Expanding horizons of cross-linguistic research on reading:

The Multilingual Eye-movement Corpus (MECO)

Noam Siegelman[1], Sascha Schroeder[2], Cengiz Acartürk[3], Hee-Don Ahn[4], Svetlana Alexeeva[5], Simona Amenta[6], Raymond Bertram[7], Rolando Bonandrini[8], Marc Brysbaert[9], Daria Chernova[5], Sara Maria Da Fonseca[10], Nicolas Dirix[9], Wouter Duyck[9], Argyro Fella[11], Ram Frost[12], Carolina A. Gattei[13,14,15], Areti Kalaitzi[10], Nayoung Kwon[16], Kaidi Lõo[17], Marco Marelli[8], Timothy C. Papadopoulos[18], Athanassios Protopapas[10], Satu Savo[7], Diego E. Shalom[13,14], Natalia Slioussar[19,5], Roni Stein[12], Longjiao Sui[9], Analí Taboh[13,14], Veronica Tønnesen[10], Kerem Alp Usal[3], and Victor Kuperman[20]

[1]Haskins Laboratories [2]University of Goettingen [3]Middle East Technical University [4]Konkuk University [5]Saint Petersburg State University [6]University of Trento [7]University of Turku [8]University of Milano-Bicocca [9]Ghent University [10]University of Oslo [11]University of Nicosia [12]The Hebrew University [13]Universidad de Buenos Aires [14]Universidad Torcuato di Tella [15]Pontificia Universidad Católica Argentina [16]University of Oregon [17]University of Tartu [18]University of Cyprus [19]Higher School of Economics (HSE) Moscow [20]McMaster University

Corresponding Author:
Noam Siegelman
Haskins Laboratories
300 George Street, Suite #900
New Haven, CT, USA, 06511
E-mail: noam.siegelman@yale.edu

Abstract

Scientific studies of language behavior need to grapple with a large diversity of languages in the world and, for reading, a further variability in writing systems. Yet, the ability to form meaningful theories of reading is contingent on the availability of cross-linguistic behavioral data. This paper offers new insights into aspects of reading behavior that are shared and those that vary systematically across languages through an investigation of eye-tracking data from 13 languages recorded during text reading. We begin with reporting a bibliometric analysis of eye-tracking studies showing that the current empirical base is insufficient for cross-linguistic comparisons. We respond to this empirical lacuna by presenting the Multilingual Eye-Movement Corpus (MECO), the product of an international multi-lab collaboration. We examine which behavioral indices differentiate between reading in written languages, and which measures are stable across languages. One of the findings is that readers of different languages vary considerably in their skipping rate (i.e., the likelihood of not fixating on a word even once) and that this variability is explained by cross-linguistic differences in word length distributions. In contrast, if readers do not skip a word, they tend to spend a similar average time viewing it. We outline the implications of these findings for theories of reading. We also describe prospective uses of the publicly available MECO data, and its further development plans.

Keywords: reading, eye-tracking, cross-linguistic research, language.

Any field of research in human cognition must account for natural variability in physiological, psychological, and behavioral traits and states of individuals. Few fields, however, also need to account for the profound and inherent variability in the very object of cognitive processing. A prime example of such a field is the study of language. A generalizable account of how language is learned, produced, comprehended, or represented in the brain or mind also needs to grapple with the world's astounding diversity of languages. In the case of *reading*, this diversity is further compounded by the variability of orthographies, i.e., solutions developed for representing speech in print (Daniels & Bright, 1996; Daniels & Share, 2018). Thus, one of the central goals of reading research is to find what universal and specific aspects exist across the written languages of the world, and subsequently, to study how these aspects influence reading development and processes (for recent reviews see, among others, Frost, 2012; Koda & Zehler, 2008; Share, 2014; Verhoeven & Perfetti, 2017). This goal brings forward extensive demands on the quantity and quality of empirical evidence and, importantly, its cross-linguistic coverage, which is not always guaranteed in an Anglo-centric scientific literature on language (Share, 2014).

It is uncontroversial that the availability of high-quality, comparable behavioral data from diverse languages and writing systems is both a driving engine and a prerequisite of meaningful and generalizable theories of reading. The history of reading research shows that the field has been propelled greatly by data that came from cross-linguistic multi-lab coordinated efforts. Consider, for instance, the Ziegler and Goswami's (2005) influential *psycholinguistic grain size* theory – a proposal that languages with inconsistent (opaque) orthographies (e.g., English) are more difficult to learn and are preferentially learned via bigger orthographic chunks than relatively consistent transparent languages (e.g., Finnish). This proposal draws on several

multilingual studies, including in particular a joint investigation of real word and non-word reading in 13 European alphabetic languages (Seymour, Aro, & Erskine, 2003).

Most research producing either cross-linguistic data or comparable single-language data so far has employed tasks revolving around single word recognition (e.g., the English Lexicon Project database of lexical decision and word naming by Balota et al., 2007[1]). Yet proficient natural reading is the reading of continuous texts to achieve comprehension, i.e., building a mental representation of the text content in one's memory and integrating it with one's prior knowledge through inferential processing (e.g., Wooley, 2011). This set of highly coordinated cognitive operations necessarily includes, but also goes far beyond, identification of individual words in the text in terms of complexity and breadth of demands on the visuo-oculomotor, perceptual, and information-processing systems in the reader (e.g., Liversedge et al., 2012; Rayner & Liversedge, 2011). For such higher-level language processing, such cross-linguistic data is a lot less evident and barely available.

In line with the goal of studying natural real-time behavior during reading for comprehension, in this study, we focus on silent reading of running texts, using eye-tracking as the experimental paradigm. Eye-tracking is the registration of eye-movements as they unfold in real-time, and its output is a demonstrably reliable and ecologically valid record of reading behavior (Kliegl et al., 2006; Rayner, 1998; Rayner et al., 2012). A rich literature shows that eye movement control is an integral part of information processing that takes place during reading (see review in Radach & Kennedy, 2013), and thus, it is reflective both of the cognitive processes of comprehension and the multiple components that underlie those processes (e.g.,

[1] The English Lexicon Project also pioneered a type of large-scale multi-lab data collection resulting in a series of mega-studies in multiple languages (see Keuleers & Balota, 2015 for review). An up-to-date list of relevant resources is maintained at http://crr.ugent.be/programs-data/megastudy-data-available.

Kennedy et al., 2000; Rayner et al., 2006; Rayner et al., 2012). One of the important advantages of eye-tracking is that it enables a fine-grained real-time account of both the temporal (when) and spatial (where) aspects of text reading. The *when* of eye movement control determines how long to fixate on a word with the eye gaze, allowing for viewing and uptake of visual and linguistic information, and when to break the fixation and initiate a saccadic movement to another location. The *where* aspect relates to decisions of which word to select as a target for the next fixation and which to skip, and what amplitude of a saccadic oculomotor movement to generate to attain this target (Radach et al., 2007; Rayner, 1998). Given vast differences in the surface characteristics of (written) languages of the world, one can expect readers of different languages to systematically vary in both the temporal and spatial dimensions of their reading behavior. An examination of such systematic patterns requires a resource of comparable eye-tracking reading data across languages.

Out of thousands of experimental studies using eye-tracking (see below), very few addressed this need for cross-linguistic comparison. One of these seminal exceptions is an eye-tracking study by Liversedge et al. (2016), which examined the eye-movements of native speakers reading closely matched written passages in three languages (Chinese, English and Finnish) representing widely different language families and writing systems. Other studies provided corpora with comparable cross-linguistic eye-tracking data in two languages. Such studies include the Dundee corpus of texts read in English and French (Pynte & Kennedy, 2006); the GECO corpus of eye-movements (Cop et al., 2017) collected from English and Dutch participants reading the same book in the original and translated version; and the Whitford and Titone's (2012) study of English-French bilinguals reading passages in both languages (see also English and German comparative data in Rau et al., 2015, and Chinese and English data in Sun

& Feng, 1999 and Feng et al., 2009). Several additional studies offer monolingual databases of eye-tracking data, including, among others, corpora in Chinese (Pan et al., 2021), English (Frank et al., 2013; Luke & Christianson, 2018), German (Kliegl et al., 2004), Hindi (Husain et al., 2014), and Russian (Laurinavichyute et al., 2019). As we show below, these and similar studies are relatively limited from the viewpoint of cross-linguistic coverage. They heavily gravitate – in line with the trend in the entire field of language research – towards alphabetic languages of Europe and especially English (Share, 2008; 2014).

Moreover, whereas all of the above studies aimed to specifically compare reading in a small number of target languages, our goal here was, for the first time, to generate a database of reading behavior across a much larger number of languages and writing systems. This database was collected using similar technology and analyzed with unified software from comparable populations of readers exposed to comparable textual stimuli. The current work thus builds upon the comparative studies cited above and extends them to investigate eye movements during reading across multiple languages.

The structure of the paper is as follows. Part I is a bibliometric analysis of scholarly publications on eye-movements in reading. We review the data available for various languages and the studies providing primary data on more than one language. Part II describes the Multilingual Eye-movement COrpus or MECO, the product of an international multi-lab collaboration of research groups in 13 countries. The goal of MECO is to supply theories of reading with primary behavioral data from a large number of diverse writing and linguistic systems. The resulting data are made freely available to empirically address a range of research questions about reading across a wide variety of languages. In Part II we also address the technological, methodological and experimental decisions that went into this corpus creation.

Part III uses MECO data to directly tackle the key theoretical goal of reading research (addressed in Liversedge et al., 2016, among others) and of this paper: quantifying similarities and differences in reading behavior across a variety of written languages. These analyses offer new insights into aspects of behavior that are shared and those that vary systematically across languages. In the General Discussion, we summarize our findings and outline limitations and plans for MECO's further development.

Part I: Bibliometric analysis of cross-linguistic reading research

To estimate the cross-linguistic coverage of studies of reading that use eye-tracking, we conducted a bibliometric analysis of 1078 papers (published from 2000 to 2018) in the Web of Science citation database[2], which were manually coded for the investigated language(s). Note that our search should not be taken as an exhaustive list, nor does it follow the accepted protocols for meta-analyses (Moher et al., 2015). Rather, it is meant to provide an estimate of the current state of the field based on a large number of papers published over the last two decades. The full bibliometric database is available at the project's OSF page (see *Data availability* section in Part II, below).

Figure 1 presents the distribution of studied languages across the 1078 papers. Note that some studies included more than one language (see below), and therefore the sum of this distribution is larger than the number of studies. Perhaps unsurprisingly, Figure 1 points to

[2] First, the following search parameters were used: TOPIC: ("reading" AND ("eye tracking" OR "eye movements")), Refined by: DOCUMENT TYPES: (ARTICLE OR REVIEW OR PROCEEDINGS PAPER); Timespan: 2000–2018; Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ES. Which returned 1956 results. Then, we manually removed papers on topics unrelated to reading of written materials (e.g., reading of emotions), papers without eye-tracking data (i.e., conducted using other paradigms) or those not reporting primary empirical data (e.g., reviews, meta-analyses).

