Journal of Memory and Language 59 (2008) 312–330

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Cross-language effects of grammatical gender in bilingual word recognition and production

Kristin Lemhöfer *, Katharina Spalek, Herbert Schriefers

NICI, Radboud University Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 20 August 2007
Revision received 9 April 2008
Available online 5 August 2008

Keywords:
Bilingual word recognition and production
Lexical decision
Picture naming
Grammatical gender in L2
Cross-language effects
Stability of L2 representations

ABSTRACT

We investigated whether bilinguals recognizing or producing noun phrases in their second language Dutch are influenced by the grammatical gender in their mother tongue German. The Dutch nouns used in the experiments were either gender-'compatible' or -'incompatible' with their German translation. Furthermore, their cognate status (form similarity with the translation) was varied. In Experiment 1, participants carried out a gender-primed lexical decision task on Dutch nouns. In Experiment 2, the same items were presented as pictures and had to be named in Dutch, either with or without their gender-marked determiners. Experiment 3 was a repetition of Experiment 2, but with a preceding training session to obtain additional and more reliable data. In all three experiments, effects of cross-language gender compatibility were obtained, especially for cognates. These results suggest that a bilingual's two gender systems interact. Additional analyses indicate that this interaction primarily affects the stability of gender representations in the second language.

Learning a second language after childhood is a difficult enterprise, and only few succeed in achieving native-like performance. One of the big stumbling blocks is grammatical gender, which remains a problem for foreign language learners at all levels (e.g., Dewaele & Véronique, 2001; Holmes & Dejean de la Bâtie, 1999; Rogers, 1987). One possible source of this problem may be the interference from the gender system of the speakers' mother tongue. The current study investigates this supposition for the domains of visual word recognition and spoken word production. Thus, we examine whether bilingual speakers are influenced by the grammatical gender of words in their *first language* (L1) during recognizing or producing words in their *second language* (L2). We will address this question for proficient, but 'unbalanced' German–Dutch bilinguals (i.e., their proficiency in their L2, Dutch, is less than native-like). Apart from the general issue of cross-language gender effects in word recognition and production, we will also investigate whether such effects depend on how similar a L2 word is to its L1 translation.

In spite of its obvious significance for second language learning, psycholinguists have only just started to experimentally investigate the influence of the native gender system during second language processing. Paris and Weber (2004) reported an experiment in which French-German bilinguals listened to German auditory questions (*Wo ist die$_{fem}$ Perle$_{fem}$?*/'Where is the pearl?') while looking at a display with several objects. The fixation patterns showed that whether competitor objects (i.e., objects with the same onset, e.g., *Perücke*—'wig') were fixated more often than control objects was co-determined by the gender of their *French translation*. All used words were cognates, that is, the translation equivalents were similar in form (e.g., *Perücke—perruque*). This result suggests that, at least in the case of cognates, the L1 translations and their gender affected auditory word recognition in L2.

A quite different conclusion was drawn by Costa, Kovacic, Franck, and Caramazza (2003). In their study, Croatian-Italian, Catalan-Spanish and Italian-French bilinguals named

* Corresponding author. Fax: +31 (24) 361 6066.
E-mail address: k.lemhofer@nici.ru.nl (K. Lemhöfer).

0749-596X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2008.06.005

pictures in gender-marked noun phrases in their respective L2. When compared to a baseline provided by monolinguals, no effects of cross-language gender-compatibility could be observed. This was taken as evidence for a complete independence of the two gender systems in bilinguals, even in quite closely related languages such as Italian and French.

In summary, the few experimental studies on this issue do not provide a conclusive answer on the question of the (in)dependence of gender systems in bilinguals. Furthermore, it is possible that results concerning cross-language gender effects are specific to the domain of language processing (e.g., word recognition vs. production) in which they are obtained. In the present study, we examined the role of native gender in second language processing for two well-investigated domains within psycholinguistic research, namely, visual word recognition and spoken word production. In direct comparison, results from word recognition and production should lead to a better understanding of word gender representation in the bilingual (and possibly also the monolingual) lexicon.

Besides establishing any potential effects of gender compatibility across languages, the present study also aims at identifying the *source* of such effects. The existing studies seem to assume explicitly or implicitly that gender is represented in the same way in L1 and L2, i.e., that for each noun in each language there is a correct and stable gender representation. Effects of cross-language gender compatibility would therefore be a consequence of "online" lexical processing, such that word processing is easier for overlapping (or compatible) relative to conflicting gender information in the two languages. However, another plausible source for gender-compatibility effects is that L2 gender representations might be unstable or even incorrect, due to an incomplete or faulty acquisition process. Such representational instability would especially affect nouns that are gender-incompatible with their translation in L1, resulting in greater processing difficulties for these nouns than for gender-compatible ones. In the present study, we will first seek to investigate whether effects of cross-language gender compatibility do exist in word recognition and production, and second, if they do, address the question whether they arise from imperfect gender acquisition in L2, or from "online" lexical competition processes.

Before we move on to the present experiments in more detail, we will give a short description of the relevant aspects of the German and Dutch gender systems and their relation. German has three classes of grammatical gender (masculine, feminine, and neuter) which differ in, among others, the singular definite determiner (nominative: *der*$_{masc}$ *Mann*$_{masc}$, *die*$_{fem}$ *Frau*$_{fem}$, *das*$_{neu}$ *Kind*$_{neu}$—'the man', 'the woman', 'the child').[1] In the past, Dutch has had the same three-way gender system, but in modern Standard Dutch, the masculine and feminine categories have practically collapsed into one, referred to as common gender (van Berkum, 1996).[2] Like in German, Dutch singular definite determiners are marked for gender, with nouns of common gender taking the determiner *de* (*de*$_{com}$ *man*$_{com}$, *de*$_{com}$ *vrouw*$_{com}$—'the man', 'the woman'), whereas neuter gender words require the definite determiner *het* (*het*$_{neu}$ *kind*$_{neu}$—'the child'). Due to the common Germanic roots of German and Dutch, many translation pairs have 'compatible' gender in the two languages, assuming that German neuter gender maps onto Dutch neuter gender, and German feminine and masculine gender onto Dutch common gender. The degree of transferability of German gender to Dutch is illustrated by the fact that among the 25 most frequent nouns in Dutch (according to the lexical database CELEX; Baayen, Piepenbrock, & Gulikers, 1995), there are only six with a gender-incompatible German translation; three of these still have the same gender if an outdated translation is assumed (e.g., 'room' is *kamer*$_{com}$ in Dutch and *Zimmer*$_{\mathrm{neu}}$ in modern German, but the old German word for it is *Kammer*$_{\mathrm{fem}}$). The high correlation between the two gender systems might lead German learners of Dutch to transfer their L1 gender knowledge to L2, even though this is not appropriate in all cases. Alternatively, given the results by Costa et al. (2003), they might acquire the gender of Dutch words 'from scratch', without taking the gender of their German equivalents into account. In the present study, we will investigate whether and to what extent the 'transfer' of word gender plays a role in L2 visual word recognition and spoken word production.

We chose German and Dutch as a language combination for several reasons. First, German and Dutch are Germanic languages with similar gender systems, which might be conducive for the transfer of gender properties from one language to the other. Because, as will be seen below, little is known about cross-language gender effects yet, we opted for a situation for which any potential effects would be most likely to occur. Second, in the monolingual literature, gender processing has been studied intensively in German and Dutch, with comparable results in both languages. In particular, gender congruency effects, which we used as tools to measure cross-language interaction at the gender level, have reliably been demonstrated in both German and Dutch (e.g., Schriefers, 1993; Schriefers & Teruel, 2000). Thus, we can be certain that speakers of these languages are sensitive to the kind of experimental manipulations employed in the present study.

As a first step, we investigated whether effects of gender-compatibility can be demonstrated at all in German–Dutch bilinguals, both in word recognition (Experiment 1) and in word production (Experiment 2). Only if such effects exist, a second step would be to clarify by an additional experiment whether they arise from imperfect gender acquisition in L2, or "online" from lexical competition between conflicting gender representations.

The participants carried out either a visual lexical decision task (Experiment 1) or a picture naming task (Experiment 2) in their L2 (Dutch) with the same word materials,

[1] We use the following indices to indicate gender: masc, masculine; fem, feminine; neu, neuter; com, common gender (in Dutch).

[2] Even though the masculine/feminine distinction is necessary for the choice of pronouns (i.e., *de koffie*. . ..*heb je hem*$_{\mathrm{masc}}$ *gehaald*? 'the coffee. . .did you get it$_{\mathrm{masc}}$?'), the masculine form is usually used for most de-words, especially in the northern language area, i.e., in the Netherlands (Klooster, 2001, pp. 352–353).

to allow for maximal comparability of the two experiments. The Dutch words were either gender-compatible or -incompatible with their German translation. Furthermore, the form similarity of the Dutch nouns with their German translation was varied: words were either dissimilar (e.g., *jurk—Kleid*, 'dress') or similar to their translation (e.g., *hond—Hund*, 'dog'). Words with a similar form and the same meaning across two languages are often referred to as *cognates*, and have been shown to be especially sensitive to cross-language influences during bilingual word recognition and production (e.g., Costa, Caramazza, & Sebastián-Gallés, 2000; Cristoffanini, Kirsner, & Milech, 1986; Dijkstra, Grainger, & van Heuven, 1999; Lemhöfer, Dijkstra, & Michel, 2004). Considering that many researchers have claimed that cognates are represented differently from non-cognates in the bilingual lexicon (de Groot & Nas, 1991; Gollan, Forster, & Frost, 1997; van Hell & de Groot, 1998), the extent of cross-talk between the gender systems of L1 and L2 might be modulated by this variable. Note that this variable was not manipulated in the bilingual studies mentioned above.

Experiment 1: Visual lexical decision

The first experiment was concerned with cross-language effects of grammatical gender during *visual word recognition* in L2, which is an issue that has to our knowledge not yet been studied. The experimental paradigm we used for the investigation of gender effects was visual lexical decision on nouns that were primed by gender-marked or gender-neutral determiners.

In the monolingual domain, it has been demonstrated that relative to a gender-neutral baseline, *invalid* (i.e., incorrect) gender cues slow down the word recognition process, while the results are mixed concerning whether or not *valid* gender primes can speed up word recognition (Gurjanov, Lukatela, Lukatela, Savic, & Turvey, 1985; Schmidt, 1986).[3] For instance, van Berkum (1996) used Dutch noun phrases with a definite or indefinite determiner (*the house/a house*), exploiting the fact that definite singular determiners in Dutch are marked for gender (*de*$_{com}$/*het*$_{neu}$), whereas the indefinite singular determiner is invariant across genders (*een*). In visual lexical decision, an overall gender (in)congruency effect could be found, with longer response latencies when target nouns were preceded by an invalid gender prime (**de*$_{com}$ *huis*$_{neu}$) than when the prime was valid (*het*$_{neu}$ *huis*$_{neu}$). However, with respect to the neutral indefinite determiner baseline (*een huis*), the result was unexpected: Participants were even faster in this condition than in the congruent-gender condition. When the incongruent condition was excluded from the experiment, there was no significant difference between the gender-neutral and the valid prime conditions.

