Chapter 5
Eye-Tracking and the Visual World Paradigm

Sanne M. Berends, Susanne M. Brouwer and Simone A. Sprenger

Abstract This chapter will focus on the use of eye-tracking in the visual world paradigm. This method can be employed to investigate a number of language comprehension issues, and we will begin with a brief overview of the history of the method and some of the applications. More centrally, we will discuss how it can be used to assess the impact of cross-linguistic interference, proficiency levels, and age of onset in L2 acquisition and L1 attrition, with an introduction to the issues that are involved in designing a study using this technique. As a case in point, we present and discuss the specific experiment employed within the multi-task, multi-language and multi-lab study on which this book is based, with special attention to the issues for analysis that arise when data from multiple systems must be combined.

Keywords Bilingualism · Grammatical gender · Eye-tracking · Second language acquisition · First language attrition

5.1 Eye-Movements and Cognition

The world is filled with visual stimuli which are constantly competing for our limited attentional resources. During visual exploration, we need to select and attend to those things that contain relevant information and ignore others. What we look at reveals a great deal about what is going on in our minds. Eye-tracking technology exploits this close temporal link between gaze and cognition to study the fast and highly automatic processes involved in language processing.

In 1974, Cooper was the first to track the eye movements of participants as they listened to short narratives while looking at a display of objects. He discovered that participants' eye gaze was drawn to objects mentioned in the narratives.

Electronic supplementary material The online version of this article (doi:10.1007/978-3-319-11529-0_5) contains supplementary material, which is available to authorized users.

M.S. Schmid et al., *Designing Research on Bilingual Development*,
SpringerBriefs in Linguistics, DOI 10.1007/978-3-319-11529-0_5

Participants were, for example, more likely to look at a picture of a lion when hearing the phrase '…when suddenly I noticed a hungry lion' than at a picture of a camera. Fixations were often initiated before the spoken word was even completed, indicating that visual and language processing are closely time-locked.

5.1.1 Gaze and Language Processing

The last decade has brought the technology to measure eye movements within reach for many research labs. Contemporary eye-trackers provide us with high-resolution quantitative evidence of a listener's visual and attentional processes (Duchowski 2002). However, while the field of psycholinguistics has seen a surge in the application of eye-tracking techniques to the study of reading since the 1980s (see, e.g., Clifton et al. 2007 for an overview) it took more than two decades until Tanenhaus et al. (1995) introduced its use in the field of auditory language processing. Since then, eye-trackers have frequently been used to study interactions between vision, attention, and the processing of spoken language.

The experimental paradigm introduced by Cooper (1974) and Tanenhaus and colleagues (1995) is known as the *visual world paradigm* (VWP, for an extensive review, see Huettig et al. 2011). Participants listen to a spoken utterance and simultaneously look at a visual scene containing various objects while their eye movements are monitored. The spoken utterance is usually related to one or more objects in the scene and the question is whether, and when, people look at these objects. When the time to launch a saccade (an eye-movement to another location) is taken into account, the point in time at which the listener's gaze is directed towards an object that has been named provides an excellent estimate of the time at which the word has been recognized (Allopenna et al. 1998). Manipulating the relationship between the objects and the linguistic input (e.g., making them harder to distinguish due to similar speech sounds) allows researchers to test theories about the way in which listeners access information in their mental lexicons.

This information can be deduced from the listeners' gaze patterns across time. In the VWP, the type of display can range from semi-realistic scenes (see Fig. 5.1 for an example) to arrays of objects (see Fig. 5.2).

Some objects are mentioned in the spoken utterance and are the targets, while others that overlap with the target to some degree function as competitors. Objects that are completely unrelated serve as distractors. The proportion of fixations on an object, time-locked to the auditory presentation of the target word, is taken to be an indication that (partial) lexical access has been achieved. For example, in a study by Allopenna et al. (1998), participants were instructed to listen to sentences such as 'Pick up the beaker; now put it below the diamond'. The names of some of the objects in the visual display were phonologically similar to the name of the target object. For example, the target object *beaker* was displayed with a competitor that phonologically overlapped at onset position (*beetle),* with a competitor that phonologically overlapped at rhyme position (*speaker*) and with a phonologically

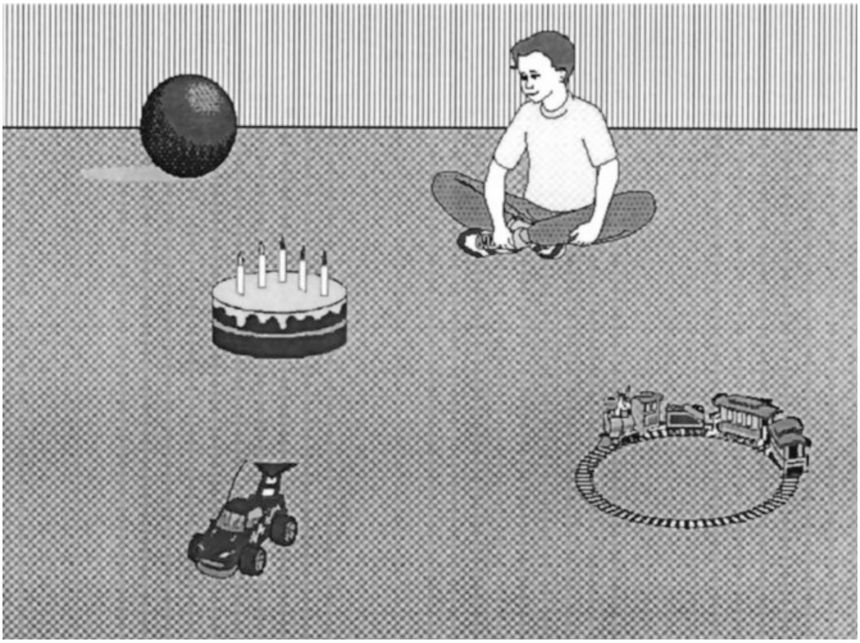

Fig. 5.1 A semi-realistic scene from an eye-tracking study (Altmann and Kamide 1999, their Fig. 1, p. 250, reprinted with kind permission from Elsevier)

unrelated distractor (*carriage)*. Results demonstrated that listeners fixated more on objects that overlapped with the target signal phonologically (*beetle, speaker*) than non-overlapping ones (*carriage*). Critically to their research question, listeners looked more often at onset competitors (*beetle*) than rhyme competitors (*speaker*), although both were fixated often enough to suggest that they were partially activated. These findings show that information at onset position is more influential in constraining lexical selection than information at final position.

The original VWP has been modulated in various ways to accommodate the demands of specific research questions. For example, McQueen and Viebahn (2007) developed a version of the paradigm in which the objects in the visual display are replaced by printed words. The benefit of this variant is that the critical stimuli do not need to be imageable, which makes it easier to design controlled sets of materials. Other changes involve the visual array, which originally contained real world objects which participants were instructed to manipulate (e.g., 'Pick up the *beaker*[1]; now put it below the diamond'), but more recently has involved objects or scenes presented on the screen with the simple instruction to look at the screen while listening to a description (e.g., 'The boy will eat the *cake'*), in order to examine effects of sentence context.

[1] Targets will be presented in italics throughout.

