Gender asymmetries in ellipsis: an experimental comparison of markedness and frequency accounts

Jon Sprouse
University of Connecticut

Troy Messick
Rutgers University

Jonathan Bobaljik
University of Connecticut

Abstract

Bobaljik & Zocca (2011) argue that ellipsis reveals the existence of (at least) two classes of gender-paired nouns, distinguished by the relationship between morphological marking and semantic specification: in the *actor/actress* class, the feminine form is specified for (semantic) sex, while the unaffixed form is semantically unspecified, exemplifying the classic markedness asymmetry (Jakobson 1932); in the *prince/princess* class, despite the same morphological relation between the words, the unaffixed form is semantically restricted to males. Bobaljik & Zocca and others (see Merchant 2014, Spathos & Sudo 2016, and Saab 2017) pursue accounts that code the difference between the classes in the linguistic representation of such nouns, incorporating (un)markedness into the explanation. By contrast, Haspelmath 2006 has suggested that differences in the relative frequency of the two forms in each pair may be the cause for the different behavior of the two classes. The frequency approach predicts that the size of the acceptability asymmetries that arise under the ellipsis test will correlate with the size of relative frequency ratio for each noun pair. In this article, we develop a formal experimental version of the Bobaljik & Zocca (2011) ellipsis test for gender asymmetries, deploy it for 16 noun pairs in English (by hypothesis 8 of each class), and then use three publicly available corpora to test this prediction. Our results suggest that the Haspelmath 2006 relative frequency hypothesis is not an empirically adequate competitor for the explanation of these asymmetries, as there is no evidence of a correlation between the size of the asymmetry effects in the acceptability judgment experiment and the size of the ratio of the relative frequency of the two forms in each pair.

Keywords: gender asymmetries, ellipsis, experimental syntax, acceptability judgments, word frequency

1. INTRODUCTION

Certain triples of nouns, in semantic fields such as kinship terms, animal names, nobility titles, and professions, among others, can be seen as organized along the lines in (1), where there is a superordinate term, which can be subdivided along a male:female sex opposition (cf. Corbett 1991 *Gender*):[1]

(1)

sibling	
brother	sister

monarch	
king	queen

sheep	
ram	ewe

However, there are also certain pairs of nouns in these same semantic fields that are organized along the lines in (2), where two terms serve to lexicalize the three possible meanings: the same term is used for the superordinate category as for the male sex, and a distinct form is used for the female sex.

(2)

lion	
lion	lion-ess

actor	
actor	actr-ess

Similar patterns obtain for other languages – the examples in (3) are from Russian and German, respectively (Jakobson 1932). In these languages, like English, the female-denoting form is marked by a suffix, and thus appears to be productively derived from the affixless form.

(3)

osël 'donkey'	
osël	osl-ica

Löwe 'lion'	
Löwe	Löw-in

In languages with grammatical gender, such as German and Russian, the affixless form is masculine gender, and the suffixed form is feminine gender, as seen in adjectival or verbal agreement, pronominal reference, and the like. Although we recognize (Modern) English has all but lost its grammatical gender system, we will continue to use the term gender when referring to the contrast in linguistic form as opposed to meaning, even in English.

Since Jakobson's (1932) seminal proposals, relations of the sort in (2)-(3) have played a prominent role in motivating the idea of *markedness* in grammar broadly, and in morphology specifically. The feminine forms are morphologically *marked* relative to the corresponding masculine forms, in the sense that the feminine forms bear a mark (i.e., the affix) which the masculine forms lack. What Jakobson argued was that there is a corresponding notion of markedness in the semantics: he argued that the feminine form is semantically marked in

[1] We use the terms *gender (feminine/masculine)* for the grammatical classifications of nouns and the terms *sex (female/male)* for the semantic classification, sometimes termed 'natural gender' (but see McConnell-Ginet 2014 for why this may be better seen as "notional" rather than "natural"). Beyond the narrow concerns of this paper, the relation between these is famously complex, and we do not aim to do justice to the many issues involved. Note in addition that at the level of analysis we pursue here, the linguistic system appears to encode two opposing values for sex, as in (1). We do not intend this binarity to imply anything about the range of values that sex (or in particular notional gender) can take, or the range of meanings that the semantic system could potentially encode.

signaling the meaning ‘sex:female’, but that the masculine form does not signal ‘sex:male’ but instead is semantically unmarked, the lack of sex specification (despite having masculine grammatical gender in gendered languages like German and Russian). In contexts where the masculine (morphologically unmarked) form appears to denote the specific semantics “male”, Jakobson proposes that some sort of logical reasoning is in the background: by selecting the unmarked form in a context where the marked form was in principle available, the speaker implies that they did not intend the marked form, and therefore did not intend the semantics of the marked form (female sex). Thus, in context, the morphologically unmarked form may come to imply the negation of the marked semantics (not-female sex), and thus the unmarked form implies male sex, but this implication is not part of the meaning of the unmarked form.

Evidence that the morphologically unmarked forms (but not the marked ones) can refer to the superordinate semantic category comes from contexts like the following. If a speaker of English is not aware of the sex of the person in question, they can utter the question in (4a), and receive the answer in (4b) without contradiction. However, they cannot utter the question in (5a) and receive the answer in (5b) without a contradiction.

(4) a. Is there an actor in that photograph?
b. Yes, namely Meryl Streep.

(5) a. Is there an actress in that photograph?
b. #Yes, namely Robert Redford.

A similar diagnostic obtains with plural forms, as has long been noted (see, e.g., Greenberg 1966:30-31, citing earlier Arabic grammarians; see Corbett 1991 Ch 9 for critical discussion and examples which pattern differently). In many languages, the plural masculine form can be used to refer to groups that contain both male and female sexes; but the feminine form can only be used to refer to a group that is exclusively female.

(6) a. Look at the photograph of those **actors** at the Academy Awards. Aren't they wearing interesting outfits?
b. Yes! I especially like what the **actress** in the middle of the photograph is wearing.

(7) a. Look at the photograph of those **actresses** at the Academy Awards. Aren't they wearing interesting outfits?
b. #Yes! I especially like what the **actor** in the middle of the photograph is wearing.

While these examples show that the unmarked (unaffixed) form *actor* can refer to members of the profession regardless of the members’ sex, Jakobson also gave examples to show that in the right context, specifically male reference can be intended. One of Jakobson’s pairs (in translation) is as follows. In (8), speaker B confirms that the animal is a lion (species) and adds the further specification as to sex. But in (9-B), *lion* appears to be used specifically to signal male sex.[2]

[2] The exchange in (9) has none of the oddity of the following. The superordinate terms in (1) can’t in the general case be used to signal the opposite gender/sex.
(i) Speaker A: Is that your sister? Is that a sheep?

(8) Speaker A: Is that a lion over there?
Speaker B: Yes, more precisely it's a lioness.

(9) Speaker A: Is that a lion-ess over there?
Speaker B: *Yes, it's a lion.
No, it's a lion.

As noted above, Jakobson's explanation for how the masculine form appears to be sometimes sex-neutral (6-8) and sometimes specifically male (9) parallels the familiar logic of Gricean reasoning. Formally, the morphologically unmarked, masculine form is not semantically marked for sex. But assuming a binary opposition between female and not-female, when femaleness is explicitly denied (as in (9-B)), the hearer is licensed to conclude that maleness is intended.

From a Jakobsonian perspective, it is surprising, then, that there are some nouns, with the same morphological unmarked:marked opposition as in (2)-(3), which fail to pattern semantically in the same way (Bobaljik & Zocca 2011, Haspelmath 2006):

(10) English: prince princ-ess
German: König König-in

Although these nouns have the same morphology as those in (2)-(3), the unmarked form fails rather strikingly to refer to the superordinate category. In these pairs, the masculine form really is male-denoting, by the same criteria that the masculine nouns in (2)-(3) are not. Compare (11)-(12) to (4)-(5):

(11) a. Is there a prince in that photograph?
b. #Yes, namely Princess Anne.

(12) a. Is there a princess in that photograph?
b. #Yes, namely Prince William.

In the recent literature, a variety of studies of various languages have converged on accounts of the above patterns which incorporate some version of markedness for one class of nouns, but which allow for additional types of representation to accommodate nouns that fail to show an asymmetry under semantic markedness diagnostics, as in (11)-(12) (e.g., Bobaljik & Zocca 2011, Merchant 2014, Spathos & Sudo 2016, and Saab 2017, among others). In these proposals, across the various classes, nouns that are morphologically marked with a (feminine derivational) suffix are also semantically marked for (female) sex.[3] In the *actress* class, the nouns that are

---

Speaker B: # No, it's my sibling. # No, it's a ram/ewe.

[3] These studies recognize a third class of nouns, not attested in English (but attested in languages like Brazilian Portuguese), in which there is no semantic markedness asymmetry under the ellipsis diagnostic to be discussed below, but in which both unmarked and marked forms behave as if they are semantically unspecified for sex, as opposed to *prince(ss)*-type nouns where both appear to be specified (further filling out the paradigm of possible options). Loosely speaking, in nouns of this third class, gender morphology behaves as if it is inflectional—supplied by

morphologically unmarked are also semantically unmarked, i.e., unspecified for sex. Jakobson's appeal to Gricean-like reasoning describes their distribution. In the *princess* class of nouns, both forms are semantically marked for sex, regardless of the morphology (and in fact, not all members of this class show affixal morphological alternations, such as *king/queen*). A noun like *prince*, despite being morphologically unmarked, is semantically specified as denoting only the male with the appropriate noble rank. It therefore cannot be used to denote the sex-neutral superordinate term (roughly: monarchical offspring).

