Quantifying relational nouns in corpora[1]

Lelia Glass

School of Modern Languages, Georgia Institute of Technology

[1]I am indebted to the two anonymous reviewers as well as audiences at the University of Düsseldorf, the University of Rochester, and the Linguistic Society of America Annual Meeting (2022) for constructive comments. Thanks also to Sebastian Löbner and Scott Grimm for inspiring discussion. Errors are mine.

Abstract

While relational nouns (*cousin*) are traditionally delineated in a binary and theory-dependent manner, this paper approximates relationality as a continuous, objective corpus metric (Percent Possessive) – allowing for lexicon-wide exploration of which nouns are more or less relational and why. Comparing across nouns and accounting for the ontological class of the noun's referent (focusing on nouns denoting artifacts, natural kinds, occupations, humans, and locations), I find that Percent Possessive is positively correlated with a noun's per-million word frequency. Comparing across different web communities, I find that a noun is more frequent, and shows a greater ratio of definite to indefinite tokens, in the community where its Percent Possessive is significantly higher. I take these findings to be consistent with the claim that a noun is more easily interpreted as relational (as measured by Percent Possessive) when human interaction with its referent is more conventional (as measured by its frequency and definite-to-indefinite ratio). Inspired by the many authors who have suggested a socio-cultural component to relationality and possession, this paper explores at scale in English how nouns reflect the conventions of the people who use them.

Keywords: relational nouns, possessives, lexical semantics, language variation, corpus linguistics

1. INTRODUCTION

Among nouns, it is common to distinguish between conceptually one-place 'sortal' nouns such as *tree*, characterizing individuals, and conceptually two-place 'relational' nouns such as *cousin*, describing a relation between individuals (Löbner 1985; Löbner 2011; De Bruin & Scha 1988; Barker 1992, 2011). This distinction has widespread consequences for lexical and compositional semantics, most notably the interpretation of possessives such as *my cousin* (Barker 1992; Partee & Borschev 1998; Vikner & Jensen 2002).

But in practice, the distinction between sortal and relational nouns remains hazy (§2). Even if *tree* and *cousin* are clear cases, it is much harder to classify *dog, horse, phone, doctor, town, life,* or the rest of the lexicon, because the literature's diagnostic criteria (various judgments of grammaticality and felicity) conflict with one another or yield gradient results.

If the empirical manifestation of relational nouns could be made precise, it would be possible to ask: which nouns are relational or not, and why? Or, if relationality is taken as a gradient notion rather than a binary one: which nouns are more or less relational, and why? Offering a chance to explore the longstanding intuition that possession has a socio-cultural dimension (Nichols 1988; Heine 1997; Stassen 2009; Ball 2011; Aikhenvald 2012; Karvovskaya 2018), one could also explore socially-conditioned diachronic and synchronic variation in lexical se-

mantics: which nouns become (more) relational over time; which nouns are (more) relational in some communities versus others, and why? But these questions cannot be addressed if the demarcation of relational nouns remains hazy.

To make progress, this paper offers an empirical method inspired by Löbner (2011) to approximate the notion of relational nouns (§3): the percentage of possessive versus non-possessive occurrences of that noun (*my cousin* versus *a cousin*) in a corpus. Percent Possessive (87% for *cousin*, less than 6% for *tree*) is argued to be a sensible proxy for relationality, aligning with other proposed classifications thereof. The relationality of a noun is thus approximated as a continuous, empirically defined quantity. Using Percent Possessive, it becomes possible to ask concretely: which nouns are more or less relational and why?

Of course, different ontological classes of nouns are more or less relational (as measured by Percent Possessive): kinship and body parts (*cousin, foot*) are more relational than artifacts (*phone*), occupations (*doctor*), or natural kinds (*tree*). Taking ontological class into account, this paper proposes (§4):

(1) **More conventional, more relational**

Within a given ontological class, a noun will be more relational (by Percent Possessive) when human interaction with its referent is more conventional.

For example, *phone* is argued to be more relational than *lamp* because, while

both are electronic artifacts with a canonical purpose, interaction with phones is more conventionalized in our society than with lamps; people typically have their own phone that they carry with them everywhere for communication, while people interact with lamps in various ways (using them at home to read, encountering them as scenery in public places). *Cat* is argued to be more relational than *horse* because, while both refer to natural kinds in the form of domestic animals, interaction with cats is more conventional in our society than with horses; many people keep cats for companions, while far fewer people keep horses for riding.

To approximate convention in corpus data, the paper uses a noun's per-million-word frequency (the more conventionally people interact with an entity, the more they might talk about it) and its ratio of non-possessive definite to indefinite tokens (the more conventionally people interact with an entity, the more they might treat its referent as discourse-familiar and thus definite). These metrics are used (§5) to compare across nouns in data from AskReddit, a large, general-interest discussion forum; and to compare across communities in data from Reddit's different specialty sub-forums, leveraging the assumption (Glass 2021) that conventions vary across such communities. Across nouns, it is found that within a given ontological class (such as 'artifact' or 'natural kind'), more frequent nouns are more often possessive (*phone* is more frequent and more often possessive than *lamp*, *cat* is more frequent and more often possessive than *horse*). Across communities, the same noun is

found to be more frequent and more often definite in the community in which it is significantly more often used as possessive (*knife* is significantly more often possessive in the Cooking subreddit compared to AskReddit, and is also more frequent and more often definite there). These findings are argued to be consistent with (1).

Such findings complement any theoretical treatment of relationality and possession (§6) – whether binary or gradient – by explaining which nouns are more or less (easily interpreted as) relational and why. More broadly (§7), the paper explores nouns as social artifacts, shaped by the conventions of the people who use them.

2. Conflicting empirical characterizations of relational nouns

Although relational nouns have inspired a vast literature, there is no consensus about how to tell whether a given noun should be considered relational or not. Here, I focus on the empirical behavior attributed to relational nouns, but the picture is actually more complicated because some theories (discussed below in §6) allow nouns to be type-shifted from relational to sortal or vice versa in a protean manner.

In some languages, different morphology is used for 'inalienable' possession (*my cousin*, typically associated with relational nouns) versus 'alienable' possession (*my tree*, associated with sortal nouns); generally, inalienable possession is less morphologically marked (Heine 1997) and is used for kinship, body parts, and sometimes culturally immanent artifacts (Nichols 1988; Heine 1997). Whether or not such morphology is taken to distinguish relational nouns in such languages

(which depends on one's assumptions; see Karvovskaya 2018 and Ortmann 2018 for discussion), this information is not marked in English.

Syntactically, Barker (1992) argues that a noun is relational if a construction using it in the *'s* genitive is synonymous with one using it with an *of* genitive (2). If the *of*-phrase sounds unnatural or has a different interpretation than the *'s* genitive (3), then the noun is not relational.

(2) Jane's cousin $\approx$ the cousin of Jane

(3) Jane's tree $\neq$?the tree of Jane

But the availability and interpretation of *of*-phrases actually depends on many factors above and beyond the head noun (Szmrecsanyi & Hinrichs 2008; Rosenbach 2014). *Of*-phrases are more available when the possessor is inanimate (*the leg of the table* sounds more natural than *?the leg of Jane*) and when the possessor is longer (*the cousin of the new student* sounds more natural than *the cousin of Jane*). Given such confounds, some researchers (Payne et al. 2013; Peters & Westerståhl 2013; Kolkmann 2016) argue that *of*-phrases do not reliably distinguish relational nouns.

As two-place predicates, relational nouns have been compared to transitive verbs such as *break*, which relates a subject/agent to an object/theme (*I broke the vase*; Barker 1992; Partee 1995; Asmuth & Gentner 2005; Gentner 2005). But

while transitive verbs are clearly identified by the presence of a syntactic object, relational nouns have no distinctive syntax (De Bruin & Scha 1988; Peters & Westerståhl 2013), except their debated interaction with *of*-phrases.

Semantically, Asmuth & Gentner (2005) suggest the 'fetch test': if one can go and fetch a noun's referent by looking at that individual alone (*tree*), the noun is sortal; if one needs further information about its relation to other individuals (*cousin*), the noun is relational. But is difficult to apply the 'fetch test' to abstract nouns such as *birthday* or *proposal*. The 'fetch test' would classify body parts (*hand*) as sortal when most authors consider them relational (Nichols 1988; Heine 1997). Finally, this test would classify *stranger* as relational, whereas Barker (2016) argues on the basis of the *of*-phrase data (*?the stranger of Jane*) that it is sortal.

