Universal Patterns of Grammatical Gender in Multilingual Large Language Models

Andrea Schröter
Centre for Language Technology
University of Copenhagen
andreaschroeter02@gmail.com

Ali Basirat
Centre for Language Technology
University of Copenhagen
alib@hum.ku.dk

Abstract

Grammatical gender is a fundamental linguistic feature that varies across languages, and its cross-linguistic correspondence has been a central question in disciplines such as cognitive science and linguistic typology. This study takes an information-theoretic approach to investigate the extent to which variational usable information about grammatical gender encoded by a large language model generalizes across languages belonging to different language families. Using mBERT as a case study, we analyze how grammatical gender is encoded and transferred across languages based on the usable information of the intermediate representations. The empirical results provide evidence that gender mechanisms are driven by abstract semantic features largely shared across languages, and that the information becomes more accessible at the higher layers of the language model.

1 Introduction

Grammatical gender is a nominal category (e.g., masculine, feminine, and neuter) that continues to challenge linguists due to the complexity of gender systems across languages and the rules governing its assignment to nouns (Corbett, 1991; Varlokosta, 2011). These rules vary cross-linguistically and cannot always be inferred from a noun's surface form. For instance, the German *das Mädchen* 'the$_{\text{NEUT}}$ girl' is grammatically neuter despite denoting a female entity, and common concepts such as *sun* differ in gender across languages, masculine in French (*le soleil*) but feminine in German (*die Sonne*).

In addition to its linguistic implications, the study of grammatical gender provides insights into cognitive science (Lucy, 1996; Bender et al., 2011; Kemmerer, 2017; Kann, 2019), assists second language learners in navigating the seemingly arbitrary rules of gender assignment (Sahai and Sharma, 2021), and helps reducing gender bias in language models (Zhou et al., 2019).

Examining grammatical gender from a typological perspective can further illuminate shared linguistic principles contributing to the assignment of grammatical gender across languages. Recent studies in computational linguistics provide clues, based on static multilingual embeddings, about the existence of universal patterns in the assignment of grammatical genders, transferable across several languages (Veeman et al., 2020). However, the linguistic depth and extent of the universal patterns of grammatical gender have remained unexplored, primarily because the multilingual word embeddings do not provide a clear mechanism in distinguishing between formal and semantic features. In particular, it is still unclear whether the linguistic patterns that drive such universalities emerge at the morphological or semantic levels and how the gender system across languages might be related at these levels (Basirat et al., 2021).

On the other hand, previous studies have shown large language models (LLMs) normally encode linguistic information in a more transparent and structured way, allowing for an access into distinct linguistic levels (Peters et al., 2018; Jawahar et al., 2019; Hewitt and Manning, 2019; Tenney et al., 2018; de Vries et al., 2020). Lower layers primarily encode surface-level and morphological features, middle layers capture syntactic structure, and higher layers represent semantic properties and more abstract linguistic features (Jawahar et al., 2019; Tenney et al., 2018). Additionally, later studies show that multilingual LLMs are capable at capturing the universal aspects of languages at their intermediate representations (Pires et al., 2019; Chi et al., 2020), including the grammatical abstractions such as gender (Sukumaran et al., 2024).

Building upon these studies, we employ mBERT (Devlin et al., 2019) to investigate the universal and language-specific aspects of grammatical gender across different linguistic levels, such as morphology and semantics. mBERT, a multilingual

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 34–46

encoder-only language model trained on a diverse set of languages, provides a structured distribution of linguistic information across its intermediate representations. This allows us to systematically examine grammatical gender at multiple linguistic levels. Moreover, its shared feature space across languages facilitates universal analyses.

Taking an information theoretic strategy, we investigate universal aspects of grammatical gender based on the amount of information transferable (generalizable) across gender systems of languages. Specifically, we extend the concept of variational-usable ($\mathcal{V}$-usable) information (Xu et al., 2020) to measure the extent to which the gender information from a source language is generalizable to a target language. A high amount of generalizable information is interpreted as evidence of structural similarities between the gender systems of the source and target languages. In addition to the cross-lingual analysis, the application of $\mathcal{V}$-usable information is also motivated as it allows us to effectively measure the intra-lingual complexity of gender systems.

Our experiments on a typologically diverse set of languages provide empirical evidence that linguistic information about gender is largely generalizable across languages with similar gender categories, while their genealogical relationship plays a secondary role. Additionally, we show that linguistically driven complexities of gender systems are reflected in the hidden representations of the language model, leading to variations in usable information in our intra-lingual analysis. Furthermore, our layer-wise analysis of usable information highlights the varying contributions of intermediate representations to gender encoding, both within and across languages. Finally, further examination of intermediate representations confirms the role of both morphology and semantics in gender representation, with semantic aspects proving to be more generalizable across languages.

Overall, this study adopts a computational approach to explore the relationships between different systems of grammatical gender based on their encoding in the intermediate representations of a large language model. Specifically, the contributions of this study include:

- Systematically evaluating how well grammatical gender information generalizes across languages with different gender systems in a multilingual large language model.
- Introducing a novel approach based on the variational usable information to investigate the generalizability of the intermediate representations of a language model for encoding grammatical gender across languages.
- Probing the intermediate representations to disentangle the roles of morphology and semantics in gender prediction.

2 Grammatical Gender

Grammatical gender is an abstract system of noun classification found in many languages, often overlapping with, or considered as subset of, noun class systems (Comrie, 1999). It is generally considered an inherent property of the noun itself (Spencer, 2002; Cucerzan and Yarowsky, 2003), with determiners, adjectives, and sometimes verbs agreeing with the noun in gender. Although grammatical gender is frequently correlated with biological sex, it is distinct from it, as evidenced by instances of gender-sex mismatches—for example, in German, *das Mädchen* ('the$_{\text{NEUT}}$ girl') is grammatically neuter despite referring to a female entity. Furthermore, grammatical gender should not be conflated with nominal declension classes (Comrie, 1999).