Though some version of markedness (or the converse—underspecification) is a component of many grammatical theories, there are competing theories that attempt to explain gender asymmetries without recourse to markedness. For example, Haspelmath 2006 proposes that the semantic asymmetries observed with the nouns above may be caused by asymmetries in the frequency of the two forms of the noun, embedding the discussion in a proposal to eliminate the concept of semantic markedness from the theory altogether. Under Haspelmath's theory, the asymmetry observed for *actress*-class words occurs because the unmarked form (*actor*) is much more frequent than the marked form (*actress*), leading, in a way that is not completely specified, to a wider semantic meaning for the unmarked form. Similarly, the lack of asymmetry for *princess*-class words is due to the unmarked (*prince*) and marked (*princess*) forms having relatively equal frequencies. While Haspelmath's short discussion is light on specifics, the relative frequency theory makes the strong prediction that as the relative frequency of the unmarked form over the marked form increases, so too should the asymmetry that we observe under the form:meaning diagnostics, like those in (4)-(9).

At a fundamental level, the markedness approach and frequency approach make different empirical predictions, and it is these we set out to test here. For the representational markedness approach, the lexical semantic representation of the masculine, typically morphologically unmarked noun in a given pair (other than *widow/widower*, on which see below) has one of two options: either it bears the specification "male" or it is unspecified.[4] Its behavior in frames such as the ones discussed above should either pattern with *prince* or with *actor*. There are two possible representations, so judgments should be categorical (in the ideal, i.e., up to speaker uncertainty, variation, and other sources of "noise"). Bobaljik & Zocca (2011) argue moreover that there is an internal unity to various semantic fields: profession nouns and animal names behave like *actor/actress*, and nobility titles and kinship nouns pattern like *prince/princess*.[5] By contrast, under a frequency approach like Haspelmath's, we expect to observe a far more gradient landscape, in which the judgments of "semantic (un)markedness" should correlate strongly with the relative frequency of the unmarked:marked or masculine:feminine members of the opposition in some suitably representative corpus.

We test these predictions by first experimentally quantifying the gender asymmetries for 16 noun pairs in English (8 putatively *actress*-class and 8 putatively *princess*-class), and then

---

context—rather than derivational—inherent to the (possibly derived) meaning of the noun. Since this does not arise in English, we leave this interesting topic aside here.

[4] The third class, mentioned in fn. 3, is distinguished by the lexical representation of the feminine member of the opposition (again, further filling out the paradigm of possible options). Since this does not arise in English, this option is not considered here.

[5] Bobaljik & Zocca argue that the explanation for this is ultimately cultural, and not that universal grammar affords these classes as pre-determined categories. See further discussion in section 5.

comparing those quantified asymmetries to the relative frequency of the noun pairs as determined using three corpora: Hyperspace Analogue to Language (Lund and Burgess 1996), SubtlexUS (Brysbaert & New 2009), and the US English portion of Worldlex (Gimenes & New 2016).

The first step for these tests is to develop a quantifiable measure of gender asymmetry. To that end, we will adopt Bobaljik & Zocca's (2011) ellipsis test for gender asymmetries. The underlying idea of the ellipsis test is that the identity requirement on ellipsis can be leveraged to uncover these asymmetries, and crucially, convert those asymmetries into unacceptability. Asymmetric nouns such as *actor/actress* display an asymmetry under ellipsis as in (5): the unmarked form (*actor*) can be the overt antecedent for an elided predicate that agrees with a female-biased name (5a), but the marked form (*actress*) cannot be the antecedent for an elided predicate that agrees with a male-biased name (5b).

(5) a. John is an actor, and Mary is too.
b. *Mary is an actress, and John is too.

For symmetric nouns such as *prince/princess*, both combinations are unacceptable under ellipsis:

(6) a. *John is a prince, and Mary is too.
b. *Mary is a princess, and John is too.

In this way, the Bobaljik & Zocca ellipsis test can be used to convert the gender asymmetry into an easily quantifiable acceptability effect that distinguishes two classes of nouns in English, while simultaneously avoiding the methodological complexities that Jakobson's original question-answer diagnostics would raise.

We have three goals in this paper. The first goal is descriptive: to develop (and deploy) a formal experimental design for the Bobaljik & Zocca (2011) ellipsis test that we can then use to empirically determine the class of 16 noun pairs in English, using both a qualitative metric (the presence/absence of an asymmetry effect) and a quantitative metric (hierarchical clustering of the noun pairs based on their ratings across conditions). The second goal is methodological: to evaluate Bobaljik & Zocca's (2011) suggestion that ellipsis permits gender mismatches that are not otherwise tolerated. The third goal is theoretical: to use the quantified gender asymmetries from the experiment to test the gradient predictions of the Haspelmath (2006) relative-frequency theory. With those goals in mind, the rest of this paper is organized as follows. In section 2 we review the Bobaljik & Zocca analysis of the ellipsis test, and develop a factorial design that isolates the effect of each of the components of the Bobaljik & Zocca analysis. This will allow us to quantify each component, and crucially, isolate the gender asymmetry effect so that we can compare it to the relative frequency for each noun pair. In section 3 we report the details of the acceptability judgment experiment, and discuss the consequences of the results for the classification of noun pairs in English and for the Bobaljik & Zocca analysis of the ellipsis test. In section 4 we report the comparison of the isolated gender asymmetry effect with the relative frequency of each noun pair as calculated from three English corpora: Hyperspace Analogue to Language (Lund and Burgess 1996), SubtlexUS (Brysbaert & New 2009), and the US English portion of Worldlex (Gimenes & New 2016). Anticipating the results slightly, we find (i) that there are three noun pairs (out of 16) that behave differently than expected (though we offer some thoughts about why this may be); (ii) that ellipsis is not required for mismatches to be

tolerated (contra Bobaljik & Zocca 2011); and (iii) that the relationship predicted by the Haspelmath (2006) relative frequency theory does not hold (instead, we find that both classes of nouns are intermixed along the range of relative frequencies[6]). This suggests that the Haspelmath relative frequency theory is not an adequate empirical competitor with the markedness theory for explaining gender asymmetry effects. Section 5 concludes with a short recap of the findings, and brief discussion of future directions for formal experimental investigations of gender asymmetry effects.

2. THE ELLIPSIS TEST FOR GENDER ASYMMETRIES

There are three components to the Bobaljik & Zocca (2011) analysis of the ellipsis test: (i) an identity requirement that holds between the antecedent and the elided material (see Merchant 2017 for a recent overview, (ii) a theory that acknowledges (at least) a three-way contrast in possible values for semantic sex, even in a grammatically two-gender system, including "unspecified" alongside "male" and "female", and (iii) some version of a principle like *Maximize Presupposition* (Heim 1991), which states that an utterance must contain the form of the noun that carries the strongest presupposition that is compatible with the given context.[7] To make the discussion concrete, we will work through the four critical sentences previously presented in section 1 before developing a formalization of the ellipsis test for our experiments.

For ellipsis constructions containing the unmarked, masculine form of an asymmetric noun such as *actor* in (7a), there are two possible resolutions, as in (7b) and (7c) (material in angled brackets is interpreted but unpronounced).

(7) a. John is an actor, and Mary is too
b. John is an actor, and Mary is <an actor> too.
c. *John is an actor, and Mary is <an actress> too.

In (7b), the unmarked form *actor* in the elided material satisfies the identity requirement. Furthermore, under the assumption that *actor* is unspecified for sex, there is no infelicity between *Mary* and *actor*. The only question is whether (7b) satisfies Maximize Presupposition – is it the strongest form compatible with the context? Although *actress* would have a stronger presupposition (i.e, female), Maximize Presupposition is subordinate to the grammatical principle that requires identity under ellipsis. Thus, the form *actress* is not part of the competition at all, and in this context, *actor* trivially satisfies Maximize Presupposition.

The converse configuration in (8a) also has two possible resolutions, as in (8b) and (8c).

(8) a. Mary is an actress, and John is too

---

[6] Thus experimentally confirming the preliminary findings in Bobaljik & Zocca (2011:155) from informal judgments and a smaller sample.

[7] Whether nominal sex is part of the asserted or presupposed meaning is not directly relevant to the issue considered here. What is relevant is that it enters into a competition of the Maximize Presuppostion / Gricean / Jakobsonian type. For concreteness, we will write below as if the issue is a matter of presuppositional meaning, although this is not a position we wish to assert. Among the references cited, see Sudo & Spathas 2016 for whom the distinction between asserted and presupposed aspects of meaning is fairly central.

b. *Mary is an actress, and John is <an actor> too.
c. #Mary is an actress, and John is <an actress> too.

In (8b), the unmarked resolution (*actor*) violates the identity requirement. In (8c), the marked resolution (*actress*) leads to infelicity because *actress* is marked female, while *John* is male-biased. Unlike (7), the sentence in (8) has no felicitous resolution, and is therefore judged unacceptable.

For symmetric nouns such as *prince/princess* in (9) and (10), there is no resolution that is both felicitous and grammatical.

(9) a. John is a prince, and Mary is too.
b. #John is a prince, and Mary is <a prince> too.
c. *John is a prince, and Mary is <a princess> too.

(10) a. Mary is a princess, and John is too.
b. *Mary is a princess, and John is <a prince> too.
c. #Mary is a princess, and John is <a princess> too.

Under the assumption that both forms of symmetric nouns are marked for sex, one of the two resolutions will be infelicitous, and the other will violate the identity requirement. Again, the key difference is in the lexical semantic representation of the morphologically unmarked terms: *prince* is semantically specified as male, while *actor* is unspecified for sex.

