Processing Grammatical Gender in Dutch:

Evidence from Eye Movements

Susanne Brouwer¹, Simone Sprenger², and Sharon Unsworth¹

¹Radboud University, Nijmegen, The Netherlands
²University of Groningen, Groningen, The Netherlands

Running head: Processing gender in Dutch

Address of correspondence:

Susanne Brouwer

Postbus 9103

6500 HD Nijmegen

T: +31 (0)24 361 2843

E: S.Brouwer@let.ru.nl
Research highlights

• The moment at which gender aids comprehension varies with production proficiency
• Successful online gender comprehension may precede successful production
• Successful gender production may trigger successful online gender comprehension
Abstract

Previous research has demonstrated that grammatical gender in Dutch is typically acquired late (e.g., Blom, Polišenská, & Weerman, 2008; Van der Velde, 2003). Most of this work used production data only, and consequently children’s knowledge of Dutch gender may have been underestimated. In this study we therefore examined whether 49 4- to 7-year old Dutch-speaking children (and 19 adult controls) were able to use gender-marking in the article preceding the object label during online sentence processing to (a) anticipate the upcoming object label or to (b) facilitate the processing of that label as it is presented. In addition, we investigated whether children’s online processing and production of gender-marking on articles were related.

In an eye-tracking task, participants were presented with sentences and visual displays with two objects, representing nouns of either the same (uninformative) or different (informative) gender. The children were divided into a non-targetlike and a targetlike group on the basis of their scores for neuter nouns in the production task. Our analyses examined whether participants could use gender-marking anticipatorily (i.e., before the onset of the noun) and facilitatively (i.e., from noun onset).

Results showed that Dutch-speaking adults and children who were successful in production used gender-marking anticipatorily. However, children who did not systematically produce gender-marked articles, only used gender-marking facilitatively.
These findings reveal that successful online comprehension may in part be possible before targetlike production is completely in place but at the same time, targetlike production may be a trigger for online comprehension to be completely successful.

Keywords: Grammatical gender; eye-tracking; online sentence processing; language development
Introduction

Studying how children process language can help us better understand children’s linguistic development. In the past fifteen years, the focus on explaining children’s production data has shifted to new techniques allowing us to access children’s online processing (Snedeker & Huang, 2015). The link between production and online processing remains poorly understood, however. The goal of the present study is therefore to use both production and online processing techniques to better understand the acquisition of grammatical gender in Dutch-speaking children.

In many languages across the world, nouns have grammatical gender. For example, in Spanish and French, masculine articles precede masculine nouns, whereas feminine articles precede feminine nouns. Some languages, like German and Russian, have a third category, neuter, and other languages even have four or more noun classes (Corbett, 1991). Dutch has a two-way gender system and makes a distinction between common and neuter gender. In this language, grammatical gender is marked on a number of agreeing elements accompanying the noun or referring to it; these include definite and demonstrative articles, relative pronouns, and adjectives (see Blom et al., 2008 for an overview). The focus in the present study is on definite articles. Common nouns are preceded by the definite article *de*, as in *de schoen* ‘the shoe’, whereas neuter nouns are preceded by the definite article *het*, as in *het huis* ‘the house’. 
Most of the prior work on Dutch grammatical gender has relied exclusively on production data and has shown that grammatical gender in Dutch is acquired relatively late by monolingual children (e.g., Blom, Polišenská, & Weerman, 2008, Bol & Kuiken, 1988, Van der Velde, 2003). More specifically, Dutch-speaking children continue to make production errors with neuter gender until at least age six, overgeneralizing common gender forms of the article (de ‘the’) to neuter nouns (*de huis instead of het huis ‘the house’; Bol & Kuiken, 1988; Polišenská, 2010; Van der Velde, 2003). This is in stark contrast to the acquisition of gender-marking in other languages, such as Spanish, French and German, where children have been shown to produce targetlike article-noun combinations from an early age (e.g., Hawkins & Franceschina, 2004; Karmiloff-Smith, 1978; Lew-Williams & Fernald, 2007; Pérez-Pereira, 1991; Van Heugten & Shi, 2009).

To better understand this cross-linguistic difference, research is needed to establish whether Dutch-speaking children really are so slow in acquiring grammatical gender, or whether this apparent delay is modality-specific. Thus far, however, almost all of the studies investigating Dutch gender acquisition (e.g., Bol & Kuiken, 1998; Polišenská, 2010; Van der Velde, 2003) have relied exclusively on production data. There are a few studies which have collected non-production data in Dutch but as we shall show presently,
these studies are limited in the age ranges of the children tested (either very young or very old) and techniques used (mostly offline).

Johnson (2005) tested 28-month-old children using a preferential looking paradigm. Children listened to sentences containing correct (e.g., het\_NEUTER huis\_NEUTER ‘the house’) or incorrect article-noun combinations (e.g., *de\_COMMON huis\_NEUTER ‘the house’) while viewing two objects (a target and a distractor) on a screen. The names of the two objects had different genders (and hence the gender-marked article was informative) or the same gender (in which case the gender-marked article was uninformative). Results showed that children shifted from distractor to target noun more quickly in contexts where gender was informative than when it was uninformative, but only for nouns preceded by common de. Thus, while toddlers at this age may be sensitive to articles (Van Heugten & Johnson, 2011), they are not yet sensitive to the grammatical gender-marking instantiated on both articles.

