





ARTICLE

Animacy effects in the English genitive alternation: comparing native speakers and EFL learner judgments with corpus data

Tanguy Dubois¹ , Jason Grafmiller² , Magali Paquot^{3,4}  and Benedikt Szmrecsanyi⁵ 

¹Institut Langage et Communication, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium; Department of Linguistics, KU Leuven, Leuven, Belgium; ²Department of English Language and Linguistics, University of Birmingham, Birmingham, UK; ³Institut Langage et Communication, Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium; ⁴Fonds de la Recherche Scientifique, Brussels, Belgium; ⁵Department of Linguistics, KU Leuven, Leuven, Belgium

Corresponding author: Tanguy Dubois; Email: tanguy.dubois@uclouvain.be

(Received 01 June 2023; Revised 21 September 2023; Accepted 22 September 2023)

Abstract

Recent years have seen a heightened interest in the interface between language use and cognition in language learners. In this study, we investigate this interface further by conducting a rating task experiment on the intuitions of 25 native speakers and 101 low–intermediate to advanced learners of English as a Foreign Language regarding the acceptability of the genitive variants (*the beauty of nature/nature's beauty*) in different contexts. These ratings were then compared against existing corpus-based statistical models that predict which variant is most likely in spoken language use with two mixed-effects linear regression models. The first model focused on the animacy of the possessor in particular, which has been found to have a different effect on native speakers and EFL learners in language use, whereas the second model tested how the ratings relate to the predictions as a whole. Results show that there is a larger discrepancy between language use and intuitions of low-proficiency learners compared to native speakers, which is partially because animate, collective, and inanimate possessors affect the intuitions and the language use of learners differently.

Keywords: rating task experiment; EFL learners; genitive alternation; usage-based SLA

1. Introduction

Recently, usage-based approaches to Second Language Acquisition (SLA) have investigated how learners of English as a Foreign Language (EFL) use the structural variants of alternation phenomena compared to native speakers (e.g. Jäschke & Plag, 2016; Kinne, 2020; Wulff & Gries, 2019). Usage-based approaches to SLA assume that language learning is driven by language experience (Ellis & Wulff, 2019, p. 41), and



the same applies to the acquisition of the probabilistic constraints on the choice between alternating variants (Wulff & Gries, 2019, p. 880). This raises the question of how the probabilistic grammar of EFL learners develops as they become more proficient. With this question in mind, Dubois et al. (2023) investigated how low–intermediate to advanced EFL learners from various L1 backgrounds (e.g. Spanish, Italian, Hindi, Chinese, and Russian) choose between the *s*-genitive (1) and the *of*-genitive (2).

- (1) the people's_{possessor} mentality_{possessum} (TLC:2_IN_11)
- (2) the mentality_{possessum} of the people_{possessor} (TLC:2_6_ME_69)

On the methodological plane, they collected 2302 genitive observations in the Trinity Lancaster Corpus (TLC, Gablasova et al., 2019), a three-million-word corpus consisting of transcribed recordings from an official spoken language examination, namely the Graded Examinations in Spoken English (GESE). They annotated the observations for the probabilistic constraints that drive the choice of genitive variant for native speakers and then analyzed whether the learners differ from the native speakers with regard to the effect of any of the constraints using mixed-effects logistic regression (e.g. Gries, 2015). In line with previous research (Gries & Wulff, 2013), they found that learners and native speakers are remarkably similar: both are sensitive to structural persistence (e.g. Bock, 1982; Szmrecsanyi, 2006), the length of the possessor and the possessum, the final sibilancy of the possessor, and the lexical density of the surrounding context (Dubois et al., 2023, p. 443).

Both native speakers and learners are also sensitive to possessor animacy, which captures the tendency for animate possessors to favour the *s*-genitive and inanimate possessors to prefer the *of*-genitive. One explanation for the effect of the animacy constraint is that speakers prefer to first produce linguistic elements that are easy to process (MacDonald, 2013). Animate referents, which are easy to retrieve from memory (Branigan et al., 2008), would therefore be produced early, which means that an animate possessor favours the *s*-genitive where the possessor comes first. While animacy is normally the strongest constraint of the genitive alternation (Rosenbach, 2014), Dubois et al. (2023) found that B1 speakers are more likely than native speakers to use the *s*-genitive when the possessor is inanimate, whereas B2 speakers are more likely than native speakers to use the *of*-genitive when the possessor is animate (Dubois et al., 2023). This finding is reminiscent of other studies on the genitive in World Englishes and the development of the English genitive over time where the strength of the animacy constraint was found to fluctuate (Heller et al., 2017; Wolk et al., 2013).

Although Dubois et al. (2023) and similar studies aim to shed light on the implicit grammatical knowledge of learners regarding constraints, there are fundamental differences between how statistical models and humans gather knowledge about distributional patterns. For example, regression models like the one used by Dubois et al. (2023) have difficulty dealing with multicollinearity, yet collinearity and redundancy are an integral part of language (Milin et al., 2016, p. 508). In general, corpus-based models are able to “reflect the characteristics of mental processes and structures yielding usage, even though we do not know the exact form of these mental representations” (Divjak & Arppe, 2013, p. 230). Put simply, corpus-based models show what speakers say, but not necessarily what they know (Ellis, 2017, p. 46; Kinne, 2020, p. 173). Applied to the genitive alternation, this means that it is unclear whether the probabilistic constraints that affect language use according to the corpus-based regression model actually derive from the implicit grammatical knowledge of

speakers (Klavan & Divjak, 2016) or whether they are simply frequent patterns of language use.

Therefore, we compare the development of EFL learners' patterns of language use according to the regression model of Dubois et al. (2023) against the acceptability judgments of learners at different proficiency levels in a rating task experiment. Whereas experimental data are typically less valid ecologically than corpus data, rating task experiments do not primarily concern language production and instead tap into the intuitions of speakers regarding the acceptability of the variants (Arppe & Järvikivi, 2007, p. 132; Gilquin & Gries, 2009; Klavan & Divjak, 2016, p. 361), which offers a different perspective on the development of probabilistic grammars for EFL learners across proficiency levels. This combination of corpus data and experimental ratings highlights the way in which cognition, operationalized as the speakers' knowledge of probabilistic constraints on variation, comes to the fore in EFL learners at different stages of language learning in "two different language usage situations, namely production and introspection" (Arppe & Järvikivi, 2007, p. 150). In our case, such an experiment would provide complementary evidence for the cognitive reality of the probabilistic constraints guiding the choice of genitive variant for native speakers and EFL learners and shed more light on the finding that learners are less sensitive to the animacy of the possessor while also providing insights on the development of the animacy constraint throughout language learning in general.

Similar experiments have been implemented successfully to complement corpus-based research for the particle placement alternation, the position of the prepositional phrase, the dative alternation, and the future marker alternation in English (e.g. Bresnan & Ford, 2010; Engel et al., 2022; Engel & Szmrecsanyi, 2023; Kinne, 2020). It has also been used to complement corpus-based models with more than two possible variants in other languages, such as Russian (Divjak & Arppe, 2013) and Estonian (Klavan, 2020). While these studies generally find that the results from corpus-based models and the experiments largely converge (Klavan & Divjak, 2016), the predictions from the corpus-based models do not correlate in a linear fashion with the acceptability of the variants according to individual speakers (Arppe & Järvikivi, 2007, pp. 149–150; Divjak et al., 2016, pp. 26–27; Kempen & Harbusch, 2005, p. 330). This is partially attributable to the fact that regression models are largely concerned with how frequently the variants occur in certain contexts. Indeed, the fact that a genitive variant is used very rarely in a certain context does not mean that speakers deem the variant unacceptable (Arppe & Järvikivi, 2007, pp. 149–150; Divjak et al., 2016, pp. 26–27; Kempen & Harbusch, 2005, p. 344).

Against this backdrop, our study investigates the following research questions:

- How does possessor animacy affect the intuitions of EFL learners at different proficiency levels in a rating task experiment compared to correlations between animacy and genitive variant in a corpus of spoken learner language?
- How does the relationship between experimental ratings and corpus data vary by the proficiency level of participants and speakers?