Regarding gender priming in bilinguals, some studies have simply looked at whether L2 learners are different from native speakers. Guillelmon and Grosjean (2001) asked native speakers of French as well as early and late English–French bilinguals to repeat the last word of an auditorily presented noun phrase (e.g., '*table*' in *la*$_{fem}$ *jolie table*$_{fem}$). The neutral baseline was a phrase with a gender-unmarked possessive pronoun (*leur jolie table*). The results showed both facilitation and inhibition effects of (congruent and incongruent) gender priming, but only for the native speakers and those bilinguals who acquired French early in life; late bilinguals did not show any effects of gender priming. Similarly, using a German auditory lexical decision task, Scherag, Demuth, Rösler, Neville, and Röder (2004) demonstrated that native speakers of German, even if they had lived abroad for years, benefited from congruent relative to incongruent gender primes (*faltiges*$_{neu}$ *Gesicht*$_{neu}$ vs. **faltiges*$_{neu}$ *Haut*$_{fem}$—'wrinkled face', 'wrinkled skin'). However, native English speakers who had lived in Germany for a long time (15 years on average) did not show such an effect of gender priming in German. Both studies indicate that the effective use of gender information in lexical access might be a function of the age at which the second language has been acquired.

While these studies demonstrated that gender processing works differently in native and certain non-native speakers of a given language, they did not look at the potential role of the first language during gender processing in L2. The present experiment investigated cross-language gender effects in a lexical decision task. In this experiment, we avoided violating the grammatical rules of the target language, Dutch, considering that the presence and a relatively high proportion of incorrect (i.e., gender-incongruent) trials makes the experiment less 'natural', and might induce experiment-specific strategies (e.g., van Berkum, 1996). Thus, no invalid gender condition was used. Rather, the manipulation of the compatibility of the noun's gender with that of its German translation opens the possibility to introduce a '*hidden incongruent*' condition. For instance, even though the Dutch determiner-noun phrase *de*$_{com}$ *jurk*$_{com}$ ('the dress') is correct, German speakers of Dutch might experience this phrase as (somewhat) incongruent because the German word for 'dress', *Kleid*, has neuter gender. In this case, an inhibition effect should be observed, similar to the one in the incongruent condition in monolingual studies.

The types of Dutch phrases we used were the same as those employed by van Berkum (1996): Participants made lexical decisions on Dutch target nouns that were either preceded by their correct (gender-marked) definite determiner (*de*$_{com}$ *jurk*$_{com}$), or by the gender-unmarked indefinite determiner (*een jurk*). The latter condition was used as a neutral baseline, with the difference between the two conditions representing the *gender priming effect*. If participants do indeed experience interference from their L1 in the 'hidden incongruent' condition, the gender priming effect should—given the robust incongruency effects in monolingual studies—be less facilitatory (or even inhibitory) for 'incompatible' nouns, relative to nouns with a gender-compatible translation in German.

Furthermore, as mentioned before, the form similarity (or cognate status) of the Dutch nouns with their German

[3] Here, we focus on visual word recognition. However, it should be noted that in the auditory domain, facilitatory priming by congruent gender information seems to be more stable (Bates, Devescovi, Hernandez, & Pizzamiglio, 1996; Bölte & Connine, 2004; Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Grosjean, Dommergues, Cornu, Guillelmon, & Besson, 1994).

translation was varied, with half of the nouns being cognates between German and Dutch (e.g., *hond—Hund*), and half of them non-cognates (*jurk—Kleid*). It is as yet unknown whether processing of word gender in L2 and its susceptibility to transfer effects from L1 is dependent on the cognate status of the respective noun. Given the sensitivity of cognates to cross-language effects, it is likely that the expected cross-language influence with respect to gender, as described above, is stronger for cognates than for non-cognates.

Methods

Participants

Twenty-four native speakers of German currently immersed in a Dutch environment participated in the experiment. Most participants were students or scientific employees of the University of Nijmegen. The data of one participant had to be excluded because she had misunderstood the instructions; another three participants were excluded because of low scores in the Dutch vocabulary test (mean% correct <70%; see section on the Vocabulary test).

The remaining 20 participants were between 23 and 37 years old (mean 28.2); 13 were female. They all reported to have normal or corrected-to-normal vision, and to be non-dyslexic. All but one were right-handed, and all participants stated that German was their dominant language. The time they had lived in the Netherlands varied from 1.5 to 11 years. A language questionnaire provided more information on the participants' language background, which is summarized in Appendix A. The participants also used other foreign languages than Dutch regularly, in particular English (18 participants). None of the participants stated using English or any other foreign language more frequently than Dutch.

The participants also carried out a Dutch language test assessing vocabulary size, which will be described in a separate section.

Stimulus materials

Words

Ninety-six Dutch nouns with a length between two and ten letters were selected from the CELEX database (Baayen et al., 1995), with 24 nouns in each of the four conditions that were formed by combining the two 2-levels factors Cognate Status and Compatibility with the German gender. None of the words had endings that are predictive with respect to Dutch word gender, with two exceptions.[4] Half of the words in each condition were *de*-words, the other half *het*-words. Because of the intention to use the same word materials in the picture naming experiment (Experiment 2), all words were concrete and could be depicted. In case of the existence of several possible German translations for a given Dutch noun, the dominant translation was used, as judged by two proficient German–Dutch bilinguals.

Words were classified as cognates when they were phonologically and/or orthographically very similar to their German translation, and as non-cognates when this similarity was small. Dutch nouns with common gender (definite determiner *de*) were regarded as gender-compatible when their German translation possessed masculine or feminine gender, and Dutch nouns with neuter gender were categorized as gender-compatible when the gender of the German translation was also neuter. 'False friends' that are (almost) identical in form, but dissimilar in meaning were not included in the material. The four groups of stimuli were matched as closely as possible on Dutch logarithmic frequency, number of letters, and number of syllables. The characteristics of the words are summarized in Table 1. There were no significant differences between the four item categories with respect to any of these variables, as analyzed in ANOVAs (all $F < 1$). All word items are listed in Appendix B.

Nonwords

Ninety-six nonwords were constructed by changing one or more letters in existing Dutch words, resulting in pronounceable and word-like letter strings. A native speaker of Dutch indicated her intuitions with respect to the potential gender (or definite determiner) of the nonwords. The selection of nonwords was then performed such that half of the selected nonwords were '*de*-nonwords', the other half had been categorized as '*het*-nonwords'. The distributions of length (3–9 letters, mean 5.3) and number of syllables (1–3 syllables, mean 1.7) of the nonwords resembled those of the words. All nonwords are listed in Appendix B.

Procedure

The complete experimental session consisted of the main experiment (the lexical decision task in Dutch), a vocabulary test in form of a non-speeded lexical decision task in Dutch, and the language questionnaire. Each session took about 30 min. Participants were paid or given course credit for their participation.

The Dutch lexical decision task began with the participant reading a written instruction, explaining that they would see a determiner and, appearing shortly after that, a letter string, and that their task was to determine whether the last letter string was a Dutch word or not. They had to do this by pressing one of two buttons as quickly and accurately as possible (with the 'yes' response assigned to the dominant hand). A practice block was presented, consisting of a total of 16 trials (eight words and eight nonwords, none of which appeared in the main experimental lists). Like in the main experiment, half of the practice trials were presented with the indefinite determiner, the other half with the definite determiner.

The main experiment consisted of 192 trials (96 words and nonwords, respectively), presented in three blocks of

[4] The two words in question, *gevangenis*$_{com}$ and *bagage*$_{com}$, were both in gender-incongruent conditions. Thus, the predictability of their gender based on their word endings, possibly facilitating gender retrieval, works against a potential critical (inhibitory) effect of gender incompatibility. It is therefore impossible that the effect in question arises as a result of gender cues in the word endings.

Table 1
Characteristics of the word materials for each of the four stimulus categories

Word category	Example (German and English translation)	Mean no. of letters	Mean no. of syllables	Mean Dutch log frequency
Cognates				
Gender-compatible	*hond*$_{\text{com}}$ (*Hund*$_{\text{masc}}$, dog)	5.1 (1.2)	1.67 (.64)	1.30 (.46)
Gender-incompatible	*auto*$_{\text{com}}$ (*Auto*$_{\text{neu}}$, car)	5.4 (1.5)	1.83 (.70)	1.27 (.46)
Non-cognates				
Gender-compatible	*vork*$_{\text{com}}$ (*Gabel*$_{\text{fem}}$, fork)	5.5 (2.0)	1.50 (.66)	1.34 (.41)
Gender-incompatible	*jurk*$_{\text{com}}$ (*Kleid*$_{\text{neu}}$, dress)	5.4 (1.6)	1.67 (.76)	1.28 (.34)
Total mean		5.4 (1.6)	1.67 (.69)	1.30 (.41)

Note: Standard deviations are given in parentheses.

64 trials each, between which participants were free to take breaks. Additionally, the first two items of each block were warming-up items (one word, one nonword) which were not included in the analyses. The participants were randomly assigned to one of four item orders (created by two different randomizations in two list versions each, where condition was counterbalanced across the two versions, i.e., the assignment of indefinite and definite determiner condition were exchanged for each item). Thus, each participant saw half of the words and nonwords with the definite determiner and the other half with the indefinite determiner; for a participant receiving the complementary list version, this assignment was reversed. Each of the four lists was presented to five of the 20 participants that entered the final analyses. The order of items in the lists was pseudo-randomized, with no more than three words or nonwords, or more than three definite or indefinite determiners in a row.

Participants were seated approximately 70 cm from the screen. Each trial began with the presentation of a fixation dot for 700 ms. After 100 ms, the dot was replaced by the determiner prime. After another 250 ms, the noun appeared on the screen to the right of the determiner, while the determiner stayed on the screen as well. The complete phrase stayed on the screen until the participant made a response, or until a deadline of 3000 ms was reached. There was no feedback on the accuracy or speed of the response. The inter-trial interval was 1000 ms. All items were presented in black 28 point upper case 'Geneva' letters on a white background. Response latencies were measured to the closest millisecond.

Dutch vocabulary test

After the main experiment, participants were asked to complete a vocabulary test in Dutch in the form of a non-speeded lexical decision task performed on the computer. The test was a Dutch version of an English vocabulary test that was originally developed by Meara (1996) and adapted by Lemhöfer et al. (2004). The present Dutch version was developed by the first author in analogy to the English test, matching the items to those in the English version with respect to length, number of syllables, frequency, syntactic class, cognate status with German, and morphological structure.

Table 2
Dutch vocabulary test scores for the participants of Experiments 1, 2, and 3

	Mean	*SD*	Minimum	Maximum
Experiment 1	83	8	71	96
Experiment 2	87	6	77	98
Experiment 3	74	8	59	88

Note: SD, standard deviation. Scores were calculated by averaging % correct values for words and nonwords.