English as the most studied language, accounting for the majority of the eye-tracking research on reading (studied in 620/1078 papers, 57.5%). Other languages with a prevalence of more than 1% of the total (i.e., 11/1078 studies or more) are (in descending order): Chinese (11%), German (9.7%), French (5.2%), Spanish (4.1%), Finnish (3.9%), Dutch (3.3%), Italian (2.1%), Japanese (1.5%) and Korean (1%). All other languages combined appear in only 5.6% of total publications. These comprise a total of 18 languages: Hebrew, Swedish, Thai (7 studies each), Arabic, Portuguese, Russian (4), Polish (3), Afrikaans, Serbo-Croatian (2), Catalan, Croatian, Greek, IsiZulu, Norwegian, Persian, Romanian, Sesotho, Uighur, and Urdu (1). Together, these results show that in the last two decades, most available data on eye movements in reading has come from English, in line with Share's (2008) criticism. With a laudable exception of Chinese and (in a much more limited way) Japanese and Korean, there is a strong bias in the field towards Indo-European languages, in line with Share's (2014) critical observation. This bias poses a serious question on the generality of any theory mainly built on data from Indo-European languages. In sum, at present, the scientific community has access to little or no eye-tracking reading data from the vast majority of the world's languages and writing systems.

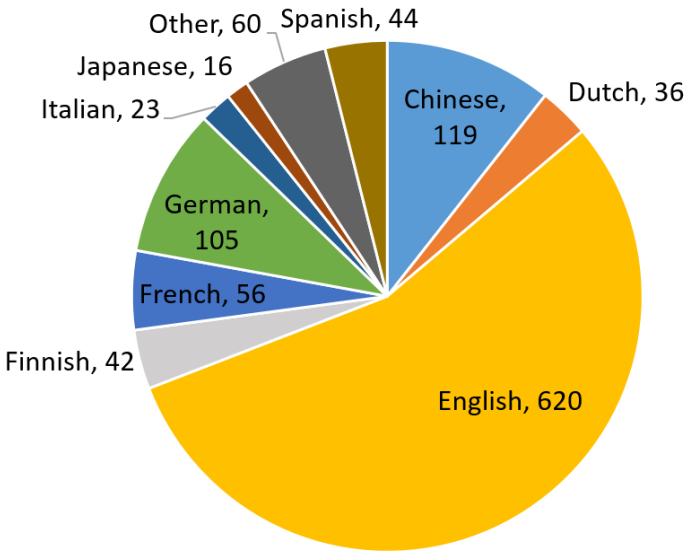

Figure 1. Distribution of investigated languages in 1078 (2000-2018) publications on eye-movements in reading.

Next, we estimated the presence of coordinated cross-linguistic studies. We found that the vast majority of studies in our bibliometric database examined only one language: 1038/1078 of studies with primary data. In other words, only 40 out of 1078 studies in the database (3.7%) conducted a direct cross-linguistic comparison. From this set of studies, 37 studies included data from two languages, and only 3 had data from three languages (Fukuda & Fukuda, 2009; Liversedge et al., 2016; Saggara & Ellis, 2013). No studies in our database report data from four languages or more.

Clearly, reading research does not have a sufficient empirical basis for investigating reading for comprehension across languages, neither in the diversity and number of represented languages nor in the availability of comparative cross-linguistic studies. Part II addresses this deficit by reporting MECO, a coordinated eye-tracking study of reading in multiple diverse languages, designed specifically for cross-linguistic comparisons.

Part II: Corpus structure and descriptive statistics

Investigated languages. Table 1 presents the languages included in the current release of MECO. At present, MECO includes samples from a total of 13 languages, selected due to the availability of partner labs, which will be complemented in the future by further contributing researchers. Table 1 lays out the diversity of the investigated languages in terms of their typological classes and genetic groups, as well as scripts, morphological types, and orthographic transparency (as classified in Dryer & Haspelmath, 2013; Seymour et al., 2003; Verhoeven & Perfetti, 2017). It also shows that many of the presently reported languages are under-studied: More than half (7/13) of the languages have an estimated prevalence of 1% or less in previous eye-tracking research, as reflected in the bibliometric search of Part I above. The present database, therefore, constitutes a considerable extension of the existing empirical data pool.

Table 1. Investigated languages and their properties.

Language	Language code	Typological Family (Branch)	Script (Script Type)	Morphological Typology	Orthographic Transparency	% of studies (2000-2018)
Dutch	DU	Indo-European (West Germanic)	Latin (alphabetic)	Synthetic, Fusional	Moderate	3.3
English	EN	Indo-European (West Germanic)	Latin (alphabetic)	Moderately Analytic	Opaque	57.5
Estonian	EE	Uralic (Finnic)	Latin (alphabetic)	Agglutinative, Fusional	Transparent	<1
Finnish	FI	Uralic (Finnic)	Latin (alphabetic)	Agglutinative, Fusional	Transparent	3.9
German	GE	Indo-European (West Germanic)	Latin (alphabetic)	Synthetic, Fusional	Moderate	9.7
Greek	GR	Indo-European (Hellenic)	Greek (alphabetic)	Synthetic, Fusional	Transparent	<1
Hebrew	HE	Semitic (North-Western Semitic)	Hebrew (abjad)	Synthetic, Fusional Semitic Morphology	Opaque	<1
Italian	IT	Indo-European (Romance)	Latin (alphabetic)	Synthetic, Fusional	Transparent	2.1
Korean	KO	Koreanic	Hangul (alphabetic)	Agglutinative	Moderate	1.0
Norwegian	NO	Indo-European (North Germanic)	Latin (alphabetic)	Synthetic, Fusional	Moderate	<1
Russian	RU	Indo-European (East Slavic)	Cyrillic (alphabetic)	Synthetic, Fusional	Moderate	<1
Spanish	SP	Indo-European (Romance)	Latin (alphabetic)	Synthetic, Fusional	Transparent	4.1
Turkish	TR	Turkic (Oghuz)	Latin (alphabetic)	Agglutinative	Transparent	<1

Notes: *% studied languages*: The estimated portion of studied languages based on the bibliometric search reported in Part I.

Participants All participating laboratories aimed to reach n=45-55 participants with usable data (see *Data editing and cleaning* below for details regarding inclusion of participants and trials), and indeed the presently available and reported data sets in most languages reached this range. In some laboratories, however, the final stages of data collection were cut short by COVID-19 related closures, therefore in two languages, the samples are smaller (n~30 each): we plan to increase these samples in the future releases of the MECO project, see Future Directions. Table 2 lists the number of participants per site, the country and institution where the data was collected, and details regarding the participants' compensation. Table 2 also includes summaries of some of the basic background information collected using the Language Experience and Proficiency Questionnaire (LEAP-Q; see *Additional questionnaires and tests*, below). This information includes age, years of education, and self-ratings of L1 proficiency in speaking, oral comprehension, and reading. Participants' full demographic information is available at the project's OSF page (see *Data availability* below). The ethics clearance was obtained by each participating site from the ethics research board of the corresponding institution or country.

Table 2. Information regarding participants in the participating sites.

Language	n	Mean Age (range)	Mean Years of Education (SD)	Mean Self-rating: speaking (SD)	Mean Self-rating: oral comp (SD)	Mean Self-rating: reading (SD)	Country	Institute	Participants' compensation	Trials after trimming, %	Data points after trimming
Dutch	45	22.69 (19-30)	16.12 (2.81)	9.47 (0.69)	9.56 (0.62)	9.6 (0.58)	Belgium	Ghent University	10 Euro/hour	67	66075
English	46	21.04 (18-28)	15.76 (1.7)	10 (0)	10 (0)	10 (0)	Canada	McMaster University	20 CAD/hour or course credit	87	83246
Estonian	52	22.23 (18-30)	14.51 (2.56)	9.31 (0.90)	9.64 (0.56)	9.46 (0.79)	Estonia	University of Tartu	Gift card worth 7.5 Euro/hour	74	58249
Finnish	49	24.29 (19-35)	15.04 (2.71)	9.67 (0.59)	9.84 (0.47)	9.82 (0.44)	Finland	University of Turku	Course credit or 2 movie tickets	91	64673
German	45	23.76 (18-39)	15.88 (2.75)	9.5 (0.69)	9.59 (0.63)	9.41 (0.72)	Germany	University of Goettingen	10 Euro/hour or course credit	83	74096
Greek	45	22.84 (18-30)	17.04 (2.5)	9 (0.88)	9.67 (0.6)	9.73 (0.58)	Cyprus	University of Cyprus	10 Euro/hour or course credit	66	60382
Hebrew	47	24.04 (18-29)	12.82 (1.37)	9.68 (0.56)	9.79 (0.41)	9.6 (0.54)	Israel	Hebrew University	40 NIS/hour or course credit	72	64786
Italian	54	22.83 (19-30)	16.72 (2.15)	9.59 (0.71)	9.76 (0.55)	9.76 (0.51)	Italy	University of Milano-Bicocca	15 Euro or course credit	76	84976
Korean	32	21.97 (19-25)	12.98 (2.13)	8.53 (1.5)	8.78 (1.31)	8.69 (1.09)	South Korea	Konkuk University	10,000 KRW	62	34685
Norwegian	42	25.69 (19-30)	15.33 (3.27)	9.31 (1.7)	9.33 (1.6)	9.21 (1.7)	Norway	University of Oslo	Volunteers	71	61548
Russian	46	24.26 (18-45)	15.45 (2.06)	9.38 (1.41)	9.69 (1.08)	9.46 (1.47)	Russia	St. Petersburg State University	Course credit/volunteers	81	67094
Spanish	48	23.04 (18-30)	19.48 (3.8)	9.73 (0.61)	9.73 (0.64)	9.58 (0.79)	Argentina	Universidad Torcuato Di Tella	8 USD	75	84942
Turkish	29	23.69 (20-29)	17.34 (2.38)	9.41 (0.73)	9.66 (0.61)	9.34 (1.59)	Turkey	Middle East Technical University	50 Turkish Liras	64	31065

Materials At each site, participants read a set of 12 texts in their first and dominant language (L1). All texts were Wikipedia-style encyclopedic entries on a variety of topics, including historical figures, events, and natural or social phenomena. Topics were chosen such that they did not rely on specialized academic knowledge and did not have a specific cultural bias making them more or less familiar to some of the participating sites. At the first stage, texts were created in English, loosely based on the Wikipedia entries. Five of the 12 texts (44 sentences in total) were chosen to serve as sources for translation. These texts were translated to the corresponding L1 from the English original by the team at each site to create translation-equivalents across all languages. The quality of translation and content similarity were ensured through back-translation from the target language to English (never done by the same person who produced the original translation) and an iterative process of introducing changes to the text in L1 and a subsequent back-translation. In a few cases, when the authors' team had professional translators with native knowledge of both English and the target language, they would evaluate the translation (made by a different person) directly in the target language. In this situation, the iterative process of aligning the source and target texts omitted the intermediate step of back-translating. See below for a quantitative evaluation of translation quality of texts across languages.

The remaining seven texts were not translated. Instead, participating sites were instructed to use non-matched texts on the same topic as English originals (e.g., country flags, beekeeping), in the same prosaic genre (i.e., encyclopedic entries), of similar length (5-12 sentences, 10-15 lines) and of comparable level of difficulty (e.g., by avoiding uncommon grammatical

constructions)[3]. These texts were typically compiled by each team using Wikipedia or similar open resources in the corresponding L1. Below we will show that the language-original texts, as far as we can attest, are similar to the English-translated texts in terms of their complexity and readability. Still, we provide English back-translations of all texts used so that users of MECO can further evaluate these and other text properties and potentially decide to focus on particular texts in their analyses based on these characteristics.