Common gender categories include masculine, feminine, neuter, and common, with Indo-European languages typically featuring masculine/feminine/neuter (e.g., German, Russian), neuter/common (e.g., Danish, Dutch), and masculine/feminine (e.g. French, Italian) gender systems. In contrast, two-gender systems are common in Afro-Asiatic languages (Corbett, 1991). The function of grammatical gender is still debated: some researchers suggest it aids in referent identification or categorization for cognitive processes such as storage and retrieval (Allassonnière-Tang and Kilarski, 2020; Contini-Morava and Kilarski, 2013; Senft, 2000; Lakoff and Johnson, 2008), while others dismiss it as 'historical junk' (Trudgill, 2011).

2.1 Gender Assignment Theories

Grammatical gender assignment can be influenced by formal features (such as morphology, phonology, or orthography) and semantics (Corbett, 1991; Sahai and Sharma, 2021) (e.g. assignment based on biological sex), or it may be entirely arbitrary (Andersson, 1992). However, the rules for gender assignment are far from clear, given their complexity and the exceptions that exist in many languages, further complicated by declension classes, inflectional morphology, and agreement involving num-

ber, case, and gender (Garbo, 2016). This continues to puzzle researchers (Fedden and Corbett, 2019), although several hypotheses have emerged.

Corbett and Fraser (2000) ascribe semantic factors a higher contribution in gender assignment, whereas Rice (2006) argue that formal and semantic features are equally important. Basirat et al. (2021) tested these theories by using character-based embeddings (formal features), context-based embeddings (semantic features), and their combination to predict grammatical gender in Russian, French, and German. Their findings revealed that formal features outperformed semantic ones as predictors of gender, and combining both did not yield significant improvements, challenging both the semantic-dominance and equality hypotheses. Similar results were reported by Sahai and Sharma (2021) who demonstrated that training a classifier using orthographic and semantic features for French results in high accuracy with orthographic features alone, but performance is further enhanced when semantic features are included.

2.2 Gender Systems Across Selected Languages

This study is based on a detailed investigation of gender transfer across seven languages from the Indo-European and Afro-Asiatic language families: Arabic, Beja, Danish, German, Greek, Italian, and Russian. In this section, we briefly overview the gender system of these languages to motivate our discussions in the following sections.

Arabic has a two-gender system (masculine/feminine) and a rich morphology, with verbs, nouns, pronouns, and adjectives agreeing in gender. Gender assignment is based on both semantic (i.e. natural gender) and morphological criteria, although the gender of inanimate nouns is often semantically arbitrary, e.g. *baab* 'door.MASC'. At the morphology level, masculine nouns are unmarked, while feminine nouns are overtly marked by suffixes, e.g., *shajar-ah* 'tree-FEM' (Alkohlani, 2016).

Beja, an Afro-Asiatic language of the Cushitic branch, classifies nouns into masculine and feminine. Gender is primarily marked on nouns through prefixes and suffixes, which agree with adjectives, pronouns, and other modifiers within the noun phrase (NP) in terms of case, number, and gender. For example, the prefix *ʔuu* agrees in case, number, and gender with the noun *gáw*: *ʔuu-gáw* 'MASC.NOM.SG.DEF-house' (Appleyard, 2007).

Danish has a two-gender system of common/neuter, with common historically formed by merging masculine and feminine. Gender is not overtly marked on the noun itself, but appears in definite NPs through determiner suffixes that agree with the noun's gender (e.g., *hus-et* 'house-NEUT.DEF.SG', *bil-en* 'car-COM.DEF.SG') or with indefinite determiners (e.g., *et hus* 'a_{NEUT} house', *en bil* 'a_{COM} car'). Adjectives also show gender agreement with the noun, but there is no gender marking in the plural (Gregersen et al., 2021).

German has three grammatical genders (masculine/feminine/neuter), and determiners and adjectives agree with the noun in gender (e.g., *ein-e schön-e Frau* 'a-FEM beautiful-FEM woman'). However, the language also has a complex case system that interacts with gender marking. Gender assignment in German is considered complex, influenced by both semantic factors or clusters (e.g., all fruits are feminine) and morphological features (e.g. all nouns with the suffix *-heit* are feminine) (Bender et al., 2011; Fedden and Corbett, 2019). Despite these patterns, gender assignment in German is often perceived as arbitrary, with many exceptions (Fedden and Corbett, 2019).

Greek uses a three-gender system of masculine/feminine/neuter in which gender assignment is predominantly based on formal features (Varlokosta, 2011; Corbett, 1991), although semantic rules exist, e.g. fruits and vegetables are often assigned neuter case. Gender is often overtly marked on the noun: for instance, masculine nouns frequently end in *-as* (e.g., *ándras*, 'man'), *-os*, or *-ís*; feminine nouns often end in *-í* (e.g., *psychí*, 'soul') or *-a*; and neuter nouns tend to end in *-o* (e.g., *moró*, 'baby'), *-í*, or *-ma*, although exceptions exist, such as the neuter noun *kréas*, 'meat'. Greek is morphologically rich, with adjectives and determiners requiring agreement in gender, number, and case, complicating the prediction of gender.

In Italian, nouns are categorized into masculine and feminine gender, where masculine nouns typically have an *o*-suffix (*il naso* '$the_{MASC.SG}$ nose'), and feminine nouns end in *-a* (*la mela* '$the_{FEM.SG}$ apple'), with few exceptions, e.g. *il pianeta* 'the planet' and *la mano* 'the hand'. However, nouns with *e*-suffixes can be either masculine or feminine. Gender is considered to be based on both formal and semantic features (Bianchi, 2013).

