Chapter 4

Researching Reading

4.1 Introduction

Reading is a relatively recent development in human history, existing for only a few thousand years (Immordino-Yang and Deacon, 2007). However, it has become an essential life skill in modern society, one that is developed over many years of exposure, formal instruction and practice. Good reading skill underpins academic achievement (for a discussion see Renandya, 2007), and for second language learners, reading is a gateway to learning new vocabulary, more colloquial language and new grammatical constructions (Wilkinson, 2012). As Huey summarised over a hundred years ago, and which is still as true today, gaining a complete understanding of reading – how we learn to read, how we become fluent readers, how to best teach reading, etc. – is an important aspiration.

> And so to completely analyse what we do when we read would almost be the acme of a psychologist's dream for it would be to describe very many of the most intricate workings of the human mind, as well as to unravel the tangled story of the most remarkable performance that human civilization learned in all of its history. (Huey, 1908, p. 6)

We know that for readers the primary goal is to identify words, ascertain their meaning and integrate them into their unfolding understanding of a sentence and/or larger discourse. However, what exactly happens when we read? How do our eyes move? Do we look at every word when we read? Eye-tracking has allowed us to gain a fairly comprehensive understanding of what happens during reading and the factors that impact it, which will be an important focus in this chapter. As we will see more explicitly in Sections 4.3–4.6, a good understanding of both of these is fundamental to our ability to create well-designed studies.

As we saw in Chapter 1, reading involves a series of ballistic eye-movements (*saccades*), brief pauses (*fixations*) and movements back to previous parts of a text (*regressions*). Saccades occur largely due to the limitations of the visual system. More precisely, vision is 'clear' around a fixation, but clarity decreases moving away from it. This situation is depicted in Figure 4.1. In example (a) when the eyes fixate 'reader', this word is clear, making it easy to identify. Moving away from this point, vision becomes progressively less sharp. In order to clearly see further words, the eyes need to advance (perform a saccade). In (b) we see that when the eyes move forward from their position in (a) to fixate the word 'really', it brings this word and the surrounding ones into a region of good visual acuity.

The decreasing visual acuity from a fixation point outwards that is demonstrated in Figure 4.1 is due to the physiological structure of the eye (Balota and Rayner, 1991). More specifically, a line of text falling on the retina of the eye can be divided into three regions: the fovea, encompassing 1° of visual angle on each side of a fixation; the parafovea, extending to 5° of visual angle on each side of fixation; and the periphery, which includes

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

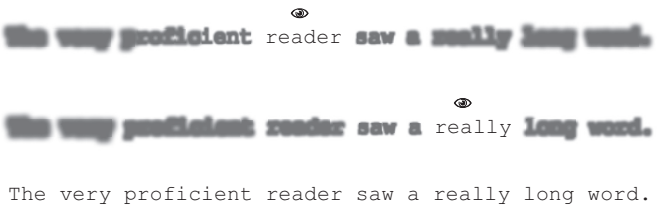

Figure 4.1 Depiction of the decrease in visual acuity moving away from a fixation point (depicted by the eye) in (a) and (b), while (c) presents an unmodified version of the same sentence.

anything beyond 5° of visual angle (Rayner, Balota and Pollatsek, 1986). The fovea is specialised for maximum visual acuity. Moving away from the fovea, visual acuity decreases. Words appearing in the parafoveal region that are closer to the fovea will be clearer than those that are further away.

In reading it makes sense to talk about visual acuity in terms of letters and spaces (referred to as 'letter spaces') instead of talking about degrees of visual angle. Although it depends on the size of the print, in reading, three to four letter spaces are roughly equivalent to 1° of visual angle (Balota and Rayner, 1991). If we apply this to Figure 4.2 and say that three letter spaces equals 1°, then in the example in (a) when fixating on 'A', the fovea would include the three letters 'B', 'C' and 'D' on both sides of the fixation. The parafovea would encompass the following twelve letter spaces, stretching from 'E' to 'P' on either side of the fixation. Finally, the periphery would include anything from the first 'X' onwards. Importantly, really good visual acuity only encompasses the fovea and 'near parafovea' (the parafoveal region closest to the fovea), which generally corresponds to the currently fixated word and the next one (although the subsequent two to three words will be clear if they are both/all very short). This helps explain why saccades in silent reading are usually seven to nine letter spaces (see Table 1.1 in Chapter 1). Normally, this saccade length corresponds to what is needed to bring the word that occurs two words after the one that is being fixated into focus.[1]

Importantly, example (a) in Figure 4.2 *only* demonstrates the constraints on visual acuity due to the structure of the eye. The area from which readers actually extract useful information is in fact asymmetric. For readers of alphabetic languages like English that are read from left to right, the effective visual field only extends three to four letters to the left of fixation and does not extend beyond the boundary of the currently fixated word, regardless of the number of letters (Rayner, Well and Pollatsek, 1980). This is the situation depicted in example (b) in Figure 4.2. The fixation is on the 'a' in the word 'reader', but useful information is only extracted as far as the onset of 'reader' and not before the word boundary. In contrast, to the right of the fixation, useful information is extracted from the fovea and parafoveal regions. In (b) this encompasses the three final letters in the word 'reader' (foveal region) and subsequent twelve letter spaces (parafoveal region) which is up to the second 'l' in 'really'. Because visual acuity decreases moving away from the foveal

[1] In the literature it is common to see the notation *n*, *n*+1 and *n*+2, where *n* indicates the word being fixated, *n*+1 the word after the fixated one, and *n*+2 the word after that. In Figures 4.1 (a) and 4.2 (b), *n* = 'reader', *n*+1 = 'saw', *n*+2 = 'a'. Saccades often go from *n* to *n*+2.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

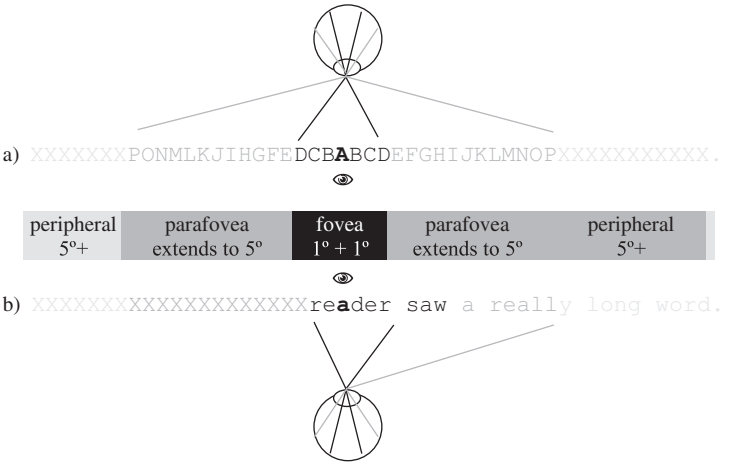

Figure 4.2 Illustration of a line of text falling on the foveal, parafoveal and peripheral regions of the retina. Example (a) shows these regions on a line of letters, while (b) demonstrates the asymmetric nature of visual acuity in reading.

region, the word 'saw' will be clearer than 'really' because it is closer to the fovea. It is important to note that visual acuity in reading is asymmetric in the direction of reading (Pollatsek et al., 1981). This means that in a language like Hebrew, where reading is right to left, parafoveal view extends to the left and not the right. Similarly, when Japanese is read vertically, the visual span is also asymmetric in the direction of reading – in this case downwards (Osaka, 2003). Thus in reading, there is asymmetric extraction of information in the direction of upcoming words (Clifton et al., 2016).

So why do we need this somewhat technical discussion about the distinction between words in the fovea and parafovea? The next few paragraphs will look at why these distinctions are important in reading, as well as their implications for designing research studies. As Figure 4.1 demonstrates, when the eyes land on the word 'reader' (in foveal view), it is seen clearly. During the fixation, the reader needs to use the visual, orthographic, phonological and potentially morphological information to identify the word and activate the lexical representation that contains information about its meaning, grammatical role, etc. The reader needs to keep track of all of this information and integrate it with similar information from preceding words.

Given all the processing work being done on the fixated word that is in foveal view (n), it seems plausible that readers might not have enough cognitive resources left to identify the next word that is in parafoveal view (n+1) – a word that is actually seen less clearly. However, evidence indicates that there is some processing of n+1 words that are in parafoveal view (for a fuller discussion see Hyönä, 2011; Rayner, 2009). Table 4.1 provides a summary of the processing effects that are evident for the word being fixated (in foveal view), as well as for the next word (in parafoveal view). As we can see from the table, letters and their corresponding phonemes are activated (orthographic and phonological decoding) for both the word in foveal view and the next one that is in parafoveal view. Morphological decoding of morphologically complex words occurs primarily in the

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

Table 4.1 Factors that have been investigated with regard to their ability to influence processing when words are in foveal (*n*) and parafoveal view (*n*+1), with 'yes' indicating that it affects processing, 'mixed' that the evidence is contradictory and '?' that it has not been reported (based on Hyönä, 2011; Rayner, 2009).

	Foveal processing	Parafoveal processing
Orthographic and phonological decoding Identifying letters and activating corresponding phonemes	Yes	Yes
Morphological decoding Identifying a word's morphemes	Yes	Mixed
Word length Longer words = more processing	Yes	Yes
Word frequency Words encountered less frequently = more processing	Yes	Mixed
Neighbourhood size	Yes	Mixed
Predictability More predictable word is in context = less processing	Yes	Mixed
Number of word meanings Words with more meanings = more processing	Yes	Mixed
Age of acquisition (AoA) Later learned words = more processing	Yes	?
Familiarity Less familiar words = more processing	Yes	?

fovea; evidence is mixed about whether it occurs for word *n*+1, and this may vary across languages. The length of words in foveal and parafoveal view influences looking times, with longer words leading to more and longer fixation times. Further, readers are able to judge the word length of upcoming words to around fifteen letter spaces to the right of a fixation. Information about upcoming word length, as well as context, may narrow down 'choices' about lexical candidates for words in parafoveal view, which can influence the length of saccades and fixation times. Put more simply, if word length and context allow a reader to make a good 'guess' about the next word, fixation time on it may be shorter because little effort may be needed to confirm the guess. In some cases the next word might be skipped altogether, thus a saccade will go to *n*+2 instead of *n*+1. This means that information about a non-fixated word can influence saccades and fixations to the words around it. Additionally, word frequency, age of acquisition (AoA; the age at which a word was learned), neighbourhood size,[2] predictability (how predictable a word is, given the prior context) and number of word meanings all affect the processing of the word in foveal

[2] Orthographic or phonological neighbours are words that differ from one another by only one letter or sound respectively. For example, the word 'hint' has the neighbours 'hilt', 'hind', 'hunt', 'mint', 'tint', 'flint', 'lint', 'pint', etc. The processing of a word has been shown to be influenced by the number of neighbours (neighbourhood size) and sometimes the frequency of its neighbours (Perea and Rosa, 2000), as well as the regularity of the grapheme-to-phoneme correspondence of a word and its neighbours, e.g. 'hint' and 'pint' (Sereno and Rayner, 2000).