For our first research question, we expect that participants find the *s*-genitive more natural when the possessor is animate, but this tendency should be weaker for B1 and B2 learners compared to native speakers in accordance with the results from the regression model of Dubois et al. (2023), henceforth referred to as the reference corpus

model. This means that the preference for the *s*-genitive with animate possessors rather than inanimate possessors should be less pronounced for B1 and B2 learners compared to native speakers. For our second research question, we expect that the intuitions from the learners in the experiment correspond to the patterns of language use. That is, the ratings from the participants in the experiment should correlate with the predictions from the reference corpus model, showing that the participants are sensitive to probabilistic constraints (Bresnan & Ford, 2010). For example, if the model predicts that the *s*-genitive is more likely in a given context, we expect participants to give higher ratings for the *s*-genitive in that same context as well.

Analysis of the ratings shows that the A2/B1 and B2 learners participating in the experiment are less sensitive to certain animacy distinctions involving animate, collective, and inanimate possessors. Overall, the intuitions from the participants in the experiment correlate with the patterns of language use of the speakers in the corpus, which provides additional evidence that the participants are sensitive to the significant probabilistic constraints included in the reference corpus model. However, this correlation is weaker for A2/B1 and B2 participants, indicating that these learners do not always produce the genitive variant they prefer in the experiment.

The structure of the study is as follows. The methodology including the materials, participants, procedure, and design of the experiment is presented in Section 2. The results from two mixed-effects linear regression models are presented in Section 3 and then discussed in Section 4.

2. Methodology

The rating task experiment was similar to the one used by Bresnan and Ford (2010) and more recently Engel et al. (2022). The participants, who were native speakers and learners of English, were shown an original corpus excerpt with both possible genitive variants. Using a slider bar, participants could indicate which variant they preferred. The position of the slider represented a rating from 0 to 100 for the *s*-genitive (*the film's name* in Figure 1). Higher scores indicated that the *s*-genitive was more natural, with a score of 100 indicating that this was the only acceptable variant in this context. Inversely, a score of 0 indicated that this variant was not acceptable at all compared to the *of*-genitive.

The experiment focused in particular on the effect of possessor animacy on the ratings as B1 and B2 learners are less sensitive to this constraint according to the reference corpus model. Dubois et al. (2023) originally distinguished between five animacy levels based on the coding scheme of Zaenen and colleagues (Zaenen et al., 2004; see also Wolk et al., 2013), namely animate, collective, inanimate, locative, and temporal possessors. Animate possessors “mostly concern humans or entities represented as humans and animals” (e.g. *god's plan*) (Dubois et al., 2023, p. 436).

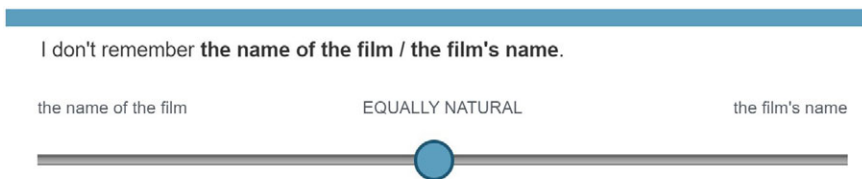


Figure 1. An example of a target experimental item presented to participants.

Collective possessors covered collective nouns such as *government* or *company* and groups of animates who display an identity as a group (e.g. *the values of the parents*). Locatives referred to locations that are fit for humans, as in *the south of France*. Possessors that were marked as temporal refer to a time period or a specific point in time (e.g. *next week*). All other possessors were considered inanimate.

Due to data sparsity issues however, the constraint was conflated into a binary distinction between animate and inanimate possessors in the reference corpus model. As a result, it is not clear whether animacy plays a less important role in the language use of B1 and B2 learners compared to native speakers because of specific animacy levels. In this respect, Dubois et al. (2023) argue that learners struggle with locative and collective possessors in particular because many possessor noun phrases can fit either category depending on the context. In example (3) for instance, copied from example (11) in Dubois et al. (2023), *South Africa* could either refer to a location or a collective group of people.

- (3) and er in May nineteen ninety four for the first time in South Africa's history
all the races voted in democratic election (TLC:2_6_IT_100)

To shed more light on the effect of animacy for our first research question, the rating task experiment was designed to test the effect of all five animacy levels on the ratings. Moreover, the experimental items stem from the 2302 observations Dubois et al. (2023) collected from the Trinity Lancaster Corpus (Gablasova et al., 2019) for the reference corpus model. This is relevant for the second research question because it allows us to compare the predictions from the reference corpus model, which reflect the patterns of language use of the speakers in the corpus, with the ratings of the participants in the experiment representing their intuitions regarding the naturalness of the variants.¹

2.1. Materials

The rating task experiment consisted of target experimental items and items copied from the Oxford Quick Placement Test, a short proficiency test consisting of 60 questions of different types, most of which require the participant to fill in the blank in a sentence or text by selecting the right word among several options (Allan, 2001). The score participants received on the test from 0 to 60 was used to determine their CEFR proficiency level from A1 to C2 (Council of Europe, 2001).

The target experimental items were selected from the observations of the reference corpus model based on a variety of criteria. As we were interested in the effect of possessor animacy on the ratings specifically, we controlled for the other effects that could influence the ratings. According to the reference corpus model, the choice of genitive variant is also influenced by the definiteness and the final sibilancy of the possessor, the length of the possessor and the possessum, the lexical density of the

¹We argue that the TLC data are representative of speakers' language use because the GESE examination is designed to resemble an authentic discussion between the learner and the examiner across four tasks that "represent a variety of settings in which spoken language is used with different degrees of formality, interactiveness, topic familiarity, and different interlocutor roles" (Gablasova et al., 2017, p. 619; see also Dubois et al., 2023).

surrounding context as well as lexical effects. To account for these effects, for the experiment we only considered observations that had definite possessors, no final sibilant, and involved lexical items that exhibited relatively minimal bias for one variant or the other.² Because the length of the constituents was difficult to control for while selecting the target items, we included the constraint as a fixed effect in the linear regression model we used to analyze the rating data (see Section 2.4). In the reference corpus model, the lexical density of the surrounding context was operationalized as the type–token ratio in the 100 words surrounding the genitive observation. Including this much context would make the experiment rather long and possibly disadvantage low-proficiency learners, so the surrounding context was reduced for the experimental items. As a consequence, the predictions from the corpus-based model would no longer be completely accurate for the corresponding corpus observation. It was nonetheless important that the predictions are accurate for our second research question on the correlation between the predictions from the corpus-based model and the ratings. Therefore, we fitted a new mixed-effects logistic regression model on the corpus data without type–token ratio, which provided a new prediction for the *s*-genitive that was based only on information available in the limited context of the experimental items. This model had an R^2_{marginal} of 0.37, an $R^2_{\text{conditional}}$ of 0.75, and a C-value of 0.95, indicating an excellent fit (Baayen, 2008). Like the original model, the updated reference corpus model without type–token ratio predicted the likelihood of the *s*-genitive on a probability range from 0% to 100% for the 2302 observations collected from the Trinity Lancaster Corpus (Gablasova et al., 2019).

Additionally, we did not consider corpus observations where the genitive variant does not match the prediction of the reference corpus model. The model might have predicted the alternative genitive variant because it did not take into account an unknown source of variation, which we did not want to influence participants' ratings. The context of the experimental item could also not contain another genitive variant to prevent priming effects (see Engel et al., 2022; Szmrecsanyi, 2006). Finally, the target items were simplified wherever possible to make sure that all participants understood each target item regardless of their proficiency level. In particular, we transformed subclauses into main clauses and we removed or replaced words that did not figure in the New General Service List (New-GSL, Brezina & Gablasova, 2015), which contains a stable vocabulary of 2,122 highly frequent English words that learners are most likely to encounter. Since the predictors in the reference corpus model related to the features of the possessor and possessum noun phrases, simplifying the possessor and possessum would again affect the accuracy of the predictions of the model, so we did not simplify these words and instead discarded observations whose possessor and possessum noun phrases did not figure in the New-GSL. The experimental items were selected from the remaining observations.