To evaluate the quality of translations in matched texts and ensure that there were no systematic differences in text readability or complexity across sites, the authors' team in each site prepared back-translations into English for all texts in all languages. Note that we had to use back-translations to estimate complexity/readability and translation equivalence because, at present, there are no (comparable) computational tools that can estimate these text-level metrics for all MECO languages. First, to estimate complexity, we tested comparability of both matched and unmatched texts across the languages in terms of readability and complexity using the back-translations. As reported in Supplementary Materials S1, a set of 10 readability and complexity metrics did not differ statistically across languages both in matched and unmatched texts. This finding suggests that the texts' complexity/readability were similar across sites and eliminates these factors as potential confounds. Second, to estimate the translation quality of the English-original texts, we quantified the text-wise cosine semantic similarity between back-translations and the English originals (using pre-trained Latent Semantic Analysis (LSA) vectors). This analysis revealed that back-translated matched texts were highly similar to the English originals (mean cosine = 0.88), significantly more than the similarity of unmatched texts to the unmatched

[3] Due to unavailability of materials, there were three cases where sites used texts on different topics than the English originals in the unmatched condition: one text in Norwegian (the text about "Shaka" changed to a text about "Coat of Arms"), and three in Turkish (the text about "Monocle" changed to a text about "Telescope"; the text about "Orange Juice" changed to a text about "Pomegranate Syrup", and the text about “World Environment Day” to that about “World Health Day”).

originals (mean cosine = 0.66, $p < .001$), and statistically on par with the similarity of back-translated Finnish texts to the English originals in the study of Liversedge et al. (2016; mean cosine = 0.93; $p > .1$; see Supplementary Materials S2 for details). Please refer to the project's OSF repository to access all back translations along with the estimates of readability/complexity and similarity to the English originals.

As a general point, we note that the decision to make some of the texts translated and others more loosely related was motivated by three considerations. First, this step ensured that the materials represent a wider natural variety of orthographic, morphological, and syntactic constructions in each language, which is not constrained by the demands of translation accuracy and sentence-by-sentence alignment of content across materials in the corpus. Second, we expected a greater diversity of texts to give rise to greater variability in individual reading strategies and patterns, which is desirable for characterizing natural reading behavior within and across languages.

A third consideration was to enable a direct investigation of a methodological issue in cross-linguistic research, namely, what degree of control over materials is needed to make a cross-linguistic comparison meaningful (see also Papadopoulos et al., 2021). It is clear that some types of analyses require a close matching of semantic equivalence across languages through translation (e.g., whether sentences with the same meaning require the same time to read in different languages, Liversedge et al., 2016). Yet it is still unclear whether, and to what extent, cross-linguistic differences in the global text contents influence eye-movement patterns over and above other well-known factors at the level of characters, morphemes, words and larger multi-word units (e.g., Schuster et al., 2016). The importance of this point is hard to overestimate. A radical methodological stand on this issue may be that any credibly cross-linguistic comparison

of oculomotor patterns must be based on semantically matched texts; otherwise, the diverging semantics of texts in different languages would present a confound. This view would invalidate virtually all existing knowledge of cross-linguistic differences in reading because only a very small portion of prior work is based on translated texts (see above). An alternative stand, however, is that semantic similarity is required only when investigating particular effects of interest and that other cross-linguistic differences in oculomotor behavior generalize regardless of the specific contents of the text[4]. By including both semantically matched and unmatched cross-linguistic materials in the design, MECO enables researchers to examine whether various comparative effects of interest generalize beyond the tight semantic control and are thus more representative of reading behavior in general (also see examples in Part III below).

Each text was followed by four yes/no comprehension questions: these were simple questions that tapped into factual knowledge obtained from the read materials and served as an attention check. The comprehension questions were similar in content across languages in matched texts, but naturally differed for non-matched texts reflecting the differences in the text content. Table 3 details the number of words and sentences in each text in each language. A word is defined in this study as a unit in writing separated by a space.

[4] Consider, for instance, a robust finding that readers of Chinese make shorter saccades than Hebrew readers and both make shorter saccades than readers of English: this fact is related to the greater spatial density of linguistic information in Chinese logographs, followed by less dense Hebrew letters and even less dense letters of the Roman alphabet (Rayner, 2009). Should this behavioral finding be trusted given that they were achieved from texts that differed in their meaning across languages? One position would be to deny the credibility of this finding because of a possible confound (differences in meaning contribute to saccade lengths in unknown ways); another would be to test the scope of the influence that the meaning of a text exercises on the eye-movement control during text reading.

Table 3. Number of sentences (#sent) and words (#word) in each text across languages.

Text #	Topic		DU	EE	EN	FI	GE	GR	HE	IT	KO	NO	RU	SP	TR
1*	Janus	#sent	10	10	10	10	10	10	10	10	10	10	10	9	10
		#word	186	131	183	128	174	189	130	185	142	177	151	210	146
2	Shaka	#sent	7	9	6	8	9	6	11	7	7	8	7	7	7
		#word	194	133	185	116	161	171	209	174	150	169	145	190	131
3*	Doping	#sent	9	9	9	10	9	9	9	9	9	9	9	9	9
		#word	185	137	187	143	190	179	151	217	167	176	155	238	156
4	Thylacine	#sent	12	13	9	7	11	9	10	6	9	9	9	7	9
		#word	206	142	182	130	180	177	168	176	158	181	190	169	167
5	World Environment Day	#sent	11	11	8	11	11	8	9	5	8	8	9	5	7
		#word	173	147	167	127	154	180	168	137	160	158	139	182	139
6	Monocle	#sent	7	10	8	8	11	8	9	8	9	8	9	11	10
		#word	180	126	152	97	153	143	151	150	143	149	165	212	142
7*	Wine Tasting	#sent	8	8	8	8	9	9	8	8	9	9	9	8	8
		#word	213	156	199	135	199	202	164	212	167	189	165	229	150
8	Orange Juice	#sent	10	7	6	8	9	6	14	7	7	11	7	7	10
		#word	161	102	136	103	132	134	165	160	130	171	150	159	126
9	Beekeeping	#sent	10	9	8	11	9	7	10	6	9	16	11	10	9
		#word	181	107	200	128	171	176	173	164	152	243	150	188	149
10	National Flag	#sent	11	10	11	13	11	11	15	8	8	11	11	10	9
		#word	187	109	180	149	181	181	201	176	168	177	164	234	127
11*	International Union for Conservation of Nature	#sent	9	8	8	8	8	8	8	8	8	9	8	7	8
		#word	196	132	176	120	172	181	140	182	125	170	164	225	139
12*	Vehicle Registration Plate	#sent	8	8	8	8	8	8	8	8	8	8	8	8	8
		#word	169	118	162	111	160	170	130	181	134	146	156	176	125

Notes: #sent: number of sentences; #word: number of words. Translated texts are marked with an asterisk, other texts were language-specific. Note that some small deviations in the number of sentences per text in matched texts are due to differences in spelling conventions (e.g., using colon or period before "For example").

Additional questionnaires and tests In addition to the reading task, participants at all sites completed a battery of individual differences tests and questionnaires. Two identical

instruments were used in all sites: (1) The non-verbal IQ test from the Culture Fair Test-3 (CFT20, Subset 3 Matrices, short version, Form A, timed at 3 minutes, Weiß, 2006), and (2) an abridged version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007). The CFT20 aimed at providing a comparable measure of non-verbal intelligence across all sites, and the LEAP-Q at collecting basic demographic and linguistic information about participants.

Furthermore, each site used a short battery of (non-eye-tracking) measures of individual differences in L1 reading and proficiency. The goal in collecting these additional measures was to enable correlational analyses of the relations between individual differences in component skills of reading and oculomotor reading behavior within samples. Given the variability in what individual-differences tests are available for specific languages, the tasks were not identical across sites. Most commonly, the tests examined participants' vocabulary size, word and pseudoword naming, phonological/morphological awareness, and other component skills of reading. The full individual-differences data from each site, along with short task descriptions, are available at the project's OSF page (see *Data availability*).

Procedure In all sites, the experimental session began with participants signing a consent form and filling out the LEAP-Q questionnaire. Then, participants proceeded to the reading task, during which their eye movements were recorded. Following the reading task, participants took the individual-differences battery, including the CFT-20 and any L1 individual-differences tests. The entire procedure lasted no more than an hour, and breaks were provided when needed.

Note that at the conclusion of the experimental session, participants in all samples (except for South Korean) proceeded to participate in an English-language eye-tracking study. The goal of that study was to create an additional eye-tracking corpus of reading in English as a non-

dominant language, which can be used to examine the L2 reading behavior of participants with different L1s. This additional study is beyond the scope of the current paper and is therefore reported elsewhere (Kuperman et al., 2021).

Apparatus and procedure Information regarding the apparatus used at the different sites and additional settings can be found in Supplementary Material S3. Eye-movements were recorded with an EyeLink Portable Duo, 1000 or 1000+ eye-tracker (SR Research, Kanata, Ontario, Canada) with a sampling rate of 1000 Hz. A chin rest and a head restraint were used to minimize head movements. Calibration was performed using a series of nine fixed targets distributed around the display, followed by a 9-point accuracy test to validate eye position. Stimuli were viewed binocularly, but eye-movement data from only one eye (the right eye in most participants) were analyzed. Prior to the presentation of the trial stimuli, a dot appeared on the monitor screen, slightly to the left (or right, in the case of Hebrew, which is a right-to-left writing system) of the first word in the passage. Once the participant had fixated on it, the trial would begin. This drift check and correction took place at the beginning of each trial, and calibration was monitored by the experimenter throughout and redone if necessary. Each of the 12 texts appeared on a separate screen. Participants were instructed to read the passages silently for comprehension and press the space bar when their reading of a passage was completed. A mono-spaced font with 1.5 spacing was used in the reading task in all languages, with a font size between 16 and 24 points (see Supplementary Materials S3). Due to inevitable differences in equipment (e.g., screen size and type) and the spatial configuration of the participating eye-tracking labs, maintaining an identical font size, distance from the screen, and screen resolution was unfeasible. Instead, we required in each lab that participants were tested in the conditions most comfortable for visual inspection of the reading materials, as established by the prior

practice of these labs and adjustments based on pilot participants (for chosen settings, see Supplementary Materials S3). For reference, in the longest matched text ('wine tasting'), the number of text lines varied from 9 to 14 (M=12.38), with a maximal number of characters per line varying from 93 to 114 (M=105.75), except for in Korean where this number was substantially smaller (60). The 12 texts were presented in the same fixed order in all languages (see text number in Table 3). Each text was followed by four yes/no comprehension questions, each showing on a separate screen one after another. Participants responded by pressing "0" for no or "1" for yes, and their answers were recorded.

Data editing and cleaning. In paragraph reading, there is often a need to correct eye fixation locations and assign fixations to text lines within a passage. This is commonly done using a manual procedure (but see Carr, 2021; Cohen, 2013, Tang et al., 2012). One of our methodological objectives was to maintain high replicability in all aspects of the experimental setup and data analysis, in line with principles of Open Science. For this reason, we opted for automatic correction of fixation locations using the *popEye* software (implemented in R, version 0.6.4, Schroeder, 2019). The *popEye* software is an integrated environment to preprocess and analyze eye-tracking data from reading experiments. During preprocessing, *popEye* assigns fixations to lines, words and letters. For the present study, an algorithm was used in which individual fixations are first grouped into sequences based on their spatial and temporal proximity. In the next step, sequences are assigned to the closest line based on their average horizontal location (see Beymer & Russell, 2005; Carr et al., 2021; Špakov et al., 2019, for similar approaches). Following this automatic procedure, the software's output was visually inspected by members of the research team to assess the quality of the resulting data. This step was necessary but may have introduced subjective judgment. We argue, however, that this

process has fewer "researcher degrees of freedom" (Simons et al., 2011) than an alternative process where fixation alignment is done fully manually.

