LINSPECTOR: Multilingual Probing Tasks for Word Representations

Gözde Gül Şahin
AIPHES and UKP Lab / TU Darmstadt
Technische Universität Darmstadt
Department of Computer Science
sahin@ukp.informatik.tu-darmstadt.de

Clara Vania*
New York University
c.vania@nyu.edu

Ilia Kuznetsov
AIPHES and UKP Lab / TU Darmstadt
Kuznetsov@ukp.informatik.tu-darmstadt.de

Iryna Gurevych
AIPHES and UKP Lab / TU Darmstadt
Gurevych@ukp.informatik.tu-darmstadt.de

Despite an ever-growing number of word representation models introduced for a large number of languages, there is a lack of a standardized technique to provide insights into what is captured by these models. Such insights would help the community to get an estimate of the downstream task performance, as well as to design more informed neural architectures, while avoiding extensive experimentation that requires substantial computational resources not all researchers have access to. A recent development in NLP is to use simple classification tasks, also called probing tasks, that test for a single linguistic feature such as part-of-speech. Existing studies mostly focus on exploring the linguistic information encoded by the continuous representations of English text. However, from a typological perspective the morphologically poor English is rather an outlier: The information encoded by the word order and function words in English is often stored on a subword, morphological level in other languages. To address this, we introduce 15 type-level

* Work done while Clara Vania was a PhD student at the ILCC / University of Edinburgh.

Submission received: 20 March 2019; revised version received: 5 November 2019; accepted for publication: 29 January 2020.

https://doi.org/10.1162/COLI_a_00376

probing tasks such as case marking, possession, word length, morphological tag count, and pseudoword identification for 24 languages. We present a reusable methodology for creation and evaluation of such tests in a multilingual setting, which is challenging because of a lack of resources, lower quality of tools, and differences among languages. We then present experiments on several diverse multilingual word embedding models, in which we relate the probing task performance for a diverse set of languages to a range of five classic NLP tasks: POS-tagging, dependency parsing, semantic role labeling, named entity recognition, and natural language inference. We find that a number of probing tests have significantly high positive correlation to the downstream tasks, especially for morphologically rich languages. We show that our tests can be used to explore word embeddings or black-box neural models for linguistic cues in a multilingual setting. We release the probing data sets and the evaluation suite LINSPECTOR with `https://github.com/UKPLab/linspector`*.*

1. Introduction

The field of natural language processing (NLP) has seen great development after replacing the traditional discrete word representations with continuous ones. Representing text with dense, low-dimensional vectors—or **embeddings**—has become the de facto approach, since these representations can encode complex relationships between the units of language and can be learned from unlabeled data, thus eliminating the need for expensive manual feature engineering. The initial success of dense representations in NLP applications has led to the development of a multitude of embedding models, which differ in terms of design objective (monolingual [Mikolov et al. 2013b], crosslingual [Ruder, Vulic, and Søgaard 2019], contextualized [Peters et al. 2018], retrofitted [Faruqui et al. 2015], multi-sense [Pilehvar et al. 2017], cross-domain [Yang, Lu, and Zheng 2017], dependency-based [Levy and Goldberg 2014]), encoding architecture, (convolution [Kim et al. 2016], linear vector operations [Bojanowski et al. 2017], bidirectional LSTM [Ling et al. 2015]), as well as in terms of the target units (words, characters, character n-grams, morphemes, phonemes).

While offering substantial benefits over the traditional feature-based representations of language, the performance of unsupervised embeddings may differ considerably depending on the language and the task. For instance, early embedding models such as word2vec (Mikolov et al. 2013b) and GloVe (Pennington, Socher, and Manning 2014) have been shown to suffer from out-of-vocabulary (OOV) issues for agglutinative languages like Turkish and Finnish (Şahin and Steedman 2018), while performing relatively well on analytic and fusional languages like English. Furthermore, there is no guarantee that a representation well-suited for some task would score similarly well at other tasks even for the same language due to the differences in the information required to solve the tasks, as demonstrated in Rogers, Ananthakrishna, and Rumshisky (2018).

Given the variety of word representations and parameter options, searching for the right word representation model for a specific language and a certain task is not trivial. Scanning the large parameter space may be extremely time consuming and computationally expensive, which poses significant challenges, especially in the lower-resource non-English academic NLP communities. To simplify the search for a good representation, and estimate the "quality" of the representations, intrinsic evaluation via similarity and analogy tasks has been proposed. Although these tasks seem to be intuitive, there are concerns regarding their consistency and correlation with downstream task performance (Linzen 2016; Schnabel et al. 2015). Furthermore, such evaluation requires manually created test sets and these are usually only available for a small number of languages. Another option to assess the quality of word representation is through extrinsic evaluation, where the word vectors are used directly in downstream tasks, such as machine translation (Ataman and Federico 2018), semantic role labeling (SRL) (Şahin and Steedman 2018) or language modeling (Vania and Lopez 2017). Although this method provides more insightful information about the end task performance, it requires expensive human annotations, computational resources, and the results are sensitive to hyperparameter choice.

To address the aforementioned problems, a few studies have introduced the idea of **probing tasks** (Köhn 2016; Shi, Padhi, and Knight 2016; Adi et al. 2017; Veldhoen, Hupkes, and Zuidema 2016; Conneau et al. 2018a); which are a set of multi-class classification problems that probe a learned word vector for a specific linguistic property, such as part-of-speech (POS), semantic, or morphological tag.[1] Probing tasks have gained a lot of attention (Belinkov et al. 2017; Bisazza and Tump 2018, among others) due to their simplicity, low computational cost, and ability to provide some insights regarding the linguistic properties that are captured by the learned representations.

The majority of the probing tests proposed so far are mostly designed for *English language only*, and operate on the *sentence-level* (e.g., *tree depth*, *word count*, *top constituent* by Conneau et al. 2018a). Although sentence-level probing may provide valuable insights for English sentence-level representations, we hypothesize that they would not be similarly beneficial in a multilingual setup for several reasons. The first reason is that the information encoded by the word order and function words in English is encoded at the morphological, subword level information in many other languages. Consider the Turkish word *katılamayanlardan*, which means "he/she is one of the folks who can not participate." In morphologically complex languages like Turkish, single tokens might already communicate a lot of information such as event, its participants, tense, person, number, polarity. In analytic languages, this information would be encoded as a multi-word clause. The second reason is the confusion of the signals: As pointed out by Tenney et al. (2019, page 10), sometimes "operating on full sentence encodings introduces confounds into the analysis, since sentence representation models must pool word representations over the entire sequence." Furthermore, we argue that such tests would carry over the statistics of the data they originate from, introducing undesired biases such as domain and majority bias. In order to address the aforementioned issues, we introduce context independent, dictionary-based *type-level* probing tasks that operate on word-level and do not contain domain or majority biases. To investigate the limitations and strengths of the proposed type-level tasks, we introduce and investigate another set of similar, but context dependent, treebank-based and thereby potentially biased *token-level* tests.

In this work:

- We extend the line of work by Conneau et al. (2018a) and Tenney et al. (2019) and introduce *15 type-level* probing tasks for *24 languages* by taking language properties into account. Our probing tasks cover a range of features: from superficial ones such as word length, to morphosyntactic features such as case marker, gender, and number; and psycholinguistic ones like pseudowords (artificial words that are phonologically

[1] We use the terms probing tasks and probing tests interchangeably throughout the article.

well-formed but have no meaning). Although languages share a large set of common probing tasks, each has a list of its own, for example, Russian and Spanish are probed for gender, whereas Turkish is probed for polarity and possession.

- We introduce a reusable, systematic methodology for creation and evaluation of such tests by utilizing the existing resources such as UniMorph (Sylak-Glassman et al. 2015; Sylak-Glassman 2016; Kirov et al. 2018), Wikipedia, and Wuggy (Keuleers and Brysbaert 2010).

- We then use the proposed probing tasks to evaluate a set of diverse multilingual embedding models and to diagnose a neural end-to-end semantic role labeling model as a case study. We statistically assess the correlation between probing and downstream task performance for a variety of downstream tasks (POS tagging, dependency parsing [DEP], SRL, named entity recognition [NER], and natural language inference [NLI]) for a set of typologically diverse languages and find that a number of probing tests have significantly high positive correlation to a number of syntactic and semantic downstream tasks, especially for morphologically rich languages.

- We introduce a set of comparable **token-level** probing tasks that additionally employs the context of the token. We analyze the type- and token-level probing tasks through a series of intrinsic and diagnostic experiments and show that they are similar with some exceptions: Token-level tasks may be influenced by *domain and majority class biases*, whereas type-level tasks may suffer in case of *lack of lexical diversity* and high *ambiguity ratios*.

- We provide comprehensive discussions for the intrinsic and extrinsic experimental results along with diagnostic and correlation study. We show that numerous factors except from the neural architectures play a role on the results such as out-of-vocabulary rates, domain similarity, statistics of both data sets (e.g., ambiguity, size), training corpora for the embeddings; as well as typology, language family, paradigm size, and morphological irregularity.

- We release the LINSPECTOR framework that contains the probing data sets along with the intrinsic and extrinsic evaluation suite: `https://github.com/UKPLab/linspector`.

We believe our evaluation suite together with probing data sets could be of great use for comparing various multilingual word representations such as automatically created crosslingual embeddings; exploring the linguistic features captured by word encoding layers of black-box neural models; systematic searching of model or architecture parameters by evaluating the models with different architectures and parameters on the proposed probing tasks; or comparing transfer learning techniques (i.e., by evaluating a set of crosslingual embeddings that are transferred or learned using different transfer learning techniques) on the proposed language-specific probing task set.

2. Related Work on Word Representation Evaluation

We begin with a review of related work on word representation evaluation. We divide the current evaluation schemes for word representations into two main categories: (1) **intrinsic**, when vectors are evaluated on a variety of benchmarks and (2) **extrinsic**, when they are evaluated on downstream NLP tasks.

2.1 Intrinsic Evaluation

A standard approach to evaluate continuous word representations is by testing them on a variety of benchmarks that measure some linguistic properties of the word. These similarity benchmarks typically consist of a set of words or word pairs that are manually annotated for some notion of relatedness (semantic, syntactic, topical, etc.). For English, some of the widely used similarity benchmarks are WordSim-353 (Finkelstein et al. 2001), MC (Miller and Charles 1991), RG (Rubenstein and Goodenough 1965), SCWS (Huang et al. 2012), rare words data set (RW) (Luong, Socher, and Manning 2013), MEN (Bruni et al. 2012), and SimLex-999 (Hill, Reichart, and Korhonen 2015). Although these benchmarks have shown to be useful for evaluating English word representations, only very few word similarity data sets exist in other languages. Human-assessed translations of WordSim-353 and SimLex-999 on three languages, Italian, German, and Russian, have been collected.[2] For the SemEval 2017 shared task, Camacho-Collados et al. (2017) introduced manually curated word-similarity data sets for English, Farsi, German, Italian, and Spanish.

Another popular benchmark for evaluating word representations is the word analogy test. This test was specifically introduced by Mikolov et al. (2013a) to evaluate word vectors trained using neural models. The main goal is to determine how syntactic and semantic relationships between words are reflected in the continuous space. Given a pair of words, *man* and *woman*, the task is to find a target word that shares the same relation with a given source word. For example, given a word *king*, one expected target word would be *queen*. The analogy task has gained considerable attention mainly because it demonstrates how "linguistic regularities" are captured by word representation models. The analogy data set of Mikolov et al. (2013a) consists of 14 categories covering both syntactic and semantic regularities. Although analogy test has become a standard evaluation benchmark, Rogers, Drozd, and Li (2017) and Linzen (2016) identified certain theoretical and practical drawbacks of this approach, which are mostly related to the consistency of the vector offset and the structure of the vector space model. Pairwise similarity benchmarks and word analogy tasks only offer a first approximation of the word embedding properties and provide limited insights into the downstream task performance. To address this limitation, Tsvetkov et al. (2015) introduced QVEC, an intrinsic word evaluation method that aligns word vector representations with hand-crafted features extracted from lexical resources, focusing on the semantic content. They showed that their evaluation score correlates strongly with performance in downstream tasks.

More recently, Rogers, Ananthakrishna, and Rumshisky (2018) proposed a comprehensive list of scores, so-called linguistic diagnostics factors, and analyzed their relation to a set of downstream tasks such as chunking, NER, or sentiment classification, using word2vec (Mikolov et al. 2013a) and GloVe (Pennington, Socher, and Manning 2014)

[2] http://leviants.com/ira.leviant/MultilingualVSMdata.html.

word representations. They extend the traditional intrinsic evaluation (word similarity and analogy) with semantics extracted from existing resources such as WordNet, and basic morphological information like shared lemma and affixes. Their findings support the previous studies that observe low correlation between word similarity/analogy and sequence-labeling downstream task performance. In addition, they observe high correlation between morphology-level intrinsic tests with such downstream tasks even for English—one of the morphologically poorest languages. Unlike probing studies that train classifiers, they rely on nearest neighbor relation as a proxy to predict the performance of word vectors similar to early word analogy works.

2.2 Extrinsic Evaluation

In general, evaluating word vectors on downstream NLP tasks is more challenging because of the time and resources needed for the implementation. The two most common approaches are to test a single representation model on several downstream tasks (Ling et al. 2015; Pennington, Socher, and Manning 2014; Bojanowski et al. 2017), or to test a number of representation models on a single task (Vania and Lopez 2017; Ataman and Federico 2018; Şahin and Steedman 2018; Gerz et al. 2018). For a more general extrinsic evaluation, we note the work of Nayak, Angeli, and Manning (2016), which introduces an evaluation suite of six downstream tasks: two tasks to assess the syntactic properties of the representations and four tasks to assess the semantic properties. Because this type of evaluation is typically task-specific, it can be conducted in multilingual settings. However, training a range of task-specific multilingual models might require significant resources, namely, training time and computational power. Apart from that, differences in the exact task formulation and the underlying data sets among languages might influence the evaluation results.

2.3 Evaluation Via Probing Task

The rise of deep learning based methods in NLP has stimulated research on the interpretability of the neural models. In particular, several recent studies analyze representations generated by neural models to get insights on what kind of linguistic information is learned by the models. Interpretability studies have been one of the emerging trends in NLP as hinted by the on-going Representation Evaluation (RepEval) (Nangia et al. 2017) and BlackBoxNLP Workshop series (Tal Linzen, Chrupała, and Alishahi 2018) organized in popular conference venues. The most common approach is to associate some linguistic properties such as POS, morphological, or semantic properties with specific representations from a trained model (hidden states or activation layer). This method, which is called **probing task** or **diagnostic classifier** (Shi, Padhi, and Knight 2016; Adi et al. 2017; Veldhoen, Hupkes, and Zuidema 2016), uses representations generated from a fully trained model with frozen weights to train a classifier predicting a particular linguistic property. The performance of this classifier is then used to measure how well the model has "learned" this particular property. A similar study has been conducted by Köhn (2015), who proposed training such classifiers for predicting syntactic features such as gender and tense, extracted from annotated dependency treebanks. Because of the unavailability of subword or contextualized embeddings at that time, the author only experimented with static word-level embeddings (word2vec, GloVe, and embeddings derived from Brown clusters) and found that they are suprisingly able to capture linguistic properties, in particular for POS information. The study assumes that the performance of this targeted word feature classifier would be directly related to the parser performance, which is later tested empirically with diagnostic classifiers on syntactic parsers for simple linguistic properties such as tense and number (Köhn 2016). Although the syntax-based classifiers in Köhn (2016) are conceptually similar to our single feature probing tasks, there are several differences. First, the training instances are created from an annotated treebank including the ambiguous words; this may introduce domain, annotator, and majority class bias unlike ours; and it may lead to inconsistent results due to unresolved ambiguity. In addition, the study is limited to the following tests: case, gender, tense, and number; and to syntactic parsing as the downstream task. Finally, it has used only three word embedding models, which can be considered too small to draw conclusions; and too similar, as their training objectives and training units are similar.