To formalize the ellipsis test, we constructed a 2x2x2 factorial design. Table 1 provides concrete examples using *actor/actress*. At a descriptive level, what we manipulated was the grammatical gender of the predicate noun and the semantic sex of the subject of the second clause (i.e., by using stereotypically biased names). We must keep in mind that gender as such is a bit misleading, since the critical factors are not the genders and sexes as such, but rather (i) whether the predicate noun in question is morphologically "unmarked" or "marked" and (ii) whether the predicate noun and the subject noun match in gender-sex or not.[8] Therefore we have

[8] Two qualifications should be made here. For the purposes of classification, along with much of the literature cited, we treat suppletive pairs like *king/queen* as if they stand in a morphological unmarked:marked relation. In addition, where morphological marking is affixation versus bare, in English and the languages investigated in the works cited, the morphologically marked form generally denotes the female sex. An exception is the pair widow/widower. Such cases are known as "markedness reversals" or "local markedness" (Tiersma 1982) in the literature. Jakobson (1932) and Bobaljik & Zocca (2011) do not discuss markedness reversals, but we included one example in the experiment. If morphological markedness alone were the sole relevant factor, we would expect *widow/widower* to show the same asymmetry as *actor/actress* when viewed from the perspective of morphological form (which is the reverse of the asymmetry when viewed from the perspective of gender). Note that we do not exclude a role for frequency in relating real-world categories to morphological markedness -- it seems reasonable to assume that cultural norms and, in the case of widows, life expectancy, play a role in determining which member of the pair will be morphologically unmarked. Where we find no role for frequency is in determining when the morphologically unmarked form will behave as if it is semantically

decided to name the first two factors MARKEDNESS and MISMATCH to better reflect the underlying logic of the design. Although this superficially appears to favor the markedness theory over the relative frequency theory, we would like to note that this is simply a terminological choice that will ultimately prove more convenient given our results. We could just as easily adopt the terminology of the relative frequency theory for our factors. The crucial question is whether the theories in question can explain the effects that obtain by manipulating the grammatical gender of the predicate NPs and the semantic gender/sex of the subject NPs in this systematic way.

sentence	MARKEDNESS	MISMATCH	ELLIPSIS
John is an actor and Bill is too.	unmarked	match	ellipsis
John is an actor and Mary is too.	unmarked	mismatch	ellipsis
Mary is an actress and Sue is too.	marked	match	ellipsis
Mary is an actress and John is too.	marked	mismatch	ellipsis
John is an actor and Bill is an actor too.	unmarked	match	non-ellipsis
John is an actor and Mary is an actor too.	unmarked	mismatch	non-ellipsis
Mary is an actress and Sue is an actress too.	marked	match	non-ellipsis
Mary is an actress and John is an actress too.	marked	mismatch	non-ellipsis

Table 1: A 2x2x2 factorial design for the ellipsis test of gender asymmetries using *actor/actress* as an example.

The primary manipulation in this design is the interaction between MARKEDNESS and MISMATCH. When the predicate NP and subject NP match in gender and sex-bias, we expect high acceptability, regardless of noun class. In this way, the match conditions form a sort of baseline for highlighting the effect of mismatch. For symmetric nouns, we expect a mismatch between the predicate NP and subject NP in sex specification and sex-bias to result in a decrease in acceptability for both the unmarked and marked form of the noun, as illustrated in the top left panel of Figure 1. This is the symmetry that characterizes symmetric nouns. For asymmetric nouns, we expect a decrease in acceptability for the marked mismatch condition, but no decrease in acceptability for the unmarked mismatch condition, as illustrated in the top right panel of Figure 1. This is the asymmetry that characterizes asymmetric nouns. In statistical terms, we expect a superadditive interaction between MARKEDNESS and MISMATCH for asymmetric nouns, but no interaction for symmetric nouns.

unmarked, in the ellipsis pattern (and under other diagnostics, extending to cover the superordinate meaning).

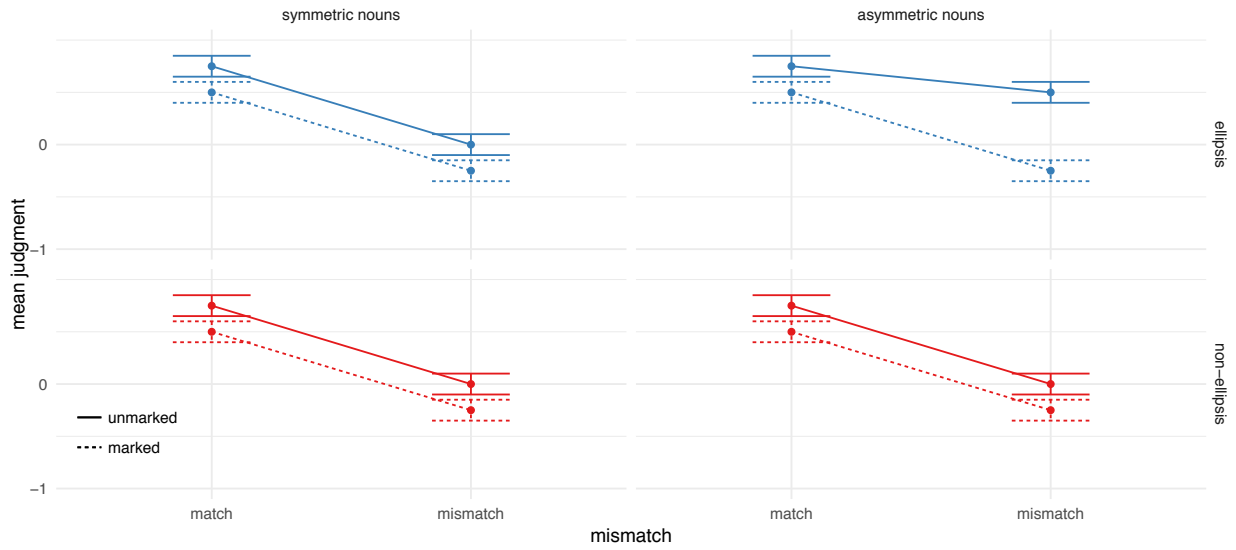

Figure 1: Expected patterns for symmetric nouns (left panel) and asymmetric nouns (right panel), under ellipsis (top row) and non-ellipsis (bottom row).

We also included a secondary manipulation in the design to test whether ellipsis is truly necessary to reveal gender asymmetries by including a third factor, ELLIPSIS, that manipulates the presence or absence of ellipsis. The Jakobsonian/Gricean analysis proposed by Bobaljik & Zocca (2011) predicts that symmetric and asymmetric nouns will show different patterns of acceptability in ellipsis constructions (top left panel versus top right panel in Figure 1), but the same pattern in non-ellipsis constructions (the bottom left and bottom right panels in Figure 1). However, Bobaljik & Zocca note that some speakers may not demonstrate the Jakobsonian/Gricean effect (i.e., some speakers may accept *Mary is an actor*). The non-ellipsis level of the factor ellipsis will test the Jakobsonian/Gricean analysis directly for the participants in our experiment.

3. THE ACCEPTABILITY JUDGMENT EXPERIMENT

There are three goals for the acceptability judgment experiment. The first is to empirically determine the class of each noun pair based on the pattern of judgments in the ellipsis conditions. We will do this in two ways: (i) by looking for the presence or absence of a superadditive effect, and (ii), by running a hierarchical clustering algorithm on the judgments for the four different conditions for each pair. The second goal is to test whether the Jakobsonian/Gricean component of the analysis proposed by Bobaljik & Zocca is needed, that is, whether the difference between asymmetric and symmetric nouns emerges reliably only under ellipsis. The third goal is to quantify the size of the asymmetry effect (the superadditive effect) for each noun pair, and then test the Haspelmath (2006) relative frequency hypothesis, which predicts that the size of the asymmetry effect will correlate with the size of the relative difference in frequency between the two forms of the noun pair. To that end, we implemented the 2x2x2 factorial design described in section 2 above in a 7-point Likert scale survey on Amazon Mechanical Turk. This section describes the details of the experiment and the results that we obtained.

3.1 *Materials*

We tested 16 noun pairs, 8 that are by hypothesis asymmetric (*actor*/*actress*, *waiter*/*waitress*, *god*/*goddess*, *widow*/*widower*, *heir*/*heiress*, *enchanter*/*enchantress*, *host*/*hostess*, *landlord*/*landlady*), and 8 that are by hypothesis symmetric (*prince*/*princess*, *king*/*queen*, *count*/*countess*, *baron*/*baroness*, *uncle*/*aunt*, *brother*/*sister*, *husband*/*wife*, *brother-in-law*/*sister-in-law*). We used the suggestion in Bobaljik & Zocca (2011) that symmetric nouns form a semantic class comprised of nobility and kinship terms to make the a priori class determinations to select our target pairs (These a priori class determinations will be quantitatively evaluated by the cluster analysis below). We constructed 8 conditions for each pair following the 2x2x2 design described in section 2 and exemplified in Tables 1 and 2. We created 8 tokens of each condition for each noun pair for a total of 1024 target items. We constructed 9 practice items that span the range of acceptability, including both agreement-centric constructions (both grammatical and ungrammatical) and other types of constructions. We also constructed 14 filler items that span the full range of acceptability, with 7 that are agreement-focused, and 7 that are of other types. The full set of materials, including the practice items and fillers, are available on the first author's website.

3.2 *Design*

We distributed the 1024 target items across experiments using a Latin Square design. Each survey contained one token of each of the 8 conditions, while each condition in a survey used a different lexical item, with four lexical items from the asymmetric class and four from the symmetric class. Each survey included the 9 practice items in the same order at the beginning of the survey (but not distinguished from the rest of the experiment), and the same 14 filler items. Each survey was 31 items long (9 practice items, 14 filler items, 8 target items). We constructed 128 distinct lists of items, and created 4 pseudorandom orders per list, for a total of 512 distinct surveys. The task was a 7-point Likert scale task. The instructions for the task are available on the first author's website.

3.3. *Participants*

3072 participants were recruited using Amazon Mechanical Turk, resulting in 6 participants for each of the 512 surveys. Participants were paid $1.00 for their participation, which, given average completion times, resulted in an hourly rate of about $12.00 per hour. The distribution of surveys and participants yielded 192 judgments per condition per noun. The large sample size is, admittedly, much more than necessary for this particular project; however, as discussed in more detail in section 5, we plan to use these results as a baseline for future studies of gender asymmetries cross-linguistically. Other languages may not allow for such a high recruitment rate, so here we take advantage of the availability of so many participants in English in order to establish well-defined baseline distributions for these noun pairs.

