Predicting Declension Class from Form and Meaning

Adina Williams[ə] **Tiago Pimentel**[ð] **Arya D. McCarthy**[3]
Hagen Blix[ʜ] **Eleanor Chodroff**[ʏ] **Ryan Cotterell**[ð,ʕ]
[ə]Facebook AI Research [ð]University of Cambridge [3]Johns Hopkins University
[ʜ]New York University [ʏ]University of York [ʕ]ETH Zürich
adinawilliams@fb.com tp472@cam.ac.uk arya@jhu.edu hagen.blix@nyu.edu
eleanor.chodroff@york.ac.uk ryan.cotterell@inf.ethz.ch

Abstract

The noun lexica of many natural languages are divided into several declension classes with characteristic morphological properties. Class membership is far from deterministic, but the phonological form of a noun and/or its meaning can often provide imperfect clues. Here, we investigate the strength of those clues. More specifically, we operationalize this by measuring how much information, in bits, we can glean about declension class from knowing the form and/or meaning of nouns. We know that form and meaning are often also indicative of grammatical gender—which, as we quantitatively verify, can itself share information with declension class—so we also control for gender. We find for two Indo-European languages (Czech and German) that form and meaning respectively share significant amounts of information with class (and contribute additional information above and beyond gender). The three-way interaction between class, form, and meaning (given gender) is also significant. Our study is important for two reasons: First, we introduce a new method that provides additional quantitative support for a classic linguistic finding that form and meaning are relevant for the classification of nouns into declensions. Secondly, we show not only that individual declensions classes vary in the strength of their clues within a language, but also that these variations themselves vary *across languages*. The code is publicly available at https://github.com/rycolab/declension-mi.

1 Introduction

To an English speaker learning German, it may come as a surprise that one cannot necessarily predict the plural form of a noun from its singular. This is because pluralizing nouns in English is relatively simple: Usually we merely add an *-s* to the end (e.g., *cat* ↦ *cat**s***). Of course, not all English nouns follow such a simple rule (e.g., *child* ↦ *child**ren***, *sheep* ↦ *sheep*, etc.), but those that do not are fairly few. German, on the other hand, has comparatively many nouns following comparatively many common, morphological rules. For example, some plurals are formed by adding a suffix to the singular: *Insekt* 'insect' ↦ *Insekt-**en***, *Hund* 'dog' ↦ *Hund-**e***, *Radio* 'radio' ↦ *Radio-**s***. For others, the plural is formed by changing a stem vowel:[1] *Mutter* 'mother' ↦ ***Mü**tter*, or *Nagel* 'nail' ↦ ***Nä**gel*. Some others form plurals with both suffixation and vowel change: *Haus* 'house' ↦ ***Hä**us-er* and *Koch* 'chef' ↦ ***Kö**ch-e*. Still others, like *Esel* 'donkey', have the same form in plural and singular. How baffling for the adult learner! And, the problem only worsens when we consider other inflectional morphology, such as case.

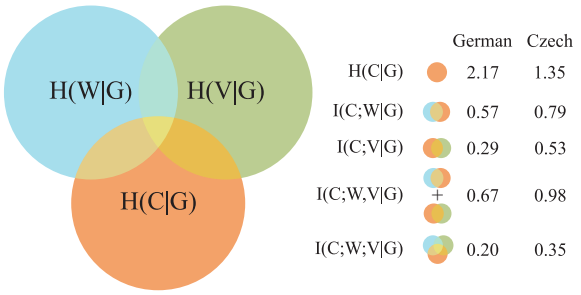


Figure 1: Declension classes, their conditional entropies (H), and their mutual information quantities (I) with form (W), meaning (V), and declension class (C), given gender (G) in German and Czech. $\mathrm{H}(W \mid G)$ and $\mathrm{H}(V \mid G)$ correspond to the overall uncertainty over forms and meaning given gender—estimating these values falls outside the scope of this paper.

Disparate plural-formation and case rules of the kind described above split nouns into **declension classes**. To know a noun's declension class is to know which morphological form it takes in which context (e.g., Benveniste 1935; Wurzel 1989; Nübling 2008; Ackerman et al. 2009; Ackerman and Malouf 2013; Beniamine and Bonami 2016; Bonami and Beniamine 2016). But, this begs the

[1]This vowel change, **umlaut**, corresponds to fronting.

question: What clues can we use to predict the class for a noun? In some languages, predicting declension class is argued to be easier if we know the noun's phonological form (Aronoff, 1992; Dressler and Thornton, 1996) or lexical semantics (Carstairs-McCarthy, 1994; Corbett and Fraser, 2000). However, semantic or phonological clues are, at best, only very imperfect hints as to class (Wurzel, 1989; Harris, 1991, 1992; Aronoff, 1992; Halle and Marantz, 1994; Corbett and Fraser, 2000; Aronoff, 2007). Given this, we quantify *how much* a noun's form and/or meaning shares with its class, and determine whether that amount of information is uniform across classes.

To do this, we measure the **mutual information** both between declension class and meaning (i.e., distributional semantic vector) and between declension class and form (i.e., orthographic form), as in Figure 1. We select two Indo-European languages (Czech and German) that have declension classes. We find that form and meaning both share significant amounts of information, in bits, with declension class in both languages. We further find that form clues are stronger than meaning clues; for form, we uncover a relatively large effect of 0.5–0.8 bits, while, for lexical semantics, a moderate one of 0.3–0.5 bits. We also measure the three-way interaction between form, meaning, and class, finding that phonology and semantics contribute overlapping information about class. Finally, we analyze individual inflection classes and uncover that the amount of information they share with form and meaning is not uniform across classes or languages.

We expect our results to have consequences, not only for NLP tasks that rely on morphological information—such as bilingual lexicon induction, morphological reinflection, and machine translation—but also for debates within linguistics on the nature of inflectional morphology.

2 Declension Classes in Language

The morphological behavior of declension classes is quite complex. Although various factors are doubtless relevant, we focus on phonological and lexical semantic ones here. We have ample reason to suspect that phonological factors might affect class predictability. In the most basic sense, the form of inflectional suffixes are often altered based on the identity of the final segment of the stem. For example, the English plural suffix is spelled as *-s* after most consonants, like in 'cats', but it gets spelled as *-es* if it appears after an *s, sh, z, ch* etc., like in 'mosses', 'rushes', 'quizzes', 'beaches' etc. Often differences in the spelling of plural affixes or declension class affixes are due to phonological rules that get noisily realized in orthography, but there might also be additional regularities that do not correspond to phonological rules but still have an impact. For example, statistical regularities over phonological segments in continuous speech guide first language acquisition (Maye et al., 2002), even over non-adjacent segments (Newport and Aslin, 2004). Probabilistic relationships have also been uncovered between the sounds in a word and the word's syntactic category (Farmer et al., 2006; Monaghan et al., 2007; Sharpe and Marantz, 2017) and between the orthographic form of a word and its argument structure valence (Williams, 2018). Thus, we expect the form of a noun to provide clues to declension class.