Qian, Qiu, and Huang (2016) investigate the effects of word inflection and typological diversity in word representation learning. They observe that language typology (word order or morphological complexity) influences how linguistic information is encoded in the representations. They also compare a standard character-level auto-encoder model to a word-level model (word2vec Skip-Gram) and find that character-level models are better at capturing morphosyntactic information. Their study highlights the importance of utilizing word form information as well as language typology.

Recent works on probing have focused on analyzing the representations learned when training for specific downstream tasks, such as machine translation (Shi, Padhi, and Knight 2016; Belinkov et al. 2017; Bisazza and Tump 2018) or dependency parsing (Vania, Grivas, and Lopez 2018). Although this approach allows probing for multilingual data, it is still task-specific and might require expensive computation for model training (e.g., machine translation typically needs a large amount parallel data for training). For a more general evaluation, Conneau et al. (2018a) and Tenney et al. (2019) each introduced a broad coverage evaluation suite to analyze representations on the sentence level with a focus on English. We build our methodology upon these recent works. However, unlike their methods, our evaluation suite is multilingual and takes language-specific features into account. Moreover, our tests are type-level, rather than sentence- (Conneau et al. 2018a) or sub-sentence-level (Tenney et al. 2019).

Finally, Belinkov and Glass (2019) recently surveyed various analysis methods in NLP and mention three important aspects for model analysis: (1) the methods (classifiers, correlations, or similarity), (2) the linguistic phenomena (sentence length, word order, syntactic, or semantic information, etc), and (3) the neural network components (embeddings or hidden states). They have also provided a non-exhaustive list of previous work which use probing task (classifier) method for analyzing representations, including word representations. For a more comprehensive list of studies on what linguistic information is captured in neural networks, we refer the readers to Belinkov and Glass (2019).

3. Probing Tasks

With our probing tasks we aim to cover the properties ranging from shallow (e.g., word length [Conneau et al. 2018a]), to deeper ones (e.g., distinguishing pseudowords from in-vocabulary words). First, we probe for morphosyntactic and morphosemantic features such as case marking, gender, tense, and number. Most probing tasks are defined for all languages, such as POS and number; however, some features are only defined for a subset of languages, for example, polarity for Portuguese and Turkish, gender for Arabic and Russian. To maintain consistency, we base the majority of our tasks on the universal grammatical classes introduced by the UniMorph project (Sylak-Glassman et al. 2015). Second, we propose tasks to evaluate a more general syntactic/semantic capability of the model such as predicting the number of morphological tags, detecting the shared or odd linguistic feature between two word forms. Finally, inspired by cognitive linguistics, we assess the ability of the embedding models to detect pseudowords (i.e., words that are phonetically similar to an existing word but have no meaning). The conceptual definitions of our probing tests are given in Section 3.1, and Section 3.2 describes the specific implementation of the probing tests used in this work.

3.1 Task Definitions

Case Marking. A substantial number of languages express the syntactic and semantic relationship between the nominal constituents and the verbs via morphological case markers. Iggesen (2013) reports that 161 out of 261 languages have at least two case markers, as shown in Table 1. Although cases may undertake different roles among languages, a type of case marking, named as *core*, *non-local*, *nuclear*, or *grammatical case*, is the most common. This category contains case markers that are used to mark the arguments of verbs such as subjects, objects, and indirect objects (Blake 2001; Comrie and Polinsky 1998). In languages with rich case marking systems, case is also commonly used to mark roles such as "location" and "instrument." Below are examples of Russian and Turkish sentences that use `Acc` and `Inst` case markers to define the patient (object affected by the action) and the instrument.

(1) a. *Mark-∅* *razbi-l-∅* *okn-o* *molotk-om*
Mark-NOM.SG break-PST-SG.M window-ACC.SG hammer-INST.SG

b. *Mark-∅* *pencere-yi* *çekiç-le* *kır-dı*
Mark-NOM.SG window-ACC.SG hammer-INST.SG break-PST.3.SG

'Mark broke the window with a hammer.'

The relation between case markers and NLP tasks such as semantic role labeling, dependency parsing, and question answering have been heavily investigated and using case marking as a feature has been shown beneficial for numerous languages and tasks (Isgüder and Adali 2014; Eryigit, Nivre, and Oflazer 2008).

Table 1
Languages with case marking.

# Case Categories	# Languages	Example
0	100	English, Spanish
2	23	Romanian, Persian
3	9	Greek
4	9	Icelandic, German, Albanian
5	12	Armenian, Serbo-Croatian, Latvian
6–7	37	Turkish, Polish, Russian, Georgian
8–9	23	Japanese
10 or more	24	Estonian, Finnish, Basque

Gender. According to Corbett (2013), more than half of the world languages do not have a gender system. The majority of the languages with a gender system, such as Spanish, French, German, and Russian, define either two (feminine, masculine) or three (+neutral) classes. Gender is a grammatical category and participates in agreement: If a language has a gender system, the gender of a noun or pronoun influences the form of its syntactic neighbors, which could be verb, adjective, determiner, numeral, or a focus particle, depending on the language. Related to NLP tasks, Hohensee and Bender (2012) showed that agreement-based features including gender can improve the quality of dependency parsing for morphologically rich languages. Bengtson and Roth (2008) demonstrate how gender can be used to improve co-reference resolution quality. In the Russian example sentence given below, the gender agreement between the subject, its adjective modifier, and the verb is shown.

(2) *Gosudarstvenn-aya duma sdela-l-a zayavlenie*
State-NOM.SG.F parliament.F make-PST-SG.F announcement

'The parliament made an announcement.'

The agreement features such as gender and number are crucial for structured grammatical analysis such as dependency parsing, co-reference resolution, as well as for grammar checking and correction, and automatic essay evaluation.

Mood. Modality of the verb, namely, the grammatical mood, is used to communicate the status of the proposition from the speaker's point of view. Some common mood categories are Indicative, Conditional, Subjunctive, Imperative-Jussive, and Potential. Many languages mark the modality of the verb with morphological affixes. German and Russian example sentences with Imperative mood feature are given below.

(3) a. *Bring-e mir das Buch*
Bring-2SG.IMP me the book

b. *Prines-i mne knigu*
Bring-2SG.IMP me book

'Bring me the book.'

Because Mood signals the factuality of the statement, it might be relevant for natural language inference and related tasks, as we demonstrate in Section 6.1; the ability of the representation to encode imperative, in turn, could be essential for interpreting the user input in dialogue systems.

Number. This feature is usually expressed by nouns, adjectives, and verbs, and, similar to gender, number is a common feature for agreement. The two most common values for gender is Singular and Plural, which are often marked by morphological affixes.

POS. We use the following eight categories defined by the UniMorph Schema: nouns, adpositions, adjectives, verbs, masdars, participles, converbs, and adverbs. A more detailed explanation for each category can be found in Sylak-Glassman (2016). POS has been one of the most prominent features of all high-level NLP tasks for decades. Throughout this work, we will use the coarse POS categories, which are universal across languages.

Person. We use the traditional six person categories that are commonly marked by morphological markers: first, second, and third person either singular or plural. This feature has strategic importance for dependency parsing, co-reference resolution, as well as high-level tasks that involve natural language understanding such as conversational agents, question answering, or multimodal applications such as generating images from sentences. An example of using person agreement to improve dependency parsing quality is shown in Hohensee and Bender (2012). Two Russian example sentences below demonstrate coordination between the personal pronoun and the verb, indicating a syntactic dependency between them.

(4) a. *Ja* *vizh-u* *ptitsu*
I.1Sg see-1Sg.Prs bird

b. *On* *vid-it* *ptitsu*
He.3Sg see-3Sg.Prs bird

(a) 'I see a bird.' (b) 'He sees a bird.'

Polarity. Some languages mark the verbs with polarity to indicate whether a statement is negative or positive. Generally, markers are used to specify the negative polarity, assuming the positive polarity by default. The verb "go" is marked with a negative marker in the Turkish sentence given below. Although this feature is not notably common across languages, it has immediate use in cases such as sentiment analysis and natural language inference, similar to negation in English.

(5) *Dün* *okul-a* *git-me-di-m*
yesterday school-Dat.Sg go-Neg-Pst-Sg

'He/she didn't go to school yesterday.'

Possession. Although the majority of the languages use adjectives such as his/her/my to express possession, some languages such as Turkish and Arabic use morphological markers on the nouns. The number of values for the feature depends on the gender system of the language. For instance, while Arabic separately marks the possession by third person singular for feminine and masculine, Turkish uses only one marker for the possession by the third person singular.

(6) *Ayakkabı-(s)ı-(n)ı* *giy-ecek*
shoe-Poss.3Sg-Acc wear-3Sg.Fut

'He/she will wear his/her shoes.'

An example sentence in Turkish with "he/she will wear his/her shoes" is given above. As can be seen, possession implicitly acts as an agreement feature (i.e., possession of the object and person of the verb must match).

Tense. We use the simplified universal definition of tense, which is encoding of the event time. Similar to previous categories, we only account for the categories and the languages that have morphological markers for tense. The most common values for tense across languages in our data set are: Past, Present, and Future. Russian and German examples with Past tense markings are given below for reference.

(7) a.

On	*kupi-l-∅*	*etot*	*dom*
He	buy-Pst-Sg.M	this	house

'He bought this house.'

b.

Auf	*dem*	*Tisch*	*lag-∅*	*ein*	*Buch*
On	the.Dat	table	lie.Pst-Sg	a	book

'There was a book on the table.'

Tense expresses the temporal order and the factuality of the events and states, and is therefore expected to contribute to inference and time-based NLP problems.

Voice. This study is only concerned with frequently occurring Active and Passive voice features that have separate morphological markers in the verb. A synthetic German example using passive voice is given below. As shown, the semantic roles of *he* (`Agent`[3]) and *house* (`Product`) are encoded differently depending on the voice of the main verb.

(8) a.

Er	*baut*	*das*	*Haus*	*.*
He.Nom	build.Act	the	house.Acc	

'He builds the house.'

b.

Das	*Haus*	*wird*	*von*	*ihm*	*gebaut*	*.*
The	house.Nom	is	by	he.Dat	build.Pass	

'The house is built by him.'

Because voice affects the encoding of the core semantic arguments, the ability of word embedding methods to represent voice information is expected to contribute to the dependency parsing and semantic role labeling and induction performance.

Tag Count. We create a test that contains tuples of surface forms and number of morphological tags (annotated according to UniMorph schema) for the token. It can be considered a simplistic approximation of the morphological information encoded in a word and is expected to cover a mixture of the linguistic aspects outlined above. For instance the Turkish word *deneyimlerine* (to their/his/her/your experiences) annotated with (N.Dat.Pl.Poss2Sg) would have the tag count of 4, whereas *deneyimler* (experiences) annotated with (N.Dat.Pl) would have the count 3. It can also be associated with the model's capability of segmenting words into morphemes, namely, **morphological segmentation**, especially for agglutinative languages like Turkish where morpheme to meaning is a one-to-one mapping. Furthermore, for fusional languages with one-to-many morpheme to meaning mapping, it can be associated with the model's ability to learn such morphemes with multiple tags as in the Spanish word *hablo*-V.Ind.1Sg.Prs (I speak), where "o" alone conveys the information about the mood, tense, and the person.

Character Bin. Here we create a test set consisting of pairs of randomly picked surface forms and the number of unicode characters they contain. For convenience, we used bins instead of real values as in Conneau et al. (2018a). The motivation behind this

[3] As per VerbNet 3.3, `https://verbs.colorado.edu/verb-index/vn3.3/`.

feature is to use number of characters as an approximation to number of morphological features, similar to previously motivated *Tag Count* test. We hypothesize that this should be possible for agglutinative languages where there is one-to-one mapping between morpheme and meaning, unlike the mapping in fusional languages. *Character Bin* can therefore be seen as a rough approximation of Tag Count with the advantage of being able to expand this resource to even more languages since it does not require any morphological tag information.

Pseudowords. Pseudowords, or nonwords, are commonly used in psycholinguistics to study lexical choices or different aspects of language acquisition. There are various ways to generate pseudowords, for example, randomly swapping two letters, randomly adding/deleting letters to/from a word; or concatenating high-frequency bigrams or trigrams. When it comes to multilingual studies, these methods have limitations such as computational time, availability of resources, and researcher's bias, as explained in detail by Keuleers and Brysbaert (2010). In this study, we use the "Wuggy" algorithm (Keuleers and Brysbaert 2010), which is the most commonly used and freely available system for multilingual pseudoword generation. It builds a grammar of bigram chains from the syllabified lexicon and generates all possible words with the grammar, both words and nonwords. It is available for German, Dutch, English, Basque, French, Spanish, and Vietnamese by default, and has been extended for Turkish (Erten, Bozsahin, and Zeyrek 2014). Some examples of generated pseudowords from our data set are given in Table 2. Because the Wuggy algorithm can generate words that sound natural, this test can be used to distinguish subword-level models that can capture semantic-level information from the ones that remain on ortography-level.

SameFeat. We choose two surface forms that share only one feature and label this form pair with the shared (same) feature. Some examples are shown in Table 3. Because features depend on the language, the number of labels and the statistics of the data set differ per language. The ability to detect shared morphological features is expected to contribute to the encoding of agreement.

OddFeat. This test is the opposite of the shared feature test. We prepare pairs of surface forms that differ only by one feature value and label them with this odd feature. Some examples are given in Table 4. Although these contrastive features are not directly linked to any simple linguistic property, we hypothesize that they can be valuable assets to compare/diagnose models for which it is important to learn

Table 2
Examples of generated pseudowords.

Language	Pseudowords
English	atlinsive, delilottent, foiry
French	souvuille, faicha, blêlament
Basque	zende, kontsiskio, anazkile, kaukasun, kaldretu
Dutch	nerstbare, openkialig, inwrannees, tedenjaaigige, wuitje
Serbian	aćejujelu, benkrilno, knjivule, haknjskim, znamaketi
German	Anstiffung, hefumtechen, Schlauben, Scheckmal, spüßten
Spanish	vuera, espisia, supencinzado, lungar, disciscir
Turkish	ular, pesteklelik, çanar, tatsazı, yalsanla

Table 3
Examples of form pairs with *only* one shared feature. *Poss3Pl*: possession by third plural person, *Poss1Sg*: possession by first singular person. Shared features shown in **bold**. Turkish positive polarity is not explicitly tagged by Unimorph. *TR*: Turkish, *RU*: Russian, *DE*: German.