In discourse, it is argued that relational nouns can appear as 'concealed question' complements to question-embedding predicates: *I found out her* {*birthday/?day*} (Heim 1979; Nathan 2006; Barker 2016). But Kalpak (2020) shows that such concealed questions depend on the discourse context and can involve nouns that are usually considered sortal: *I found out Jane's day* could make sense if we are trying to identify the date of Jane's thesis defense.

As another discourse property, Barker (2000) argues that a discourse-novel noun can only be introduced by a possessive if it is relational, not sortal, which he suggests is because the famously flexible relation between the possessor and the head

noun is provided lexically by relational nouns but contextually (and thus not inferable out-of-the-blue) for sortal nouns. In (4), the relational noun *daughter* itself explains how the daughter is related to the man, through the *daughter-of* relation, while it is unclear out of context how the sortal noun *giraffe* is related to the man.

(4) A man walked in. His {daughter/?giraffe} was with him. adapted Barker (2000)

But again, these diagnostics diverge: although *phone* is sortal by the 'fetch test' and the *of*-PP test, it would be relational by the 'walked in' test.

In sum, it is difficult to classify nouns as relational or sortal at scale. Different diagnostics disagree with each other and are confounded by other factors. Even with respect to a single diagnostic which depends on an acceptability judgment, some nouns fall into a gray zone rather than a clear pole; Seiler (2001) suggests that the distinction between alienable and inalienable possession (more or less parallel to the distinction between relational and non-relational nouns; Ortmann 2018) should be considered continuous rather than binary. As Partee & Borschev (2012) put it, 'The distinction is formally sharp, but the classification of nouns is not.'

In other words, the theoretical distinction between one-place and two-place predicates – $\lambda x[tree(x)]$ versus $\lambda x \lambda y\ [cousin(x,y)]$ – is binary. On the other hand, the facts argued to manifest that distinction – judgments of grammaticality and fe-

licity, distributional usage data – are gradient. Moreover, if one assumes (Vikner & Jensen 2002) that a given noun can take both denotations (through polysemy or type-shifting), then a noun may manifest a continuous preference for one denotation over the other. Given these facts, researchers face a methodological question about which construct – binary or gradient – should be taken as primary. If a researcher (i) begins with a theoretical binary, they might further aim to explain how it can be applied across the lexicon or how it might manifest in gradient judgment or usage data; if a researcher (ii) begins with gradient data, they might further aim to explain how it can be mapped to a theoretical binary. Either approach can be fruitful, and each one might offer complementary insights.

3. Percent Possessive as a proxy for relationality

Taking the second approach (ii) sketched above, this paper proposes to begin with a gradient, bottom-up construct as a way to illuminate the elusive theoretical binary distinction between relational and non-relational nouns. I propose to approximate the relationality of a noun continuously via the percentage of possessive tokens of that noun in a corpus, out of all tokens. For example, 87% of tokens of *cousin* are possessive in comments on the AskReddit web forum, versus <6% of tokens of *tree*. This metric, argued to align with other proposed classifications of relationality, can be easily computed to allow for large-scale study.

3.1 *Inspiration: Löbner's 'congruent' determination types*

Inspiring this definition, Löbner (2011) sorts the concepts denoted by nouns into four quadrants using two binary features, ±Relational (which I focus on here) and ±Unique (which I set aside). For Löbner, each type of noun matches with certain determiners congruent to its quadrant: +Unique nouns match with definites (*the sun*), +Relational nouns match with possessives (*my cousin*). A noun is predicted to be most frequent and pragmatically unmarked with determiners congruent to its quadrant. Using non-congruent determiners (*a sun, the cousin*) requires semantic type-shifting and pragmatic support, so is predicted to be less frequent.

Consistent with Löbner's prediction, Nissim (2004) finds in a corpus that over 90% of relational noun tokens are possessive. Using different methods in corpora of German rather than English, Horn & Kimm (2014) and Hellwig & Petersen (2015) report that only about 20% of relational noun tokens are possessive, though they still find that relational nouns are more likely to be possessive than non-relational nouns. Haspelmath (2008, 2017) finds that about 45% of corpus tokens of 'inalienable' (relational) English kinship and body-part nouns (*mother, wrist*) are possessive, versus less than 12% for 'alienable' (sortal) nouns such as *car* or *tree*. Jensen & Vikner (2004) and Kolkmann (2016) find that a large percentage of possessive tokens in corpora (71% for Jensen and Vikner, 51% for Kolkmann) are 'inherent' possessives, i.e., the relation that we infer between the possessor and the head is

supplied by a relational head noun: *my cousin* is the person in the *cousin* relation to me (while *my tree* is the tree related to me in a way that depends on the context). These studies all agree that relational nouns and possessives pattern together. But they disagree, contributing to the variable results, about which nouns should be considered relational in the first place (§2).

Stepping back, these studies all leverage the methodological assumption, also fundamental to this paper, that corpus frequencies offer insights relevant to the theory of meaning (de Marneffe & Potts 2017). A semantic theory of nouns and possessives might in principle aim purely to explain truth conditions with no prediction about corpus counts, but if one's theory embraces the variability within the vast array of attested noun types and possessive tokens, then such a theory might be informed or tested by corpus data alongside introspection and/or experiments. In particular, possessives allow multiple grammatical options (in English, *'s* versus *of*), so a corpus can illuminate the factors that determine which option is statistically preferred. The interpretation of a possessive depends to some extent on discourse context, so a corpus can provide examples of contexts that a researcher might not brainstorm alone. Possessive constructions implicate the entire lexicon of nouns, only a small portion of which could be studied by introspection, so a corpus offers data to match the scale of the phenomenon itself. Across the lexicon, some possessive tokens are argued on introspective grounds to be more or less marked, more or

less natural or in need of contextual support (for example, many theorists argue that possessed kinship nouns such as *my cousin* are less marked and require less contextual support than possessed natural kinds such as *my tree*), so a corpus allows such gradient claims – and the theories proposed to explain them – to be concretized, via the assumption that less-marked forms will be more frequent. It is for these reasons that many studies of possessive semantics (Jensen & Vikner 2004; Löbner 2011; Payne et al. 2013; Kolkmann 2016) make use of corpus data.

3.2 *Implementation*

Whereas other research has explored the percentage of possessive tokens of nouns already classified by the researcher as relational, this paper proposes to do the opposite: to start off with no assumption about which nouns are or are not relational, and then to use the percentage of possessive tokens of that noun as a continuous proxy for its degree of relationality.

Concretely, I used data from Reddit (Baumgartner et al. 2020), a large, public internet discussion platform based in the United States, written in fairly standard orthography but in a style close to that of spoken American English (Herring et al. 2013). Reddit allows users to post content (questions, articles, photos, and so on) and discuss it in threads. The site is organized into sub-forums known as subreddits, ranging from large, general-interest forums such as AskReddit (with 34 million subscribers, discussing topics such as 'What movie villain did you sympathize with the

most?' and 'What is something that people don't worry about but really should?') to smaller, niche-interest forums such as Cooking (with 2.8 million subscribers, discussing topics such as 'What the heck can I do with a 1.5L bottle of port?' and 'What's your Christmas Eve menu?').

Reddit comments from January 2018 were downloaded from the PushShift repository[2], made public by Baumgartner et al. (2020) with the cooperation of Reddit itself for scientific research. Here, I focus on 5 million words of unique (non-repeated) sentences[3] sampled from AskReddit, the largest Reddit forum dedicated to general-interest topics; below, I leverage the structure of Reddit to compare smaller communities dedicated to specific interests such as Cooking.

The AskReddit text was processed using the SpaCy dependency parser (Honnibal & Johnson 2015) in Python to extract all 255,727 two-word noun phrases (*this team, a cop, your behavior*) – excluding syntactically complex possessors (*my student's computer*) and possessees (*my new computer*) for simplicity, and also keeping only the lemmatized head nouns that occur at least ten times to reduce typos. Two-word noun phrases cannot contain modifying adjectives such as *favorite*, argued (Partee & Borschev 1998; Barker 2011; Peters & Westerståhl 2013) to contribute relational meaning, so these adjectives must be reserved for future work.

[2] `https://files.pushshift.io/reddit/comments/`

[3] By analyzing only unique sentences, we ignore repeated or re-quoted comments as well as text posted repeatedly by forum moderators or bots.

Next, for each unique lemmatized head noun, the number of possessive tokens was counted (including all possessive pronouns – *my, your,* and so on – as well as the *'s* possessive). This number was divided by the total number of tokens of that lemmatized head noun. The result is the percentage of possessive tokens of that head noun type out of all tokens thereof: for example, 87% of tokens of *cousin* are possessive compared to less than 6% of tokens of *tree*. All data and code are available through the Open Science Framework at `https://osf.io/vx6t3/`.