Trials (texts) where fixations were erroneously assigned to lines (typically due to poor calibration or software failures) were deemed unusable and were removed from the analysis. Then, participants who had less than five usable trials were removed from the analysis altogether. The number and percent of trials retained after data cleaning in each site can be found in Table 2 above. In the current release of MECO, we only report data from usable participants and trials, as determined based on the current version of *popEye*. Note that the amount of usable data, as determined by the current version of *popEye*, comprises approximately 70% of the complete data. This is in line with the estimated upper limit that can be achieved by any automated algorithm using the present setup (see Carr et al., 2021, for a comparison of different line assignment algorithms). Since the *popEye* software is under development and may improve its algorithms for correcting fixation locations, future releases of MECO may supplement the current samples with data from some of the trials or participants that are presently removed (see *Limitations and future directions* in the *General Discussion*). For the analyses below (reported in Part III), we additionally removed data points that showed either very short (< 80 ms) first fixations or very long total fixation times (top 1% of the participant-specific distribution).

Data availability. The current (and first) release of MECO includes full interest-area reports from usable participants and trials as well as full data from individual differences tests and background questionnaires. Additionally, we report data at the passage and sentence level, broken down by participant. We also include the analytical code used for Part III. The data, materials, and code are available at the project's OSF page

https://osf.io/3527a/?view_only=e01ec48ca7db41809ba9e46cda09d5f5 (this link is view-only for reviewing purposes but will be changed in the published version to the open-access link).

Reading variables. In Part III below, we consider a number of variables reflecting oculomotor behavior at the word level during reading. Note that the output of the *popEye* software includes several additional variables not discussed here, including fixation locations and information at the sentence and passage level: For future users of MECO, we provide a description of the various variables included in the database at the project's OSF page. Returning to variables used in Part III below, those defined at the word level included: *skipping*[5] (a binary index of whether the word was fixated at least once during the entire reading of the text (and not only during the first pass), labeled as *skipping*); *first fixation duration* (the duration of the first fixation landing on the word, *firstFixationDuration*); *gaze duration* (the summed duration of fixations on the word in the first pass, i.e., before the gaze leaves it for the first time, *gazeDuration*); *total fixation duration* (the summed duration of all fixations on the word, *totalFixationDuration*); *first run number of fixations* (the number of fixations on a word during the first pass, *nFixationsFirstRun*); *total number of fixations* (number of fixations on a word overall, *nFixationsTotal*); *regression* (a binary index of whether the gaze returned to the word after inspecting further textual material, i.e., to the right of the word in left-to-right orthographies, *regressionIn*); and *re-reading* (a binary index of whether the word elicited fixations after the first pass, i.e., after the gaze left the word for the first time, *rereading*). See Inhoff and Radach (1998) and Rayner (1998) for a detailed discussion of these variables. At the participant level, the following variables were defined: *comprehension accuracy* (percent of

[5] The label "skipping" may be interpreted as a support for the theoretical position that each word is meant to be fixated and that not fixating a word involves a decision, as opposed to the theoretical stance that fixations target the most salient or useful words available to the reader's perception at the moment, see Morrison (1988), Radach, Reilly & Inhoff, (2007). We use the label in an a-theoretical way, simply as a index of whether a word has been fixated or not. We thank a reviewer for raising this point.

correct responses to all 48 questions, *accuracy*), *matched comprehension accuracy* (percent of correct responses to the 20 questions in the 5 translated passages, *accuracyMatched*); and *reading rate* (in words per minute, *readingRate*), as well as mean word-level variables (e.g., participant's mean skipping rate, mean first fixation duration, etc.). While reading rate is not an oculomotor measure and is closely related to total fixation duration (though it additionally accounts for skipped words and saccade durations), we include the variable to ensure comparability of the present data with the cross-linguistic educational and psychological literature using reading rate (see review by Brysbaert, 2019). It also opens the opportunity for researchers to use our materials if they do not have access to an eye-tracker but want to collect information about reading rate and reading comprehension in their lab. Additionally, we used scores from the CFT test of non-verbal intelligence (*cft)*.

The 13-level categorical variable Language was a critical independent variable in all of our analyses. Furthermore, we considered word length in characters as a benchmark predictor of reading. Since all languages in our current corpus use spacing for segmentation, word length was defined as a number of characters between spaces, excluding punctuation marks.

Reliability. Correlational research is pointless without information about the reliability of the variables because the observed correlation between two variables depends on both the theoretical correlation and the reliability of the measured variables. The reliability of the eye-tracking data was estimated in two ways. First, we examined the reliability of the eye-tracking variables at the participant-level. For most variables, this was done by using a split-half procedure where, for each language, we examined the correlation between mean values for 'odd' and 'even' words within a participant. These reliability estimates reflect the extent to which each eye-tracking measure provides a stable measure of individual differences in each language. The

only exception to this procedure was the estimation of reliability for reading rate, which was examined by calculating the Intra-class Correlation Coefficients (ICC) across reading rates from the 12 texts in each language for each participant. Second, we estimated split-half reliability at the word token-level (i.e., the level of individual word occurrences). This was done by examining the correlation between means for 'odd' and 'even' participants within each word token for each language and eye-tracking measure. This metric represents reliability values relevant for word-level investigations (e.g., effects of length or frequency of words)[6]. For both types and for each measure, we computed both raw correlations and the Spearman-Brown corrected values (Spearman, 1910). The latter values reflect reliability estimates for the full sample size of participants/words (rather than for half of the participants/words, which are the bases for calculating uncorrected correlations). The full breakdown of reliability estimates by language is reported in Supplementary Materials S4 and S5 (which includes subject- and word-token level estimates, respectively). Below we provide a description of main findings.

The reliability of eye-tracking measures at the participant level was very high (all corrected-*r*'s > 0.93), as may be expected given the large number of words read by each participant (for related estimates and discussion, see Staub, 2021). Reliability at the word token-level was somewhat lower but still within recommended ranges for most measures and languages (see, for instance, reliabilities in GECO, Cop et al., 2017), with some eye-tracking measures (e.g., total fixation duration, skips, number of fixations) having higher reliability than others (e.g., first fixation duration, re-reading). As expected, reliability at the word token-level was

[6] Note that our word token-level estimates of reliability differ from the estimates provided in the GECO corpus (Cop et al., 2017). Cop et al.'s calculations were based on the word-type level, i.e., they averaged values across all occurrences of a word. Our choice is motivated by the fact that morphological variability of different linguistic systems greatly affects how many tokens are associated with each word type and makes the word type-level reliability less comparable across languages.

somewhat lower for sites with a smaller sample size (see, e.g., estimates for Turkish), but still in all sites the average reliability across measures was high (all mean corrected-*r*'s > 0.7).

In addition to estimates for eye-movement measures, we calculated the reliability of offline participant-level measures that were collected in all sites: CFT scores and comprehension accuracy. The former was estimated using a split-half procedure on available data collected across different languages (as the test was identical across all sites). It was found to be r=0.4 uncorrected and r=0.57 after Spearman-Brown correction. Although these values are far from perfect (which is unsurprising in a short test with only 12 items), they still point to reasonable reliability and therefore suggest that CFT scores can be used (with caution) as a metric of individual differences. In contrast, the reliability estimates for comprehension accuracy (both for all texts and matched texts only, see Supplementary Materials S4) were generally lower, with substantial variability across languages. This is expected: The goal of the comprehension question was not to provide a measure of individual differences but rather to motivate participants to attend to the texts and be used as a group-level metric. These reliability values should be taken as a warning *not* to use comprehension scores from MECO as a proxy of individual differences (at least not in most languages).

Part III – comparative analyses of reading across writing systems

We envision MECO as a resource that can generate and test hypotheses at different degrees of resolution, from a single language to a group of languages. Such groups may be defined genetically, e.g., Germanic or Romance, or typologically, e.g., morphologically agglutinative languages such as Finnish and Turkish. Finally, as demonstrated below, analyses

can be applied to the entire set of languages. Equally, units of linguistic interest to study may vary from a single character or sound to phenomena defined at the passage level. Moreover, researchers will be able to consult the data on the participant level, both within and across languages. As stated in the Introduction, this part of the paper aims to quantify differences and similarities between all 13 languages, promoting the long-standing agenda of cross-linguistic psychological research (see, among others, reviews in Frost, 2012; Liversedge et al., 2016; Verhoeven & Perfetti, 2017). This analysis offers new insights into a key theoretical question in cross-linguistic reading research: What aspects of reading behavior are shared across writing systems, and what aspects are language-specific.

Cross-linguistic variability and similarity in eye-movement behavior

This section provides an overview of reading behavior across languages. To this end, we calculated the mean values of each dependent variable for each participant in each sample. Detailed summaries are available as auxiliary files at the project's OSF page, including a breakdown of each eye-tracking variable by language. We then computed the correlations between behavioral measures of reading calculated from these by-participant means across all languages (Table 4).

Table 4. Correlation table for reading measures across languages (N = 580). Values above the diagonal show Pearson correlations; values below the diagonal show *p* values.

	1	2	3	4	5	6	7	8	9	10
1) skipping		-0.07	-0.40	-0.47	0.75	-0.60	-0.54	-0.06	-0.3	-0.13
2) first fixation duration	0.085		0.82	0.61	-0.45	0.14	0.03	-0.07	-0.04	-0.08
3) gaze duration	<0.001	<0.001		0.71	-0.66	0.67	0.29	-0.08	-0.02	-0.04
4) total fixation duration	<0.001	<0.001	<0.001		-0.87	0.43	0.80	0.42	0.67	0.04
5) reading rate	<0.001	<0.001	<0.001	<0.001		-0.56	-0.78	-0.37	-0.59	-0.08
6) number of fixations: first run	<0.001	0.001	<0.001	<0.001	<0.001		0.45	-0.08	0.00	0.02
7) number of fixations: Total	<0.001	0.413	<0.001	<0.001	<0.001	<0.001		0.57	0.88	0.10
8) regressions in	0.122	0.085	0.048	<0.001	<0.001	0.067	<0.001		0.70	0.05
9) rereading	<0.001	0.327	0.58	<0.001	<0.001	0.956	<0.001	<0.001		0.11
10) cft	0.003	0.046	0.309	0.315	0.068	0.684	0.018	0.211	0.008	

Next, we calculated the means and standard errors for all eye-movement measures and comprehension accuracy by language based on these by-participant averages (Figure 2; the values used to create this plot are available under 'auxiliary files' in the project's OSF page). A visual inspection of this figure points to substantial variability in eye-movement behavior across languages. Note that similar descriptive patterns were observed when matched, and unmatched texts were examined separately (see Supplementary Materials S6), and that by-participant means of eye-movement measures calculated for matched and unmatched texts in each language correlated very highly (mean $r = 0.90$, range: 0.68 – 0.97; see Supplementary Materials S7). This suggests that the language differences observed were not because some texts were not perfectly matched translations.

