can megastudies tell us about the orthographic structure of English words? *Quarterly Journal of Experimental Psychology*, *68*(6), 1519–1540. https://doi.org/10.1080/17470218.2014.963628

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology*, *6*, 645. https://doi.org/10.3389/fpsyg.2015.00645

Chetail, F. (2017). What do we do with what we learn? Statistical learning of orthographic regularities impacts written word processing. *Cognition*, *163*, 103-120. https://doi.org/10.1016/j.cognition.2017.02.015

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468-481. https://doi.org/10.1111/tops.12332

Clifton, C. (2015). The roles of phonology in silent reading: a selective review. In Frazier, L., & Gibson, E. (Eds.). *Explicit and implicit prosody in sentence processing: Studies in honor of Janet Dean Fodor* (Vol. 46), p. 161–176. Springer. https://doi.org/10.1007/978-3-319-12961-7_9

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological review, 100*(4), 589. https://doi.org/10.1037/0033-295X.100.4.589

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204. https://doi.org/10.1037/0033-295X.108.1.204

Conrad, M., Carreiras, M., Tamm, S., & Jacobs, A. M. (2009). Syllables and bigrams: orthographic redundancy and syllabic units affect visual word recognition at different processing levels. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 461. https://doi.org/10.1037/a0013480

Crepaldi, D., Amenta, S., Mandera, P., Keuleers, E., & Brysbaert, M. (2016). Frequency estimates from different registers explain different aspects of visual word recognition. *International Meeting of the Psychonomic Society,* Granada, Spain, 5–8 May. http://crr.ugent.be/subtlex-it/

Crepaldi, D., Rastle, K., Coltheart, M., & Nickels, L. (2010). 'Fell'primes 'fall', but does 'bell'prime 'ball'? Masked priming with irregularly-inflected primes. *Journal of Memory and Language*, *63*(1), 83-99. https://doi.org/10.1016/j.jml.2010.03.002

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Sciences*, *9*(7), 335-341. https://doi.org/10.1016/j.tics.2005.05.004

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, *9*(2), 167-188. https://doi.org/10.1207/s1532799xssr0902_4

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499-504. https://doi.org/10.1111%2F1467-9280.00392

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458. https://doi.org/10.1111%2F1467-9280.00392

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, *134*(4), 521. https://doi.org/10.1037/0096-3445.134.4.521

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, *8*(15), 1-27. https://www.jstatsoft.org/article/view/v008i15

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd Edition. Thousand Oaks, CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html

Frost, R. (2012). Towards a universal model of reading. *The Behavioral and Brain Sciences*, *35*(5), 263. https://doi.org/10.1017/S0140525X11001841

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128. https://doi.org/10.1037/bul0000210

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256. https://doi.org/10.1037/0096-3445.113.2.256

Grainger, J., & Van Heuven, W. J. B. (2004). *Modeling Letter Position Coding in Printed Word Perception.* In P. Bonin (Ed.), *Mental lexicon: "Some words to talk about words"* (p. 1–23). Nova Science Publishers.

Grainger, J., & Ziegler, J. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology, 2*, 54. https://doi.org/10.3389/fpsyg.2011.00054

Grice, J. W. (2001). A comparison of factor scores under conditions of factor obliquity. *Psychological Methods, 6*(1), 67–83. https://doi.org/10.1037/1082-989X.6.1.67

Hand, C. J., O'Donnell, P. J., & Sereno, S. C. (2012). Word-initial letters influence fixation durations during fluent reading. *Frontiers in Psychology*, *3*, 85. https://doi.org/10.3389/fpsyg.2012.00085

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review, 106*(3), 491. https://doi.org/10.1037/0033-295X.106.3.491

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662. https://doi.org/10.1037/0033-295X.111.3.662

Inhoff, A. W. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior*, *23*(5), 612-624. https://doi.org/10.1016/S0022-5371(84)90382-7

Johnston, J. C. (1978). A test of the sophisticated guessing theory of word perception. *Cognitive Psychology*, *10*(2), 123-153. https://doi.org/10.1016/0010-0285(78)90011-7

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329. https://doi.org/10.1037/0033-295X.87.4.329

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153-168. https://doi.org/10.1016/j.visres.2004.07.037

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behaviour Research Methods*, *44*(1), 287-304. https://doi.org/10.3758/s13428-011-0118-4

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184-193. https://doi.org/10.1111/cdev.12461

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: is it long-term and implicit?. *Neuroscience Letters*, *461*(2), 145-149. https://doi.org/10.1016/j.neulet.2009.06.030

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35-B42. https://doi.org/10.1016/j.cognition.2011.06.010

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and language, 115*, 104152. https://doi.org/10.1016/j.jml.2020.104152

Lima, S. D., & Inhoff, A. W. (1985). Lexical access during eye fixations in reading: Effects of word-initial letter sequence. *Journal of Experimental Psychology: Human Perception and Performance, 11*(3), 272. https://doi.org/10.1037/0096-1523.11.3.272

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods, 50*(4), 1482-1495. https://doi.org/10.3758/s13428-018-1017-8

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography-Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology, 68*(8), 1571-1583. https://doi.org/10.1080%2F17470218.2014.959709

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*(3), 485. https://doi.apa.org/doi/10.1037/a0039267

Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as Abstract Operation in Semantic Space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition, 166*, 207-224. https://doi.org/10.1016/j.cognition.2017.05.026

Massaro, D. W., Jastrzembski, J. E., & Lucas, P. A. (1981). Frequency, orthographic regularity, and lexical status in letter and word perception. In *Psychology of Learning and Motivation* (Vol. 15, pp. 163-200). Academic Press. https://doi.org/10.1016/S0079-7421(08)60175-9

Morucci, P., Bottini, R., & Crepaldi, D. (2019). Augmented modality exclusivity norms for concrete and abstract Italian property words. *Journal of Cognition*, *2*(1). http://doi.org/10.5334/joc.88

Nation, K. (2009). Form–meaning links in the development of visual word recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3665-3674. https://doi.org/10.1098/rstb.2009.0119

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*(3), 447-461. https://doi.org/10.1017/langcog.2016.20

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745-2750. https://doi.org/10.1073/pnas.0708424105

Owsowitz, S. E. (1963). *The effects of word familiarity and letter structure familiarity on the perception of words*. Santa Monica, CA: Rand Corporation Publications.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401. https://doi.org/10.1037/0096-3445.130.3.401

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56. http://dx.doi.org/10.1037/0033-295X.103.1.56

Quémart, P., Casalis, S., & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of Experimental Child Psychology, 109*(4), 478-496. https://doi.org/10.1016/j.jecp.2011.02.008

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review, 11*(6), 1090-1098. https://doi.org/10.3758/BF03196742

Rencher, A. C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician*, *46*(3), 217-225. https://doi.org/10.1080/00031305.1992.10475889

Revelle, W. (2020) psych: Procedures for Personality and Psychological Research. Northwestern University. Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych

Rice, G. A., & Robinson, D. O. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition*, *3*(5), 513-518. https://doi.org/10.3758/BF03197523

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928. https://doi.org/10.1126/science.274.5294.1926

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon, 12*(2), 263-282. https://doi.org/10.1075/ml.17009.sch

Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia, 67*(2), 147-162. https://doi.org/10.1007/s11881-016-0136-0

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523. http://dx.doi.org/10.1037/0033-295X.96.4.523

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Taft, M. & Kougious, P. (2004). The processing of morpheme-like units in monomorphemic words. *Brain and Language*, *90*, 9-16. https://doi.org/10.1016/S0093-934X(03)00415-2

Underwood, G., Clews, S., & Everatt, J. (1990). How do readers know where to look next? Local information distributions influence eye fixations. *Quarterly Journal of Experimental Psychology, 42A*, 39-65. http://doi.org/10.1080/14640749008401207

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Veldre, A., & Andrews, S. (2018). How does foveal processing difficulty affect parafoveal processing during reading?. *Journal of Memory and Language*, *103*, 74-90. https://doi.org/10.1016/j.jml.2018.08.001

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143-156. https://doi.org/10.1177%2F0956797619881134

Von Koss Torkildsen, J., Arciuli, J., & Wie, O. B. (2019). Individual differences in statistical learning predict children's reading ability in a semi-transparent orthography. *Learning and Individual Differences*, *69,* 60-68. https://doi.org/10.1016/j.lindif.2018.11.003

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2018). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science, 21*, Article e12552. https://doi.org/10.1111/desc.12552

Westbury, C., & Buchanan, L. (2002). The probability of the least likely non-length-controlled bigram affects lexical decision reaction times. *Brain and Language*, *81*(1-3), 66-78. https://doi.org/10.1006/brln.2001.2507

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review*, *8*(2), 221-243. https://doi.org/10.3758/BF03196158

Xu, J., & Taft, M. (2014). Solely soles: Inter-lemma competition in inflected word recognition. *Journal of Memory & Language*, 76, 127-140. https://doi.org/10.1016/j.jml.2014.06.008

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. https://doi.org/10.1037/0033-2909.131.1.3

Chapter V

Automatic Morpheme Identification Across Development: Magnetoencephalography (MEG) Evidence from Fast Periodic Visual Stimulation

Morphemes are the smallest linguistic units that bear meaning. For instance, a complex word like *artist* contains a stem, *art-*, and a suffix, *-ist*. Many languages are morphologically rich, meaning that that their lexicon includes a great deal of complex words, by derivation, inflection or compounding; it is estimated that 85% of the English lexicon is made up of complex words (Algeo & Algeo, 1993; Grainger & Ziegler, 2011).

In light of the role that morphological processing plays in skilled reading (Rastle, 2019), it is unsurprising that several studies in the psycholinguistic domain have focused on the sensitivity to morphological structure in visual word processing (for a review see Amenta & Crepaldi, 2012). Several theories have been proposed over the years to account for the visual identification, comprehension and reading aloud of complex words. Some of these theories dispose entirely of explicit morphological representations, and trace back the emergence of morphological effects to the appreciation of statistical regularities in the mapping between form, meaning and phonology (e.g., Baayen et al., 2011; Seidenberg, 1987). Other models affirm the existence of morphological representations, either through different, serially-arranged stages of processing (e.g., Crepaldi et al., 2010; Taft & Nguyen, 2010; Taft, 2015) or along parallel routes of processing (e.g., Grainger & Ziegler, 2011). These “localist” models of morphology build in different ways on the distinction between a level of morphological processing that is mostly based on form and one in which meaning plays a more substantial role.

Indeed, behavioural evidence has widely shown that adult readers' sensitivity to the morphological structure of words is such that even pseudocomplex words, i.e., words containing non-morphological orthographic units that overlap with existing and productive morphemes, hold a special status in visual word processing (e.g., Dawson et al., 2018; Diependaele et al., 2011; Kazanina et al., 2008; Longtin et al., 2003; Marelli et al., 2013; Rastle et al., 2004). Masked priming evidence with adults show that not only words such as *reader* prime their stem *read*, but also pseudo-morphological words such as *corner*, prime their pseudostems *corn* (as compared to a purely orthographic baseline, e.g., *brothel-BROTH*).

More recently, Grainger and Beyersmann (2017, 2021) proposed a novel view, whereby the analysis of the internal structure of words is initiated by the identification of stems as embedded, edge-aligned words. This would be a bootstrapping mechanism exploited for initiating morpho-orthographic processing, as we will discuss later in this section.

While an extensive body of research has appreciated the role of morphemes as reading units, it is far less clear at which point of reading development the ability to recognize morphemes fully matures (e.g., Beyersmann et al., 2012; Dawson et al., 2018; Quémart et al., 2011). Available evidence from behavioural studies is quite mixed in this respect. Sensitivity to morphological structure has been reported to emerge as early as seven years of age. For example, in a study by Carlisle and Stone (2005) a group of children from Grades 2 and 3 and a group from Grades 5 and 6 were administered a reading aloud task. Both groups showed faster reading times for derived words (e.g., *hilly*) than for "pseudoderived" ones, matched for number of syllables and frequency (e.g., *silly*). Similarly, Kirby et al. (2012) found that within the first 2-3 years of primary school children already display explicit morphological knowledge. Priming studies have corroborated such evidence, consistently revealing significant priming for morphologically related pairs (e.g., *golden-GOLD* in English, *kleidchen-KLEID* in German) in primary school children, as early as Grade 2 (Hasenäcker et al., 2016) or Grade 3 (Beyersmann et al., 2012).

While morpho-semantic processing appears to be present at an early developmental stage, findings are much more inconclusive when it comes to the establishing the trajectory of morpho-orthographic processing, which has been shown to regularly occur in skilled adult readers (e.g., Dawson et al., 2018; Longtin et al., 2003; Rastle et al., 2004). A few studies conducted in French (Casalis et al., 2015; Quémart et al., 2011) and Italian (Burani et al., 2002; Burani et al., 2008) have provided evidence for sensitivity to nonwords with a morphological structure in primary school children. Of course, children might try to assign some meaning to these nonwords, which are typically somewhat interpretable (e.g., *mammista*, lit. "motherist", Burani et al., 2002), particularly by individuals with a still incomplete lexicon, who might just think they have come across a novel word. In this sense, such evidence is not unequivocal support for a morpho-orthographic analysis of letter strings similar to what allows adults to see *corn* in *corner*. However, it does suggest that children can identify morphemes in unfamiliar letter strings, that is, they can access morphology pre-lexically and independently of the meaning of the letter strings where they are embedded (which, in this case, does not exist entirely).

In a cross-sectional lexical decision study investigating the emergence of sensitivity to morphemes within nonwords along development, Dawson et al. (2018) reported lower accuracy in rejecting complex nonwords (such as *earist*) compared to control nonwords (e.g., *earilt*) across all age groups that they tested (adults, older adolescents, young adolescents, children); however, only the two older age groups displayed slower reaction times to complex nonwords compared to control nonwords. This supports the conclusion that, while morphological sensitivity is already present in younger readers, it fully matures quite late in adolescence (16-17 years old).

Another host of studies investigated the emergence of the corner-corn effect itself, using the classic masked priming design with transparent (*dealer-DEAL*), opaque (*corner-CORN*) and orthographic (*dialog-DIAL*) prime-target pairs, providing quite mixed evidence about morpho-orthographic processing in developing readers. For instance, in a study in English by Beyersmann et al. (2012), children in Grades 3 and 5 showed priming only for transparent pairs, such as *golden-GOLD*, but not for opaque (e.g., *mother-MOTH*) or orthographic ones (e.g., *spinach-SPIN*). A different pattern of results emerged in a French study with third, fifth and seventh graders by Quémart et al. (2011). This experiment yielded similar effects for opaque (*baguette-BAGUE*) and transparent pairs (*tablette-TABLE*), but no priming for orthographic (*abricot-ABRI*) or semantic (*tulipe-FLEUR*) pairs. Yet another pattern emerged in Schiff et al. (2012), where both fourth and seventh graders showed strong priming when prime and target were morphologically and semantically related, and seventh graders showed also a weak priming effect for pairs that were morphologically related and semantically unrelated, displaying a pattern similar to that observed with adult readers of Hebrew in other studies (Bentin & Feldman, 1990; Frost et al., 1997).

An account of the developmental trajectory of morpho-orthographic processing has been recently proposed by Grainger and Beyersmann (2017, 2021). According to this model, beginning readers rely on previously existing representations of stems (which are often encountered as free-standing words and therefore have a more readily available visual representation) as a bootstrapping mechanism for initiating morpho-orthographic segmentation. Following consistent exposure to printed complex words, orthographic affix representations in complex words would occur later in development. Finally, the formation of affix representations in pseudocomplex words (i.e., complete morpho-orthographic processing) would only occur in the final developmental stage. At what point in development this processing stage is completed is still an open question. However, stem activation is not modulated by the presence of an affix *per se*, suggesting no specific role of morphological context in the recognition of stems, even when processing complex nonwords, such as *farmity* (e.g., Beyersmann et al., 2015). This theoretical suggestion is based on a number of studies where the classic masked priming design was extended to nonword primes (e.g., *dealness-deal* vs. *dealnuss-deal*; see, e.g., Beyersmann et al., 2015; Hasenäcker et al., 2016; Longtin & Meunier, 2005). With these stimuli, there is typically no difference between morphologically structured primes (*dealness*) and primes with an existing stem but without a suffix (*dealnuss*), contrary to the typical word prime pattern whereby *corner* yields facilitation, but *dialog* does not. These data led Grainger and Beyersmann (2017) to suggest a primary role for embedded, edge-aligned words, rather than morphological structure per se. This is nicely in line with recent experiments using a semantic task and showing that children access even the meaning of embedded stems/words (e.g., *crow* in *crown*) independently of morphology (e.g., *corn* in a pseudosuffixed word like *corner* as well as *pea* in a nonsuffixed word like *peace*; Hasenäcker et al., 2021; for similar evidence with adults, see Hasenäcker et al., 2020).

The mixed evidence described above might be due, at least in part, to issues related to commonly used behavioural paradigms, often requiring children to sit through long sessions and perform a somewhat unnatural task (e.g., primed or unprimed lexical decision), and usually yielding quite small effects. To overcome these limitations, we adopted a relatively novel, behaviour-free technique, called Fast Periodic Visual Stimulation (FPVS), paired with an oddball design, in a magnetoencephalography (MEG) study with both developing and skilled readers.

This paradigm has so far mostly been coupled with EEG recordings (see, e.g., Lochy et al., 2015, 2016; Rossion, 2014; Quek et al., 2018), and has proven very powerful in tapping into automatic visual processing through just a few minutes of passive stimulation at a fast periodic rate. Frequency-tagging paired with an oddball design allows to identify selective neural responses to stimuli with specific features. For example, stimuli may be delivered at a rate of 6 items per second (carrier frequency = 6 Hz), with an oddball stimulus inserted periodically every fifth item (6 Hz/5 = oddball frequency of 1.2 Hz) differing along a certain dimension from the stream of stimuli in which it is embedded (see Figure 1). A robust peak in the EEG (or MEG) signal at the oddball frequency indexes successful discrimination of the oddball stimuli from the base stream, suggesting the existence of a neural representation for the category exemplified by them (e.g., words in nonwords; Lochy et al., 2015, 2016) or neural sensitivity to the dimension that distinguishes between oddball and base stimuli (e.g., frequency of occurrence within the visual stream; De Rosa, Ktori et al., 2021). This paradigm has been largely employed to detect selective responses to rapidly presented faces (e.g., Rossion, 2014; Dzhelyova & Rossion, 2014; Rossion et al., 2015; Quek et al., 2018). It has also been successfully applied to psycholinguistic research, allowing to identify neural discrimination responses to words in adults (Lochy et al., 2015) and even to letter strings in preschoolers (Lochy et al., 2016), in the left occipito-temporal cortex.