3.3 *Comparison to existing lexical resources*

As discussed above (§2), classifying relational nouns is subjective and depends on one's theory. Therefore, I do not claim that Percent Possessive should be taken as the final truth about whether a noun is relational or not. Instead, I suggest it as a rough proxy that is just close enough – correlating strongly with existing datasets – to be useful.

First, Percent Possessive aligns intuitively with the WordNet ontology (Miller et al. 1990), which represents synonym, hyponym, and hypernym relations among nouns. Each noun was mapped to one of eight ontological classes using WordNet: *cousin* was placed into the 'kinship' ontological class because it inherits among its hypernyms the RELATIVE.N.01 node of the WordNet ontology; *tree* was placed in the 'natural kind' class because it inherits the hypernym LIVING_THING.N.01[4].

[4]After spot-checking the output of the code used to assign nouns to ontological classes using

By mapping each noun to a single ontological class, this approach must ignore the multiple senses of nouns such as *child* (young human vs. one's offspring) or *fan* (physical artifact versus human supporter) for the sake of simplicity. Figure 3.3 shows a box plot of the percentage of possessive tokens of all nouns in each ontological class. Notably, the two ontological classes with the greatest Percent Possessive – kinship (*cousin*) and body parts (*foot*) – are exactly the two classes often taken as inherently relational and inalienably possessed (Nichols 1988; Heine 1997), which suggests that Percent Possessive approximates relationality. Moreover, the plot shows that Percent Possessive is strikingly far higher for kinship and body part nouns versus all other nouns, which might be seen as a continuous manifestation of the binary formal distinction between relational and non-relational nouns from the literature (§2).

Percent Possessive also coheres with NomBank (Meyers et al. 2004), a database which records the argument structure of over 4000 nouns. I mapped NomBank's fine-grained classifications into three bins: relational ('DefRel' and 'ActRel' in their terms; N=242, of which 126 appear in AskReddit); abstract (attribute nouns such as *charisma*, de-adjectival nouns such as *difficulty*, and ability nouns such as *potential*; N=885, 468 of them in AskReddit); and non-relational (all others: *bot-*

WordNet, I also hand-wrote some exceptions to fix common classification errors related to sense disambiguation: for example, *mum* was originally classified as a natural kind (chrysanthemum), but I moved it in the 'kinship' class because it mainly refers to mothers in Reddit data. *Glass(es)* is a 'physical entity' in WordNet but I moved it to the 'artifact' class because it mainly refers to spectacles.

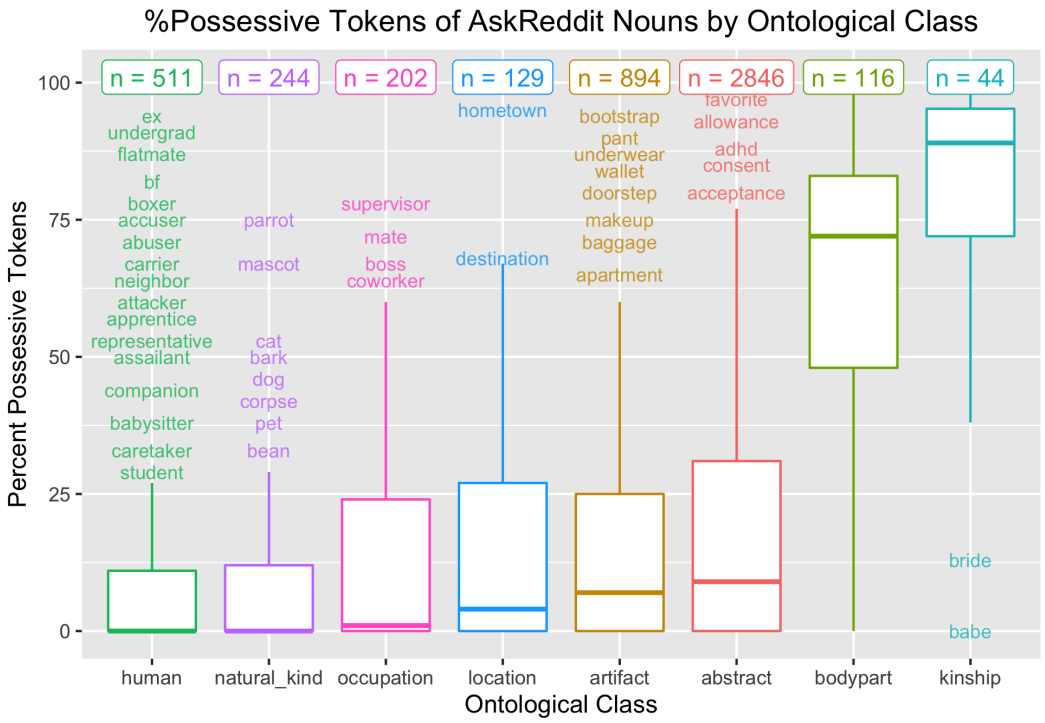

Figure 1: Box plot of the percentage of possessive tokens of all nouns in each ontological class adapted from WordNet, with outliers labeled.

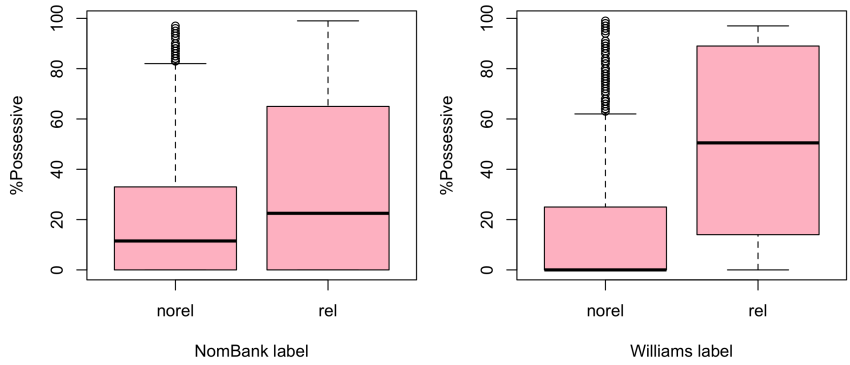

Figure 2: Percent Possessive correlates strongly with existing classifications of relational and non-relational nouns from NomBank and Williams.

tle, coastline, fugitive, goal; N=3578, of which 1800 appear in AskReddit). In a linear regression predicting a noun's Percent Possessive as a function of its bin, NomBank's 'relational' nouns (DefRel and ActRel) have a far higher Percent Possessive than its abstract or non-relational nouns (β = 12%, standard error = 2.24, $p < 0.001, R^2 = 0.01$). On average, 32% of tokens of a relational noun are possessive, compared to 20% for a non-relational noun. According to NomBank, therefore, Percent Possessive approximates relationality.

Finally, Williams (2018) offers a dataset of over ten thousand lexical items (nouns and verbs) coded as semantically one-place or two-place. I extracted the 1489 unique nouns from Williams' data that appear in AskReddit, of which 46 (*age, audience, location, uncle*) are hand-coded as relational[5]. In a linear regres-

[5]I remove from Williams' data the cases where the same noun – *sister, sibling, uncle* – is for some reason listed twice, as both relational and non-relational.

sion predicting a noun's Percent Possessive as a function of Williams' label for it, Williams' 'relational' nouns have a far higher Percent Possessive than her 'non-relational' nouns ($\beta = 34\%$, standard error = 3.60, $p < 0.001, R^2 = 0.057$). On average, 50% of tokens of a relational noun are possessive, compared to 16% for a non-relational noun. In Williams' data also, Percent Possessive coheres with relationality.

Figure 2 visualizes Percent Possessive for the nouns labeled by NomBank and Williams as relational and non-relational. The difference across these datasets (the fact that NomBank's 126 'relational' nouns have a Percent Possessive of 32% versus 50% for Williams' 46 'relational' nouns) illustrates that hand-coding nouns for relationality is highly subjective; but in both cases, Percent Possessive is much higher for the nouns labeled as relational.

In sum, I argue, Percent Possessive lines up well enough with existing datasets that it can serve as a usable proxy for relationality[6]. Moreover, while the existing datasets require laborious, subjective, categorical hand-coding, Percent Possessive has the advantages of being objective, continuous, and computed automatically.

4. Predicting the corpus distribution of relational nouns

With a usable corpus metric for the relationality of a noun, it becomes possible to test theoretically motivated predictions about which nouns should be more or less

[6]Newell & Cheung (2018) also present a large dataset of relational nouns, but – even after emails to the authors – it is not available.

relational and why.