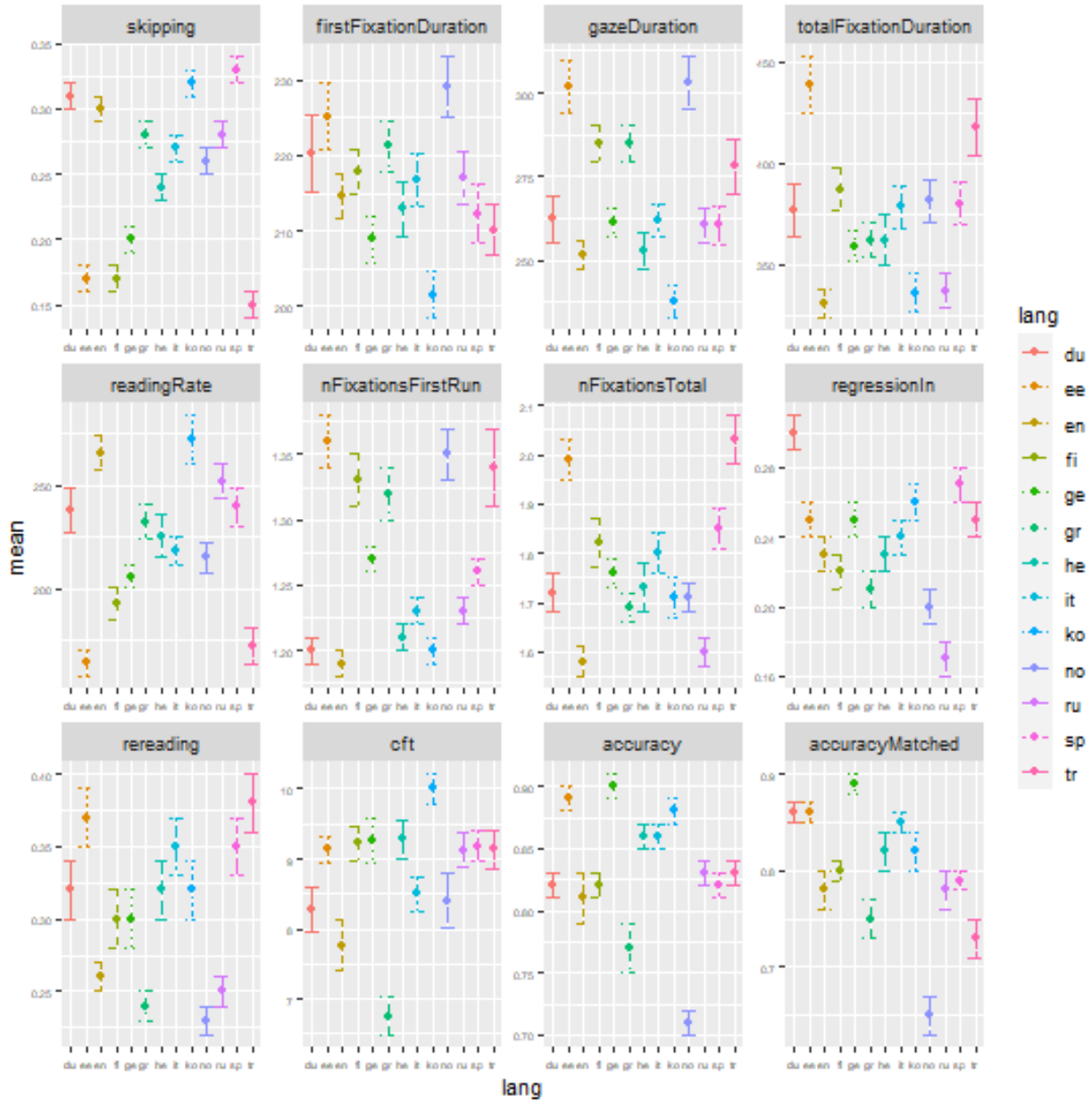

Figure 2. Means of eye-movement measures across languages. Error bars stand for ± 1 SE. accuracy: percent answers correct; accuracyMatched: percent answers correct in matched texts; cft: score in the CFT test; firstFixationDuration: first fixation duration; gazeDuration: gaze duration; nFixationsFirstRun: first run number of fixations; nFixationsTotal: total number of fixations; readingRate: reading rate; regressionIn: regression rate; rereading: likelihood of second pass; skipping: skipping rate; totalFixationDuration: total fixation duration. du: Dutch; ee: Estonian; en: English; fi: Finnish; ge: German; gr: Greek; he: Hebrew; it: Italian; ko: Korean; no: Norwegian; ru: Russian; sp: Spanish; tr: Turkish.

While a detailed analysis of specific oculomotor patterns is subject to future research, we note a few findings here. The Norwegian sample appears to stand out: these readers showed relatively lower accuracy in comprehension questions (65% in matched texts), shorter and a smaller number of fixations and a higher rate of skipping. This might indicate that this sample engaged in a relatively superficial kind of reading, investing less in the inferential and integrative processes required for comprehension than readers at other sites. Another noteworthy pattern emerged in Estonian: these readers had a large number of fixations on the words they read, along with relatively long fixations and a high re-reading rate. This stands in contrast to a typical trade-off between the number of fixations and their duration or the number of passes. A final observation is that Korean readers demonstrated short reading times and a high skipping rate, presumably due to very short words in this orthography (see below). These patterns may be important to take into account when drawing cross-linguistic comparisons.

The cross-sample variability in Figure 2 leads to the first key theoretical question we ask in this section: What behavioral measures account for the most cross-linguistic variability in reading performance? We address this central question in several complementary ways in the remainder of this paper. In the initial analysis, we fitted ordinary linear regression models to each of the dependent eye-movement variables with a 13-level factor of Language as a sole predictor. We then estimated amounts of explained variance as the effect size of Language (adjusted R^2). By far, the strongest systematic variability was found in skipping rate. Namely, 46% of individual variability in skipping rate among the MECO readers was accounted for by the language they read. Language also explained 24% of the variance in first-run number of fixations. Language explained less variance in durational measures (adjusted R^2 of first fixation

duration 5%, gaze duration 16% and total fixation time 13%)[7]. Separate analyses conducted on matched and unmatched texts revealed similar estimates of the proportion of variance explained by Language in the different eye-movement measures (see Supplementary Materials S8). Notably, in both matched and unmatched texts skipping emerged as the variable that is most strongly impacted by Language, suggesting that this effect holds regardless of the cross-linguistic comparability in the global semantic content of passages.

This pattern indicates that most cross-linguistic differences in the oculomotor behavior at the word level materialize in the spatial distribution of fixations over words (e.g., which words attract fixations and which do not). Once the word is fixated, cross-linguistic variability in how long it is viewed is substantially lower, despite the diversity of studied languages. We return to this finding below.

Another approach to identifying the relative importance of predictors of reading behavior recruits a conditional inference analysis of the MECO data. The outcome of this analysis is a decision tree, which identifies a hierarchy of reading measures that most strongly predict differences between languages. More specifically, in this analysis, by-participant mean values of all reading measures serve as input to a recursive partitioning classification tree that has language as a response variable. At each recursion, this procedure identifies the reading measure with the strongest association with the language variable (response). Then it implements a binary split of that measure on the value that offers the best binary partition of participants into classes representing languages. Inferential statistics for associations between reading variables and

[7] Comprehension accuracy appeared to vary substantially across languages as well (adjusted R^2 for all texts 29% and for matched texts 25%). A closer inspection revealed, however, that much of the variance is due to a lower comprehension accuracy of one language sample (Norwegian): when the sample is excluded, amounts of variance explained are on par with fixation durations: 18% for all texts, 14% for matched texts. See above for the discrepancy between the Norwegian and other samples.

language as the response variable, as well as the best values for partitions into classes, are estimated using the permutation test. Partitions are implemented within classes until the permutation test *p*-values for the splits are not statistically significant. For further technical details, see Matsuki et al. (2016) and references therein. We used function *ctree* from *party* package (Hothorn et al., 2006) in the statistical software platform *R*.

Figure 3 visualizes the resulting conditional regression tree. Variables that account for splits higher up in the tree are more important than those closer to the bottom, i.e., they are more strongly associated with language as a response variable. Again, skipping rate emerged as the single most important factor in accounting for cross-linguistic variability. This variable was indicated in the first recursion (from top to bottom) as the one with the strongest association with language as a response variable. Only at lower portions of the tree does an additional variable (first run number of fixations) comes into play. Durational variables did not come out as significant predictors of language as a response variable. It is noteworthy that each terminal node (representing the distribution of participants over languages in the bottom part of Figure 3) accounts for a non-trivial percent of readers from multiple languages. Thus, for example, the majority of Estonian and Finnish readers (29/52 Estonians, 25/49 Finns) were concentrated in the leftmost node in Figure 3 (i.e., had a skipping rate lower than 21.6% and less than 1.39 first pass fixations), but still more than 40% of participants from these two samples were scattered in other nodes. This suggests that there is no specific combination of oculomotor parameters (skipping rate, durations, regression rate etc.) that uniquely identifies reading in any given language. Separate analyses on matched and unmatched texts converged on skipping as the strongest predictor of cross-linguistic differences (see Supplementary Materials S9). In sum, skipping rate

and – to a smaller degree – number of first run fixations - account for the most behavioral variability between languages.

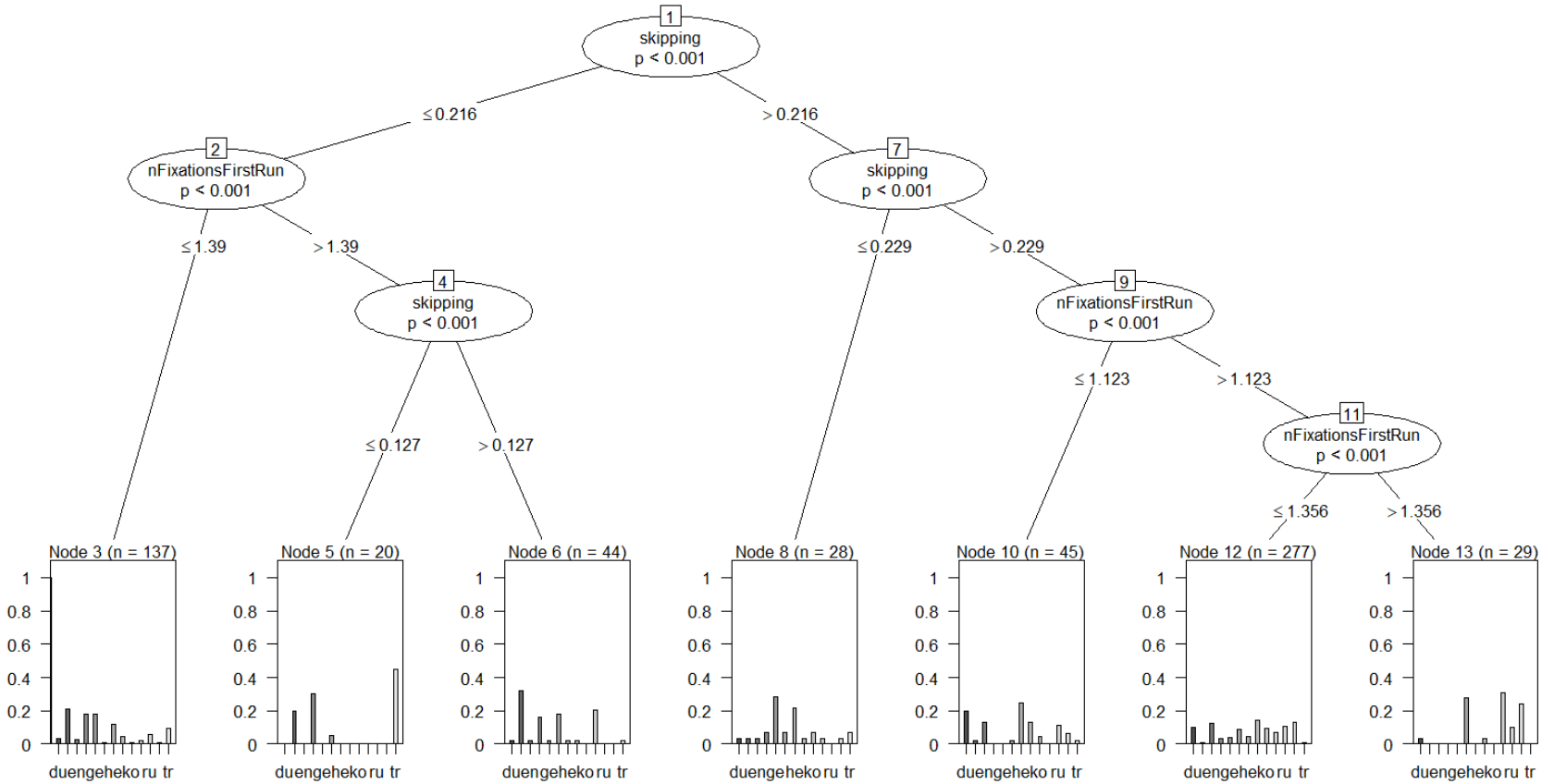

Figure 3*. Classification of participants into languages based on eye-movement measures. Best-split values of input variables are reported, along with p-values of the splits.*

The salient role of skipping in predicting cross-linguistic variability in reading performance calls for further investigation. Consistent with the literature on this aspect of oculomotor control (e.g., Brysbaert et al., 2005; Drieghe et al., 2004; Kliegl et al., 2004; Rayner & McConkie, 1976; Reilly & O'Regan, 1998; Vitu, 2011), we link skipping rate to one of the benchmark predictors of reading: word length. Specifically, we expect cross-linguistic differences in skipping rate to reflect variability in the distribution of word lengths across languages. It is a well-established finding that, within a language, longer words are skipped less often (see references above). In fact, Kuperman and colleagues (2018) show that word length has the greatest relative importance out of all predictors of skipping rate in English. Accordingly, in regard to cross-linguistic variability, we expect that written languages with shorter words on average will demonstrate proportionally higher skipping rates. While the majority of the reported languages are letter-based, Korean is an important exception. Our calculation of word length for Korean is based on syllable-based characters[8].

