Acta Psychologica 135 (2010) 150–158

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Native language effects in learning second-language grammatical gender: A training study

Kristin Lemhöfer *, Herbert Schriefers, Iris Hanique

Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 22 February 2010
Received in revised form 17 May 2010
Accepted 2 June 2010

PsyclNFO classification:
2720
2343

Keywords:
Second language acquisition
Grammatical gender
Cross-language effects
Cognates
Feedback

ABSTRACT

We investigated cross-language influences in the representation and acquisition of Dutch word gender by native speakers of German. Participants named pictures in Dutch, using gender-marked noun phrases, and were trained on this task using feedback. Nouns differed in gender compatibility and cognate status with respect to German. The results show clear effects of cross-language gender compatibility and cognate status on response accuracy, certainty, and consistency. Feedback during training reduced gender errors approximately by half, and affected the different item conditions similarly. Furthermore, relative to the initial error rates, incorrect gender responses given with great certainty were not harder to modify than those with lower certainty. The results provide insights into the nature and stability of correct and incorrect gender representations in L2, and demonstrate the pervasiveness of transfer from the first to the second language even after intensive training.

1. Introduction

"Learning Dutch is easy", is the general opinion among Germans. After only five weeks of intensive training, most German prospective students at Radboud University Nijmegen pass the required state examination of Dutch as a second language. The high degree of relatedness of the two languages indeed facilitates at least the first stages of second language acquisition. However, this relatedness also has its pitfalls, as it might also give rise to many incorrect instances of transfer from the first (L1) to the second language (L2). One example is the transfer of word gender from German to Dutch: though a majority of words, especially cognates (i.e., form-similar translation pairs), are compatible in gender between Dutch and German, there are also many that are not. Basing Dutch gender on German gender will thus cause systematic gender errors in Dutch language production.

Strangely enough, the issue of grammatical gender acquisition in L2 has not often been looked at from the perspective of gender compatibility between L1 and L2. In general, there is ample evidence in the literature that the acquisition of word gender in L2 is extremely difficult, and that only few (if any) adult learners of L2 ever reach native-like competence in this domain (Dewaele & Véronique, 2001; Holmes & Dejean de la Bâtie, 1999; Unsworth, 2008). This holds especially for languages like Dutch that have an 'arbitrary' gender system, i.e., without a close relation between the phonological form of a noun and its gender (cf. Kempe & Brooks, 2008, for experimental evidence on less successful gender acquisition where such a relation is missing). However, the reasons for the great difficulties of adult learners to acquire the gender system of their L2, as well as the nature of gender representations in the L2 lexicon, remain poorly understood. A number of studies have investigated the general role of the L1 gender system, i.e., whether L1 has grammatical gender at all, and how similar it is to that of L2, with some conflicting results (Franceschina, 2005; Sabourin, Stowe, & de Haan, 2006; White, Valenzuela, Kozlowska-MacGregor, & Leung, 2004). Only recently have some studies looked into the role of cross-language influences at the word level, however also with contradictory results: Costa, Kovacic, Franck, and Caramazza (2003) did not find any evidence for effects of cross-language gender compatibility in picture naming by Croatian-Italian, Catalan-Spanish or Italian-French bilinguals. However, Salamoura and Williams (2007) found that translation times for gender-marked adjective-noun phrases from Greek (L1) to German (L2) were longer when the respective noun was gender-incompatible between the two languages, both for cognates and for non-cognates.

In the present study, we will look at cross-language influences with respect to word gender in German learners of Dutch. German has three gender classes (masculine, feminine, and neuter), which are, among others, marked by the definite determiner (der_{masc}, die_{fem}, and das_{neu}). In Dutch, masculine and feminine gender have practically collapsed into

* Corresponding author. Donders Institute for Brain, Cognition, and Behaviour—Centre for Cognition, Radboud Universiteit Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands. Tel.: +31 24 3612630; fax: +31 24 3616066.
E-mail address: k.lemhofer@donders.ru.nl (K. Lemhöfer).

0001-6918/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2010.06.001

one common gender (Klooster, 2001), which, together with neuter gender, turns Dutch into a two-gender system. Like in German, the definite determiner in Dutch is gender-marked (de_{com} and het_{neu}). As mentioned before, many, if not most words are 'compatible' in gender between Dutch and German. By gender compatibility, we refer to a mapping of German masculine and feminine gender onto Dutch common gender, and of German neuter gender onto Dutch neuter gender.

In an earlier study (Lemhöfer, Spalek, & Schriefers, 2008), we already showed that German–Dutch bilinguals are strongly influenced by this cross-language gender compatibility when processing gender in their L2 Dutch, both in language comprehension (gender-primed lexical decision) and production (gender-marked picture naming). Dutch determiner-noun phrases with a gender-incompatible German translation equivalent gave rise to higher error rates and longer reaction times (RTs) than gender-compatible nouns. Furthermore, this held primarily for form-similar translation pairs (cognates), like *hond* (Dutch)—*Hund* (German), meaning 'dog'.

In that study, we also showed that cross-language compatibility effects are larger for relatively unstable gender representations, as measured by the consistency of responses across several item repetitions. Such a lack of stability must be a result of imperfect L2 gender acquisition. Thus, the *online* competition processes that we were originally looking for in the Lemhöfer et al. study were not, or only partly, responsible for the resulting gender compatibility effects; rather, the effects seem to predominately originate from the *acquisition* stage. Therefore, the L2 gender acquisition process has to be further investigated to better understand the factors that lead to incorrect and/or unstable gender representations. In particular, in the present article we address the question how correct and how stable L2 gender representations are for nouns that differ in cross-language form similarity and gender-compatibility. To this end, we will use the same materials as Lemhöfer et al. (2008). Because the stability of gender representations is of course not directly observable, we will use two indicators that should have a close relation to the stability concept: First, the consistency of responses across several item repetitions, as it was already used to reflect stability in the Lemhöfer et al. study, and second, participants' ratings indicating the certainty of their naming response. If these two measures indeed both reflect stability, they should be closely related to one another, and furthermore be similarly influenced by cognate status and Dutch-German gender-compatibility.

Additionally, as a natural further step, we intend to examine the possibility of giving correct L2 gender acquisition a helping hand by providing some sort of training. Apparently, given the high error rates in L2 gender production, just passively receiving correct input from the L2 environment is not sufficient for changing incorrect gender representations; it is possible that under normal circumstances, learners do not notice the discrepancy between their own incorrect gender representation and the L2 input (Schmidt, 1990). Thus, providing *explicit* feedback on incorrectly produced gender markings (in the present study, definite determiners) might already be sufficient to correct these gender errors. We will examine the general efficiency of such feedback on L2 gender production, and whether and how this effect depends on L1–L2 gender compatibility, cognate status, and on the certainty of an initial naming response. One plausible hypothesis is that incorrect gender representations of highly difficult words (e.g., gender-incompatible cognates) are harder to correct via feedback than those of easy nouns (like compatible cognates). In addition, one could expect that incorrect gender representations of which the speaker is very certain will be more 'stubborn' and harder to modify than those of which the speaker is uncertain.