3.4 *Pre-processing*

We removed participants that performed substantially differently from the other participants on the 7 non-agreement fillers from the final analysis. We chose the 7 non-agreement fillers as

"gold standard" questions because (i) the items have been pre-tested on over 100 participants in Sprouse et al. 2013 so that we know what the expected value should be (on a 7 point scale), and (ii) the non-agreement fillers are completely unrelated to the target items in this experiment. To identify the outlier participants, we calculated the difference between the expected value and reported judgment for the 7 fillers for each participant, squared those values, summed the squares, and then calculated the mean and standard deviation of all of the sum-of-squares values of the participants. We defined an outlier participant as anyone with a sum-of-squares value that is greater than 2 standard deviations from the mean. This affected 4.75% of the sample (146 out of 3072 participants). We eliminated those participants from further analysis. We then z-score transformed the results of each remaining participant in order to remove two of the most common forms of scale bias (skew and compression).

3.5 *Results*

Figure 2 reports the mean ratings for each of the 16 noun pairs in our experiment for all 8 conditions: blue lines for ellipsis conditions, red lines for non-ellipsis conditions. We have organized the figure based on the a priori class of each noun pair (not based on the empirical results of our experiment). The top row contains putative asymmetric nouns, and the bottom row contains putative symmetric nouns. Recall from section 2 that we expect asymmetric nouns to show a superadditive pattern as in the top left panel of Figure 1, and symmetric nouns to show two parallel downward sloping lines as in the right panel of Figure 1. We plot both ellipsis (blue) and non-ellipsis (red) here for completeness. In the discussions that follow, we will provide distinct ellipsis and non-ellipsis plots as appropriate.

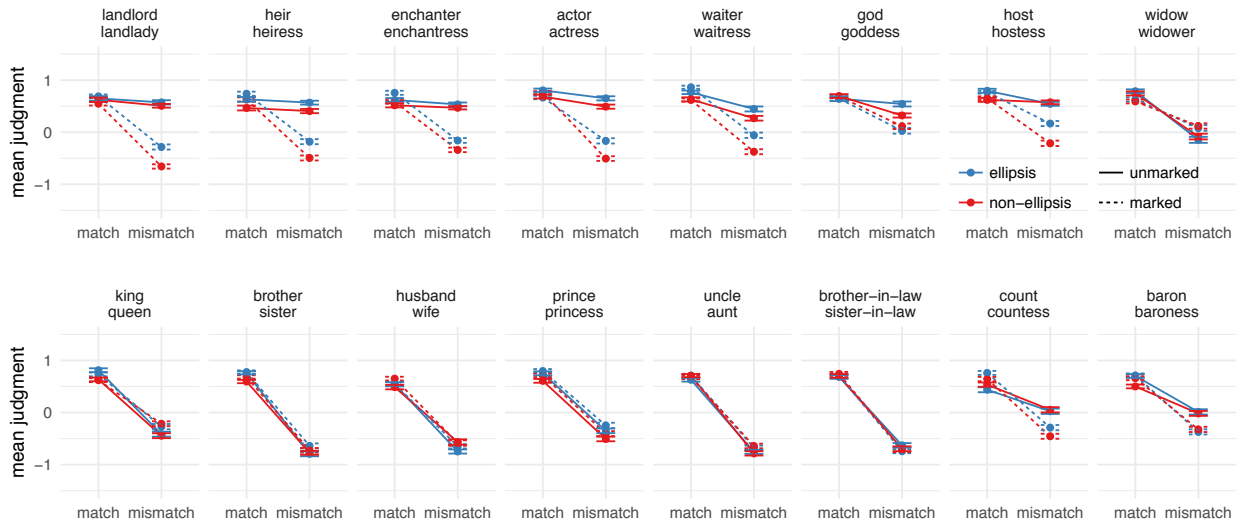

Figure 2: Mean (z-score transformed) ratings for the 2x2x2 factorial design for the 16 noun pairs. The top row contains the putative asymmetric nouns; the bottom row contains the putative symmetric nouns. Both rows are roughly internally organized by the empirical results of the experiment.

We constructed linear mixed-effects models for each noun pair with MARKEDNESS, MISMATCH, and ELLIPSIS as fixed factors, and item as a random factor (intercept only) using the lme4 package (Bates et al. 2015) for the R language (R Core Team 2015). (We could not include participants as a random effect because participants only saw one condition for each word pair. We could not include item slopes because participants only saw one item per condition). We used the lmerTest package (Kuznetsova et al. 2015) to perform omnibus ANOVAs for each noun pair using the Satterthwaite approximation of degrees of freedom. Because the omnibus ANOVAs are not part of any of the hypotheses that we are testing, we have placed the details of the omnibus ANOVAs in Appendix A. Here in the main text we focus on the specific (planned) 2x2 ANOVAs crossing MARKEDNESS and MISMATCH within each level of ELLIPSIS as required by the goals of the experiment.

3.6 *Classifying the sixteen noun pairs*

The first goal of the experiment is to classify the noun pairs as either asymmetric or symmetric according to the pattern of acceptability that they display in the four ellipsis conditions (blue lines). We plot the ellipsis conditions alone in Figure 3:

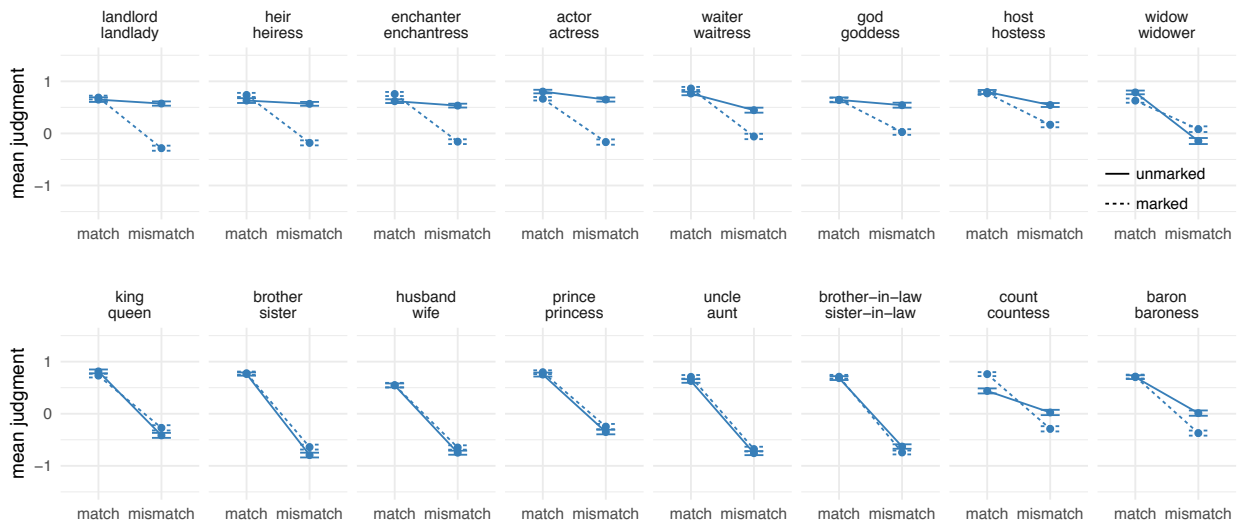

Figure 3: Mean (z-score transformed) ratings for the 2x2 factorial design for the 16 noun pairs for ellipsis conditions only. The top row contains the putative asymmetric nouns; the bottom row contains the putative symmetric nouns. Both rows are roughly internally organized by the empirical results of the experiment.

From visual inspection, we see that 7 of the 8 noun pairs in the top row appear to show the asymmetric pattern of judgments: *actor*/*actress*, *waiter*/*waitress*, *heir*/*heiress*, *enchanter*/*enchantress*, *host*/*hostess*, and *landlord*/*landlady*, and *god/goddess*. The pair *widow/widower* appears to show a small superadditive effect in the opposite direction than predicted. As noted in fn. 8 above, we included *widow/widower* because it is a well-known example of a 'markedness reversal' in that the unmarked form refers to female sex. These results seem to confirm that *widow/widower* patterns differently than the other two classes. We return to

this point briefly in section 5. We also see that two of the noun pairs in the bottom row, *count*/*countess* and *baron*/*baroness*, fail to show the symmetry pattern: *count/countess* shows a non-monotonic interaction that is similar in consequence to the asymmetry pattern; *baron/baroness* shows a small asymmetry pattern. This is a potentially surprising result given that these two nouns are nobility titles, and other nobility titles demonstrate the symmetry pattern. We speculate that this may be a reflection of less familiarity with *count(ess)* and *baron(ess)* as nobility titles (and accordingly some speaker uncertainty in the semantic representation) for the average North American AMT user, and return to this point briefly in section 5.

To quantify these visual impressions we constructed linear mixed-effects models (using treatment coding) with MARKEDNESS and MISMATCH as fixed factors and item as a random factor (intercepts-only), but only within the ellipsis conditions. We again used the lmerTest package to calculate *p*-values using the Satterthwaite approximation of degrees of freedom. We can then look for significant superadditive interactions as an index of the gender asymmetry pattern. Table 2 below lists the model estimates and *p*-values for the intercept (the mean rating of the unmarked, match condition), the simple effect of MARKEDNESS, the simple effect of MISMATCH, and the interaction of the two for each noun pair.

	INTERCEPT		MARKEDNESS		MISMATCH		INTERACTION	
word	β	*p*	β	*p*	β	*p*	β	*p*
landlady	0.65	.001	-0.04	.682	-0.07	.467	-0.90	.001
heiress	0.63	.001	-0.11	.312	-0.06	.610	-0.86	.001
enchantress	0.62	.001	-0.14	.036	-0.09	.184	-0.83	.001
actress	0.80	.001	0.14	.016	-0.15	.007	-0.68	.001
countess	0.44	.001	-0.32	.001	-0.41	.001	-0.64	.001
waitress	0.77	.001	-0.10	.106	-0.32	.001	-0.60	.001
goddess	0.64	.001	-0.00	.982	-0.10	.241	-0.52	.001
baroness	0.70	.001	-0.01	.895	-0.69	.001	-0.39	.001
hostess	0.80	.001	0.03	.744	-0.25	.007	-0.35	.007
sister-in-law	0.68	.001	-0.03	.629	-1.31	.001	-0.14	.081
aunt	0.63	.001	-0.08	.146	-1.39	.001	0.00	.986
princess	0.75	.001	-0.05	.488	-1.10	.001	0.06	.529
wife	0.54	.001	-0.01	.931	-1.29	.001	0.09	.267
sister	0.76	.001	-0.02	.783	-1.56	.001	0.14	.082
queen	0.81	.001	0.08	.255	-1.23	.001	0.23	.029
widow	0.79	.001	-0.16	.022	-0.93	.001	0.38	.001

Table 2: Results of a 2x2 linear mixed effects model for MARKEDNESS x MISMATCH for ellipsis conditions only, using the lmerTest package, and treatment coding (with match and marked as reference levels). The table is ordered based on the size of the asymmetry effect (the superadditive interaction).