L	form1	form2	SameFeat
TR	yalvaracaksınız *beg* (V.2Pl.Fut)	onadı *approve* (V.3Sg.Pst)	Polarity
TR	yolculuklarına *travel* (N.Poss3Pl.**Dat**)	düşmanıma *enemy* (N.Poss1Sg.**Dat**)	Case
TR	taşımam ***carry*** (V.1Sg.Prs.Neg)	taşıdılar ***carry*** (V.3Pl.Pst)	Lemma
TR	sarımsaklarım *garlic* (N.Pl.**Poss1Sg**.Nom)	cümlemde *sentence* (N.Sg.**Poss1Sg**.Loc)	Possession
RU	pantera *panther* (N.Nom.**Sg**)	optimisticheskogo *optimistic* (Adj.Gen.**Sg**)	Number
DE	Stofftiere *stuffed_animal* (N.**Nom**.Pl)	Tennisplatz *tennis_court* (N.**Nom**.Sg)	Case

Table 4
Examples of form pairs with *only* one different feature. Odd features shown in **bold**. Turkish positive polarity is not tagged by Unimorph. *Poss2sg*: possession by second singular person, *Poss1Sg*: possession by first singular person. Odd features shown in **bold**. Turkish positive polarity is not explicitly tagged by Unimorph. *TR*: Turkish, *RU*: Russian, *ES*: Spanish, *DE*: German.

L	form1	form2	OddFeat
TR	istemeyecek *want* (V.3Sg.Fut.**NEG**)	isteyecek *want* (V.3Sg.Fut)	Polarity
TR	seçenekler *option* (N.**Nom**.Pl)	seçeneklere *option* (N.**Dat**.Pl)	Case
TR	iyileşiyorlardı ***heal*** (V.3Pl.Pst.Prog)	geziyorlardı ***travel*** (V.3Pl.Pst.Prog)	Lemma
TR	deneyimlerine *experience* (N.Dat.Pl.**Poss2Sg**)	deneyimlerime *experience* (N.Dat.Pl.**Poss1Sg**)	Possession
RU	zashitu *defence* (N.**Acc**.Sg)	zashite *defence* (N.**Dat**.Sg)	Case
ES	legalisada *legalized* (V.Sg.Ptcp.**F**)	legalisado *legalized* (V.Sg.Ptcp.**M**)	Gender
DE	integriert ***integrate*** (V.3sg.Ind.Prs)	rechnet ***count*** (V.3Sg.Ind.Prs)	Lemma

the commonalities/differences between a pair of tokens, such as question answering, or natural language inference tasks.

3.2 Data Set Creation

In this section, we introduce the methodology for creating the type-level probing tasks, that is, tasks where surface forms are probed without the context. Afterwards, the creation process for token-level probing tasks (i.e., where surface forms to be probed are provided within a context) is described. The focus of our study is on type-level tasks;

however we provide a set of similar token-level tasks for comparison and discussion of future work.

3.3 Type-Level Probing Tasks

When searching for a data set from which to source the probing tests from, the number of languages this data set covers is of key importance. Although there is only a small number of annotated *truly* multilingual data sets, such as Universal Dependencies, unlabeled data sets are more abundant such as Wikipedia[4] and Wiktionary.[5] For type-level probing tasks, we use UniMorph 2.0 (Kirov et al. 2018), which provides a data set of inflection paradigms with universal morphology features mapped from Wiktionary for many of the world's languages. In addition to UniMorph, we use the lexicon and the software provided by Wuggy to generate pseudowords (Keuleers and Brysbaert 2010; Erten, Bozsahin, and Zeyrek 2014). Finally, we use word frequency lists extracted from Wikipedia. We follow different procedures to create data sets for each test type. Here, we briefly explain the creation process of single form feature tests such as *Tense*, *Voice*, *Mood*; paired form feature tests: *OddFeat* and *SameFeat*; followed by *Character Bin*, and pseudoword generation via Wuggy.

Single Form Feature Tests. A word annotated with UniMorph features can be used in several probing tests. For instance, the Turkish word *grubumuzdan* [from our groups] is marked with the N.Sg.Poss1Pl.Abl tag and can be used to probe the *POS*, *Case marking*, *Number*, and the *Possession* features because it has the N (Noun), Abl (Ablative), Sg (Singular), and Poss1Pl (Possession by first person plural) tags. While generating the tests, we check whether the following conditions for a language and target feature are satisfied:

- Because we need to train classifiers for the probing tests, we need large enough training data. We eliminate the language/feature pair if total number of samples for that certain feature is less than 10K.[6]
- If a feature (e.g., *case marker*), does not have more than one unique value for a given language-feature pair, it is excluded from the tests.

In addition, we perform two additional preprocessing steps: (1) removal of ambiguous forms with respect to linguistic feature, and (2) partial filtering of the infrequent words. Ambiguity is one of the core properties of the natural language, and a single word form can have multiple morphological interpretations. For instance, the German lemma *Teilnehmerin* would be inflected as *Teilnehmerinnen* as a plural noun marked either with accusative, dative, or a genitive case marker. We remove such words with multiple interpretations for the same feature. This is a deliberate design choice we make, which, although potentially causing some systematic removals for certain tasks such as German case, substantially simplifies the task architecture and guarantees fair testing. The ambiguity ratios are discussed in more detail in Section 3.5.

[4] `https://www.wikipedia.org/`.

[5] `https://www.wiktionary.org/`.

[6] In our preliminary experiments, we found that 10K is large enough to provide sufficient clues for the linguistics classifier to predict linguistic labels; and small enough to cover as many languages as possible.

UniMorph data set contains many grammatically correct but infrequent word forms such as the English “transglycosylating” or the Turkish “satrançlarımızda” [in our chesses]. To make sure that our probing tests are representative of language use, we utilize the frequent word statistics extracted from the Wikipedia dump of the corresponding language. For each probing test, the data set is compiled so that 80% of the forms are frequently encountered words. We keep a portion of “rare” words[7] and use a considerably larger proportion of frequency dictionary (e.g., we keep the first 1M words for Russian) to identify frequent words in order to keep our tests domain-independent, hence not provide any unjust advantage to embedding models trained on Wikipedia. Finally, we introduce surface forms of the “None” class, that is, forms that do not contain that test feature. For instance if the *Tense* feature is probed, the 30% of the probing data set contains nominal forms that are from the “None” class. Most NLP downstream tasks need to distinguish between a “None” class and other class labels. For instance, an SRL model needs to decide whether a token is an argument of a predicate; a dependency parser needs to decide whether two tokens can be connected by a dependency relation; or a NER needs to predict if a token is part of a named entity or not. We believe this setting provides a more realistic probing task scenario compared with having only the positive examples of a given linguistic feature.

Paired Feature Tests. Unlike for single features, we did not remove ambiguous forms for paired features tests, that is, *OddFeat* and *SameFeat*, due to the restrictive nature of the tests. For instance, while probing for the *OddFeat* between two forms, we assume that there exists a word pair differing only by one feature. Therefore, here we only consider one certain interpretation of the word form, which would share $n - 1$ features with the interpretation of the other form, where n is the total number of UniMorph features in both words.

The data set for this test type is created in two separate steps: (1) for unimorph tags and (2) for lemmas. For the *SameFeat*, we first group the words that contain the feature of interest together for the step (1). Then we split each feature group into two and sample $k = 500$ words from both groups. These word pairs are compared against each other, and included in the test set if they share the same value for the feature of interest, but differ in all other features. Because some features are tagged by default (e.g., POS), we exclude these features from the comparison process. Otherwise our data set would have no instances, since, for example, all nouns share the “N” tag. In addition to POS tags, we exclude the *Mood* feature from Finnish and Turkish, and *Interrogativity* feature from Turkish, since all verbs in UniMorph data share the same tag for those features. For (2), we follow the same steps, but check whether the lemma values are the same and others are different.[8]

While preparing the data set for the *OddFeat*, we first group the words by *lemma* tagged with the target feature for the step (1). Then we randomly sample elements from each lemma group, and perform pairwise comparison. If two sampled forms have different values for the feature (e.g., Ablative and Locative) but have the same set of values for the other features (e.g., Singular), then they are assigned this feature as the label. In addition to the features with different values, we also consider the features

[7] We have followed the 80% frequent versus 20% rare word ratio. In some exceptional cases where we can not choose 80% of the words from frequency dictionary due to a small Wikipedia, we allow a larger portion of rare words to have at least 10K instances.

[8] We perform similar preprocessing and data set balancing for all languages. The details of parameter values can be found on the project Web site.

that are not explicitly tagged. For instance, if only one of the forms has the *Possession* feature, but all features except *Possession* are shared among these two forms, we create a test pair with the value *Possession*. To generate the test pairs for the step (2), we group the words by their feature sets, namely, different forms with the exact same set of feature values will be clustered together. Then we split each group into two, and sample $k = 100$ number of forms from both halves. This procedure results in unbalanced data sets, usually dominated by the *Number* feature. In order to avoid this, we sample proportionally from such overly sized feature test pairs.

Character Bin. After removing the ambiguous forms, we created bins of numbers for character counts, since the variation was high. We used the following bins for character counts: [0–4, 5–8, 9–12, 13–16, 17–20, >20]. We applied the same bins for all languages.

Pseudo Word Test. Finally, we generated pseudo words for 9 languages. To do so, we first sampled 10K in-vocabulary words from the lexical resources provided by Wuggy. We then use those words as seeds for the Wuggy generator, and generate pseudowords by setting the maximum number of candidates per word to 5, maximal search time per word to 5 seconds; and restricting the output to match the length of sub-syllabic segments, match the letter length, match transition frequencies, and match two-thirds of sub-syllabic segments.

The sets of languages for each probing test introduced in Section 3.1 are given in Table 6. In total, we have created 15 probing tests for 24 languages, each containing 7K training, 2K development, and 1K test instances.

3.4 Token-Level Probing Tasks

Type-level probing has several advantages: It's compact and less prone to majority and domain shift effects. However, because downstream NLP tasks mostly operate on full-text data, decoupling evaluation from running text might result in less realistic performance estimates; besides, it limits the evaluation of contextualized word representations and black-box models. To investigate the limitations and the strengths of type-level tasks, we prepare a set of comparable token-level probing tasks using the modified Universal Dependency (UD) treebanks where the Morphosyntactic Descriptions have been converted to the UniMorph schema (McCarthy et al. 2018). Contrary to the type-level tasks, we do not filter out any infrequent or ambiguous surface forms; and we do not introduce a "None" class for convenience. Because the data set is annotated with the same schema as in our type-level tasks, we simply adapt our existing source code that creates single form feature tests (e.g., *Tense*, *Case*) for token-sentence pairs. Similar to the single form type-level tasks, if the total number of samples for a certain feature is less than 10K; or if a feature (e.g., *case marker*) has only one value, we exclude that feature-language pair from the tests. The created tests have the sentence, word index, and feature label information. As an example, the following line is taken from the Person-English test-language pair: *"Looks good," 0, Third person Singular*; meaning that the word at index 0 in the given sentence (*"Looks"*) has the THIRD PERSON SINGULAR label.

Following Section 3.3, we have created the same single form feature tests for all available languages with the modified UD treebank; each containing 7K training, 2K development, and 1K test instances. The current version of the token-based suite only contains category-based, morphological tests, although it can be easily extended for other probing tasks: OddFeat, SameFeat, TagCount, and CharacterBin.

3.5 Discussion on Probing Task Types

Properties and quality of the token-level probing tasks are strongly tied to the properties and the quality of resources used while creating them. To provide more insights, Table 5 provides essential statistical information for type- and token-level task sets, focusing on the languages we experiment with (explained in Section 4.1) later in this work.

Data Set Size. The agglutinative languages, Finnish and Turkish, have a higher amount of instances for type-level tasks than token-level tasks. This is due to their productive morphology that enables generating large amounts of surface forms from a single lemma. One can observe that for all fusional languages, the number of tokens in a token-level resource exceeds the number of forms available from a type-level resource, while agglutinative languages follow an opposite trend. This is due to practical reasons: Unimorph is based on Wiktionary data, and because of their agglutinative nature Finnish and Turkish allow easier generation of word forms to populate the paradigms, whereas fusional languages require manual annotation. At the same time, Russian, German, and Spanish are higher-resourced languages that offer large-scale treebanks from which the UD treebank data has been sourced.

Data Domain and Token Frequency. Type-level tasks are induced from a dictionary based resource (Wiktionary), while token-level tasks are based on existing language-specific treebanks. For instance the largest Turkish treebank (UD-IMST) is collected from daily news reports and novels, whereas the biggest Finnish treebank is a collection of manually annotated grammatical examples. Token-level tasks based on running text—especially given that treebanks are often based on a homogeneous document collection—are inevitably biased to the domain of this text, whereas type-level probing tasks are expected to be domain-neutral. In particular, dictionary-based tasks do not contain any frequency information of the surface forms, whereas token-level tasks do. Although the frequency information may be helpful in some cases (e.g., when the domains of the downstream tasks and the probing tasks are similar) it would also add a bias regarding the distribution of specific features. A token-level test would be penalized less for misclassifying rare forms, and the probing classifier might benefit from using majority class information that might depend on the domain (e.g., singular nouns are more frequently observed than plural nouns).

Table 5
Statistics for the resources used during the creation of type and token-level probing tasks. $|\pi|$: Number of inflection paradigms, #form: Number of inflected forms, N: Noun, V: Verb, A: Adjective, amb%: Ratio of ambiguous forms, #sent: Number of sentences, #token: Number of tokens, $|sent|$: Average sentence length, V (%): Vocabulary size (#token/V).

	Type-level				Token-level				
	$\lvert\pi\rvert$	#form	types	amb%	#sent	#token	$\lvert sent\rvert$	V (%)	amb%
Finnish	57K	2.5M	N, V, A	**4.87**	31K	339K	10.81	83K *(4.07)*	**17.62**
Turkish	3.5K	275K	N, V, A	**7.76**	6K	67K	11.25	22K *(3.06)*	**19.28**
Russian	28K	474K	N, V, A	**12.51**	63K	1.1M	17.89	135K *(8.29)*	**23.75**
German	15K	179K	N,V	**25.92**	14K	263K	18.74	49K *(5.39)*	**27.47**
Spanish	5.5K	383K	V	**10.75**	30K	883K	29.12	68K *(13.04)*	**35.1**

Data Quality. Wiktionary is a collaborative effort, while the UD treebanks are mostly annotated by a handful of experts. Although we cannot find an exact measure on the accuracy of Wiktionary data, a data set with a large number of collaborators may have fewer annotation artifacts than a data set created by a few experts. On the other hand, both data sets may have been affected negatively from an automatic conversion process, mostly due to converting language-specific features to universal tags.

Lexical Variety. Whereas Turkish, Finnish, and Russian type-level UniMorph data have all lexical classes, German data does not include any adjectives. Furthermore, Spanish only has verb inflections that limit the scope of probing. On the other hand, treebanks for the token-level tests are based on running text in which all lexical classes are represented.

