Investigating Language Relationships in Multilingual Sentence Encoders Through the Lens of Linguistic Typology

Rochelle Choenni
University of Amsterdam
The Institute for Logic, Language and
Computation (ILLC)
r.m.v.k.choenni@uva.nl

Ekaterina Shutova
University of Amsterdam
The Institute for Logic, Language and
Computation (ILLC)
e.shutova@uva.nl

Multilingual sentence encoders have seen much success in cross-lingual model transfer for downstream NLP tasks. The success of this transfer is, however, dependent on the model's ability to encode the patterns of cross-lingual similarity and variation. Yet, we know relatively little about the properties of individual languages or the general patterns of linguistic variation that the models encode. In this article, we investigate these questions by leveraging knowledge from the field of linguistic typology, which studies and documents structural and semantic variation across languages. We propose methods for separating language-specific subspaces within state-of-the-art multilingual sentence encoders (LASER, M-BERT, XLM, and XLM-R) with respect to a range of typological properties pertaining to lexical, morphological, and syntactic structure. Moreover, we investigate how typological information about languages is distributed across all layers of the models. Our results show interesting differences in encoding linguistic variation associated with different pretraining strategies. In addition, we propose a simple method to study how shared typological properties of languages are encoded in two state-of-the-art multilingual models—M-BERT and XLM-R. The results provide insight into their information-sharing mechanisms and suggest that these linguistic properties are encoded jointly across typologically similar languages in these models.

1. Introduction

Early work in multilingual NLP focused on creating task-specific models, and can be divided into two main approaches: language transfer (Täckström, McDonald, and Nivre

Action Editor: Zhiyuan Liu. Submission received: 30 August 2021; revised version received: 17 January 2022; accepted for publication: 30 March 2022.

https://doi.org/10.1162/coli_a_00444

2013; Tiedemann, Agić, and Nivre 2014; Banea et al. 2008) and multilingual joint learning (Ammar et al. 2016a,b; Zhou et al. 2015). The former method enables the transfer of models or data from high- to low-resource languages, hence porting information across languages, while the latter aims to leverage language interdependencies through joint learning from annotated examples in multiple languages. The inspiration for language transfer was drawn from the fact that, despite having significantly different lexica and syntactic structures, languages still tend to exhibit similarities that can be exploited (Ponti et al. 2019). In multilingual joint learning, models acquire information about multiple languages simultaneously with the assumption that these languages can support each other and enhance each other's representation and processing quality (Ammar et al. 2016a; Navigli and Ponzetto 2012), even in cases where both languages suffer from data scarcity (Khapra et al. 2011).

Despite their differences, both methods rely on the fact that there are dependencies between processing different languages from a typological perspective. For instance, some syntactic properties are universal across languages (e.g., nouns take adjectives and determiners as dependents, but not adverbs), but others are influenced by the typological features of each language (e.g., the order of these dependents with respect to the parent) (Naseem, Barzilay, and Globerson 2012). While the success of such transfer and joint models was limited (and remained inherently bilingual rather than multilingual), they paved the way for the idea that effective multilingual NLP systems could be built by efficiently handling and exploiting language similarity and variation.

In recent years, with the rise of deep learning in NLP, the development of large-scale monolingual pretraining methods for word representations (Pennington, Socher, and Manning 2014) and sentence encoders (Peters et al. 2018b; Devlin et al. 2019) has led to substantial performance improvements in a wide variety of NLP tasks. These techniques produce linguistically informed priors that allow for effective model fine-tuning to obtain task-specific text representations. Pretraining such general-purpose models, however, requires access to a vast amount of training data in a given language, and the effectiveness of fine-tuning them to a specific task depends on the availability of large datasets annotated for this task (Yogatama et al. 2019). As a result, these techniques, along with the success they bring to NLP technology, were limited to a handful of high-resource languages only, for which such datasets are available.

Aiming to extend the benefits of large-scale pretraining to low-resource languages, many studies focused on the development of models with a wider cross-lingual applicability, giving a new surge to the field of multilingual NLP. Research in this field has, thus far, led to the development of multilingual word embeddings (Ammar et al. 2016b; Chen and Cardie 2018) and sentence encoders, such as LASER (Artetxe and Schwenk 2019), Multilingual BERT (M-BERT) (Devlin et al. 2019), XLM (Lample and Conneau 2019), and XLM-R(oBERTa) (Conneau et al. 2020). These encoders are trained to project words and sentences from multiple languages into a shared multilingual semantic space, irrespective of their source language, such that their meaning can be captured more universally. Moreover, these models rely on different types of neural architectures (e.g., recurrent neural networks and Transformers) and pretraining strategies, namely, using monolingual (M-BERT, XLM-R) or cross-lingual (LASER) training objectives, or a combination thereof (XLM). Whereas models trained with cross-lingual objectives exploit parallel data for supervision, the models that rely on monolingual data are unsupervised. Having been trained on many languages, these encoders can be expected to induce shared common underlying patterns of different languages in a data-driven manner without any explicit typological guidance.

Research Questions. While work on large-scale multilingual models has met with success, enabling effective model transfer across many languages (Wu and Dredze 2019), little is known about the linguistic properties of individual languages that such models encode. Nor do we understand to what extent these models capture the patterns of cross-lingual similarity and variation. Thus far, however, little research has paid attention to investigating the linguistic properties of individual languages that pretrained multilingual representations encode. Based on our previous discussion, we derived the following hypotheses:

H(1) The large-scale pretrained general-purpose multilingual models, like the earlier tasks-specific NLP systems, (implicitly) rely on encoding typological relationships of languages.

H(2) Some of their effectiveness stems from the fact that the data-driven approach to uncovering underlying patterns enables them to more efficiently encode and share language-specific properties compared with the earlier models that had to rely on explicit typological guidance.

H(3) The different pretraining strategies used for each large-scale multilingual model (i.e., inherently monolingual vs. cross-lingual) influence the way in which the model learns to uncover shared language patterns. Hence, models might learn different typological relationships based on their training objectives.

In this work we propose methods that stem from the long line of research on interpretability of neutral models for studying language-specific properties in multilingual sentence encoders. In addition, we examine cross-lingual interaction of linguistic information within M-BERT and XLM-R, through the lens of linguistic typology. More concretely, we study the following set of questions:

Q(1) What language-specific typological properties do pretrained language models encode? (**H**(1))

Q(2) Where in the models (i.e., in which layers) is this information encoded? And is this information localizable to specific layers or rather spread across layers? (**H**(1))

Q(3) Are there systematic differences that can be ascribed to the type of pretraining strategy used? (**H**(3))

Q(4) How do multilingual models share information across a large set of typologically diverse languages? For instance, some shared properties of languages may be encoded jointly in the model, while others may be encoded separately in their individual subspaces. (**H**(2))

Methodology. Libovický, Rosa, and Fraser (2020) and Gonen et al. (2020) demonstrated that representations produced by M-BERT are projected to separate language-specific subspaces. Hence, they can be dissected into a language-neutral component, which captures the underlying meaning, and a language-specific component, which captures language identity and its linguistic properties. We use this language-specific component as the basis for our experiments, exploiting it as a means to locate the language-specific properties of languages encoded within the models.

To this end, we propose a set of 25 language-level probing tasks, which draw inspiration from the field of linguistic typology, to test for the language relationships that are encoded by these components specifically. These tasks are designed to test whether it is possible to successfully separate language-specific subspaces within multilingual encoders by the linguistic typological properties of the languages. We rely on the World Atlas of Language Structures database (Dryer and Haspelmath 2013) as a source of typological information, and investigate variation along a wide range of linguistic properties, pertaining to lexical, morphological, and syntactic structure. Using our 25 tasks, we test for which of the typological properties we are able to separate languages and in which layers of the models this information is prevalent.

We include four state-of-the-art multilingual sentence encoders in our study, namely, LASER, M-BERT, XLM, and XLM-R, that exemplify different architectures and pretraining strategies. We analyze whether these design decisions influence the linguistic organization within these encoders. To investigate how different types of language-specific information interact, we develop a simple and yet novel method to study joint encoding of linguistic information, which we refer to as **cross-neutralizing**. Using this method we test for information sharing between the language-specific subspaces, and hypothesize that these subspaces jointly encode shared properties across typologically similar languages. We test this by investigating to what extent removing language-specific information negatively affects the performance on the 25 language-level probing tasks in typologically related languages.

Contributions. Our findings can be summarized as follows:

- We find that the language-specific components of all encoders successfully capture typological properties related to word order, negation, and pronouns; however, M-BERT and XLM-R outperform LASER and XLM for a number of lexical and morphological properties.
- We find that (1) typological properties are encoded within the language-specific components across layers in M-BERT and XLM-R, but are more localizable in lower layers of LASER and XLM, and (2) the incorporation of a cross-lingual training objective contributes to the model learning an interlingua, while the use of monolingual objectives results in a partitioning to language-specific subspaces. These results indicate that there is a negative correlation between the universality of a model and its ability to retain language-specific information, regardless of architecture.
- The results of our cross-neutralizing experiments show that by localizing and removing information crucial for encoding the typological properties of one language, we are able to remove this same information from the representations of related languages (i.e., that share the same typological feature value). This indicates that the models jointly encode these typological properties across languages.

Article Structure. The research in this article lies on the intersection of work in the fields of multilingual NLP, the interpretation of neural networks, and linguistic typology. Hence, in Section 2 we start off by giving a broad overview of the core methods in each field and discuss some relevant research that has previously attempted to combine methods from these fields. In Section 3 we then introduce the various models studied in this article and explain our probing and neutralizing methods as well as outline how these methods are utilized in our experiments. In Section 4 we first evaluate the reliability of our probing tasks and then continue discussing our results for the typological probing experiments. The results of the second set of experiments pertaining to our information-sharing experiments using the cross-neutralizing methods are then outlined in Section 5. Section 6 concludes.

2. Linguistic Typology and Multilingual NLP

2.1 Linguistic Typology

Linguistic typology is a discipline that aims to study, categorize, and document the variation in the world's languages through systematic cross-linguistic comparisons (Croft 2002). These categories are not set in stone as they emerge inductively from the comparison of languages and are prone to change with the discovery of new languages (Ponti et al. 2019). For instance, one well-established sub-area in linguistic typology is that of word order typology. This branch studies the order of syntactic constituents in a language—for example, they categorize the grammatical structure in languages based on their dominant relative ordering of the Subject, Verb and Object (SVO) in clauses (Dryer 2013). From this it follows that there are 6 dominant orders that can be ascribed to a language, from most to least common: SOV, SVO, VSO, VOS, OVS, and OSV. English, like many other European languages, is grouped under the category SVO languages. For clauses to be grammatically correct in English, the subject should precede the verb, while the object follows:

$$\texttt{SVO:}\ \underbrace{\text{the dog}}_{\texttt{Subject}}\underbrace{\text{chased}}_{\texttt{Verb}}\underbrace{\text{the cat}}_{\texttt{Object}} \tag{1}$$

Although in this particular case, the object and verb can be used interchangeably without resulting in grammatical error, it is evident that this would change the meaning of the clause. On the other hand, many Asian languages (e.g. Urdu, Bengali, Hindi, Japanese, and Korean) dominantly deploy the SOV structure. In English this would translate to:

$$\texttt{SOV:}\ \underbrace{\text{the dog}}_{\texttt{Subject}}\underbrace{\text{the cat}}_{\texttt{Object}}\underbrace{\text{chased}}_{\texttt{Verb}} \tag{2}$$

Likewise, there are many structural language characteristics specified by typological linguistics, at different levels of granularity, that help distinguish and group different languages based on these varying features. Continuing in the line of word order typology, for example, they study correlations between orders in syntactic sub-domains, for example, the order of modifiers (adjectives, numerals, demonstratives, possessives, and adjuncts) in noun phrases and the order of adverbials.

Note, however, that this is an empirical science, as neatly trying to fit a multitude of languages to well-defined categories is an impossible task. There are languages (e.g., Russian) in which multiple relative orderings would technically be accepted as correct; however, one order might be more dominantly used in the language. In other languages, the correct relative order can depend on different parameters. For instance, French is predominantly a SVO language, with the exception that the SOV structure has to be used in the specific case of the object being a pronoun. Different approaches can be taken to handling similar corner cases, such as defining a "No dominant order" category, or simply ascribing the order most prevalent in the language (O'Horan et al. 2016).