Within the field of (adult) second language acquisition, most studies that have addressed the feedback issue compared different methods of feedback and input with each other (e.g., Ayoun, 2001; Carroll, Swain, & Roberge, 1992; Ellis, Loewen, & Erlam, 2006; Lyster, 2004). However, none of these studies addressed the acquisition of word gender. In this study, we will employ the probably most straightforward and simple feedback variant—simple error correction—as a starting point to examine the effect of training and the course of learning with respect to L2 word gender acquisition.

To summarize, in this study we pursue two general goals: First, to get to know more about the stability of gender representations for Dutch nouns that differ in gender-compatibility and form similarity with the L1 translation; and second, we will examine the effect of repeated feedback on the accuracy of the production of gender-marked noun phrases. We will approach these questions by measuring accuracy as well as subjective certainty (by way of ratings) for determiner-noun phrases at the beginning of the experiment, and by conducting a training session of three blocks with feedback.

2. Methods

2.1. Participants

The data reported here were collected in the context of Experiment 3 in Lemhöfer et al. (2008). The group of participants as well as the stimulus materials is thus identical to those reported in that study.

The participants were 24 native speakers of German with Dutch as a second language, most of them students at Radboud University Nijmegen. They had all normal or corrected-to-normal vision, were non-dyslectic, and had German as their only mother tongue. All participants were right-handed. They were between 19 and 41 years old (mean 24.3), 21 were female, three male. They had lived in the Netherlands between 6 months and 9.5 years (mean 2.7 years), with between 6 months and 23 years of experience with Dutch (mean 4.1 years). A language questionnaire was completed by all participants to provide more information on their language background, including their self-rated amount of experience with Dutch and their Dutch language skills. The results are summarized in Table 1. The participants also used other foreign languages than Dutch regularly, in particular English (14 participants). All participants except for one stated that Dutch was currently their most frequently used foreign language. The participants also carried out a Dutch vocabulary test, which is described in more detail in Lemhöfer et al. (2008). This test was a non-speeded Dutch lexical decision task on a high level of difficulty, including 40 very low-frequent words as well as 20 highly word-like nonwords. The average score on the test for this group of participants was 72%, calculated by averaging % correct values of words and nonwords (minimum: 52%, maximum: 88%, standard deviation: 9).

2.2. Stimulus materials

The stimulus materials were identical to those used in Experiments 2 and 3 of Lemhöfer et al. (2008), and are listed in the Appendix. Four item conditions were formed by fully crossing the two two-level factors Cognate Status (cognates vs. non-cognates) and Gender Compatibility (gender-compatible vs. incompatible between German and Dutch). For each condition, we selected 24 nouns depicted by black-on-white line drawings to be named by the participants. Twelve of these nouns had common gender, the other 12 neuter gender in Dutch. The four item conditions were matched for

Table 1
Results of the Dutch experience and proficiency ratings of the bilingual participants.

	Mean	SD
Frequency of reading literature in Dutch	5.0	1.5
Frequency of speaking Dutch	5.3	1.4
Self-rated reading experience in Dutch	4.8	1.4
Self-rated writing experience in Dutch	4.4	1.4
Self-rated speaking experience in Dutch	4.8	1.6

Note. SD = Standard Deviation. Ratings were given on a scale from 1 (low) to 7 (high).

Table 2
Characteristics of the word materials for each of the four item conditions.

Item condition	Example (German and English translation)	Mean no. of letters	Mean no. of syllables	Mean Dutch log freq.
Cognates, gender-compatible	$hond_{com}$ ($Hund_{masc}$, dog)	5.1 (1.2)	1.67 (0.6)	1.30 (0.5)
Cognates, gender-incompatible	$auto_{com}$ ($Auto_{neu}$, car)	5.4 (1.5)	1.83 (0.7)	1.27 (0.5)
Non-cognates, gender-compatible	$vork_{com}$ ($Gabel_{fem}$, fork)	5.5 (2.0)	1.50 (0.7)	1.34 (0.4)
Non-cognates, gender-incompatible	$jurk_{com}$ ($Kleid_{neu}$, dress)	5.4 (1.6)	1.67 (0.8)	1.28 (0.3)
Total mean		5.4 (1.6)	1.67 (0.7)	1.30 (0.4)

Note. Standard deviations are given in parentheses; freq. = frequency according to the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995).

Table 3
Means of % correct and certainty ratings (1–5) for the four item conditions.

Item condition	% Correct	Certainty
Cognates, gender-compatible	92 (27)	3.81 (1.10)
Cognates, gender-incompatible	30 (46)	3.59 (1.11)
Non-cognates, gender-compatible	84 (37)	3.82 (1.20)
Non-cognates, gender-incompatible	55 (50)	3.53 (1.22)
Total mean	65 (48)	3.69 (1.17)

Note. Standard deviations are given in parentheses.

word length and Dutch frequency. Item characteristics and example items are given in Table 2. Twenty-four additional pictures of the same kind were used as training and warming-up items.

2.3. Procedure

Participants were tested individually. They were seated approximately 70 cm away from a 17″ VGA monitor and separated from the experimenter by a partition wall. The stimuli were presented on the monitor at a resolution of 1024 by 768 pixels. The presentation of the stimuli was controlled by NESU software developed by the Max Planck Institute for Psycholinguistics, running on an Intel Pentium 4 computer. Participants' responses were recorded with a DAT recorder. Each experimental session lasted between 75 and 90 min in total. The experiment was preceded by the Dutch vocabulary test. At the end of the session, the participants completed the language background questionnaire.

For the *familiarization phase* of the experiment, participants received a booklet with all experimental pictures and their names, but without determiners. They were asked to write down the correct singular definite determiner (*de* or *het*) and to indicate the certainty of their answer on a scale from one to five (1 = very uncertain, 5 = very certain). Additionally, participants were asked to memorize the object names given in the booklet, and to use only those names during the remaining part of the experiment. Thus, this phase served both to familiarize participants with the pictures and their names, as well as to assess their a-priori knowledge of the nouns' gender and the certainty of that knowledge.

The subsequent *training phase* consisted of three training blocks. During each block, all pictures were presented one by one on the monitor and had to be named by the participant together with their singular definite determiner (e.g. de_{com} *hond*—'the dog'). During the first two blocks, the experimenter provided feedback on the correctness of each response. In case of an incorrect response, the experimenter provided the correct phrase. A different randomization of items was used in each block, with the restriction that semantically or phonologically similar items did not occur directly after each other.

Table 4
Results of the ANOVAs (across participants and items) of the familiarization results.