Ambiguity. In Table 5 we list the average number of ambiguous forms (i.e., forms that might be expressing more than one morphological feature bundle) for both task types. For type-level, we have averaged the ambiguity ratios over all probing tasks. For token-level tasks, we simply calculate the ratio of surface forms with different morphological tag sets to all surface forms. We notice that for agglutinative languages, where we have one-to-one morpheme to meaning mapping, the ambiguity ratios for both type- and token-level tests are lower. For fusional languages the ambiguity ratios are higher, mostly due to syncretism—one word form might encode several morphological feature bundles. In particular, for German, the average is around 26% (before removal of ambiguous entries), which means loss of considerable amount of data points. In general, token-level tests have higher ambiguity ratios, although because the tokens are provided within the context, it enables models to resolve the ambiguity.

Use Cases. Despite their similarities, type- and token-level probing tasks differ in terms of potential use cases and limitations. From the representation perspective, type-level tasks are better suited for probing context-free word embeddings (i.e., static or subword-level); whereas token-level tasks are more suitable for contextual embeddings (due to having many duplicate training and test instances when tokens are isolated from context). Token-level tasks can be used as a diagnostic probing tool for any downstream model layer that doesn't require any additional task-specific inputs (e.g., part-of-speech tags for dependency parsers, predicate flags for SRL). Type-level tasks, on the other hand, are more suited to diagnose the initial word encoding layer that generates a type of word representation in isolation, not the intermediate hidden layers that require contextual information.

Summary. In summary, both token- and type-level probing test designs come with certain implications, and the choice of the probing set depends on the task at hand. Type-level probing tasks have the advantage of containing less bias (domain, annotator, and majority class); whereas token-level tests might be sensitive to the domain biases from the underlying full-text data. On the other hand, token-level tests have the advantage of being more lexically diverse; whereas type-level tasks can be less diverse for some languages like Spanish, French, and English. In terms of data set sizes and the number of languages that can be covered, both type- and token-level probing tests are similar. Finally, type-level tasks are better suited to probe traditional word embeddings or initial word encoding layers that do not require contextual information; whereas token-level tasks are more suitable for probing contextual embeddings or intermediate layers if the layers do not require additional linguistic information.

4. Evaluation Methodology

In this section, we discuss our probing task evaluation methodology. First of all, because of the large number of languages and embedding models available, we choose a subset of each. We describe how we decide on the languages to evaluate on in Section 4.1. Next, in order to investigate the relation between probing and downstream tasks, we evaluate a set of diverse multilingual embedding models intrinsically via our probing tasks, as explained in Section 4.3, and extrinsically on several downstream tasks discussed in Section 4.4, and investigate the correlations between the corresponding task performances. Finally, in Section 4.5 we show how the proposed probing tests can be used as a diagnostic tool for black box NLP systems in a case study.

4.1 Languages

We have identified a list of languages to test our hypotheses on various research questions such as the relation between downstream and probing tasks or the information encoded in layers of black box models. For this we have considered the following criteria:

- Chosen languages should have relatively broad resource coverage (e.g., annotated data for a variety of downstream tasks),
- The set of chosen languages should have a high coverage of probing tests; and the number chosen languages should be in proportion to the number of languages that are probed for a certain test, and
- The languages should be as typologically diverse as possible in terms of linguistic properties we are probing for.

Considering these points for in-depth experimentation we have selected five languages—German, Finnish, Turkish, Spanish and Russian, which are shown in color in Table 6. Most of them have annotated resources in addition to UD treebanks, for example, data sets created for named entity recognition (NER), Natural Language Inference (NLI), and semantic role labeling (SRL). As can be seen from Table 6, all probing tests are covered and their ratio to other languages is well proportioned for each test. Our selected languages belong to diverse language families, namely, from Germanic, Uralic, Turkic, Romance, and Slavic; and are typologically diverse, that is, have representatives from agglutinative (Finnish and Turkish) and fusional (German, Spanish, Russian) languages.

4.2 Multilingual Embeddings

Following Aggarwal and Ranganathan (2016), who discussed the need for having *diverse* and *heterogeneous* samples to conduct a correlation study, we have picked multilingual embeddings that are trained with different objectives, architectures, and units; and avoided using similar models trained with a slightly different hyperparameter (e.g., word2vec trained with the same settings except dimensionality) to avoid having subclusters in our samples. Namely, in this work we experiment with the following

Table 6
List of languages for each probing task. Languages shown in colored cells are the languages we experiment on. General refers to POS, Tag Count, and Character Bin. Some of the tests with fewer numbers of languages are concatenated vertically for convenience.

CASE	MOOD	NUMBER	General	PERSON	POLARITY	TENSE	ODD FEAT	SAME FEAT
arabic	arabic	armenian	arabic	arabic	portuguese	armenian	armenian	arabic
armenian	armenian	catalan	armenian	armenian	turkish	bulgarian	czech	armenian
bulgarian	catalan	finnish	bulgarian	catalan	**POSSESSION**	catalan	finnish	bulgarian
czech	finnish	french	catalan	finnish	armenian	finnish	german	catalan
estonian	french	german	czech	french	quechua	french	hungarian	czech
finnish	german	hungarian	danish	german	turkish	german	macedonian	danish
german	hungarian	italian	estonian	hungarian	**VOICE**	hungarian	greek	dutch
hungarian	italian	macedonian	finnish	italian	arabic	italian	polish	estonian
macedonian	polish	polish	french	macedonian	bulgarian	macedonian	portuguese	finnish
greek	portuguese	portuguese	german	greek	finnish	greek	quechua	french
polish	romanian	russian	hungarian	polish	russian	polish	romanian	german
quechua	serbian	spanish	italian	portuguese	serbian	portuguese	russian	italian
russian	spanish	swedish	macedonian	quechua	swedish	quechua	serbian	macedonian
serbian		**GENDER**	greek	romanian	**PSEUDO**	romanian	spanish	greek
swedish		arabic	polish	russian	basque	russian	swedish	polish
turkish		bulgarian	portuguese	serbian	dutch	serbian	turkish	portuguese
		macedonian	quechua	spanish	english	spanish		quechua
		greek	romanian	turkish	french	turkish		romanian
		polish	russian		german			russian
		portuguese	serbian		serbian			serbian
		russian	spanish		spanish			spanish
		serbo	swedish		turkish			swedish
		spanish	turkish		vietnamese			turkish

word embedding models: word2vec (Mikolov et al. 2013a); fastText (Bojanowski et al. 2017); GloVe with Byte Pair Encoding (GloVe-BPE; 2016); supervised MUSE (Lample et al. 2018); and ELMo (Peters et al. 2018).

word2vec Among the selected representations, only *word2vec* uses word as the basic unit. We have trained a word2vec model for each of the selected languages on the latest preprocessed (tokenized, lowercased) Wikipedia dump using 300-dimensional CBOW, a window of size 10, and minimum target count of 5. We have used the implementation provided by the authors.[9]

fastText provides word representations that have subword-level information learned from character *n*-grams. In simple terms, words are represented as a linear combination of the character *n*-gram embeddings of the token's character *n*-grams. We use the embeddings distributed by fastText,[10] which are trained on preprocessed Wikipedia using CBOW with position-weights, in dimension 300, with character *n*-grams of length 5 and a window of size 5.

GloVe-BPE is another type of subword-level embedding that uses unsupervised morphological segments generated by a compression algorithm inspired by Gage (1994). We use the pretrained embeddings by Heinzerling and Strube (2018), which are trained on preprocessed Wikipedia using GloVe (Pennington, Socher, and Manning 2014). We use the Python wrapper open sourced by the authors[11] with default dictionary size of 10K and dimension 300. Because the tool provides embeddings for each segment, in case of multiple segments per token, we used the averaged vector as the word representation.

[9] https://code.google.com/archive/p/word2vec/.
[10] https://fasttext.cc/.
[11] https://github.com/bheinzerling/bpemb.

MUSE-supervised embeddings are crosslingual fastText embeddings. These embeddings are generated by aligning the monolingual fastText embeddings in a common space (in our case English) using ground-truth bilingual dictionaries. We used the aligned and mapped vectors distributed by the authors.[12] The crosslingual embeddings have the same technical properties as the *fastText* vectors described above. Because the authors only release the static embedding vector without the model, we could not generate embeddings for OOV words.

ELMo embeddings are computed on top of two-layer bidirectional language models that use characters composed using convolutional neural networks (CNN). Unlike previously introduced embedding models, ELMo provides *contextualized* embeddings, that is, the same words would have different representations when used in different contexts. However, our probing tests are type-level (as opposed to token-level), thus we only use the representations generated independently per each token both for the intrinsic and extrinsic experiments. In the scope of this study, ELMo embeddings are treated as powerful pretrained character-level *decontextualized* vectors. To highlight this important detail, we further refer to our ELMo-derived embeddings as *Decontextualized ELMo (D-ELMo)*. We use the multilingual pretrained ELMo embeddings distributed by the authors (Che et al. 2018; Fares et al. 2017),[13] which are trained with the same hyperparameter settings as the original (Peters et al. 2018) for the bidirectional language model and the character CNN. They are trained on randomly sampled 20 million words from Wikipedia dump and Common Crawl data sets and have dimensionality 1,024. We use the three-layer averaged ELMo representation for each word.

For all the experiments described in Section 4.4 and Section 4.3, we first created the vocabulary for all intrinsic and extrinsic data sets per language. Then, we generated the vectors using the embeddings that can handle OOV words, namely, *fastText*, *GloVe-BPE*, and *D-ELMo*, for each language-intrinsic and language-extrinsic pair. The static embeddings *word2vec* and *MUSE* are used as provided. Hence, for the models using these embeddings, each unknown (OOV) word is replaced by the UNK token and the same vector is used for all UNK words.

4.3 Intrinsic Evaluation: Probing Tasks

Following Conneau et al. (2018a), we use diagnostic classifiers (Shi, Padhi, and Knight 2016; Adi et al. 2017) for our main probing tests. Our diagnostic classifier is a feedforward neural network with one hidden layer, followed by a ReLU non-linearity. The classifier takes as an input a fixed trained word vector and predicts a particular label specific to the probing test. For *OddFeat* and *SameFeat*, because the input consists of two words, we first concatenate both word vectors before feeding them into the feedforward network. For all tests, we use the same hyperparameters: 300 hidden dimension and 0.5 dropout rate. We train each model for 20 epochs with early stopping (patience = 5). The input dimension vector depends on the type of pre-trained word vectors that will be evaluated. Our evaluation suite is implemented using the AllenNLP library (Gardner et al. 2018).

[12] `https://github.com/facebookresearch/MUSE`.

[13] `https://github.com/HIT-SCIR/ELMoForManyLangs`.

4.4 Extrinsic Evaluation: Downstream Tasks

We consider five tasks for our extrinsic evaluation: universal POS-tagging (POS), dependency parsing (DEP), named entity recognition (NER), semantic role labeling (SRL), and crosslingual natural language inference (XNLI). The former two tasks are useful to measure correlation of our probing test sets to downstream syntactic tasks, and the latter three provide insight into the performance on more semantic tasks. Because our main goal is to evaluate the quality of the *pre-trained* word embedding spaces, we neither update the word vectors during training nor use extra character-level information. Except for SRL, all tasks described below are trained using the models implemented in AllenNLP library.

POS Tagging. This is a classic sequence tagging task, where the goal is to assign a sequence of POS tags given the input sentence. We use data from the UD project version 2.3 (Nivre 2018), and adopt universal POS tags as our target labels. For the tagging model, we use a bidirectional LSTM encoder with 300 hidden units and 0.5 dropout. We use Adam optimizer with initial learning rate 0.001. We train each model with mini-batch size of 32 for 40 epochs, with early stopping (patience = 10). We use the accuracy as our performance metric. It must be noted that the POS-tagging downstream task is different from the POS probing task: Probing is a single-item, type-level classification task using a simple MLP classifier, whereas extrinsic POS is a sequence tagging task utilizing a more powerful Bi-LSTM architecture and operating on sentence level.

Dependency Parsing. The aim of dependency parsing is to predict syntactic dependencies between words in a sentence in the form of a tree structure. This task is especially interesting because of its deep interaction with morphology, which we will evaluate in our probing tests. We use a deep biaffine parser of Dozat and Manning (2017), which is a variant of graph-based dependency parser of McDonald, Crammer, and Pereira (2005). The parsing model takes as input a sequence of token embeddings concatenated with the corresponding universal POS embeddings. The input is then processed by a multilayer biLSTM. The output state of the final LSTM layer is then fed into four separate ReLU layers to produce four specific word representations: two for predicting the arcs (*head predictions*) and another two for predicting the dependency label (*label prediction*). The resulting four representations are used in two biaffine classifiers, one predicting the arc and another one to predict a dependency label, given a dependent/head word pair. For our experiments, we use 2-layer biLSTM with 250 hidden units, POS embedding dimension 100, and ReLU layer (for arc and label representations) with dimension 200. We train the model with mini-batch size of 128 for 30 epochs, and perform early stopping when the Label Attachment Score on the development set does not improve after five epochs.

Named Entity Recognition. The goal of this task is to label the spans of input text with entity labels, for example, *Person*, *Organization*, or *Location*. Unlike POS tagging, NER annotates text spans and not individual tokens; this is usually represented via a (Begin, Inside, Outside) BIO-like encoding. We use a standard NER architecture, a BiLSTM-CRF model where the output of BiLSTM is processed by a conditional random field to enforce global sequence-level constraints (Huang, Xu, and Yu 2015). We use a 2-layer BiLSTM with 200 hidden units and 0.5 dropout trained for 20 epochs with patience 10; the performance is measured via span-based F1 score.

Semantic Role Labeling. Semantic Role Labeling (SRL) is the automatic process of identifying predicate–argument structures and assigning meaningful labels to them. An SRL-annotated sentence with the predicate sense "buy.01: purchase" is shown below.

[Mark]$_{\text{Arg0: Buyer}}$ [bought]$_{\text{buy.01}}$ [a car]$_{\text{Arg1: Thing bought}}$ from [a retailer store]$_{\text{Arg2: Seller}}$

We consider the dependency-based, that is, CoNLL-09 style, PropBank SRL, where the goal is to label semantic argument heads with semantic roles. We use the subword-level end-to-end biLSTM based sequence tagging SRL model introduced by Şahin and Steedman (2018). It can either use pretrained embeddings as word representations, or learn task specific subword-level (character, character-*n*-gram, morphology) representations by composing word vectors via a separate bi-LSTM network. Here, we only used pretrained word embeddings concatenated with a binary predicate flag (1 if the token is predicate, 0 otherwise) and two layers of bi-LSTMs with 200 hidden dimensions on top of these representations. Finally, tokens are assigned the most probable semantic role calculated via the final softmax layer. Weight parameters are initialized orthogonally, batch size is chosen as 32, and optimized with stochastic gradient descent with adaptive learning rate initialized as 1. Gradient clipping and early stopping with patience 3 are used. We use the standard data splits and evaluate the results with the official evaluation script provided by CoNLL-09 shared task. We report the role labeling F1 scores.