2.2 Language Transfer and Multilingual Joint Learning

Language transfer approaches aim to identify and leverage language similarities. This is a complicated task, as these systems need to learn mappings between source and target languages with vastly different structures (Ponti et al. 2018). To leverage useful information from a source language, this information typically needs to be manipulated to better suit the properties of the target language first (Ponti et al. 2019). Different methods have been developed to enable such language transfer, including annotation projection, (de)lexicalized model transfer, and machine translation (Agić et al. 2014; Tiedemann 2015). In annotation projection, for instance, cross-lingual studies have resorted to word-alignment projection techniques to facilitate homogeneous use of treebanks (Hwa et al. 2005; Yarowsky, Ngai, and Wicentowski 2001; Ganchev, Gillenwater, and Taskar 2009; Smith and Eisner 2005). In these studies, word alignments are extracted from parallel corpora such that annotations for the source language can be transferred to the target language accordingly. This automatically annotated data can then be used to train a supervised model. In model transfer, on the other hand, studies attempt to train a model on a source language, delexicalize it to solve for incompatible vocabularies, and then directly apply this model to a target language instead (Zeman and Resnik 2008). This delexicalization has, for instance, been realized by taking language-agnostic (Nivre et al. 2016) or harmonized (Zhang et al. 2012) features as input. In later studies, different augmentation techniques, including multilingual representations, were integrated to better bridge the vocabulary gap (Täckström, McDonald, and Nivre 2013). The last approach is to automatically translate from source to target language, creating synthetic parallel corpora first, and then following the annotation projection paradigm to train a supervised model (Banea et al. 2008; Tiedemann, Agić, and Nivre 2014). These methods, however, rely on the availability of high-quality resources for the source languages, limiting their success to transfer from high resource languages only.

An alternative approach to leverage information from different languages is multilingual joint learning. There are two main techniques through which this is realized, parameter sharing and language vector integration. Parameter sharing is a method, commonly used in multi-task and multimodal learning, used to share certain (otherwise private) representations within a neural network framework—for example, word embeddings (Guo et al. 2016), hidden layers (Duong et al. 2015b), or attention mechanisms (Pappas and Popescu-Belis 2017)—across languages. The sharing can be realized by tying parameters of specific components of the network, for example, by enforcing minimization constraints on the distance between parameters (Duong et al. 2015a) or latent representations (Zhou et al. 2015). Another method is to induce language-specific properties to help guide joint models toward certain languages by using input language vectors (Guo et al. 2016). These are two methods in which the integration of typological information has proven useful in the past, both to guide in selecting which network components to share between which languages and to help construct language vectors (Ponti et al. 2019).

2.3 Applications of Linguistic Typology in Multilingual NLP

Several studies outline how typological information has successfully been integrated in the earlier task-specific multilingual NLP systems (O'Horan et al. 2016; Ponti et al. 2019). For instance, typological constraints have been shown effective in guiding multilingual dependency parsing (Naseem, Barzilay, and Globerson 2012) and part-of-speech (POS) tagging (Naseem, Barzilay, and Globerson 2012; Zhang et al. 2016). Naseem, Barzilay, and Globerson (2012) successfully exploit word order information in multilingual dependency parsing by enabling selective parameter sharing between source and target languages in a multilingual joint learning setting. This sharing mechanism selects source languages based on its aspects that are most relevant to the specific target language. Therefore, in some cases, using this typological information to more carefully select between languages that share similar properties for language transfer allows for more effective applications. Thus, we expect that the state-of-the-art general-purpose models implicitly rely on mechanisms to efficiently share information across typologically different languages as well.

Moreover, despite the success of multilingual encoders, much effort is still focused on improving the language-agnosticism of these models—for example, through methods such as linear projections, adversarial fine-tuning, and re-centering representations (Libovickỳ, Rosa, and Fraser 2020). The intuition behind this is that more universal representations can further boost performance on tasks such as information retrieval, where a search engine only needs to have good semantic understanding of the search query and documents (Zuccon et al. 2015). Because we are only interested in encoding general meaning without ever having to use the linguistic information in a natural language setting, in such cases, signals of cross-lingual structural variation from the source languages may hinder the task (Gerz et al. 2018). Pretrained representations are in practice, however, often used in downstream NLP tasks, such as parsing, named entity recognition (NER), and POS tagging, that require models to pick up on and reconstruct the underlying syntactic and semantic mechanisms of typologically different languages. Thus, pretraining these representations such that other models can deduce the linguistic properties of the source language is likely to improve performance in these tasks. This raises interesting questions as to what extent such models encode language-specific properties, and motivates the study of what typological information is captured in the language-specific components of multilingual models.

2.4 Probing Multilingual Models

Multilingual encoders have been successfully applied to perform zero-shot cross-lingual transfer in downstream NLP tasks, such as POS tagging and NER (van der Heijden, Abnar, and Shutova 2020), dependency and constituency parsing (Tran and Bisazza 2019; Kim, Li, and Lee 2021), text categorization (Nozza, Bianchi, and Hovy 2020), and cross-lingual natural language inference (XNLI) and question answering (XQA) (Lauscher et al. 2020). Interestingly, models trained in unsupervised monolingual tasks (M-BERT, XLM-R) exhibit competitive performance to those that rely on cross-lingual objectives and parallel data (LASER, XLM). Yet, the incorporation of cross-lingual objectives remains a popular approach, with Pires, Schlinger, and Garrette (2019) hinting at their vital role for cross-lingual transfer over divergent languages. Moreover, Huang et al. (2019) introduced Unicoder, which relies on four cross-lingual tasks. Improving on M-BERT and XLM on XNLI and XQA, the authors claim that the tasks help learn language relationships from more perspectives. This raises the question of whether multilingual encoders capture linguistic and typological properties differently, depending on the type of pretraining tasks.

To investigate this, we use techniques from the rapidly growing line of research on interpretation of neural models (Linzen, Dupoux, and Goldberg 2016; Conneau et al. 2018a; Peters et al. 2018a; Tenney et al. 2019), which has recently been extended to the multilingual setting (Chi, Hewitt, and Manning 2020; Pires, Schlinger, and Garrette 2019; Şahin et al. 2020; Ravishankar et al. 2019; Ravishankar, Øvrelid, and Velldal 2019). Ravishankar et al. (2019) and Ravishankar, Øvrelid, and Velldal (2019) study multilingual sentence encoders using probing tasks of Conneau et al. (2018a), for example, probing for universal properties such as sentence length and tree depth, but do not directly test for typological information. In a similar vein, Pires, Schlinger, and Garrette (2019) study how M-BERT generalizes across languages by testing zero-shot cross-lingual transfer in traditional downstream tasks. They only briefly touch on typology by testing generalization across typologically diverse languages in POS tagging and NER, and find that cross-lingual transfer is more effective across similar languages. They ascribe this effect to word-piece overlap, arguing that similar success on distant languages might require a cross-lingual objective. On the contrary, Karthikeyan et al. (2020) show that cross-lingual transfer can also be successful with zero lexical overlap, arguing that M-BERT's cross-lingual effectiveness stems from its ability to recognize language structure and semantics instead. In this work, we take a closer look at these emerging language structures by investigating the language-specific component across sentence representations in the models for typological properties.

2.4.1 Probing for Linguistic Properties. Several papers have already studied language relationships within multilingual models—for instance, by reconstructing phylogenetic trees to analyze preserved relations (e.g., in terms of genetic and structural differences) (Bjerva et al. 2019; Beinborn and Choenni 2020), by probing for typological properties of languages (Qian, Qiu, and Huang 2016; Şahin et al. 2020), or by studying negative interference within the models, that is, cases where competition for model capacity among multiple languages degrades performance on a given language (Wang, Lipton, and Tsvetkov 2020). To the best of our knowledge, our language-level probing tasks come closest to the work of Şahin et al. (2020), who probed non-contextualized multilingual *word* representations for linguistic properties such as case marking, gender system, and grammatical mood. We considerably expand on this work by proposing methods to study the language-specific components learned by multilingual *sentence* encoders and investigating a wider range of typological properties pertaining to lexical, morphological, and syntactic structure. Moreover, because multilingual models are inclined to learn a language identity (Wu and Dredze 2019), we also propose a paired language evaluation set-up, evaluating on languages unseen during training to prevent the model from picking up on this signal.

Working in a monolingual setting, Tenney, Das, and Pavlick (2019) studied how much each layer in BERT contributes to the encoding of linguistic information. This research is inspired by prior work showing that lower layers of a language model capture local syntax, while higher layers tend to capture more complex semantics (Peters et al. 2018a; Blevins, Levy, and Zettlemoyer 2018). Tenney, Das, and Pavlick (2019) show that the same ordering emerges in BERT, and that syntactic information is more localizable within the model, while information related to semantic tasks is scattered across many layers. We take a similar approach to test where in the model typological information is encoded and whether it is localized or is instead spread across layers.

2.4.2 Studying Shared Properties. Our work on joint information-sharing comes closest to that of Chi, Hewitt, and Manning (2020), who study shared grammatical relations in M-BERT. They use a structural probe (Hewitt and Manning 2019) to enable zero-shot transfer across languages to successfully recover syntax. Their results suggest that the probe is able to pick up on features that are jointly encoded in M-BERT across languages. We expand on this work by linking these features to linguistic typology and demonstrating that individual lexical, morphological, and syntactic properties of languages are jointly encoded across all languages that share the property. Thus, we are the first to explore how cross-lingual variation is encoded using typology and explicitly test for the joint encoding of individual properties of languages. We draw inspiration from Gonen et al. (2020) and Libovický, Rosa, and Fraser (2020), who show that M-BERT relies on a language-specific component that is similar across all representations in a language and can thus be approximated by its language centroid. They show that removing the respective centroid drastically decreases performance on language identification, while improving performance on parallel sentence retrieval, indicating stronger language-neutrality. Hence, this method successfully removes language-specific features from model representations, while still encoding the underlying meaning. These results demonstrate the existence of the language-neutral component. In subsequent work, Gonen et al. (2020) successfully decompose the representations into independent language-specific and language-neutral components through nullspace projections, thereby further supporting the existence of identifiable language components. Lastly, Wang, Lipton, and Tsvetkov (2020) investigate which shared model parameters within multilingual encoders are language-specific, using a pruning method to compare parameter similarities across languages. They find that language-specific parameters do exist, and that model parameters are better shared in the lower layers than the higher ones.

3. Methodology

In this section, we detail the methodology applied for our experiments. We provide information on the multilingual models, the data, and the methods for separating the language subspaces and for localizing and removing language-specific information.

3.1 Multilingual Sentence Encoders

LASER is a BiLSTM encoder trained with an encoder-decoder architecture and a **cross-lingual** objective—machine translation (MT). It has $L = 5$ layers with a hidden state size of $H = 512$. The encoder performs max-pooling over the last hidden states to produce sentence representations $v \in \mathbb{R}^{1024}$. The decoder LSTM is initialized with the sentence representations and trained on the task of generating sentences in a target language. Both the encoder and decoder are shared across all languages, and the input sentences are tokenized based on a joint byte-pair encoding (BPE) vocabulary. We use the pretrained model available for 93 languages. This model leverages parallel data from a combination of text corpora from the Opus Web site.[1]

M-BERT is a bidirectional Transformer with $L = 12$ and $H = 768$, trained on the **monolingual** Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks. Apart from being trained on the Wikipedia dumps of multiple languages and using a shared WordPiece vocabulary for tokenization, M-BERT is identical to its monolingual counterpart and does not contain a mechanism to explicitly encourage language-agnostic representations. We use the pretrained Multilingual Cased version that supports 104 languages. To obtain fixed-length sentence representations from the transformers, we mean-pool over hidden states. Note that it is common practice to use the hidden activation of the special `[CLS]` token as a sentence representation for classification tasks after fine-tuning. However, in this work we study the typological properties in the pretrained models, and consequently do not fine-tune the model on a downstream task. Therefore, using the hidden states from the `[CLS]` tokens as sentence representations is less suitable for this approach.

XLM is a bidirectional Transformer with $L = 12$ and $H = 1{,}024$. We use the pretrained version that uses BPE vocabulary, BERT's **monolingual** MLM objective, and introduces a new **cross-lingual** variant on this task, translation language modeling (TLM), to stimulate language-agnostic representations. In TLM two parallel sentences are concatenated and words in both target and source sentence are masked. This allows the model to leverage information from the context in either language to predict the word, thereby encouraging the alignment of representations in both languages. XLM is trained on the 15 XNLI languages only (Conneau et al. 2018b) that do not cover all languages used for probing in our work (see Appendix A). This allows us to test its ability to generalize to languages unseen during pretraining, when probing for typological features. For training on the MLM objective, XLM uses sentences from the Wikipedias of each language; for TLM it leverages parallel sentences from MultiUN, IIT Bombay corpus, the EUbookshop corpus, OpenSubtitles, Tanzil, and GlobalVoices.

Example 1: Feature 47A - Intensifiers and Reflexive Pronouns

Intensifiers and reflexive pronouns are formally *differentiated* in French and German but *identical* in form in Swedish and English. For instance, in English *“himself”* can be used both as a reflexive pronoun: *“He saw **himself** in the mirror.”*, and as an intensifier: *“He **himself** can not make that decision.”*. In German, however, reflexive pronouns and intensifiers would take a different form: *“Er sah **sich** im Spiegel.”*, compared to *“Er **selbst** kann es nicht entscheiden.”*

XLM-R is another encoder with $L = 12$ and $H = 768$, based on a robustly optimized version of BERT in terms of training regime (RoBERTa) (Liu et al. 2019). RoBERTa is trained with vastly more compute power and data retrieved from CommonCrawl, omits the NSP task, and introduces dynamic masking, that is, masked tokens change with training epochs. The XLM-R variant is trained on 100 languages and introduces the use

[1] http://opus.nlpl.eu/.

of a Sentence Piece model (SPM) for tokenization. Unlike XLM, XLM-R does not use the cross-lingual TLM objective, but is only trained on the **monolingual** MLM task.