Unsworth and Hulk (2010) used a forced choice grammaticality judgment task with 4 to 6 year olds. Children were asked to choose between the descriptions of a picture given by two different puppets, one consisting of the correct article+noun combination, e.g., het\_NEUTER boek\_NEUTER ‘the book’, and the other of an incongruent counterpart, e.g., *de\_COMMON boek\_NEUTER ‘the book’. The results showed that the children accepted the incongruent combination for neuter nouns more frequently than for common nouns, and whilst some children were better on judgment than production, the average scores on the
judgment task, at around 70%, were still not at ceiling and numerous children performed at chance level (see also Unsworth, 2013a for further discussion of these data). Another study with non-production data showing ceiling performance for 11- to 13-year-old Dutch-speaking children come from a study by Brouwer, Cornips and Hulk (2008), using a similar judgment task in which children evaluated congruent and incongruent article-noun combinations separately from each other.

In sum, the available non-production studies in Dutch have primarily examined whether children can detect errors in article-noun combinations. The children tested thus far were either relatively young (19 to 28 months) (Johnson, 2005; Van Heugten & Johnson, 2011) or relatively old (11 to 13 year olds) (Brouwer et al., 2008), resulting in effects for common gender only or ceiling effects, respectively. Whilst the study by Unsworth and Hulk (2010) targeted an age group where variability is expected (4 to 6 year olds), it employed a task (grammaticality judgment) which required children to make meta-linguistic judgments about gender incongruency; given that younger children may be unable to complete such a task (cf. Unsworth, 2013) and older children may rely on explicit knowledge in their responses, this method may lead to an over- or underestimation of children’s gender knowledge. As yet, we know very little about how Dutch-speaking children make use of grammatical gender in “real life” language processing.
Grammatical gender in Spanish has, however, been one area of interest in the child processing literature. For example, Lew-Williams and Fernald (2007) used a looking-while listening task to examine whether gender-marking on articles helps monolingual Spanish adults and children (aged 2;10 to 3;6) to identify the referents of upcoming nouns. In a similar vein to the study by Johnson (2005), participants were presented with two scenarios, one where the gender-marked article was informative, distinguishing between two referents with different genders, e.g., *la pelota* ‘the ball’ vs. *el zapato* ‘the shoe’, and one where the two referents had the same gender, e.g., *la pelota* ‘the ball’ vs. *la galleta* ‘the cookie’, and hence the gender-marked article was uninformative. Adults were found to identify the target referent more quickly when the gender-marked article preceding the noun was informative than when it was uninformative. Although children’s processing times were longer, they behaved similarly, that is, in contexts where the gender-marking on the article was informative, they were able to make use of it. These results are in line with those for production in Spanish (e.g., Pérez-Pereira, 1991), which show that children produce targetlike gender-marked articles from a very early age, arguably due to this language’s transparent gender system (Harris, 1991).

In the present study, we want to use such online processing techniques to better understand the acquisition of grammatical gender in Dutch as this language is opaque. As there are only few (online) comprehension studies on gender in Dutch-speaking children,
our first research question asks whether Dutch-speaking children use gender-marking on articles during online sentence comprehension. Following Lew-Williams & Fernald’s (2007) experimental set-up, we tested 4- to 7-year-old Dutch-speaking children, as well as an adult control group. We chose this age range for the children because previous research has shown that during this period, children are in a transition stage, moving from non-targetlike to targetlike production of the (neuter) definite article (e.g., Blom et al., 2008; Unsworth & Hulk, 2010; Van der Velde, 2003). Instead of a looking-while-listening paradigm, we used the visual-world eye-tracking technique (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see Sedivy, 2010, for an overview of the use of eye-tracking in language acquisition research), which requires no hand-coding of gaze allocation. The temporal resolution of this technique allows for real-time precision during spoken language processing. As in a looking-while-listening set-up, eye-movements are closely time-locked to the speech input and are typically taken to be related to underlying activation levels of word candidates.

In the study by Lew-Williams and Fernald (2007) the central question was whether Spanish-speaking adults and children were able to use gender-marked articles as a cue in interpreting sentences. The authors compared the reaction times (RT) in different-gender versus same-gender trials, calculating RTs to the target noun in those trials on which participants were looking at the distractor at article onset and shifted to the target picture
by the end of the noun. The RT was the latency of the first shift to the correct picture measured from article onset until 1000 ms thereafter. Their time-window for analysis thus corresponded to the eye movement pattern whilst hearing the gender-marked article and during the unfolding of the noun. In other words, their analyses focused on whether the children used this cue *facilitatively.*

Language processing is not just passively integrating information (i.e., during the unfolding of the noun) but also proactively constructing interpretations by generating predictions on the basis of the information that is being integrated (i.e., during the gender cue; e.g., Federmeier, 2007; Pickering & Garrod, 2013). This so-called predictive or anticipatory processing has been described as a central ingredient of efficient communication (Altmann & Mirković, 2009; Jaeger & Snider, 2013). For a more complete understanding of how listeners make use of gender-marking in online processing, the present study examines *predictive* as well as *facilitative* processing. To this end, we present article-adjective-noun sequences (cf. article-noun in previous work, e.g., Johnson, 2005; Lew-Williams & Fernald, 2007), allowing an analysis of anticipatory (i.e., before the noun has been uttered) and facilitative looks (i.e., while the noun is being uttered).