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

view (*n*). There are mixed findings about whether words in parafoveal view (*n*+1) that differ along these dimensions influence the processing of the fixated word.

Overall, Table 4.1 demonstrates that a word in foveal view and the subsequent word, which is in parafoveal view, are both decoded in terms of their orthographic and phonological form. There is clear evidence that the specific word properties of the word being fixated (e.g. frequency) influence the number and duration of fixations to it. What is debated is whether or not, and the extent to which, *parafoveal-on-foveal effects* exist. More precisely, do the characteristics (e.g. frequency) of the word to the right of fixation influence the duration of the fixation to the currently fixated word? This question has important implications for models of reading and eye-movement control that often make claims about whether reading is serial (one word at a time) or parallel (more than one word at a time). For example, if the frequency or the ambiguity (multiple meanings) of the next word lengthen fixations to the current word, this would provide evidence of parallel processing.

What might the debated parafoveal-on-foveal effects mean in more practical terms? Let's say we are interested in looking at how readers process new words, so we invent some words and embed them in a context, like the one in Example 4.1. We define two regions of interest (ROIs), which are the known verb 'spoke' and novel verb 'woused'. Both occur in a similar structure, and so are well matched in that regard: pronoun + verb (word vs new word) + adverb + prepositional phrase. If we find longer and more fixations associated with 'woused' compared to 'spoke', this might lead us to conclude that new words require more processing effort. However, the increased fixations on 'woused' could be due to the low frequency of the following word 'vituperatively'. There is in fact (mixed) evidence indicating that if the word after a fixated word is low frequency, reading times on the fixated word increase (for a discussion of parafoveal-on-foveal frequency effects see, Hyönä, 2011). Thus, the long reading times on 'woused' could at least in part be due to a parafoveal-on-foveal effect, which would call into question our conclusion. Ideally, when comparing ROIs, the surrounding context should be identical or extremely well matched so that an upcoming word does not 'contaminate' the fixation times for our ROI (see Chapter 3 for more on preparing stimuli).

Example 4.1 An example passage to compare processing effort (reading times) for the known word 'spoke' and the non-word 'woused', which are the ROIs (indicated with smooth underlining). The adverbs following the ROIs (indicated with wavy underlining) are not well matched in word frequency.[3]

Lisa worked for an employment agency developing advice for job seekers. She spoke extensively about her own experiences and she woused vituperatively about her previous boss.

Using the word 'vituperatively' in Example 4.1 is a bit extreme. However, it highlights an important concern in experimental design. The words next to our ROI may influence the number and duration of fixations to it. Therefore, we need to ensure that all of the factors outlined in Table 4.1, which might lead to parafoveal-on-foveal effects, are accounted for. This can be achieved in different ways. We can make sure that our ROIs are embedded in identical contexts. As we discussed in Chapter 3, this is generally achieved by creating counterbalanced 'lists' of the different versions of our stimuli. For the current question about

[3] The ROIs, which are clearly demarcated in all of the examples, would not generally be visible to participants.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

the processing of a new word like 'woused', it would mean that a sentence like 'She spoke extensively about her own experiences' would appear on one list and 'She woused extensively about her own experiences' would appear on another list. This way 'spoke' and 'wouse' occur in identical contexts, and fixation times to them should not be 'contaminated' by other factors. Alternatively, we can design contexts such that the words surrounding our ROIs are as well matched as possible on the properties listed in Table 4.1. In Example 4.1, this means that the words that occur after 'spoke' and 'woused' should be as similar as possible in terms of frequency, length, etc. Even if the words that occur after 'spoke' and 'woused' are well matched, they will not be identical, which could cause differences in fixations to our ROIs. Thus, analyses, like mixed effects modelling, that allow us to account for such differences should be considered (we discuss this further in Chapter 7).

Finally, we may want to look at reading of authentic texts like novels, newspapers, graded readers, etc. In such cases we will not be able to control factors as we normally would in a study. In these instances, analyses like mixed effects modelling that allow us to consider the potential influence of the word properties of an upcoming word will most likely need to be used. In Section 4.3 we will look at some studies examining the reading of an Agatha Christie novel, and we will consider in detail some of the methodological concerns associated with the use of authentic texts and ways to address them.

The discussion of parafoveal-on-foveal effects has demonstrated that the word to the right of a fixated word is processed, at least to some extent. Logically, if a word has been identified via parafoveal view, then the eyes do not need to fixate on it and the word can be 'skipped'. The eye-tracking record provides evidence that the processing of a skipped word takes place when a reader is fixated on the previous word: when a word is skipped, the fixation on the immediately preceding one is longer than when it is not skipped (Kliegl and Engbert, 2005). It turns out that readers only directly fixate about 70 per cent of the words in a text and skip the other 30 per cent (Schotter, Angele and Rayner, 2012). Different factors seem to lead to skipping, which Rayner (2009) provides a very thorough overview of. Content words are fixated about 85 per cent of the time, while function words are only fixated about 35 per cent of the time. The strongest predictor of skipping is word length. Words that are two to three letters long are fixated about 25 per cent of the time, while words that have eight letters or more are almost always fixated. Further, when two or three short words occur after a fixated word they might all be skipped. Words that are highly predictable from the previous context are often skipped, even when the word is long. If a high-frequency word occurs to the right of a fixated word, this can also lead to skipping. Finally, skipping may sometimes be due to oculomotor error, in other words a mislocated fixation (for an example of an oculomotor error see the example of 'overshooting' in Figure 1.1).

Why might skipping be relevant to applied linguists? We could have a research question that involves an ROI containing a word that is short, highly predictable, frequent and/or a function word. This would make our ROI susceptible to skipping. Looking at Example 4.2, we might be interested in how proficient and less proficient readers (i.e. children, low-proficiency second language learners, etc.) resolve pronouns in different contexts. As the previous discussion demonstrates, if the eyes land on the word 'because' or 'since' in any of the sentences in Example 4.2, the following pronoun 'he'/'she' is likely to be skipped because it is a function word (and short). If we simply set 'he'/'she' as our ROI, 35 per cent of our cells are likely to be missing data, which could skew our results. Further, our analysis would not reflect the processing effort for 'he'/'she' that occurred on the previous word 'because'. Thus, in cases like this one, the ROI generally includes our actual word of interest, as well as the one that precedes it.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

Example 4.2 Sentences to explore how readers process pronouns in contexts making 'he'/'she' more or less ambiguous. The word of interest is a short function word that will often be processed during fixations to the previous word; thus, the ROI (underlined) should include the word that precedes it. Sentences would generally be presented across two experimental lists (L1 = list 1; L2 = list 2).

L1(a) Jill threw the ball to Sara since she liked to play catch.
L2(a) Tom threw the ball to Sara since he liked to play catch.
L1(b) Bill sold the bike to Sally because he needed the money.
L2(b) Jill sold the bike to Sally because she needed the money.

Something else that should be considered when looking at processing in sentences like those in Example 4.2 is that our critical ROIs differ in length. The sentences with 'she' are one character longer than those with 'he'. Because word length influences processing, this length difference makes our comparison and any conclusions we draw from it problematic. We can make character length corrections, to deal with the difference – dividing a reading time measure by the number of characters. This kind of transformation is questionable when there are only small differences in length, as in this example. In such cases, it is preferable to use a residual reading time, which is a way of comparing the raw reading times for each target word to the average reading time on a per-participant basis (for a discussion as well as an explanation of calculating residual reading times see Trueswell, Tanenhaus and Garnsey, 1994).

Example 4.2 looks for an effect of the pronoun on reading times, which may only be visible on the preceding word. Here it makes sense to have a longer ROI that includes the pronoun and the word right before it. In other cases, it may be the skipping behaviour itself that is of interest. For example, we may want to investigate how predictability influences reading patterns in skilled and less skilled readers. Consequently, we may need to look at parafoveal-on-foveal processing, as well as skipping rates. Therefore, we would set two ROIs – one that only includes the word of interest, and another that includes the preceding word(s). In Example 4.3, we might hypothesise that the word 'wolf' in the idiomatic expression 'to cry wolf' (meaning to raise a false alarm) is predictable in this context. To test this, we would look at fixation times and skipping rates to the word 'wolf' in both List 1 and 2. Because predictability can influence processing on the previous word, it would be important to set 'cried' and 'read' as another ROI. Finally, when skipping rates are high, particularly in one condition and not another, it is important to consider how 'empty cells' (cells with no data because there is no fixation time) might influence the pattern of results. This is a topic that will be taken up in Chapter 7 when we discuss data analysis in more detail.

Example 4.3 Sentences exploring the influence of predictability. The hypothesised predictable word (smooth underlining) and the preceding word (wavy underlining) are the ROIs. These would generally be presented across counterbalanced experimental lists (List 1/List 2).

List 1 Joey repeatedly called the police for no reason.
He had cried wolf too many times for his accusations
to be taken seriously.
List 2 Joey repeatedly practised the sight vocabulary words.
He had read wolf too many times for him to get it wrong
on tomorrow's reading test.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

Thus far, we have had a brief overview of the mechanics of reading and what this means for eye-tracking research. The next section will provide a short summary of what eye-tracking has told us about reading. We will then turn to some methodological considerations when designing studies to investigate different phenomena in reading: known words, new words, multi-word units and syntactic integration.

4.2 What Do We Know about Reading from Eye-Tracking?

As we saw in the previous section, when we read our eyes move along a line of text, fixate a word and move to fixate another word. Broadly speaking, during fixations readers need to 'recognise' a word and 'integrate' this word into a larger sentence or context (Clifton et al., 2007). The distinction between word recognition and integration is helpful for our discussion because it can roughly be mapped onto the classification of 'early' and 'late' eye-tracking measures discussed in Chapter 3. However, it is important to recognise that not every element of reading can be classified neatly as recognition or integration. For example, a predictable word is easier to identify, but the predictability of a word relies on comprehending a context greater than this single word. That being said, by and large effects of recognition are seen in early eye-tracking measures and effects of integration are seen in late measures.