4.1 *Ontological class matters*

As previewed above (§3.3), a noun's relationality is shaped by the ontological class of its referent. Across the literature on lexical semantics and language typology (De Bruin & Scha 1988; Nichols 1988; Barker 1992; Heine 1997; Vikner & Jensen 2002; Löbner 2011; Aikhenvald 2012; Karvovskaya 2018), kinship and body parts (*cousin, foot*) are said to be the most prototypical relational nouns, and indeed Figure 3.3 finds that they are far more often possessive than other nouns. Sometimes researchers also include as relational words for parts (*edge, top*) as well as abstract nouns (*willingness*), many of which are morphologically complex and appear to retain the argument structure of an underlying verb or adjective (Barker 1992; Vikner & Jensen 2002). Location nouns (*area, country*) are not often discussed except insofar as they describe parts (*edge*). Nouns denoting humans (*girl, American*) are seen as sortal, although some of them (*child*) are arguably polysemous with a relational kinship meaning (Barker 1992).

Artifact nouns (*phone, book, car*) are a debated category, generally classified as sortal, though Vikner & Jensen (2002) and Löbner (2011) say that artifacts are easily type-shifted to a relational denotation, relating the artifact to the person who uses it for its intended purpose (*my phone*) or perhaps the person who created it (*your book*) – the 'telic' and 'agentive' roles that Pustejovsky (1995) says are inher-

ent to artifacts. The same goes for occupation nouns (*doctor*), similar to artifacts in that they are associated with a specific purpose (Vikner & Jensen 2002).

Natural kind nouns (*tree, giraffe, cloud*) are the prototypical sortal nouns, often serving as exemplars for which a possessor-head relation cannot be recovered out of the blue: *??a man walked in with his giraffe* (Barker 1992). Natural kinds exist in nature independent of humans, and humans interact with them in many different ways (Keil 1989; Bird & Tobin 2009; Levin et al. 2019), perhaps explaining why they often do not supply a salient relation to a possessor.

Stepping back, the ontological class of a noun's referent matters for its relationality. The fact that Percent Possessive varies widely across ontological classes – highest for kinship and body parts, lowest for natural kinds – is taken as evidence that Percent Possessive is a valid proxy for relationality. As labeled by WordNet (mapping each noun to a single WordNet class, thus ignoring polysemy), ontological class is used as a predictor in the statistical models to be explored below, meaning that all other findings take it into account.

4.2 *Suggestions that possession has a socio-cultural component*

It is also often suggested that relationality and possession have a socio-cultural component. The idea of possession itself – for example, the case of legal ownership – is arguably culturally situated (Heine 1997; Vikner & Jensen 2002; Aikhenvald 2012). Moreover, it is suggested that a noun behaves grammatically as more rela-

tional when it is more conventional (within a given culture) for people to interact with its referent. Such claims set aside body parts, kinship, and abstractions to focus on nouns that are traditionally considered sortal, such as those describing artifacts and natural kinds.

Among artifacts, typologists note that it is the culturally immanent ones, such as arrows, that behave as relational and/or inalienably possessed (Nichols 1988; Heine 1997; Ball 2011). In the semantics literature, Vikner & Jensen (2002) observe that an artifact may be more acceptable as a discourse-initial possessive (taken to convey relationality) if it is more conventional for people to possess it: *my car* is more sensible out-of-the-blue than *my bus*. For Löbner (2011), *my toothbrush* is easily understood as relational because people conventionally use their own toothbrush. For Jensen & Vikner (2003) and Kolkmann (2016), the possessor of an artifact can be interpreted as either its creator (*your paper*) or its user (*your shirt*), but the more likely interpretation is chosen based on socio-cultural knowledge – for example, that it's more likely for an individual person to wear a shirt than to make one.

Among natural kinds, too, culturally basic ones are more likely to be treated as relational and/or inalienably possessed (Nichols 1988; Ball 2011), including domestic animals and tobacco. For Karvovskaya (2018), the typological possessive marking of a noun such as *rabbit* is likely to depend on whether a given culture keeps rabbits as pets. Barker (1992) cites *cat* as an example of a natural kind noun

that seems 'more relational' than prototypical sortal nouns, suggesting that *cat* may be in the process of a diachronic change from sortal to relational in view of the cultural convention of keeping cats as pets.

In sum, many researchers have found examples suggesting that a noun is more relational when human interaction with its referent is more conventional. My goal is to map this insight into predictions that can be tested at the scale of the lexicon.

4.3 *Towards testable predictions: More conventional, more relational*

Combined with the Percent Possessive metric for relationality, the preceding discussion can be synthesized into a prediction:

(5) **More conventional, more relational**

Within a given ontological class, a noun will be more relational (by Percent Possessive) when human interaction with its referent is more conventional.

But even if Percent Possessive is accepted as a reasonable proxy for relationality, we would also need a way of measuring convention for (5) to be testable.

In testing (5), I exclude nouns classified by WordNet as abstractions (*willingness*), kinship (*cousin*), and body parts (*foot*) – as well as nouns absent from WordNet – because it is not clear how people interact with abstractions and because such nouns might already be considered inherently relational. The prediction is tested

only among nouns labeled as artifacts, natural kinds, locations, humans, and occupations (*phone, tree, area, boy, doctor*): traditionally considered sortal, but also to varying degrees amenable to a relational interpretation which may be grounded in human interaction with their referents.

To explore convention empirically, I propose two proxy metrics, per-million-word frequency and definite-to-indefinite ratio. I further propose two strategies, comparing across nouns and comparing across communities. I introduce each one in turn.

The first proposed proxy for convention is the per-million-word frequency of a noun. The more conventionally people interact with something, the more they might talk about that thing.

The second proposed proxy for convention is the ratio of definite (*the*) to indefinite (*a*) tokens, among all non-possessive tokens of a noun. For example, if a corpus contained three tokens of *phone* – *my phone, the phone, a phone* – then its ratio of definite to indefinite tokens is 50% (half of its non-possessive tokens are definite).

Stepping back, the definite article *the* is argued to be used with referents that are unique, salient, familiar, easily inferred, and/or uncontroversially accommodated (Strawson 1950; Heim 1982; von Heusinger 2013; Coppock & Beaver 2015). Often (Clark 1975; Lewis 1979; Spenader 2001; Roberts 2003), *the* is used with

discourse-novel referents that are nevertheless treated as discourse-familiar because they are familiar from the wider context – from society as a whole (*the summer*), from one's specific community (*the Provost* at a university), or from the knowledge evoked by the preceding discourse (*the seat* while discussing a bike). So we might expect a greater percentage of definite tokens of nouns for which humans in a given community more conventionally interact with their referent, because such referents would be more familiar in a range of discourses. Inspired by Löbner (2011), the larger idea is that certain noun types denote referents that tend to have certain discourse properties (the noun *sun* denotes a referent that is usually unique and familiar across contexts; the noun *seat* denotes a referent that is usually familiar in a bike context), in such a way that the corpus distribution of a noun's determiners can be indirectly predicted from the typical discourse properties of its referent.

The first strategy for testing (5) is to compare the frequency and definiteness ratio across different nouns within a single corpus, namely the AskReddit discussion forum. (5) would predict a positive correlation between Percent Possessive and frequency, as well as between Percent Possessive and definiteness ratio, controlling for the ontological class of the noun's referent.

The second strategy is to compare the frequency and the definiteness ratio of the same noun across different communities. While it is not easy to measure convention in absolute terms, we can explore it in relative terms by leveraging the assumption

	Possessive	**Non-Possessive**
AskReddit	**4** *Go ahead and bring **your knife** to a gun fight*	**50** *yet another celeb who has gone under **the knife** to alter their appearance*
Cooking	**97** *Press the parsley stalks with the side of **your knife***	**121** *I never had [a peeler] before and usually did it with **a knife***

Table 1: Counts of possessive and non-possessive tokens of *knife* in both AskReddit and Cooking, along with an example of each cell. A Fisher Exact Test on this contingency table shows that *knife* is significantly more often possessive in Cooking than in AskReddit.

that conventions vary across communities (Clark & Marshall 1981). Perhaps human interaction with a given noun's referent is more conventional in one community compared to another, in which case we might expect that noun to be relatively more frequent, and relatively more often definite, in that community.

This strategy uses the sub-forum structure of Reddit, which is organized into large, general-interest forums such as r/AskReddit, as well as smaller forums dedicated to specialized interests such as r/Cooking – which I take as distinct communities of practice with their own conventions (Zhang et al. 2017; Del Tredici & Fernández 2018). Following Glass (2021), I use a Fisher Exact Test to compare the possessive and non-possessive counts of the same noun in AskReddit versus in various specialized subreddits.