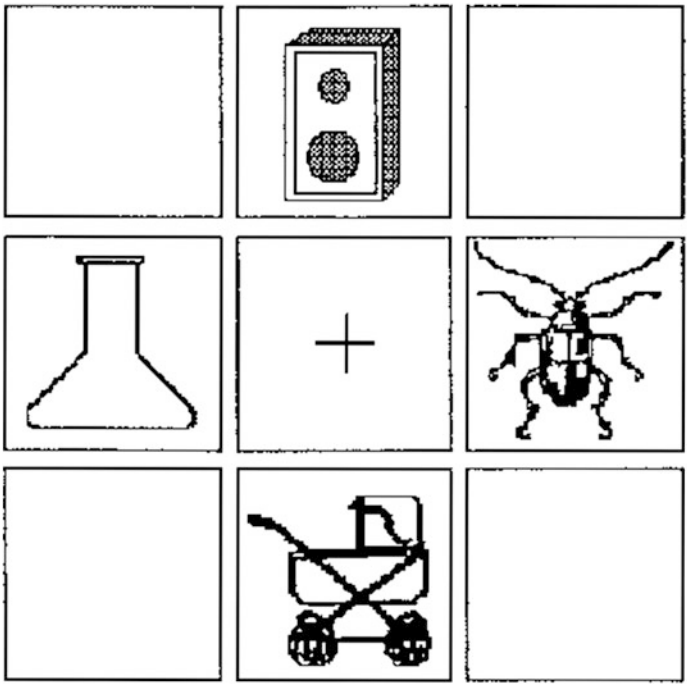

Fig. 5.2 An array of objects from an eye-tracking study (Allopenna et al. 1998, their Fig. 3, p. 427 (detail), reprinted with kind permission from Elsevier)

Until recently, the VWP was mainly used in order to answer research questions concerning monolingual language comprehension (for extensive overviews of the investigated topics in monolingual research, see Altmann 2011; Huettig et al. 2011; Tanenhaus and Trueswell 2006). Comparatively few VWP studies have been dedicated to ***L2 acquisition*** research. One of the crucial questions in this area of research is how lexical access occurs in bilinguals' native and second languages and to what extent the two languages interact during online processing. To the best of our knowledge, Spivey and Marian (1999) were the first to apply the VWP to L2 processing. In their study, Russian-English bilinguals were instructed to listen to Russian sentences (e.g., 'Poloji *marku* nije krestika'—'Put the *stamp* below the cross'). On the screen a stamp, a marker, and two unrelated distractor images were displayed. In the participants' L2, English, the object label *marker* is phonologically similar to the L1 target word *marka*, stamp. An analysis of the participants' eye movements indicated a higher proportion of eye-movements towards the between-language competitor *marker* than towards the unrelated distractors, indicating that, even if they are listening to their L1, bilinguals simultaneously activate lexical representations for words in both their L1 and their L2. The VWP can thus successfully be applied to research questions concerning L2 processing.

5.1.2 Advantages and Challenges of the Method

A first advantage of the VWP is that it is an *online* method, tapping into spoken language comprehension as it occurs and revealing aspects of processing that listeners are often completely unaware of. Second, in contrast to other (offline) methods, the listeners are not asked to perform meta-linguistic judgments that might lead to an over- or underestimation of their language abilities or encourage strategies making use of explicit knowledge. Third, the data are with a high temporal resolution (on the time-scale of milliseconds), providing precision in determining when responses to spoken language begin to differ. Fourth, the topics of interest can be investigated under natural, relatively realistic conditions in which listeners hear words, sentences, or stories that are pragmatically relevant. A last advantage is that this technique can be used for participants of most ages, since it does not require participants to read or carry out complex tasks; it has been used very successfully with preschool children in a number of studies (e.g., Lew-Williams and Fernald 2007).

A challenge lies in the fact that, as eye movements reflect the interplay of language, vision and attention, using them to study linguistic processing per se requires careful experimental design to control for possible confounds with visual processing and attention allocation. For example some colors, such as red, attract more attention than others and some positions are more frequently fixated than others, due to differences in visual processing and attention. In addition, the spoken utterances always need to be related to the visual stimuli on the screen, thereby limiting the stimulus material to either concrete, pictureable objects or relatively short written words.

5.1.3 Eye-Tracking and Grammatical Processing

The VWP has been used to investigate language comprehension on various levels of processing, ranging from discourse and sentence level to lexical processing (e.g., phonological processing and bilingual word recognition) as well as their interaction. For example, Dahan et al. (2000) used the VWP to investigate the influence of the semantic and syntactic context on spoken word recognition. Specifically, they asked whether prenominal gender cues influence the speed of spoken word recognition. In their study, French native speakers were presented with displays consisting of objects with names that shared the same phonological onset but differed in grammatical gender (e.g., *bouteille*$_{FEM,}$ bottle; *bouton*$_{MASC,}$ button). When nouns were preceded by a gender-ambiguous plural article (e.g., ‘cliquez sur les$_{\mathrm{AMB}}$ *boutons'*, ‘click on the buttons’) listeners were equally likely to fixate the target and the phonological competitor (e.g., *bouteilles*, bottles) directly following the presentation of the first few phonemes. However, when a gender-marked article preceded the noun (e.g., ‘cliquez sur le$_{\mathrm{MASC}}$ *bouton'*, ‘click on the button’), listeners more rapidly fixated the correct target and the gender-marked article eliminated the interference of the phonological competitor. These findings demonstrate that gender cues earlier in the sentence can minimize the set of possible candidates for a target noun and therefore facilitate language processing.

The same effect has since been found for article-noun combinations whose onsets did not overlap (Paris et al. 2006 for German; Lew-Williams and Fernald 2007 for Spanish, Loerts et al. 2013 for Dutch and Hopp 2013 for German). Lew-Williams and Fernald (2007) showed that the gender anticipation effect can even be found in preschoolers.

A question for L2 acquisition is whether L2 learners are, in principle, able to exploit gender marking in a similar way. This is an important question in the context of the debate on ultimate attainment among L2 learners. Of particular interest is the question of which conditions lead to nativelike processing and what role AoA and the presence and/or similarity of the gender system in the L1 play for the facilitation effect in L2. Thus far, results have been mixed, even when the same L2 is examined. Lew-Williams and Fernald (2010) and Grüter et al. (2012) did not find gender anticipation or facilitation effects for intermediate and advanced English learners of Spanish, despite the fact that Spanish has quite a transparent gender system, suggesting that a non-gender L1 precludes the acquisition of native-like L2 gender processing routines. In contrast, Dussias et al. (2013b) did find gender facilitation effects in highly proficient English speakers of Spanish. In fact, their results suggest that the presence of a gender system in the L1 may interfere with gender facilitation in the L2: A population of Italian L2 learners of Spanish that was also tested only exploited the gender cue on feminine articles and not on masculine articles. This may be explained by the fact that a greater percentage of the masculine items had opaque gender, whereas in general gender in Spanish is transparent. Another factor might have been the difference between the definite article systems of the two languages. Whereas Italian has two masculine definite articles (*il* and *lo*) and one feminine definite article (*la*), Spanish only has two definite articles (*el* for masculine and *la* for feminine).

Hopp (2013) investigated native English learners of German, which has a non-transparent gender system (see Chap. 1), and found anticipatory effects, suggesting that native-like performance on gender can be found without a gender system being present in the L1 and in the absence of phonological or form regularities of gender agreement, like those characteristic of Spanish. Loerts et al. (2013) investigated gender processing in Slavic learners of Dutch, a language which is also non-transparent. Their findings, like those of Dussias et al. (2013b), suggest that a different expression of gender in the L1 might modulate the degree to which gender marking can be used in L2 processing, since Polish learners of Dutch, who encode gender in the L1 but not on articles, showed no effect of gender facilitation.

To conclude, the factors that govern the extent to which L2 learners can acquire gender anticipation and facilitation effects during spoken language comprehension are not yet clear. While the studies that have been conducted so far suggest that it is indeed possible for L2 learners to make use of gender agreement for these purposes under certain circumstances, it is as yet unclear in which way the interaction of a range of predictors can affect the outcome. These predictors include characteristics of both the first and the second language, the level of proficiency in the second language, the AoA of the L2 learners, and potentially others.

In this context, it is also interesting whether such anticipation and facilitation effects are stable in a native language under conditions of language attrition. To our knowledge, no previous studies exist which address gender anticipation and facilitation in L1 attriters. The findings reviewed above suggest that there may be complex interactions between the bilingual's languages that determine whether cues are used to facilitate the access of upcoming information. In order to gain further insight into how gender processing is affected by bilingual development, and to elucidate the impact of the predictors named above, it is important not to confine the investigation to the later-learned L2. By comparing L2 learners with L1 attriters (individuals who have become highly proficient or dominant in the L2 after an extended period of immersion), the impact of some of these predictors—for example, proficiency and dominance—can be disentangled from others—for example AoA and the order of acquisition (see Chap. 1 for a more complete discussion). We propose that deeper insight into multilingual grammatical processing can only be gained on the basis of comparisons among participants that vary along all of these dimensions. Such investigations should thus include native controls, L2 speakers and L1 attriters from different linguistic backgrounds across a range of proficiencies and AoAs. In the following sections we will discuss the considerations that went into constructing a study on gender processing designed to achieve this goal. However, most of the considerations are sufficiently general to apply to other studies using this paradigm.