3.2 Languages and Typological Features

Consisting of 192 linguistic features annotated by typologist experts for 2,679 languages, the World Atlas of Language Structure (WALS) is the largest and most reliable publicly available typological database. In WALS, linguistic features are listed with languages and their corresponding feature values (see Example 1). However, despite its coverage, WALS is relatively sparse as only 100 languages include annotations for all features. This bears the challenge of carefully selecting which languages and features to focus on to ensure enough coverage for each task. For our experiments we manually selected 7 pairs of closely related languages:

(1) (Russian, Ukrainian)

(2) (Danish, Swedish)

(3) (Czech, Polish)

(4) (Portuguese, Spanish)

(5) (Hindi, Marathi)

(6) (Macedonian, Bulgarian)

(7) (Italian, French)

These pairs are typologically diverse and cover four language families: Germanic, Indic, Romance, and Slavic languages. They also include both high- and low-resource languages from the NLP perspective. From each pair, the sentence representations from the first languages (Russian, Danish, etc.) are used for training and the second languages (Ukrainian, Swedish, etc.) for testing. This way, we prevent the classifier from leveraging information by falling back to a language identification task. At the same time, by choosing related languages, we can ensure that similar typological properties are captured in both the training and test set. Note that, except for XLM, the encoders support all languages used.

For the typological features, we selected WALS features containing annotations for at least four of our languages and discarded features for which the chosen languages did not show typological diversity, that is, there was zero variation in feature values across languages. Moreover, we made sure that all feature values were covered by the 15 languages that XLM is trained on. As a result, we test for 25 features classified by WALS under the categories: Word order (WO), Nominal (Nom) and Verbal (Verb) categories, and Simple clauses (SC), each in a separate task (see Table 1; for detailed descriptions of the features the reader is referred to `https://wals.info/`).

3.3 Language-Level Probing Tasks to Test for Typological Properties

We develop 25 tasks to test for the typological information captured by the language-specific components of the models. By training a simple classifier to separate the language-specific subspaces within encoders based on specific typological properties, we can test whether the encoder (perhaps implicitly) relies on a similar type of linguistic typology to structure language relationships within its shared multilingual space. Given a set of input sentences per language, a dataset for each of the 25 tasks is created by annotating all sentences from a language with their corresponding feature value in WALS (see Table 2 for a task example). Hence, the annotations are at the language-level as we aim to test properties of *languages* as opposed to properties of *sentences*.

Table 1
The 25 WALS features used for probing along with their correpsonding WALS codes and categories. The multilingual sentence representations for each of these features are probed for in separate tasks. Unless indicated otherwise, all language pairs were covered. Excluded pairs: *:(1), †:(1, 3, and 6), ‡:(6 and 7), §:(2, 4, 5, and 7), $^{|}$:(5 and 6), ¶:(1, 4, 6, 7), $^{\#}$:(1–3 and 6), $^{\Diamond}$:(7), $^{\downarrow}$:(3, 5 and 7), $^{\delta}$:(5 and 7).

Code	CAT	Feature name	Example of feature value(s)
37A	Nom	Definite articles	Definite article distinct from demonstrative words (e.g., English: *the* vs. *this/that*)
38A*	Nom	Indefinite articles	Indefinite word distinct from numeral for "one"
45A†	Nom	Politeness distinctions in pronouns	None (English: *you*) or binary distinction (German *du* informal, *Sie* polite)
47A†	Nom	Intensifiers and reflexive pronouns	Identical (Eng: "himself") vs differentiated (Deu: sich vs selbst)
51A‡	Nom	Position of case affixes	e.g., case suffixes or case prefixes
70A	Verb	The morphological imperative	Special marking for 2nd singular and plural
71A	Verb	The prohibitive	2nd singular imperative + negative declarative
72A	Verb	Imperative-hortative systems	Neither type of systems
79A§	Verb	Suppletion according to tense and aspect	Suppletion according to tense (e.g., English: *go* vs. *went*)
79B§	Verb	Suppletion in imperatives and hortatives	A regular and a suppletive form alternate
81A	WO	Order of Subject, Object and Verb (SOV)	SOV, SVO, VSO, VOS, OVS or OSV
82A	WO	Order of Subject and Verb (SV)	SV or VS
83A	WO	Order of Object and Verb (OV)	OV or VO
85A	WO	Order of adposition and noun phrase	Adp-NP or NP-Adp
86A†	WO	Order of genitive and noun	Genitive-Noun or Noun-genetive
87A	WO	Order of adjective and noun	Adj-Noun or Noun-Adj
92A$^{|}$	WO	Position of polar question particles	Question particle at the beginning or end of sentence
93A¶	WO	Position of interrogative phrases in content questions	Interrogative phrases obligatorily at the beginning of sentence
95A	WO	Relationship between the OV order and the adposition and noun phrase order	OV or VO + postpositional or prepositional
97A	WO	Relationship between OV and adjective and noun order	OV or VO + Adj-Noun or Noun-Adj
115A$^{\#}$	SC	Negative indefinite pronouns and predicate negation	Negative indefinite pronouns (e.g., *"nobody"*) co-occurs with marker of predicate negation
116A$^{\Diamond}$	SC	Polar questions	Formed using question particle
143F	WO	Postverbal negative morphemes	Postverbal negative word or negative suffix
144D$^{\downarrow}$	WO	Position of negative morphemes in SVO languages	NegSVO, SNegVO, SVONeg etc.
144J$^{\delta}$	WO	Order of Subject, Verb, Negative word, and Object (SVNegO)	Separate word, no double negation

Table 2
Task example of feature 86A: *Order of Genitive and Noun*. Labels are Genitive-Noun (GN), Noun-Genitive (NG), and No Dominant Order (NDO).

Languages (ISO 639-1)	**GN**	**NG**	**NDO**
da, hi, sv, mar	☒		
cs, mk, bg			☒
pt, it, pl, es, fr		☒	

Following Pires, Schlinger, and Garrette (2019), Gonen et al. (2020), and Libovickỳ, Rosa, and Fraser (2020), we hypothesize that sentence representations from our encoders contain a language-specific component that remains constant across all sentences in a language. In the design of the language-level probing tasks we implicitly rely on this assumption by testing whether the classifier, on average, is able to correctly predict the typological feature from a large number of sentence representations in a language. The intuition behind this is that if typological information is present in the language-specific component, this information should be encoded irrespective of whether the property we test for is present in the sentence or not.

This (1) eliminates the concern for unfair distributional skews of the linguistic phenomena in our datasets, and (2) if the overall accuracy per language is close to either 100% or 0%, indicating similar performance across sentences that do and do not contain the property of interest, it demonstrates that the information indeed stems from a component that is constant across sentence representations from a language. In addition, to ensure that the classifier categorizes languages based on their typological profile instead of overfitting on the sentence meaning (e.g., perhaps data from some languages contain many similar sentences), we filtered out translations between all train and test languages.

Per language, we extract 10,000 random sentences from the Tatoeba corpora (available at: `https://tatoeba.org`), and attempt to predict the typological features from each of the 10,000 sentence representations. Table 3 provides an indication of the variation of feature values represented in our dataset. Note that paired languages do not always have the same value for the same typological feature, thus the respective tasks would not be possible to solve by falling back to a similar language identification task.

Classifier. We use an MLP with one hidden layer of 100 units, ReLU activation, and an output layer that uses the softmax function to predict class labels. The simplicity of the architecture was chosen to limit task-specific training, such that the classifier is forced to rely on information contained in the encoder representations as much as possible. We experimented with various similar architectures and hyperparameter values, but no prominent differences were observed.[2] We freeze the parameters of the sentence encoder during training such that all learning can be ascribed to the classifier P_τ. The classifier then predicts the feature values y_τ from the representations of the input

[2] Other works used more expressive models, e.g., 300 hidden units (Şahin et al. 2020) and two-layer MLPs (Tenney et al. 2019). This did not yield substantial changes in our experiments, and results using a linear classifier were slightly lower. We report results from the least expressive non-linear model tested, as high performance of M-BERT indicates that this model is in principle capable of learning the task, given an informed encoder.

Table 3
Color coding of the typological diversity of the train and test languages with respect to different features. Languages with the same color have the same feature value for that task (excluded languages are left blank).

	Train							Test						
	ru	da	cs	pt	hi	mk	it	uk	sv	pl	es	mr	bg	fr
37A	░	▓	░	▒	░	▒	▒	░	▓	░	▒	░	▒	▒
38A	░	▓	░	▒	░		▒	░	▒	░	▒	░		▒
51A	▒	▒	▒	░	▒			▒	░	▒	░	▒		
70A	▒	░	▒	▓	▒	▒	▓	▒	░	▒	▒	░	▒	▓
71A	▒	▒	▒	░	█	▓	░	▒	▒	▒	░	▓	▓	▒
72A	▒	▒	▒	▒	▒	░	▒	▒	▒	▒	▒	░	░	▒
79A	▒		▒			░		▒		▒			░	
79B	░		▒			▓		▒		▒			▓	
82A	▒	▒	▒	▒	▒	▒	░	▒	▒	░	░	▒	░	▒
86A		░	▓	▒	░	▓	▒		░	▒	▒	░	▓	▒
92A	▓	▒	▒	░			▒	░	▒	░	▒			░
93A	▒	▒	░		░			▒	▒	▒		░		
116A	▒	░	░	▒	▒	▒		▒	░	▒	░	▒	▒	
143F	▒	░	▒	▒	▒	▒	▒	▒	░	▒	▒	░	▒	░

sentences. Note that the number of sentences in the datasets depend on the number of language pairs n included in the task. For each language we have 10,000 sentences, thus, given n language pairs we use $n \times 10,000$ sentences for training. We hold out 10% of our test set for validation ($n \times 1,000$) to avoid overfitting on the train languages, and use the remaining $n \times 9,000$ sentences for testing. For all tasks we train for 20 epochs with early stopping (patience = 5), using the Adam optimizer (Kingma and Ba 2014). We set the batch size to 32 and use dropout (rate = 0.5). As some features can take $c > 2$ values, we encode the labels as one-hot vectors and obtain the non-binary predictions at test time by returning the class with the highest probability. To keep results across different tasks comparable, we perform no additional fine-tuning on the hyperparameters.

Evaluation. We report results from the language-level probing tasks using macro-averaged-F1 scores as our tasks contain class imbalances, where often only a few languages are annotated with a rare class label. Instead of smoothing these class imbalances out, we assign all classes with an equal weight as we are especially interested in these minority class predictions. Thus, this is a stricter metric for our tasks than micro-averaged-F1 scores, where majority class voting as a baseline could obtain a much higher performance on most tasks. When reporting on the performance for individual languages, we use accuracy (%) for evaluation.

3.4 Testing Across Layers

The methods described next are used to test for typological information at different layers of the models.

Individual Layers. In the first set of experiments, we will separate the language subspaces using the sentence representations from the top layer $H^{(L)}$ of the model, as these are commonly used in downstream tasks. However, each model produces a set of activations at each layer: $H^{(0)}, H^{(1)}, .., H^{(L)}$, where $H^{(L)} = [\mathbf{h}_{t_0}^{(L)}, \ldots, \mathbf{h}_{t_k}^{(L)}]$ and k is the number of tokens. To test where in the model these language-specific properties emerge and whether they evolve throughout the layers, we test sentence representations from each layer of the model. We compute per-layer sentence representations by mean-pooling over the corresponding hidden states.

Full Models. As the layer-wise approach does not take into account the interactions between different layers, we also adapt the method proposed by Tenney, Das, and Pavlick (2019) that borrows the scalar mixing technique from ELMo (Peters et al. 2018b). For each task we introduce a set of scalar parameters: λ_τ and $a_\tau^{(1)}, a_\tau^{(2)}, .., a_\tau^{(L)}$. We compute per-layer sentence representations by mean-pooling over hidden states $t_0, .., t_k$ as before:

$$\mathbf{h}_\tau^{(l)} = \text{pool}([\mathbf{h}_{t_0}^{(l)}, \mathbf{h}_{t_1}^{(l)}, .., \mathbf{h}_{t_k}^{(l)}]), \text{where } \mathbf{h}_{t_i}^{(l)} = \overrightarrow{\mathbf{h}_{t_i}^{(l)}} + \overleftarrow{\mathbf{h}_{t_i}^{(l)}} \text{ for LASER} \tag{3}$$

To pool across layers we use the mixing weights:

$$\mathbf{h}_\tau = \lambda_\tau \sum_{l=1}^{L} s_\tau^{(l)} \mathbf{h}_\tau^{(l)} \tag{4}$$

where $s_\tau = \text{softmax}(a_\tau)$. These weights a_τ are jointly learned with each task to give the probing classifier P_τ access to the full model. Note that we exclude layer 0, as token embeddings in LASER have a different dimensionality from higher layers. After training, we extract the learned coefficients from the classifier to estimate the contribution of different layers to the particular task. Higher weights are interpreted as evidence that the corresponding layer contains more information about the typological property. We report the Kullback-Leibler divergence $K(s_\tau) = KL(s_\tau || \text{Uniform})$:

$$KL(p||q) = \sum_{i=0}^{N} p(x_i) \log(\frac{p(x_i)}{q(x_i)}) \tag{5}$$

for each task as an estimation of the non-uniformity of the statistics. We interpret a higher KL divergence as an indication of a more localizable feature.