For example, Table 1 shows that *knife* is used significantly more often as pos-

sessive in the Cooking subreddit than in AskReddit[7]. The Fisher Test was used to identify 341 nouns (151 of them unique) across 33 different specialty subreddits found to be significantly more often possessive at the $p < 0.01$ level in a specialty subreddit compared to AskReddit (Table 2). Assuming that Percent Possessive approximates relationality, these nouns are more relational in the subreddits in which they are more often possessive. Assuming that a noun's frequency and definiteness ratio approximate conventional interaction with its referent, we might expect a noun to be more frequent and more often definite in the subreddits where it is more often possessive.

4.4 *The predictions*

Adopting these proposed proxies for relationality and convention, along with the strategies of comparing across nouns and across communities, we arrive at four predictions. To preview §5, all but one of them are manifested.

(6) **Across nouns: More relational, more frequent** (*empirically supported*)

There should be a positive correlation between a noun's per-million-word frequency (proxy for conventional interaction) and its percentage of possessive tokens (proxy for relationality).

[7]Because I sample 5 million words from AskReddit and 1 million words from each specialty subreddit, my code requires at least twenty total tokens of a noun in AskReddit and four total in a specialty subreddit (a five-to-one ratio), so that the sample size required to find a significant difference does not bake in any bias about the frequency of the noun across subreddits.

Subreddit	**Nouns significantly more often possessive (vs. in AskReddit)**
aquariums	area, bedroom, boy, dude, fish, floor, guy, house, office, phone, pool, sword, water
babybumps	baby, child, district, doc, doctor, employer, girl, guy, hospital, kid, left, manager, shower, side, town, water
campingandhiking	dog, girl
cars	area, book, buck, girlfriend, man, site, stigma, town
chess	city, clock, king, queen, shelf, structure
christianity	lady, letter, neighbor, people, position, side, website
cooking	area, boat, boy, cabinet, city, girlfriend, house, kitchen, knife, man, pan, phone, site
diy (do-it-yourself)	backyard, bathroom, counter, garage, girl, girlfriend, kitchen, office, site, stair, website
dogs	area, bowl, boy, building, doctor, female, girl, guy, human, male, owner, people, roommate, side, therapist, vet, website
femalefashionadvice	boyfriend, city, closet, coworker, desk, end, field, letter, library, manager, office, site, therapist, top, website
filmmakers	actor, bar, film, movie, scene, site, teacher, website
fishing	girl, man, wall
fitness	bench, doctor, gym, man, phone, surgeon, trap, wheel
gardening	backyard, county, garden, girl, landlord, pet, shower, town, tree
homebrewing	area, bar, basement, counter, dishwasher, garage, kitchen, sink, site, stove, water, website
horses	area, boy, girl, guy, horse, stall
islam	book, border, citizen, judge, kid, lawyer, people, position, slave, teacher
knitting	city, counter, customer, guy, library, town, website
makeupaddiction	bag, bathroom, book, buck, cart, city, cup, customer, doctor, girl, glass(es), site, website
martialarts	book, center, doctor, gym, left, movie, site, student, teacher, website
military	citizen, guy, office, queen, rifle, shop, woman
nba (basketball)	book, boy, ceiling, fan, friend, girl, guy, man, position, side, window
nfl (football)	book, boy, ceiling, center, corner, dumbass, fan, guy, mechanic
nursing	area, car, classmate, coworker, director, floor, hospital, manager, pump, round, water, website, world
parenting	area, baby, bear, boy, button, chicken, child, doctor, girl, glass(es), guy, hospital, kid, lawyer, shoe, side, state, world
personalfinance	area, boss, child, couch, dentist, employer, end, girl, girlfriend, internet, kid, site, state, website
photography	camera, film, house, match, picture, site, state, wall, website
running	bar, base, center, doctor, glass(es), gym, house, library, phone
skiing	center, edge, film, girl, man, movie, site, tip, website
sports	boy, hero, hotel, house, lady, plane, president, short(s), stick
teachers	boss, building, car, employer, freshman, phone, seat, site, state, student
watches	book, buck, cup, customer, girl, girlfriend, machine, man, phone, plane, site, website
weddingplanning	area, buck, child, closet, doctor, dog, film, fridge, girl, guy, kid, lady, person, side, site, venue, website

Table 2: Nouns found in a Fisher Exact Test to be used as possessive significantly more often ($p < 0.01$) in a specialty subreddit compared to AskReddit (focusing only on nouns labeled as artifacts, natural kinds, humans, occupations, or locations in WordNet).

(7) **Across communities: More relational, more frequent** (*empirically supported*)

Nouns that are significantly more often possessive (proxy for more relational) in a given subreddit should be more frequent (proxy for conventional interaction) there.

(8) **Across nouns: More relational, more often definite** (*not empirically supported*)

There should be a positive correlation between a noun's definite-to-indefinite ratio (proxy for conventional interaction) and its percentage of possessive tokens (proxy for relationality).

(9) **Across communities: More relational, more often definite** (*empirically supported*)

Nouns that are significantly more often possessive (proxy for more relational) in a given subreddit should be more often definite (proxy for conventional interaction) there.

5. TESTING THE PREDICTIONS

To compare across nouns, I use data from all two-word noun phrases in 5 million words of AskReddit, focusing on those labeled by WordNet as artifacts, natural kinds, humans, occupations, or locations. To compare across communities, I com-

pare the same noun in AskReddit versus in the specialty subreddit in which it is significantly more often possessive (for all 341 nouns given in Table 2). For each noun lemma, I gather (i) its ontological class (from WordNet, ignoring polysemy); (ii) its count and percentages of possessive versus non-possessive tokens; (iii) its per-million-word frequency; and (iv) its ratio of definite to indefinite non-possessive tokens. For (ii)–(iv), I gather this information both in AskReddit and in the specialty subreddits in which that noun is used significantly more often as possessive.

5.1 *Frequency*

First, (10) is tested across nouns.

(10) **Across nouns: More relational, more frequent** (*empirically supported*)

There should be a positive correlation between a noun's per-million-word frequency (proxy for conventional interaction) and its percentage of possessive tokens (proxy for relationality).

A series of linear regression models were run in R (R Core Team 2012) predicting Percent Possessive as a function of a noun's ontological class and its per-million-word count. One model used only ontological class, one used only per-million-word count, one used both variables as additive predictors, and one included an interaction between them. Comparing models with the Aikake Information Cri-

terion (which tries to find the best balance of data coverage and parsimony), the 'best' model includes ontological class and per-million-word count as additive predictors:

(11) lm(percentPoss $\sim$ pmw + ontType, data=d)

This model, which according to its Adjusted R Squared explains 3% of the variation in Percent Possessive, is visualized in Figure 3. Looking first at the effect of the ontological class, Percent Possessive is somewhat lower than the intercept (artifacts) for human and natural kind nouns[8]. As for the effect of frequency, a noun's Percent Possessive is positively correlated with its per-million-word count ($\beta = 0.07, SE = 0.01, t = 6.0, p < 0.001$), an effect which persists when per-million-word-count is log-transformed (as in Figure 3) to reduce its skew. These findings are consistent with (10).

Next, (12) is tested across communities.

(12) **Across communities: More relational, more frequent** (*empirically supported*)

Nouns that are significantly more often possessive (proxy for more relational) in a given subreddit should be more frequent (proxy for conven-

[8]For human nouns, $\beta = -4.7, SE = 1.1, t = -4.4, p < 0.01$; for natural kinds, $\beta = -6.1, SE = 1.4, t = -4.4, p < 0.001$

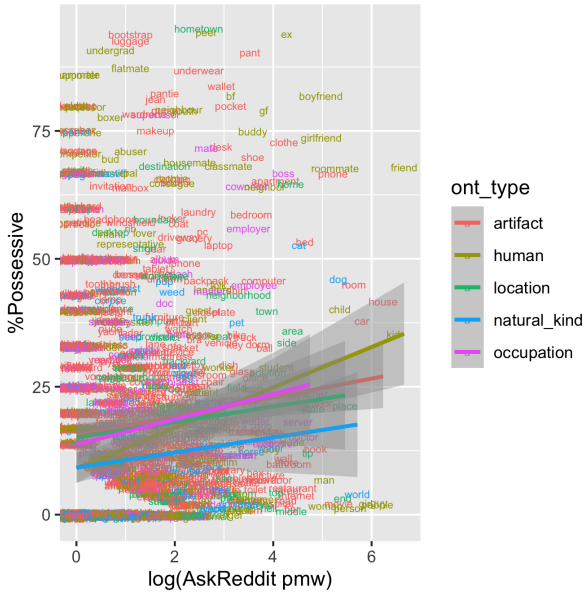

Figure 3: Percent Possessive as a function of log-transformed per-millon-word count in AskReddit, color-coded by ontological class.