5.2 General Design Issues

In this section we present a number of considerations which should be carefully taken into account when designing a VWP experiment and discuss some of the strategies that have been used to deal with each of them.

5.2.1 Fixating Visual Objects: Important Potential Confounding Factors

The types of visual displays that have been used in VWP experiments vary depending on the research question. In general, two different types of displays can be distinguished. The first type consists of arrays of line drawings (black and white or colored), or of pictures of real objects. The second display type is made up of semi-realistic scenes, which consist either of drawings of pictures presented on a computer screen or of real objects laid out on a workspace (see Figs. 5.1 and 5.2). The main difference between the two paradigms is that semi-realistic scenes give a more natural context in which the impact of world knowledge can be tested

(Henderson and Ferreira 2004), whereas in the paradigm with arrays of objects the influence of world contextual knowledge is reduced to a minimum, providing the opportunity to isolate the activation of conceptual and lexical information associated with individual words stored in the mental lexicon (Huettig et al. 2011).

Picture selection is a crucial step in developing a VWP experiment. The objects to be used in a visual display can be selected from picture databases such as the one created by Snodgrass and Vanderwart (1980). This database comprises 260 black and white line drawings which have been normed for name agreement, image agreement, familiarity and visual complexity. Since these factors might potentially influence the eye gaze, they should be taken into account when selecting the visual stimuli, making pre-normed stimuli an excellent choice (note, however, that factors such as familiarity may vary depending on the culture—an important consideration for investigations of multilingual development!).

Pictures of real objects are more detailed than black and white drawings, and can therefore enhance recognition and facilitate naming agreement. However, visual complexity differences between pictures of real objects are more difficult to control: Some pictures will be more easily recognized than others (Dussias et al. 2013a), which encourages earlier fixations when the objects are named. As this visual complexity bias might obscure any potential anticipation effect, we suggest using line drawings from the Snodgrass and Vandewart (1980) picture database in order to keep the variance in the complexity of the pictures as low as possible. As those pictures have only been normed for English naming agreement, studies investigating other languages should pilot the pictures to be used on native speakers of the languages represented in the experimental population, in order to ensure that each picture stimulus will indeed elicit the intended nouns.

The Snodgrass and Vandewart pictures are ***copyright-protected***, and in order to use them in any study or for publications it is necessary to obtain a license. Other databases of images are available (for a list of suggestions see http://www.cogsci.nl/stimulus-sets). Whatever images are selected for any given study, it is of vital importance that the researcher should consider and explore any potential copyright issues, since infringement of such rights can have serious consequences (and also make it difficult if not impossible to publish the results from the study).

Equally important is the consideration of how the ***spoken utterances*** that participants will hear may constrain the visual characteristics of the objects being employed. For example, in a study which investigates the use of gender in anticipation, it is important to reduce effects of sentence context that might also lead to anticipation because the target noun is semantically the most likely object to fit a given sentence frame. Many VWP paradigm studies therefore opt for minimally constraining contexts and present only noun phrases consisting solely of the determiner and the noun (e.g., *de appel*, 'the apple'). However, this practice is problematic for two reasons. First, since launching an eye-movement takes ca. 200 ms (Matin et al. 1993), and since determiners are considerably shorter than that in many languages, such phrases may not allow participants enough time to translate their anticipation of the upcoming noun into an actual saccade. Second, it has been proposed that L2 learners acquire frequent combinations such as determiner plus noun as chunks, and that they therefore might show a preference for the target, but for reasons that are unrelated to gender as a structural property of the noun. It is therefore preferable to extend the noun phrase by an intervening adjective (‘the ADJ apple’), where the structure of the language allows this.

This raises another problematic question, namely what type of adjective should be used: Evaluative words, such as ‘pretty’, ‘nice’ etc. may confuse participants who do not agree with the assessment—or where the description may better apply to other objects. Depicting adjectives that refer to more objective properties of the target (‘large’, ‘heavy’ etc., see Paris et al. 2006) may enhance or reduce the noticeability of the target (which then also has to be larger or heavier than the other items in the array) and thus confound the anticipation effect.

An alternative solution is suggested by Loerts et al. (2013), who investigated whether Dutch native speakers use gender marking to predict the correct referent. The visual display used in this study contained a target, a competitor and two distractors, all of them represented as colored line drawings. The competitor was either the same or a different gender and/or color as the target (while the distractors were always represented in different colors and had different genders), and the intervening adjective named the color. For example, one such array depicted a red apple (common, target), a red cake (common, competitor), a yellow lock (neuter, distractor) and a blue book (neuter, distractor). In this case, when the participant heard the phrase ‘click on the$_{\text{COM}}$ red…’, both target and competitor were equally likely to be fixated until the onset of the noun (apple vs. cake), since both were potential referents of the noun phrase. In an array where the two red objects did not share their gender and the competitor was, for example, a red book, participants were able to differentiate them at an earlier stage, despite the fact that both were represented in the same color.

Results showed that the color of the pictures interfered with the gender anticipation effect to some extent: Participants’ fixations to targets were initiated later when the target was brown as compared to other colors (red, yellow, blue, green). This finding indicates that some colors are more salient than others and therefore attract more attention. Loerts et al.’s (2013) findings revealed that the effect of the color manipulation was stronger than the gender facilitation effect in those visual displays in which the target and competitor shared color.

While inserting color adjectives between determiner and noun is therefore probably the best way of constructing noun phrases that allow the participant enough time to use gender agreement information encoded in the determiner, the color interference may override the gender facilitation effect. This, however, can be eliminated by presenting all objects within the same VW display in the same color (while making sure that color is counterbalanced across conditions).

Another aspect of the design that needs attention concerns the number of objects and the layout of the visual displays. Both of these factors can also influence the time it takes to fixate the correct object. A first factor to consider is the number of object positions on the visual display (Ferreira et al. 2013, see Figs. 5.3 and 5.4). Ferreira and colleagues studied the influence of the complexity of the visual display on the interpretation of garden-path sentences like ‘Put the book on the *chair* in the

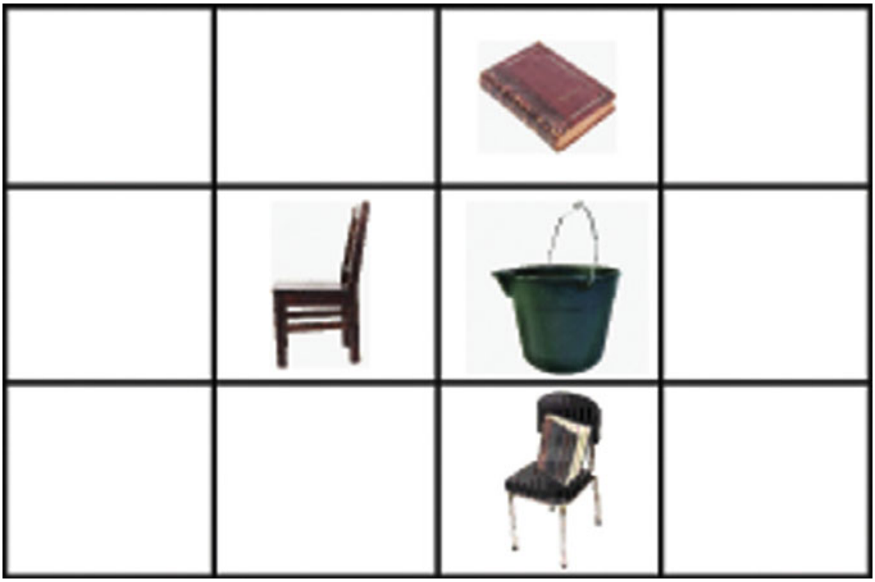

Fig. 5.3 The visual display used by Ferreira et al. (2013), their Fig. 1 (reprinted with kind permission from Elsevier)