3.5 Locating and Removing Language-Specific Information

The following two methods are used to investigate whether the language-specific components jointly encode the typological properties of related languages in Section 5.

Restructuring the Vector Space. To explicitly localize the language-specific components, we use the neutralization method from Libovický, Rosa, and Fraser (2020). We approximate the language centroid for each language in our test set $x \in$ L, by obtaining a mean language vector $\bar{\mathbf{u}}_\mathbf{x} \in \mathbb{R}^m$ from a set of S sentence representations $\{u_1, u_2.., u_S\} \in \mathbb{R}^m$ from that language (10,000 in our case):

$$\bar{\mathbf{u}}_\mathbf{x} = mean(u_1, u_2.., u_S) \tag{6}$$

The idea is that by localizing language-specific information through averaging representations, core linguistic properties remain prominent in the centroid. Simultaneously, infrequent phenomena that vary depending on sentence meaning are averaged out. We then obtain a set of language-neutral representations $v_i \in \mathbb{R}^m$ for a language i by subtracting the corresponding language centroid from the model representation u_j for a sentence j:

$$v_{ij} = u_{ij} - \bar{\mathbf{u}}_\mathbf{x}, \text{ where } i = x \tag{7}$$

This means that we remove language-specific information by re-structuring the vector space such that the average of the representations for each language is centered at the origin of the vector space. From now on we refer to this method as **self-neutralizing**. Note that we do not conduct these experiments on LASER and XLM as this method was not created for these models.

Testing for Joint Encoding. To investigate how typological properties are shared, that is, whether they are jointly encoded across languages in a localizable manner or rather in independent ways for each language, we adapt the self-neutralizing method to a cross-neutralizing scenario. Specifically, we approximate typological information from only one language (x) by computing $\bar{\mathbf{u}}_\mathbf{x}$, and subtract $\bar{\mathbf{u}}_\mathbf{x}$ from the representations of all other languages in $\mathrm{L} \setminus \{x\}$:

$$v_{ij} = u_{ij} - \bar{\mathbf{u}}_\mathbf{x}, \text{ for } i \in \mathrm{L} \setminus \{x\} \tag{8}$$

We refer to this method as **cross-neutralizing**. Each time we select a different language x for cross-neutralizing. We then test the trained classifiers on the 25 language-level probing tasks using the neutralized representations of each language $l \in \mathrm{L}$. The intuition behind this is that if the encoders were to represent languages and their properties in independent ways, we expect the performance to deteriorate only for the language x that we use for cross-neutralizing, that is, the language used to compute $\bar{\mathbf{u}}_\mathbf{x}$. In case of joint encoding of typological properties, however, we expect to see that performance (1) also deteriorates for other languages that share the same typological feature value with x (i.e., related languages) and (2) remains intact for languages that do not share the same feature value with x. Note that, by training on the unmodified representations and testing on the cross-neutralized representations, we can analyze whether our method removes crucial information that the classifier relied on. If we were to both train and test on the modified representations instead, this would only tell us whether the classifier is able to correct for the missing information.

4. Testing for Typological Information

In this section we provide results for the first set of experiments with which we study what typological information is captured in the language-specific components and where in the respective encoders this information emerges (outlined in Sections 3.3 and

Table 4
Macro-averaged-F1 scores on the test set per typological feature. Unless indicated otherwise, all language pairs were used. Excluded pairs: *:(1), †:(1, 3, and 6), ‡:(6 and 7), §:(2, 4, 5, and 7), $^{|}$:(5 and 6), ¶:(1, 4, 6, 7), $^{\#}$:(1–3 and 6), $^{\diamond}$:(7), $^{\downarrow}$:(3, 5, and 7), $^{\delta}$:(5 and 7). Highest performance per feature are **bolded**.

Code	Type	LASER	M-BERT	XLM	XLM-R	Baseline
37A	Nom	0.864	0.957	0.83	**0.997**	0.199
38A*	Nom	0.571	**0.597**	0.595	0.579	0.334
45A†	Nom	0.997	**1.0**	0.989	**1.0**	0.428
47A†	Nom	0.97	0.995	0.934	**0.999**	0.333
51A‡	Nom	0.682	**0.763**	0.752	0.762	0.375
70A	Verb	0.64	0.69	0.603	**0.695**	0.243
71A	Verb	0.347	0.522	0.452	**0.576**	0.243
72A	Verb	0.422	0.763	0.557	**0.769**	0.417
79A§	Verb	0.456	0.94	0.646	**0.978**	0.4
79B§	Verb	0.212	0.528	0.382	**0.544**	0.25
81A	WO	0.993	**1.0**	0.959	0.998	0.462
82A	WO	0.429	0.352	**0.449**	0.368	0.363
83A	WO	0.993	**1.0**	0.939	0.999	0.462
85A	WO	0.993	**1.0**	0.873	0.995	0.462
86A†	WO	0.763	0.811	0.757	**0.82**	0.166
87A	WO	0.976	**0.999**	0.944	0.998	0.416
92A$^{|}$	WO	0.212	0.16	0.231	0.206	**0.285**
93A¶	WO	0.647	0.65	0.627	**0.665**	0.25
95A	WO	0.993	**1.0**	0.96	0.999	0.462
97A	WO	0.983	0.996	0.941	**0.998**	0.243
115A$^{\#}$	SC	0.998	**1.0**	0.984	0.999	0.4
116A$^{\diamond}$	SC	0.584	0.622	0.602	**0.634**	0.4
143F	WO	0.608	0.644	0.599	**0.65**	0.364
144D$^{\downarrow}$	WO	0.978	0.998	0.979	**1.0**	0.429
144J$^{\delta}$	WO	0.983	0.996	0.954	**0.999**	0.445

3.4, respectively). First we discuss results from the top-layer representations produced by our encoders, which are commonly used in downstream tasks. Then we test and analyze the representations across all layers of the models.

4.1 Top-Layer Sentence Representations

Baseline. To test to what extent the classifier relies on information from the encoder as opposed to information learned from task-specific training, we use randomized encoders as a baseline for comparison. Following Tenney et al. (2019), we randomized the weight matrices of our pretrained models. We found that our simple classifier is unable to learn from these representations, falling back to majority class voting in all cases. Thus, the performance for all randomized encoders is identical and we report these scores under *Baseline* in Table 4.

Results. In Table 4, we report the performance over all languages per task. Note that, due to missing values, not all languages were used for each task, as indicated in the table. The results show that within all encoders we are able to separate languages based on features related to word order (e.g., 81A, 85A, 95A, and 97A), pronouns (45A, 47A), and negation (144D, 144J). For M-BERT and XLM-R, however, the classifier generally outperforms LASER and XLM when it comes to separating languages based on lexical and morphological properties, such as in the nominal (e.g., 37A, 51A) and verbal (e.g., 70-72A) category tasks. The strongest difference between encoders is observed when testing for the suppletion features (79A,B). Furthermore, for none of the encoders is the classifier capable of accurately predicting properties related to the form of questions (92A, 93A, 116A). Lastly, we find that, while obtaining a high performance for other word order tasks, the classifier fails to predict the feature 82A (*SV order*).

Evaluating the Task Set-up. To further analyze our models, we investigated the accuracy per feature broken down by language and specific feature values. Overall, we find that the classifier consistently fails to predict certain features for specific languages, as expected; this results in the per-language performance often being either very high or low (see Figure 1). This demonstrates that the classifier indeed relies on the language-specific component of the representations to capture the typological properties of languages.

Moreover, we observe that the classifier may fail both in cases where the labels for the paired test and train languages are identical and in cases where they are different. For instance, despite Bulgarian having the same label as Macedonian, the classifier based on LASER or XLM fails for Bulgarian in multiple tasks (e.g., 71A, 72A, 116A).[3] On the other hand, there are also cases where the classifier succeeds despite the test language and its most similar training language having a different label—for example, LASER for Spanish (92A), XLM for French (71A), and both LASER and M-BERT for French (82A). This demonstrates that the classifier does not merely rely on similar language identification either (see Appendix C for further analysis).

As a last test for the validity of the language-level probing tasks, we repeated experiments for properties specific to questions on a subset of our data, where only questions were used as input sentences. This resulted in a subset of $\approx 10\%$ of the full dataset per language and we obtained similar classifier performance for the features of interest across encoders, again confirming the hypothesis that the language-specific components under investigation remain constant across sentences with varying meaning in a language.

Languages and Feature Values. From Figure 1 we also see that no languages or language families were found for which an encoder always fails. Instead, low performance tends to be associated with specific features. In addition, XLM obtains performance levels similar to LASER for languages it was not pretrained on. In fact, we found no relationship between the support of language and performance, indicating that XLM successfully generalizes to unseen languages (see Appendix F). When comparing the per-language performance across encoders, we see that, although LASER and XLM exhibit a lower performance in more languages, there are specific cases in which all encoders fail. Consequently, encoders might not benefit from encoding information about these properties in their language-specific components. Such cases include, for

[3] Note that all encoders were trained on Bulgarian.

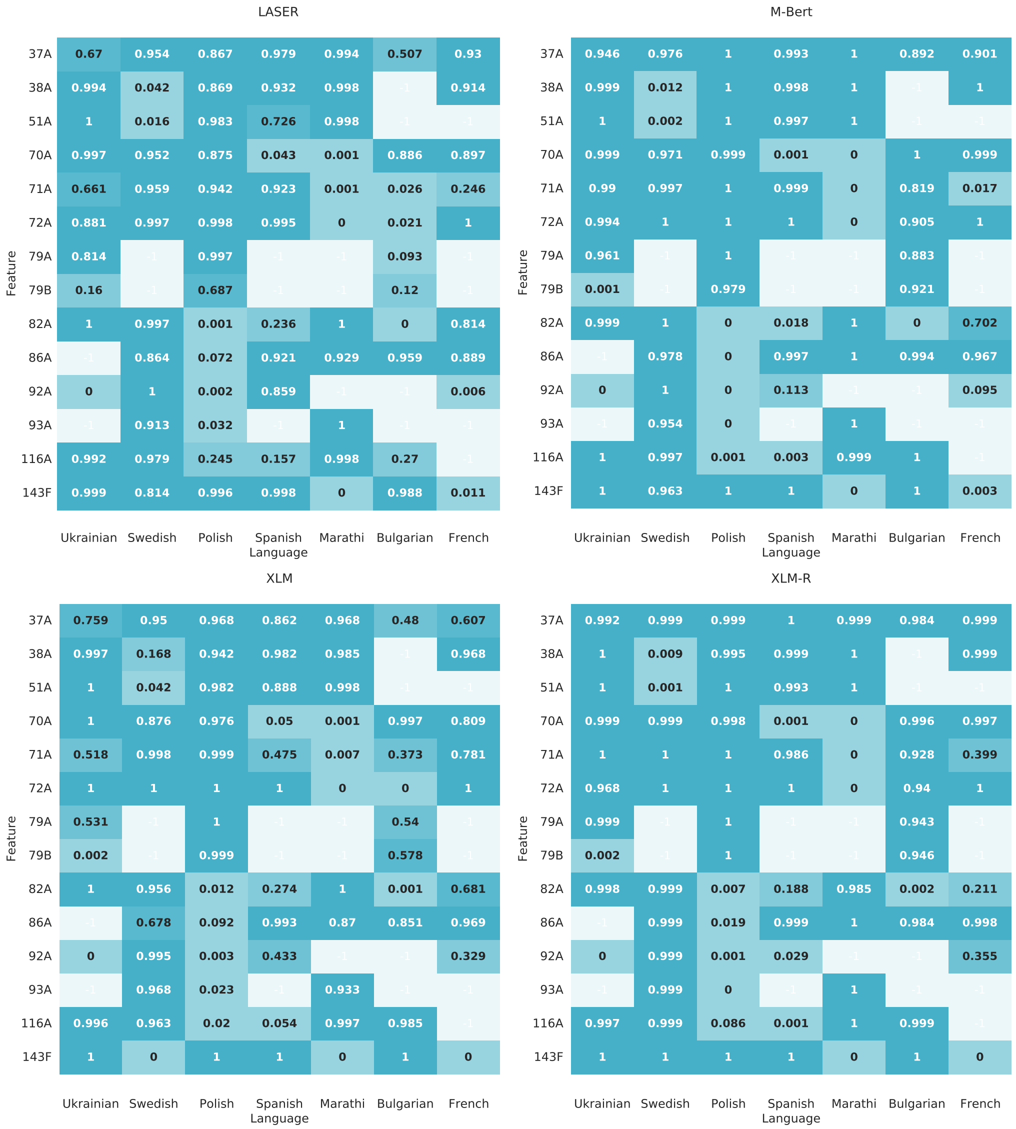

Figure 1
Heatmaps of the performance in (%) accuracy for a selected number of interesting tasks from all four multilingual encoders broken down per language.

instance, features *Indefinite articles* (`38A`) and *Postverbal negative morpheme* (`143F`) for Swedish and French, respectively.