Previous research found that acceptability ratings do not correlate in a linear fashion with the predictions from regression models, especially when the model predictions indicate that the probability of a given variant is low (Arppe & Järvikivi, 2007, pp. 149–150; Divjak et al., 2016, pp. 26–27; Kempen & Harbusch, 2005, p. 330). Therefore, the correlation between the ratings and the predictions of the reference corpus model may not be as strong when the model predicts a high likelihood of the

²This meant that we only included observations for which the possessor and possessum head lemma had an intercept adjustment in the reference corpus model that did not exceed 1 or –1.

Table 1. Distribution of target experimental items across probability bins and animacy levels. The number of observations used in the experiment for each combination of probability and animacy is shown in brackets.

Probability of the <i>s</i> -genitive	Animacy of the possessor in the experimental items		
0%–20%	Temporal (3)	Animate (3)	Locative (3)
21%–40%	Temporal (2)	Inanimate (3)	Collective (3)
41%–60%	Animate (3)	Inanimate (2)	Locative (1)
61%–80%	Temporal (1)	Inanimate (2)	Collective (3)
81%–100%	Animate (3)	Collective (3)	Locative (2)

s-genitive compared to when the *s*-genitive is not likely to occur. To properly quantify the overall correlation between the ratings and the corpus-based predictions for our second research question, we made sure to include observations from the entire probability range. The distribution of the experimental items across animacy levels is shown in Table 1, which organizes the items by the probability of the *s*-genitive according to the reference corpus model. The number of items included in the experiment for each combination of probability and possessor animacy is given in brackets. In total, there are 37 experimental items.

2.2. Participants

Power analysis using multiple simulations (Kumle et al., 2021) showed that to achieve a statistical power of 0.8, we should preferably collect data from at least 20 participants for each of the five corresponding CEFR proficiency levels featured in the TLC data, namely B1, B2, C1, C2, and native.³ The participants for the study were recruited on Prolific (Prolific, 2023), an online platform where individuals can sign up to participate in paid online experiments. These individuals fill in information about themselves upon registration to Prolific, which then allows researchers to choose who is able to participate in their experiment via prescreening.

For the purposes of the study, it was important that the participants in our experiment resemble the speakers included in the TLC data of the reference corpus model as closely as possible. Therefore, we administered the questionnaire to 25 monolingual native speakers of British English holding British nationality, who grew up and currently live in the UK (13 female and 12 male, age range: 19–67, $M = 36$, $IQR = 24$). The learner participants had to be raised monolingually with either Chinese, Hindi, Urdu, Spanish, Italian, Hungarian, Polish, or Russian as their L1, which covers the most common L1 backgrounds featured in the TLC data. More generally, including learners from several typologically diverse L1 backgrounds increases the generalizability of the findings related to learners' proficiency level and helps ensure that the results are not influenced by transfer effects from specific L1 backgrounds.

Since the proficiency level of the learner participants is determined by their score on the items from the Oxford Quick Placement Test (see Section 2.1), the questionnaire was administered to learners until we reached a satisfactory number of participants for the B1, B2, C1, and C2 levels. Given that it was particularly

³A detailed report of the power simulation is available at https://osf.io/3a6n5/?view_only=b8cda6b29d6a46aa8808d01147835f8c.

Table 2. Distribution of learner participants by mother tongue background and proficiency level

	Chinese	Hindi/Urdu	Hungarian	Italian	Polish	Russian	Spanish	Total
A2/B1	0	0	3	3	6	3	10	25
B2	3	7	5	5	4	2	4	30
C1	10	3	1	6	3	2	3	28
C2	3	3	4	3	4	0	1	18
Total	16	13	13	17	17	7	18	101

challenging to obtain enough data from B1 learners, we conflated the data from 18 B1 learners and 7 A2 learners. In total, the questionnaire was distributed to 101 EFL learners (54 female and 47 male, age range: 19–60, $M = 27$, $IQR = 15$). Table 2 presents an overview of the distribution of the EFL learners by their L1 and proficiency level.

The duration of the experiment varied considerably across participants, both among native speakers (range in minutes: 6:46–24:16, $M = 13:55$, $IQR = 5:18$) and learners (range in minutes: 11:57–94:42, $M = 24:44$, $IQR = 15:19$). Nonetheless, no submissions were excluded from further analysis because all participants performed well on the more simple questions from the Oxford Quick Placement Test aimed at beginners and intermediate learners, so it is unlikely that they answered the questions randomly.

2.3. Design and procedure

The survey was built in Qualtrics (Qualtrics, 2023) and administered online to participants recruited on Prolific. Before accessing the Qualtrics survey, the learner and native speaker participants on Prolific saw a slightly different description of the survey.⁴ As most items stemmed from a placement test (see Section 2.1), the description for the learner participants stated that we were interested in different ways of measuring how well speakers of different languages know English. Given that these items might seem particularly easy for native speakers, we explained that we also wanted to collect their data in order to test the accuracy of these measurements. This brief description also included the approximate duration of the survey and the payment participants would receive upon completion. From there, they could access the Qualtrics survey and give their consent. At the beginning of the survey, the prescreening questions were repeated. If the participants' answers were inconsistent with the information they reported on Prolific, they were automatically removed from the Qualtrics survey.

Because the experiment featured different types of questions, alternating target items and items from the Oxford Quick Placement Test could potentially have confused the participants, so questions of the same type were presented together instead. Hence, the participants answered the first 25 items from the Oxford Quick Placement Test before the target experimental items. The target items were preceded by the following instructions and an example where they could move the slider for practice:

“Below you will see sentences with two options to say the same thing, which are presented in a random order. Your task is to read each sentence carefully and choose which option sounds more natural to you. To do so, you can drag the

⁴The survey description and participant instructions are available at https://osf.io/3a6n5/?view_only=b8cda6b29d6a46aa8808d01147835f8c.

slider towards the option that sounds more natural to you. The closer you place the slider to one end, the more natural that option seems to you, and the less natural the other option. If both seem equally natural, you can place the slider closer to the middle (you will have to use the slider to go the next question). If one option seems like the only possible option, you may move the slider all the way towards that option. Please feel free to use the full range of the slider.”

Each participant saw only one observation for each combination of probability bin and animacy level shown in Table 1. Therefore, each participant saw in total 15 experimental items out of the total of 37 experimental items. Since there were up to three observations for each combination of probability and animacy level (see Table 1), the 37 experimental items were distributed across three versions of the experiment, with each version containing a different item per combination. If there were fewer than three experimental items for a given combination, the same observations would be used in multiple versions. Within each version, we made sure that there were no more than two instances of the same possessor or possessum head noun to avoid lexical priming effects (Hartsuiker et al., 2008). Participants were automatically assigned to one of the three versions after receiving the instructions. The target items were presented in pseudorandom order, and the position of the genitive variants on the scale was altered each time. This means that if the *s*-genitive figured on the right side of the slider as in Fig. 1, the *s*-genitive would figure on the left in the following item.

After the 15 target experimental items, participants answered the remaining 35 items from the Oxford Quick Placement Test. At the end of the survey, participants could leave comments or feedback about the study. The participants were then automatically returned to Prolific, and their survey was submitted for us to approve and remunerate the participant. To check whether the instructions were clear and whether the survey functioned without any technical issues, the survey was first piloted on five learner and six native speaker participants, who were made aware that they would participate in a pilot study before they started.