Such article-adjective-noun sequences have been used previously with French-speaking children (Melançon & Shi, 2015) and with Dutch-speaking adults (Loerts, Wieling, & Schmid, 2013). Loerts and colleagues found a gender effect for common but
not for neuter nouns. Perhaps surprisingly, participants appeared to “pre-select” neuter targets when the competitor shared gender as opposed to when the competitor differed in gender. It is not clear what exactly the cause of this result is, as the authors themselves acknowledge. In any case, their findings are not entirely unproblematic because of their use of four picture displays with, amongst others, colour (mis)matching competitors. Adding a colour condition seems to have overcomplicated their design and this may have contributed to their failure to find an effect for neuter gender. We will return to this issue in the Discussion. In our set-up, we will remove the effect of colour to see how robust their findings are.

Our second research question concerns the relation between the production and the comprehension of gender. Unsworth and Hulk’s (2010) study, in which both production and judgment data were collected, suggested that production may underestimate children’s underlying knowledge of grammatical gender, and this has indeed been claimed for similar findings with bilingual children (e.g., Blom & Vasic, 2011; Unsworth, 2013). To test this “comprehension-precedes-production” claim, relatively uncontroversial in child language studies (e.g., Clark, 1993, but see Hendriks & Koster, 2010 for the inverse pattern), we divide our children into targetlike and non-targetlike producers using data from an elicited production task (see Method section for more details on this division). We then compare the eye gaze patterns of these two groups on our eye-tracking task.
Our predictions are as follows. In informative contexts, we expect adults and targetlike-producing children to successfully use the available gender cue during online sentence comprehension. We predict that they use gender information as soon as possible, i.e., anticipatorily, even before the onset of the noun.

The non-targetlike producers may behave in one of two ways. If comprehension does indeed precede production, the eye gaze behavior of the non-targetlike producers should pattern like the targetlike producers. In other words, although non-targetlike children are not able to produce targetlike gender-marked articles, they may be able to use them during online sentence comprehension. This idea that errors in gender are a production-specific performance problem has also been put forward in the second language acquisition literature in the guise of the Missing Surface Inflection Hypothesis (Haznedar & Schwartz, 1997; Prévost & White, 2000). This proposal may thus also hold for monolingual acquisition. If non-targetlike producers indeed show better performance in comprehension than production, this would also indicate that previous findings showing a comparatively late development of Dutch gender may in part reflect an over-reliance on production data. However, it is also possible that the children in the non-targetlike group may not yet have acquired gender, and are thus also not able to use gender as a cue during online sentence comprehension.
Method

Participants

Nineteen native adult controls (M=22;3; SD=0;3; range=19;7 to 29;0 years) took part in this study. They were students at the University of Groningen who participated in exchange for course credits. They provided written consent. Forty-nine monolingual Dutch-speaking children (M=5;8 SD=0;2; range=4;1 to 7;11 years) participated in this study. Written consent was provided by caregivers for children.

Materials

Eye-tracking task. The speech stimuli were short Dutch sentences spoken by a female native speaker of Dutch. These sentences were questions about the visual display, in one of two forms:

(1) Zie je [het | de] [adjective] [noun]? Vind je ‘m mooi?

‘Do you see the [adjective] [noun]? Do you like it?’

(2) Waar is [het | de] [adjective] [noun]? Zie je ‘m?

‘Where is the [adjective] [noun]? Do you see it?’
Each first question ended in a target noun, which was preceded by an article (de or het) and a color adjective (gele ‘yellow’, rode ‘red’, blauwe ‘blue’, or groene ‘green’). The adjective was included for two reasons. First, it extends the period during which possible gender effects from the article can be observed. Second, it allows us to differentiate between effects of grammatical knowledge and knowledge of co-occurrence probabilities between the article and the noun (Dahan, Swingley, Tanenhaus, & Magnuson, 2000). More specifically, we need to tease apart these two alternatives to be sure that our effects are really driven by gender information (i.e., grammatical knowledge) and not by the strong tendency for nouns to co-occur with a given article. An adjective was thus added because the co-occurrence frequencies between the article and the adjective, on the one hand, and the adjective and the target, on the other, will be much lower than the co-occurrence frequency between the article and the target noun (see also Loerts et al., 2013; Mélançon & Shi, 2015).

After stimulus recording, the onsets and durations of the articles, the adjectives and the nouns were measured and averaged in Praat (Boersma & Weenink, 2005). The stimuli were then manipulated in Adobe Audition© 3.0 in order to align the onsets and durations of the critical words across stimuli. The purpose of this adjustment was to reduce between-item differences in timing and to ensure that any observed effects were not due to the sentences in question being pronounced at a slower rate. Article onset was set to 433 ms (duration:
M=113 ms, SD=30 ms), adjective onset to 547 ms (duration: M=320 ms, SD=42 ms), and noun onset to 866 ms (duration: M=726 ms, SD=62 ms). The resulting stimuli were checked by a native Dutch speaker to confirm that they sounded fluent and natural.

Eight target nouns were selected from a wordlist for Dutch-speaking 4- to 6-year olds (Damhuis, de Glopper, Broers, & Kienstra, 1992), four of which were inanimate (schoen.de ‘shoe’, lamp.de ‘lamp’, huis.het ‘house’, bed.het ‘bed’), and four of which were animate (hond.de ‘dog’, koe.de ‘cow’, paard.het ‘horse’, schaap.het ‘sheep’; within-subjects factor animacy). In both categories, half the items were of common gender and half were neuter (i.e., within-subjects factor Gender). Inanimate items were used with question type (1) and animate items were used with question type (2).

Each target noun was paired with two pictures on the visual display (see Figure 1). One of the two pictures always referred to the target noun, whereas the other picture either shared the same gender of the target noun (e.g., de.COMMON schoen.COMMON – de.COMMON lamp.COMMON) or had a different gender than the target noun (e.g., de.COMMON schoen.COMMON – het.NEUTER huis.NEUTER; i.e., within-subjects factor Condition).