There is an extensive literature making use of eye-tracking to look at both word identification and integration. About ten years ago, Clifton et al. (2007) identified one hundred articles that just looked at word integration, focusing on the effects of syntactic, semantic and pragmatic factors on sentence comprehension. The eye-tracking literature on word recognition is at least as large. Clearly, it is impossible to comprehensively review such an extensive literature in a short space. In what follows, we will simply look at some of the key findings about word recognition and integration in reading, as well as considering how and why the findings might be important for designing studies.

4.2.1 What Do We Know about Word Recognition in Reading?

As we saw in Table 4.1, there are a number of factors that influence word recognition. More precisely, these characteristics impact the cognitive effort required to recognise a word and their effects are evident in the most commonly reported eye-tracking measures, like: *first fixation duration* – the duration of the first fixation on a word, provided that the word wasn't skipped; *single fixation duration* – the duration on a word when only one fixation is made on the word; and *gaze duration* – the sum of all fixations on a word prior to moving to another word (Clifton et al., 2007). Notably, these measures are all 'early'. Chapter 3 provides a more comprehensive discussion of eye-tracking measures and their classification as 'early' or 'late'. In what follows, we will briefly look at some of the main characteristics that have been shown to impact word recognition and examine how they influence eye-movement patterns. We will also consider some additional factors that may be important in more applied reading contexts.

The amount of time spent fixating a word is influenced by its frequency, which is established using corpora like the British National Corpus (BNC) or the SUBTLEX (Brysbaert and New, 2009). Readers look longer at infrequent than frequent words (Rayner and Duffy, 1986). However, differences between high- and low-frequency words disappear after three repetitions in a text (Rayner et al., 1995). It is important to note that corpus-derived frequency measures do not tell us how often a specific person

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

has encountered a word. Thus, there can be high-frequency words that a particular reader may not have encountered very many times, as well as low-frequency words that he/she has come across a lot. To find out how familiar particular words are to individuals, we generally ask them to rate a word's familiarity on a Likert scale. When frequency is accounted for, there is still an effect of familiarity, such that less familiar words elicit more looking (Juhasz and Rayner, 2003; Williams and Morris, 2004). Further, words differ in their age of acquisition (AoA). AoA can be determined using large databases (corpora) that were created by asking people to indicate when they think they learned a word; alternatively a researcher can get these ratings for his/her particular stimuli. When frequency is accounted for, there is nevertheless an effect of AoA, with later-learned words increasing fixation times (Juhasz and Rayner, 2003, 2006).

For morphologically complex words like compounds ('backpack'), we need to consider the potential influence of the frequency of the compound as a whole, as well as that of the constituent morphemes ('back' and 'pack'). Evidence for frequency effects for each of these elements is somewhat mixed, but in general frequency effects for the different elements are found at different points in the eye-movement record, with the first morpheme having an earlier effect than the second morpheme and the whole compound (for a discussion see Hyönä, 2015; Juhasz, 2007). Another factor to consider if working with compounds is their transparency – whether we can determine the meaning of the whole based on the parts (e.g. 'backpack' vs 'hogwash'). There is limited evidence for an influence of transparency, which has only appeared in gaze duration (Hyönä, 2015; Juhasz, 2007). Significantly, as Hyönä (2015) points out, most of the research to date on multi-morphemic words has focused on compounds and has been done on a very limited set of languages (Finish, English, German and Dutch). Considerable work needs to be done to investigate the role of morphology in word identification for derived and inflected words, as well as in a wider range of languages.

Word identification effects that depend on more than the word that is currently being fixated are related to predictability and plausibility. Predictability can be conceived of in two ways. It can be a word-to-word contingency that is established using a corpus, which is often referred to as the 'transitional probability'. For example, the BNC can be used to determine how likely 'on' is following 'rely'. Alternatively, predictability of a word is related to a sentence's contextual constraint or bias. This is established via a cloze task.[4] For both types of predictability, more predictable words have decreased first fixation duration or gaze duration, and highly predictable words are often skipped altogether (Ehrlich and Rayner, 1981; McDonald and Shillcock, 2003b).

Plausibility has to do with how reasonable a word is in a particular context, which often makes use of our real-world knowledge. For example, how reasonable is it to cut carrots with a knife, axe or table? The plausibility of each of these instruments for cutting carrots would be established using ratings. Both implausible and anomalous words increase fixation times; however, an anomalous word has an immediate effect on fixations, while the effect of implausibility is evident somewhat later (Rayner et al., 2004).

A prominent factor affecting word recognition is lexical ambiguity, or the number of meanings that a word has (e.g. 'bank' – a place where you keep money; and 'bank' – the side of a river). The frequency of meanings can be balanced (relatively equal) or

[4] In a cloze task, word(s) are missing and participants are asked to provide them, e.g. 'He made many accusations, but had cried _____' versus 'He practised all of his vocabulary, and had read _____.' The percentage of 'wolf' completions in both sentences would serve as a measure of its predictability in each context.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

unbalanced (one meaning is more frequent). The frequency of the two meanings *and* the strength of the (biasing) context influence eye-movement patterns and lead to what has been termed the 'subordinate-bias effect' (Duffy, Morris and Rayner, 1988; Sereno, O'Donnell and Rayner, 2006). The basic finding is that when frequency-balanced ambiguous words occur in neutral contexts, readers look at them longer than unambiguous control words matched on length and frequency. However, if they occur in a context biased towards one of their meanings, there is no difference in looking times between frequency-balanced ambiguous words and matched control words. In contrast, there is no difference in fixation times for frequency-imbalanced ambiguous and matched control words in a neutral context. If a context biases a less-frequent meaning, fixations are longer for the ambiguous word than the control. The effect on reading times resulting from having multiple meanings seems to be true of words that share orthography and phonology ('bank'), share orthography but differ in phonology ('tear'), and share phonology but differ in orthography (e.g. 'boar'/'bore'; Carpenter and Daneman, 1981; Folk, 1999). If the ambiguous words have different parts of speech (e.g. 'duck' as a noun and 'duck' as a verb), and the subordinate meaning is the only syntactically permissible continuation of a sentence, then there appears to be no subordinate meaning 'cost' on fixations (Folk and Morris, 2003).

Words can also share lexical properties across languages. Thus, just as 'bank' has two meanings, the word 'coin' has two meanings for someone who speaks English and French. It just so happens that the two meanings come from different languages – 'coin' in English is a 'piece of metal used as money' while in French it is a 'corner'. Words like 'coin' are referred to as interlingual homographs, and share orthography (and often have similar phonologies) but have distinct meanings. Interlingual homophones share their phonology, but not their orthography and semantics. These are words like 'pool'/'poule' in English and French respectively, which we saw illustrated in the visual-world paradigm in Chapters 1 and 3. Words can also share their form in terms of phonology and/or orthography, and have more or less the same meaning across languages. These are words like 'table' in English, 'table' in French and 'テーブル' ('teburu') in Japanese,[5] and are referred to as cognates. Interlingual homographs and homophones, as well as cognates, have been used extensively to investigate whether speakers of two languages selectively activate a single language or non-selectively activate both when reading in one language. For example, if a speaker of English and French reads, 'Jen found a coin while walking', 'coin' would only lead to longer fixations than a matched control word (much like 'bank') if the meanings from both languages are activated.

In reading, the influence of shared lexical properties across languages has predominantly been studied using interlingual homographs and cognates (for a review see Whitford, Pivneva and Titone, 2016). In such studies, participants are speakers of two languages, and their processing time for interlingual homographs and cognates is compared to that of language-unique matched control words. In general, cognates have shorter fixations than control words (Balling, 2013; Cop et al., 2017b; Libben and Titone, 2009; Van Assche et al., 2011). However, this does not appear to be the case

[5] In linguistic terms, Japanese words like 'テーブル' ('teburu') 'table' are 'loanwords' because they have been borrowed into the language. Importantly, it is the overlap in form and meaning that underpins the processing advantage for cognates and *not* their linguistic origins. Because loanwords share form and meaning, they are treated as cognates in the processing literature.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

when orthography is not shared across languages (Allen and Conklin, 2017). In contrast, interlingual homographs have longer fixations than control words (Libben and Titone, 2009). For both types of words, these differences are evident in early measures of reading, but can persist through the late measures. These effects are modulated by the amount of orthographic overlap between the words in the two languages and can be influenced by the biasing strength of the sentence context (Whitford et al., 2016).

The research investigating eye-movements during reading has provided us with a good understanding of how readers identify words and how specific lexical characteristics contribute to processing effort, as well as the eye-tracking measures that are most likely to reflect this. The literature reviewed in this section has primarily focused on studies done with skilled, adult monolingual readers because reading research thus far has primarily examined this population. However, as the final part of the discussion indicates, research is beginning to look at other types of readers and situations – for example, reading in a second language. When working with and researching different populations, it is important to first have a good understanding of what 'typical' reading behaviour looks like. This allows us to see, for example, whether reading in a second language demonstrates similar effects for properties like frequency, length, etc. We could explore questions about AoA for a set of participants that learned the word 'dog' at age one but the word 'chien' in French at school at age twelve. Will 'chien' demonstrate AoA effects commensurate with words learned at twelve or one? Researchers are beginning to explore questions like these, and many others, with diverse populations of readers.

4.2.2 What Do We Know about Word Integration in Reading?

As we just saw, word identification processes are largely indexed by early eye-tracking measures. Some effects may persist and be evident in later measures as well. However, the picture is not so clear-cut for word integration and syntactic processing. Difficulties can show up at various points in the eye-tracking record (Clifton and Staub, 2011). Effects are seldom apparent in the earliest eye-tracking measures, and therefore first fixation time is rarely reported; however, they are occasionally found there (e.g. Staub, 2007). Effects can show up as increases in first pass reading times. Sometimes effects of syntactic processes may only show up as an increase in the frequency of regressions, increased go-past times and second pass reading times, or as increased times in the following region (what is called a 'spillover' effect; for a discussion see Section 3.1.2). Because effects can be found almost anywhere, it is worth inspecting data for a wide range of measures. The following provides a list of the commonly reported measures for word integration and syntactic processing (based on Clifton et al., 2007):

- first pass reading time
- go-past or regression path duration
- regressions out
- second pass reading time
- total reading time
- first fixation duration (sometimes reported, but the region should be short, often for spillover effects on an ROI of a single following word).