2.4. Statistical analysis

The ratings were analyzed as the response variable in two mixed-effects linear regression models using the *lme4* package (Bates et al., 2015) in R (version 4.2.2, R Core Team, 2022).⁵ In particular, we employed treatment coding to investigate the effect of the predictors on the ratings. In accordance with the reference corpus model, native speakers are chosen as the reference proficiency level, which allows us to compare their intuitions with those of learners at different stages of language learning. To account for idiosyncrasies and irrelevant sources of variation in the rating data, both models featured intercept adjustments for the participant and the experimental item (Baayen et al., 2008; Gries, 2015). If the *s*-genitive was presented on the left side of the slider, the participants' ratings corresponded to a score from 0 to 100 for the *of*-genitive. Because the reference corpus model predicted the likelihood of the *s*-genitive (Dubois et al., 2023), these ratings were

⁵The dataset and the analysis are available at https://osf.io/3a6n5/?view_only=b8cda6b29d6a46aa8808d01147835f8c.

transformed to also reflect a score for the *s*-genitive instead before analyzing the data.

The first model was concerned with our first research question and tested whether and how the five animacy levels influenced the ratings from the native speakers compared to the learners at different proficiency levels (Dubois et al., 2023). Hence, the first model included an interaction between the proficiency level of the participants and the animacy of the possessors in the experimental items. Collective possessors figure as the reference animacy level because this makes it easier to test the hypothesis in Dubois et al. (2023) that learners primarily struggle with collective and locative possessors. This model also featured the length of the possessor and the possessum as predictors because it was not possible to control for these constraints while designing the experiment (see Section 2.1). The length of the constituents was standardized by subtracting the median length and dividing the resulting centred value by two standard deviations (Gelman, 2008).

The second model dealt with the second research question concerning the overall relationship between the patterns of language use of the speakers in the corpus and the intuitions of participants by investigating whether and how the ratings from the native speakers and EFL learners at different proficiency levels correlated with the corpus-based predictions from the reference corpus model for the experimental items under study. Specifically, this model featured the probability of the *s*-genitive according to the reference corpus model as a predictor, which was allowed to interact with the proficiency level of the participants to reveal whether the strength of the correlation varies across participants from different proficiency levels. Importantly, the probability of the *s*-genitive was calculated based on the effect of all predictors present in the reference corpus model based on data from both native speakers and learners. This means that the prediction also incorporates the effect of the significant interaction between proficiency level and possessor animacy in the reference corpus model. Hence, if the experimental item was produced by a B1 learner in the original corpus data for example, the corpus-based prediction would capture the behaviour of B1 learners specifically. Since B1 and B2 learners are less sensitive to the animacy of the possessor than native speakers according to the reference corpus model, the corpus-based prediction for an experimental item produced by a B1 or B2 learner might not be entirely accurate for a native speaker participant and vice versa. While this is admittedly not ideal, the experimental items were selected more or less evenly from all proficiency levels in the original corpus data, so all participants had to rate items produced by speakers with a different proficiency level. Therefore, we maintain that the correlation between the predictions of the reference corpus model and the ratings from the participants provides an accurate representation of the relationship between the language use and intuitions of speakers at different proficiency levels. As for the first model on the effect of animacy, this is not an issue considering that we are mostly interested in whether the ratings for the *s*-genitive are higher or lower depending on the proficiency level of the participant.

3. Results

3.1. The animacy model

Table 3 presents the results from the first model on the effect of possessor animacy on the ratings. This model has an R^2_{marginal} of 0.15 and $R^2_{\text{conditional}}$ of 0.35.

Table 3. Effects of the individual predictors in the animacy model with collective possessors as reference level. Positive coefficients indicate an increase in ratings for the *s*-genitive. The reference level of the categorical predictors is given in brackets. (Marginally) significant *p*-values are written in bold.

Predictor	Coefficient	Std. error	<i>t</i> -value	<i>p</i> -value
(Intercept)	63.503	6.27	10.128	< 0.001
Possessor length	−1.214	5.287	−0.230	0.82
Possessum length	−15.878	5.212	−3.046	0.005
Animacy (collective)				
Animate	16.916	8.391	2.016	0.049
Inanimate	−16.456	8.889	−1.851	0.07
Locative	−17.109	9.378	−1.824	0.074
Temporal	−20.857	8.986	−2.321	0.024
Proficiency (native)				
A2 and B1	1.959	5.363	0.365	0.715
B2	−11.019	5.146	−2.141	0.033
C1	−16.116	5.218	−3.089	0.002
C2	−11.402	5.929	−1.923	0.055
Proficiency × animacy (collective)				
A2 and B1 + animate	−14.191	6.814	−2.083	0.037
A2 and B1 + inanimate	−6.64	6.813	−0.975	0.33
A2 and B1 + locative	−5.071	6.82	−0.744	0.457
A2 and B1 + temporal	−7.592	6.82	−1.113	0.266
B2 + animate	4.291	6.538	0.656	0.512
B2 + inanimate	13.391	6.535	2.049	0.041
B2 + locative	5.489	6.534	0.84	0.401
B2 + temporal	0.934	6.534	0.143	0.886
C1 + animate	9.226	6.629	1.392	0.164
C1 + inanimate	8.113	6.627	1.224	0.221
C1 + locative	7.001	6.63	1.056	0.291
C1 + temporal	8.148	6.63	1.229	0.219
C2 + animate	5.105	7.539	0.677	0.498
C2 + inanimate	3.288	7.526	0.437	0.662
C2 + locative	7.905	7.508	1.053	0.293
C2 + temporal	4.496	7.508	0.599	0.549

Whereas the length of the possessor does not influence the ratings, the model shows that the ratings for the *s*-genitive decrease when the length of the possessum increases, which echoes the finding that native speakers and learners alike tend to prefer the *of*-genitive when the possessum is long in spoken language production (Dubois et al., 2023), in conflict with the short-before-long principle (Behaghel, 1909). As animacy is part of an interaction with proficiency level, the main effects for animacy only pertain to the reference proficiency level, namely native speakers. The main effects for animacy show that native speakers give higher ratings for the *s*-genitive when the possessor is animate rather than collective, which in turn receive significantly higher ratings than temporal possessors. This pattern aligns with previous research stating that the *s*-genitive is preferred with referents that are higher on the animacy scale (Rosenbach, 2014, p. 232). The main effects for proficiency level compare the ratings for collective possessors across proficiency levels. The effects indicate that B2 and C1 learners give significantly lower ratings for the *s*-genitive than native speakers when the possessor is collective. The model features two significant interaction effects. While native speakers give higher ratings for the *s*-genitive when the possessor is animate rather than collective, this effect is much weaker for A2 and B1 learners, who give similar ratings for animate

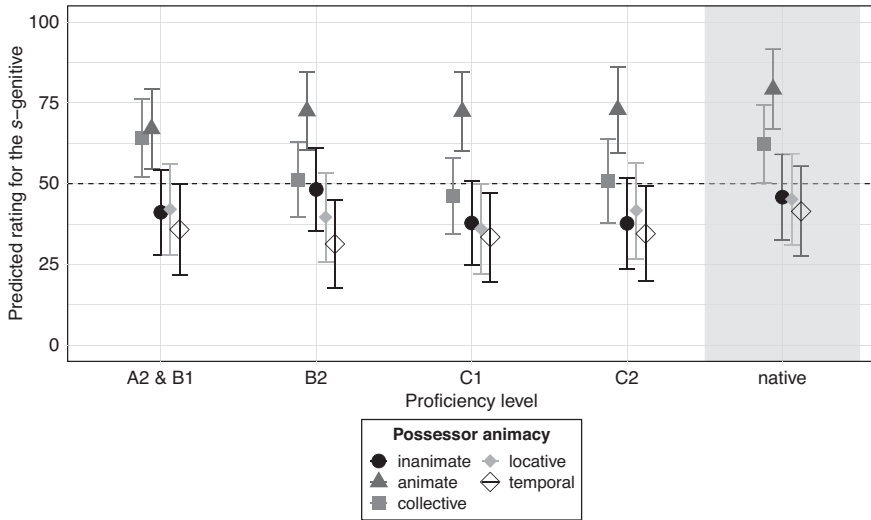


Figure 2. Partial effects plot of the interaction between possessor animacy and proficiency level. Predicted ratings on the y-axis are for the *s*-genitive. The vertical distance between the shapes reflects the effect size of the predictor. Error bars represent confidence intervals (95%). The plotted probabilities are calculated with the standardized length of the possessor and possessum at their default level, namely 0. The effects pertaining to the native speakers are shaded in grey.

and collective possessors. B2 learners, by contrast, do not distinguish between inanimate and collective possessors as much as native speakers.⁶

The effects pertaining to possessor animacy are visualized in the partial effects plot in Figure 2. Partial effects plots summarize the information from the model by changing the value of the relevant constraint, namely possessor animacy, for each of the proficiency levels while holding the other constraints in the model, namely possessor and possessum length, at their default level (Fox, 2003). In this way, the plot shows how a change in possessor animacy affects the ratings for the *s*-genitive for native speakers and learners across proficiency levels.