To exclude possible effects of phonological competition, the initial segments of the names of the two pictures did not overlap. In each pair, the first noun served as target and the other as distractor. Subsequently, each pair was quadrupled and re-ordered, such that each first noun of a pair appeared four times as the target noun of the question: twice on
the right and twice on the left side of the display. Half of them were same-gender trials and the other half were different-gender trials. This resulted in a set of 32 experimental items. The complete item set was included in both experimental lists, which only differed with respect to the allocation of a color to a specific target-distractor combination. Within lists, each color appeared equally often. The order of the experimental items was pseudo-randomized, such that neither the target’s gender and side, nor the color of an item were predictable.

The visual images of the eight target objects were line drawings retrieved from Snodgrass and Vanderwart (1980). All images were of comparable visual complexity. Object color was edited in Paint, resulting in four different versions of each image.

In addition to the experimental items, eight filler items were created that slightly deviated from the experimental items. The images in these items were multi-colored and more complex than those in the target items. The related questions focused on some funny aspect of the display instead of the color (e.g., *Zie je de gekke aap? Leuk, hè?* ‘Do you see the funny monkey? Nice, isn’t it?’). These items were intended to create a short break from the pattern created by the experimental items and to maintain children’s attention.
Production and vocabulary task. Children’s productive knowledge of grammatical gender-marking on definite articles in Dutch was tested using an elicited production task (see Unsworth & Hulk, 2010; Unsworth et al., 2014, for details; following Blom et al., 2008). This task was used to test children’s gender production of the same eight nouns that were included in the eye-tracking experiment. Children were presented with two pictures on a computer screen (e.g., a white shoe and a pink shoe). The experimenter asked the children to name the pictures on the screen using the following prompt, designed to elicit adjectival agreement with indefinite nouns: “Look! Here we see two pictures. This is a... (child: white shoe). And this is a... (child: pink shoe)”. To elicit definite articles, two additional objects (e.g., a ball and a dog) subsequently appeared in front of or next to one of the shoes. The experimenter then asked the child to complete the following prompt:
“The ball is in front of... (child: the white shoe). And the dog is next to... (child: the pink shoe)”. Each of the eight nouns was repeated four times (i.e., two with an indefinite article and two with a definite article) with different adjectives (i.e., white vs. pink, orange vs. purple, and/or small vs. large). This resulted in a set of 16 definite article items of which 8 were common and 8 were neuter nouns. Twelve filler items that tested for verb form and placement were interspersed and served as distracters. Two lists were created. The order of the lists was pseudo-randomized.

In addition to the production task, children were also tested using a Dutch version of the Peabody Picture Vocabulary Test-Third Edition (PPVT-III-NL; Dunn, Dunn, & Schlichting, 2005).

Procedure

Data collection took place at the eye-tracking lab at the University of Groningen or at Utrecht University. For the children, the three tasks were administered in the following order: production, eye-tracking and vocabulary. The production task always preceded the eye-tracking task because we did not want children to hear the correct gender production during the eye-tracking task before they had to produce the articles themselves.

Before each task, children were given verbal instructions. They were specifically instructed about how to perform an eye-tracking task (e.g., try to reduce body movements,
keep your eyes on the screen when you see a picture, etc.). Between tasks, they were praised for their performance and provided with the opportunity to ask questions. At the end of the session, they were presented with a small present (some colorful stickers). The duration of each individual task did not exceed 10 minutes, so that a typical session lasted about 30 minutes. The group of adults participated in the eye-tracking task only. Their session lasted about 10 minutes.

*Production task.* Participants were tested individually. In the production task, each child was presented with a total of 56 items in one of two lists. Each list was preceded by 5 additional practice items to familiarize the children with the task. As noted above, only the definite article data were included in the analysis here. An accuracy score was computed for each gender by dividing the total number of common/neuter nouns produced with the correct article (de/het) by the total number of nouns produced with either definite article.

*Eye-tracking task.* Participants were tested individually. Prior to the experiment, the experimenter showed each participant a picture book with all target and distractor objects. The participants were asked to name the objects in the pictures to check that they knew the relevant lexical item. If they were not sure about it, the experimenter told them the name of the picture. Each page of the picture book contained two pictures of each object, that is, both versions used in the experiment. Participants were asked to complete the following
prompt: “These are two... (participant: shoes)”. Asking for the plural form avoided any priming effects of the definite article.

Participants were seated at a comfortable distance from the computer screen. A calibration procedure prior to the experiment allowed the eye tracker (Tobii T120) to control for drifts. No headrest was used, which allowed participants to sit in a natural posture in front of the screen. All participants were instructed to keep body movements to a minimum. With respect to the experiment itself, participants were told to listen to the sentences carefully and to look at the object that appeared in the spoken utterance. They were not given an explicit task, such as clicking or pointing (cf. Huettig & Altmann, 2005). Following Lew-Williams and Fernald (2007), we presented pairs of images on a visual display, each image centered in the left or right half of the screen.

Experimental list (two levels) was varied between-subjects. In each list, 8 filler items (i.e., image pairs) and 32 experimental items were alternated in such a way that a filler item preceded a block of four experimental items. The spoken instructions were presented via a set of standard computer speakers shortly after presentation of the images. The volume was adjusted individually during the first practice item. During the complete task, participants’ eye movements were monitored at a sample frequency of 120 Hz.