Russian has a three-gender system (masculine, feminine, neuter). The language's rich morphology and complex inflection system (Parker and Sims, 2020) require adjectives and numerals

to agree with nouns in case, gender, and number. Gender is overtly marked on the noun, as seen in *zhenshchin-a* ('woman-FEM.SG.NOM') and *zhenshchin-u* ('woman-FEM.SG.ACC'). Gender assignment follows both morphological and semantic rules, such as the features [+male] and [+female] (Fraser and Corbett, 1994). However, exceptions exist, like *mužčina* ('man'), a masculine noun ending in *-a*. In the absence of semantic features, gender is assigned according to the declension class (Nikunlassi, 2000). Additionally, animacy plays a role in accusative marking for masculine, animate nouns. In such cases, the accusative form coincides with the genitive, marked by the *-a* suffix (e.g., *Ya vizhu student-a* 'I see student-MASC.SG.ACC.ANIM'), further complicating the distinction between gender classes.

3 Related Work

Veeman et al. (2020) investigate universal patterns in grammatical gender using a set of static multilingual word embeddings. Their study primarily employs a neural transfer learning approach, where the accuracy of gender classification from a source training language to a target test language serves as an indicator of similarity between their gender systems. Their findings suggest that while some factors influencing gender assignment are universal, as evidenced by successful cross-lingual transfer, others are idiosyncratic to specific language families. Similarly, Veeman and Basirat (2020) explored how different types of multilingual word embeddings capture information about grammatical gender and how well this information is transferable between languages. Their findings reveal an overlap in the encoding of gender in Swedish, Danish, and Dutch.

We extend the investigations of Veeman et al. (2020) in two key ways. First, instead of accuracy as a transferability metric, we adopt variational usable information (Xu et al., 2020), allowing for a comparative analysis of gender system complexity across languages (Ethayarajh et al., 2022). Second, we investigate gender universalities at a deeper linguistic level by analyzing the generalizability of usable information across different layers of a multilingual LLM, instead of static word embeddings.

Several studies have addressed the encoding of grammatical gender in word embeddings. For instance, Basirat and Tang (2018, 2019) study how a set of static word embeddings encode grammatical gender of Swedish nouns and Basirat et al. (2021) investigate the contribution of the formal and semantic features encoded in word embeddings into the assignment of grammatical gender. Additional approaches, including surrogate models and decision trees (Sahai and Sharma, 2021), have further illuminated the mechanisms behind gender prediction. For instance, Sukumaran et al. (2024) found that transformer models can generalize grammatical gender from minimal examples, albeit with a masculine bias.

4 Method

Our investigation of gender transfer spans both layers of a language model and languages. Specifically, we assess gender transferability across languages by measuring the information each intermediate layer provides for gender prediction. For this purpose, we adopt $\mathcal{V}$-usable information (Xu et al., 2020), an extension of Shannon mutual information (Shannon, 1948) that accounts for computational constraints. The $\mathcal{V}$-usable information in a random variable X for predicting a category Y is defined as the difference in conditional entropy between predictions based on X and a baseline prediction with no input features (denoted as Φ):

$$I_{\mathcal{V}}(Y;X) = H(Y \mid \Phi) - H(Y \mid X) \quad (1)$$

A higher value of $I_{\mathcal{V}}(Y;X)$ indicates that X significantly reduces uncertainty in predicting Y.

In our setting, X is a random vector in an embedding space formed by a hidden layer of a language model while processing nouns in a given language, and Y represents a probability vector of grammatical genders. To quantify the amount of information encoded in the embedding space of a source language i, denoted as X_i, for predicting grammatical genders in a target language j, denoted as Y_j, we extend Equation 1 as:

$$I_{\mathcal{V}}(Y_j;X_i) = H_{\mathcal{V}}(Y_j \mid \Phi) - H_{\mathcal{V}}(Y_j \mid X_i) \quad (2)$$

Intuitively, $I_{\mathcal{V}}(Y_j;X_i)$ measures the usable information that the embeddings from the source language provide for predicting gender in the target language. A high value of $I_{\mathcal{V}}(Y_j;X_i)$ suggests a strong similarity between the gender systems of the source and target languages.

In some cases that the gender systems are highly different from each other, for example when a gender category is seen in a language but not in the other (e.g., common is in Danish but not in Arabic), $I_{\mathcal{V}}(Y_j;X_i)$ can be negative or an invalid number.

We set the negative values and invalid numbers to zero to satisfy the non-negativity constraint of usable information and manage the NaN exceptions.

For a given language pair and hidden layer, we calculate the marginal entropy $H_{\mathcal{V}}(Y_j \mid \Phi)$ in Equation 2 based on the gender distribution of the target language and approximate the conditional entropy $H_{\mathcal{V}}(Y_j \mid X_i)$ using a light classifier trained on embedding-gender pairs from the source language i. The cross-entropy loss on a test sample from the target language j is then used as an estimate of $H_{\mathcal{V}}(Y_j \mid X_i)$. To address class imbalance, cross-entropy loss is weighted by the gender distribution in the source language.

5 Experiment Setup

We investigate transfer learning of grammatical gender across a typologically diverse set of languages with different grammatical gender systems and minimal lexical similarity, as outlined in Table 1. Except for Danish-German, where Danish is included to broaden gender systems, this design choice helps minimize reliance on surface-level lexical overlap. The data is sourced from Universal Dependencies (v. 2.14) (Nivre et al., 2016), where nominal gender annotations are included as part of the inflectional features. For each language, we concatenate all treebanks that include gender annotations.

Language	Family	M	F	N	C
Arabic	AA-Semitic	67	33	0	0
Beja	AA-Cushitic	75	25	0	0
Danish	IE-Germanic	0	0	31	69
German	IE-Germanic	37	41	22	0
Greek	IE-Hellenic	19	52	29	0
Italian	IE-Romance	55	45	0	0
Russian	IE-Slavic	45	35	20	0

Table 1: Gender distribution (%) in the test languages. M: masculine. F: feminine. N: neuter. C: common. AA: Afro-Asiatic. IE: Indo-European.