As suggested by the interaction effects presented in Table 3, Fig. 2 shows that the ratings for animate possessors and collective possessors are very similar at the A2/B1 level, whereas native speakers clearly give higher ratings for the *s*-genitive when the possessor is animate rather than collective. In particular, it seems that A2 and B1 learners give slightly higher ratings to collective possessors while at the same time giving lower ratings to animate possessors. Turning to the interaction effect between B2 learners and native speakers regarding collective and inanimate possessors, we can see that B2 learners give significantly lower ratings for the *s*-genitive than native speakers when the possessor is collective (see Table 3) while at the same time giving higher ratings when the possessor is inanimate. As a result, there is barely a difference between the ratings for collective possessors and inanimate possessors at the B2 level as opposed to native speakers who give higher ratings for the *s*-genitive when the possessor is collective.

⁶To ascertain that these interaction effects are not due to Spanish EFL learners, who account for about 40% of the A2 and B1 learners, we ran a model without these learners, which revealed the same patterns.

Table 4. Effects of the individual predictors in the corpus prediction model. Positive coefficients indicate an increase in ratings for the *s*-genitive. The reference level of the categorical predictors is given in brackets. (Marginally) significant *p*-values are written in bold.

Predictor	Coefficient	Std. error	<i>t</i> -value	<i>p</i> -value
(Intercept)	31.044	5.313	5.843	<.001
Corpus-based prediction	53.313	8.987	5.932	<.001
Proficiency (native)				
A2 and B1	10.063	4.732	2.127	0.034
B2	0.361	4.539	0.08	0.937
C1	-6.501	4.594	-1.415	0.158
C2	-5.209	5.221	-0.998	0.319
Proficiency x corpus-based prediction				
A2 and B1 + corpus-based prediction	-30.717	7.246	-4.239	<.001
B2 + corpus-based prediction	-13.57	6.933	-1.957	0.051
C1 + corpus-based prediction	-6.442	7.007	-0.919	0.358
C2 + corpus-based prediction	-4.255	7.944	-0.536	0.592

3.2. The corpus prediction model

Table 4 presents the results from the second model. The model has an R^2_{marginal} of 0.14 and a $R^2_{\text{conditional}}$ of 0.34.

There is a significant effect for the prediction of the reference corpus model, which shows that when the model predicts a higher likelihood for the *s*-genitive, the rating for the *s*-genitive also increases. Since the corpus-based prediction is part of a significant interaction term with proficiency level, the main effect shown in Table 4 only pertains to the reference proficiency level, namely native speakers. The interaction suggests that A2 and B1 speakers are less sensitive to the effect of the corpus-based predictions compared to native speakers, indicating that their ratings correspond less closely to the predictions of the reference corpus model. As their proficiency level increases, the ratings of the learners become more similar to those of the native speakers, and by extension, to the language use of the speakers in the corpus. Hence, while the interaction is still marginally significant for B2 learners, the coefficients gradually approach zero in the C1 and C2 levels.

The partial effects plot in Figure 3 visualizes the correlation between the ratings and the corpus-based predictions for native speakers and each learner proficiency level separately.

The figure shows that the ratings from the native speakers are most sensitive to the predictions of the reference corpus model: their slope is steeper compared to the slopes for the learners, indicating that an increase in the predicted likelihood of the *s*-genitive leads to a larger increase in ratings for the *s*-genitive. Specifically, the intercept in Table 4 indicates that native speakers give the *s*-genitive a low rating of 31 when the reference corpus model predicts the *of*-genitive, whereas they rate the *s*-genitive approximately 50 points higher when the *s*-genitive is very likely in language use. Overall, the slopes for the learners become steeper as their proficiency level increases, so they gradually approximate native speakers. The upper left panel of Fig. 3 shows that A2 and B1 learners, who differ significantly from native speakers, exhibit the flattest curve. In particular, the curve for the A2 and B1 learners seems to indicate that A2 and B1 learners give lower ratings for the *s*-genitive than native speakers when the *s*-genitive is highly likely according to the reference corpus model while also giving higher ratings than native speakers when the *s*-genitive is unlikely,

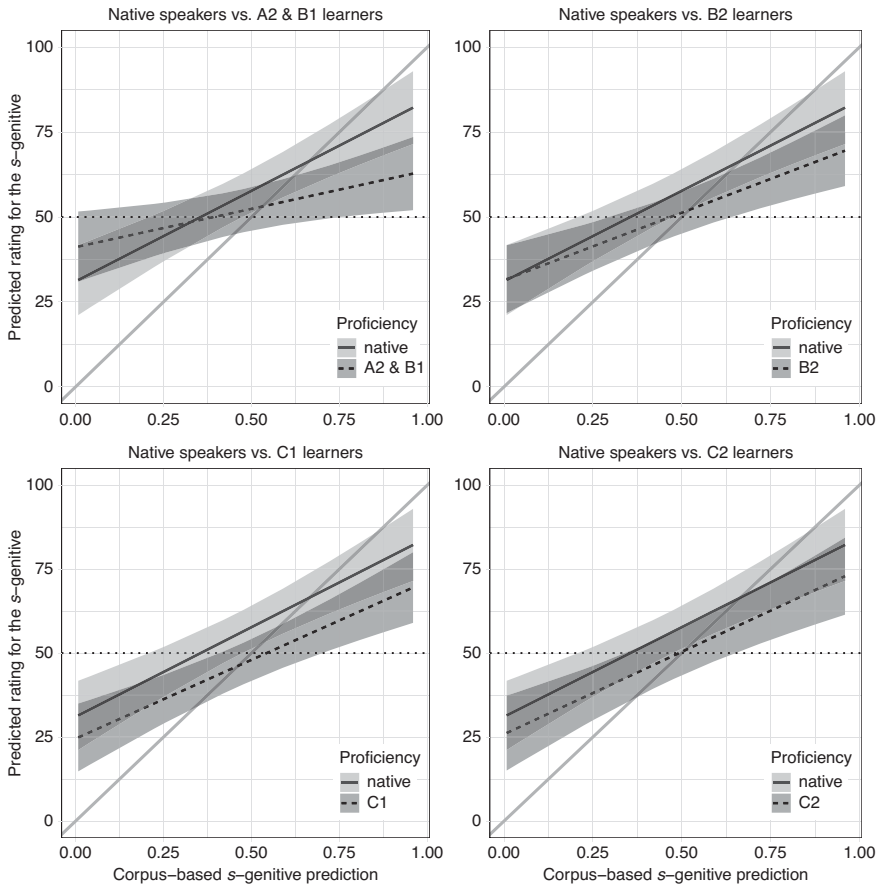


Figure 3. Partial effects plot of the interaction between proficiency level and the corpus-based prediction for the *s*-genitive for native speakers and each learner proficiency level separately. Predicted ratings on the *y*-axis are for the *s*-genitive. The grey diagonal line represents a perfect fit between corpus predictions and ratings.

i.e. when the reference corpus model predicts a preference for the *of*-genitive (see Table 4). B2 learners also rate the *s*-genitive slightly lower than native speakers when the reference corpus model predicts an *s*-genitive, but they give the same ratings as native speakers when the model predicts an *of*-genitive. Interestingly, it seems that C1 and C2 learners give slightly lower ratings for the *s*-genitive in general compared to native speakers, which might point to a slight overall preference for the *of*-genitive.⁷

The overall repeated-measures correlation between the predictions from the reference corpus model and the ratings from all participants is moderate at

⁷One could argue that the slightly lower slopes of the C1 and C2 learners align better with the corpus-based predictions as they more closely follow the grey diagonal representing a perfect fit. If learners also exhibit a slight preference for the *of*-genitive compared to native speakers in their language use, then this finding is not surprising considering that the corpus-based predictions stem from the language use of both native speakers and learners.