Why are effects found in such a wide range of measures when exploring word integration and syntactic processing? This may be in part due to experimental design. In studies of

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

word recognition, ROIs are for the most part single words. In contrast, studies of syntactic processing may have ROIs that are two to four words long, and sometimes even longer. This greater variability in the length of ROIs leads to more variability in where and when an effect will appear in the eye-tracking record (Clifton et al., 2007). Looking at the sentences in Example 4.4, if we wanted to investigate word recognition for known and unknown (in this case invented) words, we would define 'copudamets' and 'binoculars' as our ROIs in (a) and (b), and compare the processing of the two. If we were interested in how the with-phrase is integrated into the sentence in (c), we would likely define the two words 'with binoculars' as our ROI.

Example 4.4 ROIs (underlined) for word recognition (a and b) and integration (c).

(a) John feared the man with copudamets hiding in the bush.
(b) John saw the man with binoculars hiding in the bush.
(c) John saw the man with binoculars hiding in the bush.

Another reason for the variability in where effects are found is that there are more options for readers when processing difficulties that arise due to (im)plausibility, complexity or syntactic misanalysis compared to when a word is difficult to recognise (Clifton et al., 2007). In the sentences in Example 4.4, when the eyes arrive at 'with binoculars' in (b) and (c), the sentence is ambiguous because the phrase could modify two things. It may be that 'John saw the man by using binoculars himself' or 'John saw the man who had binoculars.' The eyes can go back in the text to consider these two possibilities, they can stay where they are to work out the ambiguity, or they can move ahead hoping that upcoming information will resolve the issue or that it will turn out that it doesn't matter. In (a), when the eyes arrive at the unknown 'copudamets', there is nothing in the sentence to go back to that will help resolve the difficulty. The eyes can simply stay where they are and try to puzzle out the word, or they can move forward hoping the meaning becomes clear or that it doesn't matter.

Additionally, a wider range of factors contributes to word integration in sentence processing than in word recognition. As Clifton et al. (2007) say, 'We are far from understanding how these factors are coordinated and whether their coordination is modulated by differences in a reader's abilities and strategies' (p. 367). They go on to say that 'the greater flexibility in dealing with sentence comprehension difficulty and the wide range of factors that affect it could mean that high-level processing shows up in the eye-movement record in a variety of different ways, with any one effect appearing only occasionally' (p. 367). Furthermore, word integration and sentence comprehension may be impacted more by task demands and the goals of the reader than word recognition, which would also lead to greater variability in where effects are found. Again, this means that when exploring issues of morpho-syntactic processing, we may need to explore a range of different eye-tracking measures to find effects.

Thus far in this chapter, we have explored some basic properties of the visual system and how they impact reading. We have also discussed some of the main factors known to influence reading. More particularly, we have considered two elements of reading – word recognition and integration – and looked at where and when their effects might show up in the eye-movement record. As we have seen, there is a long history of reading research using eye-tracking. In what follows we will look in greater detail at some specific examples from the literature that demonstrate how we can implement eye-tracking to investigate the

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

reading of words and meaning (Section 4.3); new words (Section 4.4); multi-word units (Section 4.5); and sentences and morpho-syntax (Section 4.6). In each section we will also focus on a different experimental paradigm and consider some methodological concerns for each: authentic texts (Section 4.3); created texts (Section 4.4); the boundary paradigm (Section 4.5); and matched and counterbalanced sentences (Section 4.6). Finally, we will look in more detail at two examples of matching participants from specific populations: second language readers (Section 4.3) and participants with an autism spectrum disorder (Section 4.6).

4.3 Words and Meaning

In this section we will consider a set of studies by Cop and her colleagues that primarily explored word recognition in monolingual English speakers and non-natives in their L1 (first language) Dutch and L2 (second language) English. Participants were asked to read the entire Agatha Christie novel *The Mysterious Affair at Styles* (Dutch title: *De zaak Styles*) while their eye-movements were monitored. Although the use of authentic materials, like a novel, adds an element of authenticity to any study, it leads to other important methodological considerations, which will be the primary focus of the discussion in this section. Authentic materials also provide a wealth of data that can be examined for a variety of things. Thus, across a series of papers, Cop and her colleagues have been able to explore a number of things like word frequency effects and the influence of cognates on reading, as well as sentence reading patterns. Cop and colleagues' four publications based on this dataset will form the basis of the discussion in this section: Cop et al. (2015a), Cop et al. (2015b), Cop et al. (2017a), Cop et al. (2017b).

4.3.1 Contextualising an Example Study: Reading an Authentic Novel

As we saw in the opening sections of this chapter, we know a considerable amount about what influences reading in well-educated, adult readers with no history of language impairments when they read in their L1. Table 4.1 listed a set of factors that have been shown to influence fixations when a word is in foveal and parafoveal view. However, considerable work remains to determine how these factors might impact reading in a much more diverse population of readers. The research by Cop and her colleagues begins to address this gap. They had monolinguals and non-native speakers read the same novel, which has allowed them to explore whether non-native speakers show similar effects in their L1 and L2, as well as being able to compare monolinguals and non-natives in their L1. With their large dataset, Cop and colleagues may eventually be able to explore many of the factors listed in Table 4.1, although more explicit manipulation may be required to explore some of them. It is important to note that the data reported in the four papers listed above all come from the 'same' reading study. For us, this means that the overview of the participants and materials is largely identical across their papers, and the discussion that follows draws on information from all of the papers.

In Cop et al.'s study, each participant read the entire Agatha Christie novel in four sessions of an hour and a half each. In the first session participants read Chapters 1 to 4, in the second 5 to 7, in the third 8 to 10 and in the fourth 11 to 13. The monolinguals read the book entirely in English, while the non-native participants read Chapters 1 to 7 in one language and 8 to 13 in the other. The order of the two languages was counterbalanced, such that half of the participants started reading the novel in Dutch and finished reading in

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

English and the other half started reading in English and finished in Dutch. Participants read the novel silently while their eye-movements were monitored. They were given multiple-choice comprehension questions to answer at the end of each chapter.

4.3.2 Matching Second Language Participants

For most eye-tracking research, whether we are investigating reading, listening, writing or any combination of these, if we are looking at the performance of a particular group, we need to show that the group itself is made up of a set of participants that are similar. The factors we match on will depend somewhat on the focus of the study. For example, it will be more important to match reading ability for a reading study and listening ability in a listening study. If a group is not well matched, this could influence eye-movement patterns and consequently our findings. For example, let's say we want to look at reading by non-native speakers. If we examine the performance of twenty participants, but do not assess and establish their proficiency, we may end up with a group with highly variable behaviour. This will make it hard to find significant results, and we may draw conclusions about non-natives more generally, when our results were driven by a subset of the participants. Thus, we need to convincingly demonstrate that our group is similar (i.e. has a similar proficiency and exposure to the L2). It is important to note that when we investigate reading by educated, adult native speakers with no history of language impairments, we generally assume (maybe incorrectly) that the group is made up of individuals who have similar reading skills. Importantly, if we want to compare groups of participants, we need to show that the two groups are the same and only differ on our manipulated factor. This means that if we want to compare high- and low-proficiency non-native speakers, the only characteristic that the groups should differ on is their proficiency. The groups should have the same number of participants, a similar gender composition, years/level of education, etc.

The studies by Cop et al. provide a very good example of matching participants. They monitored the eye-movements of fourteen English monolingual speakers and nineteen unbalanced Dutch(L1)–English(L2) bilinguals.[6] None of the participants in either group had a history of language or reading impairments. The two groups were matched in age (monolinguals M = 21.8, SD = 5.6; bilinguals M = 21.2, SD = 2.2).[7] They also had a similar level of education, as all participants were enrolled in a bachelor's or master's programme in psychology. The monolingual group had six male and seven female participants, while the bilingual group had two male and seventeen female participants. All of the bilinguals were intermediate to advanced L2 learners with a relatively late L2 age of acquisition (M = 11, SD = 2.46). Both groups of participants in the study completed a battery of language proficiency tasks, which were used to ensure that they were well matched and/or differences could be included as a factor in analyses. These tasks are outlined below:

[6] Cop et al. refer to their participants as bilinguals. It is common in the psycholinguistic literature to refer to any group of non-native speakers as 'bilinguals' and then to simply specify whether they are balanced/unbalanced, early/late, etc. bilinguals. It is important to note that because Cop et al.'s participants were recruited in Belgium it is likely that many, if not all, of them were actually multilinguals. Knowledge of additional languages could influence processing. Thus, ideally information about proficiency should be obtained for all of the languages known to participants.

[7] M means 'mean' and SD means 'standard deviation'.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

- Spelling test
  - Because there is no standardised cross-lingual spelling test, English spelling was assessed using the WRAT 4 (Wilkinson and Robertson, 2006) and Dutch spelling with the GLETSHER (Depessemier and Andries, 2009).
  - The monolinguals were tested in English, while bilinguals were tested in English and Dutch.
- Vocabulary test – LexTALE (Lemhöfer and Broersma, 2012)
  - This is an unspeeded lexical decision task (decision whether a stimulus is a word or not) that can be administered in English, Dutch and/or German. It contains a high proportion of words with a low corpus frequency. It was developed as a vocabulary test and has been validated as a measure of general proficiency for the three languages.
  - The monolinguals were tested in English, while bilinguals were tested in English and Dutch.
- Speeded lexical decision task
  - Participants were asked to classify letter strings as words or non-words as fast as possible.
  - The monolinguals were tested in English, while bilinguals were tested in English and Dutch.
- Self-report language questionnaire – based on the LEAP-Q (Marian, Blumenfeld and Kaushanskaya, 2007)
  - The questionnaire contained questions about language-switching frequency/skill, age of L2 acquisition, frequency of L2 use and reading/auditory comprehension/speaking skills in L1 and L2.
  - Completed by both the monolinguals and bilinguals.

These tests, and the analyses of them, demonstrated that the monolinguals and the bilinguals were equally proficient in their L1, but the bilinguals had relatively less exposure to their L1 than the monolinguals had to their only language. The L2 (English) proficiency of the participants was lower than their L1 (Dutch) proficiency as assessed in all tasks. Further, the tasks showed that bilinguals varied in their L2 proficiency. For example, based on the norms reported in Lemhöfer and Broersma (2012), two bilinguals could be classified as lower intermediate L2 language users, ten as upper intermediate L2 language users and seven as advanced L2 language users. Proficiency differences were therefore an important factor to consider in analyses of the data.