In addition, we analyzed specific feature values for which we cannot separate languages within the encoders. For example, we find that for LASER and XLM the classifier fails to predict the label *Maximal system* (assigned to Bulgarian and Marathi) for feature `72A` (*Imperative-hortative systems*). M-BERT and XLM-R, while failing for Marathi, obtain ±90% accuracy for Bulgarian. A similar effect for LASER is observed for other labels of verbal and nominal category tasks (e.g., *Tense* [`79A`: *Suppletion according to tense and aspect*]), and to a lesser extent also XLM (e.g., *Special imperative + special negative* [`71A`:

The prohibitive]). No such cases were identified for M-BERT or XLM-R. This observation clarifies our finding that M-BERT and XLM-R outperform LASER and XLM on the majority of nominal and verbal category tasks. Whereas LASER and XLM omit both certain feature values as well as specific language-feature combinations, M-BERT and XLM-R only discard some of the latter.

In the particular case of feature `82A` (*SV order*), the classifier always fails to predict *No dominant order*, assigned to Bulgarian, Spanish, and Polish, for all encoders. As explained in Section 2.1, this label is sometimes assigned by linguistics experts as an alternative to categorizing by the predominant order in corner cases. We speculate that the encoders are instead inclined to categorize languages by the order predominantly seen during training, without quantifying an extent, thereby forcing an order to non-dominant order languages.

Similarly, in the few cases in which ±50% accuracy is obtained, LASER specifically seems to omit the encoding of a lack of certain properties—for example, Ukrainian: *No definite or indefinite article* (`37A`), Spanish: *No case affixes or adpositional clitics* (`51A`), Polish: *No suppletive imperatives* (`79B`). Since the performance is ±50%, this information is likely not coming from the language-specific component that remains constant across representations. Hence LASER seems to organize its language-specific subspaces based on the properties that are present in the language instead.

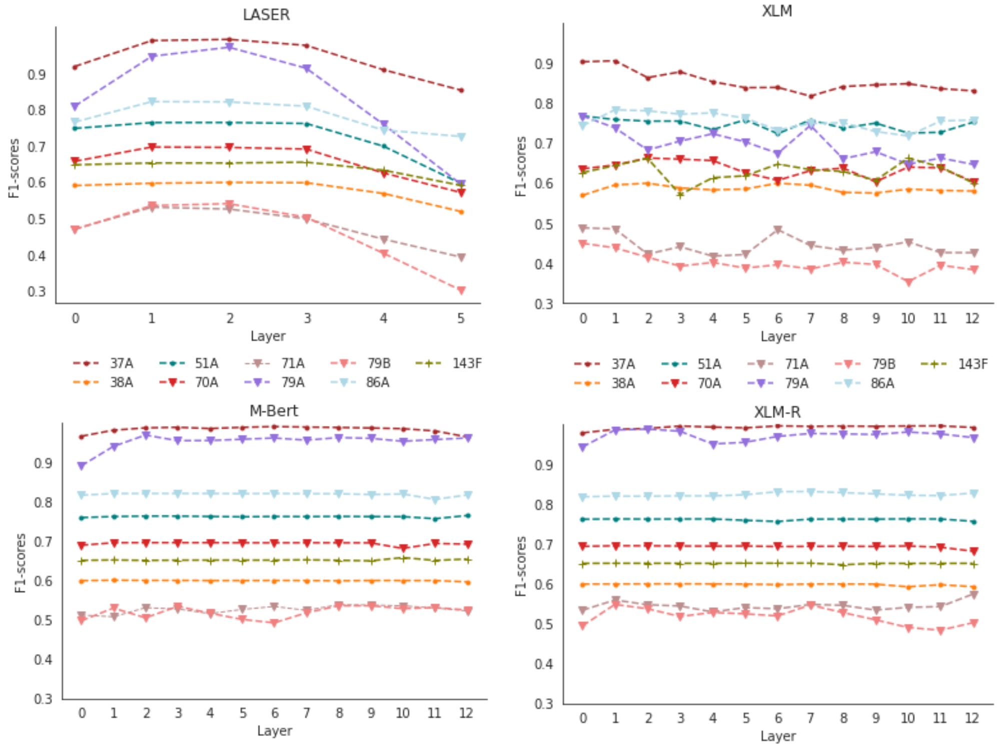

Figure 2
Macro-averaged F1-scores on probing tasks, when probing from the activation of different layers in LASER, M-BERT, XLM, and XLM-R. Layer 0 corresponds to the non-contextualized token embeddings used by these models.

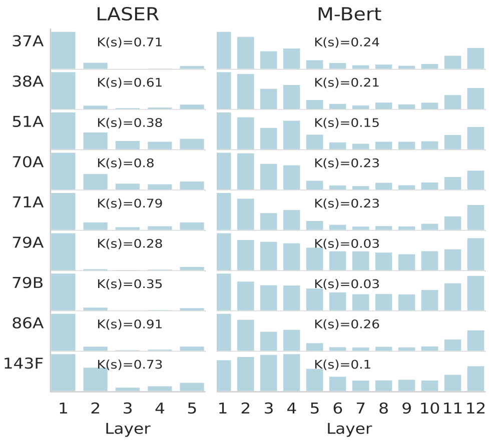

Figure 3
Learned mixing weights s_τ and corresponding KL divergences $K(s)$ for LASER and M-BERT.

4.2 Typological Information Across Layers

Results. Figure 2 shows the classifier performance when testing the different layers of the models. We find that in both models that incorporate a cross-lingual objective (LASER and XLM), the typological properties are more prevalent in lower layers of the network (1–2) and performance steadily decreases in higher layers (3+). In contrast, in M-BERT and XLM-R the performance is stable throughout all layers. This indicates that while M-BERT and XLM-R rely on language-specific components that remain constant across all layers, these components evolve throughout the layers of LASER and XLM, meaning that the latter models start to restructure their linguistic organization.

Figure 3 shows the distribution of the learned mixing weights across layers (see Appendix D for XLM and XLM-R). We find that for LASER and XLM the classifiers almost exclusively rely on information from the first layers, which is in line with our findings from the per-layer results. Given the low KL divergences across tasks, the learned weights remain more uniform for M-BERT and XLM-R. Nevertheless, we observe a trend that middle layers gradually decrease in importance, while the last few layers regain it again.

These results indicate that in models pretrained with a cross-lingual objective—LASER and XLM—typological information is localizable in the lower layers, but is lost in higher layers. For M-BERT and XLM-R, which rely on monolingual pretraining objectives, the results remain somewhat inconclusive, as we interpret the lower KL divergences as an indication of less localizable features. Thus, this information is either captured in the lower layers and correctly propagated through the higher layers, or it could be spread across the model instead.

Universality vs. Language-Specific Information. Previous research suggests that M-BERT partitions its multilingual semantic space into separate language-specific subspaces, and is thus not a true interlingua (Libovický, Rosa, and Fraser 2020; Singh et al. 2019). In Figure 4 we visualize the representations of all sentences in our test datasets from

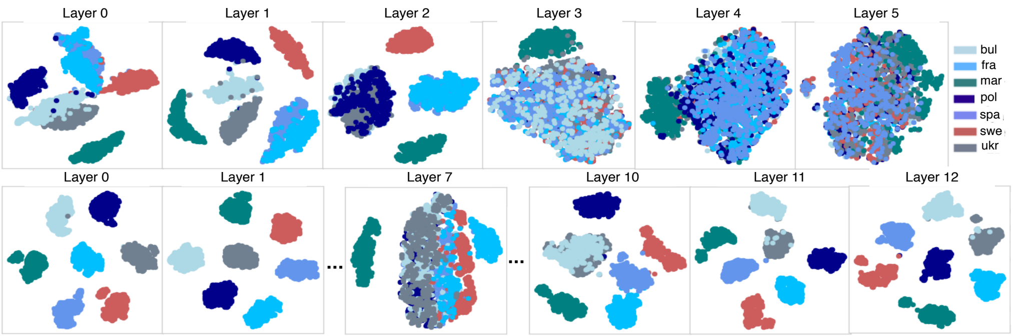

Figure 4
t-SNE plots of representations from layers of LASER (top) and M-BERT (bottom), where layer 0 corresponds to the non-contextualized token embeddings.

layers in LASER and M-BERT in a t-SNE plot.[4] In agreement with previous research, we find that in M-BERT and XLM-R, languages continue to occupy separate subspaces in the last layer (see Appendix E for XLM and XLM-R plots, which exhibit similar trends to LASER and M-BERT, respectively). Initially, LASER and XLM also appear to create a continuous language space by representing language relationships in terms of geometric distance between subspaces. However, these initial subspaces become increasingly more clustered throughout the layers, thereby creating a common, shared, interlingual space in the higher layers. Consequently, there appears to be a connection between the loss of typological information and the creation of more language-agnostic representations. Universality of LASER and XLM seems to come at the cost of retaining language-specific information.

It should be noted that all encoders at some point cluster languages by family; however, M-BERT and XLM-R recover from this (at layer 10) by projecting languages back to separate subspaces. Moreover, XLM-R does not appear to organize its space differently from M-BERT and only improves on the performance patterns also seen in M-BERT. This indicates that XLM-R simply refines the mechanism deployed by M-BERT.

Pretraining Objectives. LASER and XLM retain typological properties in higher layers to a lesser extent. Given that higher layers of a model are more tuned toward the pretraining objective, we speculate that this effect can be ascribed to their differences in the type of pretraining: LASER and XLM are trained with a cross-lingual objective vs. M-BERT and XLM-R trained on monolingual tasks only. In MT, the encoder needs to capture semantic meaning, while the decoder is responsible for reconstructing that meaning in a target language. While the decoder might benefit from typological information about the target language, the encoder has no incentive from the decoder to capture such properties of the source language. Similarly, in TLM, the model can leverage information from both languages and is explicitly stimulated to align patterns from them. On the contrary, for monolingual tasks, the model must know which language it is encoding

[4] We use the sklearn TSNE visualizer using the default parameters. Because TSNE is an expensive procedure, PCA is first applied as a simpler dimension reduction technique (as recommended in the documentation).

Table 5
The table shows the mean and standard deviation of the performance in (%) accuracy computed across languages in the test set. We report the results obtained before any neutralization and after self-neutralizing each language.

	M-BERT		XLM-R	
τ	before	after	before	after
37A	0.96 ± 0.04	0.4 ± 0.07	1.0 ± 0.01	0.41 ± 0.06
38A	0.83 ± 0.37	0.37 ± 0.03	0.83 ± 0.37	0.39 ± 0.03
45A	1.0 ± 0.0	0.58 ± 0.11	1.0 ± 0.0	0.54 ± 0.04
47A	1.0 ± 0.001	0.51 ± 0.14	1.0 ± 0.0	0.5 ± 0.04
51A	0.8 ± 0.4	0.53 ± 0.11	0.8 ± 0.4	0.5 ± 0.04
70A	0.71 ± 0.45	0.34 ± 0.09	0.71 ± 0.45	0.38 ± 0.11
71A	0.69 ± 0.43	0.36 ± 0.1	0.76 ± 0.37	0.35 ± 0.08
72A	0.84 ± 0.35	0.56 ± 0.09	0.84 ± 0.35	0.58 ± 0.09
79A	0.95 ± 0.05	0.54 ± 0.04	0.98 ± 0.03	0.55 ± 0.03
79B	0.63 ± 0.45	0.33 ± 0.07	0.65 ± 0.46	0.38 ± 0.1
81A	1.0 ± 0.0	0.57 ± 0.06	1.0 ± 0.01	0.53 ± 0.04
82A	0.53 ± 0.47	0.53 ± 0.02	0.48 ± 0.45	0.5 ± 0.08
83A	1.0 ± 0.0	0.58 ± 0.07	1.0 ± 0.0	0.5 ± 0.02
85A	1.0 ± 0.0	0.63 ± 0.11	1.0 ± 0.01	0.52 ± 0.02
86A	0.82 ± 0.37	0.35 ± 0.04	0.83 ± 0.36	0.36 ± 0.01
87A	1.0 ± 0.0	0.54 ± 0.06	1.0 ± 0.0	0.52 ± 0.03
92A	0.24 ± 0.38	0.37 ± 0.02	0.28 ± 0.39	0.36 ± 0.05
93A	0.65 ± 0.46	0.42 ± 0.03	0.67 ± 0.47	0.48 ± 0.03
95A	1.0 ± 0.0	0.54 ± 0.03	1.0 ± 0.0	0.5 ± 0.01
97A	1.0 ± 0.01	0.39 ± 0.05	1.0 ± 0.0	0.4 ± 0.08
115A	1.0 ± 0.0	0.53 ± 0.05	1.0 ± 0.0	0.52 ± 0.03
116A	0.67 ± 0.47	0.5 ± 0.03	0.68 ± 0.45	0.5 ± 0.03
143F	0.71 ± 0.45	0.52 ± 0.14	0.71 ± 0.45	0.52 ± 0.07
144D	1.0 ± 0.0	0.5 ± 0.02	1.0 ± 0.0	0.52 ± 0.03
144J	1.0 ± 0.0	0.54 ± 0.04	1.0 ± 0.0	0.54 ± 0.06

to succeed (e.g., to avoid predicting a French word for a Spanish sentence during MLM). This objective provides the model with a better incentive to retain typological properties in higher layers, as useful information can be leveraged from them to complete the tasks. Hence, cross-lingual objectives appear more suitable for training language-agnostic models. Moreover, it might not be reasonable to expect M-BERT and XLM-R to yield language-neutral representations, as their pretraining objectives do not stimulate them to learn an interlingua. This, in turn, poses challenges in zero-shot transfer on distant languages (Pires, Schlinger, and Garrette 2019) and in resource-lean scenarios (Lauscher et al. 2020).