Electrophysiological and neuroimaging techniques have allowed, over the last decades, to explore the neural bases of morpheme identification (see Leminen et al., 2019, for an extensive review). Thanks to EEG, MEG, and fMRI studies, we have gained insight into the neural underpinnings of visual identification of complex words, and more generally into morphological processing in the brain.

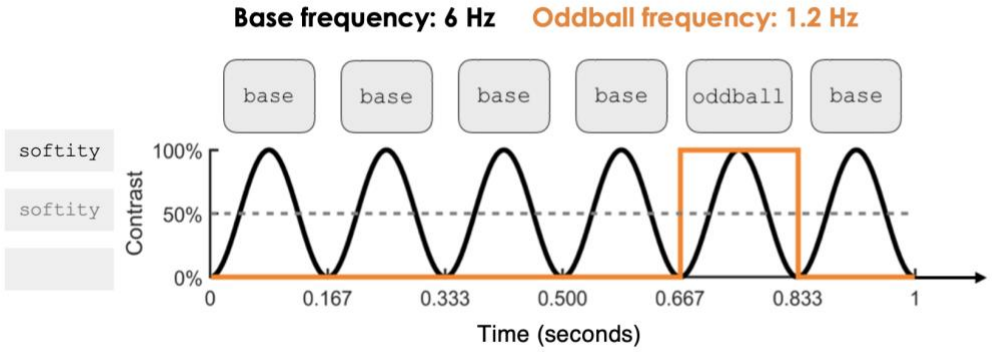


Figure 1 Schematic illustration of the Fast Periodic Visual Stimulation (FPVS) oddball paradigm. For a gradual and smooth transition between them, stimuli were presented via sinusoidal contrast modulation. In each stimulation sequence, stimuli were presented at 6 Hz (base frequency) for 60 seconds, with oddball stimuli appearing every fifth item (oddball frequency: 6/5 = 1.2 Hz). Participants engaged in an orthogonal task monitoring the colour change of a centrally presented fixation cross.

For instance, different hypotheses have been proposed regarding the processing of regular vs. irregular inflected forms (e.g., *walked* vs. *ran*). Marslen-Wilson and Tyler (1998, 2007), based on fMRI lesion studies, argued for a bihemispheric dual mechanism. On the one hand, regularly inflected forms would be processed through rule-based combinatorial processing, entailing decomposition; these processes appear to be left-lateralised, with an activation of the left inferior frontal gyrus (LIFG). A similar pattern has been found in an fMRI priming study with healthy adult participants on derivational morphology by Bozic et al. (2007), supporting the hypothesis that similar decomposition mechanisms underlie the processing of derivationally complex forms and regularly inflected ones. On the other hand, irregular inflection would yield a more broadly bilateral activation (taken to indicate access to lexical and semantic information), as shown by intact access to irregular forms by patients with left hemispheric lesion (see Marslen-Wilson & Tyler, 2007, for an overview).

Turning to derivational morphology, in an auditory lexical decision ERP study, Leminen et al. (2010) presented Finnish speaking adults with existing derived words, novel derived words, and illegal derivations. They found that both real words and legal novel derivations yielded a comparable N400-like response, suggesting successful interpretation and integration of morphemes in both cases. Illegal novel derivations, instead, elicited a larger magnitude of the N400, which was taken to reflect difficulties in the semantic integration of the constituent morphemes. These findings were overall interpreted as suggestive of simultaneous morphemic parsing and whole-word semantic access, in the auditory processing of derived stimuli.

MEG studies have attempted to address whether complex (or pseudocomplex) words yield an obligatory early decomposition stage, as established by behavioural evidence (e.g., Longtin et al., 2003; Rastle et al., 2004), or whether later semantic effects also intervene. Zweig and Pylkkänen (2009), in a lexical decision study, found that adult English speakers displayed a clearer M170 – an MEG component indicative of early decomposition processes – when presented with real derived words (either suffixed, e.g., *farmer*, or prefixed, e.g., *refill*), as opposed to opaque ones (e.g., *winter*) or morphologically simple ones (e.g., *switch*). The emergence of the M170 component was observed both in left and right occipito-temporal regions, suggesting a bilateral contribution to early stages of visual word identification. In an MEG visual lexical decision study on English suffixed words, Fruchter and Marantz (2015) found, in left temporal regions, an earlier facilitatory effect on reaction times of derivational family entropy around 240 ms, indexing decomposition processes, and a later facilitatory effect of surface frequency, around 430-450 ms, interpreted as a later recombination stage (e.g., Taft, 2004). Furthermore, a facilitation in left orbitofrontal activity around 350-500 ms was observed as an effect of semantic coherence, suggesting access to a semantic analysis of the morphemes in order to assess well-formedness.

Whiting et al. (2015) conducted a masked priming MEG study to investigate differences in the processing of simple (*walk*), complex (*farmer*), and pseudocomplex (*corner*) words. For both complex and pseudocomplex items, a similar morphological effect emerged around 330-

340 ms in the left middle temporal gyrus (MTG), diverging instead from noncomplex stimuli. A similar set of results was obtained with inflected stimuli, with real and pseudoinflected forms eliciting a similar effect around 300-370 ms in left posterior MTG and LIFG, diverging from noncomplex forms. This pattern of findings suggests that (pseudo)complex items undergo a blind decomposition process (i.e., morpho-orthographic processing), in line with behavioural accounts from masked priming.

Overall, while accounts of morphological processing in the brain are far from being homogenous, neural evidence appears to converge on an early involvement of left temporal and occipital regions, in line with an early stage of morphological analysis mostly based on form (e.g., Crepaldi et al., 2010; Taft & Nguyen-Hoan, 2010) and on converging research on the identification of visual words more generally (e.g., Dehaene et al., 2005). This is further corroborated by fMRI evidence, such as a masked priming study by Gold and Rastle (2007), in which a similar pattern of reduced activation was observed in the left posterior middle occipital gyrus for pseudomorphologically related pairs (*archer-ARCH*) and for orthographically related ones (*pulpit-PULP*), and reduced activity of the posterior face fusiform gyrus was observed specifically for pseudomorphologically related pairs.

However, all this evidence is exclusively based on adults, and fails to address the neural bases of the development of morphological sensitivity. In fact, to the best of our knowledge, there is no neuroimaging investigation of how children make use of morphology in visual word identification in their pathway towards reading proficiency.

Furthermore, all neural evidence regarding morphological processing has so far entailed highly refined and artificial experimental tasks, such as primed or unprimed lexical decision, or violation paradigms, which, as mentioned above, entail some difficulties and limitations, especially with children.

Therefore, we turned to the FPVS-oddball task described above, and we used it in conjunction with MEG recordings, to gain insight on the existence and developmental trajectories of neural representations of morphemes. In addition to its sensitivity and ease of use with children, the FPVS-oddball task is particularly suited to tag automatic visual processes, which brings two additional advantages. First, it gives little room for strategic processes to contaminate the more specific mechanisms involved in morpheme identification (similarly to masked priming). Second, it is likely to capture particularly those early stages of morphological analysis where evidence, especially in behavioural experiments, is mixed, leaving the developmental trajectory of this processing stage quite unclear.

The contrast between base and oddball stimuli was manipulated to probe selective stem and suffix identification in morphologically structured pseudowords, that is, in the absence of lexical processing. The oddball stimuli consisted of 4 types of items. The first two were aimed at assessing stem identification, either in nonwords that also contain a suffix (e.g., *softity*; Condition 1) or in nonwords that do not (e.g., *softert*, Condition 2). Symmetrically, condition 3 and 4 were focused on suffix identification, in strings that also contain a stem (e.g., *softity*) or in strings that do not (e.g., *terpity*). The base stimuli in each condition were constructed so as to lack the critical morpheme. In Condition 1, the stem+suffix oddballs (*softity*) were embedded in streams of nonstem+suffix base stimuli (*terpity*), so that the contrast between the two tracks stem identification, in the presence of a suffix. In Condition 2, the stem+nonsuffix oddballs (*softert*) were embedded in streams of nonstem+nonsuffix oddballs (*terpert*), so that the contrast tracks stem identification again, but this time in the absence of a suffix. Similarly, Condition 3 and 4 featured stem+suffix oddballs within stem+nonsuffix bases (*softity* in softert) and nonstem+suffix oddballs within nonstem+nonsuffix bases (*terpity* in *terpert*), thus tracking suffix identification. Adults were administered all four conditions, while children only underwent Condition 1 and Condition 3, in consideration of their shorter time in the MEG.

Materials & Methods

Participants

We recruited 32 skilled adult readers (age range: 18-45) and 21 developing readers (enrolled in Years 5 and 6 at the time of testing). Data from four adults and four children were eventually removed from the final sample analysed here, either for excessive head motion (greater than 5 mm for adults; greater than 11 mm for children) or due to an excessive presence of artefacts. This left us with 28 skilled adult readers (age: mean 22.93 years, sd 6.38 years) and 17 developing readers (age: mean 10.59 years, sd 0.79).

Adult participants were recruited through the Macquarie University SONA system and were offered course credit, where applicable, or monetary compensation. Children were recruited through a dedicated portal, called *Neuronauts*, and their families were awarded monetary compensation for their time. Both studies were approved by the Macquarie University Human Research Ethics Committee.

All participants were native English speakers and right-handed; none reported neurological problems, developmental issues, language difficulties, or claustrophobia. They all had normal or corrected-to-normal (through contact lenses) vision.

Stimuli

All conditions consisted of five 60-second trials, for adults, and of six 60-second trials, for children. A within-participant block design was adopted. A non-experimental, baseline condition (Condition 0) was administered to all participants. Adults were administered four experimental conditions, while children were only administered two of these (Conditions 1 and 3).

In condition 0, 4-letter words (oddball stimuli) were embedded in non-pronounceable 4-consonant strings (base stimuli). This condition was solely administered to ensure that the paradigm worked correctly; for this type of stimuli, indeed, there should be a solid discrimination response (see, e.g., Lochy et al., 2015, 2016). In condition 1, oddball stimuli were nonwords made up of a real stem and a real suffix (e.g., *softity*), which were embedded in nonwords made up of a nonstem and a real suffix (e.g., *trumess*). In condition 2, nonwords made up of a real stem and a nonsuffix (e.g., *softert*) were used as oddballs and embedded in nonwords made up of a nonstem and a nonsuffix (e.g., trumust). In condition 3, oddball nonwords were made up of a real stem and a real suffix (e.g., *softity*) and were embedded in nonwords made up of a real stem and a nonsuffix (e.g., *stopust*). Lastly, in condition 4, nonwords made up of a nonstem and a nonsuffix (e.g., *terpity*) were embedded in nonwords made up of a nonstem and a suffix (e.g., *trumust*). Sequence examples for each condition are reported in Table 1.

Contrasts in each condition were set in order to tap into stem or suffix identification. Specifically, a significant oddball response in condition 1 (and 2, in the adult sample only) would index stem identification, and in condition 3 (and 4, in the adult sample only) it would index suffix identification. The administration of two additional conditions (2 and 4) to the adult participants was intended to shed light on the role of context for the identification of morphemes – that is, whether a robust response to oddballs was present only when they could be fully broken down into the two constituents morphemes (Conditions 1 & 3), or whether morphemes would also be successfully identified when oddballs featured only one morphemic constituent (Conditions 2 & 4).

Table 1 Example stimuli delivered in a 1-second stimulation cycle in the four experimental conditions and in the non-experimental condition 0. Oddball stimuli (illustrated in italic) appeared every fifth item.

	base	**base**	**base**	**base**	***oddball***	**base**
Condition 1 *stem+suffix* in nonstem+suffix	trumess	joskive	molpory	firnure	*softity*	berfise
Condition 2 *stem+nonsuffix* in nonstem+nonsuffix	trumust	joskune	molpute	firnint	*softert*	berfere
Condition 3 *stem+suffix* in stem+nonsuffix	stopust	helpune	parkute	lastint	*softity*	townere
Condition 4 *nonstem+suffix* in nonstem+nonsuffix	trumust	joskune	molpute	firnint	*terpity*	berfere
Condition 0 *words* in nonwords	kltq	rdsc	fgnl	pdrk	*roll*	tmkj

In the adult version of the experiment, stimuli were composed of 12 items for each type: stems, nonstems, suffixes, nonsuffixes. Nonstems and nonsuffixes were created from the set of existing stems and existing suffixes, while keeping the same length, Consonant-Vowel structure, and minimising orthographic overlap with existing words (e.g., *terp* was created as a nonstem from *soft*, *ert* was created as a nonsuffix from *ity*). Stem and nonstems were 4 letters in length, while suffixes and nonsuffixes were 3-letter long. The 12 nonsuffixes were non-morphemic endings attested in English. Each set of (non)stems and (non)suffixes was divided in two subsets of 6 items; stimuli were then obtained by combining each element in one subset with each element of another. This procedure generated 72 unique combinations (6 items in the first set, times 6 items in the second set, times 2 subsets) of each type (stem+suffix, nonstem+suffix, stem+nonsuffix, nonstem+nonsuffix), yielding a total of 288 unique stimuli.

In the developmental version of the experiment, the building blocks were reduced to 6, a subset of those used for skilled adult readers. (Non)stems and (non)suffixes were combined by groups of 3, to obtain 18 (3*3*2) unique combinations of each type (stem+suffix, nonstem+suffix, stem+nonsuffix), yielding a total of 54 unique stimuli.

All building blocks (stems, nonstems, suffixed and nonsuffixes) are reported in Table 2. Statistics for stems and suffixes were obtained from two different linguistic databases: SUBTLEX-UK (Van Heuven et al., 2014) and MorphoLex (Sánchez-Gutiérrez et al., 2018). Specifically, while the former frequency database is particularly relevant for its size (over 160,000 types and 200 million tokens from English television show subtitles), the latter resource is a rich morphologically tagged database for English, allowing the extraction of metrics related to the use of items as morphemes in the language.

Stem selection

All selected stems are four-character long and have a CVCC or CCVC consonant-vowel structure. Here, we describe the features of the 12 stems used as constituents in the adult version of the experiment, a subset of which was used in the version with developing readers; the statistics related to the six stems used in the version for children are provided in square brackets. Database exploration, extraction and calculation of relevant metrics were performed using R (R Core Team, 2021).

The average SUBTLEX-UK log Zipf frequency was 5.13, with a sd of 0.43 [mean: 4.93, sd: 0.28]; the average stem token frequency in MorphoLex was 155217, with a sd of 144128.60 [mean: 132510, sd: 127152.30], while the average stem family size in MorphoLex was 15.50, with a sd of 9.64 [mean: 22.83, sd: 8.33]. Finally, the average Levenshtein distance of the 20 nearest orthographic neighbours (old20) was calculated through the R package *vwr* (version 0.3.0; Keuleers, 2013), based on the data contained in SUBTLEX-UK, the largest resource considered here: all stems had a mean old20 of 1 and a sd of 0 [mean: 1, sd: 0].

Nonstem selection

Nonstems were nonwords generated with the same length and CV structure types as the real stems, in order for the items to be orthographically and phonotactically legal, while at the same time minimizing orthographic overlap with the selected stems. The mean old20 for our nonstem selection was 1.14, with a sd of 0.21 [mean: 1.07, sd: 0.16].

Suffix selection

Twelve three-letter derivational suffixes were shortlisted from the CELEX database (Baayen et al., 1993). The CV structure types of the selected suffixes were VCC, VCV, CVC, VVC. A subset of six suffixes was used for the developmental version of the experiment. The same exploration and analysis were performed as for the above-described stem selection. The average SUBTLEX-UK log Zipf frequency was 2.41, with a sd of 0.59 [mean: 2.41, sd: 0.73].

We ensured, through MorphoLex, that all selected items were productive suffixes in the English language. The average suffix token frequency in MorphoLex was 514914.40, with a sd of 452230.50 [mean: 643484.20, sd: 523213.40], while the average suffix family size in MorphoLex was 319.25, with a sd of 226.44 [mean: 431.50, sd: 145.89]. All suffixes had a mean old20 of 1 and a sd of 0 [mean: 1, sd: 0].

Nonsuffix selection

We selected 12 three-letter clusters that occur as non-morphological endings in English, with a mean old20 of 1 and a sd of 0 [mean: 1, sd: 0].

Control condition stimuli

For the control condition, we selected 72 4-letter words (with various CV structure types, but always ending with a consonant) and 72 4-letter non-pronounceable consonant strings. A subset of 18 words and 18 consonant strings was used for the experiment with children. The average SUBTLEX-UK log Zipf frequency was 4.71, with a sd of 0.54 [mean: 4.91, sd: 0.54].

Stimuli combinations

Statistics for the stimuli used in the developmental version of the experiment, which did not feature nonstem+nonsuffix combinations, are reported in brackets. Old20 statistics were then computed for all stimuli. Stem+suffix combinations had a mean old20 of 2.32 and a sd of 0.30 [mean: 2.43, sd: 0.27], stem+nonsuffix combinations had a mean old20 of 2.49 and a sd of 0.37 [mean: 2.55, sd: 0.43], nonstem+suffix combinations had a mean old20 of 2.47 and a sd of 0.32 [mean: 2.47, sd: 0.32], and nonstem+nonsuffix combinations had a mean old20 of 2.62 and a sd of 0.31. All unique experimental stimuli can be found in Appendix B (for the adult version of the study) and in Appendix C (for the child version of the study).

Table 2 Unique stems, nonstems, suffixes and nonsuffixes combined to generate pseudoword stimuli, and the respective OLD20 statistics. The underlined items represent the subset of items used in the developmental version of the experiment.