The statistical results confirm what we see through visual inspection: *actor*/*actress*, *waiter*/*waitress*, *heir*/*heiress*, *enchanter*/*enchantress*, *host*/*hostess*, and *landlord*/*landlady*, and *god/goddess* all show significant superadditive interactions indicative of the asymmetry pattern. *Count/Countess* and *baron/baroness* also unexpectedly show significant superadditve

interactions indicative of the asymmetry pattern. *Brother/sister-in-law, uncle/aunt, prince/princess, husband, wife,* and *brother/sister* show no significant interaction, indicative of the symmetry pattern. *King*/*queen* does show a significant superadditive interaction, but this appears to be a consequence of the very high sample size that we collected, and very small differences among the conditions that align in a non-monotonic interaction pattern. W*idow/widower* also shows a significant non-monotonic interaction, but it is in the opposite direction to the predicted asymmetry pattern. This pattern is not predicted, so it is not possible to offer a definitive interpretation of this effect without additional experiments designed to tease apart potential explanations (e.g., that *widow* is *princess*-class, but perhaps *widower* is too unfamiliar to provide definitive *princess*-class results).

We can also use a hierarchical agglomerative clustering analysis to quantify the similarity among the noun pairs based on the acceptability judgments of the four conditions. This is crucially different from the analysis above. In the analysis above, we separated the pairs into two groups based on whether there is a superadditive interaction in the direction predicted by the asymmetry pattern. This cluster analysis instead looks for similarity (across the noun pairs) in the acceptability ratings of their four conditions directly. We ran a hierarchical agglomerative clustering analysis using the hclust() function in base R, with complete (or maximum) linkage, and Euclidian distances. With complete linkage, the distance between each cluster and its sister is computed based on the item within the cluster that leads to the maximum distance with the sister. This means that the dissimilarity values for each cluster (the point where the sisters join on the x-axis) is a maximum dissimilarity value for the cluster; the items within the cluster will have that dissimilarity value or less. Figure 4 shows a graphical representation (a dendrogram) of the results.

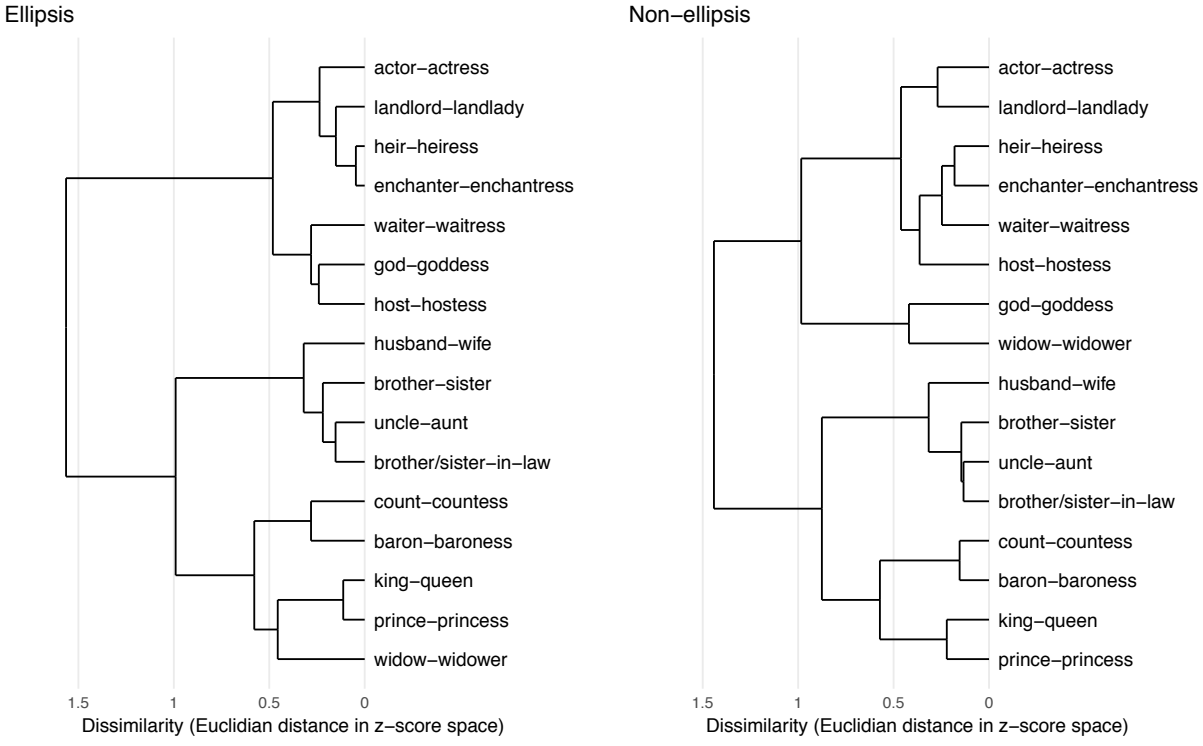

Figure 4: A hierarchical agglomerative clustering analysis of the 16 noun pairs, using complete (maximum) linkage and Euclidian distances.

For the ellipsis conditions, if we look at the two superordinate clusters, the results accord well with our a priori categorizations: kinship terms and nobility titles cluster together, and the profession-like terms cluster together. The one unexpected result is *widow/widower*, which clusters with the kinship terms and nobility titles. There is also structure within the kinship/nobility subordinate cluster: the four kinship terms cluster together, and the four nobility terms cluster together. Interestingly, *count/countess* and *baron/baroness* cluster with the kinship/nobility terms, and not with the profession-like terms. This is in contrast to the classification based on superadditive interactions above. This suggests that while *count/countess* and *baron/baroness* may show superadditive interactions, the acceptability of their four conditions is closer to the acceptability of the four conditions of the other kinship/nobility terms. This in turn suggests that these two pairs truly are in between the two categories. We also included the non-ellipsis conditions in Figure 4 in anticipation of the discussion of these conditions in section 3.7 below. There is substantial similarity in the clusters for the ellipsis and non-ellipsis conditions. The only difference for the two superordinate clusters is in the classification of *widow/widower*: *widow/widower* clusters with kinship/nobility under ellipsis, but with the other terms in non-ellipsis constructions. This is not surprising given that *widow/widower* also shows an unpredicted superadditive pattern. Given that the pair has no clear status on its own, its classification appears to be driven by a difference in the *god*/*goddess* pair, which shows a larger superadditive effect for ellipsis than non-ellipsis, and creates a cluster with *widow/widower* in the non-ellipsis conditions.

3.7 *Evaluating the role of ellipsis in revealing gender asymmetries*

The second goal of the experiment is to test the Jakobsonian/Gricean analysis proposed by Bobaljik & Zocca (2011). Their analysis predicts that the gender asymmetry pattern should disappear in the non-ellipsis conditions. The Jakobsonian/Gricean analysis predicts that the unmarked mismatch conditions for both asymmetric and symmetric nouns will violate a principle of the grammar: for symmetric nouns, there will be a gender agreement violation in the second clause (*John is a prince and Mary is a prince too*); for asymmetric nouns, there will be a Maximize Presupposition violation in the second clause (*John is an actor and Mary is an actor too*). However, based on the results of this experiment, these predictions do not appear to hold. Figure 5 shows interaction plots that isolate the non-ellipsis conditions.

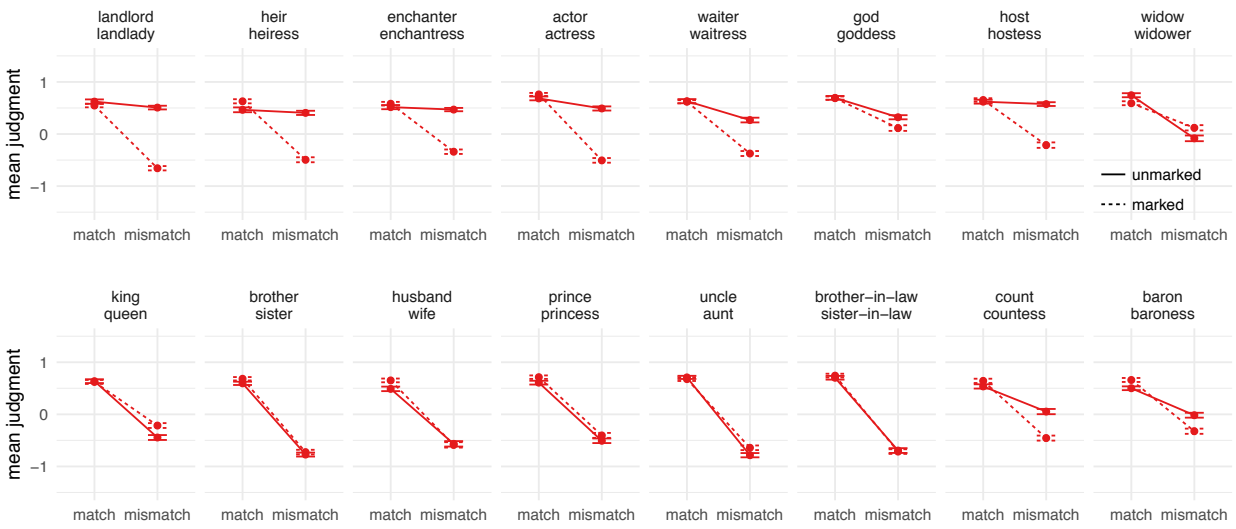

Figure 5: Mean (z-score transformed) ratings for the 2x2 factorial design for the 16 noun pairs for non-ellipsis conditions only. The top row contains the putative asymmetric nouns; the bottom row contains the putative symmetric nouns. Both rows are roughly internally organized by the empirical results of the experiment.