The test consisted of 60 items, 40 of which were words, 20 nonwords.[5] The items were between 4 and 12 letters long (mean: 7.4); the 40 words lay in a frequency range from 1 to 28 occurrences per million (mean: 6.3), according to the CELEX database (Baayen et al., 1995). The order of items in this test was the same for all participants, with no more than five words or nonwords in a row. All items are listed in Appendix C. Participants were to decide whether the presented letter string formed a correct Dutch word or not; they could take as much time for their responses as they wished. This way, the test represents a pure vocabulary test without a speed component. Furthermore, participants were instructed to respond with 'yes' only when they were sure that the item was a Dutch word; in case of uncertainty, they should press the 'no' button. Scores were calculated using a percentage correct measure, corrected for the unequal number of words and nonwords (i.e., the mean percentage of correctly recognized words and correctly rejected nonwords). The results of the vocabulary test for the participants of Experiment 1, as well as those for the participants of Experiments 2 and 3, are summarized in Table 2.

Results

Analyses of variance were run on both RTs and error rates, and on both participant and item means, with the factors Prime Type (definite vs. indefinite determiner),

[5] The reason for not keeping to the standard of a 50%–50% proportion of words and nonwords was the high difficulty of the test (i.e., the low frequency of the words), which makes it unlikely that our participants would know all of the words (in their weaker language). Under the assumption that the participants know about 75% of the presented words, the 'internal' proportion of familiar and unfamiliar items would therefore, on average, approximately be equal.

Cognate Status (cognates vs. non-cognates) and Gender Compatibility (same vs. different gender as in German). In the analysis of participants, all these factors were repeated-measures factors, while in the item analysis, Cognate Status and Gender Compatibility were between-item factors, and Prime Type a within-item factor. Four items with error rates above 50% in the indefinite determiner condition were excluded from all further analyses: *kano* ('canoe'), *muil* ('mouth'; both incompatible cognates), *boon* ('bean', compatible cognate) and *bot* ('bone', incompatible non-cognate). For the remaining items, the overall error rate was 7.2% (6.9% for nonwords, and 7.4% for words). Erroneous responses were excluded from the RT analyses, as were RTs that lay more than two standard deviations away from the participant (per experimental condition) and item (per Prime Type condition) mean. The percentage of outliers was 1.2% of the correct word responses. The mean RTs and error rates and the priming effects are shown in Table 3.

Reaction times

The results of the ANOVA on RTs are reported in Table 4. In the analysis of RTs, Cognate Status had a significant effect, with faster recognition latencies for cognates (666 ms) than for non-cognates (722 ms). Prime Type did not significantly influence RTs, but there was a main effect of Gender Compatibility: words for which the German translation equivalent was compatible in gender were recognized faster (678 ms) than gender-incompatible words (709 ms). This main effect was qualified by the crucial interaction between Gender Compatibility and Prime Type, which was significant over participants, but not over items. The triple interaction Cognate Status by Gender Compatibility by Prime Type was not significant, nor was any of the other interactions.

Even though the critical three-way interaction with Cognate Status was not significant, the data pattern for cognates and non-cognates looked qualitatively different: While there seemed to be a cross-over interaction of Gender Compatibility and Prime Type in the cognates, this was evidently not the case for non-cognates. We calculated this interaction for cognates and non-cognates separately, to investigate whether this notion was correct. In the analysis of *cognates*, the Gender Compatibility by Prime Type interaction was indeed significant. Pairwise t-tests showed that the 16 ms advantage of definite over indefinite determiner primes for compatible cognates was not significant (see Table 3 for the confidence intervals indicated by these tests). However, for incompatible cognates, response latencies were, on average, 31 ms longer for definite than for indefinite determiner primes, which was significant. For *non-cognates*, Prime Type did not interact significantly with Gender Compatibility.

Error rates

The statistical results of the error analysis are reported in Table 5. In the overall analysis of error rates, there was a facilitatory cognate effect, with fewer errors on cognates (4.9%) than on non-cognates (9.8%). There was no main effect of Prime Type or of Gender Compatibility, but the important interaction of Gender Compatibility and Prime Type was significant. However, pairwise comparisons showed that the gender priming effect, i.e., the effect of Prime Type, did not reach significance for any of the individual word conditions (see the confidence intervals in Table 3). None of the further interactions was significant. Nevertheless, in analogy to the analysis of RTs, separate analyses of cognates and non-cognates were carried out. In these analyses, none of the main effects or interactions was significant, including the interaction of Gender Compatibility by Prime Type.

Discussion

The results of the word recognition experiment show, first of all, that cognates were recognized faster and with fewer errors than non-cognates. This replicates the cognate effect, which is a standard finding in studies on bilingual visual word recognition (e.g., Caramazza & Brones, 1979; de Groot, Borgwaldt, Bos, & van den Eijnden, 2002; Dijkstra, van Jaarsveld, & ten Brinke, 1998; Lemhöfer et al., 2008) and extends it to the processing of determiner-noun phrases.

Furthermore, and more importantly given the present issue of investigation, evidence was found for cross-language gender compatibility modulating the gender priming effect. For RTs, this was mainly true for cognates: Cognates with a different gender in German (e.g., *auto*) were recognized more slowly when they followed the def-

Table 3
Mean RTs (ms) and error rates (%) and priming effects in all item conditions in Experiment 1

	RTs			Error rates		
	Definite determiner	Indefinite determiner	Priming effect[a]	Definite determiner	Indefinite determiner	Priming effect[a]
Cognates						
Gender-compatible	646 (74)	662 (88)	+16 [25]	3.0 (5.0)	5.2 (8.1)	+ 2.2 [4.1]
Gender-incompatible	693 (116)	662 (92)	– 31* [29]	6.8 (10.6)	4.5 (6.3)	– 2.3 [3.6]
Non-cognates						
Gender-compatible	702 (95)	702 (93)	0 [27]	7.9 (10.6)	10.4 (16.2)	+ 2.5 [4.6]
Gender-incompatible	746 (100)	735 (86)	– 11 [32]	10.9 (11.4)	10.0 (16.1)	– 0.9 [2.9]

Note: Standard deviations are given in rounded parentheses, the halfwidth of 95% confidence intervals is given in square parentheses. Priming effects that are significant with $p < .05$ are marked with an asterisk.

[a] Indefinite minus definite determiner condition.

Table 4
Results of the ANOVA of reaction times of Experiment 1

	*F*1			*F*2			min *F*′		
	df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Prime type	1,19	.76	.40	1,88	.93	.34	1,55	.42	.52
Cognate status*	1,19	52.61	<.001	1,88	10.75	<.001	1,107	8.92	<.01
Gender comp.*	1,19	17.06	<.001	1,88	4.53	<.05	1,106	3.58	.06
Prime type × Cognate status	1,19	.03	.87	1,88	.01	.91	1,94	.01	.92
Prime type × Gender Comp.	1,19	4.27	<.05	1,88	2.49	.12	1,86	1.57	.21
Cognate status × Gender comp.	1,19	1.00	.33	1,88	.17	.68	1,106	.15	.70
Prime Type × Cognate Status × Gender Comp.	1,19	2.31	.15	1,88	1.73	.19	1,75	.99	.32
Cognates only									
Prime Type	1,19	.75	.40	1,43	.88	.35	1,49	.40	.53
Gender Comp.	1,19	3.98	.06	1,43	1.51	.23	1,62	1.09	.30
Prime Type × Gender Comp.*	1,19	6.00	<.05	1,43	6.38	<.05	1,51	3.09	.08
Non-cognates only									
Prime Type	1,19	.27	.61	1,45	.28	.60	1,53	.14	.71
Gender Comp.	1,19	15.97	<.01	1,45	3.16	.08	1,59	2.64	.11
Prime Type × Gender Comp.	1,19	.38	.55	1,45	.03	.87	1,52	.03	.87

Note: Comp., Compatibility.
Effects that are significant ($p < .05$) in both participant and item analyses are marked with an asterisk.

Table 5
Results of the ANOVA of error rates of Experiment 1

Effect	*F*1			*F*2			min *F*′		
	df	*F*	*p*	*dfs*	*F*	*p*	*df*	*F*	*p*
Prime Type	1,19	.12	.74	1,88	.17	.68	1,50	.07	.79
Cognate Status*	1,19	7.08	<.05	1,88	8.50	<.01	1,56	3.86	.05
Gender Comp.	1,19	1.46	.24	1,88	.70	.40	1,93	.47	.51
Prime Type × Cognate Status	1,19	.13	.72	1,88	.22	.64	1,45	.08	.78
Prime Type × Gender Comp.*	1,19	4.64	<.05	1,88	4.41	<.05	1,65	2.26	.14
Cognate Status × Gender Comp.	1,19	.02	.88	1,88	.01	.93	1,92	.01	.94
Prime Type × Cognate Status × Gender Comp	1,19	.10	.75	1,88	.08	.77	1,72	.04	.83
Cognates only									
Prime Type	1,19	.00	.99	1,43	.02	.97	1,19	.00	.98
Gender Comp.	1,19	1.38	.26	1,43	.73	.40	1,62	.48	.49
Prime Type × Gender Comp.	1,19	3.22	.09	1,43	3.21	.08	1,53	1.61	.21
Non-cognates only									
Prime Type	1, 19	.19	.67	1, 45	.35	.56	1, 40	.12	.73
Gender Comp.	1, 19	.59	.45	1, 45	.21	.65	1, 64	.15	.70
Prime Type × Gender Comp.	1, 19	1.81	.19	1, 45	1.49	.23	1, 57	.82	.37

Note: Comp., Compatibility.
Effects that are significant ($p < .05$) in both participant and item analyses are marked with an asterisk.

inite determiner than after the gender-neutral indefinite determiner. In other words, inhibition of the (in a 'German' sense) 'incongruent' gender information carried by the definite determiner was observed relative to the baseline. The opposite effect for cognates with the same gender in German (facilitation of the gender prime relative to the baseline), although descriptively present, did not reach statistical significance, neither did the effects of gender priming for non-cognates. Although the result pattern in the error rates looked similar descriptively, the effects of gender priming were not strong enough to reach significance for any of the four stimulus categories.

The non-significant facilitation of gender information in the compatible condition relative to the neutral baseline is in line with the failure to find facilitatory gender priming effects for monolinguals and the same type of Dutch determiner-noun phrases by van Berkum (1996). However, the slower recognition of cognates with different gender in L1 and L2 following the gender-marked determiner indicates that our introduction of a 'hidden incongruent' condition had indeed an effect that was comparable to that of a 'real' incongruent condition in monolingual experiments. Evidently, a gender prime that is gender-congruent with respect to the target language can act as an incongruent gender prime by way of the other, task-irrelevant language.

Altogether, these data show that the grammatical gender systems in the two languages interact with each other during visual word recognition. However, this interaction has a measurable effect only on L2 nouns with a form-similar translation in L1 (i.e., cognates).