Stem	**Non Stem**	**Suffix**	**Non Suffix**	**old20 Stem**	**old20 NStem**	**old20 Suffix**	**old20 NSuffix**
help	josk	ity	ert	1.00	1.55	1.00	1.00
soft	terp	ive	une	1.00	1.00	1.00	1.00
last	firn	ory	ute	1.00	1.00	1.00	1.00
ship	bron	ure	int	1.00	1.00	1.00	1.00
stop	trum	ous	ald	1.00	1.00	1.00	1.00
hold	burk	ise	ere	1.00	1.00	1.00	1.00
park	molp	ful	sal	1.00	1.40	1.00	1.00
jump	lort	ist	arn	1.00	1.00	1.00	1.00
town	bemp	ite	ene	1.00	1.40	1.00	1.00
bird	jelt	ish	ult	1.00	1.00	1.00	1.00
farm	culp	ese	oke	1.00	1.35	1.00	1.00
milk	tand	ess	ust	1.00	1.00	1.00	1.00

Apparatus

Data were collected at the KIT-Macquarie Brain Research Laboratory (Sydney, Australia). Participants lay supine in a dimly lit and magnetically shielded room (MSR). Continuous MEG recordings were acquired using a 160-channel whole-head coaxial gradiometer system (KIT, Kanazawa Institute of Technology, Japan) at a sampling rate of 1000 Hz, with an online bandpass filter of 0.03–200 Hz. Visual stimuli were delivered through a projector (sampling rate: 60 Hz) and mirrored onto a translucent screen mounted above the participant's head, at a distance of approximately 110 cm. The experiment was controlled via a Windows desktop computer, using MATLAB (The Mathworks) and Psychtoolbox (Brainard, 1997; Kleiner et al., 2007). Parallel port triggers were used to mark the beginning and end of each trial, and a photodiode was used to check the correct delivery of oddball stimuli, through a white square in the bottom right corner of the screen.

Participants' head shape was recorded using the Polhemus FASTRAK system and digitizing pen (Colchester, VT, USA). Throughout the MEG recording session, participants wore an elastic cap with five marker coils which allowed tracking the head location relative to the MEG helmet and to measure motion over time.

Procedure

Each trial comprised a 60-second stimulation sequence (as illustrated in Figure 1), in which 360 stimuli rapidly appeared and disappeared at 6 Hz (six stimuli per second), with a contrast modulated by a sinusoidal function—each individual stimulus appeared gradually, reaching a contrast peak after 83.5 ms. Each 60-second trial thus contained 360 stimuli overall. Each oddball item appeared after five base stimuli (6 Hz/5=1.2 Hz); therefore, the stimulation sequence in each trial included 72 oddballs and 288 base items. The oddball stimuli were unique items in the adult design, whereas in the developmental design a greater number of item repetitions was present: in each trial, every oddball was delivered a total of 4 times (18*4=72). The sets of stimuli were generated through pseudo-randomisation, using in-house RStudio scripts (RStudio Team, 2021) for the adult version of the experiment, and using Mix software (Van Casteren & Davis, 2006) for the developmental version. As the process could not be entirely automatised, lists were then checked and edited manually when deemed necessary, in order to prevent repetitions of the same combinations within each stimulation sequence. Both with skilled and developing readers, we ensured that the same stimulus was not repeated within each 1-second stimulation sequence.

Overlayed to this stimulus sequence, a fixation cross (12 pixels) was constantly present at the center of the screen. The cross changed in colour (from blue to red and vice versa) randomly and participants were instructed to tap a button when they detected the colour change (Lochy et al., 2015, 2016).

In the experiment with skilled readers, visual stimuli were displayed in black Courier New font, with a fontsize of 100 px, within a white bounding box of 500*150 pixels. In the developmental version of the experiment, stimuli were slightly enlarged, and they were displayed in black Courier New bold font, with a fontsize of 110 pt, within a white bounding box of 510*170 pixels. A large font size was adopted for both skilled and developing readers due to their distance from the screen. In both versions of the experiment, stimuli were displayed over a grey background.

Responses were recorded through a fiber optic button box (fORP, Current Designs, Philadelphia, PA, USA). Accuracy in this task was very high for all participants (skilled adult readers: mean 97.83%, sd 1.84; developing readers: mean 95.64%, sd 4.84). This behavioural task was administered with the mere purpose of ensuring that participants engaged with the area in which the stimuli would be presented. Trials were separated by a 25-second break. The break ended with a 10-second countdown to the new trial. A 2-minute break was given twice between recording blocks, to allow head location measurements to be performed; one last measurement was performed at the end of the MEG recording. Overall, the MEG testing in the MSR required 45-50 minutes with adults and a maximum of 30 minutes with children.

MEG data preprocessing

Data were preprocessed in MATLAB using the FieldTrip toolbox for EEG/MEG-analysis (Oostenveld et al., 2011) and in-house functions. A lowpass filter of 100 Hz was applied; continuous MEG recordings were epoched into trials using a custom-made trial function. In trial epoching, a pre-stimulus interval and a post-stimulus interval were set, in order to avoid edge artifacts. Respectively, the first two oddball cycles (i.e., the first 1.67 seconds of stimulation) and the last one (833 ms) were cut from each trial, resulting in trials of 58.33 seconds each (see Lochy et al., 2015). Recordings were then downsampled to 250 Hz. Artifacts were not removed from individual trials; however, data from eight subjects (four adults and four children) with excessive noise artifacts (one adult) or excessive movement artifacts (three adults and four children) were discarded entirely.

Following visual inspection, noisy channels were removed based on visual inspection, and channel interpolation was performed (neighbours were defined using FieldTrip functions through a triangulation method). One dataset per condition (five trials per condition for adults, six trials per condition for children) per participant was obtained.

Frequency analysis

A very similar procedure to the one used in Lochy et al. (2015, 2016) was adopted. Each participants' trials were averaged by condition and subjected to a Fast Fourier Transform. By calculating the square root of the sum of squares of the real and imaginary parts divided by the number of data points, power spectra were then computed for each sensor. As each epoch was 58.333 seconds long, the frequency resolution was 1/58.333 = 0.0171 Hz. The spectra were then normalised by dividing the mean power spectrum of each frequency bin by the mean of the surrounding 20 bins (10 on either side, excluding immediately adjacent bins), thus obtaining a signal-to-noise ratio metric (SNR). Oddball response was defined as the mean SNR of the response at the oddball (1.2 Hz) stimulation frequency and its corresponding first three harmonics (2.4, 3.6, 4.8 Hz). So, the final dataset consisted of 22400 datapoints for the adult sample (28 participants, times 5 conditions, times 160 channels) and 8160 datapoints for the children sample (17 participants, times 3 conditions, times 160 channels).

Results

Sensor-level analysis

Region-of-interest (ROI) analysis

The first set of results presented here is based on a predefined region of interest (ROI), which reflects our focus on the left ventral occipito-temporal cortex (VOTC). Prior electrophysiological and imaging research has identified left temporal and occipital regions as responsible for visual word identification (for EEG evidence, Lochy et al., 2015, 2016) and for the visual identification of morphemes (Leminen et al., 2019, for a comprehensive review; Gold & Rastle, 2007, for fMRI evidence). Following this, we defined an ROI of 12 left occipito-temporal sensors in our participants' MEG recordings (Figure 2).

In this ROI, we checked whether SNR was significantly above 1, which is the expected value if there is no sensitivity to the oddball, and therefore power at the relevant frequency would be the same as in the surrounding bins.

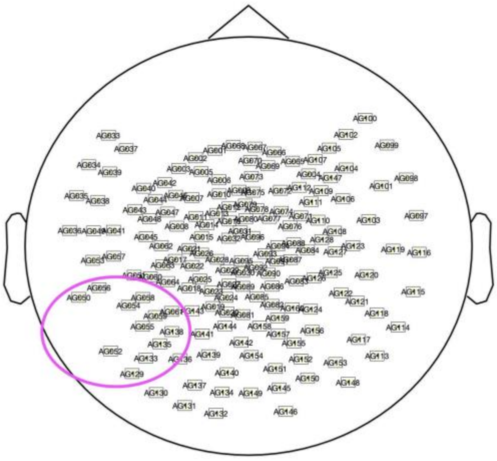


Figure 2 Visualisation of the 12 sensors comprised in the left temporo-occipital scalp ROI, on the 160-channel Yokogawa MEG layout (KIT, Kanazawa Institute, Japan).

In the sample of skilled adult readers, a robust oddball response was observed in Condition 0 (p<.001), which tracks sensitivity to words amongst nonwords. This effect was reported before (Lochy et al., 2015, 2016), and was taken as a sanity check—the paradigm must reveal the obvious sensitivity of skilled readers to existing words if we are to trust its results on morphemes.

The results in the experimental conditions are reported in Table 3 and Figure 3. A significant oddball response emerged only in Condition 3, which tapped into suffix detection in the presence of a stem (e.g., ***softity*** vs. *terpert*; see Table 3).

Similar to the adults, a robust oddball response emerged in Condition 0 also in the sample of developing readers (see Table 4 and Figure 4), which confirms the reliability of the paradigm with children. Also similar to the adults, a significant effect emerged in Condition 3, but not in Condition 1, suggesting that also developing readers are sensitive to the presence of suffixes, while we don't have evidence for sensitivity to the presence of stems.

Table 3 *T*-test results of the sensor ROI-based analysis in skilled adult readers. Mean SNR, average SNR of the oddball response in each condition; *t*, *t*-statistic from the *t*-tests; *df*, degrees of freedom; *p*, one-tailed p value.

Condition (example *oddball* in base)	**Mean SNR**	***t***	***df***	***p***
Condition 0 - word detection (*roll* in kltq)	**1.33**	**5.49**	**27**	**<.001**
Condition 1 – stem detection with suffixes (*softity* in terpity)	1.04	1.23	27	.11
Condition 2 – stem detection without suffixes (*softert* in terpert)	0.99	-0.45	27	.67
Condition 3 – suffix detection with stems (*softity* in terpert)	**1.06**	**1.69**	**27**	**.05**
Condition 4 – suffix detection without stems (*terpity* in terpert)	1.03	0.76	27	.22

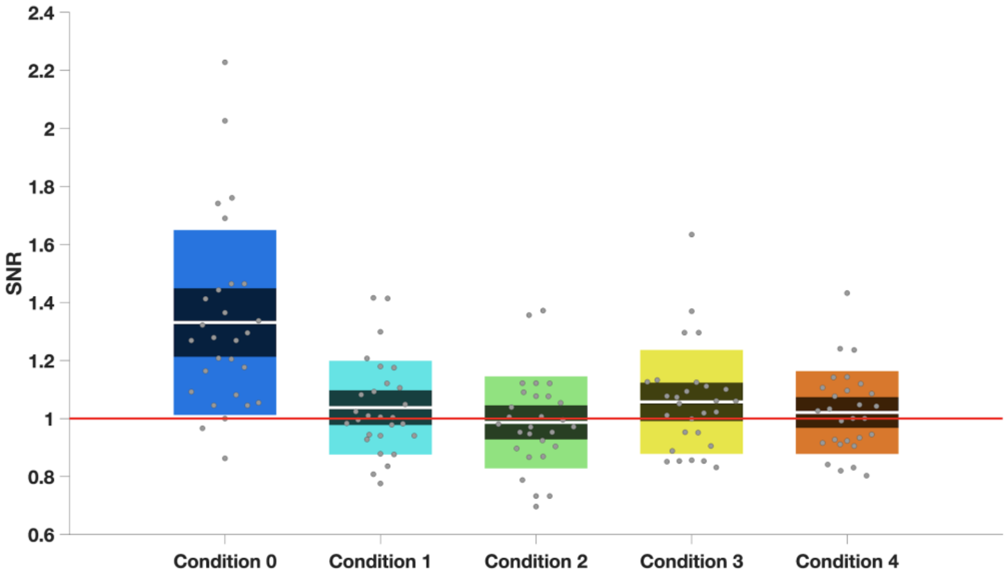


Figure 3 Visualisation of the mean SNR of the oddball response (averaged across 1.2, 2.4, 3.6, 4.8 Hz) in skilled adult readers, by condition. Non-boxplots illustrating mean (white line), standard error of the mean (darker coloured area), and standard deviation (lighter coloured area) per condition (C0–C4). Points represent individual participants (N=28). The red line represents the noise level (1), against which SNR is compared in each condition. The grey dots illustrate individual participants' average SNR of the oddball response in each condition.
The structure of the contrast between oddball (italic, bold) and base stimuli (italic) in each condition is as follows: Condition 0, ***words*** in *nonwords* (e.g., ***roll*** in *kltq*); Condition 1, ***stem+suffix*** in *nonstem+suffix* (***softity*** in *terpity*); Condition 2, ***stem+nonsuffix*** in *nonstem+nonsuffix* (***softert*** in *terpert*); Condition 3, ***stem+suffix*** in *stem+nonsuffix* (***softity*** in *terpert*); Condition 4, ***nonstem+suffix*** in *nonstem + nonsuffix* (***terpity*** in *terpert*).

Table 4 *T*-test results of the sensor ROI-based analysis in developing readers.
Mean SNR, average SNR of the oddball response in each condition; *t*, *t*-statistic from the *t*-tests; *df*, degrees of freedom; *p*, one-tailed p-value.

Condition (example *oddball* in base)	**Mean SNR**	***t***	***df***	***p***
Condition 0 - word detection (*roll* in kltq)	**1.34**	**5.54**	**16**	**<.001**
Condition 1 – stem detection with suffixes (*softity* in terpity)	1.04	0.90	16	.19
Condition 3 – suffix detection with stems (*softity* in terpert)	**1.15**	**2.63**	**16**	**.009**

These results suggest that suffixes undergo automatic morpheme identification in complex pseudowords, while this might not be the case for stems, at least as far as the MEG-FPVS paradigm employed here can tell. Also, suffix identification seems to occur only when full decomposition is possible–when the strings are entirely composed of morphemes, also in the absence of explicit lexical access. We will expand and comment more on this in the Discussion.

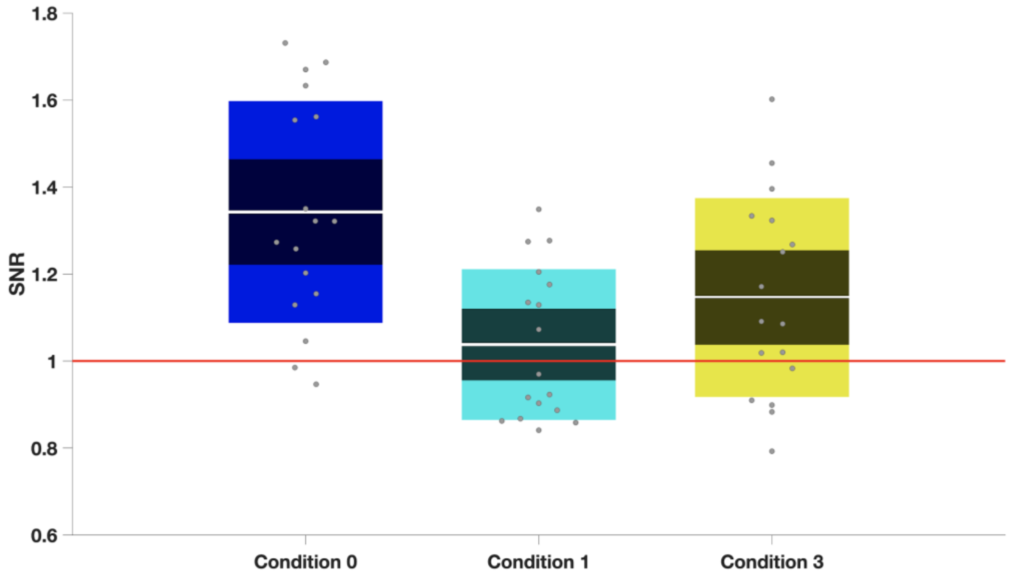


Figure 4 Visualisation of the mean SNR of the oddball response (averaged across 1.2, 2.4, 3.6, 4.8 Hz) in developing readers, by condition. Non-boxplots illustrating mean (white line), standard error of the mean (darker coloured area), and standard deviation (lighter coloured area) per condition (C0, C1, C3). Points represent individual participants (N=17). The red line represents the noise level (1), against which SNR is compared in each condition. The grey dots illustrate individual participants' average SNR of the oddball response in each condition.
The structure of the contrast between oddball (italic, bold) and base stimuli (italic) in each condition is as follows: Condition 0, ***words*** in *nonwords* (e.g., ***roll*** in *kltq*); Condition 1, ***stem+suffix*** in *nonstem+suffix* (***softity*** in *terpity*); Condition 3, ***stem+suffix*** in *stem+nonsuffix* (***softity*** in *terpert*).

Cluster-Based Permutation Analysis

Besides the theory-driven ROI analysis, a data-driven analysis approach was also adopted. Despite we are primarily interested in the visual identification of morphemes and the present paradigm emphasises quick and automatic visual access, a morphological analysis might surely trigger semantic information well outside the early stages of the ventral stream. Therefore, we wanted to assess the existence of any potential tagging of the oddball frequency at the whole-brain level. To this aim, we conducted a cluster-based permutation test at sensor level (Maris & Oostenveld, 2007), adapted for FPVS-MEG datasets, which span over space (sensors), but not time. We used a within-subject design and adopted a Montecarlo method for calculating probabilities. A minimum of two neighbouring channels was required in order for a cluster to be defined. A cluster alpha level of 0.05 was set and a one-tailed *t*-test was run (we only contemplated the hypothesis that the SNR was higher than 1). An alpha level of 0.05 was set and 5000 randomisations were performed. With this configuration, cluster-based permutation was run against an array of ones, representing the noise level in each channel (i.e., the null hypothesis).

The results for the adult skilled readers are illustrated in Figure 5. In Condition 0, which taps into whole-word identification, we found one large cluster essentially encompassing the whole posterior part of the scalp, with a slight left lateralization (*t*=416.46, *p*<.001, panel a). We also found a significant cluster in Condition 3, which yielded significant results in the ROI analysis and probes suffix identification in the presence of a stem (*t*=113.02, *p*<.001, panel b). This cluster is much smaller than in Condition 0, and extends along the midline from the vertex to the back of the brain, and then along the left ventral stream. No other significant clusters emerged, that is, there was no reliable sensitivity to the oddball stimuli in Condition 1 (designed to track stems in the presence of affixes), Condition 2 (stems in the absence of affixes) and Condition 4 (suffixes in the absence of stems). However, one significant cluster appeared in Condition 1 when we slightly relaxed the alpha criterion (0.10) at the cluster level – importantly, this does not increase the false alarm rate (Maris & Oostenveld, 2007). This cluster (*t*=69.88, *p*=.02), illustrated in panel c, is located centrally over the occipital and parietal lobes.