The experiments are based on the multilingual BERT model (mBERT) (Devlin et al., 2019) consisting of 12 layers plus an initial embedding layer each with 768 features. The model is trained on a data set including text from an extensive range of 104 languages, including all our test languages except Beja, which we have intentionally selected to assess the degree of cross-lingual transfer beyond mBERT's training languages.

Following Veeman et al. (2020), we extract embeddings from the formal representations of nouns provided in the `FORM` column of the Universal Dependencies treebanks. In cases where tokens are divided into subtokens, we average their embeddings. Since the experiments are based on cross-lingual transfer, morphosyntactic gender indicators from the source language are absent in the target language input. This design helps ensure that model performance reflects genuine cross-linguistic generalization rather than reliance on surface-level lexical cues in the target language.

For each sentence in a language, we construct a dictionary that maps the contextual embeddings of its nouns to their grammatical gender. The embeddings are extracted from all layers of mBERT, resulting in an embedding matrix of 768×13 for each occurrence of a noun in a language.

A known limitation of using mBERT in multilingual settings is that model performance can vary across languages due to differences in their representation in the pretraining data (Wu and Dredze, 2020). To mitigate this imbalance and ensure cross-linguistic comparability, we downsample the total number of nouns per language to match the smallest sample size in Arabic (5,151 nouns). For Beja, a total of 555 nouns are included.

We estimate the usable information based on 7×13 logistic regression models, corresponding to the number of languages and hidden layers. Each classifier is trained for 30 epochs with early stopping (patience of 5 epochs). We use AdamW as the optimizer with a manually tuned learning rate of 5×10^{-5}, and a learning rate scheduler that reduces the learning rate every 10 epochs by a factor of 0.1.

The train-validation-test split for all languages is 80-10-10. Training is performed on the training (80%) and validation (10%) splits, while cross-lingual entropy is estimated using the test split (10%). For each source-target language pair, $H_{\mathcal{V}}(Y_j \mid X_i)$ in Equation 2 is estimated as the average cross-entropy loss over five random seeds.

6 Results

In this section, we present and analyze the results of our experiments through 1) intra-lingual analysis of the usable information, 2) their transferability across languages, and 3) their variations across layers and languages.

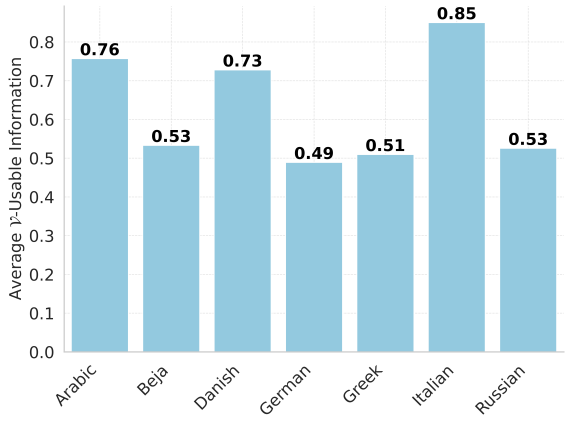

Figure 1: Averaged intra-lingual usable information.

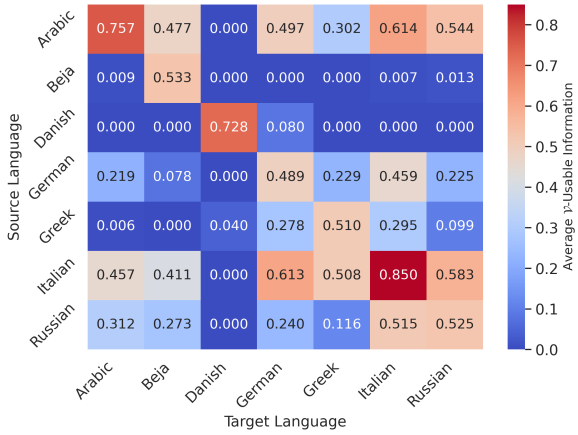

Figure 2: Averaged cross-lingual usable information.

6.1 Intra-lingual Analysis

We begin by examining the intra-lingual results of usable information for predicting gender (i.e., when the source and target languages are the same). Figure 1 presents the average usable information for predicting gender within languages across all layers of mBERT. The differences in the results can be explained by two factors: the complexity of the gender system and the quality of the intermediate representations for each of the test languages.

In general, the differences in usable information can be interpreted as variations in the complexity of a target task (Ethayarajh et al., 2022). Specifically, in the case of languages seen in the mBERT's training data, it indicates that the intermediate representations are significantly more informative about grammatical gender in Arabic, Danish, and Italian than in German, Greek, and Russian. This observation aligns with linguistic evidence, as the latter group of languages has more complex gender systems in different ways. Firstly, Arabic, Danish, and Italian have only two grammatical genders, whereas German, Greek, and Russian have three. Additionally, the former group has relatively predictable gender assignment patterns, often rooted in morphological inflections and syntactic agreements. In contrast, languages such as German, Greek, and Russian have more intricate agreement systems with numerous inflectional irregularities, making gender prediction more challenging. More details about the gender systems of the languages can be seen in Section 2.2.

Given Beja's absence from mBERT's training data, we speculate that the moderate information for its gender prediction originates from typologically related languages in pretraining, such as Arabic, which also has a two-gender system (masculine/feminine).

Beyond linguistic factors, variations in intra-lingual usable information may also be influenced by the quality of intermediate representations, which are, in turn, affected by the distribution of training data for each language in mBERT's pretraining corpus. However, this remains difficult to analyze, as the exact composition of mBERT's training data has not been publicly disclosed.