Asymmetric nouns in non-ellipsis conditions still show the asymmetry pattern. Table 3 lists the model estimates and *p*-values for the intercept (the mean rating of the unmarked, match condition), the simple effect of MARKEDNESS, the simple effect of MISMATCH, and the interaction of the two for each noun pair for the non-ellipsis conditions only (parallel to Table 2 for the ellipsis conditions).

	INTERCEPT		MARKEDNESS		MISMATCH		INTERACTION	
word	β	*p*	β	*p*	β	*p*	β	*p*
landlady	0.62	.001	0.08	.298	-0.11	.120	-1.09	.001
heiress	0.47	.001	-0.16	.087	-0.06	.511	-1.06	.001
enchantress	0.52	.001	-0.07	.218	-0.05	.374	-0.87	.001
actress	0.68	.001	-0.08	.154	-0.19	.001	-1.07	.001
countess	0.54	.001	-0.11	.108	-0.48	.001	-0.62	.001
waitress	0.63	.001	0.01	.887	-0.36	.001	-0.63	.001
goddess	0.70	.001	0.01	.926	-0.37	.001	-0.20	.119
baroness	0.50	.001	-0.16	.010	-0.52	.001	-0.47	.001
hostess	0.62	.001	-0.03	.656	-0.05	.540	-0.82	.001
sister-in-law	0.70	.001	-0.04	.563	-1.40	.001	-0.06	.545
aunt	0.71	.001	0.03	.543	-1.49	.001	0.18	.023
princess	0.61	.001	-0.10	.133	-1.12	.001	0.00	.995
wife	0.49	.001	-0.16	.013	-1.06	.001	-0.18	.049
sister	0.60	.001	-0.09	.093	-1.37	.001	-0.04	.578
queen	0.64	.001	0.01	.842	-1.08	.001	0.24	.023
widow	0.74	.001	-0.15	.039	-0.83	.001	0.36	.001

Table 3: Results of a 2x2 linear mixed effects model for MARKEDNESS x MISMATCH for non-ellipsis conditions only, using the lmerTest package, and treatment coding (with match and marked as reference levels). The table is organized to match the order of results for the ellipsis conditions in Table 2.

The presence of the asymmetric pattern suggests that participants in our study accepted sentences like *John is an actor and Mary is an actor too* for asymmetric nouns (the top right point in each of the plots in Figure 5), while rejecting this construction for symmetric nouns (**William is a prince and Anne is a prince too)*. To investigate this result in more depth, we plot the distribution of judgments for both the ellipsis and non-ellipsis versions of these conditions in Figure 6 (these are the distributions for the two top-right points in Figure 2). What we are looking for is evidence of bimodality in the non-ellipsis judgments (red lines) for the asymmetric nouns. Bimodality would suggest that there may be two populations of speakers: those that accept *Mary is an actor* and those that reject it. However, we do not see any compelling evidence of bimodality in the distributions.

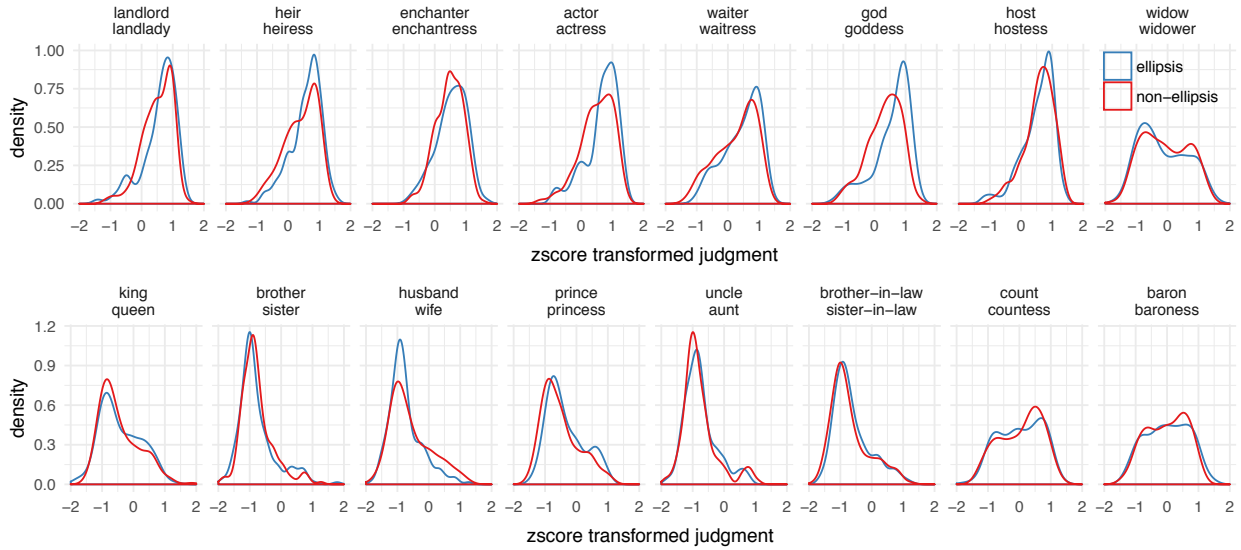

Figure 6: The distributions of judgments for the unmarked mismatch conditions for both ellipsis (*John is an actor and Mary is too*; blue) and non-ellipsis conditions (*John is an actor and Mary is an actor too*; red).

This result suggests that the role of Maximize Presupposition is less substantial than Bobaljik & Zocca assumed, and that the difference between asymmetric and symmetric nouns can be seen without ellipsis (perhaps somewhat ironically, since ellipsis served as the initial focus of our study). The asymmetry of primary interest is the difference between symmetrical nouns, which tolerate no mismatch: *#Mary is a prince*, and asymmetrical nouns, which (in principle) should allow an unmarked noun to be predicated of a female-biased subject: *Mary is an actor*. Bobaljik & Zocca argued that this difference could be seen most clearly in ellipsis, since they held that the mismatch in *Mary is an actor* disfavored when ellipsis is not involved. A Jakobsonian/Gricean logic similar to Maximize Presupposition was offered to explain the reduced acceptability of sentences like *Mary is an actor* in non-ellipsis contexts, since in those contexts (but not in ellipsis contexts) a matching alternative is available: *Mary is an actress.* Our results are

consistent with this role for Maximize Presupposition in regulating a preference when two alternatives are available, but the effect that it describes is very small; there may be a subtle preference for the matched form in non-ellipsis contexts (especially for nouns like *goddess, hostess, and waitress*), but our results indicate that it is wrong to think of sentences like *John is an actor and Mary is an actor too* as involving any kind of grammatical violation. Because ellipsis and non-ellipsis conditions both demonstrate the asymmetry effect, we will use effect sizes from both in the test of the relative frequency hypothesis presented in section 4.

As a brief aside, Figure 6 also serves to highlight the three pairs that showed unexpected results: *count/countess*, *baron/baroness*, and *widow/widower*. The less peaky, and possibly bimodal, distributions in Figure 6 suggest that participants were split in whether to treat these as gender asymmetric or not.

There was also one unpredicted result in our experiment: there is an increase in acceptability of marked mismatch conditions under ellipsis (*Mary is an actress and John is too*; bottom right points of each plot in Figure 2) compared to their non-ellipsis counterparts (*Mary is an actress and John is an actress too*) for the asymmetric nouns. These are relatively small effects. Statistically speaking, these effects should appear as a three-way interaction among all three factors in our omnibus ANOVA. However, given that these effects are so small, that three-way interaction only reaches significance for two noun pairs: *actor/actress* and *host/hostess* (see Appendix A). Neither the Bobaljik & Zocca theory nor the Haspelmath theory predicts this amelioration effect of ellipsis. To investigate this effect a little more deeply, we plot the distribution of judgments for these two conditions in Figure 7. What we see is some bimodality in both conditions, ellipsis and non-ellipsis alike, with more pronounced bimodality in the ellipsis conditions for many of the asymmetric nouns. This suggests that there may be two populations of speakers when it comes to this unexpected amelioration effect of ellipsis: those that accept *Mary is an actress and John is too*, and those that reject it. As this effect was not part of the design of the current experiment, we note it here, and set it aside for future research.

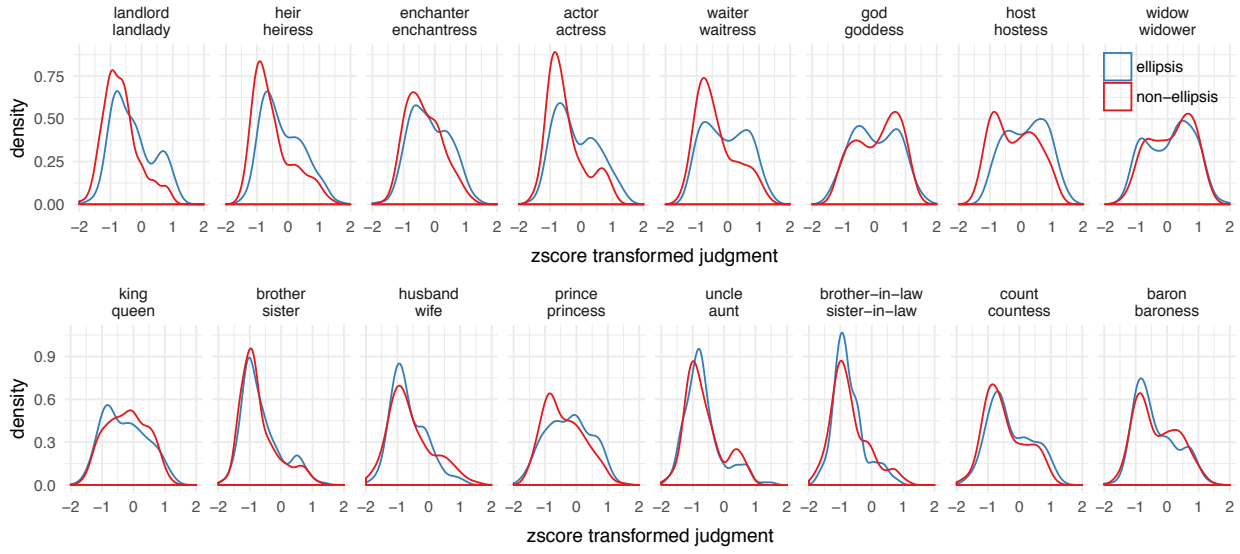

Figure 7: The distributions of judgments for the marked mismatch conditions for both ellipsis (*Mary is an actress and John is too*; blue) and non-ellipsis (*Mary is an actress and John is an actress too*; red).