Natural Language Inference. The NLI task aims to extract the relations such as Entailment, Neutral, and Contradiction between a pair of sentences—a *hypothesis* and a *premise*. This objective has been formerly addressed in the scope of the Recognizing Textual Entailment (RTE) task that used the resources provided by RTE challenge tasks which had a small size.[14] Later, a larger data set, also known as the Stanford Natural Language Inference (SNLI; Bowman et al. 2015) data set, which has been compiled from English image caption corpora and labeled via crowdsourcing, has been introduced. Some example pairs of sentences are shown in Table 7. As stated by Bowman et al. (2015) and also can be seen from Table 7, a high-performing NLI model should handle phenomena like tense, modality, and negation, which are mostly covered by our probing tasks.

MultiGenre NLI (MultiNLI; Williams, Nangia, and Bowman, 2018) is a recent data set that covers a wider variety of text styles and topics. The Crosslingual NLI (XNLI; Conneau et al. 2018b) data set has been derived from MultiNLI and is used as a benchmark for evaluating crosslingual sentence representations. This evaluation benchmark originally aimed at testing the models trained for the source language (English), on the target language, and covers 15 languages including Spanish, Turkish, Russian, and German. It should be noted that the development and test splits for each language in XNLI have been translated by professional translators. The authors also release the automatic translation of MultiNLI training split which they use to align the crosslingual sentence embeddings. Since the multilingual embeddings used in this study are not all crosslingual, here we train a separate monolingual NLI model for each language by using the automatic translation data. We use the Enhanced LSTM model (ESIM; Chen et al. 2017) with default parameters provided by AllenNLP framework. This model uses a sequential inference based on chain (bidirectional) LSTMs with attentional input encoding, enhanced with syntactic parsing information. In our experiments, we use

[14] https://aclweb.org/aclwiki/Textual_Entailment_Resource_Pool.

Table 7
Example sentence pairs taken from Williams, Nangia, and Bowman (2018).

Premise	Hypothesis	Label
Met my first girlfriend that way.	I didn’t meet my first girlfriend until later.	Contradiction
I am a lacto-vegetarian.	I enjoy eating cheese too much to abstain from dairy.	Neutral
At 8:34, the Boston Center controller received a third transmission from American 11	The Boston Center controller got a third transmission from American 11	Entailment

pre-trained word embeddings to represent both hypothesis and premise tokens. These embeddings are kept fixed during training (not updated).

4.5 Diagnostic Evaluation: A Case Study on SRL

Another proposed application of our probing tests is to diagnose the layers of a black-box NLP model. In order to do so, we used the same SRL model as described in extrinsic evaluation (see Section 4.4). This time, instead of using pretrained embeddings, we used randomly initialized character trigram embeddings. The model generates intermediate word representations by summing the weighted forward and backward hidden states from the character trigram bi-LSTM network. As the model is trained with a negative log likelihood loss for semantic roles, it is expected to learn character trigram embeddings and other model parameters that are better suited for SRL. In order to diagnose whether it *does* indeed extract morphologically relevant information during training, we save the model states for different epochs and generate the word representations via the aforementioned internal biLSTM layer and use our intrinsic evaluation suite from Section 4.3, to evaluate these representations. As preprocessing, all tokens are lowercased and marked with start and end characters. One layer of bi-LSTMs both for subword composition and argument labeling with hidden size of 200 are used. Character trigrams are randomly initialized as 200-dimension vectors. The other hyperparameters are kept the same as in Section 4.4.

5. Experiments and Results

In this section, we first discuss the data sets used for our intrinsic and extrinsic experiments. We then provide the results and briefly discuss the general patterns and exceptions observed in both experiments. It is important to note that our primary goal is to compare the performance of embedding model *instances*, and not of the embedding models per se: Given that the performance of a particular trained instance might depend on a variety of factors such as dimensionality, preprocessing details, and underlying textual corpora, a claim that a certain embedding method (e.g., word2vec) outperforms another embedding method in general would be far fetched and would fall out of the scope of our current study. Instead we provide a toolkit that allows one to empirically investigate the performance of the embedding spaces, which are by themselves treated as black box.

Table 8
Sources and statistics of our extrinsic data set. The NER data sets for Turkish and Russian are down-sampled.

Task	Language	Source	Number of tokens			OOV%	
			train	dev	test	dev	test
POS Tagging	Finnish	Finnish-TDT	162.6K	18.3K	21.K	22.99	22.3
Dependency Parsing	German	German-GSD	263.8K	12.5K	16.5K	9.67	10.76
	Russian	Russian-SynTagRus	870.5K	118.5K	117.3K	8.44	8.68
	Spanish	Spanish-AnCora	444.6K	52.3K	52.6K	4.92	4.91
	Turkish	Turkish-IMST	37.9K	10.K	10.K	24.14	23.04
NER	German	Germeval-2014 (Benikova et al. 2014)	452.9K	41.7K	96.5K	11.34	11.29
	Russian	WikiNER (Ghaddar and Langlais 2017)	169.1K	55.4K	55.2K	16.65	16.75
	Turkish	TWNERTC (Sahin et al. 2017)	272.1K	91.3K	90.9K	14.48	14.97
	Spanish	CoNLL-2002 (Sang and De Meulder 2003)	264.7K	52.9K	51.5K	7.43	5.63
	Finnish	FinNER	180.1K	13.6K	46.4K	18.9	19.7
SRL	Finnish	Finnish PropBank (Haverinen et al. 2015)	162.7K	9.2K	9.1K	22.77	23.05
	German	CoNLL-09 (Hajič et al. 2009)	648.7K	32.K	31.6K	8.43	8.69
	Spanish	CoNLL-09 (Hajič et al. 2009)	427.4K	50.4K	50.6K	6.06	6.16
	Turkish	Turkish PropBank (Şahin and Adalı 2018)	44K	9.7K	9.3K	22.79	21.82
XNLI	German	XNLI (Conneau et al. 2018b)	13.7M	77.1K	156.K	5.46	5.57
	Russian		12.3M	70.9K	143.7K	7.61	7.75
	Spanish		13.8M	81.8K	165.2K	3.17	3.15
	Turkish		10.4M	62.4K	126.6K	10.15	10.3

5.1 Data Set

For intrinsic evaluation, we use the probing data sets that have been described in Section 3.2, and experiment with the five languages: Finnish (Uralic), German (Germanic), Spanish (Romance), Russian (Slavic), and Turkish (Turkic), as discussed in Section 4.1. For POS tagging and dependency parsing, we use data sets from UD version 2.3 (Nivre 2018). For the NER data set, the Turkish and Russian data are substantially larger than the other languages. For practical reasons and fair comparison, we randomly sample 5% to 8% of the subsets of the original data sets and split them into train/dev/test sets. The details of each UD treebank and other extrinsic data set sources along with their statistics are presented in Table 8.[15]

5.2 Results

We first provide the results of intrinsic and extrinsic experiments for the type-level probing tasks. Later, we provide a comparison between token-level and type-level tests using the results of intrinsic experiments.

5.2.1 Results on Type-Level Probing Tasks. We present the type-level probing test results of the multilingual embeddings introduced in Section 4.2 for each language/test pair in Table 9. In addition, we report the baseline scores calculated with majority voting baseline for each language/test pair. According to Table 9, the majority of the tests had

[15] Finnish NER data is available from `https://github.com/mpsilfve/finer-data` and the article "A Finnish News Corpus for Named Entity Recognition" where the data set is described is reported to be under review.

Table 9
Type-Level Probing task results for all languages. **Bold** represents the best score, and *italics* represents the second best.

Finnish

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Case	30.0	49.3	59.9	83.4	*86.6*	**96.7**
Mood	50.0	62.3	67.9	84.7	*89.0*	**93.8**
Number	45.6	60.4	69.4	83.4	*90.3*	**97.4**
POS	67.9	75.3	70.3	85.7	*90.0*	**97.1**
Person	30.1	54.0	66.8	84.6	*88.8*	**94.6**
Tense	40.9	65.4	73.4	86.0	*90.6*	**94.7**
Voice	50.8	63.4	70.8	86.8	*89.6*	**95.1**
CharacterBin	44.2	45.0	44.8	52.0	*58.4*	**63.8**
TagCount	86.0	88.6	87.0	91.0	*95.0*	**98.4**
OddFeat	22.7	24.4	24.5	65.1	*76.7*	**88.4**
SameFeat	29.1	94.1	92.0	*96.9*	96.5	**98.4**

German

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Case	34.2	62.0	68.7	90.9	**95.1**	*94.0*
Mood	37.4	54.3	54.1	90.1	*91.0*	**93.9**
Number	40.1	60.4	66.8	90.7	*93.7*	**97.7**
POS	55.8	63.1	65.8	92.2	*94.9*	**96.9**
Person	52.9	65.2	60.3	90.4	*91.5*	**95.8**
Pseudo	50.0	96.7	80.1	83.2	*90.0*	**91.0**
Tense	52.9	73.1	71.5	91.5	*92.9*	**93.2**
CharacterBin	45.4	49.0	45.0	*63.0*	62.9	**70.4**
TagCount	54.9	61.5	63.1	83.0	*86.5*	**89.2**
OddFeat	22.6	37.9	34.8	65.1	*71.2*	**75.4**
SameFeat	28.4	84.5	86.5	*89.6*	**90.4**	89.0

Spanish

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Gender	34.5	67.0	74.5	98.0	*98.8*	**99.8**
Mood	52.0	67.0	66.1	89.2	*90.9*	**95.0**
Number	34.0	69.2	69.9	*95.0*	*95.0*	**99.8**
POS	70.9	85.6	84.1	97.6	*98.5*	**99.6**
Person	27.4	60.9	52.8	*92.6*	87.8	**98.6**
Pseudo	49.8	*92.3*	89.4	75.9	91.9	**94.7**
Tense	39.9	59.1	60.8	*87.1*	85.9	**95.0**
CharacterBin	50.9	55.2	55.3	*72.3*	69.6	**76.2**
TagCount	40.0	61.0	59.0	*90.8*	87.8	**95.8**
OddFeat	44.8	53.4	55.8	77.1	*78.5*	**81.7**
SameFeat	27.2	89.6	89.1	*91.1*	**93.3**	*91.1*

Russian

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Case	31.0	57.3	78.0	*80.8*	62.0	**96.7**
Gender	39.8	57.7	78.3	*95.4*	80.7	**99.3**
Number	41.1	54.7	75.7	*89.7*	74.3	**96.9**
POS	48.4	56.5	67.8	*89.7*	74.2	**98.2**
Person	31.9	49.4	72.2	*93.0*	81.0	**96.7**
Tense	43.8	56.3	73.6	*90.1*	73.6	**94.3**
Voice	47.6	62.2	66.5	**99.4**	96.1	*99.0*
CharacterBin	46.0	46.3	52.5	*68.9*	64.4	**70.9**
TagCount	53.8	60.4	68.5	*85.2*	67.9	**96.4**
OddFeat	21.8	36.9	48.2	*74.4*	55.4	**90.0**
SameFeat	29.4	84.7	90.9	*93.9*	93.6	**97.6**

Table 9
Continued.

			Turkish			
Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Case	31.1	63.5	57.4	*87.1*	85.4	**96.1**
POS	75.5	85.9	83.5	*95.9*	94.8	**98.4**
Person	30.3	52.5	52.0	*93.5*	90.5	**96.1**
Polarity	44.6	62.0	61.0	**97.3**	93.6	*96.1*
Possession	30.6	59.2	56.7	*87.1*	75.5	**92.5**
Pseudo	51.5	*90.3*	90.2	71.4	79.6	**91.7**
Tense	34.9	57.7	58.8	*89.4*	85.4	**94.7**
CharacterBin	46.1	58.1	53.6	*66.7*	*66.7*	**71.5**
TagCount	46.6	71.4	60.7	*85.6*	79.9	**89.8**
OddFeat	38.7	38.5	40.6	76.7	**79.8**	*79.0*
SameFeat	21.3	73.9	74.7	*86.9*	**90.0**	86.5

a baseline score under 50%, although some language/test pairs had higher baselines because of the data set properties such as lacking annotations for certain tags. These tests are *POS* for Finnish, Spanish, and Turkish and *TagCount* for Finnish. In addition, *SameFeat* and *OddFeat* have relatively low baseline scores consistently across languages, generally followed by *Case*.

Table 9 shows that all embedding models investigated in this work achieved their lowest score for *CharacterBin*. As none of our embedding models use characters as basic units (except for D-ELMo which uses character-level CNN), it might be difficult for them to predict the number of characters from the surface form alone. However, we note that models that use subword units such as fastText, Glove-BPE, and D-ELMo in general obtain better performance than models with words as basic units (word2vec and MUSE). In order to assess the difficulty of the tests, one can calculate the gap between the average performance of the embeddings and the baseline scores. A small gap points to a "hard-to-beat" majority vote baseline. After eliminating the tests with high baseline scores, we observe that the majority of the tests have seen improvements ranging between 50% to 200%, although recall their low baseline scores.

First of all, for probing tests we observe that all embeddings outperform the baseline for all tasks and languages. Apart from a few cases, we see that *D-ELMo* achieves the highest scores in probing for all language-test pairs, generally followed by *fastText* and *GloVe-BPE*. There are several factors that can explain why *D-ELMo* achieves the highest scores. First of all, *D-ELMo* models are trained on a different text source (subset of Wikipedia combined with CommonCrawl). Second, D-ELMo had additional training objectives compared to other traditional language modeling ones. Third, unlike other embedding models, D-ELMo is the only model with a layer operating on the character-level. It has been shown several times in separate studies (Şahin and Steedman 2018; Vania and Lopez 2017) that character-level models perform better than other subword-level models on a number of downstream tasks. Fourth, it has dimensionality of 1,024 whereas other models are of dimension 300. Finally, it may be a combination of all properties explained above. A careful investigation of the exact property of D-ELMo that grants it an advantage falls outside of the scope of this study. For the languages Finnish, Russian, and Turkish, *D-ELMo* outperforms the other embeddings by a larger margin compared with Spanish and German. *fastText* and *GloVe-BPE* perform similarly, except from Russian where *GloVe-BPE* achieves significantly higher scores than *fastText* in almost all tests, which could be due to the segmenting mechanism enabled by the BPE that can capture the morphological boundaries better than *n*-gram based *fastText* given the highly fusional nature of Russian morphological marking.

Our intrinsic experiment results show that the probing task performance and the improvement compared to the majority baseline differs depending on the language and the task, signaling that not all languages and morphological categories are equally easy to model. There exist several ways to quantitatively capture morphological complexity, for example, a recent work by Cotterell et al. (2019) plots the **morphological counting complexity** (MCC) of the languages (defined as the number of cells in a language's inflectional morphological paradigm) against a novel entropy-based *irregularity* measure to empirically demonstrate the hypothesized bound on the two complexity types: Although a language can have a large paradigm or be highly irregular, it's never both. While paradigm-based counting complexity cannot be applied to the probing tests directly because of their categorical nature, one can use the number of unique values in a respective category as a rough approximation of the complexity of this category. For instance, a weak correspondence can be seen between the number of values and the baseline performances for *Case* test—the fewer cases a language has, the higher the baseline. German with 4 cases has the majority baseline of 34.2, while Finnish with 15 cases has 30.0, as given in Table 9. However, this pattern vanishes as we move to the embedding-based models: The end performance does not seem to depend on the number of case values, for example, *D-ELMo* performs equally well for Russian (6 cases) and Finnish (15 cases). This can indeed be related to the trade-off between the number of inflection paradigms and the irregularity, discussed by Cotterell et al. (2019). The regularity of the language (e.g., Finnish) may help the embedding models to learn the patterns and lead to even higher final scores than irregular languages (e.g., German, Russian) despite much lower baseline scores due to a larger number of paradigms.