5. Testing for Information-Sharing

We now examine the interaction of linguistic typological information within M-BERT and XLM-R, using the 25 language-level probing tasks and the neutralization methods that we proposed in Section 3.5.

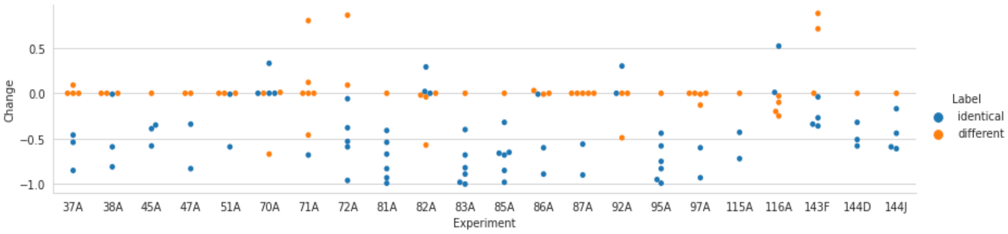

Figure 5
Change in performance for all test languages when cross-neutralizing with Spanish. Languages are categorized by an identical (blue) or different (orange) feature value from Spanish for the respective task.

5.1 Self-Neutralizing

First, we test whether the approximated language centroids $\mathbf{\bar{u}_x}$ successfully capture the typological properties of the language. We do this by testing whether self-neutralizing results in a substantial loss of information about the typological properties of the languages in our test set. We evaluate the change in performance before and after applying this method in Table 5. We observe that self-neutralizing decreases performance to chance accuracy for each language. This shows that the method successfully removes crucial typological information from the encodings.[5] Moreover, the language identity, approximated by the language centroid, is crucial for the encoding of typological properties, suggesting that typological information is largely encoded in the relative positioning of the language-specific subspaces of our models.

5.2 Cross-Neutralizing

Having confirmed that computing $\mathbf{\bar{u}_x}$ is a viable method to localize the typological properties of a language x, we apply our cross-neutralizing method. From the results, we see that depending on the language we cross-neutralize with (i.e., language x from which we compute $\mathbf{\bar{u}_x}$): (1) performance on a different set of languages is affected, and (2) this set of languages varies per task. Upon further inspection, we observe that the affected languages tend to share the same feature value as x for the respective task. Figure 5 shows the change in performance on all test languages when cross-neutralized with Spanish (see Appendix G for cross-neutralization with other languages). We categorize these languages based on whether their feature value is the same (blue) or different (orange) from the feature value of Spanish in the respective task. We indeed see that the performance on the set of languages that have the same feature value tend to deteriorate, while the performance on languages with a different feature value remains mostly constant.

Moreover, when the classifier predicts the incorrect feature value for language x, we find that the languages that share this value are affected instead (regardless of typological relationship). For instance, for task `116A`: *Polar Questions* the label *Question particle* is always incorrectly predicted for the Spanish representations (even before neutralizing).

[5] Note that Libovický, Rosa, and Fraser (2020) already confirmed that this method does not negatively effect the underlying meaning of the sentence representations.

Table 6
The average change in performance per task τ and cross-neutralizing language x for M-BERT, categorized by languages that have the same (same) or different (diff) feature value from language x. Cases for which the probing task performance on the language before neutralizing was insufficient (< 75% accuracy) are denoted in gray (it is unclear what information these centroids capture, hence we cannot reasonably expect the same trend to emerge). Note that the blank spaces indicate the cases in which x was omitted from the task due to a lack of coverage in WALS.

x	Ukrainian		Swedish		Polish		Spanish		Marathi		Bulgarian		French	
τ	same	diff.	same	diff.	same	diff.	same	diff.	same	diff.	same	diff.	same	diff.
37A	−0.45	−0.22	−0.76	−0.01	−0.62	0.03	−0.62	0.02	−0.74	0.04	−0.46	0.02	−0.15	0.01
38A	−0.8	0.0	0.11	0.0	−0.67	0.0	−0.47	0.0	−0.82	0.05	–		−0.34	0.0
45A	–		−0.4	0.0	–		−0.44	0.0	−0.61	0.0	–		−0.21	0.0
47A	–		−0.32	0.01	–		−0.58	0.0	−0.82	0.01	–		−0.18	0.0
51A	−0.22	0.14	0.18	−0.23	−0.26	0.17	−0.3	0.0	−0.75	0.44	–		–	
70A	−0.58	0.33	−0.38	0.0	−0.26	0.12	0.08	−0.22	−0.1	−0.54	−0.61	0.21	−0.51	0.05
71A	−0.15	0.04	−0.55	0.05	−0.45	0.05	−0.68	0.08	0.08	−0.06	−0.28	−0.12	−0.13	−0.11
72A	−0.07	0.05	−0.36	0.39	−0.35	0.39	−0.5	0.48	0.24	−0.18	−0.23	0.0	−0.81	0.54
79A	−0.2	0.05	–		−0.68	0.07	–		–		−0.43	0.01	–	
79B	0.14	0.07	–		−0.35	0.06	–		–		−0.45	0.13	–	
81A	−0.1	0.0	−0.62	0.0	−0.28	0.0	−0.73	0.0	−0.57	0.0	−0.46	0.0	−0.38	0.0
82A	−0.4	0.35	−0.39	0.32	0.32	−0.36	0.1	−0.16	−0.53	0.42	0.71	−0.82	−0.06	0.03
83A	−0.08	0.0	−0.6	0.0	−0.27	0.0	−0.8	0.0	−0.6	0.0	−0.41	0.0	−0.47	0.0
85A	−0.07	0.0	−0.41	0.0	−0.27	0.0	−0.69	0.0	−0.64	0.0	−0.4	0.0	−0.38	0.0
86A	–		−0.34	0.0	0.09	−0.11	−0.5	0.01	−0.84	0.0	−0.6	0.06	−0.42	0.01
87A	−0.29	0.02	−0.32	0.02	−0.16	0.02	−0.73	0.0	−0.8	0.02	−0.67	0.02	−0.34	0.0
92A	0.26	−0.05	−0.35	0.15	0.28	−0.35	0.15	−0.16	–		–		−0.06	0.1
93A	–		−0.29	0.0	0.19	−0.21	–		−0.53	0.13	–		–	
95A	−0.12	0.0	−0.65	0.0	−0.28	0.0	−0.76	0.0	−0.53	0.0	−0.47	0.0	−0.43	0.0
97A	−0.24	0.0	−0.7	0.0	−0.48	0.0	−0.76	−0.03	−0.72	−0.03	−0.87	0.0	−0.42	−0.02
115A	–		–		–		−0.57	0.0	−0.55	0.0	–		−0.28	0.0
116A	−0.47	0.4	−0.22	0.19	0.11	−0.02	0.26	−0.14	−0.27	0.28	−0.36	0.34	–	
143F	−0.75	0.66	−0.21	0.0	−0.23	0.52	−0.25	0.53	0.24	−0.03	−0.25	0.53	0.16	−0.02
144D	−0.62	0.0	−0.47	0.0	–		−0.47	0.0	–		−0.34	0.0	–	
144J	−0.74	0.0	−0.54	0.0	−0.32	0.0	−0.45	0.0	–		−0.28	0.0	–	

Consequently, when cross-neutralizing with Spanish, the performance for languages that share this feature value deteriorates (note that in Figure 5 the orange dots drop in this case). This indicates that the model encodes the feature value *Question particle* for Spanish. Thus, when we compute $\bar{\mathbf{u}}_{\mathbf{x}}$, we capture information about this feature value instead of the correct one *Interrogative word order*.

Table 6 shows the average change in performance for M-BERT, categorized by feature value, for each language with which we neutralize (see Appendix, Table H.1 for XLM-R results). The table shows that there is a clear overall pattern where the performance in languages with the same feature value suffers, while that in languages with a different feature value remains intact. These results hold true for all languages we cross-neutralize with and for both encoders. In some cases, however, we notice that cross-neutralizing on average increases performance in languages with a different feature value (e.g., x = Ukrainian for task `70A`). We speculate that removing information about the feature value of x reduces noise in the representations, allowing the classifier to pick up on the right signal.

Thus, we find that language centroids capture specific feature values in a localizable and systematically similar way across different languages, indicating that typological properties are jointly encoded across languages. We reproduced all our experiments using sentence representations from the other layers of the models and obtained similar results in all layers (see Appendix, Figure I.1).

6. Conclusion

In this work, we proposed methods for testing multilingual sentence encoders by investigating a simple classifier's ability to separate languages within the models based on a wide range of typological properties. We found that all encoders capture language relationships based on some typological properties related to word order, pronouns, and negation. However, M-BERT and XLM-R generally outperform LASER and XLM, capturing variation along a wider range of linguistic properties. This is particularly evident for features pertaining to lexical properties. Thus, M-BERT and XLM-R appear to rely on typological properties to organize their language subspaces to a greater extent. Moreover, we found that these properties are localizable to the lower layers of LASER and XLM, while in M-BERT and XLM-R they are encoded in all layers. We hypothesize that these differences can be ascribed to the models' pretraining tasks. We found a correspondence between the language independence of models, induced during cross-lingual training, and a loss of typological information, indicating that universality comes at the cost of language-specific information. While we leave correlating typological features with performance on downstream tasks for future work, these findings can guide design choices when thinking about the behavior we want the model to exhibit.

Moreover, we have shown that typological feature values are encoded jointly across languages and are localizable in their respective language centroids. In the future, we will correlate the model's ability to encode typological features with its performance in downstream tasks by progressively deteriorating the amount of typological information encoded. In addition, our method enables us to carefully select which languages we want to neutralize with respect to certain typological properties. This could inspire work on encouraging selective generalization in large-scale models based on typological knowledge, as opposed to enforcing complete language-agnosticism. Lastly, our cross-neutralizing method is easily applicable to test for joint encoding in other scenarios—for example, linguistic and visual information sharing in multimodal models.

Appendix A. Languages Supported by XLM

Bulgarian (bul), French (fra), Spanish (spa), German (deu), Greek (ell), Russian (rus), Turkish (tur), Arabic (ara), Vietnamese (vie), Thai (tha), Chinese (zho), Hindi (hin), Swahili (swa), Swedish (swe) and Urdu (urd).

These languages are typologically diverse and cover all feature values used in our tasks. Thus, while the model might not have been trained on all languages used for probing, we made sure that the model was trained on languages that contain all values we probe for. Note that all other encoders support 93 (or more) languages, including all languages used in this work.

Appendix B. Reproducibility Details

Links to Source Code and Data. The following links can be used to download the pretrained models that we study in this work:

- LASER: BiLSTM.93langs.2018-12-2

- M-BERT: Bert-Base, multilingual cased version
- XLM: xlm mlm-tlm-xnli15
- XLM-R: xlm-r.base.v0

For the Transformers we relied on the implementations from HuggingFace, and for LASER we consulted the publicly available source code on their GitHub repository. Furthermore, sentences for all languages can be downloaded from the Tatoeba Web site, and to extract typological information from WALS we used the LingTypology API.

Table B.1
Summary statistics of the model architectures: tokenization method, number of layers L, dimensionality of sentence representations *dim*, number of attention heads H, number of model parameters, vocabulary size V, and pretraining tasks used.

Model	tokenization	L	dim	H	params	V	task	languages
LASER	BPE	5	1,024	–	52M	50K	MT	93
M-BERT	WordPiece	12	768	12	172M	110K	MLM+NSP	104
XLM	BPE	12	1,024	8	250M	95K	MLM+TLM	15
XLM-R	SentencePiece	12	768	12	270M	250K	MLM	100

Number of Model Parameters. The probing classifier has a varying number of parameters, depending on the dimensionality of the sentence representations *dim* and the number of class labels in the task o_n: $params = (dim \times 100) + (100 \times o_n) + 100 + o_n$. In the scalar mixing weights experiments, another $L + 1$ weights are added to this. See Table B.1 for the number of parameters in each multilingual encoder.

Computing infrastructure. The top-layer probing experiments were run using a 2.7 GHz Intel Core i7 CPU. The other experiments required more memory and were run on the Lisa cluster, maintained by SURFsara, using a 2.10 GHz Intel Xeon Silver 4110 CPU.