4.3.3 Using Authentic Materials

When selecting authentic reading materials, it is important to consider a number of things. For one, research should be replicable. To achieve this, other researchers should be able to access the materials used in a study. When making our own materials this is straightforward; they can be made available in an appendix or upon request. When using published materials, like a book or novel, there may be copyright issues. While one researcher may receive permission to use a particular text for research purposes, this does not guarantee that others would be granted the same access. Therefore, others may not be able to replicate and/or extend the research with other populations. Ideally, when using authentic materials, we should use ones that are freely available or which can be made widely available to others. Cop et al. used a novel that is accessible as part of Project Gutenberg

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

(www.gutenberg.org), which offers over 50,000 free books in various languages. Additionally, because Cop et al. wanted to investigate Dutch–English bilingual readers, an important consideration was selecting a novel that had been translated into Dutch. Further, their chosen novel is available in other languages, which means that future research can extend Cop et al.'s work by investigating reading of the same text in other languages.

When we create materials, we can design them to fit a set of particular constraints. In contrast, when using authentic materials, we need to select a text that meets whatever criteria we feel are important for our study. For example, Cop and her colleagues wanted to use a novel that could be read within four hours. If an average adult usually reads 250–300 words per minute, or 15,000–18,000 words per hour, this means that they could consider texts that have 60,000–72,000 words. Keeping in mind that this is a very general estimate, and there will be readers who are slower than average, particularly in the L2, choosing a book that comes in under the lower limit would probably be advisable.

After coming up with a set of books that met the length criterion, they checked them for difficulty in a number of ways. First, they wanted a book in which the frequency distribution of the words in the text was as similar to natural language as possible (as established by the SUBTLEX database). They used the Kullback-Leibler test (Cover and Thomas, 1991), which measures the difference between two probability distributions, to establish the divergence between potential texts and the database. Second, they looked at the number of hapax words (words that occur only once in the SUBTLEX database), and chose a novel that had a low number of hapax words. Finally, they calculated two readability scores: the Flesch Reading Ease (Kincaid et al., 1975), which returns a score between 0 and 100 (closer to 100 is easier to read); and the SMOG grade (McLaughlin, 1969), which indicates how many years of education are needed to understand a text. The Flesch Reading Ease for the novel was 81.3, and the SMOG was 7.4.

Because the bilinguals would read part of the novel in Dutch and part in English, Cop et al. compared the novel's translations on a number of characteristics: number of words; number of word types; number of nouns; number of noun types; number of sentences; number of words per sentence; number of characters (letters) per sentence; number of content words per sentence; average word frequency; average content word frequency; and average word length. In cases where *t*-tests revealed differences in these characteristics, and Cop et al. felt that the difference was important for their research question, they included the characteristic as a variable in their analysis.

In the study, the novel was presented in black 14 pt Courier New font on a light grey background. The text appeared in paragraphs and the lines were triple spaced. A maximum of 145 words appeared on a screen, spread over a maximum of ten lines. When readers came to the end of the screen they pressed a button to move onto the next one. Calibration was done at the outset and every ten minutes, or more frequently if the experimenter deemed it necessary.

4.3.4 Data Analysis and Results

Usually when we design a reading study, we try to ensure that our ROIs do not occur at the beginning and the end of a line, or right before or after punctuation. Studies using authentic texts usually exclude fixations that occur in these positions. Thus, generally we would remove the first fixation on every line, as well as any regressive fixations immediately following this fixation, as they are most likely corrective saccades triggered when the return

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

sweep falls short of the beginning of the line (Hofmeister, Heller and Radach, 1999). All data associated with words preceded or followed by punctuation, as well as the first and last word of every line of text, should be excluded from analyses. After this, normal data cleaning procedures can be implemented, for example removing cases of track loss or other irregularities (see Chapter 7 for more on data cleaning). In addition to these procedures, Cop and colleagues removed fixations shorter than 100 ms. In their three studies focusing on individual word processing, for the most part, single fixations that differed by more than 2.5 standard deviations from the subject means per language were excluded from the dataset prior to any analyses.

Furthermore, Cop et al. constrained their datasets in certain ways to make them more manageable and to ensure that they were making relatively well-matched comparisons. When looking at individual word processing, Cop et al. usually excluded words that were identical cognates. Because cognates are known to speed processing, including them could skew the findings. Ideally, non-identical cognates should be removed as well. However, these are harder to identify through an automated process, which probably explains why they were retained in the dataset. Many of their analyses also excluded words that were not fixated and that were fixated more than once. While this helps restrict a very large dataset, the theoretical motivations for doing so should be made clear.

While not directly relevant to the focus of this section on word processing, when looking at sentence processing, Cop et al. did a number of things to match their items, and thereby restrict their dataset. As a result only 4.2 per cent of the original sentences were retained in the analyses, although the set of data still encompassed 210 sentences per participant – and were considered to be the set of optimally matched stimuli by the researchers. The following summarises their main inclusion/exclusion criteria for their study on sentence processing.

- Sentences were removed if they had
  - more than thirty-five words
  - an average word length of more than 7.4 characters
  - an average content word frequency lower than 1.56.
- Matched materials on semantic context:
  - each sentence was manually checked for translation equivalence
  - sentences that did not match were excluded from analyses.
- The final dataset only included Dutch–English sentence pairs that were matched on
  - average word length
  - number of words per sentence.

The measures that Cop and her colleagues used to analyse their data varied depending on whether individual words or whole sentences were the focus of the investigation. Although this differed slightly across studies, for investigations of word processing, Cop et al. examined the following measures: first fixation duration, single fixation duration, gaze duration, total reading time, go-past time and average skipping probability or skipping rate. When looking at the reading of the entire sentence, they explored: whole-sentence reading time including fixations and re-fixations, total number of fixations in one sentence, average fixation duration of the fixations in one sentence, average rightward saccade length per sentence, probability of making an inter-word regression towards or within a sentence, and probability of first pass skipping. For both words and sentences, the eye-tracking measures that Cop et al. used are ones that are commonly reported.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

Unsurprisingly, across a series set of studies, Cop and her colleagues had a number of interesting findings. Here we simply sum up some of their main ones. They demonstrated that monolinguals and bilinguals in their L1 have similarly sized frequency effects. However, in their L2, bilinguals exhibit a considerably larger frequency effect. A cognate processing advantage showed up in different places in the eye-tracking record, depending on language (L1/L2), whether the cognates were identical or not, and word length. For sentences, monolinguals reading in their only language and bilinguals reading in their L1 did not differ in any key ways. However, in their L2, bilinguals had longer sentence reading times, more fixations, shorter saccades and less word skipping.

Importantly for our discussion of methodology, the Cop et al. studies demonstrate the kinds of things that we need to do (e.g. appropriate line spacing and font size) to ensure that we get good data that is not contaminated by various factors (excluding the word before punctuation to remove potential sentence wrap-up effects, etc.) when using authentic materials. Their research highlights different ways we can achieve a level of experimental control for stimuli that we have not specifically designed for a study. However, doing this means that much of the data that we collect will not actually be analysed. While we will have reading time measures for all of the words in a text (except for those that are skipped), only a subset of these will be included in any analyses. Furthermore, we will almost certainly need to use statistical analyses, like mixed-effects modelling, that allow us to account for the range of variables that influence reading patterns.

4.4 New Words

The opening sections of this chapter described word recognition and integration in reading and the factors that influence them. While the discussion focused on known words, children in their L1 and people reading in their L2 will often encounter unknown or 'new' words. By definition, new words have zero frequency for the reader, are completely unfamiliar and are certainly not predictable – all factors that are known to influence processing. Since their meaning is also unknown, integrating a new word into a reader's unfolding understanding of a sentence should also be challenging. However, little research has been done to track eye-movements to new words. Thus, important questions remain about what happens when people read new words, and how the reading patterns elicited by new words relate to performance on other tasks.

If we would like to study the processing of new words, we can embed unknown and known words in counterbalanced sentences (or longer stretches of discourse) that appear in different lists, or we could present them in a single text in different but well-matched contexts and compare the processing of the two. For such a study, we could use a methodology similar to the one we will see in Section 4.6. Alternatively, we could present unknown words in an authentic text. We would simply replace some of the known words in the text with unknown or non-words, and the methodological considerations would largely be the same as those outlined in Section 4.3. In addition, it would be advisable to assess the informativeness of the context: in other words, how much the context tells the reader about the meaning of each of the unknown words, as this will likely influence looking patterns. Instead of using authentic texts, we could design texts that have particular properties. Crucially, such texts may still feel authentic to readers. By designing our own texts, we can ensure that they are suitable for our target readers,

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

for example, by only using vocabulary that is at an appropriate level. We can also manipulate and control how informative the context is for the unknown words.

Another consideration in such a study are the unknown words themselves. We could use actual but unknown words, or non-words. If we used unknown words, we would need to ensure that they were indeed unknown to the participants. This would usually be verified in a pre-test. However, by including the words in a pre-test, we would expose participants to them, as well as potentially drawing their attention to them, which could influence behaviour on the main task. Alternatively, we could use very 'word-like' non-words. In this case, we would be drawing conclusions about reading behaviour for new words when the input is not actually words. As with many experimental design decisions, there are advantages and disadvantages with our choices – we simply need to be aware of them and consider them when we draw conclusions from our data.

4.4.1 Contextualising an Example Study: Reading a Constructed Text

We will consider a study by Pellicer-Sánchez (2016) that looked at what L1 and L2 readers learn about new words when they encounter them in a constructed text. More specifically, Pellicer-Sánchez was interested in the incidental learning of new vocabulary from reading, as well as how the eye-movement pattern to new words changed over a number of occurrences. To investigate this, she compared the processing of non-words and matched known words embedded in a story. She used a set of offline vocabulary tests to look at what participants learned about the new words and examined how this related to their reading behaviour.

4.4.2 Matching Participants

Pellicer-Sánchez monitored the eye-movements of thirty-seven L2 speakers of English from various language backgrounds and thirty-six L1 speakers of English. Due to cases of drift (i.e. imprecise eye-movements indicating a deterioration of the calibration over time) in the ROIs, data from fourteen L2 participants and eleven L1 participants was discarded and not included in the analyses, leaving twenty-three and twenty-five participants in each group respectively. This is very high rate of exclusion and is *not* typical of reading studies of this nature, but was the result of particular methodological decisions (see the paper for further details). None of the participants in either group had a history of language or reading impairments. The two groups were fairly well matched in terms of age. Both cohorts were well educated, with the L1 participants being drawn from an undergraduate student population and the L2 participants were postgraduate and postdoctoral students. The L2 group had ten males and thirteen females, while L1 group had one male and twenty-four females.