We observe that the static embeddings, *word2vec* and *MUSE*, which do not have a dedicated mechanism to flexibly represent OOV words, performed similarly and had lower scores than other embedding models for most of the tests, except from *Pseudo*. Especially *MUSE* has an outstanding performance on *Pseudo* tests, compared with its performance on other tests. This is not unexpected: In the case of *Pseudo*, the OOV handling mechanism of static embeddings, which maps unseen words to the same entry, puts static models at advantage because they encode all unknown words with a single random vector, making the detection of explicit OOV items easier.

We present the results of the extrinsic experiments in Table 10. The general performance ordering of the embeddings: *D-ELMo*, *fastText*/*GloVe-BPE*, *word2vec*/*MUSE* holds for syntactic (POS, DEP) and shallow semantic tasks (SRL) for all languages, similarly to the ranking in intrinsic experiments. However, for NER and XNLI tasks, we do not observe the same trend. There might be several reasons for this discrepancy. First of all, the extrinsic tasks at hand are conceptually different: Whereas grammar-based POS, DEP, and SRL directly build upon subword information (e.g., via agreement), for NER lexical content and surface cues play a bigger role, while XNLI as a semantic task benefits from lexical information and proposition-level cues like negation, tense, and modality, rather than general subword-level phenomena. Hence, the lexical differences between the training corpora of multilingual embeddings and downstream tasks may

Table 10
Downstream tasks results for all languages. **Bold** represents the best score, and *italics* represents the second best.

			Finnish		
Task	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
SRL	62.30	57.68	60.41	*64.19*	**72.26**
DEP	79.62	79.84	80.6	*82.45*	**87.78**
POS	89.56	89.86	89.88	*92.55*	**96.56**
NER	72.96	71.17	75.69	**80.54**	*78.45*
			German		
Task	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
SRL	55.25	60.60	57.11	*61.75*	**61.85**
DEP	82.43	*82.78*	82.32	83.20	**83.46**
POS	91.82	92.14	90.59	*92.66*	**93.57**
NER	74.32	*76.13*	71.43	**78.35**	71.81
XNLI	44.03	40.08	43.55	**44.69**	*44.05*
			Spanish		
Task	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
SRL	64.49	62.78	62.34	*66.39*	**70.03**
DEP	90.17	90.26	89.99	*90.55*	**91.09**
POS	96.07	*96.58*	95.66	96.49	**97.43**
NER	77.48	**79.31**	77.36	*78.96*	77.75
XNLI	*46.75*	41.28	45.17	**46.80**	45.07
			Russian		
Task	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
DEP	90.13	*90.54*	90.16	87.41	**92.26**
POS	95.62	*96.11*	95.91	92.61	**97.84**
NER	78.38	**79.92**	75.84	64.20	*79.71*
XNLI	43.43	39.80	*43.53*	41.64	**45.05**
			Turkish		
Task	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
SRL	53.29	46.35	*53.51*	53.14	**63.38**
DEP	*57.82*	56.67	55.92	57.70	**62.97**
POS	86.52	87.35	*87.57*	86.80	**94.48**
NER	48.87	*52.21*	51.75	**52.52**	49.22
XNLI	42.79	42.93	**45.17**	*44.25*	43.81

have been more emphasized for these tasks, especially NER.[16] Another potential reason for the difference in ranking is the domain of the data underlying the respective data sets: For the majority of the languages, POS, DEP, and SRL data originates from the same treebanks and has gold (expert) annotations. On the other hand, NER and XNLI

[16] Unlike the others, *D-ELMO* has been induced from a *subset* of Wikipedia: We hypothesize that it is enough to learn good representations for grammatical phenomena in common words, but not enough to populate the entity vocabulary. Given that most NER data sets are Wikipedia-based, this could lead to lower entity vocabulary intersection compared with the other embedding spaces, and thereby to lower scores. Testing this hypothesis, however, is not trivial without access to the exact source corpora and a reliable method to identify entities in them.

data sets are generally compiled from a different, and often diverse set of resources. A third reason may be the different OOV ratios among different data sets. In order to investigate this, we have calculated the OOV ratio of development and test sets of each extrinsic task with respect to the training set, shown in Table 8. We observe that the XNLI task has the lowest OOV ratio among all other extrinsic tasks for all languages. Similarly, when OOV ratios of extrinsic tasks with respect to our static embeddings (MUSE and word2vec) are examined (shown in the Appendix), we notice that both embeddings have the lowest OOV ratio for XNLI task. These statistics could indeed explain the smaller gaps between static and subword-level models for the XNLI task. Finally, NER annotations for most of the experimented languages are of silver quality, that is, there exist many incorrect and missing labels; and the multilingual sentences provided in XNLI are automatically translated by an existing tool.

We observe that static word embedding spaces (*word2vec* and *MUSE*) rank generally higher on downstream tasks compared with the probing tasks for fusional languages (German, Spanish, Russian). We attribute this to the vocabulary difference between the extrinsic and intrinsic data sets. As mentioned in Section 3.2, our type-level probing data contains many word forms that rarely occur in Wikipedia main text, which is the primary text source for the vector space models we compare. The extrinsic data sets, on the contrary, are derived from Wikipedia and newswire, resulting in a higher lexical overlap, lower unseen word rate, and therefore better performance. We have calculated the OOV rates of intrinsic and extrinsic tasks, relative to both word2vec and MUSE embeddings and found that OOV rates for our probing tasks are indeed much higher than our extrinsic tasks supporting our hypothesis. The OOV rates are given in the Appendix.

5.2.2 Results on Token-Level Probing Tasks. In order to investigate the token-level probing tasks even more deeply, we apply the same experimental setup described in Section 4.3 to the token-level probing test suite and present the results in Table 11. Because the majority of the embedding spaces used in this study (see Section 4.2) are not contextualized (i.e., *have the same representation for the surface form independent from its surrounding words*), we only use the token itself without its context. That means that when tokens are isolated in such a way, there may be duplicates among training, development, and test sets. Therefore the results for MUSE, word2vec, GloVe-BPE, and fastText are only provided for comparison among each other, and to gain insights on some of the aspects discussed in Section 3.5. Finally to provide a more realistic use-case for the token-level probing task, we experiment with the ELMo embeddings without decontextualizing them, referred to as contextualized ELMo (C-ELMo).

To categorize our findings for token/type-level probing tasks, we use some of the aspects from our previous discussion in Section 3.5.

Data Set Size. We observe that having a smaller data set size for Turkish token-level probing tasks eliminated the possibility to probe for the "Possession" feature. We suspect that it may occur for other relatively low-resourced languages, leaving us with only the most common, generic tasks.

Tag Frequency. As discussed previously, constructing tasks on an annotated corpus may introduce biases toward frequently encountered feature values in the data set. When the gap between the majority baseline scores is examined, it can be seen that for certain features (e.g., Mood and Number) the gap is in the range of 40–60%. For instance, the "Polarity" feature for Turkish has 89% majority voting score, meaning that 89% of the instances had the "Positive" label.

Table 11
Token-Level probing task results for all languages. **Bold** represents the best score, and *italics* represents the second best.

Finnish

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo	C-ELMo
Case	35.9	79.7	69.1	85.0	97.3	*97.9*	**98**
Mood	89.7	94.9	94.3	96.7	*97.4*	97.3	**98.3**
Number	82.2	92.8	90.4	94.4	97.8	*98.3*	**98.7**
POS	29.5	69.0	72.0	68.9	71.3	*74.8*	**87.5**
Person	64.5	92.8	89.7	95.1	**97.3**	*96.8*	96.4
Tense	62.9	95.2	92.8	97.5	**98.4**	**98.4**	*98.1*
Voice	86.2	96.2	92.7	96.5	*97.7*	**98.0**	97.5

German

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo	C-ELMo
Case	32.7	56.9	52.9	53.0	*56.8*	53.4	**80.8**
Gender	38.6	72.6	69.3	66.5	*73.5*	72.8	**78.7**
Mood	96.4	98.3	98.4	98.4	**98.9**	*98.7*	**98.9**
Number	79.5	88.5	88.1	86.1	*89.5*	89.2	**94.7**
POS	20.6	76.5	82.6	74.5	76.4	*77.1*	**91.2**
Person	72.8	94.6	94.1	93.8	*95.1*	*95.1*	**97.9**
Tense	51.6	98.7	97.8	98.7	98.6	**99.2**	*98.8*

Spanish

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo	C-ELMo
Gender	58.0	98.6	95.3	97.9	99.2	*99.3*	**99.4**
Mood	92.3	97.7	97.2	95.9	**99.1**	*98.2*	97.8
Number	75.2	98.7	96.4	98.6	**99.6**	**99.6**	**99.6**
POS	19.3	78.1	83.2	78.0	*79.6*	78.9	**92**
Person	68.4	97.4	95.3	98.1	**99.2**	*99.0*	*99.0*
Tense	55.4	98.2	97.1	98.7	**99.3**	*99.2*	99.1

Russian

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo	C-ELMo
Case	31.7	75.3	47.6	68.2	78.6	*78.7*	**89.7**
Gender	45.4	86.0	64.7	87.0	**89.9**	*89.8*	89.7
Number	75.3	93.9	78.2	92.8	96.8	*97.0*	**97.4**
POS	30.3	76.4	64.5	73.3	76.5	*76.9*	**95.4**
Person	51.1	91.8	84.3	98.1	*98.6*	*98.6*	*98.8*
Tense	53.0	90.6	78.4	*97.6*	**98.7**	97.2	96.9
Voice	69.5	88.9	79.0	**95.6**	93.7	*95.4*	94.5

Turkish

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo	C-ELMo
Case	47.7	84.3	67.7	89.0	91.4	*95.4*	**96.6**
POS	35.4	71.7	68.0	75.4	*78.3*	77.4	**83.2**
Person	78.3	91.0	84.3	95.0	**97.2**	*96.8*	96.5
Polarity	89.0	95.4	92.3	98.3	**99.2**	*99.0*	98.8
Tense	55.2	87.7	76.3	94.6	**96.6**	95.9	*96.3*

Token Frequency. Earlier we hypothesized a token frequency bias when using a full-text based probing test. When we compare the performance gaps of embedding models between type and token-level tasks, we observe a substantial performance boost for the static embeddings: (MUSE and word2vec, for all languages and tasks apart from few exceptions) whereas we observe smaller gaps or performance drops for subword-based dynamic models (BPE, fasttext, and D-ELMo). The performance boost of static embeddings is again related to a lower OOV ratio in the token-level data sets.

Lexical Variety. This effect is visible from the results of the "POS" feature, compared with type-level "POS" test. First, the token-level POS baseline scores are noticeably lower; and second, all embedding spaces including D-ELMO achieve much lower scores.

Ambiguity. Token-level probing unlocks a few more probing tasks such as "Gender" for German, for which we did not have access before due to eliminating ambiguous forms. More importantly, we observe that removing ambiguous forms may have introduced a sort of bias toward some of the features, that is, simplified the task by eliminating certain feature values that always produce ambiguous surface form. This effect can be easily observed in the performance gaps between D-ELMo and C-ELMo. In other words, when a certain feature gets a performance boost by C-ELMo, this may suggest that the feature is highly ambiguous and a model that can use contextual information to resolve ambiguity outperforms the one that can't by a large margin. The following feature–language pairs demonstrate the described phenomena: German-Case, German-Number, German-Gender, and Russian-Case—which had the highest ambiguity ratios as discussed in Section 3.5. Apart from these cases, we see a similar pattern for the "POS" feature; however, this feature is also affected by the limited lexical variety of type-level tasks. Therefore, there are two phenomena responsible for the performance boost for "POS."

Finally, when embeddings are compared among each other (except from C-ELMo due to its having many duplicate training and test instances in token-level tests when tokens are isolated from context), we see a similar ranking for each language–feature pair, suggesting that type- and token-level probing tasks have many commonalities despite their differences discussed above.

6. Analysis

In this section we investigate the relation between downstream and the probing tasks more closely, and report the results with respect to language families and downstream tasks. We present the results for the diagnostic case study described in Section 4.5 and show the close connection to highly correlated probing tests. Finally, we give a brief summary of our findings related to proposed probing tasks.

6.1 Correlation

In order to calculate the relation between the downstream tasks and the probing tests, we calculate the Spearman correlation coefficient as shown in Figure 1. In addition, we calculate the two-sided p-values to test the null hypothesis, namely, whether two sets of results are uncorrelated, and interpret the results with respect to the languages and the tasks.[17]

[17] Significant correlations are given in Appendices.

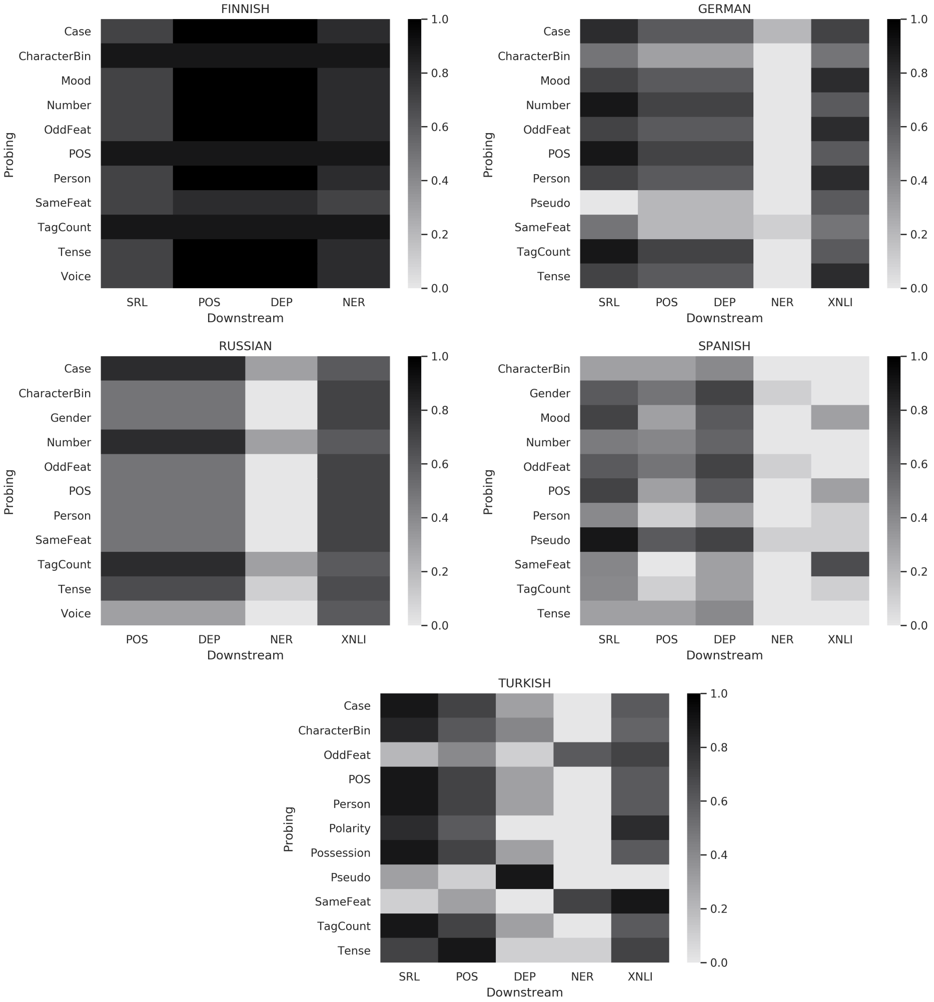

Figure 1
Spearman correlation between probing and downstream tasks for each language.