Separately for each language, we fitted a logistic regression mixed-effects model to the binary variable of whether the word was skipped, with word length as a sole predictor and by-participant and by-word random intercepts. Word lengths were centered (but not scaled), such that the intercept of a regression model estimated the predicted skipping rate for a word of average length in a given language: while given in logit units, we transformed these estimates into percentage points (i.e., estimated percent words skipped in a word of average length). The slopes of the regression models estimated an increase in skipping rate (in logit units) related to an

[8] Korean writing is based on a syllabic unit where a syllable structure of an onset, nucleus and a coda is visually represented in writing. Thus, orthographic representations of syllables like 고 and 공 consist of 2 or 3 alphabetic components that are spatially combined into a character: ㄱ and ㅗ and ㄱ, ㅗ, and ㅇ.

increase in one character. Supplementary Materials S10 includes descriptive statistics of word lengths and estimated slopes and intercepts.

The correlation between mean word length in a given language and skipping rate estimated for that length (the model intercept, transformed to percent) was negative and very strong: r = -0.88, p < 0.001. Figure 4 illustrates the finding. Korean was a language on one extreme with a mean word length of 2.92 characters (SD = 1.27) and an estimated skipping rate of 29%. This is because one Korean character typically represents 2-3 phonological elements. On the other extreme was Finnish, with the mean word length of 7.82 characters (SD = 3.90) and the estimated skipping rate of 6%. The remaining languages followed the linear trend almost perfectly[9]. Interestingly, the correlation between mean length and model slope (i.e., the rate of change in skipping rate as a function of word length) was very weak and not significant (r = 0.19, p = 0.539).

[9] The estimates for the skipping rate for an average word length in a language were not related to the number of characters subtended by one degree of visual angle, font size, or screen size. While different across testing sites, these parameters did not underlie the observed correlation between skipping rate and average word length.

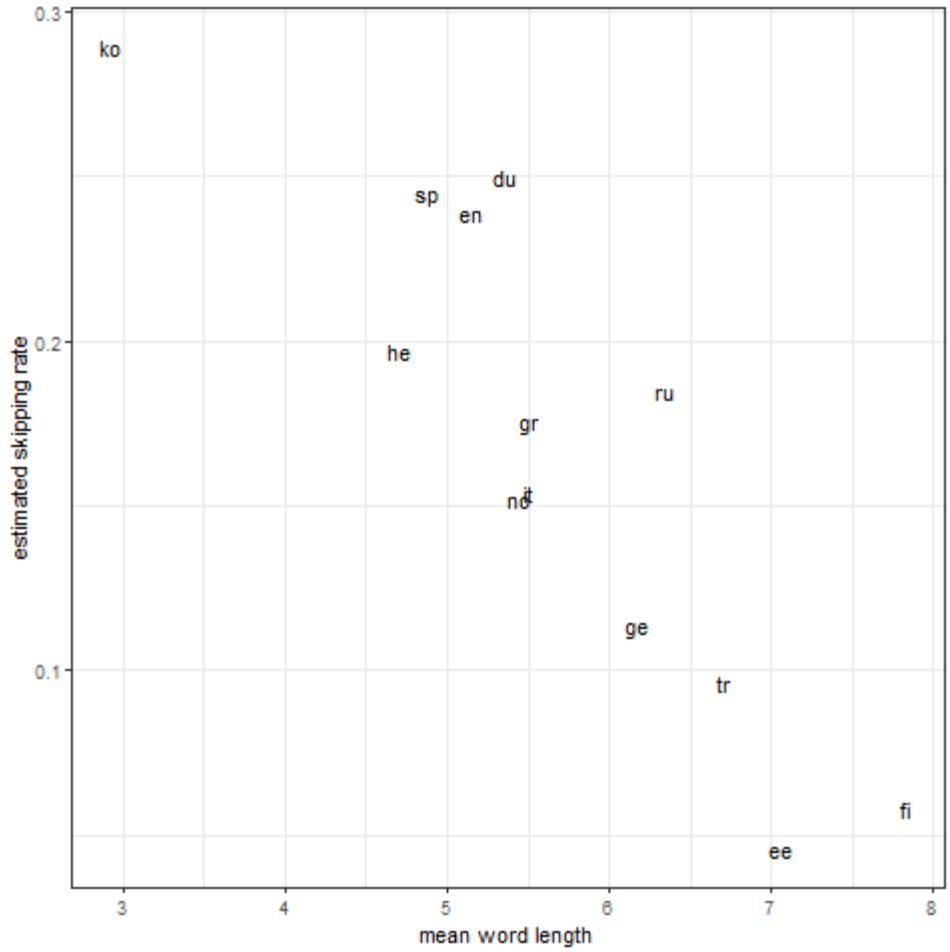

Figure 4. Estimated skipping rate for a word of an average length as a function of mean word length in language.

Taken together, these findings point to a strong role of visuo-oculomotor factors in explaining what makes eye-movement behavior vary across languages the most. It is well known that longer words elicit fewer skips within a language. We see that this finding generalizes across languages: i.e., a preference of a given written language for longer words comes with a lower skipping rate in reading. Conversely, it appears that every language responds to an increase of word length by one character with a roughly similar decrease in skipping rate, regardless of the language's overall gravitation towards longer or shorter words. This indicates a strong reliance of readers' probability of fixating vs skipping words on visual characteristics of the linguistic input.

Since this characteristic varies widely between languages (by a factor of 2.7 in mean word lengths in our sample), so does the value of skipping rate (by a factor of 5.8). At the same time, durational measures for fixated words differ much less between languages. While mean viewing times are markedly different for some pairs of languages (see Figure 2), the overall cross-linguistic variability in viewing times accounts for a relatively small amount of variance compared to the within-language variability and other eye-movement measures. Put differently, if one imagines a hypothetical reader who is equally proficient in all written languages in the present sample, most of their oculomotor accommodation to characteristics of specific languages will be driven by word lengths and will go into adjusting the rate of skipping. Once a word is fixated, cross-linguistic differences in word lengths or other characteristics will lead to a smaller adjustment in viewing time.

Our analyses so far revealed that some eye-tracking measures (i.e., skipping, and to a lesser extent, refixation) vary considerably across languages while other measures do less so. A related question is whether some languages are overall more similar to one another in terms of the eye-movement behavior of their readers. A reasonable starting point is that reading behavior may be more similar among languages that are more similar in their structure. As follows from Table 1, several languages in our sample have a similar historical origin (e.g., Germanic: Dutch, English, German, and Norwegian; and Romance: Italian and Spanish); a similar writing system type and script (e.g., there are nine alphabetic languages written in Latin-based scripts in our sample of 13); or a similar type of morphology and level of orthographic transparency. We examined whether the linguistic similarity between languages translated into similarity in the oculomotor patterns of their readers. To this end, we selected mean values of three eye-movement measures that represent different aspects of oculomotor behavior - skipping rate, gaze

duration, and total number of fixations - as a vector representing every participant. These variables were selected to reflect (with little redundancy) both the probability of fixating vs skipping a word, the time spent viewing the word in the first pass and the total effort of viewing the word (A solution that additionally included total fixation duration was also run and produced the same result: We report below the more parsimonious solution). We calculated the Euclidian distance between all pairs of (scaled) participant vectors and aggregated this participant-level data to compute an average distance between each pair of languages. This distance metric was supplied as input into a hierarchical cluster analysis using the Ward clustering criterion: the *hclust* function in R was used (Langfelder & Horvath, 2012).

Figure 5 reports the clustering solution. The first partition (from top to bottom) is between Finnish, Estonian, and Turkish, versus the remainder of languages. A lower partition on the right side of the tree (which contains the remaining ten languages) separated two clusters of five languages each (Spanish, Dutch, Korean, English and Russian; versus Greek, Norwegian, Hebrew, German and Italian). This clustering rules out several logical possibilities for behavioural commonalities. Thus, languages that have largely overlapping lexicons and broad similarities in their phonology, morphology and orthography appear to be no closer to each other in their behavioral patterns than to other languages. In particular, Germanic and Romance languages are dispersed over multiple clusters rather than grouped together. Furthermore, similarities in scripts were inconsequential: Hebrew, Korean, Russian and Greek were dispersed among languages using Latin-based scripts rather than grouped together. In fact, the only potential criterion that separated some of the clusters from others was word length, which was in turn related to skipping rate (see above): Finnish, Estonian, and Turkish, which form a distinct cluster from other languages, are the languages with the longest words in the sample. They are

also agglutinative and highly orthographically transparent, which are both factors contributing to increased word length. It is possible that a clear-cut organizing principle exists in the clustering of languages based on reading behavior, but it is masked from us due to relatively small sample sizes and possible sampling biases in the respective languages. At this point, we can only conclude that even if similarities of a linguistic nature do lead to cross-linguistic similarities in reading behavior, these tendencies are subtle.