Semantic factors too are often relevant for determining certain types of morphologically relevant classes, such as grammatical gender, which is known to be related to declension class. It has been claimed that there are only two types of gender systems: *semantic systems* (where only semantic information is required) and *formal systems* (where semantic information as well as morphological and phonological factors are relevant) (Corbett and Fraser, 2000, 294). Moreover, a large typological survey, Qian et al. (2016) finds that meaning-sensitive grammatical properties, such as gender and animacy, can be decoded well from distributional word representations for some languages, but less well for others. These examples suggest that it is worth investigating whether noun semantics provides clues about declension class.

Lastly, form and meaning might interact, as in the case of **phonaesthemes** where the sounds of words provide non-arbitrary clues about their meanings (Sapir, 1929; Wertheimer, 1958; Holland and Wertheimer, 1964; Maurer et al., 2006; Monaghan et al., 2014; D'Onofrio, 2014; Dingemanse et al., 2015; Dingemanse, 2018; Pimentel et al., 2019). Therefore, we check whether form and meaning jointly share information with declension class.

2.1 Orthography as a proxy for phonology?

We motivate an investigation into the relationship between the form of a word and its declension class by appealing at least partly to phonological motivations. However, we make the simplifying assump-

tion that phonological information is adequately captured by orthographic word forms—i.e., strings of written symbols or **graphemes**. In general, one should question this assumption (Vachek, 1945; Luelsdorff, 1987; Sproat, 2000, 2012; Neef et al., 2012). For the particular languages we investigate here, it is less problematic, as Czech and German are known to be languages with fairly "transparent" mappings between spelling and pronunciation (Matějček, 1998; Miles, 2000; Caravolas and Volín, 2001), achieving higher performance on grapheme-to-phoneme conversion than do English and other languages that have more "opaque" orthographic systems (Schlippe et al., 2012). These studies suggest that we are justified in taking orthography as a proxy for phonological form. Nonetheless, to mitigate against any phonological information being inaccurately represented in the orthographic form (e.g., vowel lengthening in German), several of our authors, who are fluent reader-annotators of our languages, checked our classes for any unexpected phonological variations. (Examples are in §3.)

2.2 Distributional Lexical Semantics

We adopt a distributional approach to lexical semantics (Harris, 1954) that relies on pretrained word embeddings for this paper. We do this for multiple reasons: First, distributional semantic approaches to create word vectors, such as WORD2VEC (Mikolov et al., 2013), have been shown to do well at extracting lexical features such as animacy and taxonomic information (Rubinstein et al., 2015) and can also recognize semantic anomaly (Vecchi et al., 2011). Second, the distributional approach to lexical meaning can be easily operationalized into a straightforward procedure for extracting "meaning" from text corpora at scale. Finally, having a continuous representation of meaning, like word vectors, enables training of machine learning classifiers.

2.3 Controlling for grammatical gender?

Grammatical gender has been found to interact with lexical semantics (Schwichtenberg and Schiller, 2004; Williams et al., 2019, 2020), and often can be determined from form (Brooks et al., 1993; Dobrin, 1998; Frigo and McDonald, 1998; Starreveld and La Heij, 2004). This means that it cannot be ignored in the present study. While the precise nature of the relationship between declension class and gender is far from clear, it is well established that the two should be distinguished (Aronoff 1992; Wiese 2000; Kürschner and Nübling 2011, *inter alia*). We first measure the amount of information shared between gender and class, according to the methods described in §4, to verify that the predicted relationship exists. We then verify that gender and class overlap in information in German and Czech to a high degree, but that we cannot reduce one to the other (see Table 3 and §6). We proceed to control for gender, and subsequently measure how much *additional* information form or meaning provides about class.

3 Data

For our study, we need orthographic forms of nouns, their associated word vectors, and their declension classes. Orthographic forms are the easiest component, as they can be found in any large text corpus or dictionary. We isolated noun **lexemes** (i.e., or syntactic category–specific representations of words) by language. We select Czech nouns from Unimorph (Kirov et al., 2018) and German nouns from Baayen et al. (1995, CELEX2). For lexical semantics, we trained 300D WORD2VEC vectors on language-specific Wikipedia.[2]

We select the nominative singular form as the donor for both orthographic and lexical semantic representations, because it is the canonical **lemma**, in these languages and also usually the **stem** for the rest of the morphological paradigm. We restrict our investigation to monomorphemic lexemes because: (i) one stem can take several affixes which would multiply its contribution to the results, and (ii) certain affixes come with their own class.[3]

Compared to form and meaning, declension class is a bit harder to come by, because it requires linguistic annotation. We associated lexemes with their classes on a by-language basis by relying on annotations from fluent speaker linguists, either for class determination (for Czech) or for verifying existing dictionary information (for German). For Czech, declension classes were derived by edit distance heuristic over affix forms, which grouped lemmata into subclasses if they received the same inflectional affixes (i.e., they constituted a morphological paradigm). If orthographic differences between two sets of suffixes in the lemma form could be accounted for by positing a phonological rule, then the two sets were collapsed into a single set; for example, in

[2]We use the GENSIM toolkit(Řehůřek and Sojka, 2010).

[3]Since these require special treatment, they are set aside.

	Original	Final	Training	Validation	Test	Average Length	# Classes
Czech	3011	2672	2138	267	267	6.26	13
German	4216	3684	2948	368	368	5.87	16

Table 1: Number of words in dataset. Counts per language-category pair are listed both before and after preprocessing, train-validation-test split, average stem length, and # of classes. Since we use 10-fold cross-validation, all instances are included in the test set at some point, and are used to estimate the cross-entropies in §5.

the "feminine *-a*" declension class, we collapsed forms for which the dative singular suffix surfaces as *-e* following a coronal continuant consonant (*figurka*:*figurce* 'figurine.DAT.SG'), *-i* following a palatal nasal (*piraňa*:*pirani* 'piranha.DAT.SG'), and as *-ě* following all other consonants (*kráva*:*krávě* 'cow.DAT.SG'). As for meaning, descriptively, gender is roughly a superset of declension classes in Czech; among the masculine classes, animacy is a critical semantic feature, whereas form seems to matter more for feminine and neuter classes. Our final tally of Czech noun contains a total of 2672 nouns in 13 declension classes.