Experiment 2: Picture naming

In Experiment 2, we investigated whether a language production experiment, involving participants drawn from the same bilingual population and the same word materials, would give rise to similar results to those observed for word recognition in Experiment 1. As pointed out by Costa and Santesteban (2004), the mechanisms underlying bilingual word recognition and production differ from each other substantially, for example in terms of the speaker's control over the relative activation of the two languages, and should therefore not be treated as "two sides of the same coin". The two domains might thus give rise to very different results concerning the interaction of the two gender systems of a bilingual.

The bilingual picture naming experiment was intended to investigate whether gender retrieval would be more difficult for Dutch nouns with a gender-incompatible German translation than for those with a gender-compatible one. As already mentioned in the introduction, Costa et al. (2003) investigated the same issue for several language pairs and concluded from their picture naming results that the two gender systems of bilinguals do *not* interact, but are independent of each other, regardless of the similarity of the two languages. However, in a study using translation from Greek (L1) to German (L2), Salamoura and Williams (2007) found evidence for a shared gender system between the two languages, with longer translation times for gender-incongruent relative to gender-congruent translation pairs, when the target phrase was marked for gender. Thus, previous evidence is not conclusive with respect to cross-language gender interaction in word production.

Experiment 2 was a Dutch picture naming experiment involving drawings of the nouns that had been used in Experiment 1. The participants were asked to name these pictures either using a noun phrase including the gender-marked definite determiner (*de$_{com}$ hond*—'the dog'), or using a bare noun (e.g., *hond*—'dog'). The choice of a different gender-unmarked baseline condition (bare noun) in this experiment compared to Experiment 1 (indefinite determiner + noun) was motivated by the requirements of the task, as will become clear from the Methods section.

Methods

Participants

Eighteen German–Dutch bilinguals taken from the same population as in Experiment 1 participated. None of them had taken part in Experiment 1. The data of two participants had to be excluded because of low scores in the Dutch vocabulary test (mean% correct <70%). The remaining 16 participants were between 22 and 46 years old (mean 28.6), 11 were female. All participants had normal or corrected-to-normal vision, and all but two participants were right-handed. German was the participants' dominant language according to their own reports. They had lived in the Netherlands for between 1 and 17 years (mean: 6.1), with between 2 and 20 years of experience with Dutch (mean: 7.5). Participants filled in the same language questionnaire as that used in Experiment 1, the results of which are summarized in Appendix A. The participants also spoke other foreign languages than Dutch, in particular English (11 participants). Two participants stated that they knew English better than Dutch. The participants also carried out the same Dutch vocabulary test as the participants of Experiment 1. The results of the vocabulary test are reported in Table 2.

Materials

For each of the 96 stimulus words of Experiment 1, line drawings of objects depicting these names were chosen from a database of the Max Planck Institute of Cognitive Neuroscience in Leipzig. Missing pictures were created using the *Google* internet image search and simplifying the resulting pictures with picture editing software. The pictures were approximately 5 by 5 cm in size, and were presented as black line drawings on a white background. Twenty-four additional pictures, 12 with Dutch names of common gender and 12 with Dutch names of neuter gender, were chosen as warming-up items to be inserted at the beginning of the experimental blocks.

Procedure

The participant was seated in a dimly lit room, separated from the experimenter by a partition wall. The visual stimuli were presented centered on a 17″ SVGA monitor at a resolution of 640 by 480 pixels. Viewing distance was about 80 cm. The presentation of the stimuli and the online collection of data were controlled by NESU software developed by the Max Planck Institute for Psycholinguistics, running on an Intel Pentium 166 MHz computer. Speech-onset latencies were measured to the closest millisecond with a voice key connected to the computer. Participants' responses were recorded with a DAT recorder.

Participants were tested individually in a session lasting about an hour. The experimental session consisted of five parts (familiarization with the picture materials, the main experiment, the language questionnaire, the vocabulary test, and an offline gender assignment task).

In the familiarization phase, the participants received a booklet with all experimental pictures and their names. They were told to study the picture names carefully and to use only those names in the experiment.

In the main experiment, as in Experiment 1, a gender-neutral baseline and a gender-marked condition were administered within participants. In the baseline condition, participants had to produce the name of the picture as a bare noun (e.g., *hond*—'dog'); in the experimental condition, pictures had to be named together with their singular definite determiner (*de$_{com}$ hond*—'the dog'). To be able to instruct participants on which phrase type had to be used for naming the pictures, the two phrase type conditions (bare nouns vs. determiner noun phrases) were administered blockwise, with short instructions given at the beginning of each block. With this necessary blocking of the conditions, the indefinite determiner *een* could not be used as the baseline condition, because all utterances in the baseline block would then start with the same word, enabling the participant to begin the utterance prematurely. Therefore, bare nouns were chosen as utterance format for the gender-neutral baseline.

The picture naming task began with a written instruction explaining the experimental procedure. There was a short practice phase, consisting of one bare noun naming block and one definite determiner noun phrase (NP) block of eight trials each. Practice items were taken from the pool of the 24 warming-up items. The main session which followed consisted of 192 experimental trials in total. The 96 experimental items were presented twice, once in the first and once in the second half of the experiment. One of the two appearances was in the bare noun condition, the other in the noun phrase condition. The items were presented in eight blocks each comprising 24 experimental items, and six additional warming-up items presented at the beginning of each block. Each block was preceded by a short Dutch instruction on the screen informing the participant whether the pictures had to be named as bare nouns or noun phrases, with alternating instructions in successive blocks. Similar to Experiment 1, two randomizations of item order were used, each of which had two versions for the counterbalancing of Phrase Type (i.e., all items presented in the bare noun condition first in one version were presented in the noun phrase condition first in the complementary version). Each of the four item orders was assigned to four participants.

At the beginning of each trial during the main experiment, a fixation dot was presented for 800 ms. After a blank interval of 200 ms, the picture was presented for 2500 ms; response registration was possible for 3000 ms from picture onset. The inter-trial interval was 750 ms.

After the main experiment, participants filled in the language questionnaire and carried out the vocabulary test. In the vocabulary test, ten warming-up items were included at the beginning to make participants familiar with the lexical decision task. The same experimental software (NESU) was used as in the main experiment. Participants were instructed to press the button at the side of their dominant hand for the 'yes' (or 'word') response and the other button for 'no' (or 'not a word').

Finally, participants were given a list on paper with all experimental nouns written below each other in alphabetic order, and were asked to write the correct definite determiner in front of each word. This additional test was intended to clarify whether gender errors that had occurred in the main experiment also occurred without time pressure.

Results

The statistical procedure was identical to that of Experiment 1, except that the factor Prime Type was now replaced by Phrase Type (definite determiner NP vs. bare noun).

The overall error rate was 28.5%. These errors included disfluencies and self-corrections, selection of the wrong determiner or noun, missing reactions, or voice key errors, and were excluded from the RT analyses, as were outlier RTs. The percentage of outliers was 0.2% of the correct trials. The mean RTs, error rates, and the differences between noun phrase naming and bare noun naming are shown in Table 6.

Reaction times

Note that as a consequence of the high error percentages, the statistical power in the analyses of the remaining correct RTs was extremely reduced. The statistical values of the ANOVA of RTs are given in Table 7.

When analyzed across participants, there was a significant effect of Cognate Status, with faster naming latencies for cognates (1121 ms) than for non-cognates (1163 ms), which was not significant by items.[6] Phrase Type also had a significant influence on RTs, with faster naming latencies in the bare noun naming (1057 ms) than in the determiner NP naming condition (1228 ms). Critically, the interaction of Phrase Type and Gender Compatibility was not significant, nor was any of the other effects and interactions.

In analogy to Experiment 1, we proceeded by carrying out separate analyses for cognates and non-cognates, especially to investigate the critical Phrase Type by Gender Compatibility interaction. However, this interaction was not significant for either of the two word types, even though there was a trend for non-cognates.

Error rates

The results of the ANOVA on error rates are reported in Table 8. There was no significant main effect of Cognate Status, but the effect of Phrase Type was significant, with a mean of 17% errors in the bare noun naming condition and 41% in the determiner noun phrase condition. There was also a main effect of Gender Compatibility: More errors were made on nouns with incompatible gender in German (36%) than on nouns with compatible gender (21%). Importantly, the interaction of Phrase Type and Gender Compatibility was significant, which was further qualified by a significant three way interaction with the factor Cognate Status.

The interactions of Cognate Status were further investigated by means of separate analyses for cognate and non-cognate nouns. The results for *cognates* and *non-cognates* followed the same pattern: In all cases, Phrase Type significantly influenced error rates, with more errors in the production of determiner-noun phrases than in the production of bare nouns. The crucial gender congruency effect, visible in the interaction of Phrase Type and Gender Compatibility, was significant for both cognates and non-cognates. Paired comparisons (see confidence intervals in Table 6) showed that in spite of the larger effect of Phrase Type for incompatible relative to compatible nouns, it was significant for all noun types.

Online gender errors

The difference between the overall error rates in the noun phrase and bare noun naming condition, should, in principle, reflect errors in word gender, assuming that other error sources (disfluencies, erroneous naming of the noun, voice key artifacts) affected both naming condi-

[6] Note that in the RT analysis, some items in some conditions did not enter the item analysis at all because of error rates of 100% (for a respective item and condition), affecting the degrees of freedom of F2.

Table 6
Mean RTs (in ms) and Error Rates (in%) and Phrase Type Effects in all Item Conditions in Experiment 2

	RTs			Error rates		
	Noun phrase	Bare noun	Phrase Type effect[a]	Noun phrase	Bare noun	Phrase Type effect[a]
Cognates						
Gender-compatible	1178 (177)	1002 (115)	176* [57]	21.6 (9.8)	16.4 (10.3)	5.2* [3.8]
Gender-incompatible	1261 (281)	1044 (126)	217* [121]	62.5 (14.7)	16.2 (9.6)	46.3* [7.0]
Non-cognates						
Gender-compatible	1193 (178)	1091 (130)	102* [64]	28.4 (15.8)	16.4 (11.9)	12.0* [6.1]
Gender-incompatible	1281 (203)	1089 (126)	192* [82]	49.5 (14.1)	17.2 (12.0)	32.3* [12.5]

Note: Standard deviations are given in rounded parentheses, the halfwidth of 95% confidence intervals is given in square parentheses. Phrase Type effects that are significant with $p < .05$ are marked with an asterisk.

[a] Noun phrase minus bare noun condition.