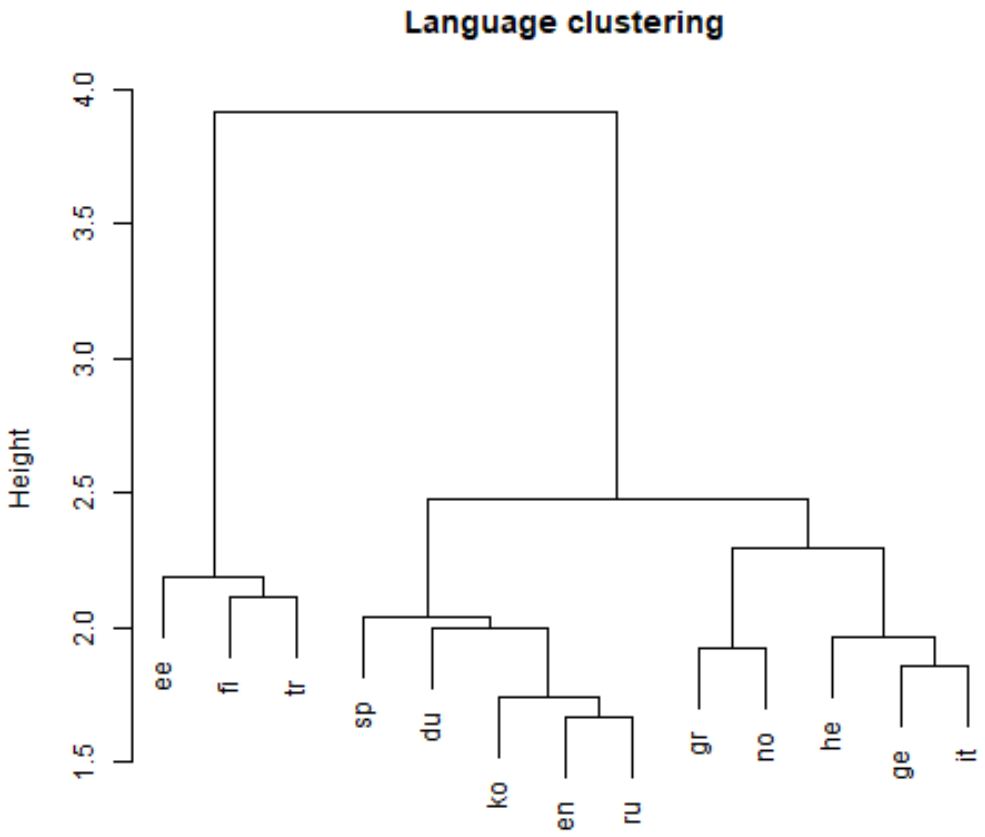

Figure 5. Hierarchical clustering of languages based on eye movements.

General Discussion

The inspiration for this paper is that empirical science both drives and is driven by accessibility to high-quality and large-scale data. The Open Science movement in the Cognitive Sciences adopted this notion, leading to a constantly growing number of collaborative multi-lab studies aimed at providing theories with such data (e.g., Hagger et al., 2016; Open Science Collaboration, 2015; ManyBabies Consortium, 2020). However, in addition to typical requirements from multi-lab investigations, a collaborative study of *reading* must also additionally reflect the striking diversity of languages (which vary in their phonology, morphology, and syntax), including written languages (which embody a range of solutions as to how to reflect speech in print). This is essential because theories of reading that claim any degree of cross-linguistic coverage must be tested using comparable data from multiple languages. Such data should be obtained using comparable designs across languages, both in format, content, task, and data collection methods.

The present paper provides the field of reading with such necessary data. We specifically focus on eye-tracking methodology to study reading, which is arguably the most ecologically valid and temporally sensitive record of reading behavior, and indeed eye-movements are part and parcel of reading itself. We start by examining whether the need for cross-linguistic data has already been satisfied in studies of eye-movements during reading by using a bibliometric analysis of relevant publications over the last two decades to estimate the field's cross-linguistic coverage. The analysis reported in Part I revealed clear biases towards a handful of languages: except Chinese, well-represented languages tend to be alphabetic and Roman script-based, and European (mostly Indo-European, with an expected further bias towards English). Moreover, the

number of studies that conducted a coordinated examination of more than one language is very small, and no study has covered more than three languages.

In Part II of this paper, we introduce the Multilingual Eye-movement COrpus (MECO): a collaborative international project aimed at addressing the need for comparable cross-linguistic data. MECO comprises eye-tracking data for reading in the first (dominant) language, reading in English (a non-dominant language for all but one sample), and a battery of individual differences tests both in the readers' first language and in English. In the current first release of MECO, we report first language reading data from laboratories in 13 countries and languages. These 13 languages exemplify a typologically wide range of phonological, morphological, and syntactic systems, originating from multiple language families. MECO thus makes possible a direct comparison between different writing systems (alphabets and abjads) and scripts (alphabetic Roman and non-Roman based, Hebrew, and Hangul). Reading materials were 12 encyclopedic texts, including both translation-equivalent and untranslated materials. Participants were university students in their respective countries, with the language of testing as their dominant language. The MECO eye-movement record includes information on a broad range of oculomotor measures. It is further supplemented by data on comprehension accuracy, demographic and linguistic background as well as tests of individual differences, some of which were shared across all samples and others were specific to each language. In the spirit of Open Science, the MECO data, materials and code are made available to promote cross-linguistic collaborative research on reading and advance reproducibility. Therefore, MECO constitutes a valuable tool to address novel reading research questions across a wide variety of languages without the need to collect (eye-tracking) data. It is also accessible to researchers working on

less-studied languages who may not have the necessary equipment to run eye-tracking experiments at their disposal.

In Part III of the paper, we demonstrate the utility of the MECO data by providing a comparative analysis aiming at characterizing similarities and differences in cross-linguistic reading patterns based on all languages in the corpus. The main finding is that the oculomotor measure differentiating reading behaviors across languages the most is skipping rate. That is, languages differ in the rate of likelihood in which readers tend to fixate on a word at least once versus skipping it altogether. In turn, we find that skipping rate in a language is very strongly determined by the average word length in that language, with languages gravitating towards longer words (e.g., Finnish, Estonian, or Turkish) showing an overall lower skipping rate than those with shorter words (e.g., Korean or Hebrew). These systematic patterns were observed in both semantically matched and unmatched texts, suggesting that they are robust to natural variability in topics and propositional contents. Remarkably, neither the differences in word length nor other linguistic characteristics of the current set of languages showed a noticeable systematic influence on any other oculomotor measure. In particular, there were minor differences in fixation durations across languages (either first fixation durations or total fixation duration). In all languages, if readers select a word for fixation, they tend to spend a similar time viewing it on average. This suggests that viewing times are mainly representative of core language processes rather than surface characteristics of languages on the linguistic levels examined in this paper. Of course, this finding does not imply that multi-lingual investigations of reading times (and all other behavioral measures of reading) are unnecessary. Our results pertain to the effect of general linguistic features on reading times and do not necessarily extend to other phenomena that are language-specific and could (and should) be investigated cross-linguistically.

It may be tempting to entirely couch the discussion of the cross-linguistic impact of word length and skipping rate in visual terms, with the count of characters and the space they occupy on a screen driving the oculomotor planning and execution (see references above). It is important to realize, however, that cross-linguistic differences in word lengths reflect fundamental properties of written languages (see discussion in Liversedge et al., 2016). Whether a writing system that a language adopts chooses its symbols to reflect all individual sounds (alphabetic), some types of sounds (consonantal alphabets or abjads), syllables (syllabaries), or entire words (logographic) has a profound impact on word lengths in a system. Similarly, how characters of a language package phonological information visually affects word length too (see Korean *Hangul*). Other factors of influence include orthographic transparency (the degree of completeness and consistency with which orthographic words reflect words' phonology), the use of function words (e.g., articles, prepositions), and the type of morphology (e.g., agglutinative like Finnish or Turkish where markers of syntactic functions are affixed to the word versus isolating languages like English where they are expressed as separate function words).

Moreover, specific orthographic conventions within a language affect word length. For instance, Hebrew does not allow single-letter orthographic words. In the same vein, German, Dutch, Norwegian and Finnish allow very long unspaced compounds, while English introduces spaces between some constituents. Thus, word length and skipping rate as its behavioral counterpart are strongly related to the architecture of a written language and its relationship with the oral language. With the present representation of languages, we still do not have sufficient cross-language coverage and variability to address a systematic influence of specific linguistic features like script, typological family, morphological type or orthographic transparency. This question can be addressed as a wider range of languages is added to MECO.

Other noteworthy findings concern the surprising lack of similarity in reading behavior between written languages that are genetically or typologically related. That is, the clustering solution based on major predictors of eye-movement behavior grouped together languages in a way that, to our knowledge, does not reflect any accepted classification of either oral languages or writing systems (Daniels & Bright, 1996; Dryer & Haspelmath, 2013). This finding hints at a possibility that behavioral patterns during reading are mostly guided by features of input texts that are not accounted for and do not easily translate into existing language classifications. If so, a new, behaviorally relevant map of language structures may be required.

Our conclusions are reached based on texts of which some were translated from English to other languages of MECO, and some were not (though they were constructed to represent the same topic and genre). This manipulation pursued the goal of determining how essential close semantic matching is for different cross-linguistic investigations. This question is methodologically critical. If translated texts are imperative for reaching a sound comparative conclusion about any aspect of reading, no previous data are valid material for cross-linguistic comparisons unless based on translations.

A full exploration of what cross-linguistic reading behaviour patterns hold in both matched and unmatched materials is beyond the scope of this first paper. Still, in all the analyses of MECO data above, highly similar results are obtained in matched and unmatched materials. Namely, we found similar descriptive patterns in matched and unmatched texts, similar estimates of variance explained by Language when estimated based on reading of matched texts, unmatched texts, or their combination, and a similar role of skipping as the strongest predictor of cross-linguistic differences. Thus, in pursuing the present set of analytical goals mostly tied to the level of the word, we did not find matching through translation to be a relevant factor, at least

not within the genre of encyclopedic expository passages. We further showed high correlations between by-participant means of eye-movement measures computed on matched and unmatched texts in the various sites (see Supplementary Materials S2). This suggests that for a certain range of research questions and phenomena (in particular, those that employ by-participant means of different eye-movement measures), the requirement of a close semantic matching across languages may be relaxed. We emphatically do not imply that *all* questions can be answered without resorting to translated texts. To give one example, Liversedge et al. (2016) demonstrated that a fruitful study of global, cumulative eye-movement patterns at the sentence and passage level demands a thorough semantic matching across languages. For such questions, researchers should use the matched portion of the MECO corpus only. Importantly, since it may not be clear a priori which aspects of reading are or are not critically bound to the close semantic similarity between materials across languages, MECO can serve as a testbed for addressing this question and thus guide design decisions in future work.

Limitations and future directions

We view MECO as a living organism that undergoes evolution, partly as a way to remedy its limitations, and thus discuss current limitations and future directions jointly. An important goal for the MECO project is to expand its coverage of individual languages, language groups and writing systems. This will correct the current over-representation of alphabetic languages and languages that use spacing for overt segmentation of characters into words or syllables. A future release of MECO will contain an additional range of languages.

An additional limitation is that currently, there is no systematic way to disentangle behaviors specific to this sample (e.g., selectivity of their university, variability in reading proficiency within a country, and testing procedures in specific labs) from behaviors dictated by particulars of that language. While we view this as an inherent limitation of cross-language research, one way to mitigate some of it is to collect multiple samples for each language. Within-language samples can represent regional varieties or differences in educational or social backgrounds of readers and differences between universities on how participants are asked to read. Importantly, we invite additional collaborators to participate in this multi-lab initiative. Both new languages and additional samples from currently included languages are welcome and are critical for expanding the present resource and increasing its variability and reliability. We hope that public access to all MECO materials and procedures will facilitate this expansion. Guidelines for how to participate are provided at the MECO project website, www.meco-read.com.

A related limitation has to do with statistical power. Current samples (mostly, 45-55 participants) afford sufficient power for some types of analyses but limited power for others. The exact power estimates obviously depend on the unit and type of analysis and the expected effect size, but the general estimates are as follows. MECO is expected to provide sufficient statistical power to observe effects of a median size ($|d| \geq 0.4$) in sentence-level analyses, even when only two language samples are considered and even when matched and unmatched texts are examined separately. This is because each of such conditions would generally meet the 80% power requirement of 40 participants x 40 observations estimated in Brysbaert and Stevens (2018). By extension, sufficient power is also expected at the word-level, where a much larger number of observations is available in each language sample, even if split into matched or unmatched texts.