The non-native speaker participants in Pellicer-Sánchez's study were much more diverse than those in the Cop et al. studies reviewed above. Here, the participants came from eleven different language backgrounds that used different scripts (alphabetic languages; logographic languages; syllabic languages or abugidas). When examining performance by participants from same- and different-script languages, it is important to consider how this might impact speed of reading and reading patterns. This can be done by comparing global reading measures to see if same- and different-script readers perform in a similar fashion. All of the non-native speakers had spent a minimum of twelve

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

months and a maximum of six years living in an English-speaking environment (M = 2.4, SD = 1.7). They were advanced learners who had met a university entry requirement of English proficiency (6.0 or above on the International English Language Testing System [IELTS] or equivalent examination). At the beginning of the experiment L2 participants completed a self-rating questionnaire of proficiency (on a scale from one to ten, with ten being native-like). The mean values for all skills (reading, writing, listening and speaking) were all above seven and crucially all participants rated their reading skills at seven or above.

The discussion in Section 4.3.2 goes into more detail about matching and assessing participants who are non-native speakers, and presents the battery of language proficiency tasks that Cop and her colleagues used in their research to demonstrate that the non-native participants were well matched and/or to provide proficiency metrics that could be used as variables in analyses. Ideally researchers should use some of these tasks, or other similar ones, when conducting research with non-native participants.

4.4.3 Using Constructed Materials

When we create our own text, we can ensure that it addresses our research questions while at the same time making sure that it is appropriate for the target group of participants. For her study, Pellicer-Sánchez wrote a 2,300-word story. She carefully controlled the vocabulary in the story to help ensure that the acquisition of the unknown words would not be hindered by lack of knowledge of the remaining words in the text. Thus, 97 per cent of the words in the story belonged to the 3,000 most frequent words of the British National Corpus (BNC; determined by Compleat Lexical Tutor; www.lextutor.ca). Only four words (0.17%) were from the 5,000 to 9,000 frequency bands. These were considered adequate percentages to assure participants' comprehension.

Embedded in the story were six non-words and six control words (real known words), all of which were repeated eight times. This meant that forty-eight words in the text were unknown (2%), while the remaining were known (98%). The non-words were evenly spread throughout the text to make sure that there was a balanced distribution of unknown items throughout the story. To ensure that the new words would be unknown to all participants, non-words (i.e. invented letter strings that look like real words in English) were used. Non-words came from the Compleat Lexical Tutor (Cobb, n.d.), and modified to suit the required length (two syllables, six letters). They all replaced high-frequency (1,000–3,000 frequency band from the BNC), concrete nouns in the story. Further, Pellicer-Sánchez wanted to compare reading behaviour for the new and known words, and she wanted to make certain that any observed effect for the non-words was not simply a practice or repetition effect. Thus, she included six known words, also repeated eight times in the story, that had the same characteristics as the non-words (nouns, six letters and two syllables) and were from the same high-frequency band.

While the text was designed to make sure that the non-words were equally guessable from context, Pellicer-Sánchez wanted to confirm that this was indeed the case. She conducted a separate norming study with eighty-seven native speakers of English divided across eight groups. Group 1 read the first context in which each of the non-words appeared (including the non-word sentence, the previous sentence and the following sentence), Group 2 read contexts one and two, Group 3 read contexts one through three and so forth. Participants were asked to read the context and guess the meaning of the

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

non-words. The results showed that the vast majority of participants provided the same guesses for the non-words (93%–98% agreement for each non-word), indicating that the context was equally informative for all of the unknown words.

The story was presented using 18 pt Courier New font with double line spacing. The text was divided over twenty-five screens, all of which were eight lines long and contained 82–103 words. A screen contained a maximum of two non-words. The position of non-words and known words was carefully controlled so that none of them appeared in initial or final position in a line or sentence. Following the story, there was a short, twelve-question true/false task to check that participants had read for comprehension. None of the questions contained a non-word. Before beginning the main study, to familiarise them with the task, participants read a short 423-word story, which also contained unknown words, and which was followed by comprehension questions. Calibration was done before and after the practice story and halfway through the experiment.

Pellicer-Sánchez's study also involved a set of three vocabulary tests: form recognition, meaning recall and meaning recognition. These were carried out after the reading task. The first vocabulary test assessed participants' ability to recognise the correct form of the non-words. A multiple-choice task presented four different options, and participants were asked to select the correct spelling of the target item. The second test measured participants' ability to recall the meaning of the non-words. Participants were shown items one by one and were asked to say everything they knew about the meaning of the item. A third measure of the form–meaning link (i.e. meaning recognition) was included to capture knowledge below the level of meaning recall. For each non-word, participants were given four possible meanings to choose from, as well as an 'I don't know' option. Careful attention was given to the design of distractors, which were all semantically related to the content of the story (otherwise their discrimination would have been too easy) and were all of the same word class. For all three tasks, participants had to indicate on a scale from one to four (one = very uncertain, four = very certain) how certain they were of their responses. Example 4.5 shows a sample of the created story with two non-words, as well as examples of the vocabulary knowledge tasks.

Example 4.5 The stimuli from the Pellicer-Sánchez (2016) study. Part A shows an example screen containing the two new words, 'holter' and 'soters', which were ROIs (underlined). Parts B and C show the form and meaning recognition task for these two items. Note that the meaning recall task is not depicted here as it allowed participants to provide unconstrained responses.

Part A

Hugo grew up in the holter with the other boys. It was not a very nice place for children. Nobody would like to grow up there. It was a very old place and not very comfortable. The boys never received any care at all. They never had enough warm soters or food. Life there was hard, but living alone in the outside world would be even harder. But living there was not free. Having a place to sleep and a bit of food to eat had a price. They had to work hard if they wanted to stay there.

Part B

Choose the right spelling for the following six words that have appeared in the story (only one is correct) and indicate in the scale on the right how certain you are of your response (1 = very uncertain, 4 = very certain).

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

1.	a) hotler	b) holter	c) houter	d) houler	**1**	**2**	**3**	**4**
2.	a) solers	b) soters	c) sorels	d) sorets	**1**	**2**	**3**	**4**

Part C

Select one of the five options. Only one is the correct definition. If you don't know the meaning of the word, please select option 'e'. Indicate on the scale on the right how certain you are of your response (1 = very uncertain, 4 = very certain)

1. holter **1 2 3 4**
   a) basement
   b) workhouse
   c) prison
   d) food hall
   e) I don't know.

2. soters **1 2 3 4**
   a) shoes
   b) clothes
   c) dishes
   d) tools
   e) I don't know.

4.4.4 Data Analysis and Results

Pellicer-Sánchez began by cleaning her data. Single fixation durations shorter than 100 ms and longer than 800 ms were discarded. (Note that for ROIs with more than one word, an 800 ms cut-off may be too short and overly conservative. For known, single words it is the standard cut-off.) Fixation counts greater than or equal to ten for an ROI were also discarded. This resulted in the loss of 5 per cent of the L2 data (218 fixations out of the total 3,824 fixations) and 6.5 per cent of the L1 data (227 fixations out of the total 3,262 fixations). Pellicer-Sánchez then examined the reading behaviour for the new and known words using the following eye-tracking measures: first fixation duration, gaze duration, number of fixations and total reading time. She explored the eye-tracking data for both the L1 and L2 participants, examining the influence of word/non-word status and number of occurrences. She also analysed the results of the vocabulary post-tests and explored the relationship between reading time measures and vocabulary knowledge.

Results of the study showed that both L1 and L2 participants learned new words from reading. For both groups, non-word reading got faster as the number of encounters increased, and non-word reading looked like that of already known words by the eighth occurrence. It also demonstrated that spending more time reading the non-words led to better meaning recall.

Importantly, the discussion of the Pellicer-Sánchez study highlights some of the methodological considerations when investigating the processing of unknown words, as well as a number of things we should consider when designing our own texts for eye-tracking (e.g. placement of critical words on the page, font size, line spacing).

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

4.5 Multi-Word Units

Eye-tracking while reading can be used to explore the processing of sequences of words, which are referred to in the literature by a number of names, such as 'formulaic sequences', 'multi-word units', 'multi-word expressions', etc. Eye-tracking has been used successfully to investigate idioms ('kick the bucket'), compounds ('teddy bear'), collocations ('strong coffee'), binomials ('fish and chips') and lexical bundles ('I don't want to'). Carrol and Conklin (2014) provide a detailed discussion of some of the challenges posed by ROIs that span multiple words. The main difficulty comes from the fact that such sequences are made up of individual words that at the same time form a unit. This is clearest in the case of idioms like 'kick the bucket', where 'kick' + 'the' + 'bucket' = 'die', while none of the words on their own mean 'die'. To reflect the fact that both the individual words and the whole sequence can contribute to processing, it is advisable to establish ROIs that allow us to analyse the processing of both. There are also some additional variables that should be considered when investigating multi-word units, which we will talk about in the next section.

4.5.1 ROIs and Additional Variables for Multi-Word Units

Two eye-tracking studies on idioms demonstrate different ways of achieving a balance between an exploration of the parts and the whole of multi-word sequences. In an investigation of the processing of three-word idioms by native and non-native speakers, Carrol, Conklin and Gyllstad (2016) defined two ROIs: the whole idiom, and the final word of the idiom. Because the final word may be predictable, and therefore elicit shorter fixations and more skipping, it was thought to be important to consider it by itself. For the whole idiom, Carrol et al. looked at first pass reading time, total reading time and fixation count, while for the final word they analysed likelihood of skipping, first fixation duration, gaze duration, total reading time and regression path duration. In a study on longer idioms, Siyanova-Chanturia, Conklin and Schmitt (2011) used a cloze task to determine the recognition point of idioms – the place where an idiom becomes recognisable. For example, if a native speaker encounters 'the straw that broke' they will recognise the idiom having seen the initial four words. Thus, the words up to and including 'broke' occur before the recognition point, while 'the camel's back' occur after the recognition point. Siyanova-Chanturia et al. defined three ROIs: the whole idiom; words up to the recognition point; and the words after the recognition point. For all three ROIs they analysed first pass reading time, total reading time and fixation count. Because all of their ROIs were made up of multiple words, the same eye-tracking measures were suitable for all of them. In contrast, because one of Carrol et al.'s ROIs was a single word, different measures were appropriate when looking at it.