For German, nouns came morphologically parsed and lemmatized, as well as coded for class (Baayen et al., 1995, CELEX2, v.2.5). We use CELEX2 to isolate monomorphemic noun lexemes and bin them into classes. CELEX2 declesion classes are more fine-grained than traditional descriptions of declension class; mappings between CELEX2 classes and traditional linguistic descriptions of declension class (Alexiadou and Müller, 2008) are provided in Table 4 in the Appendix. CELEX2 declension class encoding is compound and includes: (i) the number prefix (the first slot 'S' is for singular, and the second 'P' for plural), (ii) the morphological form identifier—zero refers to non-existent forms (e.g., plural is zero for *singularia tantum* nouns), and other numbers refer to a form identifier of morphological paradigm (e.g., genitive applies an additional suffix for singular masculine nouns, but never for feminines)—and (iii) an optional 'u' identifier, which refers to vowel umlaut, if present. More details of the German preprocessing steps are in the Appendix. In the final tally, we consider a total of 16 declension classes, which can be broken into 3 types of singular and 7 types of plural, summing to a total of 3684 nouns.

After associating nouns with forms, meanings, and classes, we perform exclusions: Because frequency affects class entropy (Parker and Sims, 2015), we removed all classes with fewer than 20 lexemes.[4] We subsequently removed all lexemes which did not appear in our WORD2VEC models trained on Wikipedia dumps. The remaining lexemes were split into 10 folds for cross-validation: One for testing, another for validation, and the remaining 8 for training. Table 1 shows train-validation-test splits, average length of nouns, and number of declension classes, by language. Table 5 in the Appendix provides final noun lexeme counts by declension class.

4 Methods

Notation. We define each lexeme in a language as a triple. Specifically, the i^{th} triple consists of an orthographic word form $\mathbf{w}_i$, a distributional semantic vector $\mathbf{v}_i$ that encodes the lexeme's semantics, and a declension class c_i. These triples follow a (unknown) probability distribution $p(\mathbf{w}, \mathbf{v}, c)$—which can be marginalized to obtain marginal distributions, e.g. $p(c)$. We take the space of word forms to be the Kleene closure over a language's alphabet Σ; thus, we have $\mathbf{w}_i \in \Sigma^*$. Our distributional semantic space is a high-dimensional real vector space $\mathbb{R}^d$ where $\mathbf{v}_i \in \mathbb{R}^d$. The space of declension classes is language-specific and contains as many elements as the language has classes, i.e., $\mathcal{C} = \{1, \ldots, K\}$ where $c_i \in \mathcal{C}$. For each noun, a gender g_i from a language-specific space of genders $\mathcal{G}$ is associated with the lexeme. In both Czech and German, $\mathcal{G}$ contains three genders: feminine, masculine, and neuter. We also consider four random variables: an $\mathbb{R}^d$-valued random variable V, a Σ^*-valued random variable W, a $\mathcal{C}$-valued random variable C and a $\mathcal{G}$-valued random variable G.

Bipartite Mutual Information. Bipartite MI (or, simply MI) is a symmetric quantity that measures how much information (in bits) two random variables share. In the case of C (declension class) and W (orthographic form), we have

$$\mathrm{I}(C; W) = \mathrm{H}(C) - \mathrm{H}(C \mid W) \qquad (1)$$

[4] We ran another version of our models that included all the original classes and observed no notable differences.

As can be seen, MI is the difference between an unconditional and a conditional entropy. The unconditional entropy is defined as

$$\mathrm{H}(C) = -\sum_{c \in \mathcal{C}} p(c) \log p(c) \tag{2}$$

and the conditional entropy is defined as

$$\mathrm{H}(C \mid W) = -\sum_{c \in \mathcal{C}} \sum_{\mathbf{w} \in \Sigma^*} p(c, \mathbf{w}) \log p(c \mid \mathbf{w}) \tag{3}$$

A good estimate of $\mathrm{I}(C; W)$ will naturally encode how much the orthographic word form tells us about its corresponding lexeme's declension class. Likewise, to measure the interaction between declension class and lexical semantics, we also consider the bipartite mutual information $\mathrm{I}(C; V)$.

Tripartite Mutual Information. To consider the interaction between three random variables at once, we need to generalize MI to three classes. One can calculate tripartite MI is as follows:

$$\mathrm{I}(C; W; V) = \mathrm{I}(C; W) - \mathrm{I}(C; W \mid V) \tag{4}$$

As can be seen, tripartite MI is the difference between a bipartite MI and a conditional bipartite MI. The conditional bipartite MI is defined as

$$\mathrm{I}(C; W \mid V) = \mathrm{H}(C \mid V) - \mathrm{H}(C \mid W, V) \tag{5}$$

In plainspeak, Equation 4 is the difference between how much C and W interact and how much they interact after "controlling" for V. [5]

Controlling for Gender. Working with mutual information also gives us a natural way to control for quantities that we know influence meaning and form. We do this by considering conditional MI. We consider both bipartite and tripartite conditional mutual information. These are defined as follows:

$$\mathrm{I}(C; W \mid G) = \mathrm{H}(C \mid G) - \mathrm{H}(C \mid W, G) \tag{6a}$$

$$\mathrm{I}(C; W; V \mid G) = \mathrm{I}(C; W \mid G) - \mathrm{I}(C; W \mid V, G) \tag{6b}$$

[5] We emphasize here the subtle, but important, distinction between $\mathrm{I}(C; W; V)$ and $\mathrm{I}(C; W, V)$. (The difference in notation lies in the comma replacing the semicolon.) While the first (tripartite MI) measures the ammount of (redundant) information shared by the three variables, the second (bipartite) measures the (total) information that class shares with either the form *or* the lexical semantics.

Estimating these quantities tells us how much C and W (and, in the case of tripartite MI, V also) interact after we take G (the grammatical gender) out of the picture. Figure 1 provides a graphical summary for this section until this point.

Normalization. To further contextualize our results, we consider two normalization schemes for MI. Normalizing renders MI estimates across languages more directly comparable (Gates et al., 2019). We consider the **normalized mutual information**, i.e., which *fraction* of the unconditional entropy is the mutual information:

$$\mathrm{NMI}(C; W) = \frac{\mathrm{I}(C; W)}{\min\{\mathrm{H}(C), \mathrm{H}(W)\}} \tag{7}$$

In practice, $\mathrm{H}(C) \ll \mathrm{H}(W)$ in most cases and normalized mutual information is more appropriately termed the **uncertainty coefficient** (Theil, 1970):

$$\mathrm{U}(C \mid W) = \frac{\mathrm{I}(C; W)}{\mathrm{H}(C)} \tag{8}$$

This can be computed from any mutual information equation, and will yield a percentage of the entropy that the mutual information accounts for—a more interpretable notion of the predictability between class and form or meaning.

5 Computation and Approximation

In order to estimate the mutual information quantities of interest per §4, we need to estimate a variety of entropies. We derive our mutual information estimates from a corpus $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{w}_i, c_i)\}_{i=1}^{N}$.

5.1 Plug-in Estmation of Entropy

The most straight-forward quantity to estimate is $\mathrm{H}(C)$. Given a corpus, we may use plug-in estimation: We compute the empirical distribution over declension classes from $\mathcal{D}$. Then, we plug that empirical distribution over declension classes $\mathcal{C}$ into the formula for entropy in Equation 2. This estimator is biased (Paninski, 2003), but is a suitable choice given because we have only a few declension classes and a large amount of data. Future work will explore whether better estimators (Miller, 1955; Hutter, 2001; Archer et al., 2013, 2014) affect the conclusions of studies such as this one.