6.2 Cross-lingual Analysis

Figure 2 summarizes the usable information for predicting gender in a target language based on the information gained from a source language. The results, averaged over the intermediate layers of mBERT, provide clear evidence about the varying transferability of gender across languages. Danish and Beja have the least generalizable systems, while Arabic and Italian demonstrate the highest.

The poor cross-lingual performance of Danish can be attributed to its unique common/neuter gender system, the only such system in our study. However, transfer to Beja appears more feasible despite its absence from mBERT's training data. Notably, Arabic $\rightarrow$ Beja achieves relatively strong transfer, followed by Italian. This pattern likely reflects the structural similarity of their gender systems, with Arabic benefiting additionally from genetic relatedness and language contact with Beja (Vanhove, 2012). Moreover, the limited orthographic overlap between Beja and other test languages indicates that this successful transfer cannot be attributed to surface formal features at the tokenization level; rather, it is likely due to deeper cross-linguistic representations in mBERT that capture

universal patterns of gender assignment (Veeman et al., 2020). This effect may also be strengthened by indirect transfer from typologically related languages present in mBERT's training data and by loanwords from Arabic.

The cross-lingual results in Figure 2 indicate that gender information generalizes more effectively from Arabic and Italian to other languages, except for Danish, which has entirely different gender categories. Arabic transfers best to Italian, as both languages share the same gender categories (i.e., masculine and feminine), and moderately well to languages that also include a neuter gender. Similarly, Italian transfers well to languages with both masculine and feminine genders. This suggests a strong alignment between the masculine and feminine genders in Arabic and Italian and their counterparts in other languages. Still further investigation is needed to explain special cases, such as Italian → German, where cross-lingual transfer is more informative than mono-lingual.

Surprisingly, both German and Russian provide nearly the same amount of information for predicting gender in Italian as they do in their own monolingual settings. This indicates strong structural similarities between these languages, which is likely the result of partially similar morphosyntactic gender agreement in these languages, as discussed in Section 2.2 and their alignment in the masculine and feminine categories, as discussed earlier in this section. A deep investigation of this phenomenon falls outside the scope of this paper.

For Greek as a source language, moderate transfer is achieved to German and Italian, with an average score of 0.3, and fairly low results on other languages. The low transfer to Arabic and Beja can be due to the differences in the number of grammatical genders and the distant genealogical relationship between these languages and Greek. The near-zero transfer to Russian is likely due to differences in declension systems, agreement rules, and the higher number of exceptions and irregularities in Russian, which is also reflected in the transfer from Russian to Greek.

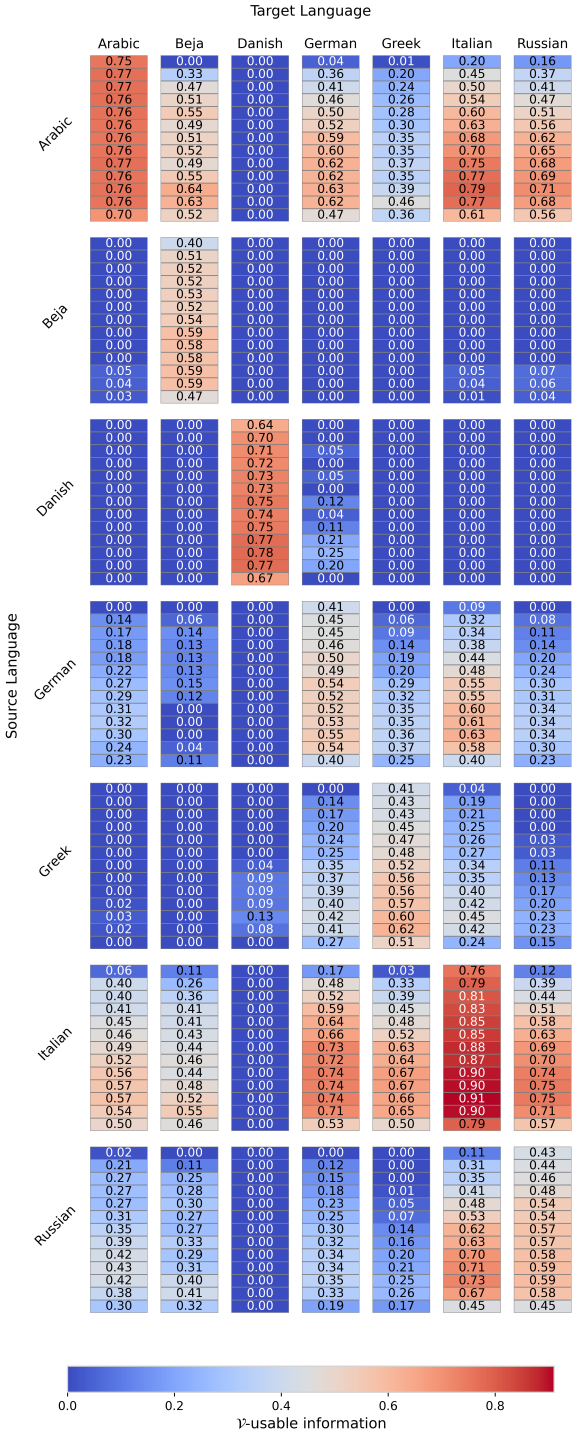

Figure 3: Mean usable information across language pairs and layers. Layers are ordered top-down, from the embedding layer to layer 12.

6.3 Layer-wise Analysis

In this section, we provide detailed analyses of usable information in the intermediate representations for predicting gender across languages. The results across layers and languages are represented in Figure 3. The unnormalized results, including the negative usable information and standard deviations, are also visualized in Appendix A.

The usable information for gender prediction increases from the initial layer, peaks between Layers 9-11, and slightly decreases at the final layer. This trend is visible in both monolingual and cross-lingual settings. One exception, however, is German → Beja, where the trend experiences a

drop after Layer 6. Further distinct visualizations of the trends by each language are in Appendix B.