4. THE RELATIVE FREQUENCY HYPOTHESIS

The Haspelmath (2006) relative frequency hypothesis proposes that the asymmetric acceptability patterns for asymmetric nouns can be explained by the relative frequency of the marked and unmarked forms. Haspelmath says "to really explain what is going on, we need to refer to a variety of factors, among them clearly frequency of use: in the pair *dog/bitch*, *bitch* has a much lower proportional frequency than *queen* has in the pair *king/queen*, so it is not surprising that it behaves more like a hyponym of *dog*." Though Haspelmath leaves the causal mechanism unstated, the idea seems to be that pairs with large differences in relative frequency are more likely to display the asymmetry pattern (as in *actor/actress*); and pairs with a smaller difference in relative frequency are more likely to display the symmetry pattern (as in *king/queen*). We can test this prediction by looking for a correlation between relative frequency of the two forms and the size of the superadditive interaction in the acceptability judgment experiment.

We retrieved the frequency of the noun forms in our study from three publicly available corpora: the Hyperspace Analogue to Language (Lund & Burgess 1996), which consists of 131 million words taken from Usenet groups in February 1995; the SubtlexUS corpus (Brysbaert & New 2009), which consists of 71 million words taken from the subtitles of US films and television programs; and the US English portion of Worldlex (Gimenes & New 2016), which consists of 104.2 million words collected from twitter, blogs, and various publicly accessible news webpages. We did not include *count/countess* and *host/hostess* in the analysis because *count* and *host* can be both nouns and verbs, and the corpora that we used are not tagged for part of speech. We also did not include *brother-in-law/sister-in-law* because the corpora that we consulted did not treat these as single items. We then calculated the log-transformed relative frequency of the marked form (e.g., *actress*) to the unmarked form (e.g., *actor*). We log-transformed the relative frequencies because this normalizes the logarithmic distribution of word frequencies in natural language. This also has the added benefit of making the numbers easy to interpret. The sign indicates the direction of the relative frequency of the marked form (e.g., *actress*): a negative sign indicates that the marked form is less frequent than the unmarked form (e.g., *actress* < *actor*); zero indicates that the two frequencies are equal (e.g., *actress* = *actor*); and a positive sign indicates that the marked form is more frequent than the unmarked form (e.g., *actress* > *actor*). The magnitude of the log relative frequency indicates the order of magnitude of the relative difference: -1 means that the marked form was 1/10 as frequent, -2 means that the marked form was 1/100 as frequent, -3 means that the marked form was 1/1000 as frequent, etc. We calculated the relative frequency in this direction (marked-to-unmarked) because Haspelmath (2006) phrases the relative frequency hypothesis in terms of the "low proportional frequency" of the marked item in the pair. It would be informationally equivalent to calculate the relative frequency in the other direction (unmarked-to-marked); and because we are working with log-transformed relative frequencies, the result would simply be a sign change. Nonetheless, we choose to keep it in Haspelmath's (2006) terms for maximum compatibility with his formulation of the relative frequency hypothesis.

The relative frequency hypothesis states that marked forms (e.g., *actress*) that are relatively less frequent than the unmarked (e.g., *actor*) should lead to the gender asymmetry pattern. The mathematical prediction is thus that negative log relative frequencies should show larger superadditive judgment effects, and positive log relative frequencies should show smaller (or no) superadditive judgment effects. In other words, we are looking for a negative correlation between log relative frequency and the superadditive judgment effects (a downward sloping line

if both quantities are plotted from smallest to largest). Figure 8 plots the correlation between the log relative frequencies and the size of the asymmetry effect from the acceptability judgment experiment (defined as the size of the MARKEDNESS x MISMATCH interaction) for each corpus (columns) and for both ellipsis and non-ellipsis conditions (rows).

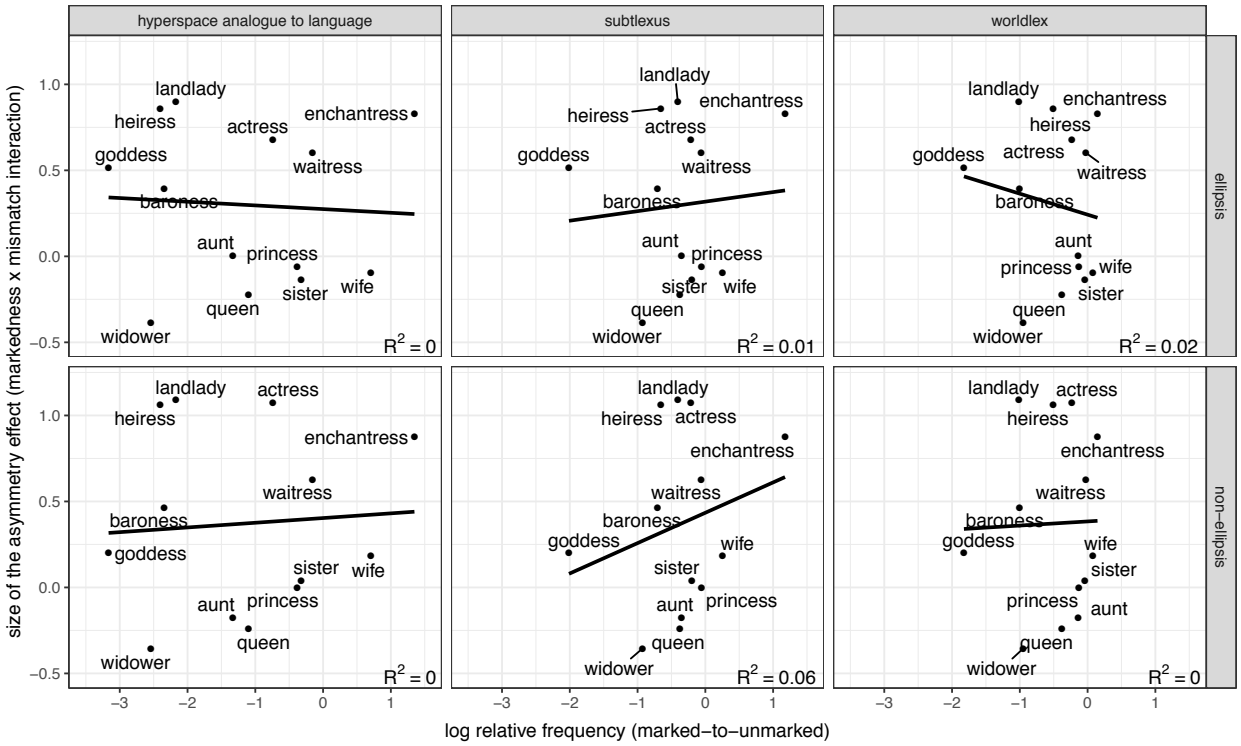

Figure 8: The correlation of log relative frequency between the marked (e.g., *actress*) and unmarked (e.g., *actor*) forms of each noun and the size of the asymmetry effect, defined as the size of the MARKEDNESS x MISMATCH interaction.

The Haspelmath (2006) hypothesis predicts a strong negative correlation between log relative frequency and the asymmetry effect: as the log relative frequency of the marked-to-unmarked form decreases, the size of the asymmetry effect should increase. But this is not what we find. The line of best fit is only negative in two of the six panels in Figure 8; in the other four, it is positive, contrary to the predictions of the relative frequency hypothesis. Furthermore, in all of the panels, the amount of variance explained by the line of best fit is exceedingly small, from less than 0.5% to 6% as indicated by the R^2 values in the bottom right of each plot. These small R^2 values suggest that there is not much of a linear relationship between log relative frequency and the size of the asymmetry effects. Although the two classes of nouns are nicely separated along the y-axis (with asymmetric nouns higher on the y-axis), the two classes of nouns appear to be well-mixed along the x-axis. In short, there is no clustering of the noun classes in different relative frequency ranges, suggesting neither a continuous nor a categorical separation of the noun classes based on relative frequency. This suggests that the relative frequency hypothesis cannot be a substantial component of the explanation of the gender asymmetry effects. Although this does not prove that the markedness-based theory is correct, it does suggest that relative

frequency is not an empirically adequate competitor with semantic markedness for the explanation of gender asymmetry effects.

5. A NOTE ON INDETERMINATE CATEGORIES IN A (BINARY) MARKEDNESS APPROACH

Our results broadly support the categorical (in fact, binary) markedness approach over the more gradient frequency approach. However, *baron/baroness*, *count/countess*, and *widow/widower* do not quite fit neatly into the two patterns predicted by the markedness approach: *widow/widower* ultimately displays a superadditve pattern that is distinct from both classes, and is classified differently by the cluster analysis in the ellipsis and non-ellipsis conditions. *Count/countess* and *baron/baroness* are nobility titles, but pattern with both classes: they show a superadditive judgment pattern like the profession class, and they cluster with the kinship/nobility class in terms of the acceptability of the four conditions. The question we raise briefly in this section is whether a binary markedness system can account for noun pairs that do not fit neatly within either of the predicted patterns. We believe the answer is yes, but it requires a closer look at the role of semantic fields in the markedness system.