6.1.1 Language-Related Findings.

Finnish. We observed the highest correlations with a p-value of 0.1 in Finnish language.[18] According to the calculated p-values, all proposed tests, except from *SameFeat*, had a statistically significant correlation with POS, DEP, SRL, and NER for Finnish. As already shown in Table 5, in type-level statistics columns, Finnish data had the lowest ambiguity ratio, highest number of surface forms and paradigms, and the highest lexical diversity, which leads to strong correlations to downstream tasks. Furthermore, being an agglutinative language with high MCC but lower irregularity, it allows encoding

[18] Because the number of samples, i.e., number of embeddings, for the correlation analysis are 5, we use a high p-value of 0.1.

considerable amount of syntactic-semantic information on type-level, which is another explanation for strong correlations of the proposed tasks.

Turkish. For Turkish, we found strong correlations for all single feature tests for syntactic tasks (except from DEP), and registered relatively high correlation between *Polarity*, *SameFeat*, *OddFeat*, *Tense*, and XNLI task. Although Turkish is typologically similar to Finnish and has a similar degree of MCC and irregularity, we observe weaker correlations for POS and inconclusive correlations for DEP. According to Table 5, the data that Turkish probing tasks are originated from are only around 10% of Finnish data, with a slightly higher ambiguity ratio. Although the lexical variety in terms of word classes is similar, smaller data size and more importantly smaller number of paradigms (i.e., less variety encoded on type-level) may have influenced the correlation scores negatively. Finally, the respective treebanks of both languages have been sourced from different domains. While the Finnish treebank is based on a grammar book, Turkish data is a combination of domains ranging from news data to children stories. It would not be unexpected to have higher correlations between tasks from similar domains as in Finnish.

German. For German, we observed high correlation with a p-value of 0.1 for *Number*, *POS*, and *TagCount* tests, whereas *Case*, *Mood*, *OddFeat*, *Person*, and *Tense* have statistically significant correlation with a p-value of 0.2 for SRL. For German, the correlation pattern of *Case*, *Number*, *POS*, and *TagCount* repeated for syntactic and shallow semantic tasks: POS, DEP, and SRL, whereas XNLI correlated well with *Case*, *Mood*, *OddFeat*, *Person*, and *Tense*. We observe that the correlating tasks are similar to those of agglutinative languages in general (except from *CharacterBin*—explained below), however weaker. The weaker correlations may be the result of the highly ambiguous nature of German data (especially *Case* and *POS*), and less lexical diversity, *which are both common among fusional languages*, as shown in Table 5. In addition, the paradigm sizes for German nouns and verbs are only 29 and 8 (for reference Turkish has 120 and 100 respectively) (Cotterell et al. 2018).

Russian. For Russian, we find that *Case*, *Number*, and *TagCount* have high correlations with a p-value of 0.2 for syntactic tasks, similar to German, whereas XNLI correlated better with the other features such as *SameFeat*, *OddFeat*, *Person*, and *Tense*. Russian is among fusional languages like German, with similar MCC and irregularity values. One exception is the high irregularity of Russian verbs (with a score of 1.67) compared with German verbs that have a score of 0.77 (Cotterell et al. 2018). This could explain the weaker correlations for verb-related probing tasks such as *Person* and *Tense*, as can be seen more clearly in Figure 1.

Spanish. For Spanish, there was no clear correlation pattern, except from the *Pseudo* test that had a strong correlation to SRL and DEP extrinsic tasks, with $p = 0.2$. Lack of correlations can be attributed to the lack of lexical variety in Spanish probing tasks discussed in Section 3.5. In addition, Spanish is one of the languages with the highest gap of OOV ratio between its extrinsic and intrinsic data sets. For instance, MUSE has an OOV ratio of 2.71% in training split of Spanish dependency treebank, while having a minimum of 48% for the probing tasks as given in the Appendix. This could lead to unnatural performance gaps between static and other embeddings, and this can affect the correlation.

Summary. We observed that a large number of probing tasks had high correlation to syntactic tasks especially for *agglutinative* languages: Turkish and Finnish. This result can be connected to several of our previous observations discussed in Section 3.5, namely, regarding the *lexical variety* and the *ambiguity ratio*. In Section 3.5, we demonstrated that the probing tests for these languages had the richest lexical variety with the coverage of nouns, verbs, and adjectives; and had the lowest ambiguity ratio mostly due to the property of one morpheme encoding only one morphological feature. These observations suggest that the instances in the proposed tests are satisfactory representatives of the language. Apart from these observations, another reason is the amount of information encoded by morphological features on the type-level, hinted at by the regularity of the languages (Cotterell et al. 2019). As shown in Cotterell et al. (2018), Finnish and Turkish have much lower irregularity scores compared with the other languages we have experimented with. It can be interpreted as: The more regularity a language has, the more information can be incorporated into a "single unit/word" without introducing more ambiguity, which makes it easier to capture more syntactic information in a single form.

On the other hand, we found a smaller set of probing tasks with high correlation for *fusional* languages, especially for syntactic tasks. This can be connected to the amount of syntactic information that can be encoded in a word, which can be linked to the number of paradigms. For instance, whereas Turkish has 120 noun paradigms, German and Russian only have 29 and 25, respectively (Cotterell et al. 2018), suggesting that the amount of information encoded in nouns are indeed limited. In addition, lack of *lexical variety* (having only verbs) has been observed to have the biggest impact on Spanish, which caused the weakest correlations among all languages. Furthermore, we observed that a set of common probing tests have higher correlation to certain downstream tasks among most languages, such as *Case*, *POS*, *Number*, and *TagCount* to syntactic tasks; and *OddFeat*, *SameFeat*, *Tense* to XNLI; we haven't detected any strong correlation to NER for almost all languages. Case is among the features that signals syntactic and semantic connection between nominals and verbs, as discussed in Section 3.1; hence it has been expected to correlate well with syntactic and shallow semantic tasks. *POS* is an obvious syntactic feature, whereas *Number* is not. However, *Number* is among the common grammatical agreement features and provides clues for linking syntactically related words that should agree on number (e.g., linking subject and verb in dependency parsing). The result of the *TagCount* test suggests that simplistic approximation of morphology may be a good predictor for syntactic task performance. We also note that there are cases where a correlation was hypothesized but never observed or has been found weak, such as simple morphological feature probing tasks, for example, *Case*, *Tense*, *Number*, and Turkish dependency parsing. Similarly, in some cases, a correlation is found although not hypothesized, such as Turkish NER correlating well with *SameFeat* feature. These suggest that there is a certain amount of noise in the correlation measurements that may be a result of many different factors such as small number of data points (embedding models). Although our results reveal certain patterns, obtaining the data points (language - extrinsic score - probing score) is computationally expensive, which limits the precision of the correlation tests.

6.1.2 Downstream Task-Related Findings.

SRL. For all languages, SRL is found to correlate with the highest number of probing tasks. This finding is intuitive because SRL performance is dependent on more complex linguistic phenomena compared with other tasks. Regardless of the languages families, we find that SRL has high correlations with *Case* and *POS*, generally followed by *Person* and *Tense* tests. This finding is on par with the traditional language-independent features used for SRL back in the feature-engineering days (Hajič et al. 2009). In addition to those tests, for agglutinative languages, we find high correlation for *CharacterBin* and *TagCount* (explained later in this section). In addition, SRL has high correlation to *Possession* and *Polarity*, which only exists for Turkish. As discussed in Section 3.1, *Possession* provides a link between the possession of the object and person of the verb. This is highly relevant to SRL, which aims to detect the arguments of the predicates, hence it may help especially for the argument identification subtask. *Polarity* can be directly linked to SRL due to its having labeled negation arguments (*ArgM-NEG*). We see that *Mood* is a common highly correlated test for fusional languages, whereas *Number* only correlates with German SRL, and *Pseudo* only correlates for Spanish SRL. *Mood* can be considered highly relevant to SRL for several reasons. First, an imperative predicate generally implies that the subject (which usually undertakes the Agent or Patient role) is missing. Furthermore, other Mood values such as *conditional* provide valuable links between the main predicate and the dependent clause, which is also labeled as an argument *(ArgM-PRD)*.

POS and DEP. They can be considered easier tasks compared with SRL, where superficial linguistic cues would be enough to decide on local classes. However, these cues are not expected to be distinct from SRL, rather, they are a subset of it. Confirming this, for German we see that the set of highly correlated features are reduced to the subset *Case*, *Number*, *POS*, and *TagCount*. Another hint to support this hypothesis is the decline in the correlation scores of *CharacterBin* and *TagCount* in POS and DEP for agglutinative languages. This finding suggests that instead of a feature that distantly approximates the morphological features of a given word such as *TagCount*, a feature focused on a single linguistic phenomenon has higher correlation with more syntactic tasks.

NER. Except for Finnish NER, none of NER tasks had significantly high correlations to our probing tests. Whereas POS, DEP, and SRL represent different levels of grammatical analysis that are correlated with morphological phenomena, NER is a surface-level semantic task for which the lexical content of the target and surrounding tokens is by far more important than the morphological markers evaluated by our probing tests. The observed correlations to morphological probing tests are therefore weak and irregular among the languages.

XNLI. For XNLI, we observe a noticeable pattern consistently among almost all languages, which is high correlation to *Mood*, *Polarity*, *Tense*, and *Person*, which is in agreement with the original study by Williams, Nangia, and Bowman (2018); and high correlation with one of our paired tests (usually *SameFeat*) that resembles the NLI task, in a way that both tasks aim to capture the commonalities between a pair of tokens. However, our probing tasks mostly capture morphological commonalities and differences, which might only constitute a subset of the phenomena relevant for NLI. As discussed in Section 5.2, the largest overlap between the vocabulary of static embeddings is to the vocabulary of XNLI task, which leads to the low OOV ratios and smaller performance gaps between static and subword-level models for the XNLI task. This may have caused an unfair shift in rankings and, subsequently, the correlations; this deserves a separate dedicated study.

Furthermore, we notice that *CharacterBin* and *TagCount* are redundant tests for agglutinative languages. This is due to these languages having one to one morpheme/tag mapping, which suggests that the number of characters is also a good indicator for number of tags. Because other languages are fusional, namely, exhibit a one to many relation between morphemes and tags, these tests do not relate to each other, as can be seen from the correlation matrices of German and Russian, for which the *TagCount* has high correlation scores unlike the *CharacterBin*.

6.2 Diagnostic Task

In this section, we demonstrate the results and analysis of the diagnostic case study for Finnish and Turkish. As described in Section 4.5, we train an end-to-end SRL model that only uses character trigrams as input. We first probe the word encoding layer with the suggested type-level probing tests for three consecutive epochs, where we see a large improvement in SRL F1 scores. We probe Finnish for epochs 2, 8, and 20, which had F1 scores of 29.60, 46.64, and 57.93, respectively. Similarly, we probe the encoding layer of Turkish SRL for epochs 4, 6, and 14, with 15.55, 43.06, and 54.11 F1 scores, respectively. Then we use the token-level tasks to diagnose the encoding layer together with the consequent LSTM layer.

6.2.1 Type-Level Diagnosis. In the case of Finnish, we have seen large improvements on *Case*, *Mood*, *Person*, and *Voice*, while seeing a drop or a constant score for the features *CharacterBin*, *Number*, *POS*, and *TagCount*, as shown in Figure 2. These results suggest that the encoding layer captures more of *Case*, *Mood*, *Person*, and *Voice* information throughout the training for an SRL objective—these probing tests are also found to have significantly high correlation to Finnish SRL in our correlation study. Interestingly, we observe constant or lower scores in correlated features such as *CharacterBin* and *Number*. We note that, even if these tests provide a predictive performance on the SRL task, not all neural models are capable of learning all correlated features discussed in the previous section. This could be due to the lack of capacity of the neural model, or these features getting captured easily during the very early stages of the training. Because the aim of this section is to demonstrate a case study, a thorough comparison and in-depth investigation of the root causes is not in the scope of this work.

Since the F1 improvements are more pronounced for Turkish, we see a more clear pattern in the probing task improvements. Similar to our results for Finnish, we have encountered considerably high improvements for *Case*, *Person*, *Polarity*, *POS*, *Possession*, and *Tense* features, whereas no improvement has been seen for *CharacterBin*, *TagCount*, or *Pseudo*. Again, all tests with increasing scores had been shown to have significantly high correlation to Turkish SRL, whereas *Pseudo* had no significant correlation. Other correlated features with non-increasing scores can be explained similar to the case for Finnish.

6.2.2 Token-Level Diagnosis. We show the results of our diagnostic case study with token-level probing tasks in Figure 3. Probing of the Finnish encoding layer shows that there is a significant performance improvement in *Case* and *POS* features, followed by *Person*, *Tense*, and *Voice* across subsequent epochs. Although the performance boost for *Case* and *POS* are still visible in intermediate layer results, we notice that *Person* and *Tense* features may have been forgotten in the next layer. This suggests that *Case* and *POS* features may be the crucial factors to predict semantic roles since they are conveyed to the next layer.