The structure of an individual trial was as follows: each item was preceded by a fixation cross that appeared in the middle of the display. To ensure that the subjects’ attention was
on the display, the subsequent presentation of the images was contingent on a minimum gaze duration of 1000 ms on the fixation cross. A given set of images was presented for two seconds before the spoken instructions would start (preview time). The set remained on the screen during the complete duration of the spoken stimuli (ca. 3000 ms), plus an additional two seconds (search time).

**Vocabulary task.** After the eye-tracking experiment, children completed the vocabulary task. The PPVT-III-NL (Dunn et al., 2005) is a standardised receptive vocabulary test requiring children to point to one of four pictures based on a single word given by the experimenter. The procedure and scoring followed the standard outlined in the manual. This task took approximately 10 minutes.

**Results**

**Production and vocabulary tasks**

To understand whether performance in online comprehension is directly related to performance in production, we used the production data to divide our children into two groups for subsequent analysis of the eye-tracking data. Given the previous findings that children primarily experience difficulties with knowledge of neuter and not common gender (e.g., Blom et al., 2008; Unsworth & Hulk, 2010; Van der Velde, 2003), we grouped the children according to their accuracy scores for neuter nouns in the production task. Given the extensive between-child variability in the acquisition of Dutch gender, this division will likely provide a better picture of the acquisition process than grouping the children according to age.
For this division we aimed to approximate an equal number of participants in each group. However, using a median-split would have meant that some children with the same percentage correct would be categorized in the targetlike and others in the non-targetlike group. We therefore decided to use a cut-off point such that children with this same percentage all fell within the same group. The targetlike production group thus consisted of 26 children with an accuracy percentage of 75% or higher (i.e., minimally 6 out of 8 neuter items correct) and the non-targetlike production group consisted of 23 children with an accuracy percentage of 62.5% or lower (i.e., maximally 5 out of 8 neuter items correct).

The means and standard deviations on the production task of the two groups are presented in Table 1, along with their age and standard score on the PPVT-III-NL (Dunn et al., 2005). Independent samples t-test showed that the two groups significantly differed in their production scores on the neuter nouns ($t(47)=-4.43; p<.0001$) and in age ($t(47)=-16.23; p<.0001$). The targetlike production group scored better on the neuter nouns and were older than the non-targetlike production group. No differences were found on common nouns and PPVT scores between the two groups ($p>.05$). Pearson correlations showed that production scores on the neuter items correlated positively with age ($r(49)=.60; p<.001$) and PPVT scores ($r(49)=.38; p<.01$). This indicates that the higher the children’s age and vocabulary scores, the better the performance on the production of neuter items.
Table 1: *Overview of the two child groups, divided according to their production scores on the neuter nouns.*

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>M_{COMMON}</th>
<th>M_{NEUTER}</th>
<th>Age</th>
<th>Range</th>
<th>PPVT score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targetlike</td>
<td>26</td>
<td>90 (17)</td>
<td>91 (9)</td>
<td>6;3</td>
<td>4;1-7;11</td>
<td>115 (9)</td>
</tr>
<tr>
<td>Non-targetlike</td>
<td>23</td>
<td>77 (28)</td>
<td>24 (26)</td>
<td>5;0</td>
<td>4;1-7;3</td>
<td>106 (13)</td>
</tr>
</tbody>
</table>

SDs are reported between parentheses.

**Eye-tracking task**

Figure 2 and 3 show the time course of fixation proportions to targets from article onset for adults (Fig. 2) and for children in the non-targetlike and the targetlike production group (Fig. 3). The latency to plan and launch a saccade has been estimated at 200 ms (Matin, Shao, & Boff, 1993). This means that the earliest point at which it is expected that fixations are driven by acoustic information from the article is around 200 ms after article onset. We therefore conducted our analyses from this point on. We arcsine-transformed the mean fixation proportions to targets in our analyses (cf. Cohen & Cohen, 1983).

We conducted four types of analyses on our data. First, to measure whether adults and children are able to *anticipate* the upcoming noun on the basis of the article, we conducted a 3 (Group) x 2 (Condition) x 2 (Gender) mixed ANOVA with the arcsine-transformed mean fixation proportions from article onset until noun onset (taking into account the 200 ms it takes to plan and launch a saccade) as the dependent variable. Group (adults vs. targetlike production vs. non-targetlike production) was included as between-subjects variable and Condition (same vs. different) and Gender (common vs. neuter) as within-subjects variables.
Secondly, to measure whether the recognition of the noun is facilitated by the article in adults and in children, we conducted a 3 (Group) x 2 (Condition) x 2 (Gender) mixed ANOVA with the arcsine-transformed mean fixation proportions as the dependent variable. The time window initiated from noun onset until the time that the same and different gender functions converged\(^2\). Following Dahan et al. (2000), we used visual inspection to decide the endpoint of analysis. Figures 2 and 3 show that the convergence of the functions were dependent on Group, that is, the same and different lines of adults converged earlier than those of the children. We therefore chose to include the data from noun onset until 1000 ms for adults, from noun onset until 1200 ms for target-like children and from noun onset until 1400 ms for non-targetlike children. Note that similar timing differences between adults and children can be found in previous eye-tracking research (e.g., Lew-Williams & Fernald, 2007). In our model, Group (adults vs. targetlike production vs. non-targetlike production) was included as a between-subjects variable and Condition (same vs. different) and Gender (common vs. neuter) as within-subjects variables. For both of these mixed ANOVAs, effect sizes were computed using partial eta-squared ($\eta_p^2$).