$r = 0.38$ ($p < 0.001$) (Bakdash & Marusich, 2017). If we consider that ratings above 50 represent a preference for the *s*-genitive, we can also examine how often the participants' intuitions and the predictions from the reference corpus model agree on the same genitive variant in a categorical way. Hence, for each of the participant proficiency levels, we selected a small, balanced sample consisting of 100 observations for which the reference corpus model predicts that the *s*-genitive is more likely and 100 for which it predicts that the *of*-genitive is more likely. The intuitions of the native speakers correspond to the predicted variant in 80% of the cases. By comparison, the intuitions from participants at the A2 and B1 levels match the predicted variant in only 68% of the cases. As learners become more proficient, their intuitions become more similar to actual language use, with an agreement of 73% at the B2 level, 77% at the C1 level, and 81% at the C2 level.

4. Discussion

4.1. (Methodological) summary

The present study investigated the interface between implicit grammatical knowledge and language. To do so, we compared the intuitions of native speakers and EFL learners regarding the naturalness of the genitive variants in a rating task experiment against the use of the genitive variants in language production, for which we took as a reference point the corpus-based regression model of Dubois et al. (2023). The ratings of 25 native speakers and 101 EFL learners from various L1 backgrounds with a proficiency level ranging from A2 to C2 were analyzed with two separate mixed-effects linear regression models. The first model focused on our first research question and compared how possessor animacy, operationalized as a five-level distinction between animate, collective, inanimate, locative, and temporal possessors, influences the intuitions of native speakers and EFL learners in the rating task experiment compared to the language use of the speakers in the corpus. The second model was aimed at our second research question, which investigates how the intuitions from the participants in the experiment relate to the patterns of language use in general. Therefore, the second model tested how the predictions from the reference corpus model, which reflect spoken language use, correlate with the ratings of the participants in the experiment as this would indicate that native speakers and EFL learners are generally sensitive to probabilistic constraints driving the choice of variant (Bresnan & Ford, 2010).

4.2. RQ1: How does possessor animacy affect the intuitions of EFL learners at different proficiency levels in a rating task experiment compared to correlations between animacy and genitive variant in a corpus of spoken learner language?

In accordance with previous research (Rosenbach, 2014, p. 232), the results from the first model show that native speakers find the *s*-genitive most natural when the possessor is animate, followed by collective, inanimate, locative, and temporal possessors. By contrast, A2 and B1 learners find the *s*-genitive more or less equally natural with animate and collective possessors. Figure 2 suggests that this is due to A2 and B1 learners giving slightly higher ratings for collective possessors while at the same time giving lower ratings for animate possessors. B2 learners, on the other hand, find the *s*-genitive equally natural when the possessor is collective or inanimate,

whereas native speakers clearly prefer the *s*-genitive when the possessor is collective. Figure 2 suggests that this is again due to a combination of B2 learners giving slightly lower ratings for collective possessors (see Table 3) and higher ratings for inanimate possessors.

At first glance, the lack of distinctions between certain animacy levels in the ratings of A2/B1 and B2 participants in the experiment reflects the tendency for B1 and B2 learners to rely less on the animacy constraint than native speakers in spoken language production (Dubois et al., 2023). There are, however, important differences. According to Dubois et al. (2023), B1 and B2 learners are less sensitive to animacy because they have difficulty determining the animacy of possessors such as *region*, *province*, *world*, and *India* in specific contexts where these possessors can be interpreted as either locatives, which take the *of*-genitive, or collectives, which prefer the *s*-genitive. As a result, there is a trade-off between these two categories, causing learners to use the *s*-genitive more often with locative possessors and the *of*-genitive more often with collective possessors. Although the experiment certainly suggests that B2 learners find the *s*-genitive less natural than native speakers when the possessor is collective up until the C1 level (see Table 3), this does not apply to the A2 and B1 learners. The participants also do not appear to experience difficulties with locative possessors and instead A2/B1 and B2 learners struggle with animate and inanimate possessors respectively. This is surprising considering that the easiest strategy for learners would be to rely on the very strong animacy constraint in a categorical way by always using the *s*-genitive when the possessor is animate and the *of*-genitive when the possessor is inanimate, as suggested by prescriptivists (Murphy, 2012; cited in Heller et al., 2017, p. 10).

Moreover, while a trade-off between locative and collective possessors seems plausible, it is unclear how learners would fail to differentiate between animate and collective possessors or collective and inanimate possessors, unless there are important distinctions beyond our five-level annotation of animacy that we did not take into account. In general, animacy is difficult to operationalize in all its facets, since the linguistic animacy distinctions are language-specific rather than biological and not diachronically stable (Wolk et al., 2013; Zaenen et al., 2004). Rather than a trade-off, the more likely explanation is that learners struggle with each of the categories independently. According to Dubois et al. (2023), the animacy constraint requires learners to first determine the animacy of the possessor noun phrase in the context. Once this has been determined, they can produce the preferred genitive variant for this animacy category. For a possessor like *dog* for example, learners would first have to determine that it is an animate possessor. Since animate possessors prefer the *s*-genitive, they would be more likely to produce an *s*-genitive with *dog* as a possessor. With this in mind, learners' struggle with independent animacy categories could either result from the fact that the learners have not yet fine-tuned how these larger animacy categories affect the choice of variant, e.g. collectives prefer the *s*-genitive, but not as much as animate possessors, or they simply do not have enough experience with certain individual possessor noun phrases to determine their animacy in a consistent way to begin with. The latter explanation seems more plausible since what makes the animacy constraint more complicated than other constraints is the multitude of nouns that have to be categorized correctly based on the context without overt marking (see Dubois et al., 2023).

In fact, low-proficiency learners might not even consider the animacy of the possessor when choosing the genitive variant because learners do not start out with

knowledge of these categories and their associations with the variants. Instead, they are exposed to genitive exemplars with thousands of different nouns, each of which has its own preferred genitive variant. Once they learn how many of these nouns pattern with the genitive variants, they may recognize that nouns behaving the same way share the same animacy, which then gradually influences language production (see Abbot-Smith & Tomasello, 2006, p. 284; Azazil, 2020, p. 420; Divjak et al., 2021, p. 78).

In such an account, the animacy constraint would capture item-specific effects in the early stages of acquisition, so the most plausible explanation for the differences between native speakers and low-proficiency learners regarding the animate, inanimate, and collective possessors is that the nouns included in the experiment for these categories are somehow problematic. According to usage-based approaches, language learning is driven by input (Ellis & Wulff, 2019), so the frequency of possessor nouns can be expected to influence their learnability considerably. To compare the frequency of the possessor nouns across animacy levels, we checked their ranking in the New-GSL (Brezina & Gablasova, 2015). The expectation is that the animate, collective, and inanimate possessors feature less frequent nouns on average than locative and temporal possessors, so they should be found lower on the list. The average ranking of animate (593), collective (688), and inanimate (970) possessors in the list is indeed lower on average than temporal possessors (289). Most locative possessors are country names, which are not included in the New-GSL, but it is highly likely that learners are also very familiar with words such as America and India, so overall it seems that temporal and locative possessors were represented by easier nouns in the experiment. Another difficulty with collective possessors in the experiment is that they mostly feature singular collective nouns with plural reference such as *association*, *community*, and *government*, which are potentially confusing for learners (e.g. Dziemianko, 2008). Nonetheless, one would need to adopt a more sophisticated approach to collect more conclusive evidence considering that the overall frequency of a noun does not correlate perfectly with how frequently it is encountered as a possessor in genitive exemplars.