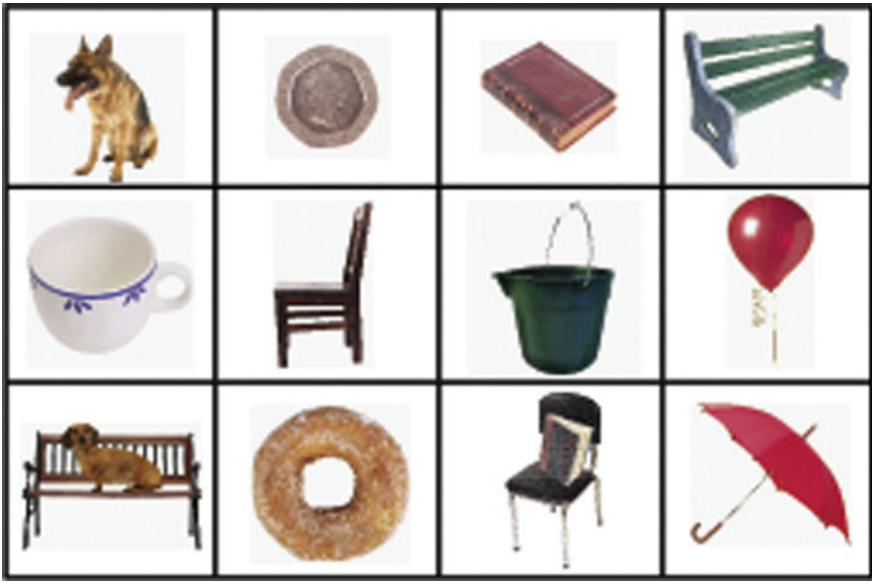

Fig. 5.4 The visual display used by Ferreira et al. (2013), their Fig. 4 (reprinted with kind permission from Elsevier)

bucket.' Here, the prepositional phrase *on the chair* is temporarily ambiguous because it can either be interpreted as being the goal (the location where the book is to be placed) or the modifier (the location from which it is to be removed, in order to be placed in the bucket). In the unambiguous equivalent 'Put the book that's on the *chair* in the bucket', the same prepositional phrase *on the chair* can only be interpreted as the modifier. In a visual context with a book on a chair, a single book, an empty chair and an empty bucket participants were less likely to fixate the empty chair even in an unambiguous sentence. This can be interpreted as suggesting that when two books are present in the display, the listener is more likely to assume that the chair serves as a location, identifying the appropriate book, i.e. that the modifier interpretation of the prepositional phrase is favored.

However, when the single book is replaced by an unrelated object, garden path sentences typically elicit more gazes towards the incorrect goal, that is, the empty chair (Tanenhaus et al. 1995) as compared to the unambiguous counterpart. Ferreira and colleagues also used instructions which did or did not contain garden-paths. In addition, the complexity of the visual displays was manipulated by presenting participants with 4 (Fig. 5.3) or 12 (Fig. 5.4) objects. The findings indicated that it is more difficult to construct an online interpretation when the visual display is more complex (i.e. contains more objects). Looks to the target (i.e., the book on the chair) increased much later, which suggests that the visual search takes longer in complex visual contexts. Furthermore, the classical garden-path effect disappeared due to the delay. This suggests that it is undesirable to use an unnecessarily complex display. In order to ensure comparable results between eye tracking studies with similar research questions, it is therefore advisable to use the same number of objects in the visual display as other studies in the field.

Participants also have clear preferences for particular screen positions, depending on the reading direction in their native language. Readers whose script runs from left-to-right and from top-to-bottom will have the tendency to first direct their gaze to the upper left corner of a visual display. To control for this bias, the positions of the objects on the visual display should always be counterbalanced so that each condition (target, competitor, distractors) is presented equally often in each position and that each individual object is displayed in all positions and all conditions. To avoid repetitions of the same array for the same participant, the various uses of each object can be distributed across different lists (so that, for example, in one list, *apple* is used as target, in another as competitor, and so on); at the level of the entire experiment, the effects of position should then be balanced.

5.2.2 *Presenting Auditory Stimuli: Important Potential Confounding Factors*

In addition to the effects of the visual display, eye movements can also be influenced by inadvertent properties of the ***auditory stimuli***. When selecting the critical words for an experiment on gender processing, several important factors must be controlled in order to avoid confounds in the experiment. One of these factors is ***phonological overlap*** between the target object and the other objects presented on the same visual display. Recall the experiment by Allopenna et al. (1998) described earlier, in which participants were found to direct their eye gaze more towards the phonological onset competitor than to the rhyme and unrelated competitor. However, participants also fixated the rhyme competitor more than the unrelated competitor. Studies investigating anticipatory gazes should therefore avoid phonological overlap at either onset or rhyme as much as possible.

Furthermore, ***word frequency*** can also modulate lexical access speed and therefore impact on fixation time. For example, presenting a comparatively low-frequency auditory target like *bench* alongside a high-frequency phonological competitor (e.g., *bed*), a low-frequency phonological competitor (e.g., *bell*) and an unrelated distractor elicits more fixations to the high- than to the low-frequency competitor (Dahan et al. 2001). Importantly, this effect is not limited to cases with phonological competition. Dahan and colleagues also found frequency effects in eye movements toward targets without phonologically related competitors. Eye gaze latencies towards targets with high frequency names (e.g., *horse*) were faster than for targets with low frequency names (e.g., *horn*). It is therefore extremely important that the lexical frequency of all items in an array be stringently controlled. For studies of L2 acquisition, it may furthermore be advisable to select only items of comparatively high frequency, since this will reduce the chance of participants being unfamiliar with certain words or their gender.[2]

When carrying out an L2 acquisition study, cross-linguistic ***gender overlap*** of the target noun can also potentially affect the results. Weber and Paris (2004) presented French learners of L2 German with German instructions to click on a target object, and the target noun was preceded by a gender-marked article. Participants saw visual displays with a target and a competitor, as well as two distractors. The target and the competitor always had phonologically overlapping onsets not only in the language of the experiment (German, e.g. *Perle* 'pearl' vs. *Perücke* 'wig') but also in their French translation equivalents (*perle, perruque*). In addition, whereas the gender of the target was always shared between the German noun and the French equivalent (e.g., *die*$_{FEM}$ *Perle*, *la*$_{FEM}$ *perle*), the competitor's gender was manipulated in such a way that it either had the same (e.g., *die*$_{FEM}$ *Perücke*, *la*$_{FEM}$ *perruque*) or a different gender (e.g., *die*$_{FEM}$ *Kanone*, *le*$_{MASC}$ *canon*) across both languages. In the latter condition, where the competitor had the same gender as the target in the language used (German), earlier fixations on the target could only be ascribed to the influence of the participants' first language, French. Such a control condition is thus necessary to ensure that anticipatory effects are, indeed, based on L2 gender only. The influence of crosslinguistic competition was confirmed by the finding that Weber and Paris' French learners of L2 German only

[2] Irrespective of the frequency of the chosen items, it is highly advisable to ensure that all participants know all of the words, see Sect. 5.3.3.

showed an effect of competition when the gender of the competitor matched that of the target in *both* languages. The result suggests that participants were unable to eliminate the L1 gender while listening to instructions in the L2. Neglecting to include this manipulation in the design would therefore have produced misleading results.

5.2.3 Controlling Timing

One of the advantages of the VWP is that it allows tracking eye movements over time, to see how and when the auditory stimuli direct attention toward elements of the visual display. The fine-grained temporal resolution of eye-movements makes ***accurate timing of the recordings*** of these stimuli essential across multiple trials. This is a more challenging task for experiments using spoken language than for designs that rely on written language (see also Chap. 6). We recommend a procedure in which all stimuli (which for the purpose of the gender anticipation task take the form of sentences such as 'click on DET ADJ NOUN' or 'where is DET ADJ NOUN') have the same timing across relevant regions. To achieve this goal, we used the following procedure: Each sentence was recorded three to five times by a female native speaker who spoke a standard version of the target language and had considerable elocution training. From these recordings, the best exemplar for each stimulus sentence was selected, based on normal speaking rate and naturalistic prosody (stress, rhythm and intonation).