Effect		F_1 (1,23)	p (F_1)	F_2 (1,92)	p (F_2)
Cognate status	% correct	47.22	<0.001	4.31	0.04
	certainty	0.38	0.55	0.06	0.81
Gender compatibility	% correct	146.33	<0.001	129.13	<0.001
	certainty	46.00	<0.001	7.28	0.008
Cognate status × Gender compatibility	% correct	115.40	<0.001	16.71	<0.001
	certainty	0.91	0.35	0.13	0.72

3. Results

3.1. Familiarization phase: Accuracy and certainty

In the first phase of the experiment, participants wrote the gender-marked definite determiners in front of the given nouns and rated the certainty of their response. Table 3 shows the mean percentages of correct answers, as well as the mean certainty scores for the four item conditions.

As can be seen from the table, both response accuracy and certainty were influenced by cognate status and cross-language gender compatibility: Percentages of correct responses were highest for gender-compatible cognates, lowest for incompatible cognates, and intermediate for non-cognates (with fewer correct responses on gender-incompatible non-cognates than on gender-compatible ones). In contrast, subjective certainty (regardless of accuracy), which we regard as an indicator of the stability of gender representations, was influenced only by gender compatibility (with lower certainty ratings for incompatible relative to compatible items), but not by cognate status. ANOVA's[1] performed across participants and items supported this descriptive pattern: While the effects of Cognate Status, Gender Compatibility, and their interaction were all significant for % correct, only the Gender Compatibility effect was significant for the certainty ratings (see Table 4). Pairwise comparisons with one-tailed t-tests confirmed that all four item conditions differed from each other with respect to % correct (all $p<0.03$). However, in terms of certainty ratings, only gender-compatible and incompatible items differed from each other (for both cognates and non-cognates, all $p<0.05$), but within the compatibility groups, cognates and non-cognates were statistically indistinguishable (all $p>0.16$). Thus, L2 learners do not only make more gender errors when producing nouns with an L1-incompatible gender in L2 (and even more so for cognates), but they are also subjectively less certain about the gender of these nouns. Form similarity of the translation equivalents only additionally influenced the percentage of correct responses, but not subjective certainty.

3.2. Certainty categories

In order to capture both the certainty and the correctness of a response in one measure, we created a combined scale of 'representational strength', with very certain, but incorrect representations at one end, uncertain ones in the middle, and very certain correct responses at the other end. The 'certain and correct' end of the scale can be regarded as the end where native speakers' gender representations lie. Fig. 1 shows how the original certainty scores and correctness of the response were combined into one measure.

[1] Given that the ratings are ordinal data, non-parametric tests (Wilcoxon Matched-Pairs tests on the medians) were also carried out, which yielded highly similar results as the ANOVA.

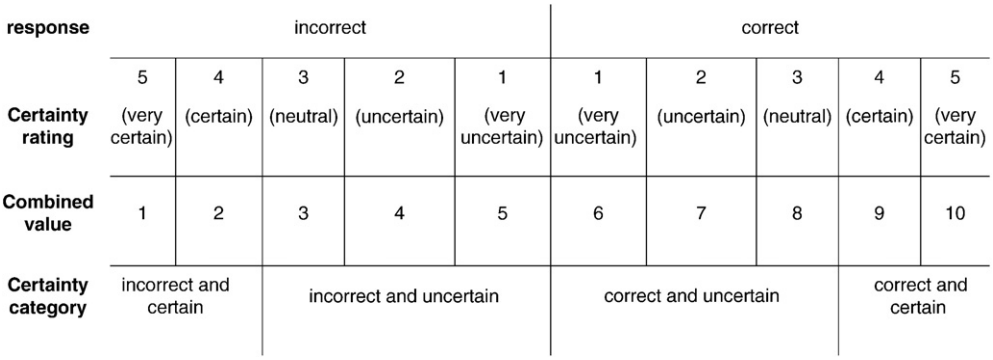

Fig. 1. Combination of certainty scores and correctness of the response into one measure.

Table 5
Distribution of items across the four certainty categories.

Certainty category	Incorrect and certain	Incorrect and uncertain	Correct and uncertain	Correct and certain
% of compatible cognates	1.0	6.6	26.2	**66.1**
% of incompatible cognates	**41.7**	28.6	12.7	17.0
% of compatible non-cognates	8.2	7.6	27.4	**56.8**
% of incompatible non-cognates	23.8	21.5	20.1	**34.5**
% of total	18.7	16.1	21.6	**43.6**

Note. Bold numbers indicate the certainty category that contains the highest percentage of items.

We divided the combined scale into four segments (see Fig. 1), forming the categories 'incorrect and certain' (combined values 1 and 2), 'incorrect and uncertain' (values 3, 4, and 5), 'correct and uncertain' (values 6, 7, and 8), and 'correct and certain' (values 9 and 10). Table 5 shows the percentages of items (per item condition and overall) that fall into these categories.

As can be seen from Table 5, the distribution of data across the four certainty categories differs a lot for the four item conditions. For instance, while an average of 66% of the compatible cognates fall within the 'correct & certain' category and only 1 % in the 'incorrect & certain' category, the pattern is reversed for the gender-incompatible cognates (42% incorrect—certain, 17% correct—certain). We will use the certainty categories again when looking at the course of the training.

3.3. Response consistency before the first feedback

Besides the participants' performance at the single moment of familiarization, we examined to what degree participants produced identical or variable gender assignments for a given item across two repetitions, depending on item condition. Recall that we assumed that besides the participants' certainty ratings, response consistency represents a second indicator of representational stability. To measure response consistency, we used the response agreement between the very first gender assignment during familiarization and the one during the first block of training (i.e., before the participant had received any feedback). Two different responses to a given item by a given participant in these two blocks would point to a lack of stability of the item's gender representation. Table 6 shows the mean percentages of consistently correct and incorrect responses for the four item conditions.

Table 6
Mean percentages (across participants) of consistent responses across familiarization and the first block of training.

Item condition	% consistent when correct in familiarization	% consistent when incorrect in familiarization	% consistent total[a]
Cognates, gender-compatible	93.7 (5.8)	33.8 (41.1)	89.9 (8.3)
Cognates, gender-incompatible	63.4 (24.7)	81.2 (11.2)	78.3 (10.0)
Non-cognates, gender-compatible	91.0 (6.5)	72.3 (12.4)	87.5 (7.3)
Non-cognates, gender-incompatible	82.0 (12.6)	71.3 (12.4)	77.8 (10.6)
Total mean	82.5 (18.6)	66.4 (30.2)	83.4 (10.5)

Note. Standard deviations are given in parentheses.
[a] Note that this column is not an average of the previous two, given that the proportions of correct and incorrect responses were not equal.

As evident in Table 6, the percentage of consistently correct responses is influenced by both gender compatibility and cognate status in the same way as the error rates in the familiarization phase: When considering the initially correct responses only, the chance of an item being also correct in the next repetition (first training block) was highest for gender-compatible cognates, followed by compatible non-cognates and gender-incompatible non-cognates, and it was lowest for gender-incompatible cognates (all pairwise comparisons by means of *t*-tests[2] were significant with $p<0.01$ apart from that of compatible cognates and compatible non-cognates, $t(23)=1.83$, $p=0.08$).