5.2 Model-based Estimation of Entropy

In contrast, estimating $\mathrm{H}(C \mid W)$ is non-trivial. We cannot simply apply plug-in estimation because

we cannot compute the infinite sum over Σ^* that is required. Instead, we follow previous work (Brown et al., 1992; Pimentel et al., 2019) in using the cross-entropy upper bound to approximate $\mathrm{H}(C \mid W)$ with a model. More formally, for any probability distribution $q(c \mid \mathbf{w})$, we estimate

$$\begin{aligned} \mathrm{H}(C \mid W) &\leq \mathrm{H}_q(C \mid W) \\ &= -\sum_{c \in \mathcal{C}} \sum_{\mathbf{w} \in \Sigma^*} p(c, \mathbf{w}) \log q(c \mid \mathbf{w}) \end{aligned} \tag{9}$$

To circumvent the need for infinite sums, we use a held-out sample $\tilde{\mathcal{D}} = \{(\tilde{\mathbf{v}}_i, \tilde{\mathbf{w}}_i, \tilde{c}_i)\}_{i=1}^{M}$ disjoint from $\mathcal{D}$ to approximate the true cross-entropy $\mathrm{H}_q(C \mid W)$ with the following quantity

$$\hat{\mathrm{H}}_q(C \mid W) = -\frac{1}{M} \sum_{i=1}^{M} \log q\left(\tilde{c}_i \mid \tilde{\mathbf{w}}_i\right) \tag{10}$$

where we assume the held-out data is distributed according to the true distribution p. We note that $\hat{\mathrm{H}}_q(C \mid W) \rightarrow \mathrm{H}_q(C \mid W)$ as $M \rightarrow \infty$. While the exposition above focuses on learning a distribution $q(c \mid \mathbf{w})$ for classes and forms to approximate $\mathrm{H}(C \mid W)$, the same methodology can be used to estimate all necessary conditional entropies.

Form and gender: $q(c \mid \mathbf{w}, g)$. We train two LSTM classifiers (Hochreiter and Schmidhuber, 1996)—one for each language. The last hidden state of the LSTM models is fed into a linear layer and then a softmax non-linearity to obtain probability distributions over classes. To condition our model on gender classes, we embedd each gender and feed it into each LSTM's initial hidden state.

Meaning and gender: $q(c \mid \mathbf{v}, g)$. We trained a simple multilayer perceptron (MLP) classifier to predict the declension class, given the WORD2VEC representation. When conditioning on gender, we again embedded each class, concatenating these embeddings with the WORD2VEC ones before feeding the result into the MLP.

Form, meaning, and gender: $q(c \mid \mathbf{w}, \mathbf{v}, g)$. We again trained two LSTM classifiers, but this time, also conditioned on meaning (i.e., WORD2VEC). We avoided overfitting by reducing the WORD2VEC dimensionality from its original 300 dimensions to k with language-specific PCAs. We then linearly transformed them to match the hidden size of the LSTMs, and fed them in. To also condition on gender, we followed the same procedures, but used half of each LSTM's initial hidden state for each vector (i.e., WORD2VEC and gender one-hot embeddings).

Optimization. All classifiers were trained using Adam (Kingma and Ba, 2015) and code was implemented using PyTorch. Hyperparameters—number of training epochs, hidden sizes, PCA compression dimension (k), and number of layers—were optimized using Bayesian optimization with a Gaussian process prior (Snoek et al., 2012). For each experiment, fifty models were trained to maximize expected improvement on the validation set.

5.3 An Empirical Lower Bound on MI

With our empirical approximations of the desired entropy measures, we can calculate the desired approximated MI values, e.g.,

$$\mathrm{I}(C; W \mid G) \approx \hat{\mathrm{H}}(C \mid G) - \hat{\mathrm{H}}_q(C \mid W, G) \tag{11}$$

where $\hat{\mathrm{H}}(C \mid G)$ is the plug-in estimation of the entropy. Such an approximation, though, is not ideal, since we do not know if the true MI is approximated by above or below. Nonetheless, we use plug-in estimation, which underestimates entropy, and $\mathrm{H}_q(C \mid W, G)$ is estimated with a cross-entropy upperbound, we have

$$\begin{aligned} \mathrm{I}(C; W \mid G) &= \mathrm{H}(C \mid G) - \mathrm{H}(C \mid W, G) \\ &\gtrapprox \hat{\mathrm{H}}(C \mid G) - \mathrm{H}(C \mid W, G) \\ &\gtrapprox \hat{\mathrm{H}}(C \mid G) - \hat{\mathrm{H}}_q(C \mid W, G) \end{aligned} \tag{12}$$

We note that these lower bounds are *exact* when taking an expectation under the true distribution p. We cannot make a similar statement about tripartite MI, though, since it is computed as the difference of two mutual information quantities, both of which are lower-bounded in their approximations.

6 Results

Our main experimental results are presented in Table 2. We find that both form and lexical semantics significantly interact with declension class in both Czech and German. We observe that our estimates of $\mathrm{I}(C; W \mid G)$ is larger (0.5–0.8 bits) than our estimates of $\mathrm{I}(C; V \mid G)$ (0.3–0.5 bits). We also observe that the MI estimates in Czech are higher than in German. However, we caution that the estimates for the two languages are not fully comparable because they hail from models trained on different amounts of data. The tripartite MI estimates between class, form, and meaning, were relatively

	Form & Declension Class (LSTM)				Meaning & Declension Class (MLP)			
	$\mathrm{H}(C \mid G)$	$\mathrm{H}_Q(C \mid W, G)$	$\mathrm{I}(C; W \mid G)$	$U(C \mid W, G)$	$\mathrm{H}(C \mid G)$	$\mathrm{H}_Q(C \mid V, G)$	$\mathrm{I}(C; V \mid G)$	$U(C \mid V, G)$
Czech	1.35	0.56	**0.79**	58.8%	1.35	0.82	**0.53**	39.4%
German	2.17	1.60	**0.57**	26.4%	2.17	1.88	**0.29**	13.6%
	Both (Form and Meaning) & Declension Class				**Tripartite MI (LSTM)**			
	$\mathrm{H}(C \mid G)$	$\mathrm{H}_Q(C \mid W, V, G)$	$\mathrm{I}(C; W, V \mid G)$	$U(C \mid W, V, G)$	$\mathrm{I}(C; W \mid G)$	$\mathrm{I}(C; W \mid V, G)$	$\mathrm{I}(C; W; V \mid G)$	$U(C \mid W; V, G)$
Czech	1.35	0.37	**0.98**	72.6%	0.79	0.44	**0.35**	25.9%
German	2.17	1.50	**0.67**	30.8%	0.57	0.37	**0.20**	9.2%

Table 2: MI between form and class (top-left), meaning and class (top-right), both form and meaning and class (bottom-left), and tripartite MI (bottom-right). All values are calculated given gender, and bold if significant.