The results for the developing readers are illustrated in Figure 6. For Condition 0, we found one significant cluster (*t*=347.17, *p*<.001, panel a), which is largely left-lateralised and extends over temporo-parieto-occipital sensors, mostly overlapping with the selected ROI, and thus likely reflecting an involvement of VOTC areas in response to the presentation of words. While this would need to be confirmed through source-level analysis, a VOTC location appears plausible and in line with the aforementioned neurophysiological evidence on visual word recognition (Leminen et al., 2019; Gold & Rastle, 2007). A significant occipital cluster, mostly located around the midline, emerged in Condition 1 (*t*=74.90, *p*=.007, panel b), in which stem identification in the presence of suffixes is tracked. In contrast with the ROI analysis results, no significant cluster emerged for Condition 3 (suffixes in the presence of stems). However, following the same procedure as for the adult readers, one large cluster in Condition 3 emerged when we used a relaxed alpha criterion (0.10) at the cluster level. This cluster (*t*=126.11, *p*=.008), illustrated in panel c, extends bilaterally over temporal and occipital areas.

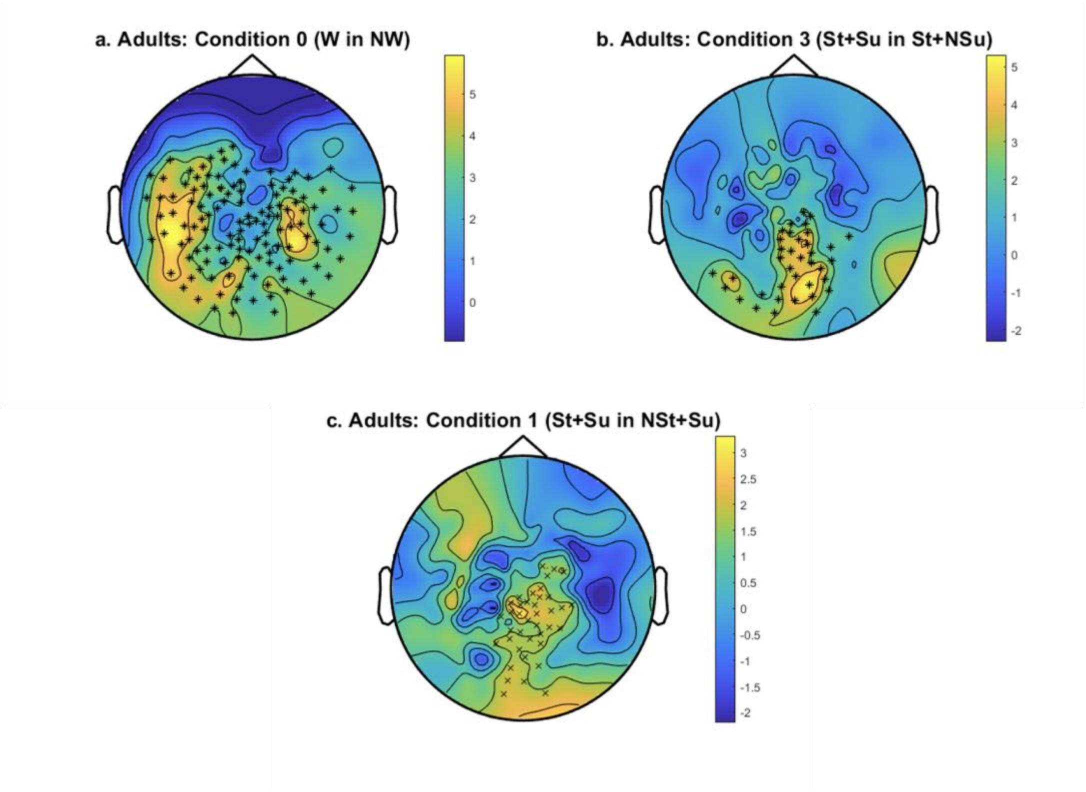


Figure 5 Sensor-level clusters in which a significant oddball response emerged, by condition. Data from skilled adult readers. Panel a: Large temporo-parieto-occipital cluster (mostly left-lateralised, with a right-lateralised part) indicating widespread identification of words in nonwords, in Condition 0; p<.001, cluster alpha level = 0.05. Panel b: Left and central occipital cluster for the identification of stem+suffix oddballs in stem+nonsuffix base stimuli, in Condition 3; p<.001, cluster alpha level = 0.05. Panel c: Central temporo-occipital cluster for the identification of stem+suffix oddballs in nonstem+suffix base stimuli, in Condition 1; p=.02, cluster alpha level = 0.10.

Colourbars represent SNR on a continuous scale (blue = low, yellow = high).

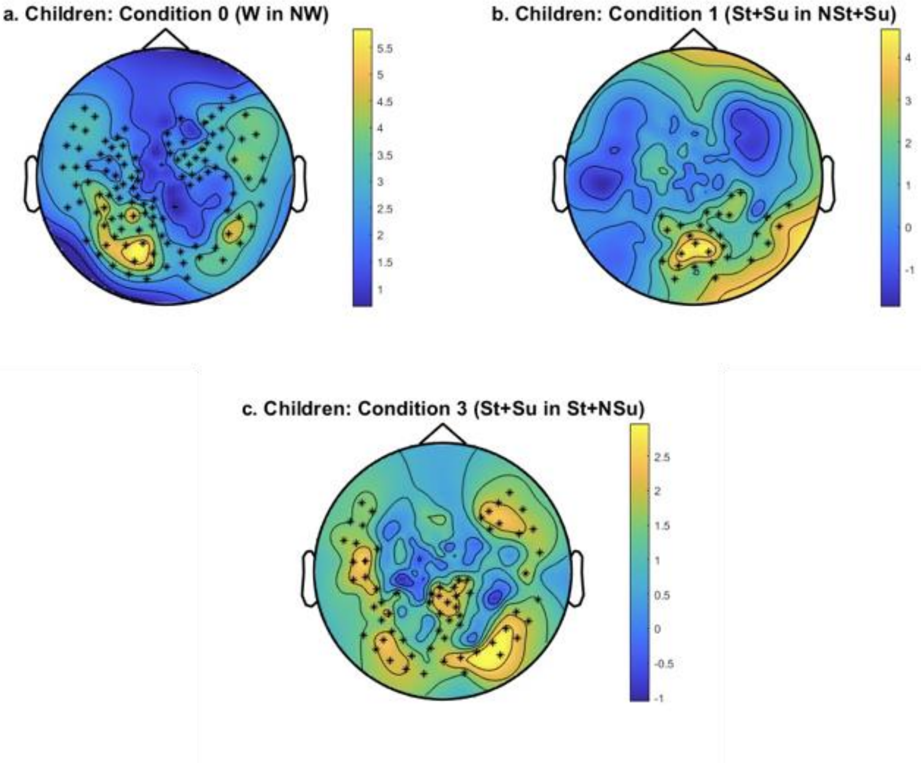


Figure 6 Clusters in which a significant oddball response emerged, by condition. Data from developing readers. Panel a: Temporo-parieto-occipital cluster, largely left-lateralised, for the identification of words in nonwords, in Condition 0; p<.001, cluster alpha level = 0.05. Panel b: Central occipital cluster for the identification of stem+suffix oddballs in nonstem+suffix base stimuli, in Condition 1; p=.007, cluster alpha level = 0.05. Panel c: Central temporo-occipital cluster for the identification of stem+suffix oddballs in stem+nonsuffix base stimuli, in Condition 3; p=.008, cluster alpha level = 0.10.

Colourbars represent SNR on a continuous scale (blue = low, yellow = high).

Discussion

As described in the Results section, our ROI findings reveal successful word identification (Condition 0), as well as successful identification of suffixes when presented in oddballs which could be completely broken down into two morphemic constituents (Condition 3), both in developing and in adult readers, in left occipito-temporal sensors, which reflect our focus on VOTC. Using a cluster-based permutation approach to sensor-level analysis, we found, once again, a significant temporo-parieto-occipital cluster for Condition 0, in both adults and children. This cluster was quite large and emerged bilaterally for the skilled adult readers, and was instead smaller and largely left-lateralised for the developing ones. As far as morpheme identification is concerned, an occipital cluster along the midline and left VOTC emerged for Condition 3 (suffixes in the presence of stems) in the adult participants, while an occipital, mostly centrally located, cluster emerged for condition 1 (stems in the presence of suffixes) in the developing readers.

Condition 0, tracking identification of words in streams of consonant nonwords, importantly revealed that the paradigm worked correctly, confirming that the FPVS technique, which has already been quite extensively paired with EEG recordings, can also successfully be employed in MEG studies investigating visual word identification. Besides, with the exception of Lochy et al.'s (2016) EEG study with preschoolers, to the best of our knowledge, no FPVS studies have been conducted in the psycholinguistic domain with developing readers. It is indeed worth noting that this paradigm offers a number of advantages, which should be considered by researchers seeking to investigate language processing in children. First, it is a behaviour-free technique, which makes it suitable also for the more challenging experimental populations, including very young children; furthermore, the stimulation is rapid and implicit, thus preventing the use of strategies by the participants. Lastly, specifically in the word identification condition, our findings confirm how powerful this paradigm is, as a robust response is elicited with a relatively small number of trials (see also Lochy et al., 2015, 2016) and that this also occurs in developing readers.

Indeed, the ROI-based analysis of Condition 0 showed a very strong response to the presentation of words, suggesting that the selected sensor-level region reflects an involvement of the VOTC. However, the cluster-based permutation analysis revealed the presence of large clusters across the posterior part of the scalp, mostly left-lateralised. This suggests that even automatic and implicit word identification triggers full processing, possibly even including a semantic stage.

Remarkably, the patterns of results observed for word identification were very similar across our developing and adult participants. These findings suggest that 5th and 6th graders, with just a few years of reading instructions, have already built up a highly sophisticated visual word identification system, roughly comparable to that of adults, at least in terms of automatic and implicit word identification

For what concerns the core research question of the present study, whether automatic identification of morphemes emerges in complex pseudowords, the pattern of results that we get is – perhaps unsurprisingly – less clear and robust than for words.

We did find evidence for automatic morpheme identification, but this does not seem to be widespread and solid across different types of stimuli. In particular, the ROI analysis conducted on the sensors covering VOTC revealed a significant response to suffixes, when presented in oddballs that can be fully broken down into morphemes. The fact that the system shows sensitivity to suffixes is perhaps not so surprising; they appear across different words and tend to be frequently recurring clusters of letters. As such, they are surely salient units in the language, both semantically (as they convey systematic meaning) and perceptually/orthographically (as frequent chunks). However, we see solid signs of suffix identification only in the context of morphological nonwords, which suggests that the process is not simply a catch of frequent and salient units, but involves a comprehensive (morphological) analysis of the whole strings. Therefore, this result sits well with the wealth of studies showing that the brain and the cognitive system make heavy use of morphology during reading and visual word identification (e.g., Amenta & Crepaldi, 2012; Feldman et al., 2000; Leminen et al., 2019; Marslen-Wilson et al., 2008; Rastle et al., 2000; Whiting et al., 2015). On the other hand, it is somewhat in contrast with more recent evidence showing that masked priming with nonwords is not modulated by the presence of suffixes (e.g., Lisi's studies). We will get back to this issue below.

Interestingly, stems did not elicit a significant response in the ROI analyses. Overall, we only found significant evidence for stem identification in the cluster-based permutation analysis on the developing readers.

This is quite surprising for a few reasons. First, stems are often encountered as whole words in English; from this point of view, they might be even more perceptually salient than suffixes, given that the surrounding blank spaces might work as "chunking cues" that help the system identify these items as important functional units (e.g., Grainger & Beyersmann, 2017). Furthermore, stems are more informative about word identity, allowing to narrow the lexical and semantic interpretation of a word more than a suffix does *per se*. For example, upon encountering *dark*-, a reader can reliably predict the general meaning of the rest of that word; instead, many different words end in -*ness*. We may therefore hypothesise that, at least for what concerns the signal captured by FPVS, the frequency of a morpheme may be more relevant than its informativeness.

These data also point quite naturally to theories of complex word processing that place more emphasis on affixes than on stems. This tenet was quite popular in the early days of psycholinguistic research into complex word identification, when the dominant view was that affixes are immediately identified and then stripped away, so that the stem is isolated and undergo the process of lexical identification proper (e.g., Taft & Forster, 1975). In fact, the emphasis in this account is rather more on the stem than on the suffix, as the former is the key for lexical access. However, the very first step in the processing of complex words is indeed the identification (and elimination) of the affix, quite in line with the present data.

The idea of "affix stripping" has been progressively abandoned, in favour of more explicit representational accounts (e.g., Taft, 1994). In these models, the original asymmetry between affixes and stems has become more blurred, and has mostly taken the form of whether affixes have an explicit representation at the most central levels of morphological processing (often called "lemma"). For example, Crepaldi et al. (2010) take the stance that affixes are not represented as lemmas, while Taft and Nguyen-Hoan (2010) suggest so; the present data seem to provide support for the latter. Where most theories agree is that affixes and stems are both represented more peripherally, early in the visual identification system, at a morpho-orthographic stage of processing (e.g., Crepaldi et al., 2010; Longtin et al., 2003; Rastle et al., 2004). Since the present paradigm insists particularly on automatic, quick visual processing, this may be the stage where the data described here have emerged. This would reconcile the contrast between Crepaldi et al.'s and Taft's theories, but would not explain the fundamental asymmetry between stems and suffixes that is apparent here.

Taking a fairly different perspective, Grainger and Beyersmann (2017) suggested that what is typically interpreted as morpho-orthographic processing may in fact reflect a mechanism of embedded word/stem identification that is not, per se, genuinely morphological, i.e., it would operate independently of the presence of an affix. This account fits perfectly with the recent observation that when lexical competition is partialled out in priming experiments, that is, when nonwords are used, affixed and non-affixed primes provide the same amount of facilitation (*farmald-FARM* = *farmness-FARM*). However, it is at odds with two aspects of the present results, (i) that affixes are identified, but stems are not and (ii) that morpheme identification proceeds much better in strings that are entirely decomposable.

From a developmental perspective, the pattern of results reveals similarities between adults and children: in both samples, a much stronger response emerges for words compared to morphemes, and, overall, suffixes appear to be more reliably identified than stems. A direct comparison was neither sought nor warranted by the differences in the experimental design; however, our findings suggest that children in Grades 5-6 have already developed an efficient adult-like visual identification system, thus corroborating our findings for words in Condition 0. As we noted above, this is at odds with the priming literature with children, particularly on morpho-orthographic effects. While some studies have provided data in line with the conclusion that one could draw here, that is, that children show a morpho-orthographic priming profile early on, similarly to adults (Quémart et al., 2011), others have noted that children don't show morpho-orthographic processing until quite late in development (Beyersmann et al., 2012; Dawson et al., 2018; Schiff et al., 2012). More than a theoretical issue, this might actually be an effect of the task; while masked priming is a fairly difficult paradigm for children, and therefore provides noisy evidence (although note that Dawson et al., 2018, is not based on masked priming), the present paradigm – completely implicit, behaviour-free and based on neural entrainment – might provide a better way to look into the developing reading system. Note, however, that a developmental trajectory that brings children to adult-like performance later on during reading acquisition was also highlighted in other studies. For example, in a reading aloud task, Burani et al. (2008) found that very young or less skilled readers benefit from the morphological structure of a word, as this allows them to access the constituents, by-passing whole-word access. On the other hand, more skilled readers, either developing or adult ones, would show less sensitivity to this aspect, as lexical access does not pose any additional difficulty for them.

Our findings are also somewhat informative as to the areas that appear to be involved in the response to morphemes in the adult readers. Indeed, the involvement of the VOTC in the identification of suffixes is not only supported by the ROI analysis, but by the clustering analysis as well. The fact that a response is elicited by morphemes in a region which is primarily devoted to visual word identification (and to visual identification more in general) suggests that suffixes are likely processed as visual units, at least at this stage of processing. This sits well with theories that state the existence of a level of morphological analysis that is mostly based on form (e.g., Crepaldi et al., 2010; Grainger and Ziegler, 2011; Xu and Taft, 2014). At this level of analysis, morphemes are primarily seen as frequent, statistically associated clusters of letters, perhaps not so differently from what happens in other domains of vision (e.g., Vidal et al., 2021). It is well known that neural circuitry in the ventral stream is particularly apt at finding regularities in the co-occurrence of lower-level units, to then build higher-level representations that take advantage of these regularities (e.g., Dehaene et al., 2005; Tkačik et al., 2010). This property is particularly convenient in the domain of visual word identification, which is characterised exactly by lower-level units (i.e., letters) that bind together in higher-level objects (i.e., morphemes and words). In this context, it should not be surprising that morphemes are captured as chunks of strongly associated letters.

Experimental evidence in support of this "statistical learning" view of visual word identification, and morphology in particular, is growing. For example, Chetail (2017) asked her participants to familiarise themselves with an artificial lexicon made up of pseudo-characters. The lexicon was such that some bigrams were particularly frequent; when participants were involved in a wordlikeness task with entirely novel stimuli, those that contained the frequent bigrams were judged as more word-like. So, even in a completely unfamiliar novel lexicon, made up of completely unfamiliar pseudo-characters, a few minutes of exposure were sufficient for participants to develop sensitivity to small clusters of particularly high frequency. With a similar design and experiment, Lelonkiewicz et al. (2020) were able to reproduce effects that emerged in morphological nonwords (e.g., Crepaldi et al., 2010; Taft & Forster, 1975) with an artificial lexicon that was entirely devoid of any phonological or semantic ties, that is, a set of purely visual, non-linguistic entities made up of sequences of pseudocharacters. These data suggest that at least part of the morphological effects that we observe with genuine linguistic material can be reproduced in purely visual, non-linguistic systems.