Table 7
Results of the ANOVA of RTs of Experiment 2

Effect	F1			F2			min F′		
	df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Phrase Type	1,15	43.41	<.001	1,91	94.68	<.001	1,31	29.76	<.001
Cognate Status	1,15	7.77	<.01	1,91	1.72	.19	1,105	1.41	.24
Gender Comp.	1,15	11.69	<.01	1,91	1.52	.22	1,105	1.35	.25
Phrase Type × Cognate Status	1,15	2.37	.15	1,91	.19	.66	1,102	.18	.68
Phrase Type × Gender Comp.	1,15	2.60	.13	1,91	2.28	.14	1,57	1.21	.28
Cognate Status × Gender Comp.	1,15	.29	.60	1,91	.08	.78	1,101	.06	.80
Phrase Type × Cognate Status × Gender Comp	1,15	.66	.43	1,91	1.73	.19	1,28	.48	.50
Cognates only									
Phrase Type*	1,15	34.22	<.001	1,45	73.0	<.001	1,30	23.30	<.001
Gender Comp.	1,15	4.28	.06	1,45	.36	.55	1,52	.33	.57
Phrase Type × Gender Comp.	1,15	.51	.49	1,45	.03	.87	1,50	.03	.86
Non-cognates only									
Phrase Type*	1,15	29.39	<.001	1,46	33.77	<.001	1,42	15.71	<.001
Gender Comp.	1,15	9.21	<.01	1,46	1.53	.22	1,58	1.31	.26
Phrase Type × Gender Comp.	1,15	4.45	.05	1,46	3.12	.08	1,53	1.83	.18

Note: Comp., Compatibility.
Effects that are significant ($p < .05$) in both participant and item analyses are marked with an asterisk.

Table 8
Results of the ANOVA of Error Rates of Experiment 2

Effect	F1			F2			min F′		
	df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Phrase Type	1,15	157.52	<.001	1,92	138.90	<.001	1,56	73.81	<.001
Cognate Status	1,15	.30	.59	1,92	.17	.68	1,76	.11	.74
Gender Comp.	1,15	85.71	<.001	1,92	24.68	<.001	1,101	19.16	<.001
Phrase Type x Cognate Status	1,15	1.85	.19	1,92	.80	.37	1,88	.56	.46
Phrase Type × Gender Comp.*	1,15	118.40	<.001	1,92	57.13	<.001	1,83	38.53	<.001
Cognate Status × Gender Comp.	1,15	13.73	<.01	1,92	2.22	.14	1,107	1.91	.17
Phrase Type × CognateStatus × Gender Comp.*	1,15	25.64	<.001	1,92	6.56	<.01	1,104	5.22	<.05
Cognates only									
Phrase Type*	1,15	193.70	<.001	1,46	72.37	<.001	1,61	52.69	<.001
Gender Comp.*	1,15	98.77	<.001	1,46	17.36	<.001	1,58	14.76	<.001
Phrase Type × Gender Comp.*	1,15	119.56	<.001	1,46	46.08	<.001	1,61	33.26	<.001
Non-cognates only									
Phrase Type*	1,15	65.80	<.001	1,46	66.71	<.001	1,45	33.13	<.001
Gender Comp.*	1,15	25.35	<.001	1,46	7.58	<.01	1,61	5.84	<.05
Phrase Type × Gender Comp.*	1,15	40.24	<.001	1,46	14.04	<.001	1,61	10.41	<.01

Note: Comp., Compatibility.
Effects that are significant ($p < .05$) in both participant and item analyses are marked with an asterisk.

tions equally. However, since this assumption need not necessarily be true, we also investigated gender errors (i.e., responses with an incorrect determiner) separately. Evidently, this type of error occurred in the noun phrase condition only. Table 9 shows the distribution of these and other types of errors and repeats the difference scores in the total error rates from Table 6 for comparison. As can be seen, there is indeed a considerable similarity between the gender error rates in the noun phrase condition, and the difference scores between noun phrase and bare noun condition in the overall error rates. Moreover, the results of the statistical analysis of the gender errors were basically identical to those of the Phrase Type effects in the overall error analysis reported above.

Offline gender errors

To investigate whether participants, if given enough time, knew the correct gender of the items on which they had produced an error in the picture naming task, an offline gender assignment task on the experimental items had been administered as a last part of Experiment 2. The results are also shown in Table 9. The pattern of results was highly similar to those of the online task, with strong gender compatibility effects (especially for cognates), and with a comparable level of error percentages.

Discussion

The results of Experiment 2 suggest that the two gender systems of L1 and L2 also interact during spoken word production, as evident from the analysis of error rates: More errors, and specifically more gender errors, were produced for gender-incompatible Dutch nouns than for compatible ones. Even though this effect was larger for cognates, it was also significant for non-cognates.

However, the error rates in the production task of Experiment 2 were extremely high (40.5% in the noun phrase condition), so that the number of valid trials that entered the RT analyses was, in some conditions, very low. It is quite plausible that this loss of statistical power was responsible for the failure to find any effects of gender compatibility in the RT analysis. Even though high error rates are not rare for second language tasks that involve the processing of quite sophisticated aspects of the non-native language, such as word gender (for studies obtaining error rates of 40% or more in some conditions in various second language tasks, see, e.g., de Groot, Delmaar, & Lupker, 2000; Dijkstra, Timmermans, & Schriefers, 2000; Holm & Dodd, 1996; Holmes & Dejean de la Bâtie, 1999), in Experiment 3 we attempted to find a means to reduce error rates, to investigate whether the effects observed in the error rates in Experiment 2 would then also occur for RTs.

Furthermore, Experiment 3 was set up such that we could address the issue of the origin of effects of cross-language gender compatibility established in Experiment 2. The high error rates both from the online and the offline task in the incompatible conditions in Experiment 2 suggest that participants were extremely unsure about the gender of items from these word categories. It seems thus plausible that the difficulties with which gender-incompatible nouns are processed arise from problems during L2 gender acquisition, leading to unstable gender representations, rather than from online lexical competition processes between conflicting gender information.

To examine the validity of this account, as well as to reduce error rates, we conducted Experiment 3. Experiment 3 was a replication of Experiment 2, but with a new participant population that was trained on the items' gender beforehand. First, if any potential RT effects were concealed by the high error rates in Experiment 2, the lowering of error rates by pre-experimental training should let these 'true' RT effects emerge. Second, with a training consisting of the participants repeatedly producing the items in gender-marked phrases (and receiving feedback on this performance), a measure became available that indicated the *stability* of the gender representation for a given item. If a noun has a stable lexical gender representation (as it is the case in L1, for example), its gender should reliably be produced correctly in all, or at least the majority, of repetitions during the training. In contrast, highly variable

Table 9
Frequency of Occurrence of Error Types (in % of the absolute number of trials) per Experimental Condition in Main Experiment and Offline Task

		Noun phrase	Bare noun	Difference[a]	Difference in total error score (from Table 6)	Offline gender errors
Cognates						
Gender-compatible	Noun selection	8.6	8.3	0.3		
	Gender	4.9	0	4.9	5.2	6.3
	Other	8.1	8.9	−0.8		
Gender-incompatible	Noun selection	6.8	7.3	−0.5		
	Gender	46.4	0	46.4	46.3	53.7
	Other	9.4	8.1	1.3		
Non-cognates						
Gender-compatible	Noun selection	8.6	9.1	−0.5		
	Gender	9.4	0	9.4	12.0	10.4
	Other	10.4	10.2	0.2		
Gender-incompatible	Noun selection	7.6	7.0	0.6		
	Gender	28.9	0	28.9	32.3	30.7
	Other	13.0	7.3	5.7		

Note: "Other" errors comprise voice key artifacts (e.g., too early ['smacks'] or too late triggering of the voice key), omissions, and disfluencies.
[a] noun phrase minus bare noun condition.

performance across training trials points at unstable representations. Thus, we will analyze the data of the main experiment, that follows the training, not only in terms of general RTs and error rates, but we will also investigate whether representational stability played an additional role in the occurrence of gender compatibility effects.

Experiment 3: Picture naming with previous gender training

Methods

Twenty-two German–Dutch bilinguals, mostly foreign students at Nijmegen University (aged 19–41, mean 24.5, 19 of them female), took part. Due to the difficulty to find another group of highly proficient participants, their proficiency in Dutch was somewhat lower than that of the previous groups (mean score in the vocabulary test: 74% correct; see Table 3). This difference was intended to be compensated by the training (see below) and the larger number of participants (22 vs. 16 in Experiment 2).

During the picture familiarization phase (where each picture was shown together with its assigned name), the participants' gender knowledge was simultaneously assessed by asking them to write down the definite determiner for each noun. Additionally, participants were asked to indicate the certainty of their answer on a scale from 1 ("not certain at all") to 5 ("very certain"). After that, a training phase followed. During this training, participants saw the 96 pictures on the computer screen and had to name them together with their definite determiner. In case of an incorrect response, participants were corrected by the experimenter. This training was repeated three times, yielding three training blocks with feedback. After the training, the main experiment was conducted in the same way as Experiment 2. [7]

Results

Overall

The mean error rate was 14.4%. An additional 2% of RTs was excluded as outliers. Table 10 shows the mean RTs and error rates, and the statistical results of the ANOVA are given in Tables 11 and 12.

First of all, inspection of the means shows that the pre-experimental training was effective in terms of lowering the error rates: Even though the participant group was slightly less proficient than the previous one, their overall error rate in the main experiment was approximately halved compared to Experiment 2. Furthermore, RTs in this experiment were about 100 ms shorter than in the previous experiment.

The analysis of RTs showed that very similarly to Experiment 2, there were significant main effects of Phrase Type and Gender Compatibility. More importantly, Phrase Type significantly interacted with Gender Compatibility. Even though this interaction was not further modified by Cognate Status, separate analyses for cognates and non-cognates revealed that it was primarily carried by cognates, as it was non-significant for non-cognates.

The results for the error rates were essentially parallel to those of RTs. There was an overall Phrase Type by Gender Compatibility interaction, which, this time, was significantly modulated by Cognate Status. The separate analyses for cognates and non-cognates showed that the Phrase Type by Gender Compatibility interaction was highly significant for cognates, whereas it was only significant by participants for non-cognates.

In summary, the overall analysis of Experiment 3 confirmed the supposition that the lack of gender compatibility effects with respect to RTs in Experiment 2 was due to the high rate of errors, that had to be excluded from the RT analysis. After lowering the error rates in Experiment 3, both error rates and RTs showed the expected effects: Gender-marked phrases were produced with more difficulty when the noun had a gender-incompatible translation in German. This time, however, the effect was restricted to cognates.

Stability analysis

In order to examine the role of representational stability of word gender with respect to our observed effects, we analyzed the data of Experiment 3 in more detail. Each noun had been produced together with its gender four times before the actual main experiment (once during the familiarization phase, and three times during training). Thus, the number of times where the definite determiner was correctly produced for a given noun can serve as an independent measure of the 'stability' of the noun's gender representation. For each participant, we used this criterion to apply a median split to each (Cognate Status by Gender Compatibility) item condition, dividing the data into those that had relatively stable representations in a participant, and those that were characterized by highly variable performance already prior to the main experiment (see Appendix D for more details). If ('online') cross-language gender competition was the main factor underlying our gender compatibility effects, this competition should also affect gender-incompatible, but stably represented nouns (and thus be visible in the RT patterns). If, on the other hand, incorrect and unstable representations for (some) incompatible nouns were responsible for our effects, these effects should disappear when looking at items that are represented correctly and in a stable way.