	$\mathrm{H}(C)$	$\mathrm{H}(C \mid G)$	$\mathrm{I}(C; G)$	$U(C \mid G)$
Czech	2.75	1.35	**1.40**	50.8%
German	2.88	2.17	**0.71**	24.6%

Table 3: MI between class and gender $\mathrm{I}(C; G)$: $\mathrm{H}(C)$ is class entropy, $\mathrm{H}(C \mid G)$ is class entropy given gender, $U(C; G)$ is the uncertainty coefficient.

small (0.2–0.35 bits) for both languages. We interpret this finding as showing that much of the information contributed by form is not redundant with information contributed by meaning—although a substantial ammount is. All results in this section were significant for both languages, according to a Welch (1947)'s t-test, which yielded $p < 0.01$ after Benjamini and Hochberg (1995) correction.[6]

As a final sanity check, we measure mutual information between class and gender $\mathrm{I}(C; G)$ (see Table 3). In both cases, the mutual information between class and gender is significant. MIs ranged from approximately $3/4$ of a bit in German to up to 1.4 bits in Czech, nearly 25% and nearly 51% of the remaining entropy of class, respectively. Like the quantities discussed in §4, this MI can also be estimated using simple plug-in estimation. Remember, if class were entirely reducible to gender, conditional entropy of class given gender would be zero. This is not the case: Although the conditional entropy of class given gender is lower for Czech (1.35 bits) than for German (2.17 bits), in neither case is declension class informationally equivalent to the language's grammatical gender system.

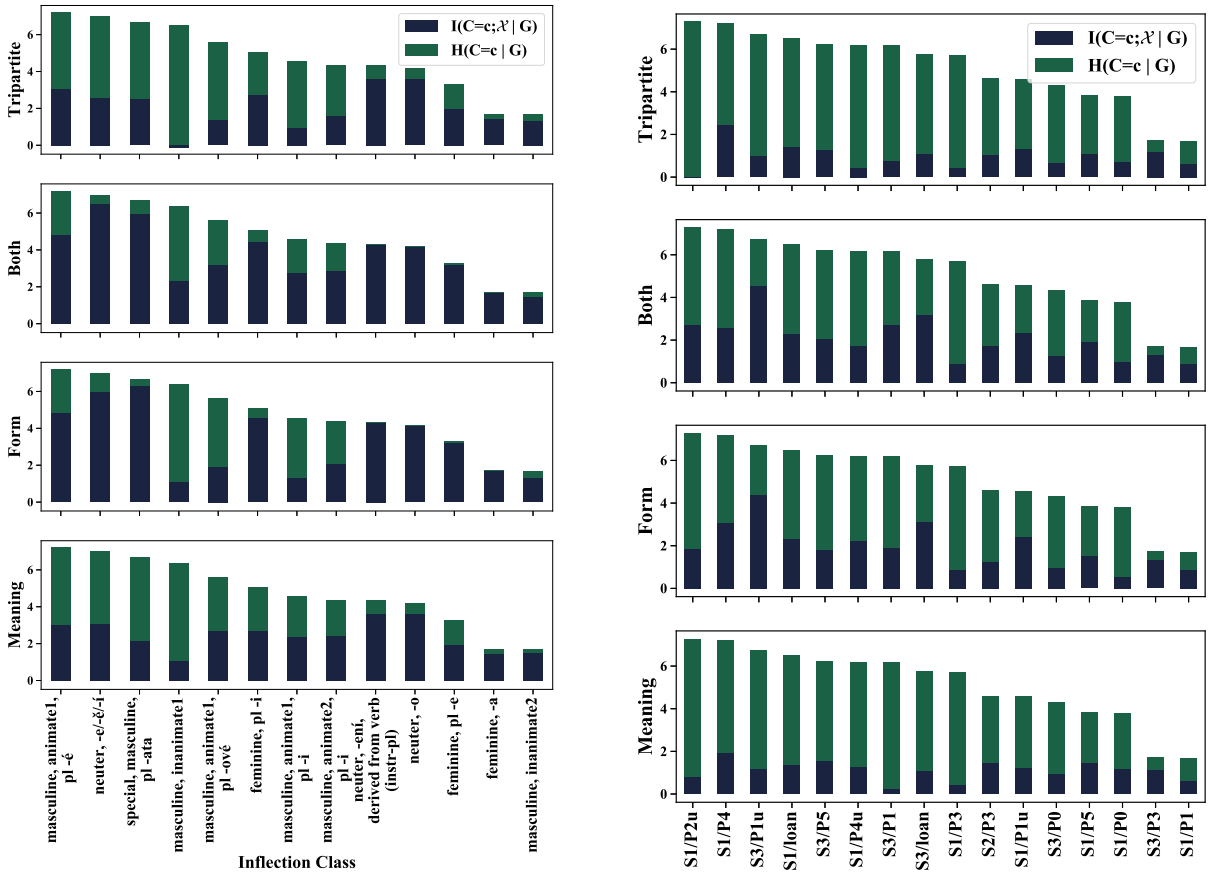
7 Discussion and Analysis

Next, we ask whether individual declension classes *differ* in how idiosyncratic they are, e.g., does any one German declension class share less information with form than the others? To address this, we qualitatively inspect per-class pointwise mutual information (PMI) in Figure 2a–2b. See Table 5 in the Appendix for the five highest and lowest surprisal examples per model. Several qualitative trends were observed: (i) classes show a decent amount of variability, (ii) unconditional entropy for each class is inversely proportional to the class' size, (iii) PMI is higher on average for Czech than German, and (iv) classes that have high $\mathrm{PMI}(C; V \mid G)$ usually have high $\mathrm{PMI}(C; W \mid G)$ (with notable exceptions we discuss below).

Czech. In general, masculine classes have smaller $\mathrm{PMI}(C = c; W \mid G)$ than feminine or neuter ones of comparable size—the exception being 'special, masculine, plural *-ata*'. This class ends exclusively in *-e* or *-ě*, which might contribute to that class' higher $\mathrm{PMI}(C = c; W \mid G)$. That $\mathrm{PMI}(C = c; W \mid G)$ is high for feminine and neuter classes suggests that the overall $\mathrm{I}(C; W \mid G)$ results might be largely driven by these classes, which predominantly end in vowels. We also note that the high $\mathrm{PMI}(C = c; W \mid G)$ for feminine 'plural *-e*', might be driven by the many Latin or Greek loan words present in this class.