It appears less clear why in the children's data (with respect to stem identification), and partly in the adults' data (with respect to suffix identification), a cluster for morpheme identification emerges quite centrally in occipital sensors. While a direct comparison cannot be drawn with EEG studies, this location might be suggestive of a classic N400 evoked response in ERP experiments, thus reflecting some degree of semantic processing. However, MEG and fMRI studies locate the source of an N400-like response in the left temporal lobe (e.g., Halgren et al., 2002; Van Petten & Luka, 2006). As far as our data are concerned, further investigation on the brain source of this signal can help shed light on these sensor-level findings. We may cautiously assume that the largely central cluster for stem identification and the widespread bilateral cluster for suffix identification (when a relaxed alpha criterion is used at the cluster level) observed in the developing readers reflect a kind of processing which is less specific to morpho-orthographic units, perhaps suggestive of a more generic lexical/semantic response. This would agree with accounts that suggest morpho-semantic processing to mature earlier along development than morpho-orthography, which is instead thought to emerge only at the last stage of reading development (Grainger & Beyersmann, 2017), and to be still in maturation during adolescence (Dawson et al., 2018).

To conclude, we were able to show that sensitivity to morphological structure, also in the absence of explicit meaning assignment, has already sufficiently matured by Grades 5 and 6 to show up in the implicit, behaviour-free paradigm that we adopted here, FPVS-MEG. Moreover, the present results suggest that: (i) suffixes are identified, while stems are not (or at least, less so), and (ii) morpheme identification is stronger in strings that are entirely made up of morphemes (i.e., when both a stem and suffix are present). Signs of this identification process appears in sensors that are compatible with a neural source in VOTC, in line with accounts of (early) morphological identification as a predominantly visual process, perhaps connected to language-agnostic, statistical learning mechanisms (e.g., Crepaldi et al., 2010; Lelonkiewicz et al., 2020; Rastle & Davis, 2003; Vidal et al., 2021). However, morpheme identification shows up also in sensors that capture activity in other brain areas, compatible with deeper processing, perhaps up to the semantic level; this suggests that the FPVS-MEG paradigm adopted here is potentially able to also track this kind of processing, in addition to more peripheral, implicit visual mechanisms.

References

Algeo, J., & Algeo, A. S. (Eds.). (1993). *Fifty years among the new words: A dictionary of neologisms 1941-1991*. Cambridge: Cambridge University Press.

Amenta, S., & Crepaldi, D. (2012). Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in Psychology, 3,* 232. https://doi.org/10.3389/fpsyg.2012.00232

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438. https://doi.org/10.1037/a0023851

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). The CELEX Lexical Database (CD-ROM) Philadelphia. *PA: Linguistic Data Consortium, University of Pennsylvania*. https://doi.org/10.35111/gs6s-gm48

Bentin, S., & Feldman, L. B. (1990). The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew. *Quarterly Journal of Experimental Psychology*, *42*(4), 693-711. https://doi.org/10.1080/14640749008401245

Beyersmann, E., Casalis, S., Ziegler, J. C., & Grainger, J. (2015). Language proficiency and morpho-orthographic segmentation. *Psychonomic Bulletin & Review*, *22*(4), 1054-1061. https://doi.org/10.3758/s13423-014-0752-9

Beyersmann, E., Castles, A., & Coltheart, M. (2012). Morphological processing during visual word recognition in developing readers: Evidence from masked priming. *The Quarterly Journal of Experimental Psychology*, *65*(7), 1306-1326. https://doi.org/10.1080/17470218.2012.656661

Bozic, M., Marslen-Wilson, W. D., Stamatakis, E. A., Davis, M. H., & Tyler, L. K. (2007). Differentiating morphology, form, and meaning: Neural correlates of morphological complexity. *Journal of Cognitive Neuroscience*, *19*(9), 1464-1475. https://doi.org/10.1162/jocn.2007.19.9.1464

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433-436. https://doi.org/10.1163/156856897X00357

Burani, C., Marcolini, S., De Luca, M., & Zoccolotti, P. (2008). Morpheme-based reading aloud: Evidence from dyslexic and skilled Italian readers. *Cognition*, *108*(1), 243-262. 10.1016/j.cognition.2007.12.010

Burani, C., Marcolini, S., & Stella, G. (2002). How early does morpholexical reading develop in readers of a shallow orthography?. *Brain and Language*, *81*(1-3), 568-586. https://doi.org/10.1006/brln.2001.2548

Carlisle, J. F., & Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly, 40*(4), 428-449. https://doi.org/10.1598/RRQ.40.4.3

Casalis, S., Quémart, P., & Duncan, L. G. (2015). How language affects children's use of derivational morphology in visual word and pseudoword processing: Evidence from a cross-language study. *Frontiers in Psychology, 6*, 452. https://doi.org/10.3389/fpsyg.2015.00452

Crepaldi, D., Rastle, K., & Davis, C. J. (2010). Morphemes in their place: Evidence for position specific identification of suffixes. *Memory & Cognition*, *38*(3), 312-321. https://doi.org/10.3758/MC.38.3.312

Dawson, N., Rastle, K., & Ricketts, J. (2018). Morphological effects in visual word recognition: Children, adolescents, and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 645. http://dx.doi.org/10.1037/xlm0000485

De Rosa, M., Ktori, M., Vidal, Y., Bottini, R., & Crepaldi, D. (2020, September 21). Implicit Statistical Learning in Fast Periodic Visual Stimulation. https://doi.org/10.31219/osf.io/n85wc

Dzhelyova, M., & Rossion, B. (2014). Supra-additive contribution of shape and surface information to individual face discrimination as revealed by fast periodic visual stimulation. *Journal of Vision*, *14*(14), 15-15. https://doi.org/10.1167/14.14.15

Feldman, L. B. (2000). Are morphological effects distinguishable from the effects of shared meaning and shared form?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1431. https://doi.org/10.1037/0278-7393.26.6.1431

Frost, R., Forster, K. I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 829. https://doi.org/10.1037/0278-7393.23.4.829

Fruchter, J., & Marantz, A. (2015). Decomposition, lookup, and recombination: MEG evidence for the full decomposition model of complex visual word recognition. *Brain and Language*, *143*, 81-96. https://doi.org/10.1016/j.bandl.2015.03.001

Gold, B. T., & Rastle, K. (2007). Neural correlates of morphological decomposition during visual word recognition. *Journal of Cognitive Neuroscience*, *19*(12), 1983-1993. https://doi.org/10.1162/jocn.2007.19.12.1983

Grainger, J., & Beyersmann, E. (2017). Edge-aligned embedded word activation initiates morpho-orthographic segmentation. In *Psychology of Learning and Motivation* (Vol. 67, pp. 285-317). Academic Press. https://doi.org/10.1016/bs.plm.2017.03.009

Grainger, J., & Beyersmann, E. (2021). Effects of lexicality and pseudo-morphological complexity on embedded word priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(3), 518–531. https://doi.org/10.1037/xlm0000878

Grainger, J., & Ziegler, J. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology*, *2*, 54. https://doi.org/10.3389/fpsyg.2011.00054

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage*, *17*(3), 1101-1116. https://doi.org/10.1006/nimg.2002.1268

Hasenäcker, J., Beyersmann, E., & Schroeder, S. (2016). Masked morphological priming in German-speaking adults and children: Evidence from response time distributions. *Frontiers in Psychology*, *7*, 929. https://doi.org/10.3389/fpsyg.2016.00929

Hasenäcker, J., Beyersmann, E., & Schroeder, S. (2020). Morphological priming in children: Disentangling the effects of school-grade and reading skill. *Scientific Studies of Reading*, *24*(6), 484-499. https://doi.org/10.1080/10888438.2020.1729768

Hasenäcker, J., Solaja, O., & Crepaldi, D. (2020). Food in the corner and money in the cashews: Semantic activation of embedded stems in the presence or absence of a morphological structure. *Psychonomic Bulletin & Review*, *27*(1), 155-161. https://doi.org/10.3758/s13423-019-01664-z

Hasenäcker, J., Solaja, O., & Crepaldi, D. (2021). Does morphological structure modulate access to embedded word meaning in child readers? *Memory & Cognition*, 1-14. https://doi.org/10.3758/s13421-021-01164-3

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. R package. https://CRAN.R-project.org/package=vwr

Kirby, J. R., Deacon, S. H., Bowers, P. N., Izenberg, L., Wade-Woolley, L., & Parrila, R. (2012). Children's morphological awareness and reading ability. *Reading and Writing*, *25*(2), 389-410. https://doi.org/10.1007/s11145-010-9276-5

Kleiner, M., Brainard, D. H., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36 ECVP Abstract Supplement*.

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language*, *115*, 104152. https://doi.org/10.1016/j.jml.2020.104152

Leminen, A., Leminen, M. M., & Krause, C. M. (2010). Time course of the neural processing of spoken derived words: an event-related potential study. *Neuroreport*, *21*(14), 948-952. https://doi.org/10.1097/WNR.0b013e32833e4b90

Leminen, A., Smolka, E., Dunabeitia, J. A., & Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex*, *116*, 4-44. https://doi.org/10.1016/j.cortex.2018.08.016

Lochy, A., Van Belle, G., & Rossion, B. (2015). A robust index of lexical representation in the left occipito-temporal cortex as evidenced by EEG responses to fast periodic visual stimulation. *Neuropsychologia*, *66*, 18-31. https://doi.org/10.1016/j.neuropsychologia.2014.11.007

Lochy, A., Van Reybroeck, M., & Rossion, B. (2016). Left cortical specialization for visual letter strings predicts rudimentary knowledge of letter-sound association in preschoolers. *Proceedings of the National Academy of Sciences*, *113*(30), 8544-8549. https://doi.org/10.1073/pnas.1520366113

Longtin, C. M., & Meunier, F. (2005). Morphological decomposition in early visual word processing. *Journal of Memory and Language*, *53*(1), 26-41. https://doi.org/10.1080/01690960701588004

Longtin, C. M., Segui, J., & Hallé, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *18*(3), 313-334. https://doi.org/10.1080/01690960244000036

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marslen-Wilson, W. D, & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Sciences*, *2*(11), 428-435. https://doi.org/10.1016/s1364-6613(98)01239-x

Marslen-Wilson, W. D., & Tyler, L. K. (2007). Morphology, language and the brain: the decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 823-836. https://doi.org/10.1098/rstb.2007.2091

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*. https://doi.org/10.1155/2011/156869

Quek, G. L., Liu-Shuang, J., Goffaux, V., & Rossion, B. (2018). Ultra-coarse, single-glance human face detection in a dynamic visual stream. *NeuroImage*, *176*, 465-476. https://doi.org/10.1016/j.neuroimage.2018.04.034

Quémart, P., Casalis, S., & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of Experimental Child Psychology*, *109*(4), 478-496. https://doi.org/10.1016/j.jecp.2011.02.008

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rastle, K. (2019). The place of morphology in learning to read in English. *Cortex*, *116*, 45 54. https://doi.org/10.1016/j.cortex.2018.02.008

Rastle, K. & Davis, M. H. (2003). Reading morphologically-complex words: Some thoughts from masked priming. In: Kinoshita, S., & Lupker, S. J. (Eds.) *Masked priming: state of the art*. Hove: Psychology Press.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090-1098. https://doi.org/10.3758/BF03196742

Rossion, B. (2014). Understanding individual face discrimination by means of fast periodic visual stimulation. *Experimental Brain Research*, *232*(6), 1599-1621. https://doi.org/10.1007/s00221-014-3934-9

Rossion, B., Torfs, K., Jacques, C., & Liu-Shuang, J. (2015). Fast periodic presentation of natural images reveals a robust face-selective electrophysiological response in the human brain. *Journal of Vision, 15*(1), 18-18. https://doi.org/10.1167/15.1.18

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, *50*(4), 1568-1580. https://doi.org/10.3758/s13428-017-0981-8

Schiff, R., Raveh, M., & Fighel, A. (2012). The development of the Hebrew mental lexicon: When morphological representations become devoid of their meaning. *Scientific Studies of Reading*, *16*(5), 383-403. https://doi.org/10.1080/10888438.2011.571327

Seidenberg, M. S. (1987). *Sublexical structures in visual word recognition: Access units or orthographic redundancy?* In M. Coltheart (Ed.), *Attention and performance 12: The Psychology of Reading* (p. 245–263). Lawrence Erlbaum Associates, Inc.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and cognitive processes*, *9*(3), 271-294.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology Section A*, *57*(4), 745-765. https://doi.org/10.1080/02724980343000477

Taft, M. (2015). *The nature of lexical representation in visual word recognition.* In A. Pollatsek & R. Treiman (Eds.), *Oxford library of psychology. The Oxford handbook of reading* (p. 99–113). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.013.6

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638-647. https://doi.org/10.1016/S0022-5371(75)80051-X

Taft, M., & Nguyen-Hoan, M. (2010). A sticky stick? The locus of morphological representation in the lexicon. *Language and Cognitive Processes*, *25*(2), 277-296. https://doi.org/10.1080/01690960903043261

Tkačik, G., Prentice, J. S., Victor, J. D., & Balasubramanian, V. (2010). Local statistics in natural scenes predict the saliency of synthetic textures. *Proceedings of the National Academy of Sciences*, *107*(42), 18149-18154. https://doi.org/10.1073/pnas.0914916107

Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, *38*(4), 584-589. https://doi.org/10.3758/BF03193889

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190. https://doi.org/10.1080/17470218.2013.850521

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language*, *97*(3), 279-293. https://doi.org/10.1016/j.bandl.2005.11.003

Whiting, C., Shtyrov, Y., & Marslen-Wilson, W. (2015). Real-time functional architecture of visual word recognition. *Journal of Cognitive Neuroscience*, *27*(2), 246-265. https://doi.org/10.1162/jocn_a_00699

Vidal, Y., Viviani, E., Zoccolan, D., & Crepaldi, D. (2021). A general-purpose mechanism of visual feature association in visual word identification and beyond. *Current Biology*. https://doi.org/10.1016/j.cub.2020.12.017

Zweig, E., & Pylkkänen, L. (2009). A visual M170 effect of morphological complexity. *Language and Cognitive Processes*, *24*(3), 412-439. https://doi.org/10.1080/01690960802180420

Chapter VI

General Discussion

The general research question that this thesis has addressed concerns the ability to extract orthographic and morpho-orthographic regularities during reading, as well as the role of such mechanism in reading development. Thus far, research on sensitivity to the statistical properties of written language has been conducted largely with artificial experimental designs and has not focused on natural reading, neglecting the potential impact of such mechanisms on an activity that we carry out routinely across most of our lifespan.

Indeed, if statistical learning is exploited for all visual processing, including orthographic processing, we might expect to see this mechanism at play also during text reading, where it might emerge in the form of sensitivity to clusters of frequently co-occurring letters (n-grams), for instance. Furthermore, if this learning mechanism is used to form and reinforce orthographic representations during the early stages of reading development, we should also see signs of it in relatively young readers.

To investigate these issues with an ecologically valid approach, we carried out an eye-tracking study using a multiline text reading task. This study addressed a gap in the literature, providing a large developmental eye-tracking corpus of natural reading, called *EyeReadIt* (presented in Chapter II of this thesis). This resource was validated through the analysis of developmental trends in reading behaviour, such as reading rate (number of words read per minute), which increases as a function of age, or fixation duration, which, instead, decreases as readers become more skilled. Moreover, we looked for some "benchmark" lexical effects, such as those of word length and word frequency, on eye-tracking measures. Results from both sets of analyses replicate well-established findings and, in fact, extend their scope to the natural reading of multi-line texts (most of these effects were originally obtained with isolated sentences). Additionally, the analysis was extended to an investigation of developing readers' sensitivity to morphological structure during natural reading. Our findings demonstrate the added value that a similar developmental eye-tracking resource can provide to this field of research.

As a further development, *EyeReadIt* stimulated the creation of a methodological study, in collaboration with Dr Jon Carr (Chapter III of this thesis). Here, we assessed eight existing algorithms and two novel ones that we proposed, for the automatised correction of fixations. Eye-tracking researchers, especially those working with multiline text reading, may indeed find that some fixations in their data are misaligned, due to a set of phenomena related to "vertical drift". These ten methods were validated on some of the data from *EyeReadIt*, as well as on simulated data, allowing us to provide researchers with guidance in the selection of the most appropriate algorithm (or algorithms) for the different phenomena characterising the eye-movement recordings. In particular, one of the novel algorithms that we proposed, which is based on Dynamic Time Warping (Sakoe & Chiba, 1978; Vintsyuk, 1968), proved very promising for the correction of vertical drift in noisy eye-tracking data, such as children's.

Crucially, *EyeReadIt* was also used to address our question about the emergence of sensitivity to the statistical properties of the written language across reading development (Chapter IV of this thesis). Such question was operationalised through the use of n-gram (bigram, trigram, quadrigram) frequency metrics (minimal, average, maximal frequency); this differs from the current literature, which has largely been confined to the use of average bigram frequency, with mixed evidence about its role (e.g., Chetail, 2015; Schmalz & Mulatti, 2017). Our results showed an effect of n-gram frequency metrics, particularly of average and maximum frequency for bigrams and trigrams, across developing and adult readers, with longer durations in all three eye-tracking measures considered (first-of-many-fixations duration, gaze duration, and total reading time). Importantly, at a lexical and post-lexical processing stage, such effects were also significantly modulated by children's grade, with younger readers displaying greater sensitivity to the n-gram frequency statistics of the texts they are presented with, compared to adults. These findings contribute to a theoretical advancement with respect to the role of sublexical regularities in reading development, in that they inform us that readers of a transparent orthography like Italian appear to make use of such statistical cues from quite a young age. While the direction of the effects that we observed may appear counterintuitive, with higher maximal frequency n-grams eliciting longer reading times, this is taken to signal that n-grams represent highly salient perceptual orthographic units, especially in young developing readers, who might be relying more significantly than adults on sublexical processing, while still perfecting whole-word representations (see, e.g., Castles et al., 2007; Nation, 2009).

After assessing the sensitivity to orthographic regularities through eye tracking, we shifted our focus to the cognitive (and neural) underpinnings of the developmental trajectories of sensitivity to morphological structure. This has been addressed in an MEG study (Chapter V), using an FVPS-oddball paradigm, conducted at Macquarie University (Sydney, Australia) in collaboration with Dr Elisabeth Beyersmann, Prof Paul Sowman, and Prof Anne Castles. While it has been established that priming for opaque pairs (*corner-CORN*) occurs in skilled adult readers (e.g., Longtin et al., 2003; Rastle et al., 2004), evidence concerning developing readers is far more mixed. Several studies have failed to find evidence for similar morpho-orthographic processing in very young readers (e.g., Beyersmann et al., 2015; Hasenäcker et al., 2016); it has thus been proposed that this only fully matures at quite a late stage of reading development (e.g., Dawson et al., 2018; Grainger & Beyersmann, 2017). However, sensitivity to words' morphological structure in the absence of whole-word semantic access has at times been reported in children as well (for evidence with primary school children, see: Quémart et al., 2011; Burani et al., 2002, 2008; for partial evidence of morpho-orthographic processing in 7th graders, see Schiff et al., 2012).