For the purposes of the gender experiment, there are three regions preceding the noun in the stimulus sentence: verb plus preposition (preamble), determiner (moment at which relevant information is presented), and adjective (intervening region). We established the average duration of all of these regions and subsequently adjusted them to this average length in every one of the recordings, so that the onset of the noun always occurred at exactly the same moment in each sentence. The purpose of this adjustment was to avoid a larger facilitation effect in those sentences that were pronounced at a slower rate and thus reduce between-item jitter in timing as much as possible.

A second important issue to keep in mind in relation to timing is when the auditory and the visual stimuli should be presented relative to each other. In the VWP, the presentation of the visual display often starts at or shortly before the onset of the utterance. Previous work has shown that the amount of ***preview time*** can affect the likelihood of fixations to particular objects (Huettig and McQueen 2007; Ferreira et al. 2013). Ferreira and colleagues manipulated the amount of preview time in their study on the interpretation of garden path sentences discussed above. They found the classical garden-path effect when the visual display contained four objects and was presented three seconds before the instructions initiated. However, the effect disappeared when the visual and auditory information were presented simultaneously (i.e., preview time was reduced to zero). The authors suggest that preview time may allow participants to build better expectations of what may be referred to in the upcoming utterance. The primary goal of the preview time should be to allow the participants to construct a spatial representation of the VW array, as longer times may lead to strategies and expectations that influence the results. The ability to predict upcoming information might also increase over the course of an experiment as experience in the task grows. We therefore recommend as short a preview time as is consistent with the participants being able to carry out the task. In our case we chose to give no preview time, the most extreme option, as both our visual display and our instructions were comparatively simple.

5.2.4 Summary of General Considerations

In sum, we recommend a research design that takes into account the following factors:

- properties of the lexical items: word frequency (based on large speech/written corpora), phonological overlap (at onset or at rhyme) between targets, competitors and unrelated items, controlled gender overlap between items in the target language and in the L1 of L2 speakers where applicable (in our case Polish and Russian)
- properties of the picture stimuli: reliability of naming, effects of color, size and complexity
- properties of the auditory stimuli: equal duration of the relevant regions preceding the noun (e.g. imperative/preposition, determiner, adjective) for all recorded stimuli, achieved through adjusting the length of the recorded segments
- properties of the visual world scene and presentation: locations of the target and competitor objects counterbalanced across the regions of the scene which are employed, a preview time that is consistent with the participants being able to carry out the task.

These design principles are quite general and apply to both the Dutch and the German version of the experiment we present below. However, the actual implementation of the design differed slightly with respect to the target nouns, carrier sentences and counterbalancing results due to differences in gender and case systems between Dutch and German, a problem that will occur in any multi-language study. In the following, we therefore provide descriptions of how the general design was adjusted for the Dutch and German stimuli, while maintaining enough similarity for comparison across languages.

5.3 The Present Experiment

5.3.1 Rationale of the Experiment

In our study, we were interested in the effects of the characteristics of both a bilinguals' languages on gender processing. In addition we were interested in how the salience of the gender system affects L2 acquisition and L1 attrition: While both Dutch and German have opaque gender systems, which are generally not predictable on the basis of morphological or phonological characteristics of the word, gender is more salient in German due to the interaction with case marking (see Chap. 1 for a full discussion). We therefore compared adult learners of Dutch and German with various types of L1, L1 attriters and predominantly monolingual speakers of these languages. Our goal was to assess the effects of language dominance, order of acquisition, proficiency and AoA on the use of gender agreement to facilitate information retrieval during online processing. In order to measure the listeners' response to grammatical gender information in Dutch and German noun phrases, we manipulated gender overlap between the objects displayed, as in the studies cited above.

In each trial participants saw four line-drawings on a screen and heard a sentence directing their attention to one of these objects (e.g., 'Klik op het$_{\text{NEUT}}$ groene blad$_{\text{NEUT}}$', 'Click on the green leaf'). Crucially, in this sentence the gender-marked determiner must agree with the gender of the noun; if listeners are able to anticipate the correct noun based on the gender information provided by the determiner, this demonstrates a quick and automatic use of gender information in comprehension. Two types of display were used, one where there was no gender competitor (see Fig. 5.5, Panel A for an example from the Dutch experiment) and one that contained two potential target objects, the actual target and a gender competitor (see Fig. 5.5, Panel B). The difference between the two types of displays represents the within-subject factor ***gender competition***.

If listeners make use of the gender information encoded in the definite determiner to direct their gaze to the target as quickly as possible, they should settle on the *leaf* sooner in panel A than in panel B, since there is no competition from a gender-congruent competitor (in panel B, *eye* represents such a competitor). Such a difference in the visual selection of the target object can be taken as evidence for active use of gender information in comprehension. In order to get reliable estimates of the time it takes to select the visual targets, the effects have to be independent from the specific materials used. Therefore, the visual and the auditory stimuli, as well as the visual displays, had to be selected and created with care.

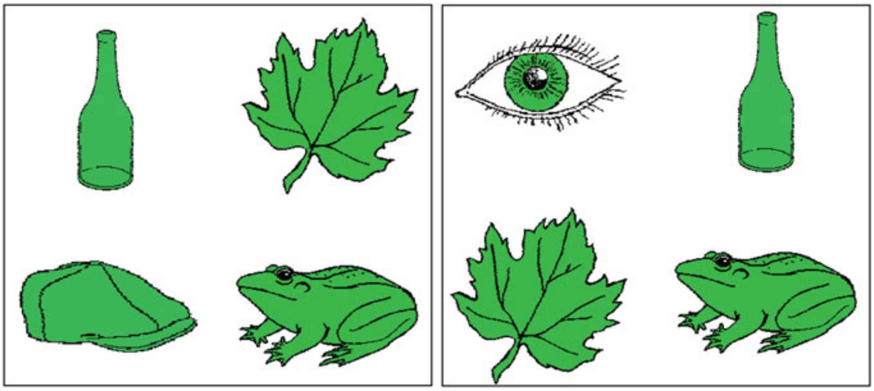

Fig. 5.5 Example of two visual stimulus arrays in the gender competition task for the spoken Dutch stimulus ‘Klik op het$_{\text{neu}}$ groene *blad*$_{\text{neu}}$’ (Click on the green leaf). In panel A (*left*) the objects other than the target provide no gender competition as they are all common gender, whereas in panel B (*right*) one of the objects (i.e., het$_{\text{neu}}$ oog$_{\text{neu}}$, ‘the eye’) is a gender competitor for the target

5.3.2 Materials

48 highly frequent Dutch nouns and 48 highly frequent German nouns referring to pictureable objects were selected as targets. While the Dutch nouns were evenly spread over the two genders encoded in the language (e.g., common: *de emmer* ‘bucket’, 24 items; neuter: *het potlood* ‘pencil’, 24 items), the German nouns were limited to two of the three genders (e.g., masculine: *der Hut* ‘hat’, 24 items and neuter: *das Kleid* ‘dress’, 24 items). No feminine German nouns were included, as the feminine singular definite article *die* is ambiguous with the plural definite article for all genders. The Dutch nouns were selected based on the Spoken Dutch Corpus’ estimates of frequency of occurrence (Oostdijk 2000) and the German nouns were selected from a basic vocabulary list for German learners (Oehler and Gerretsen 1973). This was the first within-subject factor in the experiment: ***grammatical gender***.