With respect to meaning, recall that masculine declension classes reflect animacy status: 'animate1' contains nouns referring mostly to humans, as well as a few animals (*kocour* 'tomcat', *čolek* 'newt'), 'animate2' mostly animals with a few humans (*syn* 'son', *křest'an* 'Christian'), 'inanimate1' contains many plants, staple foods (*chléb* 'bread', *ocet* 'vinegar') and meaningful places (*domov* 'home', *kostel* 'church'), and 'inanimate2' contains many basic inanimate nouns (*kámen* 'stone'). Of these masculine classes, 'inanimate1' has a lower $\mathrm{PMI}(C = c; V \mid G)$ than its class size alone might lead us to predict. Feminine and neuter classes show no clear pattern, although neuter classes '-eni' and '-o' have comparatively

[6] A Welch (1947)'s t-test differs from Student (1908)'s t-test in that the latter assumes equal variances, and the former does not, making it preferable (see Delacre et al. 2017).



(a) Czech (b) German

Figure 2: Pointwise MI for declension classes. PMI for each random variable $\mathcal{X} \in \{W, V, \{W, V\}, \{W; V\}\}$ are plotted for classes increasing in size (towards the right): $I(C = c; V|G)$ (bottom), $I(C = c; W|G)$ (bottom middle), $I(C = c; V, W|G)$ (top middle), and tripartite $I(C = c; V; W|G)$ (top).

high $\mathrm{PMI}(C = c; V \mid G)$.

For $\mathrm{PMI}(C = c; V; W \mid G)$, we observe that 'masculine, inanimate1' is the smallest quantity, followed by most other masculine classes (e.g., masculine animate classes with *-ové* or *-i* plurals) for which $\mathrm{PMI}(C = c; W \mid G)$ was also low. Among non-masculine classes, we observe that feminine 'pl *-i*' and the neuter classes *-o* and *-ení* show higher tripartite PMI. The latter two classes have relatively high PMI across the board.

German. $\mathrm{PMI}(C = c; W \mid G)$ for classes containing words with umlautable vowels (i.e., S3/P1u, S1/P1u) or loan words (i.e., S3/loan) tends to be high; in the prior case, our models seem able to separate umlautable from non-umlautable vowels, and in the latter case, loan word orthography from native orthography. $\mathrm{PMI}(C = c; V \mid G)$ quantities are roughly equivalent across classes of different size, with the exception of three classes: S1/P4, S3/P1, and S1/P3. S1/P4 consists of highly semantically variable nouns, ranging from relational noun lexemes (e.g., *Glied* 'member', *Weib* 'wife', *Bild* 'picture') to masses (e.g., *Reis* 'rice'), which perhaps explains its relatively high $\mathrm{PMI}(C = c; V \mid G)$. For S1/P3 and S3/P1, $\mathrm{PMI}(C = c; V \mid G)$ is low, and we observe that both declension classes idiosyncratically group clusters of semantically similar nouns: S1/P3 contains "exotic" birds (*Papagei* 'parrot', *Pfau* 'peacock'), but also nouns ending in *-or*, (*Trakt**or*** 'tractor', *Past**or*** 'pastor'), whereas S3/P1 contains very few nouns, such as names of months (*März*, 'March', *Mai* 'May') and names of mythological beasts (e.g., *Sphinx*, *Alp*).

Tripartite PMI is fairly idiosyncratic in German: The lowest quantity comes from the smallest class, S1/P2u. S1/P3, a class with low $\mathrm{PMI}(C = c; V \mid G)$ from above, also has low tripartite PMI. We speculate that this class could be a sort of 'catch-all' class with no clear regularities. The highest tripartite PMI comes from S1/P4, which also had high $\mathrm{PMI}(C = c; V \mid G)$. The result suggests that submorphemic meaning bearing units, or phonaes-

themes might be present; taking inspiration from Pimentel et al. 2019, which aims to automatically discover such units, we observe that many words in S1/P4 contain letters {*d, e, g, i, l*}, often in identically ordered orthographic sequences, such as ***Bild***, ***Bie****st*, ***Feld***, ***Geld***, ***Glied***, ***Kind***, ***Lei****b*, ***Lied***, *Sch****ild***, ***Vie****ch*, ***Wei****b*, etc. While these letters are common in German orthography, their noticeable presence suggests further elucidation of declension classes in the context of phonaesthemes could be warranted.

8 Conclusion

We adduce new evidence that declension class membership is not wholly idiosyncratic nor fully deterministic based on form or meaning in Czech and German. We measure several mutual information quantities that range from 0.2 bits to nearly a bit. Despite their relatively small magnitudes, our measured mutual information between class and form accounted for between 25% and 60% of the class' entropy, even after relevant controls, and MI between class and meaning accounted for between 13% and nearly 40%. We analyze results per-class, and find that classes vary in how much information they share with meaning and form. We also observe that classes that have high $\mathrm{PMI}(C = c; V \mid G)$ often have high $\mathrm{PMI}(C = c; W \mid G)$, with a few noted exceptions that have specific orthographic (e.g., German umlauted plurals), or semantic (e.g., Czech masculine animacy) properties. In sum, this paper has proposed a new information-theoretic method for quantifying the strength of morphological relationships, and applied it to declension class. We verify and build on existing linguistic findings, by showing that the mutual information quantities between declension class, orthographic form, and lexical semantics are statistically significant.

Acknowledgments

Thanks as well to Guy Tabachnik for informative discussions on Czech phonology, to Jacob Eisenstein for useful questions about irregularity, and to Andrea Sims and Jeff Parker for advice on citation forms. Thanks to Ana Paula Seraphim for helping beautify Figure 1.

References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. *Analogy in Grammar: Form and Acquisition*, pages 54–82.

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, pages 429–464.

Artemis Alexiadou and Gereon Müller. 2008. Class Features as Probes. In Asaf Bachrach and Andrew Nevins, editors, *Inflectional Identity*, volume 18 of *Oxford Studies in Theoretical Linguistics*, pages 101–155. Oxford University Press, Oxford.

Evan Archer, Il Memming Park, and Jonathan W. Pillow. 2013. Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755.

Evan Archer, Il Memming Park, and Jonathan W. Pillow. 2014. Bayesian entropy estimation for countable discrete distributions. *The Journal of Machine Learning Research*, 15(1):2833–2868.

Mark Aronoff. 1992. Noun classes in Arapesh. In *Yearbook of Morphology 1991*, pages 21–32. Springer.

Mark Aronoff. 2007. In the beginning was the word. *Language*, 83(4):803–830.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX2 lexical database. *Linguistic Data Consortium*.

Sacha Beniamine and Olivier Bonami. 2016. A comprehensive view on inflectional classification. In *Annual Meeting of the Linguistic Association of Great Britain*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Émile Benveniste. 1935. *Origines de la formation des noms en indo-européen*, volume 1. Adrien-Maisonneuve Paris.

Olivier Bonami and Sarah Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Patricia J. Brooks, Martin D. S. Braine, Lisa Catalano, Ruth E. Brody, and Vicki Sudhalter. 1993. Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, 32(1):76–95.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Markéta Caravolas and Jan Volín. 2001. Phonological spelling errors among dyslexic children learning a transparent orthography: the case of Czech. *Dyslexia*, 7(4):229–245.