In order to assess whether developing readers display sensitivity to the presence of morphemes in pseudowords, we used FPVS – a behaviour-free paradigm, that allowed us to tap into automatic and implicit morpheme identification, while MEG recordings were acquired from 17 children in Grades 5 and 6, and 28 adults. Our findings revealed a similar pattern of morpheme identification between children and adults. For both populations, an ROI-based analysis conducted at sensor level over left VOTC showed a significant response to suffixes, when presented in oddball stimuli that can be fully decomposed into a stem and a suffix (*softity*). However, this is not the case for the same type of oddballs when the contrast tracks stem identification. The adult version of the study featured two additional conditions in which oddballs contained just one morpheme, and could thus not be fully broken down into their constituents. In both additional conditions, no significant oddball response emerged, indicating failure to identify stems or suffixes in non-morphologically structured pseudowords, and thus revealing an essential role of context for morpheme identification.

Further sensor-level analysis of the MEG data was carried out through cluster-based permutation, which yielded a consistent pattern of results in adults, whereby a robust response to suffixes in complex oddballs emerges in an occipitally located (mostly central, partially left-lateralised) cluster. As far as the developing readers are concerned, instead, the clustering analysis revealed a significant central occipital cluster in response to the condition tracking identification of stems in fully segmentable oddballs (*softity*).

Overall, suffixes appear to be reliably identified both by children and by adults, when full decomposition of the oddballs is warranted. This suggests that by 10-12 years of age readers are already equipped with an adult-like pattern of visual identification of morphemes, at least for what concerns the level of processing that FPVS taps upon. Our results regarding the brain areas that are responsible for morpheme identification suggest an involvement of the left VOTC, in line with a few neuroimaging accounts (see Gold & Rastle, 2007; Leminen et al., 2019, for a review). The cluster-based permutation analysis yielded instead less straightforward findings, revealing a robust oddball response in centrally located occipital sensors, especially for stem identification in children (and partially for suffix identification in adults). This may be suggestive of a response that is less specifically tuned to morpho-orthographic processing, and perhaps more generally lexical/semantic. If this were the case, such findings would support some classic views on the development of morphological sensitivity, according to which morpho-semantic processing occurs earlier along development (e.g., Grainger & Beyersmann, 2017). However, caution is warranted in the interpretation of these results, as source-level analysis will hopefully provide further insight as to the location of our effects across developmental stages.

Conclusive remarks and implications

The studies presented in this thesis aimed at providing a research contribution as to the emergence, across reading development, of sensitivity to the statistical properties of sublexical units. First, we were able to answer a refined question – whether sensitivity to n-gram frequency emerges in reading – adopting an ecologically valid approach to the developmental study of eye movements in reading. Furthermore, we employed an implicit MEG paradigm to investigate an underspecified aspect of morpho-orthographic processing—its neural underpinnings at different stages of reading development.

On a theoretical level, our findings suggest that both n-grams and morphemes are exploited as reading units, also at quite early stages of reading development. As outlined in the Introduction (Chapter I of this thesis), sensitivity to orthographic and morpho-orthographic regularities supports the account that children must be relying on a statistical learning mechanism, which likely allows them to improve their reading abilities. Hence, our findings have a twofold implication for current major models of orthographic and morpho-orthographic processing on the one hand, and for models of reading acquisition on the other.

We showed that children are sensitive to statistics of letter co-occurrences quite early during their pathway towards becoming proficient readers. This resonates with several models of visual word identification that state the existence of an early stage of processing where morphemes (and, possibly, other frequent letter clusters) are identified independently of their semantic contribution (e.g., Crepaldi et al., 2010; Taft & Nguyen-Hoan, 2010), or with models that highlight the sensitivity of the system (or at least part of it) to letter statistics more generally (e.g., Grainger & Ziegler, 2011; Grainger & Beyersmann, 2017).

However, the role of *within*-level regularities is still underspecified, in a developmental framework. Research has highlighted the role of sensitivity to linguistic regularities for skilled reading. As skilled reading builds up on an increasingly more efficient mapping between graphemes and phonemes, or between print and meaning, the focus of such research has so far mostly concerned *between*-level regularities. For instance, Share's (1995) self-teaching hypothesis relies on the acquisition of letter-to-sound mappings, through a phonological recoding process; the core of Ziegler and Goswami's (2005) theory is that the shared grain size between orthographic (also at a sublexical level) and phonological units supports successful grapheme-to-phoneme mapping.

In the present work, we revealed the saliency, across development, of statistical regularities at the orthographic level – independently of whether these units map onto specific sound or meaning entities –, showing that developing readers exploit such regularities to process written language from quite an early stage, at least in a language with a very transparent orthography like Italian. Also, children appear to rely on letter statistics more markedly than adults, whose reading system, we would suggest, is perhaps already skilled and efficient enough to afford greater reliance on whole-word representations. We thus suggest that this aspect ought to be taken into account for the advancement of literacy acquisition research, besides and beyond the obvious role played by regularities across different linguistic levels.

Of course, the idea that visual word identification, and morphology in particular, might be based on the many statistical cues that orthographic systems offer is not entirely new (e.g., Baayen et al., 2011; Rueckl et al., 1997; Rumelhart et al., 1986; Seidenberg, 1987). However, these accounts focused mostly on regularities *between* levels of processing (e.g., form and meaning, orthography and phonology), while the results presented here (and several others that emerged recently; e.g., Chetail, 2017; Lelonkiewicz et al., 2020; Vidal et al., 2021) stress the fact that statistical learning might also be happening *within* levels of processing (orthography, in the present case), so that the relevant functional units (e.g., word and morpheme representations, but possibly also representations for other statistically associated chunks that do not correspond to meaning or phonological units) get established based on this mechanism.

As outlined in the first section of this Discussion, developmental research is still unclear about sensitivity to a specific type of between-level regularities, that is, morpho-orthographic processing. Differently from some of the prior behavioural findings, we showed that, at a neural level, English-speaking fifth and sixth graders displayed a similar pattern of morpheme identification as their adult counterparts, in an automatic and implicit task. In addition to this, we provided evidence for the role of context towards successful identification of morphemes in pseudowords[1].

Furthermore, we showed that suffixes were successfully identified by children, as well, suggesting that children aged 10-12 may have already reached a higher proficiency stage, in

[1] Note that this conclusion is especially warranted for the adult participants, as they were subjected to two additional experimental conditions aimed precisely at shedding light on this aspect. Oddball pseudowords were always made up of a stem and a suffix in the two conditions shared by the child and adult version of the experiment.

terms of morpho-orthographic processing, than has been proposed based on behavioural evidence (e.g., Dawson et al., 2018). These results inform a rich debate around the role of stems and affixes in visual word identification, in both adults (e.g., Crepaldi et al., 2010; Feldman et al., 2009; Xu & Taft, 2013) and children (e.g., Grainger & Beyersmann, 2017), suggesting a predominant role for suffixes (vs. stems) and for the morphological structure of letter strings (vs. the mere identification of frequent chunks or embedded words). As we employed for the first time a technique like FPVS, paired with MEG, to the study of morphological sensitivity in children, we are aware that our findings, albeit novel, warrant further investigation. However, if our general pattern of results were to be corroborated by future findings, this may inform a reconsideration of the developmental stage at which morpho-orthographic sensitivity fully forms.

Finally, from a methodological perspective, this work provides two valuable methodological resources for the eye-tracking and reading community: i. a large developmental eye-tracking corpus of natural reading in Italian, *EyeReadIt* (Chapter II), which is being finalised and will soon be freely available; ii. a series of algorithms for the automatised correction of vertical drift in eye-tracking data (Chapter III), which is already publicly available.

References

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49. https://doi.org/10.1109/tassp.1978.1163055

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychological Review*, *118*(3), 438.https://doi.org/10.1037/a0023851

Beyersmann, E., Casalis, S., Ziegler, J. C., & Grainger, J. (2015). Language proficiency and morpho-orthographic segmentation. *Psychonomic Bulletin & Review*, *22*(4), 1054-1061. https://doi.org/10.3758/s13423-014-0752-9

Burani, C., Marcolini, S., De Luca, M., & Zoccolotti, P. (2008). Morpheme-based reading aloud: Evidence from dyslexic and skilled Italian readers. *Cognition*, *108*(1), 243-262. 10.1016/j.cognition.2007.12.010

Burani, C., Marcolini, S., & Stella, G. (2002). How early does morpholexical reading develop in readers of a shallow orthography?. *Brain and Language*, *81*(1-3), 568-586. https://doi.org/10.1006/brln.2001.2548

Castles, A., Davis, C., Cavalot, P., & Forster, K. (2007). Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology, 97*(3), 165-182. https://doi.org/10.1016/j.jecp.2007.01.006

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology, 6*, 645. https://doi.org/10.3389/fpsyg.2015.00645

Chetail, F. (2017). What do we do with what we learn? Statistical learning of orthographic regularities impacts written word processing. *Cognition, 163*, 103-120. https://doi.org/10.1016/j.cognition.2017.02.015

Crepaldi, D., Rastle, K., & Davis, C. J. (2010). Morphemes in their place: Evidence for position specific identification of suffixes. *Memory & Cognition*, *38*(3), 312-321. https://doi.org/10.3758/MC.38.3.312

Dawson, N., Rastle, K., & Ricketts, J. (2018). Morphological effects in visual word recognition: Children, adolescents, and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 645. http://dx.doi.org/10.1037/xlm0000485

Feldman, L. B., O'Connor, P. A., del Prado, M., & Martín, F. (2009). Early morphological processing is morphosemantic and not simply morpho-orthographic: A violation of form-then-meaning accounts of word recognition. *Psychonomic Bulletin & Review, 16*, 684–691. https://doi.org/10.3758/PBR.16.4.684.

Gold, B. T., & Rastle, K. (2007). Neural correlates of morphological decomposition during visual word recognition. *Journal of Cognitive Neuroscience, 19*(12), 1983-1993. https://doi.org/10.1162/jocn.2007.19.12.1983

Grainger, J., & Beyersmann, E. (2017). Edge-aligned embedded word activation initiates morpho-orthographic segmentation. In *Psychology of Learning and Motivation* (Vol.67, pp. 285-317). Academic Press. https://doi.org/10.1016/bs.plm.2017.03.009

Grainger, J., & Ziegler, J. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology*, *2*, 54. https://doi.org/10.3389/fpsyg.2011.00054

Hasenäcker, J., Beyersmann, E., & Schroeder, S. (2016). Masked morphological priming in

German-speaking adults and children: Evidence from response time distributions. *Frontiers in Psychology*, *7*, 929. https://doi.org/10.3389/fpsyg.2016.00929

Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language, 115,* 104152. https://doi.org/10.1016/j.jml.2020.104152

Leminen, A., Smolka, E., Dunabeitia, J. A., & Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex, 116*, 4-44. https://doi.org/10.1016/j.cortex.2018.08.016

Longtin, C. M., Segui, J., & Hallé, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *18*(3), 313-334. https://doi.org/10.1080/01690960244000036

Nation, K. (2009). Form–meaning links in the development of visual word recognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3665-3674. https://doi.org/10.1098/rstb.2009.0119

Quémart, P., Casalis, S., & Colé, P. (2011). The role of form and meaning in the processing of written morphology: A priming study in French developing readers. *Journal of Experimental Child Psychology*, *109*(4), 478-496. https://doi.org/10.1016/j.jecp.2011.02.008

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090-1098. https://doi.org/10.3758/BF03196742

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S., & Mars, F. (1997). Morphological priming, fragment completion and connectionist networks. *Journal of Memory and Language*, *36*, 382–405. https://doi.org/10.1006/jmla.1996.2489

Rumelhart, D. E., McClelland, J. L., & PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, Volume 1. Cambridge, MA: MIT Press.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26*(1), 43–49. https://doi.org/10.1109/tassp.1978.1163055

Schiff, R., Raveh, M., & Fighel, A. (2012). The development of the Hebrew mental lexicon: When morphological representations become devoid of their meaning. *Scientific Studies of Reading, 16*(5), 383-403. https://doi.org/10.1080/10888438.2011.571327

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon*, *12*(2), 263-282. https://doi.org/10.1075/ml.17009.sch

Seidenberg, M. S. (1987). *Sublexical structures in visual word recognition: Access units or orthographic redundancy?* In M. Coltheart (Ed.), *Attention and performance 12: The Psychology of Reading* (p. 245–263). Lawrence Erlbaum Associates, Inc.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Taft, M., & Nguyen-Hoan, M. (2010). A sticky stick? The locus of morphological representation in the lexicon. *Language and Cognitive Processes*, *25*(2), 277-296.

https://doi.org/10.1080/01690960903043261

Vidal, Y., Viviani, E., Zoccolan, D., & Crepaldi, D. (2021). A general-purpose mechanism of visual feature association in visual word identification and beyond. *Current Biology*. https://doi.org/10.1016/j.cub.2020.12.017

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics, 4* (1), 52–57. https://doi.org/10.1007/BF01074755

Xu, J., & Taft, M. (2014). Solely soles: Inter-lemma competition in inflected word recognition. *Journal of Memory & Language*, 76, 127-140. https://doi.org/10.1016/j.jml.2014.06.008

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. https://doi.org/10.1037/0033-2909.131.1.3

Appendix A

Materials used as stimuli for *EyeReadIt* (Chapter II of this Thesis). The column "Set" specifies whether the passage belongs Set A or B (the practice passage was common to A and B). "Number" illustrates the order in which the passages were displayed. "Trial type" distinguishes between practice and experimental trials. "Passage" reports the content of the passage as it was displayed. In the column "Author; Book; Story" is reported, as applicable, the information about the author, book, and story for each passage, in Italian, and in brackets an English translation is provided; website references are also provided in this column.

Set	**Number**	**Trial type**	**Passage**	**Author; Book; Story**
A,B	1	Practice	La bambina bussò alla porta. "Nonnina, posso entrare?" - chiese. Il lupo, nascosto tra le coperte, rispose: "Vieni! Tira il paletto ed entra". "Ma che voce che hai, nonna!" - si stupì la bambina. "È per salutarti meglio, tesoro." - rispose il lupo. "E che occhi grandi che hai!" "È per guardarti meglio, cara." "Ma che mani grandi che hai!" "Per accarezzarti meglio, bambina mia!" "Ma che bocca grande che hai!" "Per mangiarti meglio!" - ruggì il lupo.	Cappuccetto Rosso (Little Red Riding Hood) Retrieved from: http://culturabile.it/wp-content/uploads/cappuccetto%20rosso.txt

Set	Number	Trial type	Passage	Author; Book; Story
A	2	Experimental	C’erano una volta tre Orsi, che vivevano in una casina nel bosco. C’era Babbo Orso grosso grosso, con una voce grossa grossa; c’era Mamma Orsa grossa la metà, con una voce grossa la metà; e c’era un Orsetto piccolo piccolo con una voce piccola piccola. Una mattina i tre Orsi facevano colazione e Mamma Orsa disse: "La pappa è troppo calda, ora. Andiamo a fare una passeggiata nel bosco, mentre la pappa diventa fredda". Così i tre Orsi andarono a fare una passeggiata nel bosco. Mentre erano via, arrivò una piccola bimba chiamata Riccidoro. Quando vide la casetta nel bosco, si domandò chi mai potesse vivere là dentro, e picchiò alla porta. Nessuno rispose, e la bimba picchiò ancora. Nessuno rispose: Riccidoro allora aprì la porta ed entrò. E là, nella piccola stanza, vide una tavola apparecchiata per tre.	Fratelli Grimm; Fiabe; Riccidoro (Brothers Grimm; Fairy Tales; Goldilocks) Retrieved from: http://www.lefiabe.com/grimm/riccidoro.htm

Set	Number	Trial type	Passage	Author; Book; Story
A	3	Experimental	Così il soldato viveva allegramente, andava a teatro, passeggiava nel giardino reale di Parigi e dava ai poveri tanto denaro, e questo era ben fatto. Lo sapeva bene dai tempi passati, quanto fosse brutto non avere neppure un soldo. Ora era ricco e aveva abiti eleganti e si trovò tantissimi amici, tutti a ripetergli quanto era simpatico, un vero cavaliere, e questo al soldato faceva molto piacere. Ma spendendo ogni giorno dei soldi e non guadagnandone mai, alla fine rimase con i soli spiccioli e fu costretto a trasferirsi, dalle splendide stanze in cui aveva abitato, in una piccolissima cameretta, proprio sotto il tetto, e dovette pulirsi da sé gli stivali e cucirli con un ago, e nessuno dei suoi amici andò a trovarlo, perché vi erano troppe scale da fare.	H.C. Andersen; Fiabe; L'acciarino (H.C. Andersen; Fairy Tales; The Tinderbox) Retrieved from: http://archigianni.altervista.org/senza_titolo1_00001b.htm
A	4	Experimental	Fabio si trovava un giorno nella foresta, e aveva appena finito di tagliare legna all'incirca sufficiente per caricare i suoi asini, quando vide una fitta polvere che si alzava in aria e avanzava verso di lui. Guarda attentamente e distingue un numeroso gruppo di persone a cavallo che arrivavano a buona andatura. Per quanto nel paese non si parlasse di briganti, Fabio, tuttavia, sospettò che questi cavalieri potessero esserlo. Senza considerare ciò che sarebbe capitato ai suoi asini, pensò a salvare sé stesso. Salì su un grosso albero i cui rami si diramavano in cerchio, tanto vicini gli uni agli altri da essere separati solo da uno spazio piccolissimo.	Le mille e una notte; Alì Babà e i quaranta ladroni (One Thousand and One Nights; Ali Baba and the Forty Thieves) Retrieved from: http://www.nuvolotta.altervista.org/milleuna/alibaba4.htm