Thirdly, we examined the time course of exactly when in time the different groups started using the gender cue for anticipation or facilitation of the target noun. For each group separately, we conducted a 2 x 2 Repeated Measures ANOVAs on the arcsine transformed mean proportion fixations for Condition (same vs. different) and Gender
(common vs. neuter) at each 100-ms time bin. To minimize the possibility that differences measured by these multiple statistical comparisons might have arisen by chance, we consider the earliest time of anticipation or facilitation when the first window of five consecutive time windows are significant at a level of \( p < 0.05 \) (for a similar type of time window analysis, see Borovsky, Elman, & Fernald, 2012).

Fourthly, we examined whether children’s performance on the eye-tracking task correlated with their standardized vocabulary scores and with age (in months) using a Pearson \( r \) correlation test. Performance on the eye-tracking task was divided into anticipation ability (i.e., arcsine-transformed target fixations in different minus same gender trials in the anticipation window) and facilitation ability (i.e., arcsine-transformed target fixations in different minus same gender trials in the facilitation window).

1. Anticipation

The analysis on anticipation effects showed a main effect of Condition \( (F(1,65)=5.84; p<.05, \eta^2=.08) \) and a significant interaction between Condition and Group \( (F(2,65)=3.91; p<.05, \eta^2=.11) \). The main effect of gender and its interactions did not reach significance \( (p>.05) \). Paired \( t \)-tests showed an effect of Condition for adults \( (t(18)=-2.44; p<.05, \text{Cohen's } d=-0.55) \) and for the targetlike production group \( (t(25)=-2.56; p<.05, \text{Cohen's } d=-0.52) \), indicating that both groups demonstrated more fixations to the target noun in the different-gender (adults: \( M_{\text{SAME-DIFF}}=-.08; SE_{\text{SAME-DIFF}}=.03 \); targetlike group: \( M_{\text{SAME-DIFF}}=-.08; SE_{\text{SAME-DIFF}}=.03 \)) than the same-gender condition. The effect of Condition did not reach significance in the non-targetlike production group \( (p>.05) \).
2. Facilitation

The analysis on facilitation effects showed a main effect of Condition \( (F(1,65)=31.29; p<.0001, \text{partial } \eta^2=.33) \), a main effect of Group \( (F(2,65)=27.18; p<.0001, \text{partial } \eta^2=.46) \), and a significant interaction between Condition and Gender \( (F(1,65)=8.83; p=.01, \text{partial } \eta^2=.12) \). Post-hoc Bonferroni tests on the effect of Group showed that the overall eye gaze pattern (i.e., proportionally more looks to target nouns) differed for adults compared to both child groups (both \( p \text{'s}<.0001 \)), whereas the eye gaze data of the child production groups patterned similarly \( (p>.05) \). This indicates that the proportion of adults’ looks to targets was higher in general than those of the children. To unpack the Condition by Gender interaction, a paired \( t \)-test showed that the facilitation effect of Condition is significant for the neuter nouns \( (M_{\text{SAME-DIFF}}=-.09; \text{SE}_{\text{SAME-DIFF}}=.02) \), but not for the common nouns \( (M_{\text{SAME-DIFF}}=.006; \text{SE}_{\text{SAME-DIFF}}=.02; t(67)=3.60; p<.01, \text{Cohen’s } d=0.28) \).

3. Time course

The results of the time course analysis are reported in Table 2. This analysis showed that adults started using the gender cue to anticipate the upcoming noun around 500 ms after article onset. Note that making a saccade takes about 200 ms, hence adults only needed to hear about 300 ms of the article and the adjective to anticipate which noun is coming up. The data of the targetlike production group of children showed a similar pattern as the adults. The effect of Condition also initiated around 500 ms.

In addition, Gender also seems to play an important role for the targetlike production group. More specifically, as can be seen in Figures 2 and 3, the effect of Condition is larger for the neuter than for the common nouns. Finally, the non-targetlike production group
showed that they are overall slower than the adults and the targetlike production group in demonstrating any effects. These children needed around 800 ms of the acoustic signal to process, which indicates that they only use gender as a facilitative and not as an anticipatory cue.

4. Eye movements, vocabulary scores and age

The results of the Pearson $r$ correlations tests revealed no significant relations between performance in the eye-tracking task (measured as anticipation or as facilitation) and the children’s vocabulary scores (all $p$’s > .1). This is likely because the children in our sample generally had very high scores on the vocabulary task, within and often above age-appropriate norms ($M=112$, $SD=12$; see also Table 1). The correlation between anticipation and age (in months) showed a marginally significant relation ($r(49)=.271$; $p=.061$). However, plotting the data revealed that this marginal relationship was mainly driven by two outliers. When these outliers were removed, the marginal relation disappeared ($r(47)=.172$; $p>.1$). The relation between facilitation ability and age (in months) was not significant ($p>.1$). Overall, we thus conclude that vocabulary scores and age are not the determining factors on our outcome measures (anticipation and facilitation).
Figure 2: Time course of gaze allocation to targets in the same and different-gender condition for adults, overall and split by gender (common vs. neuter). The vertical dotted line represents the end of the time window analysis.
Figure 3: Time course of gaze allocation to targets in the same and different-gender condition for children, overall, divided into non-targetlike and targetlike production groups, and split by gender (common vs. neuter). The vertical dotted line represents the end of the time window analysis.
Table 2: *F*-scores for Condition, Gender, and Condition by Gender, measured from 200 ms after article onset per group.