The increasing trend in usable information highlights the importance of the number of transformer layers in encoding gender information generalizable across languages. To further assess the significance of higher contribution of top layers to cross-lingual gender transfer, we group the model's layers into two groups: (1) the lower layers, including the embedding layer (0) and Layers 1–6, and (2) the upper layers (6–12). For each source-target language pair, we compute the mean performance of these two groups and test whether the top layers are significantly more informative than the lower layers. We conduct a paired t-test and a Wilcoxon signed-rank test, comparing the mean performance of the lower and upper layers, for settings with a positive sum of usable information across layers.

Both statistical tests show that the usable information scores are significantly higher in the upper layers compared to the lower layers (paired t-test: $t = -7.93$, $p = 1.12 \times 10^{-9} < 0.001$; Wilcoxon: $W = 11.00$, $p = 4.22 \times 10^{-7} < 0.001$). These results suggest that cross-lingual gender transfer is primarily driven by linguistic features encoded in the middle to late layers, indicating that semantic features contribute more to gender assignment than formal features encoded in lower layers (Corbett and Fraser, 2000; Tenney et al., 2018).

The increasing trend in the usable information persists even in the monolingual setting, where the classifier is trained and tested on the same language (paired t-test: $t = -3.52$, $p = 0.006 < 0.01$; Wilcoxon: $W = 1.0$, $p = 0.03 < 0.05$). This pattern is also observed in languages where gender is explicitly marked morphologically on nouns through their formal features. One example is Italian, with few exceptions in the marking of feminine and masculine nouns, as mentioned in Section 2.2 (mean lower layer score = 0.82, mean upper layer score = 0.88). These results support the semantic-dominance hypothesis proposed by Corbett and Fraser (2000) and are consistent with the findings of Sahai and Sharma (2021) for French, which suggest that while orthographic and formal features alone yield high accuracy, performance improves further when semantic features are incorporated.

Notably, we observe a consistent performance drop in the final layer (see Figure 3). A possible explanation is that the last layer of mBERT encodes more abstract linguistic knowledge and long-range dependencies, which may be less relevant for gender prediction (Puccetti et al., 2021; Peters et al., 2018). Similar declines in the final layers' performance are also reported in general for higher-level linguistic probing tasks (Kunz and Kuhlmann, 2022).

7 Conclusions

This study explores the cross-linguistic transferability of grammatical gender in multilingual language models, focusing on the extent to which gender information generalizes across languages with different gender systems. Using variational-usable ($\mathcal{V}$-usable) information, we quantify how grammatical gender is encoded within and across languages in mBERT. Our findings reveal that gender information is more transferable between languages that share similar gender categories, whereas genealogical relationships play a secondary role.

Through intra-lingual analysis, we demonstrate that the complexity of a language's gender system is reflected in the amount of usable information available in the intermediate representation of mBERT. Our cross-lingual results highlight that languages with two-gender systems, such as Arabic and Italian, exhibit the highest transferability, particularly to languages with similar gender distinctions. In contrast, languages with more complex gender systems, such as German and Russian, show reduced transfer due to the added complexity of declension systems and irregularities in gender assignment.

A layer-wise analysis further reveals that intermediate representations in mBERT play a critical role in encoding gender information. Gender distinctions are captured more effectively in the middle-to-upper layers, supporting the idea that semantic information is more generalizable across languages than purely morphological features.

Overall, our findings contribute to a deeper understanding of how grammatical gender is represented in multilingual LLMs and offer insights into universal aspects of grammatical gender. Future research could extend this analysis to other multilingual models (e.g., mGPT, BLOOM) and investigate additional factors influencing gender transfer, such as word frequency effects, training data composition, and finer-grained linguistic features. Expanding the study to a broader range of languages beyond Indo-European and Afro-Asiatic families would further enhance our understanding of cross-linguistic gender representation.

Limitations

A limitation of this study is the relatively small selection of languages analyzed. To better generalize cross-linguistic patterns in grammatical gender assignment, it is crucial to evaluate transfer learning across a more diverse set of languages, particularly from underrepresented language families. Additionally, our analysis is based solely on mBERT, an encoder-only model, which may limit the scope of the findings. Expanding the study to include additional multilingual language models, such as mT5, BLOOM, or mGPT, could provide more reliable and comprehensive insights into the transferability of grammatical gender. Another potential limitation is the uneven distribution of training data across languages in mBERT, which may influence gender transferability. Low-resource languages likely have weaker representations, affecting gender predictability. A broader investigation of training data composition and its impact on gender encoding would help disentangle model-specific biases from linguistic typology.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback on this paper. We also acknowledge the Danish e-Infrastructure Consortium (DeiC) for providing computational resources through UCloud, supported under the Linguistic Universals in Language Models project.

References

Fatima A Alkohlani. 2016. The problematic issue of grammatical gender in arabic as a foreign language. *Journal of Language and Cultural Education*, 4(1):17–28.

Marc Allassonnière-Tang and Marcin Kilarski. 2020. Functions of gender and numeral classifiers in nepali. *Poznan Studies in Contemporary Linguistics*, 56(1):113–168.

Anders-Börje Andersson. 1992. Second language learners' acquisition of grammatical gender in swedish.

David Appleyard. 2007. Beja morphology. *Morphologies of Asia and Africa*, 1:447–481.

Ali Basirat, Marc Allassonnière-Tang, and Aleksandrs Berdicevskis. 2021. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard*, 7(1):20200048.

Ali Basirat and Marc Tang. 2018. Lexical and morpho-syntactic features in word embeddings-a case study of nouns in swedish. In *Special Session on Natural Language Processing in Artificial Intelligence*, pages 663–674. SCITEPRESS-Science and Technology Publications.