Set	Number	Trial type	Passage	Author; Book; Story
A	5	Experimental	Alcuni piatti ricoperti dalla loro campana d'argento furono posati simmetricamente sulla tovaglia e noi prendemmo posto a tavola. Il pane e il vino brillavano per la loro assenza e l'acqua, benché fosse limpida e fresca, non era troppo gradita a Lorenzo. Tra le vivande che ci furono servite c'erano diverse qualità di pesci cucinati accuratamente, ma di altre, peraltro eccellenti, non avrei nemmeno saputo dire se fossero animali o vegetali. Su ogni piatto era incisa la lettera N circondata da un motto quanto mai adatto a quel battello sottomarino. La lettera N era senza dubbio l'iniziale del nome dell'enigmatico personaggio che comandava negli abissi.	J. Verne; Ventimila leghe sotto i mari (J. Verne, Twenty Thousand Leagues Under the Seas) Retrieved from: http://www.nemoischia.it/wp-content/uploads/2013/10/20mlaleghe.pdf
A	6	Experimental	Il capitano alzatosi più presto del solito era sceso alla spiaggia col suo coltellaccio dondolante sotto le larghe falde del suo abito blu, il cannocchiale sotto l'ascella, e il cappello buttato indietro sulla nuca. Vedo ancora il suo alito ondeggiare in aria dietro a lui come fumo mentre egli si allontanava rapidamente. L'ultimo suono che giunse alle mie orecchie mentre egli girava dietro la grande rupe fu un potente sbuffo di ira, come se egli ancora fosse agitato dal pensiero del dottor Rossi. Mia madre era in quel momento di sopra col papà; ed io stavo apparecchiando la tavola per la colazione del capitano, quando l'uscio della sala si aprì, ed uno sconosciuto si fece avanti. Era pallido come cera; due dita gli mancavano alla mano sinistra; e, per quanto portasse un coltellaccio, non pareva troppo aggressivo.	R.L. Stevenson; L'isola del tesoro (R.L. Stevenson; Treasure Island) Retrieved from: https://www.lingq.com/el/lesson/capitolo-2-1-1111670/

Set	Number	Trial type	Passage	Author; Book; Story
A	7	Experimental	In estate vi è in quella valle un visitatore che gli Indiani non conoscono. È un grande lupo dalla meravigliosa pelliccia, simile agli altri lupi, e tuttavia diverso da loro. Arriva solitario dal ridente paese dei boschi e scende fino a una radura tra gli alberi. Là un fiume chiaro fluisce da sacchi marciti di pelle di alce e si disperde a terra; lunghe erbe e muschi lo ricoprono e nascondono al sole il suo giallo splendore. E là egli rimane per qualche tempo silenzioso, ululando una volta sola, a lungo e tristemente, prima di partire. Non sempre è solo. Quando vengono le lunghe notti d'inverno e i lupi seguono il loro cibo nelle vallate più basse, lo si può vedere correre alla testa del branco nella pallida luce lunare o dell'aurora boreale.	J. London; Il richiamo della foresta (J. London; The Call of the Wild) Retrieved from: https://ita.calameo.com/read/002611183bdbb50b159ff
B	2	Experimental	C’era una volta un vecchio asino che aveva lavorato sodo per tutta la vita. Ormai non era più capace di portare pesi e si stancava facilmente, per questo il suo padrone aveva deciso di relegarlo in un angolo della stalla ad aspettare la morte. L’asino però non voleva trascorrere così gli ultimi anni della sua vita. Decise di andarsene a Brema, dove sperava di poter vivere facendo il musicista. Si era incamminato da poco quando incontrò un cane, magro e ansimante. "Come mai hai il fiatone?" gli chiese. "Sono dovuto scappare in tutta fretta per salvare la pelle" gli rispose il cane. "Il mio padrone voleva uccidermi, perché ora che sono vecchio non gli servo più".	Fratelli Grimm; Fiabe; I musicanti di Brema (Brothers Grimm; Fairy Tales; The Bremen Town Musicians) Retrieved from: https://www.fiaberella.it/i-musicanti-di-brema/

Set	Number	Trial type	Passage	Author; Book; Story
B	3	Experimental	Ida mise i fiori nel lettino della bambola, li coprì per bene con la coperta e disse che dovevano stare tranquilli: avrebbe preparato del tè per loro, così sarebbero guariti e si sarebbero alzati di nuovo l'indomani. Poi tirò le tende vicino al lettino per evitare che il sole li disturbasse. Per tutta la sera non poté fare a meno di pensare a quello che lo studente le aveva raccontato, e quando lei stessa dovette andare a letto, guardò prima dietro le tendine della finestra dove c'erano i bei fiori della sua mamma, i giacinti e i tulipani, e sussurrò piano piano: "So bene che dovete andare al ballo questa notte"; i fiori fecero finta di niente, non mossero neppure una foglia, ma Ida sapeva bene quello che diceva.	H.C. Andersen; Fiabe; I fiori della piccola Ida (H.C. Andersen; Fairy Tales; Little Ida’s Flowers) Retrieved from: https://www.andersenstories.com/it/andersen_fiabe/i_fiori_della_piccola_ida
B	4	Experimental	L’uomo con la giacca blu portava la bisaccia come gli altri, si avvicinò alla roccia, molto vicino all'albero su cui Giovanni si era rifugiato; e, dopo essersi fatto strada attraverso gli arbusti, pronunciò queste parole: "Sesamo, apriti", così distintamente che Giovanni le sentì. Appena il capo dei ladri le ebbe pronunciate, si aprì una porta; e, dopo aver fatto passare tutti i suoi uomini davanti a sé ed averli fatti entrare tutti, entrò anche lui, e la porta si chiuse. I ladri restarono a lungo nella rupe; e Giovanni, temendo che qualcuno di loro o tutti insieme uscissero mentre egli lasciava il suo nascondiglio per fuggire, fu costretto a rimanere sull'albero e ad aspettare con pazienza.	Le mille e una notte; Alì Babà e i quaranta ladroni (One Thousand and One Nights; Ali Baba and the Forty Thieves) Retrieved from: http://www.nuvolotta.altervista.org/milleuna/alibaba4.htm

Set	Number	Trial type	Passage	Author; Book; Story
B	5	Experimental	Verso mezzogiorno entrai dal capo con qualche bibita rinfrescante, e medicine. Egli si trovava ancora nel medesimo stato, forse un tantino sollevato, e appariva insieme debole ed eccitato. "Giacomo" disse "tu sei l'unico, qui, che valga qualcosa; e tu sai come io sono sempre stato buono con te. Non c'è stato mese che non ti abbia pagato i tuoi quattro euro. E ora tu vedi, amico mio, come sono malandato e abbandonato da tutti. Giacomo, tu mi devi dare un bicchierino di rum; è vero che me lo dai, mio piccolo amico?". "Il medico..." presi a dire. Ma egli mi tagliò la parola con una voce fiacca ma appassionata. "I medici sono una massa di scope: e quel medico, che vuoi che sappia, lui, di gente di mare? Io sono stato in paesi dove si arrostiva, e i miei compagni la febbre gialla te li faceva cascar come mosche, e i terremoti facevano ondeggiare la terra come un mare: ebbene, che può sapere il medico di paesi simili?"	R.L. Stevenson; L'isola del tesoro (R.L. Stevenson; Treasure Island) Retrieved from: https://www.lingq.com/el/lesson/capitolo-2-1-1111670/

Set	Number	Trial type	Passage	Author; Book; Story
B	6	Experimental	Fido non era né un cane casalingo né un cane da canile. Il reame era tutto suo. Si tuffava nella vasca o andava a caccia con i figli del giudice; scortava Marta e Alice, le figlie del giudice, durante lunghe passeggiate mattutine o crepuscolari; e, nelle serate invernali, stava sdraiato ai piedi del giudice davanti al camino scoppiettante della biblioteca. Si lasciava cavalcare dai nipotini del giudice o li faceva rotolare sull'erba, e sorvegliava i loro passi nelle loro avventurose escursioni alla fontana nel cortile delle scuderie e anche più in là, verso i prati e i cespugli. Andava deciso fra i segugi e ignorava Tito e Isabella nel modo più assoluto, perché era un re: un re di tutto ciò che camminava, strisciava o volava nella proprietà del giudice Bianchi, compresi gli uomini.	J. London; Il richiamo della foresta (J. London; The Call of the Wild) Retrieved from: https://ita.calameo.com/read/002611183bdbb50b159ff
B	7	Experimental	Allora nelle società e nelle pubblicazioni scientifiche scoppiò una polemica interminabile tra quelli che credevano al fenomeno e gli increduli. La questione accese gli spiriti, i giornalisti di parte scientifica in lotta con gli umoristi versarono fiumi d'inchiostro. La battaglia continuò per sei mesi con alterna fortuna ed esito incerto. Ma a poco a poco l'umorismo sconfisse la scienza e la faccenda del mostro si concluse tra le risate universali. Così nei primi mesi dell'anno l'argomento sembrava ormai dimenticato, quando accaddero altri strani fatti che vennero ben presto a conoscenza del pubblico. Allora il fenomeno apparve sotto una luce nuova: non si trattava più di un problema scientifico da risolvere, bensì di un pericolo serio e reale dal quale bisognava difendersi.	J. Verne; Ventimila leghe sotto i mari (J. Verne; Twenty Thousand Leagues Under the Seas) Retrieved from: http://www.nemoischia.it/wp-content/uploads/2013/10/20mlaleghe.pdf

Appendix B

The table illustrates the complete set of unique combinations of stems/nonstems and suffixes/nonsuffixes (N=288) used as experimental stimuli for the adult version of the FPVS-MEG study (Chapter V of this Thesis).

The table header is to be interpreted as follows: “(Non)stem” refers to the stem (or nonstem) used as first constituent of the pseudoword. “(Non)suffix” refers to the suffix (or nonsuffix) used as second constituent of the pseudoword. “(Non)stem ID” and “(Non)suffix ID” are the unique identifiers for (non)stems and (non)suffixes, used for the creation of semi-randomised lists. “Stimulus” is the pseudoword combination used as stimulus. “Stimulus ID” is the unique identifier of the stimulus, resulting from the two constituents’ IDs and from a description of the combination used for the stimulus. “Global old20” refers to the stimulus’ old20 metric, as obtained from SUBTLEX-UK (Van Heuven et al., 2014), using the old20 function from the R package *vwr* (Keuleers, 2013).

Colour code: light blue = stem+suffix combinations (N=72); light green = stem+nonsuffix combinations (N=72); yellow = nonstem+suffix combinations (N=72); orange = nonstem+nonsuffix combinations (N=72).

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
soft	ity	2	1	softity	0201_StemSuffix	2.50
ship	ity	4	1	shipity	0401_StemSuffix	2.25
stop	ity	5	1	stopity	0501_StemSuffix	2.65
hold	ity	6	1	holdity	0601_StemSuffix	2.30
jump	ity	8	1	jumpity	0801_StemSuffix	2.20
farm	ity	11	1	farmity	1101_StemSuffix	2.10
soft	ous	2	5	softous	0205_StemSuffix	2.75
ship	ous	4	5	shipous	0405_StemSuffix	2.85
stop	ous	5	5	stopous	0505_StemSuffix	2.35
hold	ous	6	5	holdous	0605_StemSuffix	2.45
jump	ous	8	5	jumpous	0805_StemSuffix	2.60
farm	ous	11	5	farmous	1105_StemSuffix	1.95
soft	ful	2	7	softful	0207_StemSuffix	2.85
ship	ful	4	7	shipful	0407_StemSuffix	2.45
stop	ful	5	7	stopful	0507_StemSuffix	2.75
hold	ful	6	7	holdful	0607_StemSuffix	2.60
jump	ful	8	7	jumpful	0807_StemSuffix	2.95
farm	ful	11	7	farmful	1107_StemSuffix	2.40
soft	ite	2	9	softite	0209_StemSuffix	2.50
ship	ite	4	9	shipite	0409_StemSuffix	2.45
stop	ite	5	9	stopite	0509_StemSuffix	2.30
hold	ite	6	9	holdite	0609_StemSuffix	2.05
jump	ite	8	9	jumpite	0809_StemSuffix	2.55
farm	ite	11	9	farmite	1109_StemSuffix	1.95

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
soft	ese	2	11	softese	0211_StemSuffix	2.20
ship	ese	4	11	shipese	0411_StemSuffix	2.45
stop	ese	5	11	stopese	0511_StemSuffix	2.40
hold	ese	6	11	holdese	0611_StemSuffix	2.20
jump	ese	8	11	jumpese	0811_StemSuffix	2.75
farm	ese	11	11	farmese	1111_StemSuffix	1.90
soft	ess	2	12	softess	0212_StemSuffix	2.10
ship	ess	4	12	shipess	0412_StemSuffix	2.00
stop	ess	5	12	stopess	0512_StemSuffix	1.95
hold	ess	6	12	holdess	0612_StemSuffix	1.90
jump	ess	8	12	jumpess	0812_StemSuffix	2.55
farm	ess	11	12	farmess	1112_StemSuffix	1.95
help	ive	1	2	helpive	0102_StemSuffix	2.55
last	ive	3	2	lastive	0302_StemSuffix	1.90
park	ive	7	2	parkive	0702_StemSuffix	1.95
town	ive	9	2	townive	0902_StemSuffix	2.45
bird	ive	10	2	birdive	1002_StemSuffix	2.40
milk	ive	12	2	milkive	1202_StemSuffix	2.25
help	ory	1	3	helpory	0103_StemSuffix	2.75
last	ory	3	3	lastory	0303_StemSuffix	1.95
park	ory	7	3	parkory	0703_StemSuffix	2.20
town	ory	9	3	townory	0903_StemSuffix	2.75
bird	ory	10	3	birdory	1003_StemSuffix	2.75
milk	ory	12	3	milkory	1203_StemSuffix	2.20

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
help	ure	1	4	helpure	0104_StemSuffix	2.80
last	ure	3	4	lasture	0304_StemSuffix	1.95
park	ure	7	4	parkure	0704_StemSuffix	2.35
town	ure	9	4	townure	0904_StemSuffix	2.60
bird	ure	10	4	birdure	1004_StemSuffix	2.65
milk	ure	12	4	milkure	1204_StemSuffix	2.60
help	ise	1	6	helpise	0106_StemSuffix	2.35
last	ise	3	6	lastise	0306_StemSuffix	2.00
park	ise	7	6	parkise	0706_StemSuffix	1.90
town	ise	9	6	townise	0906_StemSuffix	2.30
bird	ise	10	6	birdise	1006_StemSuffix	2.25
milk	ise	12	6	milkise	1206_StemSuffix	2.05
help	ist	1	8	helpist	0108_StemSuffix	2.35
last	ist	3	8	lastist	0308_StemSuffix	1.90
park	ist	7	8	parkist	0708_StemSuffix	1.95
town	ist	9	8	townist	0908_StemSuffix	2.50
bird	ist	10	8	birdist	1008_StemSuffix	2.40
milk	ist	12	8	milkist	1208_StemSuffix	2.45
help	ish	1	10	helpish	0110_StemSuffix	2.15
last	ish	3	10	lastish	0310_StemSuffix	2.00
park	ish	7	10	parkish	0710_StemSuffix	1.85
town	ish	9	10	townish	0910_StemSuffix	2.05
bird	ish	10	10	birdish	1010_StemSuffix	2.10
milk	ish	12	10	milkish	1210_StemSuffix	2.25

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
soft	ert	2	1	softert	0201_StemNonsuffix	2.25
ship	ert	4	1	shipert	0401_StemNonsuffix	2.00
stop	ert	5	1	stopert	0501_StemNonsuffix	2.00
hold	ert	6	1	holdert	0601_StemNonsuffix	1.90
jump	ert	8	1	jumpert	0801_StemNonsuffix	1.95
farm	ert	11	1	farmert	1101_StemNonsuffix	1.80
soft	ald	2	5	softald	0205_StemNonsuffix	2.90
ship	ald	4	5	shipald	0405_StemNonsuffix	2.80
stop	ald	5	5	stopald	0505_StemNonsuffix	2.85
hold	ald	6	5	holdald	0605_StemNonsuffix	2.75
jump	ald	8	5	jumpald	0805_StemNonsuffix	2.95
farm	ald	11	5	farmald	1105_StemNonsuffix	2.20
soft	sal	2	7	softsal	0207_StemNonsuffix	2.90
ship	sal	4	7	shipsal	0407_StemNonsuffix	2.70
stop	sal	5	7	stopsal	0507_StemNonsuffix	2.60
hold	sal	6	7	holdsal	0607_StemNonsuffix	2.80
jump	sal	8	7	jumpsal	0807_StemNonsuffix	2.90
farm	sal	11	7	farmsal	1107_StemNonsuffix	2.75
soft	ene	2	9	softene	0209_StemNonsuffix	2.05
ship	ene	4	9	shipene	0409_StemNonsuffix	2.55
stop	ene	5	9	stopene	0509_StemNonsuffix	2.60
hold	ene	6	9	holdene	0609_StemNonsuffix	1.95
jump	ene	8	9	jumpene	0809_StemNonsuffix	2.65
farm	ene	11	9	farmene	1109_StemNonsuffix	1.95

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
soft	oke	2	11	softoke	0211_StemNonsuffix	2.75
ship	oke	4	11	shipoke	0411_StemNonsuffix	2.75
stop	oke	5	11	stopoke	0511_StemNonsuffix	2.30
hold	oke	6	11	holdoke	0611_StemNonsuffix	2.70
jump	oke	8	11	jumpoke	0811_StemNonsuffix	2.95
farm	oke	11	11	farmoke	1111_StemNonsuffix	2.60
soft	ust	2	12	softust	0212_StemNonsuffix	2.75
ship	ust	4	12	shipust	0412_StemNonsuffix	2.85
stop	ust	5	12	stopust	0512_StemNonsuffix	2.70
hold	ust	6	12	holdust	0612_StemNonsuffix	2.50
jump	ust	8	12	jumpust	0812_StemNonsuffix	2.85
farm	ust	11	12	farmust	1112_StemNonsuffix	2.70
help	une	1	2	helpune	0102_StemNonsuffix	2.80
last	une	3	2	lastune	0302_StemNonsuffix	2.55
park	une	7	2	parkune	0702_StemNonsuffix	2.50
town	une	9	2	townune	0902_StemNonsuffix	2.75
bird	une	10	2	birdune	1002_StemNonsuffix	2.65
milk	une	12	2	milkune	1202_StemNonsuffix	2.65
help	ute	1	3	helpute	0103_StemNonsuffix	2.70
last	ute	3	3	lastute	0303_StemNonsuffix	2.70
park	ute	7	3	parkute	0703_StemNonsuffix	2.60
town	ute	9	3	townute	0903_StemNonsuffix	2.80
bird	ute	10	3	birdute	1003_StemNonsuffix	2.65
milk	ute	12	3	milkute	1203_StemNonsuffix	2.85