This pattern was reversed when looking at the initially incorrect responses: When an item of the 'easiest' condition, gender-compatible cognates, was responded to incorrectly during familiarization, the chance of it being incorrect again in block 1 of training was only 34%, while this chance was 81% for the most difficult condition (incompatible cognates), with the non-cognates in between. Paired *t*-tests showed that statistically, the value for gender-compatible cognates was different from all others ($p<0.001$), while the other three values did not differ from each other (all $p>0.14$).

Overall, the data presented up to now show that the four item conditions give rise to different degrees of difficulty with gender assignment, with the least difficulties for gender-compatible cognates, most difficulties for gender-incompatible cognates, and the two non-cognate conditions in between. These differences in gender processing were reflected in error rates, certainty ratings, and the level of consistency across two item repetitions.

Given that we introduced both the consistency of responses as well as the certainty ratings from the familiarization phase as indicators of representational stability, we also investigated in how far these two measures were related (which should be the case when they reflect the same underlying concept). To this means, we again collapsed the five certainty rating points (but regardless of accuracy) into two rating categories "uncertain" (rating responses 1 to 3) and

[2] Because the categorization into consistent and inconsistent responses was variable across participants for a given items, these analyses were calculated across participants only.

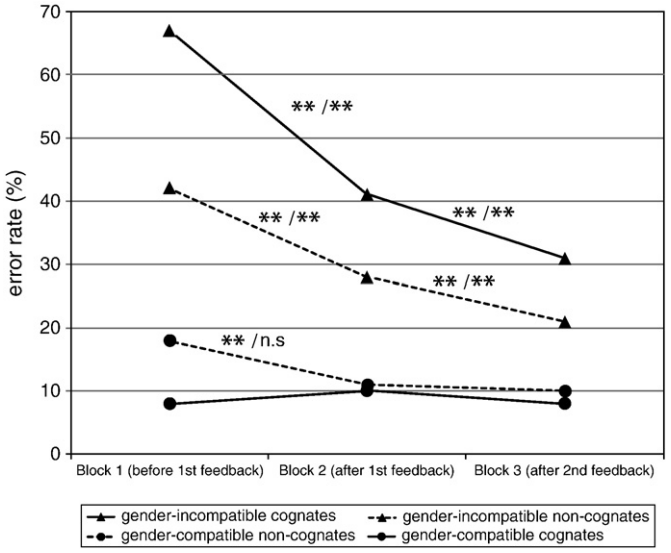

Fig. 2. Error rates across the three training blocks for the four item conditions. *Note.* The asterisks indicate the significance (across participants / across items) of the error rate change between two consecutive blocks, with *: $p<0.05$, **: $p<0.01$, n.s. = non-significant. Contrasts that did not reach significance at all are unmarked.

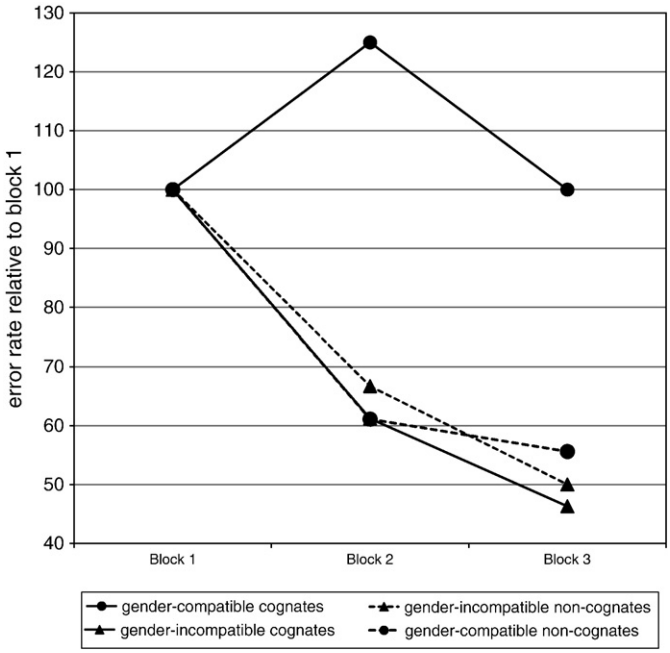

Fig. 3. Course of error rates normalized for the error rate in block 1 (= 100%) for the four item conditions.

"certain" (responses 4 and 5), and calculated the percentage of consistent responses for each participant and rating category. Means calculated across participants show that "uncertain" responses were consistent across the two repetitions in 71% of the cases, compared to 90% for "certain" responses (recall that the chance baseline for consistency is 50%). This difference was highly significant in a repeated-measures one-factorial ANOVA ($F(1, 23) = 71.34$, $p<.001$). The estimation of the proportion of variance in the consistency rates explained by the certainty category (eta-squared) was 0.76. In other words, the rated certainty and the consistency of responses shared 76% of their variance, indicating that they are highly related.

3.4. Training phase: Item conditions

In the second phase of the experiment, participants were provided with feedback on their production of gender-marked noun phrases (definite determiner and noun) after each trial during the first two of three training blocks, with each block comprising the complete item set. The question was whether participants' performance would benefit from the feedback at all, and if so, whether this benefit would differ for gender-compatible and incompatible cognates and non-cognates.

Fig. 2 shows the error rates for the three blocks during training for the four item conditions. As can be seen from the figure, the error rates indeed dropped with each proceeding block for each item condition apart from gender-compatible cognates, which remained at an equally low error rate level (block 1: 8%, block 2: 10%, block 3: 8%). The overall error rate decreased from 34% in block 1 to 22% in block 2 and 17% in block 3.

A repeated-measures ANOVA on the error rates was conducted with the factors Block (3 levels), Cognate Status (2) and Gender Compatibility (2). The results can be found in Table 7. Most crucially, the triple interaction of all three factors was significant, indicating that the slope of the curves is different for the different item conditions. Therefore, the Block effect was analyzed for the four item conditions separately, showing that it was significant for all conditions (all $p<0.001$ across participants, all $p<0.02$ across items) except gender-compatible cognates (both $p>0.28$). Paired contrasts were carried out to further examine the drops in error rates from block 1 to block 2, and from blocks 2 to 3. The asterisks in Fig. 2 indicate which error rate changes were significant. They show that for both gender-incompatible item conditions (cognates and non-cognates), error rates dropped significantly after each block. For gender-compatible non-cognates, only the first feedback was effective in reducing error rates, and finally, for the 'easiest' condition with the least errors, gender-compatible cognates, there was no significant learning effect at all.