Appendix C. Similar Language Identification

For each test language we removed the most similar language from the train set and retrained the classifiers to test whether it is still able to predict the right label without being able to fall back to similar language identification. Each time we replace the most similar train language from each pair with German such that we still train on the same number of languages (e.g., when testing Spanish we replace Portuguese with German, for French we replace Italian with German, etc). We obtained the following results (for the classifier trained on top of XLM), shown in Table C.1, which confirm that in most cases it is still able to succeed with a high accuracy (> 85%).

Table C.1
For each language under investigation we see: (1) the number of times that the classifier is still able to succeed after removing the most similar language (note that we excluded languages from the total for which we did not succeed in the first place) and (2) the average accuracy across the tasks for which the classifier still succeeded.

Language	# times the classifier succeeds	Avg. accuracy across tasks
French	7/7	86.5%
Swedish	9/13	91.3%
Polish	10/15	85.0%
Ukrainian	6/7	99.5%
Bulgarian	5/6	86.6%
Spanish	11/12	87.3%
Marathi	0/3	N/A

Appendix D. Learned Mixing Weights

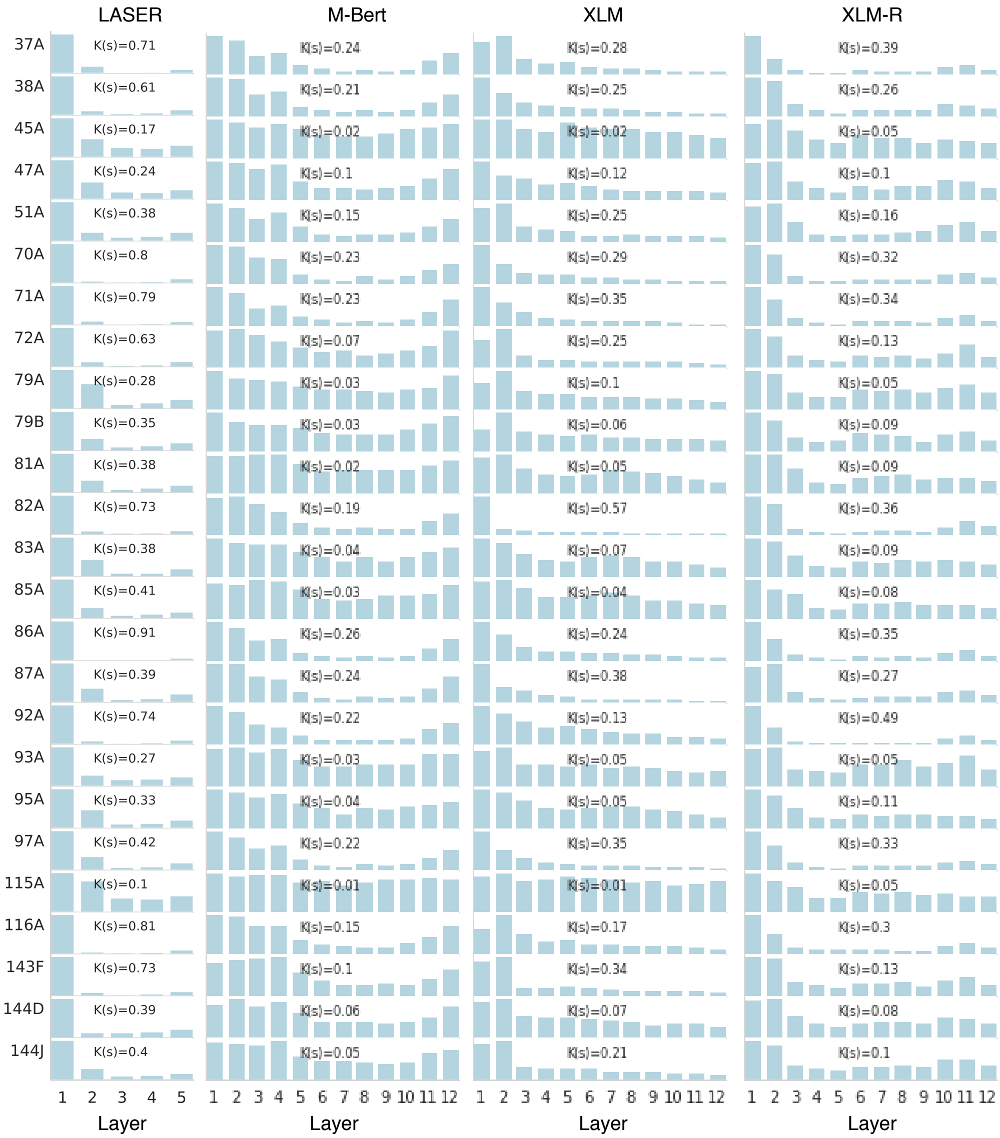

Figure D.1
Learned mixing weights s_τ for each encoder and the corresponding KL divergence $K(s)$ for all 25 tasks. We see that LASER and XLM exhibit the same pattern, where higher layers become less important. In M-BERT and XLM-R, on the other hand, layers from ± 10 and up seem to regain importance again.

Appendix E. t-SNE Plots per Layer

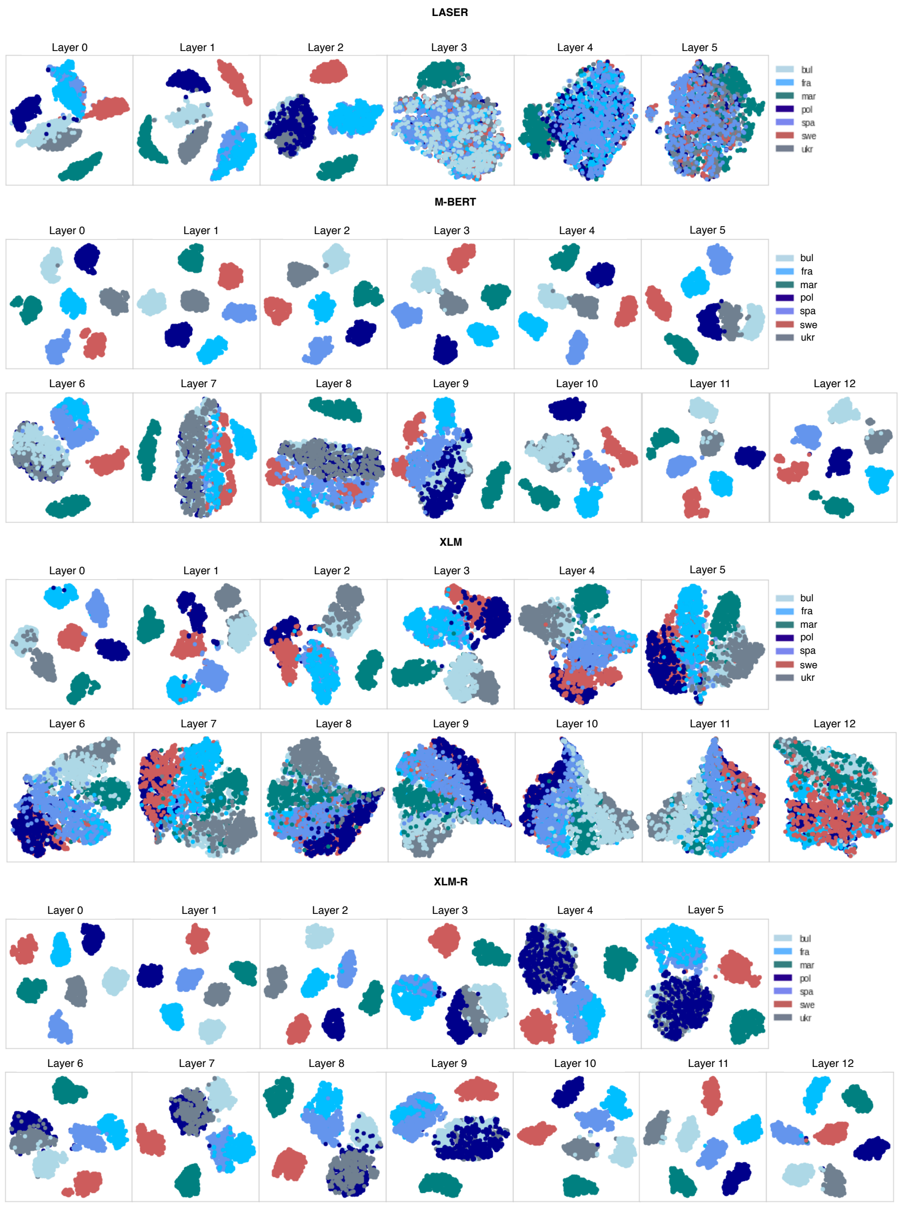

Figure E.1
t-SNE visualizations of the sentence representations retrieved from the different layers of LASER, M-BERT, XLM, and XLM-R, where layer 0 corresponds to the non-contextualized token embeddings (made using PCA with $k = 10$). Whereas LASER and XLM project all languages to a shared space in their last layers, M-BERT and XLM-R project the representations back to language-specific subspaces. Note that a similar trend is observed when only plotting the representations for the languages that XLM is trained on.

Appendix F. XLM Generalization to Unseen Languages

Table F.1
Macro-averaged-F1 scores for LASER and XLM computed separately over the set of XNLI languages that are not supported by XLM (Ukrainian, Polish, and Marathi) (non-XNLI). Note that LASER was trained on all languages and is thus used as a comparison to the scores obtained by XLM. We see that, in general, XLM obtains similar, and sometimes better, scores compared to LASER, despite not having been trained on the languages.

WALS code	LASER non-XNLI	XLM non-XNLI
37A	0.305003	0.315463
38A	0.325383	0.329041
45A	0.498301	0.494767
47A	0.481012	0.498201
51A	0.498408	0.498352
70A	0.257501	0.266342
71A	0.221012	0.209911
72A	0.386056	0.400814
79A	0.475045	0.433473
79B	0.198133	0.221936
81A	0.991734	0.963265
82A	0.399975	0.411855
83A	0.991061	0.944318
85A	0.991632	0.970614
86A	0.360055	0.363591
87A	0.495345	1.000000
92A	0.000754	0.000975
93A	0.367556	0.340872
95A	0.991335	0.985749
97A	0.659789	0.655212
115A	0.499235	0.497361
116A	0.426288	0.400575
143F	0.400558	0.400814
144D	0.499834	0.499734
144J	0.499083	1.000000

Appendix G. Cross-Neutralizing Results for M-BERT

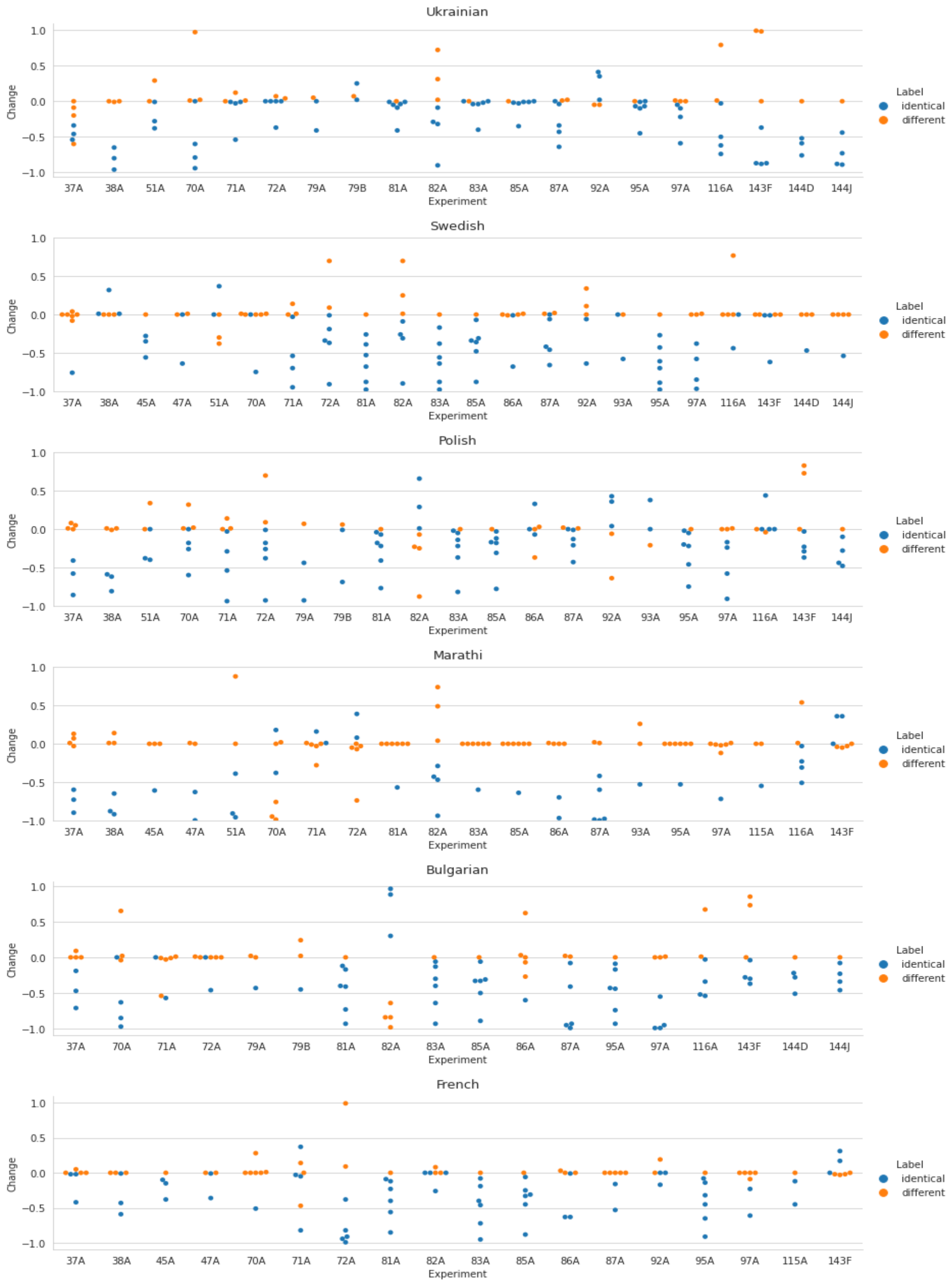

Figure G.1
Change in performance after cross-neutralizing with the other test languages for M-BERT. The performance change for all 25 probing tasks is shown per language used for cross-neutralizing.