Note, however, that analyses where each participant contributes only one data point and languages are examined individually or are compared pairwise may be characterized by limited power. In future releases of MECO, this will be addressed in two ways. First, some laboratories will increase their samples, especially ones where pandemic-related closures thwarted data collection. A second increase in the number of observations may come from further developing the *popEye* software package (Schroeder, 2019), which we use here for automatic analysis of eye fixation locations (thus avoiding an error-prone manual process). Refined algorithms may also reduce the number of observations the software excludes due to minor deficiencies of calibration or head movements. Despite these limitations, it is clear that even this first release of MECO provides unprecedented statistical power for cross-language analyses, with more than 500 participants providing almost 800,000 data points.

As stated above, MECO is a resource that can be used for pursuing a variety of research questions. For instance, additional work needs to be done on investigating cross-linguistic effects of benchmark predictors of reading behavior, including the effects of word length, frequency and predictability in context (Rayner, 2009). We expect future MECO updates to incorporate frequency and predictability estimates for most languages and provide analyses that will compare benchmark effects across languages and writing systems. Also, our present focus was on the word-level reading behavior across languages: a follow-up is recommended to tackle cross-linguistic variability at the sentence and passage level. An additional examination of spatial information regarding landing positions and amplitudes of saccadic eye-movements will shed further light on the "where" aspect of oculomotor control. Furthermore, we expect analyses relating skill tests in individual languages to the patterns of reading behavior to contribute to the growing literature on individual differences in reading (e.g., Radach & Kennedy, 2013). Another

example of an interesting prospective study derives from the finding that skipping vs. fixation probability account for most of the cross-linguistic differences. Further research may look at whether and how these differences reflect trade-offs between spatial fixation density and reading times (e.g., Radach & Heller, 2000), possibly based on orthographic or morphological complexity of individual languages[10].

A final future direction that we outline here departs from the present focus on a bird's-eye view of reading behavior across written languages of the world. We believe that MECO can also be useful for more particular questions of psychological and linguistic interest and a more specific examination of individual languages and language groups.

Open Practices Statement

MECO's data and materials are made available at the project's OSF page – see 'data availability' section above for details.

Acknowledgements

Research reported in this publication was supported by the following grants: The Social Sciences and Humanities Research Council of Canada Partnered Research Training Grant, 895-2016-1008 (PI: G. Libben); the Canada Research Chair (Tier 2; PI: V. Kuperman); the CFI Leaders Opportunity Fund (PI: V. Kuperman); Concerted research action BOF13/GOA/032 of Ghent University; FWO Project (PI: M. Brysbaert); Estonian Research Council Mobilitas Pluss postdoctoral researcher grant MOBJD408 (PI: K. Lõo); ERC Advanced grant, project 692502-

[10] We thank Ralph Radach for this and many other helpful suggestions.

L2STAT (PI: R. Frost), the Israel Science Foundation (ISF) grant 48/20 (PI: N. Siegelman), and Saint Petersburg State University grant ID 75288744, 121050600033-7 (PI: N. Slioussar).

We wish to thank the following individuals: Mariam Bekhet, Paige Cater, John Connolly, Melda Coskun Karadag, Connie Imbault, Alyssa Janes, Shani Kahta, Minji Kang, Evgenia-Peristera Kouki, Elizaveta Kuzmina, Nadia Lana, Sean McCarron, Kelly Nisbet, Victoria Ong, Anat Prior, Eva Saks, Elisabet Service, Anna Swain, Heather Wild, Sophia Yang, and Laoura Ziaka. Thanks are due to Ralph Radach and two anonymous reviewers for valuable comments to earlier drafts.

References

Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2018). SPALEX: A Spanish lexical decision database from a massive online data collection. *Frontiers in Psychology*, *9*, 2156.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Beymer, D., & Russell, D. M. (2005). WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1913-1916).

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, 104047.

Brysbaert, M., Drieghe, D., & Vitu, F. (2005). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 53-77). Oxford, UK: Oxford University Press.

Brysbaert, M., Keuleers, E. & Mandera, P. (2019). Recognition times for 54 thousand Dutch words: data from the Dutch crowdsourcing project. *Psychologica Belgica, 59(1)*, 281–300.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of Cognition, 1(1):9.

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*, 1-24.

Cohen, A. L. (2013). Software for the automatic correction of recorded eye fixation locations in reading experiments. *Behavior Research Methods, 45*(3), 679-683.

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. Retrieved from http://expsy.ugent.be/downloads/geco/

Daniels, P. T., & Bright, W. (Eds.). (1996). *The world's writing systems*. Oxford University Press on Demand.

Daniels, P. T., & Share, D. L. (2018). Writing system variation and its consequences for reading and dyslexia. *Scientific Studies of Reading*, *22*(1), 101-116.

Drieghe, D., Brysbaert, M., Desmet, T., & De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology*, *16*(1-2), 79-103.

Dryer, Matthew S. & Haspelmath, Martin (eds.) (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2021-08-30.)

Feng, G., Miller, K., Shu, H., & Zhang, H. (2009). Orthography and the development of reading processes: An eye-movement study of Chinese and English. *Child Development*, *80*(3), 720-735.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... & Pallier, C. (2010). The French lexicon project: lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*(2), 488-496.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods, 45*(4), 1182-1190.

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(5), 263-279.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 104.

Fukuda, R., & Fukuda, T. (2009). Comparison of reading capacity for Japanese, German, and English. *Perceptual and Motor Skills, 108*(1), 281-296.

Furnes, B., & Samuelsson, S. (2011). Phonological awareness and rapid automatized naming predicting early development in reading and spelling: Results from a cross-linguistic longitudinal study. *Learning and Individual Differences, 21*(1), 85-95.

Geva, E., & Wang, M. (2001). The development of basic reading skills in children: A cross-language perspective. *Annual Review of Applied Linguistics, 21*, 182-204.

Goswami, U. (2002). Phonology, reading development, and dyslexia: A cross-linguistic perspective. *Annals of Dyslexia, 52*(1), 139-163.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... & Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics, 15*(3), 651-674.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, *8*(2), 1-12.

Inhoff, A. W., & Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. In: G. Underwood (ed.), *Eye Guidance in Reading and Scene Perception*, (pp. 29–53). Oxford: Elsevier.Katz, L., Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In Frost, R., Katz, L. (Eds.), *Orthography, Phonology, Morphology, and Meaning* (Vol. 94, pp. 67–84). Amsterdam, The Netherlands: Elsevier Science.

Kennedy, A., Heller, D., Pynte, J., & Radach, R. (Eds.). (2000). Reading as a perceptual process. Elsevier.

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287-304.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262-284.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12.

Koda, K., & Zehler, A. M. (Eds.). (2008). *Learning to read across languages: Cross-linguistic relationships in first-and second-language literacy development*. Routledge.

Kuo, L. J., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist, 41*(3), 161-180.

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.

Kuperman, V., Siegelman, N., Schreuder, S… Usal, K. (2021). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus (MECO). Under revision in *Studies in Second Language Acquisition.*

Langfelder, P., & Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *Journal of statistical software, 46(11)*.

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., & Kliegl, R. (2019). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods, 51*(3), 1161-1178.

Liversedge, S. P., Blythe, H. I., & Drieghe, D. (2012). Beyond isolated word recognition. *Behavioral and Brain Sciences*, *35*(5), 293-294.

Liversedge, S. P., Drieghe, D., Li, X., Yan, G., Bai, X., & Hyönä, J. (2016). Universality in eye movements and reading: A trilingual investigation. *Cognition*, *147*, 1-20.

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods, 50*(2), 826-833.

Mandera, P., Keuleers, E., & Brysbaert, M. (2019). Recognition times for 62 thousand English words: data from the English crowdsourcing project. *Behavior Research Methods, 52*, 741 760.

ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24-52.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-967.

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading, 20*(1), 20-33.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews, 4*(1), 1-9.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Organisation for Economic Co-operation and Development. (2013). *Technical report of the survey of adult skills (PIAAC).*

Pan, Yan, Richter, Shu, & Kliegl (2021). The Beijing Sentence Corpus: A simplified Chinese sentence corpus with eye movement data and predictability norms. *Behavior Research Methods.*

Papadopoulos, T. C., Csépe, V., Aro, M., Caravolas, M., Diakidoy, I. A., & Olive, T. (2021). Methodological Issues in Literacy Research Across Languages: Evidence From Alphabetic Orthographies. Reading Research Quarterly, 56, S351-S370.

Pynte, J., & Kennedy, A. (2006). An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research*, *46*(22), 3786-3801.

Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. Quarterly Journal of Experimental Psychology, 66(3), 429-452.

Radach, R., Reilly, R., & Inhoff, A. (2007). Models of oculomotor control in reading: Toward a theoretical foundation of current debates. Eye Movements, 237-269.

Rau, A. K., Moll, K., Snowling, M. J., & Landerl, K. (2015). Effects of orthographic consistency on eye movement behavior: German and English children and adults process the same words differently. *Journal of Experimental Child Psychology*, *130*, 92-105.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457-1506.

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.

Rayner, K. & Liversedge, S.P. (2011). Linguistic and cognitive influences on eye movements during reading. In S. Liversedge, I. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook of Eye Movements* (pp. 751-766). Oxford, UK: Oxford University Press.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research, 16,* 829-837.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C., Jr. (2012). *Psychology of reading*. New York, NY: Psychology Press.

Reilly, R. G., & O'Regan, J. K. (1998). Eye movement control during reading: A simulation of some word-targeting strategies. *Vision Research, 38(2),* 303-317.

Sabatini, J. (2015). Understanding the basic reading skills of US adults: Reading components in the PIAAC literacy survey. *ETS Center for Research on Human Capital and Education*.

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition, 35*(2), 261-290.

Schroeder, S. (2019). *popEye: Analysis of eye-tracking data from reading experiments*. R package version 0.6.4.

Schröter, P., & Schroeder, S. (2017). The developmental lexicon project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods, 49*(6), 2183-2203.

Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M., & Richlan, F. (2016). Words in context: The effects of length, frequency, and predictability on brain responses during natural reading. *Cerebral Cortex, 26(10),* 3889-3904.

Seymour, P. H., Aro, M., Erskine, J. M., & Collaboration with COST Action A8 Network. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143-174.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, *134*(4), 584.

Share, D. L. (2014). Alphabetism in reading science. *Frontiers in Psychology*, *5*, 752.

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.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295.

Staub, A. (2021). How reliable are individual differences in eye movements in reading?. Journal of Memory and Language, 116, 104190.

Sun, F., & Feng, D. (1999). Eye movements in reading Chinese and English text. In J. Wang, A. W. Inhoff, & H.-C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 189–205). Mahwah, NJ: Lawrence Erlbaum.

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2014). The Chinese lexicon project: a repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*(1), 263-273.

Tang, S., Reilly, R. G., & Vorstius, C. (2012). EyeMap: a software system for visualizing and analyzing eye movement data in reading. *Behavior research methods, 44(2)*, 420-438.

Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, *50*(5), 1763-1777.

Verhoeven, L., & Perfetti, C. (Eds.). (2017). *Learning to read across languages and writing systems*. Cambridge University Press.

Vitu, F. (2011). On the role of visual and oculomotor processes in reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 731–749). Oxford University Press.Weiß, R. H. (2006). *Grundintelligenzskala 2 mit Wortschatztest and Zahlenfolgetest* [Basic intelligence scale 2 with vocabulary knowledge test and sequential number test]. Göttingen, Germany: Hogrefe.

Whitford, V., & Titone, D. (2012). Second-language experience modulates first-and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review*, *19*(1), 73-80.

Woolley, G. (2011). *Reading comprehension: Assisting children with learning difficulties*. Dordrecht, The Netherlands: Springer International.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*(4), 992-1003.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., ... & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*(4), 551-559.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin*, 131(1), 3-29.