In the analyses presented so far, the decrease in error rate may be related to the size of the error rate in the first block: When more errors are made before the training, there is more room for improvement, thus a potentially steeper drop in error rates. In our data, the amount of reduction of error rates in blocks 2 and 3 indeed follows the ordering of error rates in the first block. Therefore, we also applied a normalization procedure in order to examine whether the effect of learning was truly different for the different item groups, regardless of their initial error rates. To this end, we used the error rate in the first block as a reference for each participant (i.e., we set it at 100%) and calculated the error data of blocks 2 and 3 relative to this baseline. For

Table 7
Results of the ANOVAs (across participants) of the error rates across training blocks.

Effect	F_1	df(F_1)	p (F_1)	F_2	df(F_2)	p (F_2)
Block	58.05	2, 46	<0.001	79.85	2, 184	<0.001
Cognate status	24.04	1, 23	<0.001	4.19	1, 92	<0.05
Gender compatibility	239.08	1, 23	<0.001	94.53	1, 92	<0.001
Training block × Cognate status	1.85	2, 46	0.17	0.83	2, 184	0.44
Training block × Gender compatibility	41.84	2, 46	<0.001	43.02	2, 184	<0.001
Cognate status × Gender compatibility	52.63	1, 23	<0.001	13.19	1, 92	<0.001
Training block × Cognate status × Gender Compatibility	14.46	2, 46	<0.001	10.35	2, 184	<0.001

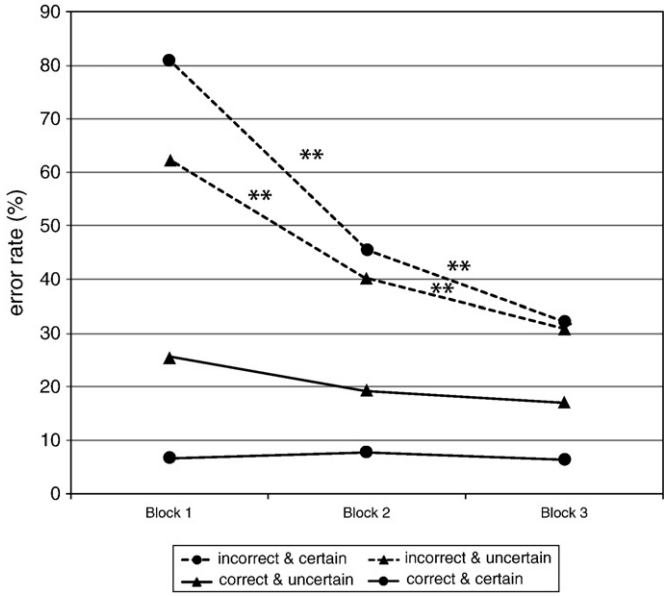

Fig. 4. Course of error rates across the three training blocks for the four Certainty Categories. *Note.* The certainty categories (correctness and certainty) are based on the familiarization phase *before* the start of training. Asterisks indicate the significance of the error rate change between two consecutive blocks, with *: $p<0.05$, **: $p<0.01$. Contrasts that did not reach significance at all are unmarked.

example, a normalized value of 50% in block 2 would mean that the error rate for a given item condition in block 2 was reduced by half compared to block 1.

Fig. 3 shows the data that result from this normalization procedure. The condition that stands out immediately are the compatible cognates, for which there was a rise in error rates rather than a drop from block 1 to block 2 (and then a drop again to block 3). However, because the error rates in this condition were very low in the first place, this change was only small in absolute terms (and not significant, as reported above), but looks large when expressed in percentages.

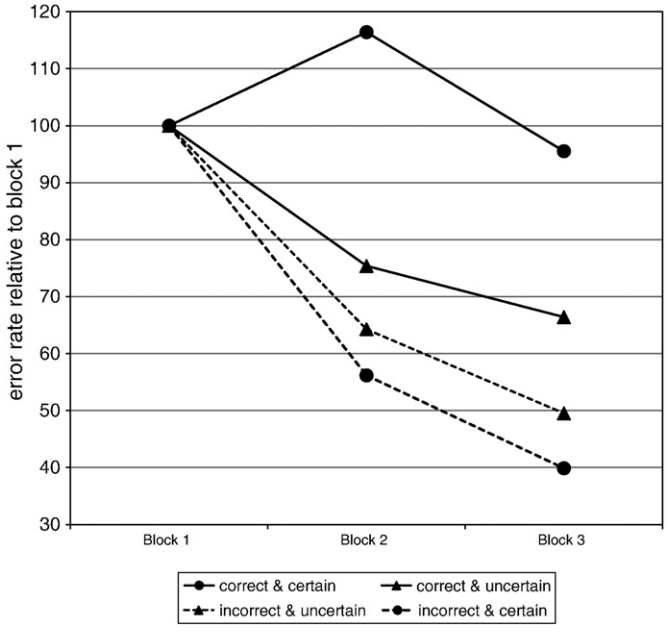

Fig. 5. Course of error rates normalized for the error rate in block 1 (=100%) for the four Certainty Categories.

After applying this normalization procedure to every participant and item condition, an ANOVA across participants was carried out on the data of blocks 2 and 3[3] in order to assess whether the learning effect was different for the different item conditions. Because of the exceptional status of the compatible cognate condition and because the earlier analyses (see Fig. 2) had already shown that the learning (i.e., block) effect was not significant for this condition, we included only the remaining three conditions in the analysis, using a 3-level factor 'Item Condition'. This analysis showed a main effect of the 2-level factor Block ($F(1,23)=7.39$, $p=0.01$), again indicating that there was a significant improvement from blocks 2 to 3. However, neither the main effect of Item Condition ($F<1$), nor the interaction of Block and Item Condition ($F(2,46)=1.47$, $p=0.24$) were significant. Thus, when error rates in blocks 2 and 3 are expressed as percentages of the error rate in block 1, there is no difference in learning between the conditions (except for compatible cognates).

To summarize the training data on the different item categories, the analyses showed that performance for all item conditions except that for compatible cognates improved significantly through training. In absolute terms, the more difficult item categories (gender-incompatible cognates and non-cognates) improved the most. However, relative to the error rate in the first training block, the improvement of the three 'difficult' categories was identical.

3.5. Training phase: Certainty categories

A different way to look at the data is to divide the items according to their 'representational strength' in the familiarization phase, rather than using distinctions based on cognate status and gender compatibility. We did that already when forming correctness and certainty response categories as in Fig. 1 and Table 5. Now, the question is whether (initially) correct and incorrect items, and items that were initially responded to with high or low certainty, behaved differently in the course of training.

Given the extremely unbalanced distribution of items across the 16 cells in Table 5, analyses that simultaneously take both classifications (item condition and certainty category) into account are not possible. Therefore, in the following, we will drop the original categorization of items according to cognate status and gender compatibility, and examine the learning data based on their certainty category only.

Fig. 4 shows the error rates for each certainty category across the three training blocks. Keep in mind that certainty category refers to the correctness and subjective certainty of the response in the *familiarization phase*, i.e. before block 1 of training. Thus, it is possible that an item within the 'correct—certain' or 'correct—uncertain' category was responded to *incorrectly* in training block 1.