Tables 13 and 14 show the mean RTs and error rates for the groups of 'stable' and 'unstable' items. Because the split procedure categorized the same items differently for different participants, this analysis could only be performed by participants. The statistical results of the separate analyses of stable and unstable items are summarized in Table 15 (for RTs) and Table 16 (for error rates).

In the following summary, we report only the critical Phrase Type by Gender Compatibility interaction (see Tables 15 and 16 for the other results). In the RT analysis, this

[7] List assignment was such that Phrase Type Condition remained counterbalanced, with the same number of participants assigned to each counterbalanced list version; five participants were assigned to each list version of the first randomization, and six to each list version of the second randomization.

Table 10
Mean RTs (ms) and Error rates (%) and Phrase Type Effects in all Item Conditions in Experiment 3

	RTs			Error rates		
	Noun phrase	Bare noun	Phrase Type effect[a]	Noun phrase	Bare noun	Phrase Type effect[a]
Cognates						
Gender-compatible	1050 (152)	906 (103)	144* [56]	9.8 (8.0)	5.7 (6.0)	4.1* [3.1]
Gender-incompatible	1190 (201)	983 (92)	207* [74]	31.6 (18.6)	5.1 (4.4)	26.5* [7.8]
Non-cognates						
Gender-compatible	1078 (136)	970 (100)	108* [40]	17.0 (14.3)	10.6 (9.6)	6.4* [3.9]
Gender-incompatible	1143 (162)	1020 (101)	123* [54]	23.5 (14.8)	11.9 (9.4)	11.6* [5.8]

Note: Standard deviations are given in rounded parentheses, the halfwidth of 95% confidence intervals is given in square parentheses. Phrase Type effects that are significant with $p < .05$ are marked with an asterisk.

Table 11
Results of the ANOVA of RTs of Experiment 3

Effect	*F*1			*F*2			min *F*′		
	df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Phrase Type*	1,21	47.20	<.001	1,92	144.95	<.001	1,36	35.61	<.001
Cognate Status	1,21	3.08	.09	1,92	.53	.47	1,112	.45	.50
Gender Comp.*	1,21	68.63	<.001	1,92	13.73	<.001	1,113	11.44	<.001
Phrase Type × Cognate Status*	1,21	7.13	<.05	1,92	7.59	<.01	1,66	3.68	.06
Phrase Type × Gender Comp.*	1,21	9.55	<.01	1,92	4.04	<.05	1,104	2.84	.10
Cognate Status × Gender Comp.	1,21	3.36	.08	1,92	.75	.39	1,113	.61	.44
Phrase Type × Cognate Status × Gender Comp.	1,21	1.07	.31	1,92	1.27	.26	1,61	.58	.45
Cognates only									
Phrase Type*	1,21	39.22	<.001	1,46	117.89	<.001	1,36	29.43	<.001
Gender Comp.*	1,21	39.94	<.001	1,46	7.97	<.01	1,61	6.64	<.05
Phrase Type × Gender Comp.*	1,21	5.01	<.05	1,46	5.29	<.05	1,56	2.57	.11
Non-cognates only									
Phrase Type*	1,21	36.73	<.001	1,46	40.22	<.001	1,56	19.20	<.001
Gender Comp.*	1,21	11.34	<.01	1,46	5.84	<.05	1,67	3.85	.05
Phrase Type × Gender Comp.	1,21	.31	.59	1,46	.37	.55	1,54	.17	.68

Note: Comp., Compatibility.
Effects that are significant ($p < .05$) in both participant and item analyses are marked with an asterisk.

Table 12
Results of the ANOVA of Error Rates of Experiment 3

Effect	*F*1			*F*2			min *F*′		
	df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Phrase Type*	1,21	37.20	<.001	1,92	87.80	<.001	1,41	26.13	<.001
Cognate Status	1,21	2.61	.12	1,92	1.48	.23	1,94	.94	.33
Gender Comp.*	1,21	50.67	<.001	1,92	10.66	<.01	1,113	8.81	<.01
Phrase Type × Cognate Status*	1,21	16.42	<.001	1,92	5.97	<.05	1,108	4.38	<.05
Phrase Type × Gender Comp.*	1,21	28.62	<.001	1,92	27.95	<.001	1,69	14.14	<.001
Cognate Status × Gender Comp.	1,21	7.00	<.05	1,92	2.30	.13	1,110	1.73	.19
Phrase Type × CognateStatus × Gender Comp.*	1,21	29.88	<.001	1,92	11.01	<.001	1,108	8.05	<.01
Cognates only									
Phrase Type*	1,21	51.86	<.001	1,46	64.96	<.001	1,53	28.84	<.001
Gender Comp.*	1,21	37.21	<.001	1,46	12.90	<.001	1,66	9.58	<.01
Phrase Type × Gender Comp.*	1,21	35.43	<.001	1,46	34.47	<.001	1,58	17.47	<.001
Non-cognates only									
Phrase Type*	1,21	17.41	<.001	1,46	25.91	<.001	1,49	10.41	<.01
Gender Comp.	1,21	6.35	<.05	1,46	1.38	.25	1,62	1.13	.29
Phrase Type × Gender Comp.	1,21	6.45	<.05	1,46	2.09	.16	1,66	1.58	.21

Note: Comp., Compatibility.
Effects that are significant ($p < .05$) in both participant and item analyses are marked with an asterisk.

interaction was not significant for 'stable' items, but it was for 'unstable' ones. The triple interaction with Cognate Status was not significant in either case. When analyzing cognates and non-cognates separately, the Phrase Type by Gender Compatibility interaction remained (almost) significant for both cognates and non-cognates in the case of

Table 13
Mean RTs for 'Stable' and 'Unstable' items in Experiment 3

	Nouns with few pre-experiment errors ('stable' items)			Nouns with many pre-experiment errors ('unstable' items)		
	Noun phrase	Bare noun	Phrase Type effect[a]	Noun phrase	Bare noun	Phrase Type effect[a]
Cognates						
Gender-compatible	987 (163)	896 (126)	91* [52]	1151 (242)	974 (140)	177* [109]
Gender-incompatible	1124 (197)	988 (139)	136* [63]	1318 (258)	972 (95)	346* [108]
Non-cognates						
Gender-compatible	1025 (134)	970 (147)	55* [53]	1173 (184)	1004 (140)	169* [69]
Gender-incompatible	1068 (181)	1024 (143)	44 [86]	1268 (204)	1033 (158)	235* [50]

Note: Standard deviations are given in rounded parentheses, the halfwidth of 95% confidence intervals is given in square parentheses. Phrase Type effects that are significant with $p < .05$ are marked with an asterisk.
[a] Noun phrase minus bare noun condition.

Table 14
Mean error rates for 'Stable' and 'Unstable' items in Experiment 3

	Nouns with few pre-experiment errors ('stable' items)			Nouns with many pre-experiment errors ('unstable' items)		
	Noun phrase	Bare noun	Phrase Type effect[a]	Noun phrase	Bare noun	Phrase Type effect[a]
Cognates						
Gender-compatible	4.0 (7.1)	3.1 (5.0)	0.9 [2.6]	18.2 (14.9)	10.0 (11.3)	8.2* [6.1]
Gender-incompatible	18.4 (20.1)	4.7 (4.5)	13.7* [8.8]	45.9 (19.9)	5.4 (6.6)	40.5* [8.5]
Non-cognates						
Gender-compatible	9.3 (11.0)	6.6 (9.1)	2.7 [3.9]	26.8 (21.7)	15.0 (14.6)	11.8* [7.6]
Gender-incompatible	12.9 (14.8)	9.4 (7.9)	3.5 [6.1]	34.4 (19.1)	14.6 (14.7)	19.8* [8.1]

Note: Standard deviations are given in rounded parentheses, the halfwidth of 95% confidence intervals is given in square parentheses. Phrase Type effects that are significant with $p < .05$ are marked with an asterisk.
[a] Noun phrase minus bare noun condition.

Table 15
Results of the ANOVA of RTs for Stable and Unstable items in Experiment 3

Effect	'Stable' items			'Unstable' items		
	df	*F*	*p*	*df*	*F*	*p*
Phrase Type	1,21	11.33	<.01	1,21	68.95	<.001
Cognate Status	1,21	2.33	.14	1,21	.46	.50
Gender Comp.	1,21	22.53	<.001	1,21	9.23	<.01
Phrase Type × Cognate Status	1,21	9.69	<.01	1,21	3.71	.07
Phrase Type × Gender Comp.	1,21	.53	.47	1,21	14.69	<.001
Cognate Status × Gender Comp.	1,21	3.36	.08	1,21	.21	.65
Phrase Type × Cognate Status × Gender Comp.	1,21	1.29	.27	1,21	1.62	.28
Cognates only						
Phrase Type	1,21	20.54	<.001	1,21	46.83	<.001
Gender Comp.	1,21	25.45	<.001	1,21	5.99	<.05
Phrase Type × Gender Comp.	1,21	3.51	.08	1,21	5.63	<.05
Non-cognates only						
Phrase Type	1,21	3.24	.09	1,21	70.36	<.001
Gender Comp.	1,21	3.29	.08	1,21	3.93	.06
Phrase Type × Gender Comp.	1,21	.06	.81	1,21	4.18	.05

Note: Comp. = Compatibility.

Table 16
Results of the ANOVA of error rates for Stable and Unstable items in Experiment 3

Effect	'Stable' items			'Unstable' items		
	df	*F*	*p*	*df*	*F*	*p*
Phrase Type	1,21	6.26	<.05	1,21	68.97	<.001
Cognate Status	1,21	2.14	.16	1,21	1.31	.27
Gender Comp.	1,21	12.52	<.01	1,21	22.86	<.001
Phrase Type × Cognate Status	1,21	3.34	.08	1,21	5.94	<.05
Phrase Type × Gender Comp.	1,21	6.47	<.05	1,21	34.31	<.001
Cognate Status × Gender Comp.	1,21	2.32	.14	1,21	4.58	<.05
Phrase Type × Cognate Status × Gender Comp.	1,21	13.00	<.01	1,21	23.80	<.001
Cognates only						
Phrase Type	1,21	8.84	<.01	1,21	83.21	<.001
Gender Comp.	1,21	19.57	<.001	1,21	17.37	<.001
Phrase Type × Gender Comp.	1,21	11.16	<.01	1,21	46.54	<.001
Non-cognates only						
Phrase Type	1,21	2.01	.17	1,21	23.12	<.001
Gender Comp.	1,21	1.60	.22	1,21	3.00	.10
Phrase Type × Gender Comp.	1,21	.10	.75	1,21	4.70	<.05

Note: Comp., Compatibility.

unstable items, while it was not significant for stable items (neither cognates nor non-cognates).