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
help	int	1	4	helpint	0104_StemNonsuffix	2.25
last	int	3	4	lastint	0304_StemNonsuffix	1.90
park	int	7	4	parkint	0704_StemNonsuffix	1.85
town	int	9	4	townint	0904_StemNonsuffix	2.40
bird	int	10	4	birdint	1004_StemNonsuffix	1.95
milk	int	12	4	milkint	1204_StemNonsuffix	2.05
help	ere	1	6	helpere	0106_StemNonsuffix	2.20
last	ere	3	6	lastere	0306_StemNonsuffix	1.95
park	ere	7	6	parkere	0706_StemNonsuffix	1.90
town	ere	9	6	townere	0906_StemNonsuffix	1.90
bird	ere	10	6	birdere	1006_StemNonsuffix	1.90
milk	ere	12	6	milkere	1206_StemNonsuffix	1.90
help	arn	1	8	helparn	0108_StemNonsuffix	2.55
last	arn	3	8	lastarn	0308_StemNonsuffix	2.05
park	arn	7	8	parkarn	0708_StemNonsuffix	2.15
town	arn	9	8	townarn	0908_StemNonsuffix	2.80
bird	arn	10	8	birdarn	1008_StemNonsuffix	2.60
milk	arn	12	8	milkarn	1208_StemNonsuffix	2.30
help	ult	1	10	helpult	0110_StemNonsuffix	2.90
last	ult	3	10	lastult	0310_StemNonsuffix	2.90
park	ult	7	10	parkult	0710_StemNonsuffix	2.80
town	ult	9	10	townult	0910_StemNonsuffix	2.90
bird	ult	10	10	birdult	1010_StemNonsuffix	2.95
milk	ult	12	10	milkult	1210_StemNonsuffix	3.00

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
terp	ity	2	1	terpity	0201_NonstemSuffix	2.25
bron	ity	4	1	bronity	0401_NonstemSuffix	2.45
trum	ity	5	1	trumity	0501_NonstemSuffix	2.50
burk	ity	6	1	burkity	0601_NonstemSuffix	2.40
lort	ity	8	1	lortity	0801_NonstemSuffix	2.65
culp	ity	11	1	culpity	1101_NonstemSuffix	2.55
terp	ous	2	5	terpous	0205_NonstemSuffix	2.20
bron	ous	4	5	bronous	0405_NonstemSuffix	2.15
trum	ous	5	5	trumous	0505_NonstemSuffix	2.20
burk	ous	6	5	burkous	0605_NonstemSuffix	2.40
lort	ous	8	5	lortous	0805_NonstemSuffix	1.95
culp	ous	11	5	culpous	1105_NonstemSuffix	2.70
terp	ful	2	7	terpful	0207_NonstemSuffix	2.75
bron	ful	4	7	bronful	0407_NonstemSuffix	2.75
trum	ful	5	7	trumful	0507_NonstemSuffix	2.65
burk	ful	6	7	burkful	0607_NonstemSuffix	2.85
lort	ful	8	7	lortful	0807_NonstemSuffix	2.70
culp	ful	11	7	culpful	1107_NonstemSuffix	2.70
terp	ite	2	9	terpite	0209_NonstemSuffix	2.00
bron	ite	4	9	bronite	0409_NonstemSuffix	1.90
trum	ite	5	9	trumite	0509_NonstemSuffix	2.70
burk	ite	6	9	burkite	0609_NonstemSuffix	2.35
lort	ite	8	9	lortite	0809_NonstemSuffix	2.45
culp	ite	11	9	culpite	1109_NonstemSuffix	2.55

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
terp	ese	2	11	terpese	0211_NonstemSuffix	2.30
bron	ese	4	11	bronese	0411_NonstemSuffix	2.00
trum	ese	5	11	trumese	0511_NonstemSuffix	2.70
burk	ese	6	11	burkese	0611_NonstemSuffix	2.10
lort	ese	8	11	lortese	0811_NonstemSuffix	2.45
culp	ese	11	11	culpese	1111_NonstemSuffix	2.75
terp	ess	2	12	terpess	0212_NonstemSuffix	2.05
bron	ess	4	12	broness	0412_NonstemSuffix	1.95
trum	ess	5	12	trumess	0512_NonstemSuffix	2.40
burk	ess	6	12	burkess	0612_NonstemSuffix	1.80
lort	ess	8	12	lortess	0812_NonstemSuffix	2.00
culp	ess	11	12	culpess	1112_NonstemSuffix	2.70
josk	ive	1	2	joskive	0102_NonstemSuffix	2.80
firn	ive	3	2	firnive	0302_NonstemSuffix	2.65
molp	ive	7	2	molpive	0702_NonstemSuffix	2.70
bemp	ive	9	2	bempive	0902_NonstemSuffix	2.70
jelt	ive	10	2	jeltive	1002_NonstemSuffix	2.40
tand	ive	12	2	tandive	1202_NonstemSuffix	2.50
josk	ory	1	3	joskory	0103_NonstemSuffix	2.95
firn	ory	3	3	firnory	0303_NonstemSuffix	2.85
molp	ory	7	3	molpory	0703_NonstemSuffix	2.80
bemp	ory	9	3	bempory	0903_NonstemSuffix	2.85
jelt	ory	10	3	jeltory	1003_NonstemSuffix	2.95
tand	ory	12	3	tandory	1203_NonstemSuffix	2.10

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
josk	ure	1	4	joskure	0104_NonstemSuffix	2.85
firn	ure	3	4	firnure	0304_NonstemSuffix	2.70
molp	ure	7	4	molpure	0704_NonstemSuffix	2.90
bemp	ure	9	4	bempure	0904_NonstemSuffix	2.70
jelt	ure	10	4	jelture	1004_NonstemSuffix	2.25
tand	ure	12	4	tandure	1204_NonstemSuffix	2.50
josk	ise	1	6	joskise	0106_NonstemSuffix	2.90
firn	ise	3	6	firnise	0306_NonstemSuffix	2.15
molp	ise	7	6	molpise	0706_NonstemSuffix	2.60
bemp	ise	9	6	bempise	0906_NonstemSuffix	2.70
jelt	ise	10	6	jeltise	1006_NonstemSuffix	2.75
tand	ise	12	6	tandise	1206_NonstemSuffix	2.20
josk	ist	1	8	joskist	0108_NonstemSuffix	2.85
firn	ist	3	8	firnist	0308_NonstemSuffix	2.25
molp	ist	7	8	molpist	0708_NonstemSuffix	2.75
bemp	ist	9	8	bempist	0908_NonstemSuffix	2.65
jelt	ist	10	8	jeltist	1008_NonstemSuffix	2.60
tand	ist	12	8	tandist	1208_NonstemSuffix	2.20
josk	ish	1	10	joskish	0110_NonstemSuffix	2.75
firn	ish	3	10	firnish	0310_NonstemSuffix	1.85
molp	ish	7	10	molpish	0710_NonstemSuffix	2.10
bemp	ish	9	10	bempish	0910_NonstemSuffix	2.60
jelt	ish	10	10	jeltish	1010_NonstemSuffix	2.15
tand	ish	12	10	tandish	1210_NonstemSuffix	1.90

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
terp	ert	2	1	terpert	0201_NonstemNonsuffix	2.25
bron	ert	4	1	bronert	0401_NonstemNonsuffix	2.00
trum	ert	5	1	trumert	0501_NonstemNonsuffix	2.30
burk	ert	6	1	burkert	0601_NonstemNonsuffix	1.95
lort	ert	8	1	lortert	0801_NonstemNonsuffix	2.25
culp	ert	11	1	culpert	1101_NonstemNonsuffix	1.95
terp	ald	2	5	terpald	0205_NonstemNonsuffix	2.85
bron	ald	4	5	bronald	0405_NonstemNonsuffix	2.35
trum	ald	5	5	trumald	0505_NonstemNonsuffix	2.85
burk	ald	6	5	burkald	0605_NonstemNonsuffix	2.60
lort	ald	8	5	lortald	0805_NonstemNonsuffix	2.60
culp	ald	11	5	culpald	1105_NonstemNonsuffix	2.80
terp	sal	2	7	terpsal	0207_NonstemNonsuffix	2.75
bron	sal	4	7	bronsal	0407_NonstemNonsuffix	2.50
trum	sal	5	7	trumsal	0507_NonstemNonsuffix	2.85
burk	sal	6	7	burksal	0607_NonstemNonsuffix	2.60
lort	sal	8	7	lortsal	0807_NonstemNonsuffix	2.70
culp	sal	11	7	culpsal	1107_NonstemNonsuffix	2.90
terp	ene	2	9	terpene	0209_NonstemNonsuffix	2.10
bron	ene	4	9	bronene	0409_NonstemNonsuffix	2.40
trum	ene	5	9	trumene	0509_NonstemNonsuffix	2.65
burk	ene	6	9	burkene	0609_NonstemNonsuffix	2.25
lort	ene	8	9	lortene	0809_NonstemNonsuffix	2.05
culp	ene	11	9	culpene	1109_NonstemNonsuffix	2.65

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
terp	oke	2	11	terpoke	0211_NonstemNonsuffix	2.90
bron	oke	4	11	bronoke	0411_NonstemNonsuffix	2.10
trum	oke	5	11	trumoke	0511_NonstemNonsuffix	2.95
burk	oke	6	11	burkoke	0611_NonstemNonsuffix	2.85
lort	oke	8	11	lortoke	0811_NonstemNonsuffix	2.90
culp	oke	11	11	culpoke	1111_NonstemNonsuffix	2.80
terp	ust	2	12	terpust	0212_NonstemNonsuffix	2.75
bron	ust	4	12	bronust	0412_NonstemNonsuffix	2.45
trum	ust	5	12	trumust	0512_NonstemNonsuffix	2.70
burk	ust	6	12	burkust	0612_NonstemNonsuffix	2.65
lort	ust	8	12	lortust	0812_NonstemNonsuffix	2.80
culp	ust	11	12	culpust	1112_NonstemNonsuffix	2.80
josk	une	1	2	joskune	0102_NonstemNonsuffix	2.95
firn	une	3	2	firnune	0302_NonstemNonsuffix	2.90
molp	une	7	2	molpune	0702_NonstemNonsuffix	2.85
bemp	une	9	2	bempune	0902_NonstemNonsuffix	2.85
jelt	une	10	2	jeltune	1002_NonstemNonsuffix	2.70
tand	une	12	2	tandune	1202_NonstemNonsuffix	2.85
josk	ute	1	3	joskute	0103_NonstemNonsuffix	2.90
firn	ute	3	3	firnute	0303_NonstemNonsuffix	2.75
molp	ute	7	3	molpute	0703_NonstemNonsuffix	2.80
bemp	ute	9	3	bempute	0903_NonstemNonsuffix	2.70
jelt	ute	10	3	jeltute	1003_NonstemNonsuffix	2.85
tand	ute	12	3	tandute	1203_NonstemNonsuffix	2.70

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
josk	int	1	4	joskint	0104_NonstemNonsuffix	2.40
firn	int	3	4	firnint	0304_NonstemNonsuffix	2.50
molp	int	7	4	molpint	0704_NonstemNonsuffix	2.65
bemp	int	9	4	bempint	0904_NonstemNonsuffix	2.75
jelt	int	10	4	jeltint	1004_NonstemNonsuffix	2.35
tand	int	12	4	tandint	1204_NonstemNonsuffix	2.00
josk	ere	1	6	joskere	0106_NonstemNonsuffix	2.85
firn	ere	3	6	firnere	0306_NonstemNonsuffix	2.40
molp	ere	7	6	molpere	0706_NonstemNonsuffix	2.70
bemp	ere	9	6	bempere	0906_NonstemNonsuffix	2.15
jelt	ere	10	6	jeltere	1006_NonstemNonsuffix	2.40
tand	ere	12	6	tandere	1206_NonstemNonsuffix	1.95
josk	arn	1	8	joskarn	0108_NonstemNonsuffix	2.95
firn	arn	3	8	firnarn	0308_NonstemNonsuffix	2.80
molp	arn	7	8	molparn	0708_NonstemNonsuffix	2.80
bemp	arn	9	8	bemparn	0908_NonstemNonsuffix	2.80
jelt	arn	10	8	jeltarn	1008_NonstemNonsuffix	2.90
tand	arn	12	8	tandarn	1208_NonstemNonsuffix	2.05
josk	ult	1	10	joskult	0110_NonstemNonsuffix	2.90
firn	ult	3	10	firnult	0310_NonstemNonsuffix	2.95
molp	ult	7	10	molpult	0710_NonstemNonsuffix	2.95
bemp	ult	9	10	bempult	0910_NonstemNonsuffix	3.00
jelt	ult	10	10	jeltult	1010_NonstemNonsuffix	3.00
tand	ult	12	10	tandult	1210_NonstemNonsuffix	2.90

Appendix C

The table illustrates the complete set of unique combinations of stems/nonstems and suffixes/nonsuffixes (N=54) used as experimental stimuli for the child version of the FPVS-MEG study (Chapter V of this Thesis).
The table header is to be interpreted as follows: "(Non)stem" refers to the stem (or nonstem) used as first constituent of the pseudoword. "(Non)suffix" refers to the suffix (or nonsuffix) used as second constituent of the pseudoword. "(Non)stem ID" and "(Non)suffix ID" are the unique identifiers for (non)stems and (non)suffixes, used for the creation of semi-randomised lists. "Stimulus" is the pseudoword combination used as stimulus. "Stimulus ID" is the unique identifier of the stimulus, resulting from the two constituents' IDs and from a description of the combination used for the stimulus. "Global old20" refers to the stimulus' old20 metric, as obtained from SUBTLEX-UK (Van Heuven et al., 2014), using the old20 function from the R package *vwr* (Keuleers, 2013).

Colour code: light blue = stem+suffix combinations (N=18); light green = stem+nonsuffix combinations (N=18); yellow = nonstem+suffix combinations (N=18).

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
soft	ity	2	1	softity	0201_StemSuffix	2.50
ship	ity	4	1	shipity	0401_StemSuffix	2.25
hold	ity	6	1	holdity	0601_StemSuffix	2.30
soft	ous	2	5	softous	0205_StemSuffix	2.75
ship	ous	4	5	shipous	0405_StemSuffix	2.85
hold	ous	6	5	holdous	0605_StemSuffix	2.45
soft	ful	2	7	softful	0207_StemSuffix	2.85
ship	ful	4	7	shipful	0407_StemSuffix	2.45
hold	ful	6	7	holdful	0607_StemSuffix	2.60
town	ory	9	3	townory	0903_StemSuffix	2.75
bird	ory	10	3	birdory	1003_StemSuffix	2.75
milk	ory	12	3	milkory	1203_StemSuffix	2.20
town	ise	9	6	townise	0906_StemSuffix	2.30
bird	ise	10	6	birdise	1006_StemSuffix	2.25
milk	ise	12	6	milkise	1206_StemSuffix	2.05
town	ish	9	10	townish	0910_StemSuffix	2.05
bird	ish	10	10	birdish	1010_StemSuffix	2.10
milk	ish	12	10	milkish	1210_StemSuffix	2.25
soft	ert	2	1	softert	0201_StemNonsuffix	2.25
ship	ert	4	1	shipert	0401_StemNonsuffix	2.00
hold	ert	6	1	holdert	0601_StemNonsuffix	1.90
soft	ald	2	5	softald	0205_StemNonsuffix	2.90
ship	ald	4	5	shipald	0405_StemNonsuffix	2.80
hold	ald	6	5	holdald	0605_StemNonsuffix	2.75

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
soft	sal	2	7	softsal	0207_StemNonsuffix	2.90
ship	sal	4	7	shipsal	0407_StemNonsuffix	2.70
hold	sal	6	7	holdsal	0607_StemNonsuffix	2.80
town	ute	9	3	townute	0903_StemNonsuffix	2.80
bird	ute	10	3	birdute	1003_StemNonsuffix	2.65
milk	ute	12	3	milkute	1203_StemNonsuffix	2.85
town	ere	9	6	townere	0906_StemNonsuffix	1.90
bird	ere	10	6	birdere	1006_StemNonsuffix	1.90
milk	ere	12	6	milkere	1206_StemNonsuffix	1.90
town	ult	9	10	townult	0910_StemNonsuffix	2.90
bird	ult	10	10	birdult	1010_StemNonsuffix	2.95
milk	ult	12	10	milkult	1210_StemNonsuffix	3.00
terp	ity	2	1	terpity	0201_NonstemSuffix	2.25
bron	ity	4	1	bronity	0401_NonstemSuffix	2.45
burk	ity	6	1	burkity	0601_NonstemSuffix	2.40
terp	ous	2	5	terpous	0205_NonstemSuffix	2.20
bron	ous	4	5	bronous	0405_NonstemSuffix	2.15
burk	ous	6	5	burkous	0605_NonstemSuffix	2.40
terp	ful	2	7	terpful	0207_NonstemSuffix	2.75
bron	ful	4	7	bronful	0407_NonstemSuffix	2.75
burk	ful	6	7	burkful	0607_NonstemSuffix	2.85
bemp	ory	9	3	bempory	0903_NonstemSuffix	2.85
jelt	ory	10	3	jeltory	1003_NonstemSuffix	2.95
tand	ory	12	3	tandory	1203_NonstemSuffix	2.10
bemp	ise	9	6	bempise	0906_NonstemSuffix	2.70
jelt	ise	10	6	jeltise	1006_NonstemSuffix	2.75

(Non)stem	(Non)suffix	(Non)stem ID	(Non)suffix ID	Stimulus	Stimulus ID	Global old20
tand	ise	12	6	tandise	1206_NonstemSuffix	2.20
bemp	ish	9	10	bempish	0910_NonstemSuffix	2.60
jelt	ish	10	10	jeltish	1010_NonstemSuffix	2.15
tand	ish	12	10	tandish	1210_NonstemSuffix	1.90