In addition to considerations about what would be the best ROIs for multi-word sequences, there are some additional variables that may need to be considered when designing materials and/or in the analysis:

- Frequencies of the individual words and of the whole sequence
  These are established with a corpus.
- Cloze probability and/or recognition point
  Both are established with a cloze task and tell us how predictable a final word(s) might be and where the sequence is recognised.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

- Transitional probability (forward and backward)
  This is established with a corpus and tells us how likely it is to see one word in the context of another. Forward transitional probability calculates the likelihood of one word following another (how likely is 'on' after 'rely'), while backward transitional probability calculates the likelihood of a preceding word (how likely that the word is 'rely' given the following word 'on').
- Association strength
  This tells us how strongly words are associated in memory and is established with databases such as the Edinburgh Associative Thesaurus (http://konect.uni-koblenz.de/networks/eat) and the University of South Florida Free Association Norms (w3.usf.edu/FreeAssociation). Association strength is mainly considered as a variable in investigations of collocations and binomials.
- Transparency and/or compositionality
  This tells us whether a sequence's meaning is computable from its parts. For example, 'kick the bucket' is not transparent, while 'fish and chips' is. Transparency and compositionality are generally established using rating and norming studies, and are assessed when the experimental materials contain items for which the meaning is not fully transparent, with idioms being the quintessential case.
- Mutual information score (MI score)
  The MI score shows the relationship between how many times a particular word combination appears in a corpus, relative to the expected frequency of co-occurrence by chance based on the individual word frequencies and the size of the corpus. MI scores are mainly considered as a factor in studies on collocations.

In general, studies of multi-word sequences have the same design and methodological considerations as studies investigating the processing of single words and morpho-syntax. Thus, a study on multi-word sequences would be very similar to those considered in the other sections in this chapter. In other words, we can present multi-word sequences in matched sentence pairs (similar to Section 4.6), we can embed them in a longer constructed reading context (like in Section 4.4), or we can study them as they occur in natural texts (as in Section 4.3). Because these experimental methods are presented in other sections, here we will consider a different experimental technique – the boundary paradigm.

4.5.2 Contextualising an Example Study: Using the Boundary Paradigm

A study by Cutter, Drieghe and Liversedge (2014) explored whether spaced compounds are processed like a large lexical unit during reading. To do this they used the boundary paradigm. The boundary paradigm involves a 'gaze-contingent display' (see Chapter 3 where we discuss types of triggers in eye-tracking studies), such that the stimulus that is being presented is updated as a function of where a participant is fixating. It allows researchers to control the input that the processing system receives to determine how different kinds of information are used. More specifically, there is a display that is triggered to change when the eyes cross a boundary that is invisible to participants. This means that what is viewed via parafoveal preview can be different from what will appear after the eyes cross the boundary. The previewed stimuli can be a number of different things, like a nonsense word in the Cutter et al. study, which is illustrated in Example 4.6. The paradigm allows us to compare how different types of preview affect the fixations on a target, which in turn allows us to establish how much is understood when a word appears in the parafovea and periphery. If having the 'wrong' information doesn't influence fixations relative to when the information is 'correct', then the

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

information in the parafovea or periphery was not processed. In contrast, if a change in the stimulus changes patterns of fixations, it indicates that information in the parafovea affects processing.

Example 4.6 Stimuli from Cutter et al. (2014) making use of the boundary paradigm. The vertical black line represents the position of the invisible boundary. In the study there were four different types of preview. In all cases, when the eye crossed the boundary, the preview was replaced with the correct version of the spaced compound ('teddy bear'). Their three regions of analysis were: *n*, *n*+1 and *n*+2.

ROIs	*n*	*n*+1	*n*+2
(a) The small child gently cuddled his	fluffy	teddy	bear while ...
(b) The small child gently cuddled his	fluffy	teddy	hocu while ...
(c) The small child gently cuddled his	fluffy	fohbg	bear while ...
(d) The small child gently cuddled his	fluffy	fohbg	hocu while ...

In the example from Cutter and colleagues, there is a boundary after the word 'fluffy', which directly precedes the compound. Before participants' eyes crossed this, they had a preview of one of four different types of stimuli: (a) both words of the compound; (b) unmodified first word and modified second word; (c) modified first word and unmodified second word; and (d) both words modified. In all cases, once the boundary was crossed the compound 'teddy bear' appeared. Their ROIs were the word right before the boundary (*n* = 'fluffy'), the two words of the compound (*n*+1 = 'teddy', *n*+2 = 'bear') and the whole compound ('teddy bear'). They were particularly interested in a processing advantage for *n*+1 and *n*+2, which would provide an indication that compounds are processed as a lexical unit during reading.

4.5.3 Matching Participants and Controlling Materials

All of the participants in the study were adult, native English speakers with no history of language impairments, and were therefore assumed to have a similar reading skill. (For discussions of matching participants in populations where reading skill is assumed to vary see Sections 4.3.2 and 4.6.2). Cutter et al. tested sixty-one participants, but only analysed the data for forty-four of them. The data from the other seventeen participants was removed before any analyses were carried out because these participants had noticed the changes in the display. Typically, participants who notice changes in boundary paradigm studies are excluded, and the rate of exclusion in this study is not unusual.

The study made use of forty spaced compounds, which were matched on a number of characteristics. Words occurring in the pre-boundary position (*n*) were matched for length and frequency, as were the words occurring in the first position of the compound (*n*+1), in the second position (*n*+2), as well as the whole compound. The mean forward transitional probability of the compounds was 0.42 in the BNC, which means that the first constituent appeared as part of the spaced compound 42 per cent of the time within the corpus. Cutter and his colleagues performed a cloze task on a similar set of

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

participants to verify that the compounds were equally predictable. When participants were given the sentence up to the word *n*, the compound was produced 33 per cent of the time, and when they were given up to the word *n*+1, they produced the second word of the compound 97 per cent of the time. This indicates that the compound as a whole was not overly predictable, but the second word of the compound was after having seen the first one.

The non-words for the preview condition were generated using an algorithm that replaced letters in the actual words with other letters, thereby preserving the word shape of the original words. In the sentences themselves, the compounds were placed at least two words from the end of the sentence. Crucially, the cloze task showed that the sentences were equivalent in 'predicting' the compound. In the study, the forty experimental stimuli were interspersed among sixty-nine filler items. One-third of the trials had a yes/no question to ensure that the participants were reading for comprehension. Although it is not explicitly stated in the paper, it is assumed that the items were presented across four counterbalanced lists, such that a participant saw ten items in each of the four conditions (unchanged preview, modified first word, modified second word, modified both words).

4.5.4 Data Analysis and Results

As in other studies, Cutter et al. started by cleaning their data. They removed trials where the boundary change happened early and when the boundary change finished more than 10 ms after fixation onset. Trials in which participants blinked during a critical region were removed. Finally, they only analysed trials in which there was a fixation on word *n*. Together, their exclusions account for 44 per cent of the original data. It is important to note that this is considerably more data loss than we see with other paradigms, but is in-line with studies using the boundary paradigm.

Cutter and colleagues had four ROIs: *n*, *n*+1, *n*+2 and the whole compound. Thus, their analysis considered the parts of the compound and the compound as a whole. For all of the ROIs they looked at gaze duration, go-past time and single fixation duration. They also considered first fixation duration for *n*, *n*+1 and *n*+2, but not for the whole compound because this measure is only appropriate for short (single-word) regions. Finally, they only looked at skipping probability for *n*+1, *n*+2, as skipping is irrelevant before the compound occurs. Cutter et al. found that readers spent less time looking at *n*+1 when they had seen 'bear' in a preview condition. This processing advantage was evident when the word 'teddy' was present, but not its matched non-word 'fohbg' had been visible. This was taken as an indication that the second word of a spaced compound is processed as part of a larger lexical unit during natural reading. However, this only occurs when the first constituent is visible and indicates that a compound is present.

The Cutter et al. study introduces a paradigm that may be somewhat unfamiliar. It has been used extensively by Rayner and many of his colleagues to look at effects of parafoveal processing and how expectancies for upcoming words influence processing. Generally in such studies ROIs only include *n* and *n*+1. Because the Cutter et al. study was interested in the processing of compounds, *n*+2 was also an important region of analysis. Further, the discussion in this section highlights a range of variables relating to the individual words that make up the unit, as well as ones that relate to the whole, that we should consider when investigating multi-word units. Importantly, we should establish ROIs that consider the parts as well as the whole of the multi-word unit. Finally, we

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

have seen that different eye-tracking measures are appropriate for ROIs of single versus multiple words.

4.6 Sentences and Morpho-Syntax

Thus far, we have focused on studies looking at word processing. In this section, we are going to turn to a study by Howard, Liversedge and Benson (2017) on syntactic processing. Notably, the study investigated readers who were on the autism spectrum and compared them to typically developed readers. While the main points of interest in this section concern how the participants and materials were matched, we will also consider other methodological aspects of the study.

4.6.1 Contextualising an Example Study: Counterbalanced Materials

As we saw in Example 4.4, a sentence like 'John saw the man with binoculars hiding in the bush' is ambiguous. The prepositional with-phrase can modify how the seeing was done 'with binoculars', or it can modify the man, 'the one with the binoculars'. In the literature on syntax this is called an attachment ambiguity. In other words, a modifier can attach to a higher node in a syntactic tree (referred to as high attachment) or to a lower node in the tree (referred to as low attachment). In our example, the 'seeing' interpretation reflects high attachment, while the 'man' interpretation reflects low attachment. Additionally, it is thought that verbs can have an attachment preference (i.e. a verb may be more likely to have high attachment). Further, real-world knowledge can play a role in attachment ambiguity. More specifically, 'with binoculars' is ambiguous because binoculars can be used for seeing and they are something a person can have. In contrast, in a sentence like 'John feared the man with binoculars', real-world knowledge tells us that the prepositional phrase cannot modify how the fearing was done; thus it can only refer to the man.

The Howard et al. (2017) study looked at how participants with and without an autism spectrum disorder read sentences containing an ambiguous prepositional phrase using verbs that had a high-attachment preference when there was and was not real-world knowledge to help disambiguate the sentence. Howard and her colleagues were trying to determine (1) whether the syntactic preferences (high attachment) held by typically developing readers are true of participants with an autism spectrum disorder, and if the time course of any disruption due to ambiguity is the same for both groups of readers; and (2) whether readers with an autism spectrum disorder use real-world knowledge during reading.

4.6.2 Matching Participants from a Special Population

As we discussed in Section 4.3.2, when we are studying a particular group of participants whose language and reading skills may vary, we need measures that show that the group itself is made up of a set of participants who are similar and/or that allow us to account for any participant differences in our analyses. An excellent example of the care that should be taken when comparing two groups comes from the Howard et al. study. First, they ensured that the group of autism spectrum disorder participants was similar. They did this by means of a standardised test – module 4 of the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012). Using this test, they showed that all but four of their participants met the autism spectrum cut-off criteria. When these four participants

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

Table 4.2 Summary of the ways in which Howard and colleagues (2017) showed that their typically developed participants and those with an autism spectrum disorder only varied on the factor of autism. The first five characteristics demonstrate that the groups were the same, while the sixth shows that they differed on autism spectrum traits.