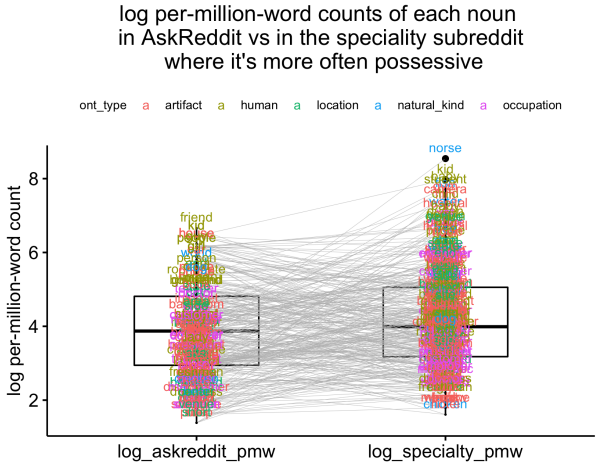

Figure 4: Paired visualization of the per-million-word count of the same noun in AskReddit versus in the specialty subreddit in which it is significantly more often possessive.

tional interaction) there.

A Wilcoxon test for paired samples was used to compare the per-million-word count of each noun in AskReddit versus in the specialty subreddit in which it is more often possessive. This test ($V = 25222, p < 0.05$), visualized in Figure 3 and replicated to a $p < 0.01$ significance level in a paired t test, finds that nouns are more frequent in the subreddits in which they are significantly more often possessive, consistent with (12)[9].

5.2 *Percentage of definite tokens*

Turning to the predictions related to definiteness, (13) is tested across nouns.

(13) **Across nouns: More relational, more often definite** (*not empirically supported*)

There should be a positive correlation between a noun's definite-to-indefinite ratio (proxy for conventional interaction) and its percentage of possessive tokens (proxy for relationality).

A series of linear regression models were run in R (R Core Team 2012) pre-

[9]The code used to identify these 341 nouns requires twenty total instances for the noun in 5 million words sampled from AskReddit and four instances in the 1 million words sampled from each specialty subreddit, so the method for identifying these nouns is designed not to bias the frequency comparison (otherwise, the minimum counts required to find a significant difference in the Fisher Test across skewed samples could potentially bake in a difference in frequency). Therefore, the observed difference in frequency is a true, non-spurious finding.

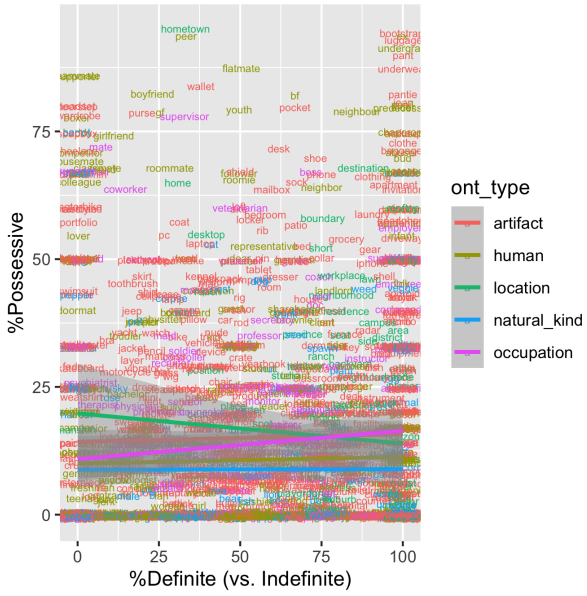

Figure 5: Percent Possessive as a function of the percentage of definite versus indefinite tokens in AskReddit, color-coded by ontological class.

dicting Percent Possessive as a function of a noun's ontological class and its per-million-word count (shown above to be a significant predictors) as well as its percentage of definite versus indefinite non-possessive tokens. Models were run using subsets, additive combinations, and interactions of these independent variables. Across such models, contrary to (13), Percent Definite does not significantly predict Percent Possessive.

Next, (14) is tested across communities.

(14) **Across communities: More relational, more often definite** (*empirically supported*)

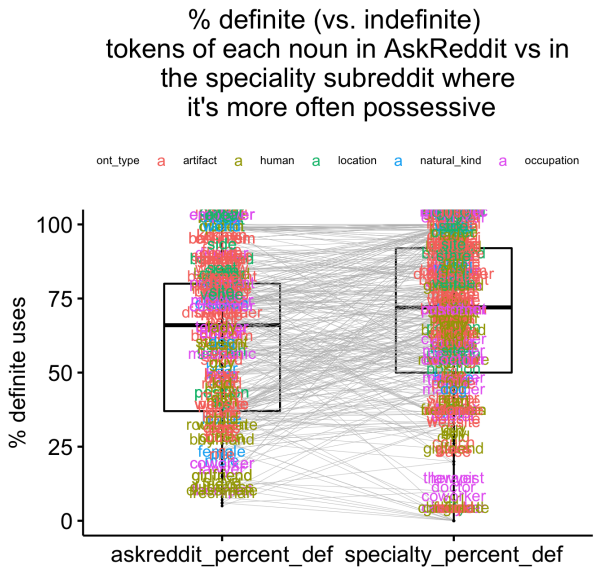

Figure 6: Paired visualization of the percentage of definite tokens of the same noun in AskReddit versus in the specialty subreddit in which it is significantly more often possessive.

> Nouns that are significantly more often possessive (proxy for more relational) in a given subreddit should be more often definite (proxy for conventional interaction) there.

A Wilcoxon Test for paired samples was used to compare the percentage of definite versus indefinite tokens of each noun in AskReddit versus in the specialty subreddit in which it is more often possessive. The Wilcoxon test ($V = 17622, p <$ 0.001, median percent definite = 66% in AskReddit versus 72% in the specialty subreddit), visualized in Figure 6 and replicated in a paired *t* test, finds that nouns are indeed more often definite in the specialty subreddits in which they are more often possessive, consistent with (14).

5.3 *Discussion*

This corpus study finds evidence consistent with three of the four predictions intended to quantify the overarching hypothesis that a noun should be more relational (as measured by Percent Possessive) when human interaction with its referent is more conventional (as measured by frequency and definiteness ratio). As predicted, there is a positive correlation across nouns between a noun's Percent Possessive and its frequency. Also as predicted, across communities, the same noun is more frequent and more often definite in the specialty subreddit in which it is more often possessive.

As for why Percent Possessive fails to correlate with definite-to-indefinite ratio across nouns, I suggest that the percentage of definite tokens of a noun is an imperfect proxy for conventional interaction with its referent. The chance of a noun being used as definite versus indefinite depends not just on the conventional familiarity of its referent but also on whether the referent can be considered unique in context (Löbner 2011; Coppock & Beaver 2015).

Among the nouns that are often used as definite but rarely as possessive, we find names for participants in a discourse-given event (*perpetrator, culprit, interviewer*, unique in context); and locations that seem to prefer the *of*-form to the *'s* possessive because their possessor is usually inanimate (*outskirts, outback, rooftop, forefront*). All these examples are 100% definite and 0% possessive in their AskReddit tokens. In these cases, I would suggest that the noun's high definite-to-indefinite ratio does not indicate the conventionality of human interaction with its referent.

In other words, there are many factors that determine a noun's definite-to-indefinite ratio beyond conventional interaction, so I argue that the overall hypothesis – that a noun is more relational when human interaction with its referent is more conventional – can still be true even if the predicted relation between definiteness ratio and Percent Possessive is not manifested. In a more controlled comparison of the same noun across communities, holding constant many of the other factors contributing to a noun's definite-to-indefinite ratio, the predicted effect is found.

5.4 *Examples*

I turn to some examples illustrating these findings. Across nouns, *phone* is far more frequent in AskReddit than *lamp* – both artifacts; occurring 184 versus 2 times per million (15)–(16) – which I take to approximate the intuition that human interaction with phones is far more conventional than with lamps. (*Phone* is also more often definite than *lamp*: 75% of *phone*'s non-possessive tokens are definite, versus 29% for *lamp*). As predicted, *phone* is also far more often possessive than *lamp* (67% versus 42% Percent Possessive).

(15) As a 16 yo, I shouldnt need restrictions on how long i'm using **my phone**, right? (r/AskReddit)

(16) We had no furniture, just a tv, maybe **a lamp** or so. (r/AskReddit)

Similarly, *dog* is far more frequent in AskReddit than *horse* – both natural kinds; occurring 205 versus 35 times per million; (17)–(18) – which I take to approximate the intuition that human interaction with dogs is far more conventional than with horses. (*Dog* is also more often definite than *horse*: 56% of *dog*'s non-possessive tokens are definite, versus 30% for *horse*). As predicted, *dog* is also far more often possessive than *horse* (46% versus 12% Percent Possessive).