We analyzed the data using a 4 (Certainty Category) by 3 (Block) repeated measures ANOVA. This analysis was calculated across participants only, because each item could fall into different certainty categories for different participants. The ANOVA showed significant effects of Certainty Category ($F(3,69)=103.7$, $p<0.001$), Block ($F(2,46)=61.91$, $p<0.001$), and a significant interaction of the two ($F(6,138)=21.58$, $p<0.001$). Again, the interaction was further examined by testing the Block effect for each certainty category separately. The results show that only words that had been responded to incorrectly in the familiarization phase showed a significant learning effect (i.e., block effect) during the subsequent training (both $p<0.001$), while the items that had been correct during familiarization (but might have been incorrect in the first training block) did not show such an effect (both $p>0.15$). Again, the significances of the pairwise contrasts between blocks 1 and 2, and 2 and 3, are shown as asterisks in Fig. 4.

[3] Block 1 cannot be included because it has a variance of 0.

Again, as can be seen from Fig. 4, the items with the highest error rates in block 1 show the steepest drop in error rate, just as observed before for the incompatible cognates. Thus, the data were normalized again, by setting the error rate value of block 1 to 100% and expressing the error rates in blocks 2 and 3 as a percentage of this baseline. Fig. 5 shows the normalized values for the three blocks.

Similarly to the compatible cognate condition before, the certainty category with the lowest error rates in block 1 (correct and certain) stands out because of a rise in error rates in block 2 rather than a drop. Because the previous analysis (see above and Fig. 4) already showed that the two 'correct' categories did not show significant learning, we included only the two 'incorrect' categories in the analysis of the normalized values (and only blocks 2 and 3, as before). Here, neither the effect of Certainty Category ($F(1,23) = 2.67$, $p = 0.12$) nor the interaction of Certainty Category and Block was significant ($F < 1$), indicating that the reduction of errors relative to block 1 was identical for the two 'incorrect' categories. However, there was a main effect of Block ($F(1,23) = 20.35$, $p < 0.001$), confirming again that there was significant improvement between blocks 2 and 3.

These results of the training data for the certainty categories show that significant learning took place after both instances of feedback, but this learning effect occurred only for nouns that had been responded to incorrectly in the familiarization phase before the training. The certainty with which this incorrect response was given did not further modulate the learning effect.

4. Discussion

The current study aimed to investigate the correctness, stability, and modifiability of Dutch gender representations in German learners of Dutch, using nouns that differ in cognate status and compatibility with the German gender. To our knowledge, it is the first study that combines the issue of L1–L2 transfer (in this case, transfer of L1 word gender to L2 translation equivalents) with that of learnability and the flexibility of representations.

We will first have a look at performance levels before the training phase, before discussing the effect of feedback during training.

4.1. Familiarization phase: Accuracy and certainty

The performance data from the first encounter of our participants with the experimental materials during the familiarization phase show that the correctness and certainty of gender representations depended on compatibility with the L1 gender and cognate status. This is in line with the results by Lemhöfer et al. (2008) and Salamoura and Williams (2007). In particular, gender-incompatible cognates posed a major problem to German learners of Dutch, with only 30% correct on average, while compatible cognates represented the other extreme (92% correct). For non-cognates, the difference between gender-compatible and incompatible nouns was somewhat smaller (84% vs. 55%), but still clearly present. The results indicate again that German learners of Dutch as L2 tend to transfer German gender to Dutch nouns, in particular when the two translations are form-similar (cognates), but to a lesser extent also when they are not (non-cognates). The latter is quite remarkable, because the high incidence of gender-compatibility between Dutch and German holds for translation pairs with common roots (and thus high form-similarity, i.e. cognates) only. For words unrelated in terms of etymology (and in form, like *vork—Gabel*, 'fork'), gender compatibility is only coincidental. Nevertheless, German–Dutch bilinguals were influenced by German word gender not only in the case of cognates, where the mapping is systematic, but also in the case of non-cognates. In both cases, speakers, when acquiring the gender of a noun in their L2, also seem to activate the L1 translation equivalent including its grammatical gender, which is then in many cases transferred to the L2 noun (for a more detailed discussion of cross-language gender transfer, see Lemhöfer et al., 2008).

As a result of this L1–L2 transfer, the ranking of item conditions in terms of their gender assignment difficulty—with the highest difficulty for gender-incompatible cognates, followed by gender-incompatible non-cognates, gender-compatible cognates, and finally compatible cognates as the easiest item condition—returned in all measures and all phases of the present study. Gender-compatible cognates were not only responded to correctly more often than the other item groups, but participants were also most certain of their (correct) response for these items. In contrast, gender-incompatible cognates were shifted towards the 'certain and incorrect' part of the scale, i.e., participants have in many cases established a fairly stable, but incorrect gender representation for these nouns. This is confirmed also by the data shown in Table 5: Most incompatible cognates fall within the category 'incorrect—certain'. This illustrates the strong incorrect, L1-based bias for gender representations for these words.

4.2. Familiarization phase and begin of training: Response consistency and stability

One characteristic of second language performance that is absent in the native language is inconsistent behavior: a language error made once might not be made next time, or the other way around, an item that has been produced correctly before might suddenly be produced incorrectly. This is indeed confirmed by our consistency data (Table 6), that capture to what degree participants were consistent across two responses for a given item. The condition with the fewest errors, that of gender-compatible cognates, also produced the most consistent correct responses, while the most difficult condition (incompatible cognates) contained by far the fewest consistent correct responses. Thus, even when an item of this difficult condition was assigned its correct gender when it was first encountered (i.e., during familiarization), it had a fairly high chance (37%) of being produced *incorrectly* next time (i.e., in block 1 of training), compared to only 6% of such a chance for (initially correctly produced) compatible cognates. The incorrect responses show the mirrored pattern: When an 'easy' item (compatible cognate) was incorrect, its chance of being incorrect again in the next block was only 34%, while this chance was 81% for a 'difficult' item (incompatible cognate) and about 72% for the items of 'medium' difficulty, the non-cognates.

Together with the certainty ratings (see Table 3) which turned out to be highly correlated with response consistency (see Section 3.3), these data suggest that gender representations for L2 nouns with an incompatible L1 gender do not only tend to be incorrect, but are also relatively unstable, i.e. characterized by variable performance and low subjective certainty. The problem is therefore not only one of 'blind' incorrect transfer of L1 gender, because in that case, responses should *always* be incorrect and provided with a relatively high degree of certainty. Rather, in many cases, L2 speakers seem to switch back and forth between the incorrect L1-based gender and the correct one, the latter being the gender value that they must have encountered many times in the L2 input they receive every day. In contrast, for gender-compatible nouns, the degree of stability of the gender representation is much higher (89% of consistent responses, compared to 78% for incompatible nouns), possibly as a result of the fact that L1 bias and past experience with L2 converge onto the same gender value.