In the stability analysis of error rates, a slightly different pattern emerged: Here, the interaction of Phrase Type and Gender Compatibility reached significance not only for the unstable, but also for stable items. In both cases, this interaction was further modified by Cognate Status. The separate analyses for cognates and non-cognates revealed that Phrase Type interacted with Gender Compatibility for unstable cognates, unstable non-cognates, and stable cognates, but not for stable non-cognates.

Discussion

Experiment 3 was a replication of Experiment 2, except that participants were trained on the gender of the experimental items beforehand. This modification allowed us to obtain reliable RT data in addition to the error data of Experiment 2, and it enabled us to investigate the origin of gender compatibility effects.

First, error rates were effectively reduced in Experiment 3 compared to Experiment 2, and this reduction of errors led to the expected RT effects: Compared to the gender-neutral bare noun baseline, gender-marked phrases containing nouns with a gender-compatible German translation were not only produced more accurately (in both experiments), but also faster (in Experiment 3) than gender-incompatible noun phrases. Thus, both Experiment 2 and 3 provide evidence for the production of gender-marked Dutch (L2) noun phrases being influenced by the gender-compatibility with the German (L1) translation of the noun. Additionally, in both experiments, this effect of Gender Compatibility was modulated by Cognate Status, as it was larger for cognates than for non-cognates.

Our data are in line with those of Salamoura and Williams (2007), who demonstrated effects of cross-language gender compatibility in a translation task from Greek (L1) to German (L2). In that study, the effect was not significantly modulated by cognate status, but it was descriptively larger for cognates as well, pointing in the same direction as our results, indicating that our findings might be generalizable to different tasks and language combinations.

In contrast, at first glance, our results appear to be at variance with those by Costa et al. (2003), who did not find an influence of word gender in L1 on the production of gender-marked noun phrases in L2 in either RTs or error rates. However, the stability analysis as well as several important differences between the study of Costa et al. and ours might help to explain the difference in results. First, as pointed out by Salamoura and Williams (2007), the target languages used in Costa et al.'s study (Italian, Spanish, Catalan and French) are Romance languages that have special properties with respect to word gender. These properties lead to, for instance, a missing (monolingual) gender congruency effect in these languages (Alario & Caramazza 2002; Miozzo & Caramazza, 1999), which might explain why Costa et al. did not find *cross-language* gender congruency effects in these languages either. As a second difference, Costa et al. did not systematically manipulate cognate status (but rather collapsed cognates and non-cognates in one analysis), and their materials included fewer cognates (up to 30%). Given that the effects in the present study were primarily carried by cognates, the higher percentage of cognates (50%) in our experiments probably strengthened the co-activation of the native language (but note that because of the high level of similarity of Dutch and German, such a high percentage of cognates is not 'unnatural' or disproportionate). Finally, the low error rates in Costa et al.'s study (up to 10%) compared to the present experiments (up to 60% in Experiment 2 and 30% in Experiment 3) suggest that L2 gender representations might have been more stable in Costa et al.'s participants (possibly due to overall proficiency, or to a more transparent gender system) than in ours. In the light of what we now know about stability, this is likely to have led to smaller gender compatibility effects.

The additional analysis of the data with respect to the stability of individual gender representations, as measured by the level of performance in the pre-experimental training, indeed suggested that this factor plays a decisive role. The observed gender compatibility effects turned out to be primarily carried by nouns (particularly cognate nouns) with unstable gender representations, especially where RTs are concerned. For these unstable nouns, the effect was robust even for non-cognates. The effect for nouns with relatively stable gender representations was much weaker and restricted to cognates (and reached statistical significance only in error rates).

The implications of these results for a theoretical account of cross-language effects of gender compatibility will be discussed in General discussion.

General discussion

The present study explored fairly unknown territory, both because of the unsettled issue of cross-language interactions of word gender in bilingual language processing, but also due to the direct comparison of word recognition and production. At present, bridging the gap between these two domains of language research is probably one of the major challenges of psycholinguistics (e.g., Schiller & Meyer, 2003).

The present study was set up with two goals in mind. First, given the small number and equivocal set of results in previous studies, we intended to examine whether effects of cross-language gender compatibility could be established for an identical set of L2 nouns in both word recognition and production. Second, in case these effects would indeed occur, we aimed at identifying their origin, in particular by distinguishing an 'online' account from an acquisition-based one.

In both word comprehension (Experiment 1) and production (Experiments 2 and 3) in L2, we did indeed find robust effects of gender compatibility with L1: gender-marked phrases were processed with more difficulty when the gender of the Dutch nouns was incompatible with their German translation. This was particularly true for cognates.

These results could fairly easily be accommodated within a 'classical' model of lexical processing, where word gender is represented as an abstract lexical feature of nouns, e.g., in the form of 'gender nodes' (Levelt, Roelofs, & Meyer, 1999). Adapted to the bilingual situation, the remaining question would be whether these gender representations are separate or shared between languages. Cross-language effects of word gender as found in the present study would then point at a shared gender system (Salamoura & Williams, 2007), while the absence of such effects would be more in line with two separate and independent systems of gender representation (Costa et al., 2003).

When assuming this account, our results would support the first view: When an L2 noun has to be processed together with its gender, either in recognition or in production, its L1 translation equivalent and its gender become active as well. In case of gender incompatibility, the two conflicting gender representations will subsequently compete with each other, hampering further processing. The co-activation of the translation equivalent (and its gender) is larger for cognates than for non-cognates, which is in line with findings showing that due to their large form overlap, cognate translations are co-activated to a larger degree than non-cognates, both during L2 word production (Christoffels, Firk, & Schiller, 2007; Costa et al., 2000) and word recognition (e.g., Cristoffanini et al., 1986; Sánchez Casas, Davis, & García Albea, 1992).

However, when looking at our stability analysis, applying 'classical' models of monolingual language processing in such a simple and straightforward way does not seem justified. Bilinguals, especially unbalanced bilinguals who have acquired their L2 as adults, fundamentally differ from monolinguals in important respects. Adult speakers usually possess *stable and correct* lexical representations in their L1, with, under natural circumstances, constant and almost error-free language performance. Therefore, getting to know something about the native language system requires its experimental manipulation and sometimes 'destabilization' by external influences, such as distractor words, primes, time pressure, etc., causing variations in reaction times (and, to a lesser degree, error rates) from which theoretical conclusions can be drawn. In contrast, L2 speakers' performance is highly variable and only rarely error-free, even outside the experimental setting. This is true even for highly experienced L2 speakers, as only a small minority of adult L2 learners ever reaches native-like proficiency (Birdsong & Molis, 2001; Johnson & Newport, 1989). As word gender is presumably one of the domains that is most difficult to learn in a language with an opaque gender system like Dutch, Dutch learners' lexical representations of gender are often weak, unstable or even incorrect. In consequence, rather than reflecting 'online' competition processes in an otherwise stable lexical network, the increased RTs and error rates for gender-incompatible Dutch nouns in our experiments might be due to unstable and incorrect gender representations in the L2 lexicon (or, in other words, to imperfect gender *acquisition*).

The additional analysis of Experiment 3 generally supports this hypothesis for word production. The effects of gender compatibility either disappeared (concerning RTs) or became substantially weaker (concerning error rates) for items with a relatively 'stable' gender representation. Thus, the primary mechanism causing the effects found in the present experiments seems to be an increased difficulty to *acquire* correct (and stable) gender representations. This difficulty affects those L2 nouns which possess a different gender in L1, and even more so, when they are also similar in form to their translation (cognates). Apparently, native speakers of German learning Dutch base the acquisition of the L2 gender system on their native language, which is, on the whole, a fairly successful strategy. However, this strategy adversely affects those words for which there is no cross-language gender compatibility. A part of these nouns will either be incorrectly represented with the L1-compatible gender in L2, or the links to their genders will be weak and unstable, resulting in variable outcomes when production of a gender-marked element is required. In our data, the imperfect representation of these 'unstable' nouns is reflected both in high error rates during noun phrase production, and in higher RTs for correctly produced gender-incompatible nouns.

That the gender compatibility effects observed here were substantially stronger for 'unstable' gender representations does of course not rule out that 'online' competition between two conflicting gender representations might also exist. In fact, when looking at cognates only, there was an effect in the error rates even for the 'stable' group, and a trend for the same items in the RTs. However, it should be noted that because the stability split was a relative one, the 'stable' group might still have been fairly unstable, especially in the most difficult item category, that of incompatible cognates. Independent of this question, our data show that first, in contrast to L1, gender representations in L2 are often incorrect and unstable; and second, the greater this instability, the greater the effect of cross-language gender compatibility is.

Incorrect and unstable noun-to-gender links might (partly) also have been responsible for the results of Experiment 1, that is, for the longer response times for nouns preceded by, in a 'German' sense, incongruent gender primes. However, it does not seem possible to carry out a stability analysis for this experiment as well, in a similar way as we did for word production. The crucial point here is that the task dimension in the lexical decision experiment (word/nonword) is different from the dimension of interest (word gender), and that consecutively assessing the participants' performance on both dimensions cannot be accomplished without one influencing the other. For instance, suppose that the stability of the word gender representations would be assessed *before* the lexical decision experiment, e.g., by asking participants to repeatedly carry out some form of gender assignment for the critical nouns. Obviously, only words could be included in this task. However, the prior presentation of words, but not nonwords would turn the lexical decision task into an episodic memory task rather than a word recognition task. On the other hand, the possibility of measuring word gender stability *after* the lexical decision task is ruled out by the fact that all nouns are presented together with their *correct* gender-marked determiners during the lexical decision task (and, as mentioned before, there are good reasons to include only correct phrases). This would very likely influence the subsequent stability measurement.[8]

The similar results of Salamoura and Williams (2007) with respect to word production involving Greek and German, and of Paris and Weber (2004) on word recognition in French-German bilinguals indicate that our findings may not be limited to highly related languages such as Dutch and German. Rather, one important factor might be the opaque nature of the gender system in the second language: In absence of reliable form-related cues (i.e., word

[8] One could argue that the problem of measuring stability in visual word recognition could be solved by replacing the lexical decision task by a gender decision task. However, in our view, gender decision is a highly unnatural task focusing the participants' attention on word gender in a way that does not normally occur during visual word recognition.

endings as in Spanish or Italian) for word gender (like in German or Dutch), the learner tends to use L1 gender information, regardless of how closely L1 is related to L2. By contrast, when easy-to-learn rules govern the assignment of grammatical gender, L1 influences might be overruled, or might not even arise in the first place. Further research is needed to confirm this assumption and to explore the use of gender cues in L2 in general (see also Bordag, Opitz, & Pechmann, 2006, for a demonstration of an *increased* use of gender cues by L2 learners of German as opposed to native speakers).