<table>
<thead>
<tr>
<th>Time window</th>
<th>Adults</th>
<th>Non-targetlike production group</th>
<th>Targetlike production group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cond</td>
<td>Gend</td>
<td>Cond*Gend</td>
</tr>
<tr>
<td>200-299</td>
<td>0.23</td>
<td>0.80</td>
<td>0.03</td>
</tr>
<tr>
<td>300-399</td>
<td>0.76</td>
<td>1.06</td>
<td>0.28</td>
</tr>
<tr>
<td>400-499</td>
<td>2.32</td>
<td>0.20</td>
<td>0.69</td>
</tr>
<tr>
<td>500-599</td>
<td>7.44*</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>600-699</td>
<td>18.91**</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>700-799</td>
<td>29.63***</td>
<td>0.18</td>
<td>0.62</td>
</tr>
<tr>
<td>800-899</td>
<td>9.28*</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>900-999</td>
<td>5.99*</td>
<td>1.75</td>
<td>0.01</td>
</tr>
<tr>
<td>1000-1099</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1100-1199</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1200-1299</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1300-1399</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Cond = Condition (same vs. different gender); Gend = Gender (common vs. neuter nouns); #p<.05; *p<.01; **p<.001; ***p<.0001; Shaded cells are significant effects (at least five consecutive time windows)*
Discussion

The aim of the current study was twofold. First, we examined whether Dutch-speaking children and adult controls use grammatical gender during online sentence comprehension. Secondly, we investigated whether children’s online eye gaze behavior was related to their production of gender-marking on articles with the same target nouns. In order to do this, we tested the same group of children on an elicited production task (following Blom et al., 2008) and an online comprehension task (following the experimental set-up of Lew-Williams & Fernald, 2007). The children were divided into two groups, i.e., a non-targetlike and a targetlike production group, on the basis of their accuracy scores for the neuter nouns of the production task. Online processing was tested in an eye-tracking task. Dutch sentences were heard in combination with visual displays showing two pictures representing nouns of either the same (uninformative) or different gender (informative). We predicted that if our participants had knowledge of grammatical gender and more specifically, how this is marked on definite articles, then they should be better able to shift their gaze to the target object when presented with an informative gender-marked article as opposed to an uninformative gender-marked article. This prediction was investigated by not only examining whether participants could use the gender cue facilitatively (i.e., from noun onset), as has been done in most previous studies on grammatical gender processing, but also whether they could use gender to anticipate the upcoming noun (i.e., before the onset of the noun).

The results of the native-speaker adults showed, as predicted, that they were able to anticipate the upcoming noun on the basis of the article. This result is in line with previous work on gender processing with native-speaker adults in other languages (e.g., Dahan et
al., 2000; Grüter, Lew-Williams, & Fernald, 2012; Lew-Williams & Fernald, 2007). The processing advantage furthermore persisted during the unfolding of the noun. These findings were supported by our time course analysis, which revealed that adults started using the gender cue around 300 ms until 800 ms after article onset (taking the 200 ms saccade into account).

The behavior of the targetlike producing children, who had high accuracy scores on the neuter nouns in the production task, patterned like the adults’. Before the onset of the noun, this group also showed proportionally more looks to the target noun in the different-gender condition, where the article was informative, than in the same-gender condition. The time course analysis demonstrated that the anticipatory effect for this group initiated at the same time as the adults (around 300 ms) and persisted until around 1200 ms after article onset. The reason why the effect persisted longer in the children than in the adult group is probably due to children’s slower speech processing speed and cognitive resource limitations (Snedeker & Huang, 2015). Similar delays have been found in previous work comparing adult and child data (e.g., Lew-Williams & Fernald, 2007).

The non-targetlike producing children, whose scores on the neuter nouns of the production task were poor, behaved differently from the adults and the targetlike production group. These children were not able to use the gender cue anticipatorily, that is, they did not make use of gender-marking before the onset of the noun. However, they were able to use the gender cue in a facilitative way, just as the adults and the targetlike production group did: around 600 ms after article onset, the non-targetlike production group benefitted from the informativeness of the gender cue.
Our first research question, whether 4-7 year old monolingual children can use gender during online sentence comprehension, can thus be answered positively. The results revealed that Dutch-speaking children (and adults) are able to make use of the gender cue during online processing. This is striking as the evidence available in the input to a Dutch-learning children that Dutch has grammatical gender, and more specifically, how this is marked on neuter nouns, is rather limited (Unsworth et al., 2014). These results thus indicate that in an opaque system like Dutch, learning to use gender-marking during comprehension may not necessarily take as long as previous research, based on production data only, has suggested.