Two noun phrase constructions were created: (1) definite determiner-adjective-noun (e.g., *het gele hek* ‘the yellow fence’; *der blaue Hut* ‘the blue hat’), and (2) indefinite determiner-adjective-noun combinations (e.g., *een geel hek* ‘a yellow fence’; *ein gelbes Kleid* ‘a yellow dress’). This was the within-subject factor ***gender structure***. In the definite noun phrase condition, gender information is marked on the definite determiner in both languages (e.g., *de*$_{com}$ vs. *het*$_{neu}$; *der*$_{mas}$ vs. *das*$_{neu)}$. In the indefinite noun phrase condition gender is carried by the adjective (e.g., *gele*$_{com}$ vs. *geel*$_{\text{neu}}$; *blauer*$_{masc}$ vs. *blaues*$_{neu}$). The adjectives used were *yellow*, *blue*, *green*, and *brown*. The Dutch noun phrase combinations always followed the sentence frame ‘Klik op…’ (Click on…). However, it seemed best to present the German nouns in the nominative case form, since this is considered to be the default form and is the more frequent one, to make the experimental sentences more comparable cross-linguistically. Therefore, the German noun phrases were embedded in the Wh-question ‘Wo ist…’ (Where is…), since the German equivalent of ‘Click on’ would require the accusative form. The distribution of gender across the objects on the experimental trials was either 1:3 (i.e., the gender of the target versus the gender of the unrelated competitor and both distractors) or 2:2 (i.e., the gender of the target and the related competitor versus the gender of the distractors).

The Dutch sentences were recorded by a female native speaker of Dutch living in the Netherlands and the German sentences were recorded by a female native speaker of German living in Germany. Both speakers had professional experience as radio presenter or voice over. Recordings (16 bits, 44.1 kHz) were made in a sound-attenuating double-walled booth. The duration of the preambles, the determiner and the adjective were measured and averaged across all sentences in PRAAT (Boersma and Weenink 2012). All auditory stimuli were then manipulated in Adobe Audition© 3.0 to adjust the duration of the three regions from the onset of the gender cue to the following values: determiner (Dutch: 858 ms; German: 445 ms), adjective (Dutch: 974 ms; German: 660 ms) and noun (Dutch: 1351 ms; German: 1060 ms). The stimuli were also equalized to the same rms level of 65 dB and modified to fade in and out by means of PRAAT.

In addition, for each language 12 pictures of naturally red-colored objects were selected and used to construct 24 filler displays. In contrast to the experimental displays, the filler displays contained a target, two objects with the same grammatical gender and only one of a different gender. They were included to prevent participants from strategically looking for the odd one out, since the experimental target item was always different in gender from at least two and sometimes three of the other objects. In the filler displays the target had two same gender competitors and the distractor was the odd one out. Finally, six Dutch and five German practice displays were constructed in different colors.

The words used in the two language experiments were paired with 48 pictures selected from the Snodgrass and Vanderwart (1980) standardized picture set of black and white line drawings. Color was then added to the line drawings. The possible word-picture pairs were constrained, as they must plausibly appear in either yellow, blue, green or brown in the real world (in addition to the 12 naturally red fillers) and the pictures were colored accordingly (e.g., *sun* was colored yellow; *apple* was colored green). A naming pretest, in which participants were asked to name individually presented pictures on a computer screen, confirmed that the pictures were highly recognizable and elicited the correct target noun. Objects that did not consistently elicit the expected noun were revised or excluded. All pictures are included in the online supplementary material.

With these 48 pictures, 96 array combinations were created by positioning the objects in two-by-two grids (see Fig. 5.5); there were two versions of each array which varied as to competitor (gender matched or not); these were counterbalanced across lists to prevent repetition effects. The four objects included a target, a competitor and two distractors. All four objects always had the same color so that color did not add any information for identification of the target; nor would one of the objects be more noticeable due to its color. The objects in the display were positioned equidistant from the center of the screen. The competitor's gender was either congruent or incongruent with respect to the target object (e.g., Dutch *vliegtuig*$_{\text{neu}}$ 'airplane' vs. *lepel*$_{\text{com}}$ 'spoon' for the target *potlood*$_{\text{neu}}$ 'pencil'; German *Boot*$_{\text{neu}}$ 'boat' vs. *Löffel*$_{\text{masc}}$ 'spoon' for the target object *Haus*$_{\text{neu}}$ 'house'). The genders of the distractors were always different from that of the target. In order to exclude the possibility of phonological competition, the onset of the target item never overlapped with the onset of any of the other object names in the Dutch experiment. However, in the German experiment, we could not avoid phonological overlap completely (16 items in list 1 and 18 in list 2). The lists of array combinations of targets, gender congruent and incongruent competitors and distractors are included in the online supplementary material.

For each language two pseudo-randomized item lists were created in which targets were counterbalanced to appear either with a gender-congruent or a gender-incongruent competitor, and each participant was assigned to one list. On each list, the positions of target objects on the screen were counterbalanced. That is, the target appeared with equal probability in the four quadrants of the screen over the course of an experimental run. There was no repetition of colors on subsequent trials. Both lists were divided into two blocks, so that per block each picture appeared once as target, once as competitor and twice as distractor. One block contained the auditory stimuli in which the determiner was definite, and one had the indefinite constructions.

To sum up, each participant saw 96 target arrays and 24 filler arrays, in a factorial design varying

- ***grammatical gender*** (two levels: common vs. neuter for Dutch and masculine vs. neuter for German),
- ***structure*** (two levels: definite vs., indefinite), and
- ***gender competition*** (two levels: competitor vs. no competitor)

combining to 9 conditions, each with 12 exemplars per participant.

5.3.3 Procedure

Before beginning the experiment proper, participants were presented with a series of two pictures on a computer screen and simultaneously heard a bare noun corresponding to one of the pictures presented. Their task was to click on the corresponding picture on the screen. This task ensured that all participants were familiar with the names of all objects. Participants heard only the bare noun in order to avoid any priming of the article.

For the eye-tracking experiment, participants were seated in front of a computer screen while their eye movements were monitored by an eye-tracker. The presentation of the auditory and the visual stimuli was controlled with E-Prime (Schneider et al. 2002), which could be used at all labs where data was collected. Prior to the experiment a 9-point calibration procedure was performed; by directing the participant's gaze to specific regions of the screen, the eye-tracker is able to check whether the data being gathered corresponds to the actual location which the participant should be fixating to establish accuracy. Participants were then instructed to use the computer mouse to click on the object in the visual display representing the target item they heard in the sentence. They were asked to respond as fast and as accurately as possible. Each trial started with a central fixation cross, displayed for 500 ms in order to avoid baseline effects, followed by a visual display with four objects. The spoken sentence started simultaneously with the onset of the display. When participants clicked on an object, they initiated the next trial. Both reaction times and eye gaze data were recorded.

Each participant saw two stimulus blocks, with either definite determiners or indefinite determiners used in the auditory descriptions. The order of blocks was counterbalanced over participants. Prior to each block, participants performed a practice session of two or three trials to become acquainted with the (change in) task and procedure. Each block lasted approximately 10 min and participants were given a short break between the two blocks. The total duration of this testing session was approximately 30 min.

5.4 Data Recording and Analysis

5.4.1 Eye-Tracking Devices

As will usually be the case in studies aiming to collect data on different languages and from different populations, and therefore at different testing sites, the systems available for the experiment differed between locations and labs (see Chap. 3). For this experiment, data were obtained from SMI, SR Research and Tobii eye-trackers, each of which generates different formats of output, as will be discussed in more detail below. In addition, the physical set-up experienced by the participants differed from head-mounted through desktop-mounted to built-in eye-tracking systems. A full list of the variants and the populations which were tested on each set-up is given in Table 5.1.