Ali Basirat and Marc Tang. 2019. Linguistic information in word embeddings. In *Agents and Artificial Intelligence: 10th International Conference, ICAART 2018, Funchal, Madeira, Portugal, January 16–18, 2018, Revised Selected Papers 10*, pages 492–513. Springer.

Andrea Bender, Sieghard Beller, and Karl Christoph Klauer. 2011. Grammatical gender in german: A case for linguistic relativity? *Quarterly Journal of Experimental Psychology*, 64(9):1821–1835.

Giulia Bianchi. 2013. Gender in italian–german bilinguals: A comparison with german l2 learners of italian. *Bilingualism: Language and Cognition*, 16(3):538–557.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Bernard Comrie. 1999. Grammatical gender systems: a linguist's assessment. *Journal of Psycholinguistic research*, 28:457–466.

Ellen Contini-Morava and Marcin Kilarski. 2013. Functions of nominal classification. *Language sciences*, 40:263–299.

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

Greville G Corbett and Norman M Fraser. 2000. Gender assignment: a typology and a model. In *Systems of Nominal Classification (Language, Culture and Cognition 4)*, pages 293–325. Cambridge University Press.

Silviu Cucerzan and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Sebastian Fedden and Greville G Corbett. 2019. The continuing challenge of the german gender system. In *International Symposium of morphology*.

Norman M Fraser and Greville G Corbett. 1994. Gender, animacy, and declensional class assignment: A unified account for russian. In *Yearbook of morphology 1994*, pages 123–150. Springer.

Francesca Di Garbo. 2016. Exploring grammatical complexity crosslinguistically : The case of gender. *Linguistic Discovery*, 14:46–85.

Frans Gregersen, Leonie Cornips, and Ditte Boeg Thomsen. 2021. The acquisition of grammatical gender of determiners in danish monolingual and bilingual children: An experimental study. *Journal of Germanic Linguistics*, 33(2):147–178.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Katharina Kann. 2019. Grammatical gender, neowhorfianism, and word embeddings: A data-driven approach to linguistic relativity. *arXiv preprint arXiv:1910.09729*.

David Kemmerer. 2017. Categories of object concepts across languages and brains: the relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition and Neuroscience*, 32(4):401–424.

Jenny Kunz and Marco Kuhlmann. 2022. Where does linguistic information emerge in neural language models? measuring gains and contributions across layers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4664–4676.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

John A Lucy. 1996. *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis*. Cambridge University Press.

Ahti Nikunlassi. 2000. On gender assignment in russian. *Trends in linguistic studies and monographs*, 124:771–792.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeff Parker and Andrea D Sims. 2020. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of russian nouns. *The complexities of morphology*, pages 23–51.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How do bert embeddings organize linguistic knowledge? In *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pages 48–57.

Curt Rice. 2006. Optimizing gender. *Lingua*, 116(9):1394–1417.

Saumya Sahai and Dravyansh Sharma. 2021. Predicting and explaining french grammatical gender. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 90–96.

Gunter Senft. 2000. What do we really know about nominal classification systems? In *Systems of nominal classification*, pages 11–49. Cambridge University Press.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Andrew Spencer. 2002. Gender as an inflectional category. *Journal of Linguistics*, 38(2):279–312.

Priyanka Sukumaran, Conor Houghton, and Nina Kazanina. 2024. Investigating grammatical abstraction in language models using few-shot learning of novel noun gender. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 747–765, St. Julian's, Malta. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. *International Conference on Learning Representations*.

Peter Trudgill. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press, USA.

Martine Vanhove. 2012. Roots and patterns in beja (cushitic): The issue of language contact with arabic. In Martine Vanhove, Thomas Stolz, Hitomi Otsuka, and Aina Urdze, editors, *Morphologies in contact*, pages 311–326. Akademie Verlag, Berlin.

Spyridoula Varlokosta. 2011. The role of morphology in grammatical gender assignment. *Morphology and its interfaces*, 178:321.

Hartger Veeman, Marc Allassonnière-Tang, Aleksandrs Berdicevskis, and Ali Basirat. 2020. Cross-lingual embeddings reveal universal and lineage-specific patterns in grammatical gender assignment. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 265–275, Online. Association for Computational Linguistics.

Hartger Veeman and Ali Basirat. 2020. An exploration of the encoding of grammatical gender in word embeddings. *arXiv preprint arXiv:2008.01946*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *International Conference on Learning Representations*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. 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 5276–5284, Hong Kong, China. Association for Computational Linguistics.

A Appendix

Original results for the averaged $\mathcal{V}$-usable information across layers for cross-lingual transfer between language pairs. Negative values were converted to zero to respect the boundaries of $\mathcal{V}$-usable information (see Section 6).

B Appendix

Averaged $\mathcal{V}$-usable information across layers for each source language, illustrating transfer scores to all target languages in the study.