(17) Currently watching Netflix with **my dog** on my lap. (r/AskReddit)

(18) I live in Texas and I've never ridden **a horse** here. (r/AskReddit)

Across communities, *knife* is far more often possessive in the Cooking subreddit than in AskReddit (44% versus 7%). As predicted, *knife* is far more frequent in the Cooking subreddit than in AskReddit – occurring 218 versus 11 times per million, exemplified in (19)–(20) — which I take to approximate the intuition that interaction with knives is far more conventional for cooks than for laypeople. *Knife* is also more often definite in the Cooking subreddit than in AskReddit (44% versus 32% of its non-possessive tokens are definite), which I take as further evidence for the same point.

(19) Flailing **the knife** on the stone or rubbing a stone on **the knife** are inefficient and unsafe for a beginner. (r/Cooking)

(20) Cut to 15 minutes later she's screaming in the kitchen holding **a knife**. I think **the knife** was just a coincidence, she's not a murderer. (r/AskReddit)

Finally, across communities, *horse* is far more often possessive in the Horses subreddit than in AskReddit (37% versus 12%). As predicted, *horse* is vastly more frequent in the Horses subreddit than in AskReddit – occurring 5129 versus 35

times per million (21)–(22) – which I take to approximate the intuition that human interaction with horses is far more conventional for equestrians than for laypeople. *Horse* is also more often definite in the Horses subreddit than in AskReddit (44% versus 30% of its non-possessive tokens are definite), which I take as further evidence for the same claim.

(21) **My horse** is still barefoot and never needed shoes before, during, and after having white line disease. (r/Horses)

(22) I live in Texas and I've never ridden **a horse** here. (r/AskReddit)

As illustrated by these examples, the Reddit corpus study finds evidence across nouns that Percent Possessive is positively correlated with a noun's per-million word frequency. Across communities, a noun is more frequent and more often definite in the community where its Percent Possessive is significantly higher. I take these findings to be consistent with the claim that a noun is more relational in a gradient sense (as measured by Percent Possessive) when human interaction with its referent is more conventional (as measured by its frequency and definite-to-indefinite ratio).

6. THEORETICAL CONSEQUENCES

Approximating a continuous construct of relationality via Percent Possessive, this paper has explored which nouns are more or less relational and why. Framed as binary rather than continuous, the same question pervades the literature on the semantics of possessive constructions: a researcher must decide whether to give two different analyses for *my cousin* versus *my tree* – in which case they must also decide which other nouns behave like *cousin* or like *tree* – or whether to propose a unified semantics for both *my cousin* and *my tree* while explaining their differences pragmatically. Any of these approaches can capture the facts, but each one also leaves open the question of which nouns should be analyzed in which way(s) and why, and so can be complemented by an answer to that question like the one offered here.

As mentioned above (§2), a researcher must also decide whether to begin from a binary distinction between two-place versus one-place predicates; or from the gradient data manifested in grammaticality judgments, across the lexicon, and in corpus data. If one begins with a formal binary, one might also aim to explain how it connects to the gradient data observed across the lexicon and in usage, or vice versa. Here, I review the literature's formal approaches to relationality and possession and explore – in some cases extrapolating beyond the authors' own claims – how each one could be linked to the gradient findings presented above.

6.1 *Two types of nouns, two analyses for possessives*

For Barker (1992), a relational noun such as *cousin* is a two-place predicate – $\lambda x \lambda y[cousin(x,y)]$ – and the possessor saturates one of its arguments. In contrast, a sortal noun such as *tree* is a one-place predicate – $\lambda x[tree(x)]$ – and the possessor is related to *tree* by a free variable, named R or π, which supplied by the possessive morphology and saturated by context. Inspired by the languages that use different morphology for inalienable versus alienable possessives, this approach reflects the intuition that possessed relational nouns (*my cousin*) provide a possessor-head relation lexically, whereas possessed sortal nouns (*my tree*) find their possessor-head relation in context (Ortmann 2018). This analysis assumes that nouns have to be somehow classified as relational or sortal, or perhaps (as Barker 1992 suggests for *child*) polysemous between the two. The distinction between one-place and two-place predicates is binary, although one could make a gradient prediction that possessive tokens of two-place predicates would be more frequent than possessive tokens of one-place predicates.

In this framework, Percent Possessive could serve as a continuous indicator of a noun's classification, and could help to quantify the relative frequency of different senses of nouns treated as polysemous (for example, 40% of AskReddit tokens of *child* are possessive compared to 84% of tokens of *daughter*, which might illustrate the frequency of *child*'s relational sense). Moreover, conventional interaction with

a noun's referent could help to explain these facts.

6.2 *Two types of nouns, one analysis for possessives*

For Vikner & Jensen (2002) and those inspired by them, all possessives are built from relational nouns, which are either inherently relational or type-shifted to a relational denotation. *Cousin* is relational and its possessor saturates one of its arguments. *Phone* is sortal, so in order to be possessed, it must be type-shifted to a relation that holds between an individual and the phone that they use for communication, drawing on encyclopedic information ('qualia structure' in a rich lexical representation inspired by Pustejovsky 1995) about the typical use of such an artifact. *Tree* is also a sortal noun, so it must also be type-shifted to a relation between an individual and the tree that they 'control' (in some way that must be interpreted contextually) – a last-resort type-shift (predicted to be infrequent and in need of contextual support) for nouns whose qualia structure does not provide any other relation. This approach strives for a unified analysis of possessives while maintaining a distinction between nouns that are inherently relational versus those that have to be type-shifted more or less easily to a relational meaning.

This analysis also assumes that nouns have to be classified as relational or sortal, and further that sortal nouns must be classified with respect to the relation-adding type-shifters that (tend to) combine with them. Here too, Percent Possessive could convey a noun's classification as well as its propensity for type-shifting, and con-

ventional interaction with its referent could help to explain these facts.

6.3 *One type of noun, one analysis for possessives*

For Payne et al. (2013) and Peters & Westerståhl (2013), all nouns are sortal and all possessives introduce a free variable *R* which relates the noun to its possessor and is supplied by some combination of lexical and contextual factors. On this view, *cousin* and *tree* are both just one-place predicates, but our lexical and/or encyclopedic knowledge easily supplies a salient possession relation for *my cousin*, whereas we have to look further at the context to find a suitable relation for *my tree*. This analysis embraces a continuous approach to relationality and a unified account of all possessives. It does not require nouns to be classified as relational versus sortal, but it still leaves open the question of which nouns provide more or less information to saturate the free variable *R*. Percent Possessive could quantify that cline, and the most likely relation between the possessor and the head noun could be supplied and explained by conventional interaction with its referent.

7. CONCLUSION

Facing the binary, theory-specific distinction between prototypical relational nouns such as *cousin* and prototypical sortal nouns such as *tree*, this paper offers Percent Possessive to reframe relationality as a continuous, objective corpus metric, and uses it to investigate at scale which nouns are more or less relational and why.

Across nouns, Percent Possessive is found to correlate with a noun's frequency; across communities, the same noun is found to be more frequent and more often definite in the community in which it is more often possessive. Expanding the suggestion that relationality and possession are grounded in culture, these findings are taken as evidence for the claim that, taking ontological class into account, a noun is more relational when human interaction with its referent is more conventional.

Stepping back, this paper illustrates the value as well as the challenge of approximating abstractions such as relationality and convention in corpus data. On the one hand, the proxy measurements of Percent Possessive, per-million-word frequency, and definite-to-indefinite ratio are of questionable validity as stand-ins for relationality and convention. On the other hand, these measurements allow for large-scale hypothesis testing, which is impossible if relationality and convention remain abstract.

This paper also constitutes an attempt to study lexical semantics at the scale of the lexicon. Any time words are classified with respect to some property that applies clearly to prototypical examples, there is a challenge to be found in explaining which further words fall into which class and why. Here, that explanation advances the larger idea that grammar is social: the syntactic distribution of a noun is linked to the conventions of the people who interact with its referent.

Author's address:
School of Modern Languages
Georgia Institute of Technology
613 Cherry St NW, Atlanta, GA 30313
United States
lelia.glass@modlangs.gatech.edu

REFERENCES

Aikhenvald, Alexandra Y. 2012. Possession and ownership: A cross-linguistic perspective (introduction). In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *Possession and ownership: A cross-linguistic typology*, 1–64. Oxford: Oxford University Press.

Asmuth, Jennifer A. & Dedre Gentner. 2005. Context sensitivity of relational nouns. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 27, 163–168.