In summary, the present study shows that German–Dutch bilinguals are influenced by word gender in L1 when processing gender-marked noun phrases in L2. Even though, at first sight, it seems straightforward to interpret this effect as a consequence of a shared gender system in a bilingual lexicon that otherwise resembles the monolingual one, such an interpretation does not appear to be correct. Rather, additional data obtained in Experiment 3 indicate that the cross-language gender compatibility effect was mainly due to participants 'not knowing for sure' which gender some nouns have, i.e., that they were primarily a consequence of imperfect gender acquisition in L2. This imperfect acquisition is heavily biased by L1, resulting in processing difficulties only for those Dutch nouns (and particularly cognates) that are incompatible with their German gender, both during recognizing and producing gender-marked noun phrases in L2. More generally, the present study shows that monolingual concepts of syntactic processing are not always appropriate for bilinguals, and new psycholinguistic models and approaches have to be developed that take the unstable, variable and 'probabilistic' (Birdsong, 2004) nature of L2 representations into account.

Acknowledgments

A part of the work on the manuscript was carried out while Kristin Lemhöfer was at the Max Planck Institute for Psycholinguistics, Nijmegen. Katharina Spalek is now at the Department of German Language and Linguistics, Humboldt University, Berlin. We thank Iris Hanique for carrying out Experiment 3 as part of her Bachelor thesis. Many thanks to Ton Dijkstra who provided valuable suggestions and comments during various stages of this research, and to Wido La Heij who commented on earlier versions of the manuscript.

Appendix A. Language Background of Participants of Experiments 1, 2, and 3 as Reported in the Language Questionnaire

	Experiment 1	Experiment 2	Experiment 3
Mean number of years of experience with Dutch	5.4 (2.9)	7.3 (4.9)	4.4 (4.8)
Self-ratings:			
Mean frequency of reading literature in Dutch	4.7 (1.8)	4.8 (1.5)	5.1 (1.5)
Mean frequency of speaking Dutch	6.2 (1.4)	6.1 (1.0)	5.5 (1.2)
Mean self-rated reading experience in Dutch	5.3 (1.2)	5.3 (1.1)	4.9 (1.5)
Mean self-rated writing experience in Dutch	4.9 (1.6)	5.1 (1.1)	4.6 (1.3)
Mean self-rated speaking experience in Dutch	6.1 (1.3)	6.0 (0.7)	5.1 (1.4)

Note. Standard deviations are given in parentheses. Self-ratings are measured on a scale from 1 (low) to 7 (high).

Appendix B. Stimulus Materials in the Main Experiments 1, 2, and 3

Words

For each of the 24 test items in a condition, the following information is given: *Dutch word*, Dutch definite determiner, German translation, German definite determiner, English translation.

Dutch-German cognates, congruent gender

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.

Dutch–German cognates, incongruent gender

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.

Dutch–German non-cognates, congruent gender

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.

Dutch–German non-cognates, incongruent gender

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.

Nonwords (Used in Experiment 1 only)

Nonwords used with the definite determiner 'de'

baag, boop, borie, brapel, brimte, fleus, foop, gemise, giemerij, gof, gorm, groeder, grofine, halmoer, holk, kaneur, kinker, kolie, korvel, kreuker, lan, lapiek, maap, machtade, mieg, mool, moom, morant, nergel, nug, pergist, pialing, pistoop, poei, pors, slieg, soem, sterator, strief, stulerij, vazel, vlietage, vloop, voeve, zaas, zeuk, zil, zorm.

Nonwords used with the definite determiner 'het'

aril, aspitaat, begel, blankenis, brous, bruur, dijf, dink, dorief, drof, dron, foriet, gedeik, getonkel, gittel, grokbord, hantuig, heken, jal, jonief, kantuik, kert, marake, melaat, merp, minul, mog, moreld, nerk, nijk, pakreel, palken, pandsel, peiniek, rapa, schak, tanaat, taneet, uim, vek, vinton, vlein, weeld, wooi, zensel, ziep, zord, zwaloon.

Appendix C. Words and Nonwords used in the Dutch Vocabulary test

Words

acteur, affiniteit, avonturier, bretel, chagrijnig, doop, doornat, dronkenschap, exploitatie, fornuis, geloei, gelovig, geraakt, getint, hengel, kazerne, knullig, laakbaar, martelaar, matig, mikken, nopen, normatief, onbekwaam, onledig, paars, paviljoen, publiekelijk, retorisch, riant, romig, rups, slaags, stagnatie, toetsing, verguld, verspilling, voornemen, woelig, zetelen.

Nonwords

aanhekking, compromeet, etaal, flajoen, futeur, haperie, joutbaag, klengel, kluiper, leurig, maliteit, markatief, ontpelen, proom, speven, starkatie, streuren, vertediseren, vlut, zolf.

Appendix D. Details of the Stability Analysis of Experiment 3

The criterion for the median split was the total number of correct responses for each item in the familiarization and training phase (0-4). As a second criterion (when several items had the same number of correct responses score), the certainty rating in the familiarization phase was used (with which participants had indicated the certainty of their gender assignment). When the gender response given during familiarization was incorrect, the rating was not used; when it was correct, it was used as a subordinate criterion (with items with high certainty ratings being regarded as more 'stable' than those with low ratings).

Because of the smaller cell sizes after applying the median split, errors were now excluded pairwise for the RT analysis (i.e., in the noun phrase and in the bare noun condition), to be sure that the same items per participant contributed to both conditions. This led to the exclusion of 24% of the data. Outliers were not excluded, as this would lead to an even higher exclusion rate.

A median split was used in each (participant by Cognate Status by Compatibility) cell. Items in the middle of the criterion distribution that were indistinguishable from each other were assigned to the "stable" group. This led to a distribution of 43.8% "unstable" and 56.2% "stable" cases in the total number of (3210) correct trials used for the RT analysis, and of 44.6 vs. 55.4% of all 4224 trials in the error analysis.

References

Alario, F.-X., & Caramazza, A. (2002). The production of determiners: Evidence from French. *Cognition, 82*(3), 179–223.

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

Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in Italian. *Perception & Psychophysics, 58*(7), 992–1004.

Birdsong, D. (2004). Second language acquisition and ultimate attainment. In A. Davies & C. Elder (Eds.), *Handbook of Applied Linguistics* (pp. 82–105). London: Blackwell.

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language, 44*(2), 235–249.

Bölte, J., & Connine, C. M. (2004). Grammatical gender in spoken word recognition in German. *Perception & Psychophysics, 66*(6), 1018–1032.

Bordag, D., Opitz, A., & Pechmann, T. (2006). Gender processing in first and second languages: The role of noun termination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1090–1101.

Caramazza, A., & Brones, I. (1979). Lexical access in bilinguals. *Bulletin of the Psychonomic Society, 13*(4), 212–214.

Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research, 1147*, 192–208.

Costa, A., Caramazza, A., & Sebastián-Gallés, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1283–1296.

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*(3), 181–200.

Costa, A., & Santesteban, M. (2004). Bilingual word perception and production: Two sides of the same coin? *Trends in Cognitive Sciences, 8*(6), 253.

Cristoffanini, P., Kirsner, K., & Milech, D. (1986). Bilingual lexical representation: The status of Spanish–English cognates. *Quarterly Journal of Experimental Psychology, 38A*(3), 367–393.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language, 42*(4), 465–480.

de Groot, A. M. B., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language, 47*(1), 91–124.

de Groot, A. M. B., Delmaar, P., & Lupker, S. J. (2000). The processing of interlexical homographs in translation recognition and lexical decision: Support for nonselective access to bilingual memory. *Quarterly Journal of Experimental Psychology, 53A*(2), 397–428.

de Groot, A. M. B., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language, 30*(1), 90–123.

Dewaele, J.-M., & Véronique, D. (2001). Gender assignment and gender agreement in advanced French interlanguage: A cross-sectional study. *Bilingualism: Language and Cognition, 4*(3), 275–297.

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41*(4), 496–518.

Dijkstra, T., Timmermans, M., & Schriefers, H. (2000). On being blinded by your other language: Effects of task demands on interlingual homograph recognition. *Journal of Memory and Language, 42*, 445–464.

Dijkstra, T., van Jaarsveld, H., & ten Brinke, S. (1998). Interlingual homograph recognition: Effects of task demands and language intermixing. *Bilingualism: Language and Cognition, 1*(1), 51–66.

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: Masked priming with cognates and noncognates in Hebrew–English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1122–1139.

Grosjean, F., Dommergues, J.-Y., Cornu, E., Guillelmon, D., & Besson, C. (1994). The gender-marking effect in spoken word recognition. *Perception & Psychophysics, 56*(5), 590–598.

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory & Cognition, 29*(3), 503–511.

Gurjanov, M., Lukatela, G., Lukatela, K., Savic, M., & Turvey, M. (1985). Grammatical priming of inflected nouns by the gender of possessive adjectives. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 692–701.

Holm, A., & Dodd, B. (1996). The effect of first written language on the acquisition of English literacy. *Cognition, 59*(2), 119–147.

Holmes, V. M., & Dejean de la Bâtie, B. (1999). Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics, 20*(4), 479–506.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*(1), 60–99.

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., Dijkstra, T., & Michel, M. C. (2004). Three languages, one ECHO: Cognate effects in trilingual word recognition. *Language and Cognitive Processes, 19*(5), 585–611.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A mega-study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 12–31.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–75.

Meara, P. M. (1996). *English Vocabulary Tests: 10k. Unpublished manuscript. Swansea: Center for Applied Language Studies.*

Miozzo, M., & Caramazza, A. (1999). The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 907–922.

Paris, G., Weber, A. (2004). The role of gender information in spoken-word recognition in a non-native language. Paper presented at the *AMLaP (10th annual conference on architecture and mechanisms of language processing), Aix en Provence, France.*

Rogers, M. (1987). Learners' difficulties with grammatical gender in German as a foreign language. *Applied Linguistics, 8*(1), 48–74.

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*(3), 257–275.

Sánchez Casas, R. M., Davis, C. W., & García Albea, J. E. (1992). Bilingual lexical processing: Exploring the cognate/non-cognate distinction. *European Journal of Cognitive Psychology, 4*(4), 293–310.

Schiller, N. O., & Meyer, A. S. (Eds.). (2003). *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Berlin: Mouton e Gruyter.

Scherag, A., Demuth, L., Rösler, F., Neville, H. J., & Röder, B. (2004). The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects. *Cognition, 93*(3), B97–B108.

Schmidt, R. (1986). Was weiss der Artikel vom Hauptwort? Ein Beitrag zur Verarbeitung syntaktischer Beziehungen beim Lesen [What does the article know about the noun?]. A contribution to the processing of syntactical relations in reading. *Zeitschrift für Experimentelle und Angewandte Psychologie, 33*(1), 150–163.

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 841–850.

Schriefers, H., & Teruel, E. (2000). Grammatical gender in noun phrase production: The gender interference effect in German. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1368–1377.

van Berkum, J. J. A. (1996). *The psycholinguistics of grammatical gender*. The Netherlands: University of Nijmegen, Nijmegen.

van Hell, J. G., & de Groot, A. M. B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1*(3), 193–211.