In the light of such high levels of errors and instability of gender representations, the aim of a successful L2 acquisition process would of course be the 're-stabilization' of incorrect representations towards the correct state. One simple possible method to reach this aim might be to give explicit and repeated feedback on erroneous responses, which is what we did in the training phase of the experiment, and which will be discussed in the following.

4.3. Training phase

In the training phase, participants repeatedly produced noun phrases that consisted of the object names and their gender-marked definite determiners in three blocks, and received feedback after each response. We examined to what degree the L2 speakers were able to learn the correct gender responses for the different item conditions. In contrast to what we hypothesized, the difficult (incompatible) items showed the biggest learning effects in absolute numbers (incompatible cognates, followed by incompatible non-cognates and compatible non-cognates; see Fig. 2). The easy condition of compatible cognates showed no significant learning, probably due to a floor effect (i.e., low error rates of only 8 % in training block 1, possibly leaving too little room for improvement). However, overall, the error rates were reduced by half after both feedbacks, with the largest effect after the first feedback, but also a substantial reduction in error rates after the second feedback (from 22% to 17% across conditions). Thus, erroneous gender representations are in many cases modifiable by simple corrective feedback.

The learning data were normalized to examine whether the size of the learning effect was independent of its different starting levels (i.e., error rates in training block 1). When expressed as percentages relative to the initial error rate level, the learning rate was statistically identical in all item conditions except for the compatible cognates, where no learning occurred (see Fig. 3). Cognate status and gender compatibility therefore did not play a role in the (relative) extent of improvement, except that gender-compatible cognates displayed such a high accuracy that there was no room for learning.

A further question had been whether the effectiveness of the feedback would be different depending on whether the initial gender representation (as assessed in the familiarization phase before training) was correct, and on how certain the participants were of the accuracy of that representation. First, the data (see Fig. 4) and their statistical analysis show that only the items that were also incorrect before the training (during familiarization) showed any training effect at all. Thus, at this point, response consistency comes into play again: Consistently incorrect responses across two item repetitions (dashed lines in Fig. 4) showed more learning than inconsistent responses that were first correct, but then incorrect (solid lines in Fig. 4). The failure to reduce inconsistent errors by feedback might point to a memory problem: In cases where a speaker gives variable gender assignments for a given noun, he or she might have difficulty remembering their last response to an item to which the feedback was given, which might make feedback ineffective—speakers would remember only the fact that they received negative feedback on their last response, but not the incorrect response itself that triggered that feedback. In contrast, this problem of remembering the to-be corrected responses should occur less frequently for consistently incorrect responses. Alternatively, especially in the 'correct certain' case, the non-significant improvement of performance for initially correct items might also (partly) be due to a floor effect given the low absolute error rates, similarly to what has been observed for gender-compatible cognates.

Second, with respect to the comparison between 'certain' and 'uncertain' items, our hypothesis had been that incorrectly acquired gender representations might be harder to modify when the participant was relatively certain of that (incorrect) representation. Because the initially correct responses did not show any learning at all, the test of this hypothesis has to focus on the two 'incorrect' conditions. Contrary to the hypothesis, the drop in error rates was descriptively steeper for 'incorrect and certain' items than for 'incorrect and uncertain' ones (see Fig. 4). However, the analysis of the normalized error rates showed that this difference in error rate change was not significant. This appears to be at odds with the account proposed above, namely that responses that are more consistent—and thus, more stable and probably also more 'certain'—are easier to modify than unstable ones due to a higher rate of remembering the 'to-be-corrected' response. However, even though not reaching statistical significance ($p = .12$ across both block transitions), the same trend seemed to be present in the data here, with larger improvements for 'incorrect certain' than for 'incorrect uncertain' items. Note that due to the unbalanced distribution of the certainty categories, the two categories under discussion here comprised only 16% (incorrect certain) and 19% (incorrect uncertain) of all trials (see Table 5), which might have lead to a reduced statistical power. Therefore, we have to conclude that statistically and in relative terms, the certainty of an incorrect gender representation did not play a role with respect to its modifiability; incorrect responses given with great certainty did not behave more 'stubbornly' when corrective feedback was given.

4.4. Post-training state (block 3)

Besides looking at the extent of improvement and whether it differs for the various item types, we should also consider the final error rates in block 3. In spite of the effective feedback, the error rates for the four different item conditions do not converge on the same level in the last block. Rather, they remain in their original order (see Figs. 1 and 3). The error rates for the most difficult item condition, gender-incompatible cognates, remain at above 30% even after two instances of feedback. Learning, though clearly present, was thus far from complete, and given that learning is usually asymptotic (Ritter & Schooler, 2001), it can be assumed that further feedback would have had only little additional effect. The limits of learning in the present situation are also illustrated by the fact that none of the participants showed a 100% learning rate. Thus, two instances of corrective feedback were never sufficient to (re-)acquire correct gender representation for all 96 used words. This illustrates once again the arduous nature of the acquisition of grammatical gender in Dutch by native speakers of German (and probably also by speakers of other native languages), and gives a hint as to the origins of 'fossilized' errors in L2: certain errors (e.g., those arising from L1 to L2 conflicts) seem to be hard to correct by explicit feedback. Findings from the L2 acquisition literature suggest that it is even less likely that these errors will be corrected by the implicit kind of non-corrective input non-native speakers usually receive from their L2 environment (Carroll et al., 1992; Ellis et al., 2006).

4.5. Summary and conclusions

In sum, the present study went beyond the mere question of whether word gender as such is learnable at all, for instance, when it is absent in L1 (Franceschina, 2005; Keating, 2009; White et al., 2004). Rather, we studied the nature and flexibility of existing L2 gender representations and found that they are strongly influenced by L1. Our results show that those gender representations that conflict with the corresponding L1 ones are more often represented incorrectly and in a more unstable way, give rise to variable performance, and remain relatively difficult for the L2 learners even after training. These difficulties are not restricted to form-similar gender-incompatible translations pairs (gender-incompatible cognates), but occur also for dissimilar translations (non-cognates), for which there is no systematic mapping between Dutch and German gender. Thus, the transfer from L1 to L2 is not purely a result of L2 learners' exploitation of existing L1–L2 correlations, but happens also in cases where such correlations are missing. This is in line with the fact that similar cross-language effects have also been found for Greek and German, two languages with largely unrelated lexical and gender systems (Salamoura & Williams, 2007).

Acknowledgements

The research reported in this article was part of the third author's Bachelor Thesis at the Donders Centre for Cognition. She is now at the

Centre for Language Studies, Radboud University, Nijmegen, and at the Max Planck Institute for Psycholinguistics, Nijmegen. Parts of the study were supported by a VENI-grant by the Netherlands Organization for Scientific Research (NWO) to Kristin Lemhöfer (project number 016.084.015). We would like to thank Matthieu Koppen for valuable statistical advice.