4.3. RQ2: How does the relationship between experimental ratings and corpus data vary by the proficiency level of participants and speakers?

The second model showed that there is a positive correlation between the predictions from the reference corpus model and the ratings in the experiment: when the reference corpus model predicts a high likelihood for the *s*-genitive, the participants in the experiment also give higher ratings for the *s*-genitive. These results suggest that the significant constraints on which the reference corpus model is based (see Dubois et al., 2023), i.e., the constraints that drive spoken language use, are indeed part of speakers' implicit linguistic knowledge as posited by the Probabilistic Grammar framework (Bresnan & Ford, 2010; Engel et al., 2022). However, the correlation between the predictions from the reference corpus model and the ratings is weaker for the A2/B1 learners and to a lesser extent the B2 learners compared to native speakers.

This difference could be explained in two ways. On the one hand, the linear regression model might have inferred that the ratings from the A2/B1 and B2 learners correspond less well to the predictions from the reference corpus model because these

learners tended to stick to the middle range of the rating scale, as implied by their flatter curve in Fig. 3. One reason for this might be for example that A2/B1 and B2 learners are less confident about their choices than native speakers. In this case, their intuitions about the naturalness of the variants would in principle be similar to those of native speakers, but they did not want to give ratings on the extreme ends of the slider given that a high rating for one variant implies that the other variant is much less natural in our experiment design. In other words, although learners might have a clear preference for one variant that aligns with the intuition of native speakers, they might not have wanted to exclude the other variant completely. To test this possibility, we compared to what extent the ratings of A2/B1, B2, and native speakers deviated from the middle point of the rating scale, namely 50. For example, if a learner gave a rating of 75 for the *s*-genitive, the deviation from 50 is 25. If A2/B1 and B2 learners stick more to the middle range of the rating scale, their deviations from 50 should on average be smaller than those of the native speakers. Wilcoxon tests revealed that neither A2/B1 learners ($p = 0.99$) nor B2 learners ($p = 0.34$) gave ratings that are significantly closer to the middle point of the scale than native speakers. Hence, the interaction effect is probably not due to these learners sticking to the middle range of the rating scale.

This supports the second possible explanation, which is that the ratings from A2/B1 and B2 learners correspond less well to the predictions of the reference corpus model compared to native speakers because these learners do not always share the same intuitions as native speakers. More specifically, the intuitions from the native speakers in the experiment appear to correspond more closely to how native speakers and learners use the genitive variants in spoken language production. The question arises as to whether this difference in intuitions appears in specific contexts. In this respect, the partial effects plot in Fig. 3 shows that the intuitions of A2 and B1 learners diverge from those of native speakers both in contexts where speakers would use the *s*-genitive and in contexts where they would use the *of*-genitive in language production. In *s*-genitive contexts, A2 and B1 learners find the *of*-genitive more natural than native speakers in the experiment. In *of*-genitive contexts, A2 and B1 learners find the *s*-genitive more natural than native speakers. B2 learners share the same intuitions as native speakers in *of*-genitive contexts, but they differ from native speakers in *s*-genitive contexts, where they find the *of*-genitive slightly more natural.

4.4. Synopsis

While there are subtle differences in the way in which the animacy constraint affects the intuitions and spoken language production of A2/B1 and B2 learners, both sources of data point to the fact that low-proficiency learners are generally less sensitive than native speakers to certain animacy levels (see Dubois et al., 2023). In general, the results from our study are reminiscent of the findings of Wolk et al. (2013) that the effect of animacy is weakening over time in different alternation phenomena and more importantly Heller et al. (2017), who found that the effect of animacy on the choice of genitive variant is also weaker for speakers from different World Englishes compared to native varieties. In line with Dubois et al. (2023), we argue that there are differences between learners and native speakers with regard to possessor animacy because the constraint requires learners to pick up on the statistical association between the genitive variants and many different possessor

nouns. As a result, low-proficiency learners lack the necessary language exposure to develop abstract animacy categories they can rely on during language production like native speakers.

However, it is important to keep in mind that the models present the aggregated results across the individual participants. It is therefore unclear whether the results from the model derive from the A2/B1 and B2 learners as a group or by some individuals in particular who developed different strategies from the other learners at the same proficiency level. This is all the more relevant in the context of learner language as it features even more individual variability compared to native speakers (Jäschke & Plag, 2016, pp. 511–512; Kinne, 2020, p. 128). Inspired by Kinne (2020), we therefore calculated Cronbach's alpha on the 15 ratings of each participant to get an idea of how much individual variability there is at each proficiency level. As expected, the native speakers and the C2 learners are the most consistent in their ratings with a Cronbach's α of 0.95, followed by the C1 learners (Cronbach's $\alpha = 0.94$) and the B2 learners (Cronbach's $\alpha = 0.9$), all of whom exhibit excellent internal consistency above 0.9 (Gliem & Gliem, 2003). By comparison, the A2 and B1 learners are much less consistent with a Cronbach's α of 0.72. This implies that the A2 and B1 learners are indeed not very homogenous with respect to how natural they find the genitive variants with different possessors in various contexts.

A closer look at the ratings from individual A2/B1 participants showed that 10 out of 25 participants (40%) already follow the expected pattern whereby *s*-genitives are generally considered more natural with animate possessors than with collective possessors. In the B2 group, a similar pattern emerges: half the participants (15 out of 30) find the *s*-genitive more natural with collective possessors than inanimate possessors like native speakers. Since learners build an interlanguage grammar that accommodates the language input they have received in a rational way (Ellis, 2006), the different strategies of learners ultimately depend on the exemplars they have encountered in their previous language experience (Ellis & Wulff, 2019). Because low-proficiency learners have less experience with the language, their interlanguage grammar is built on fewer exemplars that can be categorized in different ways, which naturally leads to more variability across learners.

More generally, we found that the intuitions from the participants correlate with the patterns of spoken language use, which provides further evidence that native speakers and EFL learners alike are sensitive to the implicit probabilistic constraints that drive language use, as posited by the Probabilistic Grammar framework (Bresnan & Ford, 2010; Engel et al., 2022). However, this correlation is slightly weaker for the A2/B1 and B2 learners, which implies that these learners are more likely than native speakers to produce a genitive variant that they do not find most natural in a given context. As learners become more proficient, their intuitions and language use become more similar.

To some extent, this result can be explained by the fact that low-proficiency participants are less consistent in their ratings, which could attenuate the relevant statistical association in the model and, by extension, the regression slope in Fig. 3.⁸ This is also partially due to the effect of the animacy constraint. In accordance with previous literature, the reference corpus model found that the animacy of the possessor is one of the strongest determinants of genitive choice in spoken language

⁸We thank an anonymous reviewer for this observation.

production (Rosenbach, 2014). Hence, the prediction of the reference corpus model was largely determined by the effects of possessor animacy, with animate and collective possessors strongly favouring the *s*-genitive and inanimate, locative, and temporal possessors disfavoring the *s*-genitive (Dubois et al., 2023). The fact that the effect of the animacy constraint influences the language use and the intuitions of the A2/B1 and B2 learners slightly differently contributes to why the intuitions of these learners correspond less well to the predictions of the reference corpus model.

However, the effect of the animacy constraint does not overlap with the findings from the corpus-prediction model well enough to account for this finding in its entirety. Another explanation is that spoken language involves different kinds of time-sensitive, processing-related factors that influence language production and might cause a speaker to produce a variant that they find less natural but that is easier to process in a given context. These restrictions did not apply during the experiment, so the participants could always choose the variant they found more natural. This could lead to differences between the intuitions of the participants and their spoken language use, especially for low-proficiency learners for whom processing-related considerations during spoken language production are arguably more important given that their language production is in all likelihood less efficient and automatic than that of more advanced learners and native speakers (Segalowitz, 2003). In reality however, most processing-related constraints are controlled for during the experiment and previous research showed that even low-intermediate learners are surprisingly similar to native speakers when it comes to the actual use of the genitive variants in spoken language production (Dubois et al., 2023; see also Gries & Wulff, 2013).