Andrew Carstairs-McCarthy. 1994. Inflection classes, gender, and the principle of contrast. *Language*, pages 737–788.

Greville G. Corbett and Norman M. Fraser. 2000. Gender assignment: a typology and a model. *Systems of Nominal Classification*, 4:293–325.

Marie Delacre, Daniel Lakens, and Christophe Leys. 2017. Why psychologists should by default use Welchs t-test instead of Students t-test. *International Review of Social Psychology*, 30(1).

Mark Dingemanse. 2018. Redrawing the margins of language: Lessons from research on ideophones. *Glossa*, 3(1).

Mark Dingemanse, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. 2015. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10):603 – 615.

Lise M. Dobrin. 1998. The morphosyntactic reality of phonological form. In *Yearbook of Morphology 1997*, pages 59–81. Springer.

Annette D'Onofrio. 2014. Phonetic detail and dimensionality in sound-shape correspondences: Refining the *bouba-kiki* paradigm. *Language and Speech*, 57(3):367–393.

Wolfgang U. Dressler and Anna M. Thornton. 1996. Italian nominal inflection. *Wiener Linguistische Gazette*, 55(57):1–26.

Thomas A. Farmer, Morten H. Christiansen, and Padraic Monaghan. 2006. Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103(32):12203–12208.

Lenore Frigo and Janet L. McDonald. 1998. Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, 39(2):218–245.

Alexander J. Gates, Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. 2019. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, 9(8574).

Morris Halle and Alec Marantz. 1994. Some key features of distributed morphology. *MIT Working Papers in Linguistics*, 21(275):88.

James W. Harris. 1991. The exponence of gender in Spanish. *Linguistic Inquiry*, 22(1):27–62.

James W. Harris. 1992. The form classes of Spanish substantives. In *Yearbook of Morphology 1991*, pages 65–88. Springer.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Sepp Hochreiter and Jürgen Schmidhuber. 1996. LSTM can solve hard long time lag problems. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, pages 473–479. MIT Press.

Morris K. Holland and Michael Wertheimer. 1964. Some physiognomic aspects of naming, or, maluma and takete revisited. *Perceptual and Motor Skills*, 19(1):111–117.

Marcus Hutter. 2001. Distribution of mutual information. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 399–406. MIT Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sebastian Kürschner and Damaris Nübling. 2011. The interaction of gender and declension in Germanic languages. *Folia Linguistica*, 45(2):355–388.

Philip A. Luelsdorff. 1987. *Orthography and phonology*. John Benjamins Publishing.

Z Matějček. 1998. Reading in Czech. part I: Tests of reading in a phonetically highly consistent spelling system. *Dyslexia*, 4(3):145–154.

Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. 2006. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, 9(3):316–322.

Jessica Maye, Janet F. Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):101–111.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Elaine Miles. 2000. Dyslexia may show a different face in different languages. *Dyslexia*, 6(3):193–201.

George Miller. 1955. Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*, pages 95–100.

Padraic Monaghan, Morten H. Christiansen, and Nick Chater. 2007. The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, 55(4):259–305.

Padraic Monaghan, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369:20130299.

Martin Neef, Anneke Neijt, and Richard Sproat. 2012. *The relation of writing to spoken language*, volume 460. De Gruyter.

Elissa L. Newport and Richard N. Aslin. 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2):127 – 162.

Damaris Nübling. 2008. Was tun mit Flexionsklassen? Deklinationsklassen und ihr Wandel im Deutschen und seinen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, pages 282–330.

Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Jeff Parker and Andrea D. Sims. 2015. On the interaction of implicative structure and type frequency in inflectional systems. In *The 1st International Quantitative Morphology Meeting*.

Tiago Pimentel, Arya D. McCarthy, Damiá Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Investigating language universal and specific properties in word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin, Germany. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 726–730, Beijing, China. Association for Computational Linguistics.

Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225.

T. Schlippe, S. Ochs, and T. Schultz. 2012. Grapheme-to-phoneme model generation for Indo-European languages. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4801–4804.

Beate Schwichtenberg and Niels O. Schiller. 2004. Semantic gender assignment regularities in German. *Brain and Language*, 90(1-3):326–337.

Victoria Sharpe and Alec Marantz. 2017. Revisiting form typicality of nouns and verbs. *The Mental Lexicon*, 12(2):159–180.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959.

Richard Sproat. 2000. *A Computational Theory of Writing Systems*. Cambridge University Press.

Richard Sproat. 2012. The consistency of the orthographically relevant level in Dutch. *The Relation of Writing to Spoken Language*, pages 35–46.

Peter Starreveld and Wido La Heij. 2004. Phonological facilitation of grammatical gender retrieval. *Language and Cognitive Processes*, 19(6):677–711.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Henri Theil. 1970. On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76(1):103–154.

Josef Vachek. 1945. Some remarks on writing and phonetic transcription. *Acta Linguistica*, 5(1):86–93.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9. Association for Computational Linguistics.

Bernard L. Welch. 1947. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Michael Wertheimer. 1958. The relation between the sound of a word and its meaning. *The American Journal of Psychology*, 71(2):412–415.

Bernd Wiese. 2000. Warum flexionsklassen? über die deutsche substantivdeklination. In Rolf Thieroff, Matthias Tamrat, Nanna Fuhrhop, and Oliver Teuber, editors, *Deutsche Grammatik in Theorie und Praxis*. De Gruyter.

Adina Williams. 2018. *Representing Relationality: MEG Studies on Argument Structure*. Ph.D. thesis, New York University.

Adina Williams, Damián Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5733–5738, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2020. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*.

Wolfgang Ullrich Wurzel. 1989. *Inflectional morphology and naturalness*, volume 9. Springer Science & Business Media.

A Further Notes on Preprocessing

The breakdown of our declension classes is given in Table 4. We will first discuss more details about our preprocessing for German, and then for Czech.

German. After extracting declension classes from CELEX2, we made some additional preprocessing decisions for German, usually based on orthographic or other considerations. For example, we combined the classes S1 with S4 classes, P1 with P7, and P6 with P3 because the difference between each member of any of these pairs lies solely in spelling (a final <s> is doubled in the spelling when GEN.SG *-(e)s*, or the PL *-(e)n* is attached).

Whether a given singular, say S1, becomes inflected as P1 or P2—or, for that matter, the corresponding umlauted versions of these plural classes—is phonologically conditioned (Alexiadou and Müller, 2008). If the stem ends in a trochee whose second syllable consists of schwa plus /n/, /l/, or /r/, the schwa is not realized, i.e., it gets P2, otherwise it gets P1. For this phonological reason, we also chose to collapse P1 and P2.