Participants with an autism spectrum disorder	Typically developed participants
Group composition to demonstrate that they have a similar size and gender make-up.	
19 adults (3 females)	18 adults (4 females)
Comparison of age to show that the groups are the same, $t(35) = 0.94$, $p = 0.354$.	
ages 18–52; M = 31.37 years SD = 10.45	ages 20–52; M = 28.33 years SD = 9.18
Comparison of various IQ measures from a standardised test, the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999) to show that the groups did not differ in IQ.	
verbal IQ $t(35) = 0.58$, $p = 0.621$	
M = 118.32 SD = 11.06	M = 116.50 SD = 11.07
performance IQ $t(34) = 0.99$, $p = 0.331$	
M = 116.21 SD = 14.75	M = 111.94 SD = 11.45
full-scale IQ $t(35) = 0.87$, $p = 0.389$	
M = 119.42 SD = 11.89	M = 116.17 SD = 10.79
Comparison of expressive language ability using the Clinical Evaluation of Language Fundamentals II (Semel, Wiig and Secord, 2003), showing the groups did not differ, $t(35) = 0.53$, $p = 0.599$.	
M = 86.95 SD = 6.22	M = 88.00 SD = 5.87
Comparison of general reading ability with the York Assessment of Reading Comprehension (Stothard et al., 2010), demonstrating that the groups did not differ in single-word reading or reading a passage for comprehension	
single-word reading $t(33) = 0.51$, $p = 0.614$	
M = 68.17 SD = 2.12	M = 67.74 SD = 2.96
passage comprehension $t(34) = 0.35$, $p = 0.727$	
M = 8.97 SD = 2.03	M = 9.19 SD = 1.77
Comparison of autism spectrum traits using the standardised Autism-Spectrum Quotient (Baron-Cohen et al., 2001), to show that the groups differed, $t(32) = 9.24$, $p = 0.001$.	
M = 37.37 SD = 6.10	M = 15.61 SD = 8.03

were excluded from the analyses, the pattern of results did not change. Further, Howard et al. established that their participant groups were essentially the same and only differed on whether they had an autism spectrum disorder or not. All of their participants were English native speakers with normal or corrected to normal vision. Table 4.2 shows that the groups were well matched on the number of participants, age and gender. A set of standardised tests was used to establish that the groups were the same, as well as demonstrating their critical differences. It is important to note that the groups' characteristics were compared using *t*-tests; thus differences or the lack of them were attested statistically.

4.6.3 Matching and Controlling Materials

The Howard et al. study employed a standard experimental design in which everything was held constant, while a particular factor was manipulated and compared (see

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

Example 4.7). First, they ensured that all of their verbs had a high-attachment preference – in other words that the preference was for the prepositional phrase to modify how the action was done. To test this, they asked a set of participants who were from a similar population as their typically developed participants to complete a cloze task: 'Charlie demolished the dilapidated house with _____.' Ninety-eight per cent of the completions modified how the demolishing was done rather than house, confirming that the sentences had a high-attachment preference.

Example 4.7 Experimental sentences from Howard et al. (2017). The lines delineating the main regions of interest (pre-target, target, post-target) would not have been visible to participants. The target region contains the word that was manipulated for real-world knowledge and would encourage a high attachment interpretation as in (a) or a low attachment interpretation as in (b).[8]

	pre-target	target	post-target	
(a) Charlie demolished the dilapidated house with	a huge	fence	last	year.
(b) Charlie demolished the dilapidated house with	a huge	crane	last	year.

Using the SUBTLEX database (Brysbaert and New, 2009), Howard et al. ensured that the nouns appearing in the target region were matched on a number of properties that have been shown to influence reading time: length, frequency, number of orthographic neighbours, mean bigram frequency, number of morphemes and number of syllables (a bigram is a sequence of two adjacent elements and typically refers to adjacent letters, syllables or words). They used *t*-tests to statistically demonstrate that there was no difference in the target nouns in the two conditions (e.g. 'crane' and 'fence' in Example 4.7). To make sure that two conditions did not differ in plausibility, Howard et al. carried out a plausibility rating task with a set of participants from a similar population to their typically developed participants, but who did not take part in the previous cloze task or the main study. They asked the participants to rate how likely it was that an event described in a sentence would occur. To prevent the low attachment of the prepositional phrase acting as a confounding variable, which could cause participants to rate these sentences as less likely, they reformulated the sentences to be an unambiguous description of the events depicted in the sentences: 'that a crane would be used to demolish a huge, dilapidated house' and 'that a huge, dilapidated house that has a fence, would be demolished'. Their rating study showed that the events described in the two conditions did not differ in plausibility.

At the end of their norming procedure, Howard and colleagues had forty-four sentence pairs like those in Example 4.7. The sentences were presented across counterbalanced lists such that each list contained only one version of a pair. All participants saw eighty-eight sentences; of these, forty-four contained an ambiguous prepositional phrase (22 high attachment, 22 low attachment). In addition, ten practice sentences appeared before the

[8] Howard et al. (2017) had additional regions of interest, but here we consider the three that were the main focus of the investigation.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

experimental sentences. The stimuli were presented in a random order, with 50 per cent of them having a following yes/no comprehension question.

4.6.4 Data Analysis and Results

Howard and colleagues used fairly standard data cleaning procedures. They removed fixations of less than 80 ms and more than 800 ms (less than 1% of the original fixation data). Trials with a blink or other disruption were removed (7.95% of the data). Data points that were more than 2.5 standard deviations away from the mean, which was computed individually for each participant per condition, were excluded (less than 3% of the data from each fixation measure). The fixation measures were log transformed and linear mixed effects modelling was used to analyse the data. For skipping and regression rate, logistic linear effects models were computed.

To determine whether there were any basic differences between the two groups when reading syntactically ambiguous sentences, Howard et al. conducted an analysis of global reading measures using mean fixation duration, mean fixation count and total sentence reading time. They then analysed the pre-target region, the target region and the post-target region using the following eye-tracking measures: skipping rate, first fixation duration, single fixation duration, gaze duration, total time, second pass time, regressions in and first pass regressions out.

Howard and colleagues found that both groups of readers demonstrated a comparable reading disruption for low-attachment sentences. This suggests that the participants with an autism spectrum disorder have a syntactic preference for high-attachment sentences that is similar to that of the typically developed participants. Their similar performance also shows that the two groups make use of real-world knowledge in similar ways during reading. However, the participants with an autism spectrum disorder skipped target words less often and took longer to read sentences in second pass reading time, suggesting that they adopt a more cautious reading strategy and take longer to evaluate their sentence interpretation.

Even though this study looked at sentence-level processes, the same methodological concerns apply here as elsewhere. We need to consider whether our stimuli are well matched and designed so that any effects are not contaminated by variables that have nothing to do with our research questions. We also need to ensure that if we are looking at different populations of participants, they are well matched on all of variables except the one that is being manipulated. The Howard et al. study provides us with an example of good practice on all of these counts.

4.7 Conclusions and Final Checklist

There are a number of factors that have been shown to impact word recognition and integration in reading. When designing stimuli for a study or selecting an authentic text we need to carefully consider these variables. In other words, one of the most important aspects in designing a reading study is 'matching' our ROIs and the material that appears around them, and/or having good measures for word and sentence characteristics, so that they can be accounted for in our analyses. Equally important is the 'matching' of our participants. Thus, we need to have appropriate tasks to establish that a cohort of participants is similar, and/or that provide measures that can be used as a variable in an analysis. In this chapter, we considered a variety of different kinds of texts

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

and tasks that we can use to explore reading: authentic texts (Section 4.3); created texts (Section 4.4); the boundary paradigm (Section 4.5); and matched and counterbalanced sentences (Section 4.6). In Chapter 6, we will consider eye-tracking in a wide range of reading contexts: language testing (Section 6.2), writing (Section 6.3), corpus linguistics (Section 6.4), translation (Section 6.5), computer-mediated communication (Section 6.6) and literary linguistics (Section 6.7).

Finally, we opened this chapter with a quote from Huey's (1908) influential chapter on 'The mysteries and problems of reading', so it seems only appropriate to close with another. As Huey says, the average reader does not understand how we read. However, hopefully this chapter has given us, as researchers, valuable insight into reading and provided us with the means to uncover more of the 'miracle' of reading.

> Real reading is still the noblest of the arts, the medium by which there still come to us the loftiest inspirations, the highest ideals, the purest feelings that have been allowed mankind,—a God-gift indeed, this written word and the power to interpret it. And reading itself, as a psycho-physiological process, is almost as good as a miracle. To the average reader the process by which he gets his pages read is not understood. . . (Huey, 1908, p. 5)

Final checklist

• What is the focus of your research?
  • word recognition
  • for individual words
  • for multi-word sequences
  • word integration (morpho-syntactic processing)
	⇨ The answer to this will dictate the kind of matching that needs to be done for the materials, as well as the eye-tracking measures that should be considered.

• What variables should be considered when designing stimuli? • Stimuli need to be well matched and/or differences in stimuli characteristics should be accounted for in analyses.	⇨ For individual words, consider the variables from Table 4.1. ⇨ For multi-word sequences, consider the variables listed in Section 4.5.1. ⇨ For word integration, matching will vary depending on the research focus.
• What eye-tracking measures should be considered?	⇨ For individual words, typical measures are: first fixation duration, single fixation duration and gaze duration. ⇨ For multi-word sequences, establish ROIs that encompass the individual words and the whole sequence. Different measures will be appropriate depending on the size of the ROI. ⇨ For word integration, effects could show up in a range of measures, so output data for a variety of them.
• What do you need to do to 'match' participants?	⇨ This varies depending on the target population. Look at the tasks other researchers have used for guidance. For an example of matching non-native speakers see Section 4.3.2 and those with an autism spectrum disorder Section 4.6.2.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press

(cont.)

Final checklist	
• What paradigm is most appropriate for your research question, and what do you need to consider for each design?	⇨ There are a range of techniques to consider, each having a set of methodological concerns. For a discussion of a few of them, see the following: • authentic texts, see Section 4.3 • creating texts, see Section 4.4 • boundary paradigm, see Section 4.5 • matched, counterbalanced items, see Section 4.6.

https://doi.org/10.1017/9781108233279.005 Published online by Cambridge University Press