Ball, Christopher. 2011. Inalienability in social relations: Language, possession, and exchange in Amazonia. *Language in Society* 307–341.

Barker, Chris. 1992. *Possessive descriptions*. Ph.D. thesis, University of California, Santa Cruz.

Barker, Chris. 2000. Definite possessives and discourse novelty. *Theoretical Linguistics* 26(3), 211–228.

Barker, Chris. 2011. Possessives and relational nouns. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, 1109–1130. Berlin: De Gruyter Mouton.

Barker, Chris. 2016. Why relational nominals make good concealed questions. *Lingua* 182, 12–29.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire & Jeremy Blackburn. 2020. The PushShift Reddit dataset. In *Proceedings of the International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Web and Social Media*, vol. 14, 830–839.

Bird, Alexander & Emma Tobin. 2009. Natural kinds. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, online (no page numbers). Palo Alto, CA: Stanford University.

Clark, Herbert H. 1975. Bridging. In Roger C. Schank & Bonnie Lynn Nash-Webber (eds.), *Theoretical issues in natural language processing*, 169–174. New York: Association for Computing Machinery.

Clark, Herbert H. & Catherine R. Marshall. 1981. Definite knowledge and mutual knowledge. In Aravind Joshi, Bonnie Webber & Ivan Sag (eds.), *Elements of discourse understanding*, 10–63. Cambridge: Cambridge University Press.

Coppock, Elizabeth & David Beaver. 2015. Definiteness and determinacy. *Linguistics and Philosophy* 38(5), 377–435.

De Bruin, Jos & Remko Scha. 1988. The interpretation of relational nouns. In *26th Annual Meeting of the Association for Computational Linguistics*, 25–32.

Del Tredici, Marco & Raquel Fernández. 2018. Semantic variation in online communities of practice. In Claire Gardent & Christian Retoré (eds.), *Proceedings*

of the 12th International Conference on Computational Semantics (IWCS 2017), 1–13.

Gentner, Dedre. 2005. The development of relational category knowledge. In Lisa Gershkoff-Stowe & David H. Rakison (eds.), *Building object categories in developmental time*, 245–275. Philadelphia, PA: Taylor & Francis Psychology Press.

Glass, Lelia. 2021. English verbs can omit their objects when they describe routines. *English Language and Linguistics* 26(1), 49–73.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1), 1–33.

Haspelmath, Martin. 2017. Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft* 36(2), 193–231.

Heim, Irene. 1979. Concealed questions. In *Semantics from different points of view*, 51–60. Springer.

Heim, Irene. 1982. *The semantics of definite and indefinite noun phrases*. Ph.D. thesis, University of Massachusetts, Amherst.

Heine, Bernd. 1997. *Possession: Cognitive sources, forces, and grammaticalization*. No. 83 in Cambridge Studies in Linguistics. Cambridge, U.K.: Cambridge University Press.

Hellwig, Oliver & Wiebke Petersen. 2015. Correlation between lexical and determination types. In *Proceedings of the German Society for Computational Linguistics*, 130–137.

Herring, Susan C., Dieter Stein & Tuija Virtanen. 2013. Introduction to the pragmatics of computer-mediated communication. In Susan C. Herring, Dieter Stein & Tuija Virtanen (eds.), *Pragmatics of computer-mediated communication*, 3–32. Berlin: Mouton De Gruyter.

von Heusinger, Klaus. 2013. The salience theory of definiteness. In Allessandro Capone, Francesco Lo Piparo & Marco Carapezza (eds.), *Perspectives on linguistic pragmatics*, 349–374. Cham: Springer.

Honnibal, Matthew & Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378.

Horn, Christian & Nicolas Kimm. 2014. Nominal concept types in german fictional texts. In Thomas Gamershlag, Doris Gerland, Rainer Osswald & Wiebke Petersen (eds.), *Frames and concept types*, 343–362. Cham: Springer.

Jensen, Per Anker & Carl Vikner. 2003. Producer interpretations of the English pre-nominal genitive. In Matthias Weisgerber (ed.), *Proceedings of Sinn und Bedeutung 7*, 173–183. Konstanz: Universität Konstanz.

Jensen, Per Anker & Carl Vikner. 2004. The English pre-nominal genitive and lexical semantics. In *Possessives and beyond: Semantics and syntax*, 3–27. Amherst, MA: Graduate Linguistic Student Association Publications.

Kalpak, Hana. 2020. The semantics and pragmatics of nouns in concealed questions. *Semantics and Pragmatics* 13(7), 1–25.

Karvovskaya, Lena. 2018. *The typology and formal semantics of adnominal possession*. Ph.D. thesis, University of Leiden, Leiden.

Keil, Frank C. 1989. *Concepts, Kinds, and Cognitive Development*. Cambridge, Massachusetts: MIT Press.

Kolkmann, Julia. 2016. *The pragmatics of possession: Issues in the interpretation of pre-nominal possessives in English*. Ph.D. thesis, University of Manchester, Manchester, UK.

Levin, Beth, Lelia Glass & Dan Jurafsky. 2019. Systematicity in the semantics of noun compounds: The role of artifacts vs. natural kinds. *Linguistics* 57(3), 429–471.

Lewis, David. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 3(8), 339–359.

Löbner, Sebastian. 1985. Definites. *Journal of Semantics* 4, 279–326.

Löbner, Sebastian. 2011. Concept types and determination. *Journal of Semantics* 28(3), 279–333.

de Marneffe, Marie-Catherine & Christopher Potts. 2017. Developing linguistic theories using annotated corpora. In Nancy Ide & James Pustejovsky (eds.), *The handbook of linguistic annotation*, 411–438. Dordrecht: Springer.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young & Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL*, 24–31.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4), 235–244.

Nathan, Lance Edward. 2006. *On the interpretation of concealed questions*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Newell, Edward & Jackie Chi Kit Cheung. 2018. Constructing a lexicon of relational nouns. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 3405–3410.

Nichols, Johanna. 1988. On alienable and inalienable possession. In William Shipley (ed.), *In honor of Mary Haas: From the Haas Festival conference on Native American linguistics*, 557–610. Berlin: De Gruyter Mouton.

Nissim, Malvina. 2004. Lexical information and choice of determiners. In *Possessives and beyond: Semantics and syntax*, 133–152. Amherst, MA: Graduate Linguistic Student Association Publications.

Ortmann, Albert. 2018. Connecting the typology and semantics of nominal possession: alienability splits and the morphology–semantics interface. *Morphology* 28(1), 99–144.

Partee, Barbara. 1995. Lexical semantics and compositionality. In Lila Gleitman Daniel Osersohn & Mark Liberman (eds.), *Invitation to cognitive science*, 311–360. Cambridge, MA: MIT Press.

Partee, Barbara H. & Vladimir Borschev. 1998. Integrating lexical and formal semantics: Genitives, relational nouns, and type-shifting. In Robin Cooper & Tamaz Gamkrelidze (eds.), *Proceedings of the Second Tbilisi Symposium on Language, Logic, and Computation*, 229–241.

Partee, Barbara H. & Vladimir Borschev. 2012. Sortal, relational, and functional interpretations of nouns and Russian container constructions. *Journal of Semantics* 29(4), 445–486.

Payne, John, Geoffrey K. Pullum, Barbara C. Scholz & Eva Berlage. 2013. Anaphoric 'one' and its implications. *Language* 794–829.

Peters, Stanley & Dag Westerståhl. 2013. The semantics of possessives. *Language* 713–759.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge: MIT Press.

R Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, Craige. 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy* (26), 287–350.

Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18(2), 215–262.

Seiler, Hansjakob. 2001. The operational basis of possession: A dimensional approach revisited. In Irène Baron, Michael Herslund & Finn Sørensen (eds.), *Typological Studies in Language*, vol. 47, 27–40. Amsterdam: John Benjamins.

Spenader, Jennifer. 2001. *Presuppositions in spoken discourse*. Ph.D. thesis, Stockholm University, Stockholm.

Stassen, Leon. 2009. *Predicative possession*. Oxford: Oxford University Press.

Strawson, Peter Frederick. 1950. On referring. *Mind* (59), 320–334.

Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English. In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: John Benjamins.

Vikner, Carl & Per Anker Jensen. 2002. A semantic analysis of the English genitive: Interaction of lexical and formal semantics. *Studia Linguistica* 56(2), 191–226.

Williams, Adina. 2018. *Representing relationality: MEG studies on argument structure*. Ph.D. thesis, New York University, New York.

Zhang, Justine, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky & Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In Winter Mason, Alice Marwick & Sandra González-Bailón (eds.), *Proceedings of the International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Weblogs and Social Media*, 377–386. Palo Alto, CA: The AAAI Press.