Appendix H. Averaged Performance Change Over Languages for XLM-R

Table H.1
The average change in performance per task τ and cross-neutralizing language x for XLM-R categorized by languages that have the same and those that have a different feature value from language x. Cases for which the probe performance on the language before neutralizing was insufficient (< 75% accuracy) are denoted in gray (it is unclear what information these centroids capture, hence we cannot reasonably expect the same trend to emerge). Note, the blank spaces indicate the cases in which x was omitted from the task due to a lack of coverage in WALS.

x	Ukrainian		Swedish		Polish		Spanish		Marathi		Bulgarian		French	
τ	same	diff.	same	diff.	same	diff.	same	diff.	same	diff.	same	diff.	same	diff.
37A	−0.26	0.0	−0.74	−0.02	−0.77	0.01	−0.81	0.0	−0.74	0.01	−0.54	0.0	−0.22	0.0
38A	−0.76	0.02	0.13	−0.0	−0.5	0.01	−0.52	0.0	−0.88	0.0	–		−0.2	0.0
45A	–		−0.72	0.0	–		−0.34	0.0	−0.52	0.0	–		−0.28	0.0
47A	–		−0.48	0.0	–		−0.72	0.0	−0.64	0.0	–		−0.24	0.0
51A	−0.81	0.51	0.24	−0.14	−0.17	0.31	−0.27	0.0	−0.53	0.5	–		–	
70A	−0.63	0.02	−0.36	−0.01	−0.16	0.02	0.11	−0.31	0.1	−0.57	−0.6	0.03	−0.62	0.02
71A	−0.66	−0.08	−0.68	0.0	−0.72	0.01	−0.68	0.11	0.08	0.0	−0.27	0.04	−0.18	−0.19
72A	−0.09	0.03	−0.32	0.04	−0.77	0.39	−0.33	0.04	0.24	−0.61	−0.2	0.0	−0.69	0.28
79A	−0.24	0.05	–		−0.72	0.07	–		–		−0.35	0.01	–	
79B	0.29	−0.03	–		−0.22	0.03	–		–		−0.44	0.13	–	
81A	−0.09	0.0	−0.81	0.0	−0.42	0.0	−0.51	0.0	−0.5	0.0	−0.57	0.0	−0.52	0.0
82A	−0.52	0.7	−0.72	0.73	0.29	−0.25	0.25	−0.24	−0.48	0.63	0.8	−0.84	−0.08	0.02
83A	−0.14	0.0	−0.81	0.0	−0.41	0.0	−0.52	0.0	−0.48	0.0	−0.52	0.0	−0.55	0.0
85A	−0.19	0.0	−0.84	0.0	−0.32	0.0	−0.49	0.0	−0.49	0.0	−0.47	0.0	−0.55	0.0
86A	–		−0.58	0.1	0.09	−0.02	−0.56	0.0	−0.8	−0.02	−0.62	0.06	−0.3	−0.01
87A	−0.88	0.01	−0.28	0.01	−0.19	0.01	−0.75	0.0	−0.58	0.01	−0.44	0.01	−0.28	0.0
92A	0.7	−0.26	−0.38	0.26	0.4	−0.12	0.13	−0.0	–		–		0.1	−0.08
93A	–		−0.26	0.0	0.22	0.0	–		−0.49	0.5	–		–	
95A	−0.13	0.01	−0.83	0.01	−0.43	0.01	−0.52	0.01	−0.48	0.0	−0.56	0.01	−0.52	0.01
97A	−0.89	0.03	−0.69	0.03	−0.22	0.03	−0.7	0.01	−0.72	−0.0	−0.64	0.03	−0.26	−0.01
115A	–		–		–		−0.6	0.0	−0.51	0.0	–		−0.32	0.0
116A	−0.14	0.02	−0.26	0.23	0.11	−0.02	0.22	−0.32	−0.56	0.49	−0.45	0.45	–	
143F	−0.13	0.2	−0.19	0.0	−0.3	0.29	−0.67	0.34	0.48	−0.97	−0.74	0.34	0.14	−0.09
144D	−0.16	0.0	−0.52	0.0	–		−0.53	0.0	–		−0.75	0.0	–	
144J	−0.14	0.0	−0.58	0.0	−0.36	0.0	−0.55	0.0	–		−0.81	0.0	–	

Appendix I. Cross-Neutralizing Results for M-BERT Across Layers

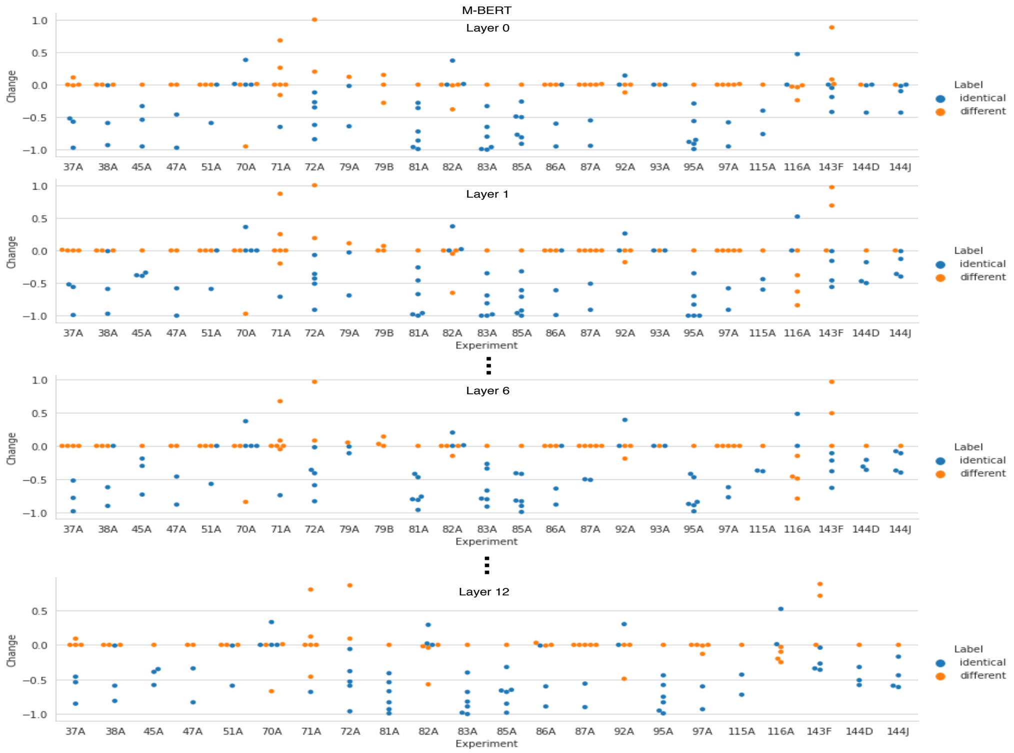

Figure I.1
The change in performance for all test languages when cross-neutralizing M-BERT representations with a language centroid computed from the Spanish sentences. Languages are categorized by whether they had the same or different feature value from that of Spanish for the respective tasks.

References

Agić, Željko, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 13–24. https://doi.org/10.3115/v1/W14-4203

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016a. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444. https://doi.org/10.1162/tacl_a_00109

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016b. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. https://doi.org/10.1162/tacl_a_00288

Banea, Carmen, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135. https://doi.org/10.3115/1613715.1613734

Beinborn, Lisa and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603. https://doi.org/10.1162/coli_a_00382

Bjerva, Johannes, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389. https://doi.org/10.1162/coli_a_00351

Blevins, Terra, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs Encode Soft Hierarchical Syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19. https://doi.org/10.18653/v1/P18-2003

Chen, Xilun and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D18-1024

Chi, Ethan A., 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, ACL 2020*, pages 5564–5577. https://doi.org/10.18653/v1/2020.acl-main.493

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. https://doi.org/10.18653/v1/P18-1198

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. https://doi.org/10.18653/v1/D18-1269

Croft, William. 2002. *Typology and Universals*, Cambridge University Press. https://doi.org/10.1017/CBO9780511840579

Devlin, Jacob, 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.

Dryer, Matthew S. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dryer, Matthew S. and Martin Haspelmath, editors. 2013. *WALS Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Duong, Long, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. 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 845–850. https://doi.org/10.3115/v1/P15-2139

Duong, Long, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348. https://doi.org/10.18653/v1/D15-1040

Ganchev, Kuzman, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 369–377. https://doi.org/10.3115/1687878.1687931

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in*

Natural Language Processing, pages 316–327. https://doi.org/10.18653/v1/D18-1029

Gonen, Hila, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from Multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56. https://doi.org/10.18653/v1/2020.blackboxnlp-1.5

Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu. 2016. A universal framework for inductive transfer parsing across multi-typed treebanks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 12–22.

Hewitt, John 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, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Huang, Haoyang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494. https://doi.org/10.18653/v1/D19-1252

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325. https://doi.org/10.1017/S1351324905003840

Karthikeyan, K., Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Khapra, Mitesh M., Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for WSD. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 561–569.

Kim, Taeuk, Bowen Li, and Sang-goo Lee. 2021. Multilingual chart-based constituency parse extraction from pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 454–463. https://doi.org/10.18653/v1/2021.findings-emnlp.41

Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NIPS)*, pages 7059–7069.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499. https://doi.org/10.18653/v1/2020.emnlp-main.363

Libovický, Jindřich, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1663–1674. https://doi.org/10.18653/v1/2020.findings-emnlp.150

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Naseem, Tahira, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637.

Navigli, Roberto and Simone Paolo Ponzetto. 2012. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1399–1410.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 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.

Nozza, Debora, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

O'Horan, Helen, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308.

Pappas, Nikolaos and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *IJCNLP (1)*, pages 1015–1025.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, Matthew, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018a. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509. https://doi.org/10.18653/v1/D18-1179

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pires, Telmo, 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. https://doi.org/10.18653/v1/P19-1493

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601. https://doi.org/10.1162/coli_a_00357

Ponti, Edoardo Maria, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1531–1542. https://doi.org/10.18653/v1/P18-1142

Qian, Peng, Xipeng Qiu, and Xuan-Jing 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. https://doi.org/10.18653/v1/P16-1140

Ravishankar, Vinit, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47.

Ravishankar, Vinit, Lilja Øvrelid, and Erik Velldal. 2019. Probing multilingual sentence representations with X-PROBE. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168. https://doi.org/10.18653/v1/W19-4318

Şahin, Gözde Gül, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385. https://doi.org/10.1162/coli_a_00376

Singh, Jasdeep, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55. https://doi.org/10.18653/v1/D19-6106

Smith, Noah A. and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 354–362. https://doi.org/10.3115/1219840.1219884

Täckström, Oscar, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. `https://doi.org/10.18653/v1/P19-1452`

Tenney, Ian, Patrick Xia, Berlin Chen, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Tiedemann, Jörg. 2015. Cross-lingual dependency parsing with universal dependencies and predicted POS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Tiedemann, Jörg, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*, pages 130–140. `https://doi.org/10.3115/v1/W14-1614`

Tran, Ke M. and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 281–288. `https://doi.org/10.18653/v1/D19-6132`

van der Heijden, Niels, Samira Abnar, and Ekaterina Shutova. 2020. A comparison of architectures and pretraining methods for contextualized multilingual word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9090–9097. `https://doi.org/10.1609/aaai.v34i05.6443`

Wang, Zirui, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Wu, Shijie and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. 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 833–844. `https://doi.org/10.18653/v1/D19-1077`

Yarowsky, David, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8. `https://doi.org/10.3115/1072133.1072187`

Yogatama, Dani, Cyprien de Masson d'Autume, Jerome Connor, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Zeman, Daniel and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Zhang, Yuan, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of NAACL-HLT*, pages 1307–1317. `https://doi.org/10.18653/v1/N16-1156`

Zhang, Yuan, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal POS tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1368–1378.

Zhou, Guangyou, Tingting He, Jun Zhao, and Wensheng Wu. 2015. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1426–1432.

Zuccon, Guido, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*, pages 1–8. `https://doi.org/10.1145/2838931.2838936`