Therefore, we cannot exclude the alternative possibility that low-proficiency learners differ from native speakers because of factors that are specific to the rating task experiment. In this regard, the intuitions feature much unexplained variability that does not correspond to the patterns of spoken language use, especially at the lower-proficiency levels. On the one hand, this implies that some of the constraints driving language use are not relevant during introspective tasks like the rating task experiment. Unfortunately, it is not possible to test the relevance of each individual constraint since the corpus-based prediction results from the combined effect of all significant constraints, not to mention most constraints are controlled for apart from possessor animacy. On the other hand, this divergence between corpus-based predictions and ratings is to be expected given that introspection yields slightly different results than corpus data (Arppe & Järvikivi, 2007, p. 152).

More generally, rating task experiments offer the participant more freedom than forced-choice tasks for example, so learners might have a different interpretation of what sounds natural to them as they complete the experiment. On a related note, the ratings of especially low-proficiency learners could be influenced by prescriptivism and their previous teaching experience, meaning that they could give higher ratings for the variant they were taught is 'right' although this does not align with their implicit grammatical knowledge. Overall, it is not entirely clear to what extent the ratings capture implicit grammatical knowledge as opposed to strategic behaviour, or whether it is natural for participants to gauge the acceptability of the variants on such a fine-grained scale as in the present design to begin with (see Stefanowitsch, 2006, p. 73). Future research should therefore investigate whether there are differences between what these ratings capture at low-proficiency levels compared to higher proficiency levels and native speakers. Nonetheless, it is reassuring that the strength of the correlation between corpus-based predictions and ratings for native speakers

and learners across several proficiency levels and L1 backgrounds is comparable to recent studies that adopt the same approach for native speakers of English exclusively (see Engel, 2022, p. 176; Engel & Szmrecsanyi, 2023, p. 370).

Data availability statement. The data, the code with the analyses, the survey description, and participant instructions are available at https://osf.io/3a6n5/?view_only=b8cda6b29d6a46aa8808d01147835f8c.

Competing interest. The authors declare none.

Financial support. This work was supported by the Fonds de la Recherche Scientifique - FNRS under Grant(s) n° 40005897.

References

- Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, 23(3), 275–290.
- Allan, D. (2001). *Oxford Placement Test 1*. Oxford University Press.
- Arppe, A., & Järvikivi, J. (2007). Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory*, 3(2), 131–159.
- Azail, L. (2020). Frequency effects in the L2 acquisition of the catenative verb construction – evidence from experimental and corpus data. *Cognitive Linguistics*, 31(3), 417–451.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, 8, 456.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25, 110–142.
- Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 89(1), 1–47.
- Branigan, H. P., Pickering, M. J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118(2), 172–189.
- Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language: Journal of the Linguistic Society of America*, 86(1), 168–213.
- Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.
- Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Divjak, D., & Arppe, A. (2013). Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics*, 24(2), 221–274.
- Divjak, D., Dąbrowska, E., & Arppe, A. (2016). Machine Meets Man: Evaluating the Psychological Reality of Corpus-based Probabilistic Models. *Cognitive Linguistics*, 27(1), 1–33.
- Divjak, D., Milin, P., Ez-zizi, A., Józefowski, J., & Adam, C. (2021). What is learned from exposure: An error-driven approach to productivity in language. *Language, Cognition and Neuroscience*, 36(1), 60–83.
- Dubois, T., Paquot, M., & Szmrecsanyi, B. (2023). Alternation phenomena and language proficiency: The genitive alternation in the spoken language of EFL learners. *Corpus Linguistics and Linguistic Theory*, 19(3), 427–450.
- Dziemiątko, A. (2008). Subject-Verb Concord with Collective Nouns or the Count-Mass Distinction: Which is More Difficult for Polish Learners of English? *Poznań Studies in Contemporary Linguistics*, 44(4), 449–468.
- Ellis, N. C. (2006). Cognitive perspectives on SLA: The associative-cognitive CREED. *AILA Review*, 19, 100–121.

- Ellis, N. C. (2017). Cognition, Corpora, and Computing: Triangulating Research in Usage-based Language Learning. *Language Learning*, 67(S1), 40–65.
- Ellis, N. C., & Wulff, S. (2019). Cognitive approaches to Second Language Acquisition. In J. W. Schwieter & A. Benati (Eds.), *The Cambridge Handbook of Language Learning* (pp. 41–61). Cambridge University Press.
- Engel, A. (2022). *The register-specificity of probabilistic grammars in English and Dutch—Combining corpus analysis and experimentation*. KU Leuven.
- Engel, A., Grafmiller, J., Rosseel, L., & Szmrecsanyi, B. (2022). Assessing the complexity of lectal competence: The register-specificity of the dative alternation after *give*. *Cognitive Linguistics*, 33(4), 727–766.
- Engel, A., & Szmrecsanyi, B. (2023). Variable grammars are variable across registers: Future temporal reference in English. *Language Variation and Change*, 34(3), 355–378.
- Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8, 1–27.
- Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, 5(2), 126–158.
- Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2017). Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style. *Applied Linguistics*, 38(5), 613–637.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865–2873.
- Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1), 1–26.
- Gliem, J. A., & Gliem, R. R. (2003). Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-type Scales. In *Proceedings of the 2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education* (pp. 82–88).
- Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125.
- Gries, S. T., & Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners: Towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, 18(3), 327–356.
- Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58(2), 214–238.
- Heller, B., Szmrecsanyi, B., & Grafmiller, J. (2017). Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics*, 45(1), 3–27.
- Jäschke, K., & Plag, I. (2016). The dative alternation in German–English interlanguage. *Studies in Second Language Acquisition*, 38(3), 485–521.
- Kempen, G., & Harbusch, K. (2005). The Relationship between Grammaticality Ratings and Corpus Frequencies: A Case Study into Word Order Variability in the Midfield of German Clauses. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives* (pp. 329–350). De Gruyter Mouton.
- Kinne, A. (2020). *Particle Placement in English L1 and L2 Academic Writing*. Presses universitaires de Louvain.
- Klavan, J. (2020). Pitting corpus-based classification models against each other: A case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory*, 16(2), 363–391.
- Klavan, J., & Divjak, D. (2016). The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica*, 50(2), 355–384.
- Kumle, L., Vö, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, 53, 2528–2543.
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, Article no. 226, 1–16.
- Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, 27(4), 507–526.
- Murphy, R. (2012). *English Grammar in Use. A self-study reference and practice book for intermediate learners of English*. Cambridge University Press.
- Prolific. (2023). *Prolific*. Prolific. <https://www.prolific.com>.
- Qualtrics. (2023). *Qualtrics*. Qualtrics. <https://www.qualtrics.com>.

- R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rosenbach, A. (2014). English genitive variation—The state of the art. *English Language and Linguistics*, 18(2), 215–262.
- Segalowitz, N. (2003). Automaticity and Second Languages. In Catherine J. Doughty & Michael H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 382–408). John Wiley & Sons, Ltd.
- Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory*, 2(1), 61–77.
- Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Mouton de Gruyter.
- Wolk, C., Bresnan, J., Rosenbach, A., & Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructural variation and change. *Diachronica*, 30(3), 382–419.
- Wulff, S., & Gries, S. T. (2019). Particle Placement in Learner Language. *Language Learning*, 69(4), 873–910.
- Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M. C., & Wasow, T. (2004). Animacy encoding in English: Why and how. In B. Webber & D. Byron (Eds.), *Proceedings of the 2004 ACL Workshop on Discourse Annotation, Barcelona, July 2004* (pp. 118–125). Association for Computational Linguistics.

Cite this article: Dubois, T., Grafmiller, J., Paquot, M., & Szmrecsanyi, B. (2024). Animacy effects in the English genitive alternation: comparing native speakers and EFL learner judgments with corpus data, *Language and Cognition* 16: 624–646. <https://doi.org/10.1017/langcog.2023.51>