Appendix A. List of materials

For each of the 24 nouns in an item condition, the following information is given: **Dutch word**, Dutch definite determiner, German translation, German definite determiner, English translation.

Gender-compatible cognates

hond, de, Hund, der, dog; **vleugel,** de, Flügel, der, wing; **nagel,** de, Nagel, der, nail; **muis,** de, Maus, die, mouse; **villa,** de, Villa, die, villa; **boon,** de, Bohne, die, bean; **worst,** de, Wurst, die, sausage; **mantel,** de, Mantel, der, coat; **trompet,** de, Trompete, die, trumpet; **banaan,** de, Banane, die, banana; **bloem,** de, Blume, die, flower; **ezel,** de, Esel, der, donkey; **been,** het, Bein, das, leg; **geweer,** het, Gewehr, das, rifle; **podium,** het, Podium, das, stage; **net,** het, Netz, das, net; **juweel,** het, Juwel, das, jewel; **roer,** het, Ruder, das, rudder; **orkest,** het, Orchester, das, orchestra; **hemd,** het, Hemd, das, shirt; **skelet,** het, Skelett, das, skeleton; **stadion,** het, Stadion, das, stadium; **oor,** het, Ohr, das, ear; **pakket,** het, Paket, das, parcel.

Gender-incompatible cognates

auto, de, Auto, das, car; **gevangenis,** de, Gefängnis, das, prison; **datum,** de, Datum, das, date; **hoorn,** de, Horn, das, horn; **kabel,** de, Kabel, das, cable; **bijl,** de, Beil, das, ax; **muil,** de, Maul, das, mouth (of an animal); **taxi,** de, Taxi, das, taxi; **krokodil,** de, Krokodil, das, crocodile; **kameel,** de, Kamel, das, camel; **knie,** de, Knie, das, knee; **kano,** de, Kanu, das, canoe; **zand,** het, Sand, der, sand; **pistool,** het, Pistole, die, pistol; **kanaal,** het, Kanal, der, canal; **cijfer,** het, Ziffer, die, digit; **balkon,** het, Balkon, der, balcony; **strand,** het, Strand, der, beach; **spek,** het, Speck, der, bacon; **masker,** het, Maske, die, mask; **kompas,** het, Kompass, der, compass; **orgel,** het, Orgel, die, organ; **adres,** het, Adresse, die, address; **altaar,** het, Altar, der, altar.

Gender-compatible non-cognates

tuin, de, Garten, der, garden; **druppel,** de, Tropfen, der, drop; **vijver,** de, Teich, der, pond; **mand,** de, Korb, der, basket; **schuur,** de, Scheune, die, barn; **ui,** de, Zwiebel, die, onion; **laan,** de, Allee, die, avenue; **trui,** de, Pullover, der, jumper; **paddestoel,** de, Pilz, der, mushroom; **vlinder,** de, Schmetterling, der, butterfly; **krant,** de, Zeitung, die, newspaper; **vork,** de, Gabel, die, fork; **raam,** het, Fenster, das, window; **schilderij,** het, Gemälde, das, painting; **varken,** het, Schwein, das, pig; **wiel,** het, Rad, das, wheel; **konijn,** het, Kaninchen, das, rabbit; **zeil,** het, Segel, das, sail; **brein,** het, Gehirn, das, brain; **cadeau,** het, Geschenk, das, present; **vierkant,** het, Rechteck, das, rectangle; **gewricht,** het, Gelenk, das, joint; **gat,** het, Loch, das, hole; **spook,** het, Gespenst, das, ghost.

Gender-incompatible non-cognates

fiets, de, Fahrrad, das, bike; **poort,** de, Tor, das, gate; **groente,** de, Gemüse, das, vegetable; **pijp,** de, Rohr, das, pipe; **lucifer,** de, Streichholz, das, match; **tent,** de, Zelt, das, tent; **beurs,** de, Portemonnaie, das, peurs; **bagage,** de, Gepäck, das, baggage; **pleister,** de, Pflaster, das, plaster; **korrel,** de, Korn, das, grain; **jurk,** de, Kleid, das, dress; **piano,** de, Klavier, das, piano; **bos,** het, Wald, der, forest; **horloge,** het, Armbanduhr, die, watch; **plafond,** het, Decke, die, ceiling; **bot,** het, Knochen, der, bone; **perron,** het, Bahnsteig, der, platform; **blik,** het, Dose, die, tin; **hert,** het, Hirsch, der, deer; **potlood,** het, Bleistift, der, pencil; **fornuis,** het, Herd, der, stove; **krat,** het, Kasten, der, crate; **pak,** het, Anzug, der, suit; **litteken,** het, Narbe, die, scar.

References

Ayoun, D. (2001). The role of negative and positive feedback in the second language acquisition of the passé composé and the imparfait. *Modern Language Journal*, *85*, 226–243.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2) [CD-ROM].* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Carroll, S., Swain, M., & Roberge, Y. (1992). The role of feedback in adult second language acquisition: Error correction and morphological generalizations. *Applied Psycholinguistics*, *13*, 173–198.

Costa, A., Kovacic, D., Franck, J., & Caramazza, A. (2003). On the autonomy of the grammatical gender systems of the two languages of a bilingual. *Bilingualism: Language and Cognition*, *6*, 181–200.

Dewaele, J. -M., & Véronique, D. (2001). Gender assignment and gender agreement in advanced French interlanguage: A cross-sectional study. *Bilingualism: Language and Cognition*, *4*, 275–297.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*, 339–368.

Franceschina, F. (2005). *Fossilized second language grammars: The acquisition of grammatical gender.* Philadelphia: Benjamins.

Holmes, V. M., & Dejean de la Bâtie, B. (1999). Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics*, *20*, 479–506.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning*, *59*, 503–535.

Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning*, *58*, 703–746.

Klooster, W. (2001). Grammatica van het hedendaags Nederlands. Een volledig overzicht. [Grammar of contemporary Dutch. A complete overview.] Den Haag, The Netherlands: SDU.

Lemhöfer, K., Spalek, K., & Schriefers, H. (2008). Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language*, *59*, 312–330.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, *26*, 399–432.

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In N. J. Smelser & P.B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 8602–8605). Amsterdam: Elsevier.

Sabourin, L., Stowe, L. A., & de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, *22*, 1–29.

Salamoura, A., & Williams, J. N. (2007). The representation of grammatical gender in the bilingual lexicon: evidence from Greek and German. *Bilingualism: Language and Cognition*, *10*, 257–275.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.

Unsworth, S. (2008). Age and input in the acquisition of grammatical gender in Dutch. *Second Language Research*, *24*, 365–395.

White, L., Valenzuela, E., Kozlowska-MacGregor, M., & Leung, Y. -K. I. (2004). Gender and number agreement in nonnative Spanish. *Applied Psycholinguistics*, *25*, 105–133.