Our second research question focused on whether there is a relation between production and online processing of gender. This question gives us insight into whether poor gender producers are nevertheless able to process gender cues during online sentence processing. Our data revealed that children who produced targetlike articles with neuter nouns at a rate of 63% or less were able to make use of the gender cue facilitatively but not anticipatorily. How to interpret this discrepancy? If one considers facilitation as a successful way of using gender during processing, one could argue that the previous production data on the monolingual acquisition of Dutch gender may have underestimated the knowledge of these children. In other words, as for many other aspects of language, targetlike comprehension precedes targetlike production abilities. However, one could also argue that these children’s online comprehension abilities are not yet completely targetlike due to the lack of anticipation skills in this group. In either case, our results extend the previous offline comprehension data (Unsworth & Hulk, 2010; Unsworth, 2013) by showing that children as young as 4 years old are able to make (at least some) use of gender-marking.
Our findings show that children are able to use gender facilitatively before they can actively and correctly produce it (the non-targetlike production group). However, once they have mastered this knowledge (targetlike production group), they can also use gender anticipatorily. These findings give us insight into the developmental path via which Dutch gender is acquired. They reveal that these children do not simply have a production-specific problem, as has been suggested for similar data with bilingual and second language learners (Blom & Vasic, 2011; Unsworth, 2008; White, Valenzuela, Kozlowska-Macgregor, & Leung, 2004). Rather, the interplay between the development of targetlike comprehension and targetlike production is more complex. More specifically, it seems that whilst targetlike online comprehension may in part be possible before targetlike production is completely in place, targetlike production may be a trigger for online comprehension to be completely successful. Longitudinal data would be necessary to test this claim, but this account is in line with what has been claimed for the relationship between production and online processing of grammatical gender in (L2) adults). More specifically, the quality of lexical representation (Perfetti, 2007; Perfetti & Hart, 2001, qtd in Kaan, 2014), which includes gender knowledge, has been linked to L2 adults’ ability to use gender-marking predictively in online processing (Hopp, 2013). Our data suggest that the same may hold for (monolingual) children. Whilst a detailed account of the relationship between comprehension and production and how this develops is beyond the scope of the present paper, it is worth noting that current models in this area also relate abilities in comprehension to prior experience, both in terms of linguistic input and children’s own language production (e.g., Chang, Dell, & Bock, 2006; MacDonald, 2013).
Another interesting finding, which we did not anticipate at the outset of this experiment, is the asymmetry in eye gaze behavior on common versus neuter nouns. This pattern occurred in the facilitative time window for all our groups. In particular, we found that the neuter article *het* was more informative than the common article *de*. As discussed in the Introduction, there are important differences between the distribution and use of *de* and *het* which may explain this result. Perhaps neuter gender is a more informative cue because it arguably has fewer form-to-meaning mappings than common gender (but see Roodenburg & Hulk, 2008). Also, the definite article *de*, used with singular common nouns, is also used for neuter nouns in the plural, making it a less reliable cue. This means that upon hearing the definite article *de*, the listener will – all things being equal – not yet know whether the upcoming noun will be singular or plural. Even in a context in which only single items are displayed, this feature may discourage the active use of *de* as a cue. However, it is also possible to argue that *het* is anticipatorily less useful, given that *het* can be used to precede diminutives, or as a personal or indefinite pronoun (Loerts et al., 2014; Roodenburg & Hulk, 2008). A final explanation for the (late) interaction between condition by gender in the targetlike production group has been given by an anonymous reviewer who suggested that there is a higher proportion of target fixations on same-gender trials in the common gender trials, but a higher proportion of target fixations on different-gender trials in the neuter gender trials starting approximately 1000 ms after target word onset.

Note that the difference between *de* and *het* that we observe is not consistent with Johnson’s (2005) study, which found that toddlers were only sensitive to common gender-marking. This is likely due to the difference in age of the children between studies. Johnson’s study involved much younger children (28 months old) who routinely substitute
de for het (overgeneralization). Common de might thus be easier to use or recognize than neuter het at that age. We therefore believe that children might first go through a stage in which they can only use de as a gender cue (see e.g., Cornips & Hulk, 2008; Unsworth, 2008). However, once they get older and start to make use of het in their production, producing fewer overgeneralisations, they are also able to make use of het as a gender cue in their online language processing. First, children can only use the neuter gender cue in a facilitative way, and later, when overgeneralizations in their production hardly occur and production is close to perfect, they are able to use the neuter gender cue anticipatorily.

Our results go against Loerts et al.’s (2014) findings. In their study on gender processing in monolingual Dutch adults, these authors found that only common gender-marking and not neuter gender-marking affected the processing of the subsequent noun. As outlined in the Introduction, the set-up of their study was more complex than ours because it contained four items in the visual array. We argue, however, that the discrepancy between the two sets of results is not simply an effect of a less/more complex display: the number of additional factors involved in their study (use of definites, indefinites, colour) make it difficult to explain their findings, as the authors themselves acknowledge. It thus remains unclear whether their set-up was truly measuring gender effects, making it hard to compare the results of their study with ours.

In conclusion, this study has shown that Dutch adults and 4-to 7-year old children are able to use gender-marking during online sentence comprehension. Importantly, children use gender facilitatively before they can successfully produce it. However, once production is at targetlike level, they can also use gender-marking anticipatorily. Targetlike production thus seems to function as a trigger for online comprehension to be successful.
Footnotes

1. A number of explanations have been given as to why gender is acquired comparatively late in Dutch, including frequency, saliency and transparency (see Polišenskà, 2010 for detailed discussion).

2. Note that our facilitation analysis differs in the investigated time window from Lew-Williams and Fernald’s (2007) analysis. They examined eye gaze behavior from article onset until noun offset, whereas we looked from noun onset. However, as their dependent variable is also different from ours (i.e., they only included trials on which participants were looking at the distractor at article onset and shifted to the target picture by the end of the noun, thereby excluding many trials), it is anyway hard to be completely comparable.
Acknowledgements

We would like to thank Casey Lew-Williams for sharing the experimental set-up of Lew-Williams and Fernald (2007). We would also like to thank Elleke Baakman, Laurette Gerts, Jidde Jacobi, Vera Nees, Nathalie Urbanus, Angela van de Weg, and Rosanne Westra for their assistance with running this experiment. Furthermore, we like to thank Natalie van Eerden for lending her voice for the recordings. Finally, we would like to acknowledge our discussions with Pim Mak at an earlier stage of this project.
References