Some technical differences notwithstanding, all of these systems are designed to answer the simple question "when does a participant look where?" That is, they provide us with x-y-coordinates of the participants' gaze, measured at a high temporal resolution (60–500 Hz, see Table 5.1), time-locked to the spoken stimuli. All systems analyze the reflection of infrared light from the eye in real time in order to determine the location of the pupil as well as the corneal reflection. The combination of these parameters allows the software to determine the participants' direction of gaze. Even though each set-up differed with respect to the position of the infrared emitter and the camera (and therefore with respect to its sensitivity for particular forms of movement artifacts), each of them has reliably been applied in

Table 5.1 Sampling rates of the various eye-trackers used in our project

Eye-tracker	Location	Sampling rate (Hz)	Set-up	Participant group
Eyelink II	Chicago	250	Head-mounted	Dutch attriters
	Toronto	250	Head-mounted	Dutch and German attriters
Eyelink 1000	Hamburg	250	Remote	German learners and natives
	Leiden	500	Remote	Dutch learners and natives
	London (ON)	500	Desktop-mount	Dutch attriters
SMI	Berlin	60	Remote	German learners and natives
Tobii T60	New York	60	Remote	German attriters
Tobii T120	Groningen	120	Remote	Dutch learners and natives

experimental settings in which timing information is essential. Given our within-subject design, and given our relatively large regions of interest, possible differences between eye-trackers were not expected to affect the results.

All eye-trackers were linked to and controlled by the same experiment software using somewhat different interfaces (E-Prime software, Psychology Software Tools, Pittsburgh, PA). For example, communication between E-prime and the Tobii eye-trackers was established by a set of software extensions that link the eye-tracker server with E-prime. The E-prime extensions for the Tobii package can be obtained from the Psychology Software Tools, Inc website (http://www.pstnet.com/downloads/eet/EETVersion6.pdf). The procedure built in a check for accuracy which helped ensure that the data from different set-ups was comparable; each recording session started with a calibration procedure that mapped the signal of the eye-tracker to the dimensions of the computer screen and the visual field of the participant.

5.4.2 *Dependent and Independent Measures*

For our project, blinks and saccades were discarded, as the focus was on fixations. Specifically, we were interested in the time course along which the listeners' gaze was directed at various objects in the visual scene. To operationalize this, we created four different visual regions of interest by splitting the screen into four quadrants, containing the four objects. Subsequently, we computed the average probability with which each object/region of interest was fixated within a given time bin (bin size of 50 ms), across a time span beginning 200 ms after the gender cue was heard. This starting time was chosen because of estimates that it takes approximately 200 ms to program and launch a saccadic eye movement (e.g., Matin et al. 1993).

In the analysis, objects were classified as targets, competitors and distractors. For the stimuli with gender marking on the determiner, we examined gaze proportions in the time window until noun onset to see if fixations reflected gender anticipation when no competitor was present.

5.4.3 Combining Data from Different Eye-Tracking Systems

Eye-trackers provide information about gaze location across time. Typically, this information is coded in terms of x- and y-coordinates, with x and y being the pixel dimensions of the experiment screen (where [0, 0] typically identifies the upper left corner). Time is logged based on the eye-tracker's internal clock time. The accuracy with which the gaze location is determined depends on the eye-tracker's sampling rate. In visual world experiments, sampling rates typically vary between 60 (one measurement every 16.667 ms) and 500 Hz. (one measurement every 2 ms). The sampling rates of the different eye-trackers used in our project are listed in Table 5.1. Given an average fixation duration of approximately 330 ms in scene viewing, in combination with relatively large regions of interest (i.e., screen quadrants) and time bins of 50 ms, all these systems provide ample accuracy for the study of spoken word recognition.

In order to be interpretable, the eye-tracking output must contain information in addition to gaze location and time. First, there must be a way to relate the eye-tracker's clock time to the timing of the experiment (e.g., onset of presentation of the picture and/or the spoken stimulus). Second, each sample must be coded for the experimental item, the trial, and the levels of one or more independent variables. Finally, each set of samples must specify the participant, and, if applicable, experimental list.

The output formats that have been chosen by the various manufacturers differ profoundly in the way in which this information is coded. Eye link data are in European Data Format (EDF), which is a standard binary file format for medical time series data. EDF files start with a header that contains general experiment information, as well as information about the calibration procedure. The header is followed by data records, of which there are various types. For our analyses, we relied on *messages* (i.e., a timestamp and some pre-specified string that is generated by the experiment software at certain events) and *samples* (i.e., a timestamp, followed by x- and y-coordinates of one eye and a measure of pupil dilation). The messages were used to mark important events in the time course of a trial (e.g., sound onset) and to provide information about the levels of the independent variables (per trial). We combined all three types of information by means of a custom-made script that we applied to the ASCII-converted files, which re-arranged the information for easier statistical analysis. That is, each sample was coded for time, experimental list, participant, trial, item, levels of the independent variables, and time of sound onset.[3] The data were then stored in a data table format with one row per sample and one variable per column.

Tobii gaze data are already in data table format when exported from the Tobii eye-tracker server. Each row represents a sample taken by the eye-tracker. In addition to timestamp information in milliseconds, the eye-tracker provides information about the x- and y-positions of *both* eyes, a measure of pupil dilation, and a number of parameters related to the validity of the measurement. A detailed overview of these variables can be found in the Tobii manual (http://www.tobii.com/Global/Analysis/Downloads/User_Manuals_and_Guides/Tobii_T60_T120_EyeTracker_UserManual.pdf). The sample information was automatically complemented with the relevant trial coding (i.e., the independent variables, see the section on EDF output). Importantly, the onset of the auditory instruction and visual display was marked by means of an additional variable, making it possible to align the eye-tracking measurements with the timing of the spoken and visual stimuli. The tutorial section of the E-prime extensions for the Tobii package includes instructions and examples of how to use inline scripts in E-prime to create such integrated gaze data files.

The data that were acquired by means of an SMI eye-tracker were in SMI's IDF (iView Data File) format. Similar to the EDF files, IDF files are binary files that consist of a header, followed by data records. Like EDF files, IDF files must be converted to ASCII for further processing. The actual data in IDF files can be considered a hybrid of the EDF and Tobii formats, as the data records are arranged in the form of a data table (one sample per row, marked as SMP), interspersed with messages (e.g., a timestamp in milliseconds, followed by the marker MSG and the string *start recording*). With respect to the variables, the output is similar to that of Tobii. That is, each sample includes a timestamp, x- and y-positions of *both* eyes, a measure of pupil dilation, and a number of parameters related to the validity of the measurement. In addition, trial number is included. Independent measure coding was added post hoc, by merging the corresponding E-prime output with the sample data.

Once the original data format had been standardized for all types of equipment and output files, further data processing proceeded along the following lines. First, we verified that each data file was in data table format, coded for experiment, experimental list, experimental block, participant, participant-related independent variables (i.e., first language, presence of gender in L1, age of arrival), trial, trial timing, item, and experimentally controlled independent variables (related either to our hypotheses or to counterbalancing). A further issue concerns the coding of missing data across the different file types. This is important, as each time a participant blinks (or, more accurately, each time the system loses track of the eye), the eye-tracker will report missing data. The amount of data lost to either blinks or a

[3] Recall that our audio files were standardized with respect to the length of the fragment that precedes the target.

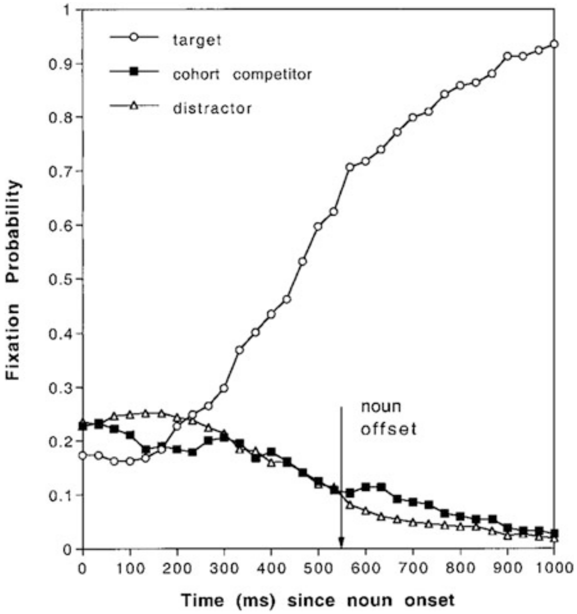

Fig. 5.6 Probabilities of fixation for the target object, the competitor object and the distractor objects (divided by two) across the time course of the noun (Dahan et al. 2000, their Fig. 5, p. 475, reprinted with kind permission from Elsevier)

poor signal is an important indicator of the overall validity of the acquired data and should therefore be subject to analysis by itself.