To a first approximation, Bobaljik & Zocca suggested that semantic fields correlate in different ways with semantic (under)specification for sex. This claim is substantially borne out by these results, modulo the three exceptions noted above. Crucially, semantic fields (such as profession names, kinship terms, and nobility titles) are not a part of the linguistic representations we assume, but are instead cultural constructs. Bobaljik & Zocca suggest that the reason why nobility nouns, like *prince*, show symmetrical behavior is that sex is intimately tied up in the cultural contexts in which these terms are used.[9] Very loosely speaking, when a *prince* is mentioned, it is culturally relevant or salient that it is a male. By contrast, it seems likely that in many contexts in which a profession is mentioned, it is the profession itself that is culturally relevant, and sex less so. We do not know precisely how speakers acquire the fine details of lexical meaning, but we assume that language-external factors such as these play a role in shaping speakers' decisions as to whether or not to assign to a morphologically unmarked noun the semantic property "male" or to leave it underspecified. And once a range of nouns have been specified, it may be possible for learners to generalize from certain semantic patterns to new lexical items (e.g., to generalize that all nobility titles are specified for sex).

From this perspective, we can well imagine cross-linguistic variation,[10] as well as speaker uncertainty about individual lexical items, in particular where the terms denote concepts that speakers rarely encounter. Princesses and queens are well represented in the popular media, even in an ostensibly democratic republic such as the US. But counts, countesses, barons, and baronesses are probably less familiar. A speaker may easily wonder whether these are hereditary noble titles, like *prince(ss)*, or more like professional titles such as "doctor" or "chair", whose

[9] For example, in some Western European countries, there are not only clear gender-based asymmetries in the rules of succession, but also in the assignment of titles: the wife of a king takes the title of queen, but the husband of a queen does not become a king.

[10] Which we indeed find: the examples in Greenberg (1966:30-31) cited at the beginning of the article involve Spanish and Arabic, where (at least some) kinship terms pattern with the *actress* class rather than the *princess* class, distinctly from English. Both Spanish and Arabic use the plural of the noun meaning 'father' to translate 'parents', whereas in English, kinship terms as a class seem to cluster in the *princess* class.

most salient aspect is a rank of some sort, rather than the gender. We might therefore expect, for nouns of this sort, uncertainty in speakers' judgments. This uncertainty could arise in a number of ways (all of which tend to lead to non-normal looking distributions; see Dillon et al. 2017). Looking again at Figure 6, for these two noun pairs, as well as for *widow/widower*, which reverses markedness for gender and includes the fairly uncommon masculine term, this indeed appears to be what we find. Unfortunately, some of the design features of our specific experiment make it difficult to tease apart the different sources for this non-normality (e.g., one judgment per pair per participant); therefore we must leave the precise mechanism for future study. That said, it seems as though the indeterminate results for these three pairs is well-within the range of expectation for the markedness theory given the specific semantic fields that these pairs instantiate.

6. Conclusion

Our goal in this study was to develop a formal experimental version of the Bobaljik & Zocca (2011) ellipsis test for gender asymmetries, and use that experiment to (i) empirically classify 16 nouns in English, (ii) test the Bobaljik & Zocca (2011) Jakobsonian/Gricean analysis of the ellipsis test, and (iii) test the Haspelmath (2006) relative frequency hypothesis of gender asymmetries. Our results revealed that 7 of the nouns we tested were clearly asymmetrical, 6 were clearly symmetrical, and 3 were difficult to classify. The results also suggested that there is no evidence that a Jakobsonian/Gricean maxim or Maximize Presupposition effect rules out sentences such as *John is an actor and Mary is an actor too.* The results also revealed an unpredicted amelioration in sentences such as *Mary is an actress and John is too* (compared to *Mary is an actress and John is an actress too*). Finally, the results suggest that the Haspelmath 2006 relative frequency hypothesis is not an empirically adequate competitor for the explanation of these asymmetries, as there is no evidence of a correlation between the size of the asymmetry effects in the acceptability judgment experiment (defined as a superadditive interaction) and the log relative frequency of the marked-to-unmarked forms of the nouns (as retrieved from three publicly available corpora). Though this is not direct evidence for markedness, the underperformance of a popular competitor theory does help to whittle down the potential explanations for these asymmetries.

In addition to the theoretical contributions of this study, we also collected a large data set of judgments for 16 nouns and 8 conditions (in a 2x2x2 design) with 192 judgments per noun per condition. This large data set may be useful in future studies of gender asymmetries, as it can be used to establish an expected distribution for both asymmetric and symmetric nouns across these conditions. One obvious next step for this study is to test gender asymmetries in other languages, in particular Brazilian Portuguese, which Bobaljik & Zocca (2011) suggest may have three classes of nouns instead of two. Another obvious next step is to test other functionally-oriented explanations of the gender asymmetry effect beyond relative frequency, as the general goal of reducing the number of objects in the theory is an important one in the course of science.

Appendix A – Results of the omnibus ANOVAs

Table A1: The three-way ANOVA for MARKEDNESS x MISMATCH x ELLIPSIS for each noun pair (using treatment coding with unmarked, match, and ellipsis as reference levels).

	intercept		markedness		mismatch		ellipsis		mk x mm		mk x el		mm x el		three-way	
	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*
landlady	0.65	.001	-0.04	.640	-0.07	.403	0.04	.756	-0.90	.001	-0.12	.339	-0.04	.723	-0.19	.275
heiress	0.63	.001	-0.11	.269	-0.06	.573	0.11	.118	-0.86	.001	0.05	.730	0.00	.976	-0.20	.313
enchantress	0.62	.001	-0.14	.012	-0.09	.120	0.14	.060	-0.83	.001	-0.07	.353	0.04	.620	-0.04	.684
actress	0.80	.001	0.14	.013	-0.15	.006	-0.14	.032	-0.68	.001	0.21	.006	-0.04	.609	-0.39	.000
countess	0.44	.001	-0.32	.000	-0.41	.001	0.32	.142	-0.64	.001	-0.22	.024	-0.07	.458	0.02	.874
waitress	0.77	.001	-0.10	.182	-0.32	.001	0.10	.068	-0.60	.001	-0.11	.288	-0.04	.685	-0.03	.854
goddess	0.64	.001	-0.01	.982	-0.10	.244	0.01	.545	-0.52	.001	-0.01	.937	-0.27	.029	0.32	.070
baroness	0.70	.001	-0.01	.874	-0.69	.001	0.01	.004	-0.39	.001	0.15	.123	0.17	.075	-0.07	.597
hostess	0.80	.001	0.03	.726	-0.25	.003	-0.03	.031	-0.35	.003	0.06	.590	0.21	.075	-0.47	.005
sister-in-law	0.68	.001	-0.03	.667	-1.31	.001	0.03	.768	-0.14	.116	0.01	.877	-0.08	.347	0.08	.521
aunt	0.63	.001	-0.08	.144	-1.39	.001	0.08	.138	0.01	.986	-0.11	.145	-0.11	.159	0.18	.105
princess	0.75	.001	-0.05	.494	-1.10	.001	0.05	.035	0.06	.536	0.06	.531	-0.01	.885	-0.06	.656
wife	0.54	.001	-0.01	.932	-1.29	.001	0.01	.371	0.09	.278	0.16	.071	0.24	.007	-0.28	.025
sister	0.76	.001	-0.02	.776	-1.56	.001	0.02	.002	0.14	.073	0.07	.341	0.19	.014	-0.18	.098
queen	0.81	.001	0.08	.256	-1.23	.001	-0.08	.015	0.23	.027	0.07	.508	0.15	.142	0.02	.903
landlady	0.79	.001	0.16	.026	-0.93	.001	-0.16	.537	0.38	.001	0.01	.968	0.11	.282	-0.03	.832

References

Bobaljik, Jonathan David & Cynthia Levart Zocca. 2011. Gender markedness: the anatomy of a counter-example. *Morphology* 21(2), 141–166 .

Brysbaert, Marc. & Boris New. 2009. Moving beyond Kucera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods* 41 (4), 977–990.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Douglas Bates, Martin Maechler, Ben Bolker, & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

Gimenes, Manuel, & Boris New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods* 48(3), 963–972.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1), 25–70.

Heim, Irene. 1991. Artikel und Definitheit. In Armin von Stechow & Dieter Wunderlich (eds.), *Semantik: Ein internationals Handbuch der zeitgenössichen Forschung*. Berlin: Mouton de Gruyter.

Jakobson, Roman. 1932. Zur Struktur des russischen Verbums. *Charisteria Gvilelmo Mathesio qvinqvagenario a discipulis et Circuli Lingvistici Pragensis soladibus oblata: 74-84.* Prague. [English translation published as: Jakobson, Roman (1984) Structure of the Russian verb, in Linda R. Waugh & Morris Halle (eds.), *Roman Jakobson: Russian and Slavic Grammar*. *Studies 1931-1981*, 1–14. Berlin: Mouton de Gruyter.]

Kuznetsova, Alexandra, Per Bruun Brockhoff, & Rune Haubo Bojesen Christensen. 2017. lmerTest: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82(13), 1–26.

Lund, Kevin, & Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods* 28(2), 203–208.

Merchant, Jason. 2014. Gender mismatches under nominal ellipsis. *Lingua* 151, 9–32.

Merchant, Jason. 2016. Ellipsis: A survey of analytical approaches. To appear in Jeroen van Cranenbroeck & Tanja Temmerman (eds.), *The Oxford Handbook of Ellipsis*. Oxford: Oxford University Press.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Saab, Andrés. 2017. Nominal ellipsis. To appear in Jeroen van Cranenbroeck & Tanja Temmerman (eds.), *The Oxford Handbook of Ellipsis*. Oxford: Oxford University Press.

Sudo, Yasutada and Giogros Spathas. 2016. Gendered nouns and nominal ellipsis in Greek. Ms., University College London and Universität Stuttgart/Humboldt Universität zu Berlin.