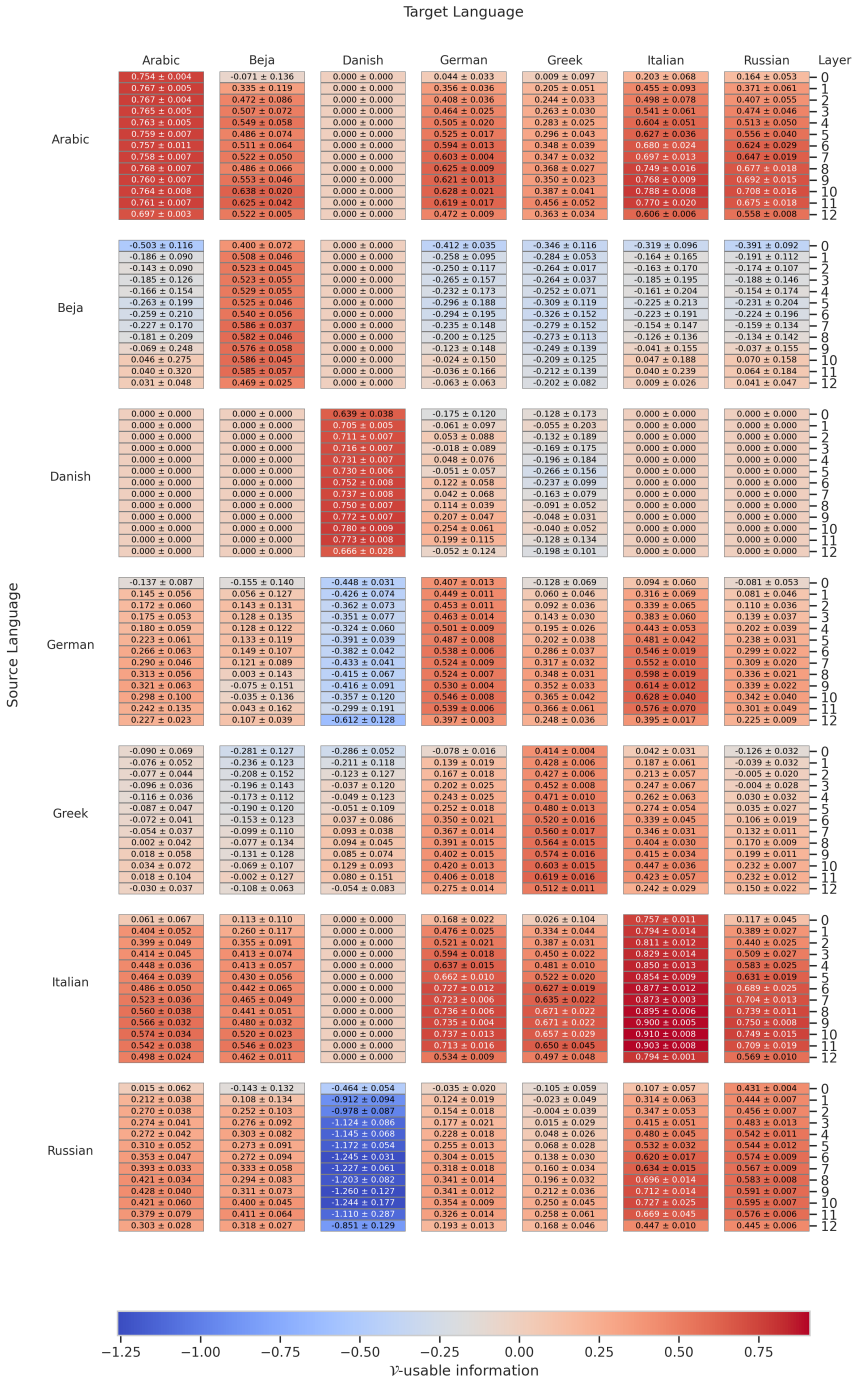

Figure 4: Mean usable information across language pairs and mBERT layers.

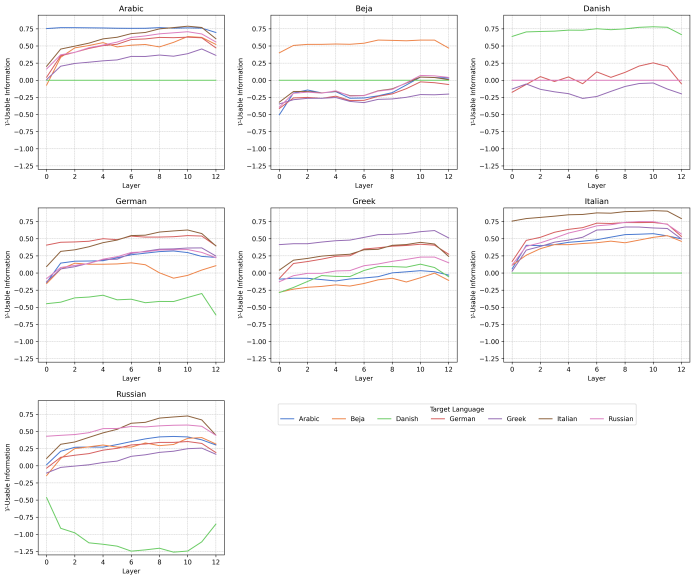

Figure 5: Averaged $\mathcal{V}$-usable information across mBERT layers for each source language, with transfer scores to all target languages.

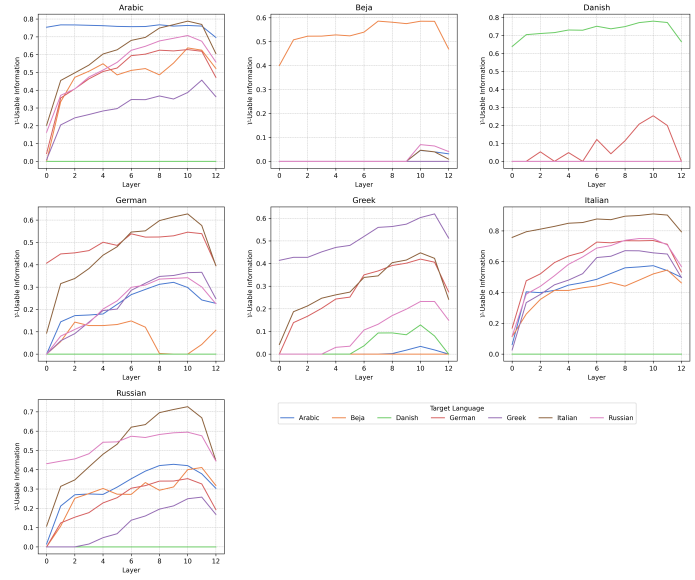

Figure 6: Averaged $\mathcal{V}$-usable information across mBERT layers for each source language, with transfer scores to all target languages, after setting negative values to zero.