We also collapsed all loan classes (i.e., those with P8–P10) under one plural class 'Loan'. This choice resulted in us merging loans with Greek plurals (like P9, *Myth-os* / ***Myth-en***) with those with Latin plurals (like P8, *Maxim-um* / ***Maxim-a*** and P10, *Trauma* / ***Trauma-ta***). This choice might have unintended consequences on the results, as the orthography of Latin and Greek differ substantially from each other, as well as from the native German orthography, and might be affecting our measure of higher form-based MI for S1/Loan and S3/Loan classes in Table 3 of the main text. One could reasonably make a different choice, and instead remove these examples from consideration, as we did for classes with fewer than 20 lemmata.

Czech. The preprocessing for Czech was a bit less involved, since the classes were derived from an edit-distance heuristic. A fluent speaker-linguist identified major noun classes by grouping together nouns with shared suffixes in the surface (orthographic) form. If the differences between two sets of suffixes in the surface form could then be accounted for by positing a basic phonological rule—for example, vowel shortening in monosyllabic words—then the two sets were collapsed.

Among masculine nouns, four large classes were identified that seemed to range from "very animate" to "very inanimate." The morphological divisions between these classes were very systematic, but there was substantial overlap: dat.sg and loc.sg differentiated 'animate1' from 'animate2', 'inanimate1' and 'inanimate2'; acc.sg, nom.pl and voc.pl differentiated 'animate2' from 'inanimate1' and 'inanimate2', and gen.sg differentiated 'inanimate1' from 'inanimate2' (see Figure 3. Further subdivisions were made within the two animate classes for the apparent idiosyncratic nominative plural suffix, and within the 'inanimate2' class, where nouns took either *-u* or *-e* as the genitive singular suffix. This division may have once reflected a final palatal on nouns taking *-e* in the genitive singular case, but this distinction has since been lost. All nouns in the 'inanimate2' "soft" class end in coronal consonants, whereas nouns in the 'inanimate1' "hard" class have a variety of final consonants.

Among feminine nouns, the 'feminine -a' class contained all feminine words that ended in *-a* in the nominative singular form. (Note that there exist masculine nouns ending in *-a*, but these did not pattern with the 'feminine -a' class). The 'feminine pl -e' class contained feminine nouns ending in *-e*, *-ě*, or a consonant, and as the name suggests, had the suffix *-e* in the nominative plural form. The 'feminine pl -i' class contained feminine nouns ending in a consonant and had the suffix *-i* in the nominative plural form. No feminine nouns ended in a dorsal consonant.

Among neuter nouns, all words ended in a vowel.

		animate1	animate2	inanimate1	inanimate2
Singular	nom	-	-	-	-
	gen	a	a	a	u
	acc	a	a	-	-
	dat	ovi	u	u	u
	loc	ovi	u	u	u
	instr	em	em	em	em
	voc	e	e	e	e
Plural	nom	i	i	y	y
	gen	ů	ů	ů	ů
	acc	y	y	y	y
	dat	ům	ům	ům	ům
	loc	ech	ech	ech	ech
	instr	y	y	y	y
	voc	i	i	i	i

Figure 3: Czech paradigm for masculine nouns.

B Some prototypical examples

To explore which examples, across classes might be most prototypical, we samples the top five highest and lowest suprisal examples. The results are

German				Czech		
class	#	classic class	gender(s)	class	#	gender
S1/P1	1157	Decl I	MSC, NEUT	masculine, inanimate2	823	MSC
S3/P3	1105	Decl VI	FEM	feminine, *-a*	818	FEM
S1/P0	264	Singularia Tantum	MSC, NEUT, FEM	feminine, pl *-e*	275	FEM
S1/P5	256	"default *-s* PL"	MSC, NEUT, FEM	neuter, *-o*	149	NEUT
S3/P0	184	Singularia Tantum	MSC, NEUT, FEM	neuter, *-ení*	133	NEUT
S1/P1u	154	Decl II	MSC	masculine, animate2, pl *-i*)	130	MSC
S2/P3	151	Decl V	MSC	masculine, animate1, pl *-i*)	112	MSC
S1/P3	70	Decl IV	MSC, NEUT	feminine, pl *-i*	80	FEM
S3/loan	67	Loanwords	MSC, NEUT, FEM	masculine, animate1, pl *-ové*	55	MSC
S3/P1	11	Decl VIII	FEM	masculine, inanimate1	32	MSC
S1/P4u	51	Decl III	MSC, NEUT	special, masculine, pl *-ata*	26	MSC
S3/P5	49	"default *-s* PL"	MSC, NEUT, FEM	neuter, *-e/-ě/-í*	21	NEUT
S1/loan	41	Loanwords	MSC, NEUT	masculine, animate1, pl *-é*	18	MSC
S3/P1u	35	Decl VII	FEM			
S1/P4	25	Decl III	MSC, NEUT			
S1/P2u	24	Decl II	MSC, phon.			
Total	3684					2672

Table 4: Declension Classes. 'class' refers to the declension class identifier, '#' refers to the number of lexemes in each declension class, and 'gender' refers to the gender(s) present in each class. German declension classes came from CELEX2, for which 'S' refers to a noun's singular form, 'P' refers to its plural, 'classic class' refers to the conception of class from *Brockhaus Wahrig Wörterbuch*.

Czech			German		
stem	class	$\mathrm{H}(C \mid W)$	stem	class	$\mathrm{H}(C \mid W)$
azalka	feminine, *-a*	$6.1x10^{-5}$	*Kalesche*	FEM, 6, S3P3	0.013
matamatika	feminine, *-a*	$6.2x10^{-5}$	*Tabelle*	FEM, 6, S3P3	0.013
čtvrtka	feminine, *-a*	$6.6x10^{-5}$	*Stelze*	FEM, 6, S3P3	0.014
paprika	feminine, *-a*	$6.7x10^{-5}$	*Lende*	FEM, 6, S3P3	0.014
matoda	feminine, *-a*	$6.7x10^{-5}$	*Gamasche*	FEM, 6, S3P3	0.015
ptakopysk	masculine, animate1, pl *-i*	1.34	*Karton*	MSC, 1, S1P5	2.03
špendlík	masculine, inanimate2	1.34	*Humus*	MSC, ?, S3P0	2.06
hospodář	neuter, *-ení*, derived from verb (instr-pl)	1.36	*Mufti*	MSC, 1, S1P5	2.19
dudlík	masculine, inanimate2	1.39	*Magma*	NEU, ?, S1P10	2.23
záznamník	masculine, inanimate2	1.48	*Los*	NEU, 1, S1P1	2.43

Table 5: Five highest and lowest surprisal examples given form and meaning (w2v) by language.

in Table 5. We observe that the lowest surprisal from form for each language generally come from a single class for each language: feminine, *-a* for Czech and S3/P3 for German. These two classes were among the largest, having lower class entropy, and both contained feminine nouns. Forms with higher surprisal generally came from several smaller classes, and were predominately masculine. This sample size is small however, so it remains to be investigated whether this tendency in our data belies a genuine statistically significant relationship between gender, class size, and surprisal.