In order to combine data from various sources, we downsampled the gaze data to the lowest sampling rate available for each analysis. For example, if German participants were included, the lowest sampling rate was 60 Hz or one sample every 16.7 ms. Data from other eye-trackers was converted to the same rate. This step was primarily necessary for making average images of the data, since the larger bins described below for the analysis are essentially a still more stringent downsampling. The data was further minimized by excluding information that was not used in the analyses; for example, only fixations were selected and saccades and blinks were excluded. Moreover, critical items were selected and practice items and fillers were eliminated.

Third, for all trials we aligned the sound onset times to zero to establish a common time frame for all trials. Average gaze locations were aggregated per participant per trial in 50 ms time bins. Each gaze location with respect to region of interest was categorized (i.e., either one of the four quadrants of the screen or the center of the screen[4]), and with respect to the type of object (i.e., target object, competitor object or one of two distractor objects).

Fourth, probability of fixation was computed for each type of object across time, operationalized as the percentage of gaze per participant per object type per time

[4] Since each trial was preceded by a fixation cross, participants tended to fixate the center of the screen in the beginning of a trial.

bin. The averages of these percentages across participants were used to plot the data (see Fig. 5.6 for an example from Dahan et al. 2000). The y-axis represents the proportions of fixations to the different types of objects and the x-axis the time in milliseconds. The values of the y-axis range from zero to one as the data is proportional. The zero point of the x-axis represents the onset of the noun. The vertical line in the plot corresponds to the offset of the noun.

5.4.4 Statistical Approaches

In VWP studies, participants' fixations on specific regions of interest are measured across the time course of an experimental trial. These regions of interest are defined on the basis of the different objects on the visual display. For example, in the case of a 4-object display, the monitor might be divided into a grid of four quadrants. On each trial each quadrant is classified as target, competitor or distractor. The typical question for the analyses is whether different regions of interest significantly differ in their likelihoods and timing of being looked at during different experimental conditions.

Currently, there is no consensus on the best way to inferentially test the observed differences. The difficulty with eye-tracking data is that fixations in any given region and at any given time are categorical (i.e. 0 or 1), whereas the independent measure, i.e. time, is continuous. A variety of analyses have been used to examine differences in proportion fixation (see special issue 59 of the *Journal of Memory and Language,* 2008). Traditional models compare proportional fixation to different objects over time using ANOVA or t-tests. The problem with such models is that they violate the underlying statistical assumptions of these tests (Barr 2008). Recently, researchers have therefore proposed more sophisticated statistical techniques such as multi-level logistic regression (Barr 2008), growth curve analyses (Mirman et al. 2008) and generalized additive models (Wood 2006).

Suggestions for Further Reading

Dussias, P.E., J.V. Kroff, and C. Gerfen. 2013a. Visual world eye-tracking. In *Research methods in second language psycholinguistics*, ed. J. Jegerski, and B. van Patten, 93–126. New York: Routledge.

Ferreira, F., M. Tanenhaus. 2008. Language-vision interaction [Special issue]. *Journal of Memory and Language*, 57(4).

Huettig, F., J. Rommers, and A.S. Meyer. 2011a. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica* 137: 151–171.

References

Allopenna, P.D., J.S. Magnuson, and M.K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38: 419–439.

Altmann, G.T.M. 2011. The mediation of eye movements by spoken language. In *The oxford handbook of eye movements*, ed. S.P. Liversedge, I.D. Gilchrist, and S. Everling, 979–1004. Oxford: Oxford University Press.

Altmann, G.T.M., and Y. Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73: 247–264.

Barr, D.J. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language* 59: 457–474.

Boersma, P., D. Weenink. 2012. Praat: Doing phonetics by computer (Version 5.3.08) [Computer software]. http://www.praat.org/.

Clifton, C., A. Staub, and K. Rayner. 2007. Eye movements in reading words and sentences. In *Eye movements: A window on mind and brain*, ed. R.P.G. van Gompel, M.H. Fischer, W.S. Murray, and R.L. Hill, 341–372. Amsterdam: Elsevier.

Cooper, R.M. 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology* 6: 84–107.

Dahan, D., J.S. Magnuson, and M.K. Tanenhaus. 2001. Time course of frequency effects in spoken word recognition: Evidence from eye movements. *Cognitive Psychology* 42: 317–367.

Dahan, D., D. Swingley, M.K. Tanenhaus, and J.S. Magnuson. 2000. Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language* 42: 465–480.

Duchowski, A.T. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34(4): 455–470.

Dussias, P.E., J.V. Kroff, and C. Gerfen. 2013a. Visual world eye-tracking. In *Research methods in second language psycholinguistics*, ed. J. Jegerski, and B. VanPatten, 93–126. New York: Routledge.

Dussias, P.E., J.R. Valdés Kroff, R.E. Guzzardo Tamargo, and C. Gerfen. 2013b. When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition* 35: 353–387.

Ferreira, F., A. Foucart, and P.E. Engelhardt. 2013. Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, 69 (3), 165−182.

Grüter, T., C. Lew-Williams, and A. Fernald. 2012. Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research* 28: 191–215.

Henderson, J.M., and F. Ferreira. 2004. Scene perception for psycholinguists. In *The interface of language, vision, and action*, ed. J.M. Henderson, and F. Ferreira, 1–58. New York: Psychology Press.

Hopp, H. 2013. Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research* 29(1): 33–56.

Huettig, F., and J.M. McQueen. 2007. The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language* 57(4): 460–482.

Huettig, F., J. Rommers, and A.S. Meyer. 2011b. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica* 137: 151–171.

Lew-Williams, C., and A. Fernald. 2007. Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science* 33: 193–198.

Lew-Williams, C., and A. Fernald. 2010. Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language* 63: 447–464.

Loerts, H., M. Wieling, and M.S. Schmid. 2013. Neuter is not common in Dutch: Eye movements reveal asymmetrical gender processing. *Journal of Psycholinguistic Research* 42(6): 551–570.

Matin, E., K.C. Shao, and K.R. Boff. 1993. Saccadic overhead: Information-processing time with and without saccades. *Perception and Psychophysics* 53: 372–380.

Mirman, D., J.A. Dixon, and J.S. Magnuson. 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language* 59: 475–494.

McQueen, J.M., and M.C. Viebahn. 2007. Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology* 60: 661–671.

Oehler, H., and O.S. Gerretsen. 1973. *Duitse woordenschat—Alfabetische basisvocabulaire met systematische uitbreiding*. Groningen: Wolters Noordhoff.

Oostdijk, N. 2000. The spoken Dutch corpus. Overview and first evaluation. In Gavralidou, M., G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer (eds.), *Proceedings of the second international conference on language resources and evaluation*, 887−893. Paris: ELRA.

Paris, G., A. Weber, and M.W. Crocker. 2006. *German morphosyntactic gender and lexical access*. Poster presented at the 12th annual conference on architectures and mechanisms for language processing (AMLaP 2006), Nijmegen, Netherlands.

Schneider, W., A. Eschman, and A. Zuccolotto. 2002. *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.

Snodgrass, J.G., and M. Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6: 174–215.

Spivey, M.J., and V. Marian. 1999. Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science* 10: 281–284.

Tanenhaus, M.K., and J.C. Trueswell. 2006. Eye movements and spoken language comprehension. In *Handbook of psycholinguistics*, 2nd ed, ed. M. Traxler, and M. Gernsbacher, 86–900. Amsterdam: Elsevier.

Tanenhaus, M.K., M.J. Spivey-Knowlton, K.M. Eberhard, and J.C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268: 1632–1634.

Weber, A., G. Paris. 2004. *The origin of the linguistic gender effect in spoken-word recognition: Evidence from non-native listening*. Poster presented at the 26th Annual Meeting of the Cognitive Science Society (CogSci 2004), Chicago, IL.

Wood, S. 2006. *Generalized additive models: An introduction with R*. Boca Raton: Chapman & Hall/CRC Press.