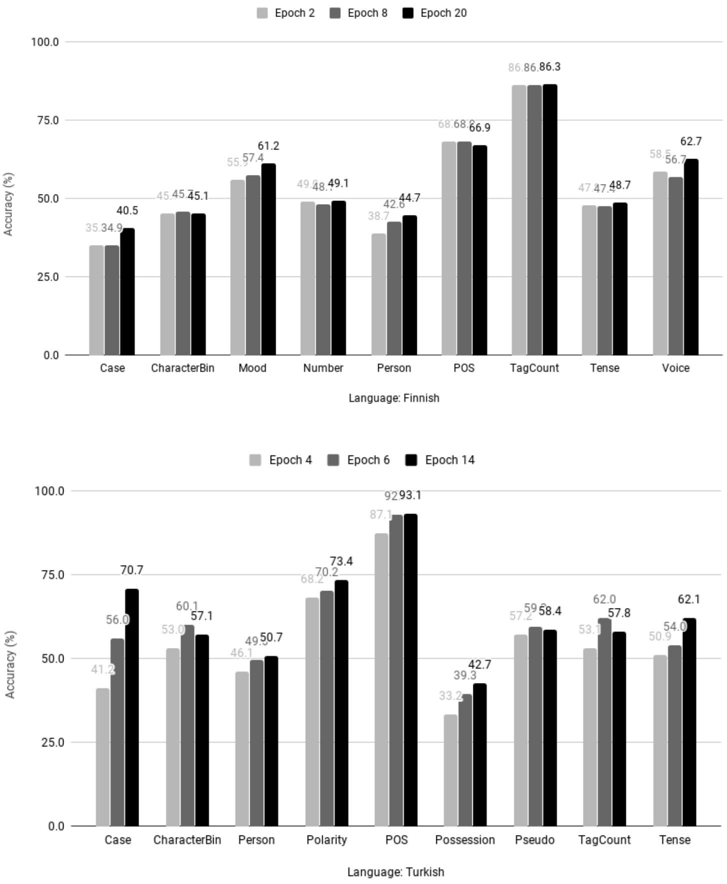

Figure 2
Type-level probing tests to diagnose encoding layer of pretrained Turkish and Finnish SRL models.

We see a similar pattern for the Turkish encoding layer, where there is a rise in accuracy scores for *Case*, *POS*, and *Tense*, followed by modest improvements in *Person* and *Mood* scores. Unlike Finnish, which only transferred the *Case* and *POS* features to the next layer, Turkish seems to convey *Person* and *Tense* features along with *Mood*. Although these features don't seem to be as crucial as the *Case* and *POS* features that have been predominantly used as linguistic inputs to SRL systems, they may provide implicit clues. For instance, a predicate with a first person singular tag is more likely to have an argument with an agent role; whereas a predicate with a third person singular tag is less likely since most predicates in passive voice will be tagged with third person singular. Similarly, a predicate tagged with future tense may be more likely to have a goal argument. In other words, these features are not mutually exclusive, and *cannot be* mutually exclusive.

6.2.3 Discussion on Type- and Token-Level Diagnosis. Next, we examine the word encoding diagnostic results for type and token-level tasks for comparison, starting with Finnish.

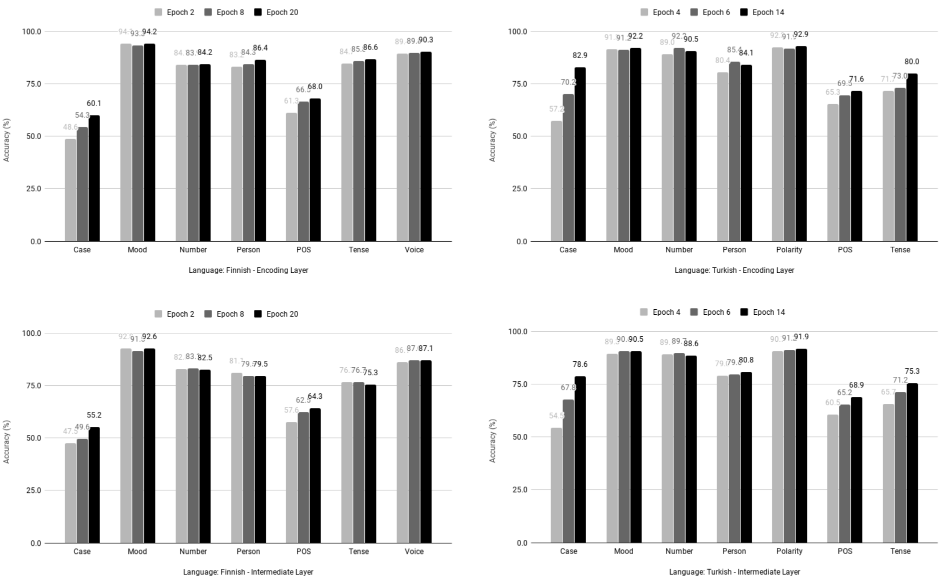

Figure 3
Token-level probing tests to diagnose encoding and intermediate layers of pretrained Turkish and Finnish SRL models. **Top Left:** Finnish-Encoding; **Top Right:** Turkish-Encoding; **Bottom Left:** Finnish-Intermediate; **Bottom Right:** Turkish-Intermediate.

First of all, almost all common features show the same increasing pattern for both languages, hinting at the similarity of both tasks, with exceptions on *Mood* and *POS* for Finnish and *Person* and *Polarity* for Turkish. We hypothesized earlier that token-level tasks may be biased toward frequently occurring tag values. The difference in *Mood* features may be a result of this observation. As discussed in Section 6.1.2, *Mood* can be a distinguishing feature for SRL. Because there are more sentences with "Indicative" predicates, the *Mood* feature in token-level probing will have a bias toward the Indicative tag, resulting with inconclusive trajectory; while a more clear pattern is noticeable in type-level *Mood* feature. Next, we see the opposite for the *POS* feature, where we have a clear pattern in token-level tasks, as opposed to a vague pattern in type-level tests. As we discussed earlier, this may be due to the limited lexical variety in type-level tasks, unlike token-level tasks. When we move on with the Turkish results, we see the same tag frequency effect as in *Mood* for the Finnish pair in *Person* and *Polarity* features. Even though *POS* has a clearer pattern again in token-level diagnostics, the same pattern is still visible in type-level, suggesting that the lexical variety was not as severe as in Finnish case. In order to encourage researchers to conduct similar multilingual diagnostic studies, we have also released a more convenient, online diagnostics platform that uses the proposed probing tasks (Eichler, Sahin, and Gurevych 2019).

6.3 Summary of Experimental Findings

Here we summarize our general findings on the proposed probing tasks from the experiments (see Section 5.2) and analysis (see Section 6.1–6.2.1) sections.

- For all languages, the general ranking of the embeddings is: D-ELMo, fastText/GloVe-BPE, word2vec/MUSE for most type-level intrinsic tasks and the extrinsic tasks POS, DEP, and SRL. Although the same pattern is visible in token-level intrinsic experimental results, these embeddings are mostly surpassed by C-ELMo.
- Static word embedding spaces (word2vec and MUSE) generally rank higher on downstream tasks compared with probing tasks, due to having lower OOV ratios in downstream tasks.
- Low majority voting baseline scores in probing tasks generally mirror the MCC of the language (i.e., the more *Case* categories, the lower the baseline). The trade-off between MCC and the morphological irregularity (i.e., the higher the MCC, the lower irregularity) later yields higher probing accuracies despite the low baselines.
- Type-level probing tasks contain less domain and majority class bias, whereas the statistics of the resource (e.g., tag and token frequencies) have a direct impact on token-level tasks. However, token-level tests are generally more lexically diverse and more convenient to probe contextualized embedding models or intermediate hidden layers of black-box models. Removing ambiguous forms from a type-level probing task mostly results in weaker correlations as revealed by diagnostic and correlation studies; and mostly impacts the fusional languages where one morpheme may correspond to several morphological features. Despite these differences, the same ranking of various embeddings on both probing task types; and similar performance trajectories among epochs of SRL models for some common probing tasks hint at the commonalities between type and token level tasks:
- The calculated correlations were positive.
- The set of correlated tests were generally small for fusional languages, and large for agglutinative languages.
- The set of correlated tests varied with the *complexity* of the downstream task; however, the correlation pattern was common across similar tasks (e.g., *SRL had a large set of correlated tests, while POS tagging has only a subset of it)*.
- The set of correlated tests varied with the *requirements* of the downstream task (e.g., paired tests like *SameFeat* had strong correlation to XNLI, but had weak ties to syntactic tasks).
- We observe commonalities among the correlated probing tests for Finnish, Turkish, German, and Russian. For instance, the correlation between *Case*, *POS*, *Person*, *Tense*, *TagCount*, and the downstream tasks were higher than the other probing tests. This suggests that the findings are transferable, hence the proposed probing tests can be used for other languages.
- We also observe that language-specific tests are beneficial, that is, have significantly high correlation such as *Polarity* for Turkish, and some tests could be impactful for a language family, for example, *CharacterBin* for agglutinative languages, *Pseudo* for Spanish.

- For almost all languages, there is a lack of correlation between the probing tests and NER performance, which we attribute to the shallow, grammar-agnostic nature of the NER task.
- Apart from linguistic properties, data set statistics play a crucial role in the results (e.g., domain similarity for Finnish; lexical variety for Spanish; low OOV ratio for all XNLI tasks) and may add noise to correlation study. When the number of data points is not enough to reliably estimate the correlation, the noise should be interpreted carefully.
- There is a strong connection between the correlated tests and the morphological features captured throughout the epochs of black-box Finnish and Turkish SRL models, suggesting that diagnostics can be a useful application of the probing tasks.

To follow up on our discussion on strong baselines, and the difficulty of the tests from Section 5.2, we find that correlations neither depend on how strong the initial baseline is, nor how low the accuracy scores for this test are. For instance, POS tagging has strong baselines for many languages, however, it is also one of our most correlated tests. Moreover, "hard" tests with low scores, such as *OddFeat* and *CharacterBin*, behaved like any other tests (i.e., had low-to-high levels of correlation for different language/downstream task pairs).

7. Conclusion

In this study we have introduced 15 type-level probing tests for a total of 24 languages, where the target linguistic phenomena differ depending on the typological properties of the language (e.g., *Case*, *Polarity* for Turkish; *Gender* for Russian and German). These tests are proposed as an exploratory tool to reveal the underlying linguistic properties captured by a word embedding or a layer of a neural model trained for a downstream task. Furthermore, we introduce a methodology for creation and evaluation of such tests that can be easily extended to other data sets and languages. We release the framework LINSPECTOR (`https://github.com/UKPLab/linspector`), which consists of the data sets for probing tasks along with an easy-to-use probing and downstream evaluation suite based on AllenNLP.

We have performed an exhaustive set of intrinsic and extrinsic experiments with a diverse set of pretrained multilingual embeddings for five typologically diverse languages: German, Spanish, Russian, Turkish, and Finnish. We found that evaluated embeddings provide a varying range of improvement over the baselines. Our statistical analysis of intrinsic and extrinsic experimental results showed that the proposed probing tasks are positively correlated to the majority of the downstream tasks. In general, the number of correlated probing tests was higher for agglutinative languages, especially for syntactic tasks. We showed that the sets of correlated tests differ depending on the type of the downstream task. For instance, XNLI performance is strongly correlated with the *SameFeat* probing accuracy, whereas SRL is correlated well with the *Case*. We observed *Case*, *POS*, *Person*, *Tense*, and *TagCount* to have significantly high correlations for the majority of the analyzed languages and tasks; in addition, language-specific tests such as *Possession* were found to correlate well in cases when they were applicable. Furthermore, the results of our diagnostic case study, where we probe encoding and an intermediate layer of a black-box neural model, showed strong connections to the correlated tests. All these findings suggest that the proposed probing tests can be used to estimate the predictive performance of an input representation on a downstream task, as well as to explore the strengths and weaknesses of existing neural models, or to understand the relation between a model parametrization and its ability to capture linguistic information (e.g., *how the performance on probing tests changes after increasing the model size)*.

We have shown that data set statistics (e.g., *out-of-vocabulary ratio, data set size)* are a major factor influencing the results in addition to linguistic properties of the language (e. g., *typology, paradigm size, regularity)*. This can sometimes introduce noise and yield inconclusive correlations. Finally, investigation of *token-level* tasks revealed that *type-level* tasks contain less domain and majority class bias compared with token-level tasks, whereas token-level tests are generally more lexically diverse. In addition, removal of ambiguous forms from a type-level probing task may result in weaker correlations.

Appendix A. Out-of-Vocabulary Analysis

Table A.1
Training and Development OOV Rate for Intrinsic Task with MUSE.

Feature	Finnish		German		Russian		Spanish		Turkish	
	train	dev	train	dev	train	dev	train	dev	train	dev
Case	61.9	59.7	56.1	56.5	61.7	62.8	–	–	44.8	44.9
Gender	–	–	–	–	63.4	64.2	47.9	48.8	–	–
Mood	55.6	56.6	78.0	76.3	–	–	51.7	49.7	–	–
Number	59.7	58.6	59.7	60.3	64.5	66.2	49.6	50.5	–	–
POS	71.5	71.6	72.1	70.2	73.7	71.9	62.7	63.3	46.0	44.7
Person	55.4	58.0	78.4	78.0	67.4	67.3	51.1	52.5	57.3	56.1
Polarity	–	–	–	–	–	–	–	–	56.8	56.4
Possession	–	–	–	–	–	–	–	–	45.3	45.5
Pseudo	–	–	46.5	49.2	–	–	50.8	50.6	52.2	51.0
Tense	55.2	55.6	72.4	72.8	67.5	68.3	50.8	49.0	50.8	52.5
Voice	54.8	54.6	–	–	83.0	82.2	–	–	–	–
CharacterBin	67.9	68.4	59.7	58.0	68.0	68.2	63.9	63.8	44.1	44.3
TagCount	67.9	68.4	59.7	58.0	68.0	68.2	63.9	63.8	44.1	44.3
OddFeat	87.2	88.7	50.9	49.4	60.3	60.0	64.3	65.2	62.2	61.0
SameFeat	91.7	92.2	50.7	49.7	65.6	64.8	83.4	83.8	78.7	78.4

Table A.2
Training and Development OOV Rate for Intrinsic Task with word2vec.

Feature	Finnish		German		Russian		Spanish		Turkish	
	train	dev	train	dev	train	dev	train	dev	train	dev
Case	19.7	18.6	34.5	33.5	18.6	18.6	–	–	21.7	20.1
Gender	–	–	–	–	18.3	17.9	22.3	21.4	–	–
Mood	19.4	20.4	66.7	65.5	–	–	38.6	37.9	–	–
Number	19.4	18.9	38.2	37.1	18.7	18.2	31.2	30.9	–	–
POS	20.3	19.8	45.5	43.4	36.3	36.6	37.8	37.4	21.8	19.9
Person	19.2	20.6	66.5	68.1	19.1	19.9	39.0	38.7	35.6	35.4
Polarity	–	–	–	–	–	–	–	–	35.5	33.8
Possession	–	–	–	–	–	–	–	–	21.3	21.9
Pseudo	–	–	51.6	52.1	–	–	42.1	42.0	47.9	47.5
Tense	20.0	17.9	59.5	58.0	18.7	20.0	37.8	35.6	27.6	27.8
Voice	18.9	20.0	–	–	54.6	51.5	–	–	–	–
CharacterBin	20.3	21.4	39.9	41.0	33.9	34.0	35.7	36.0	21.1	21.5
TagCount	20.3	21.4	39.9	41.0	33.9	34.0	35.7	36.0	21.1	21.5
OddFeat	75.7	77.2	38.4	39.5	29.7	29.6	46.3	44.8	49.3	50.2
SameFeat	85.1	85.9	38.7	38.3	37.0	38.0	73.1	73.5	70.6	71.0

Table A.3
OOV Rate for Extrinsic Task with MUSE and word2vec Embeddings.

Task	Language	MUSE			word2vec		
		train	dev	test	train	dev	test
NER	Finnish	16.33	14.44	16.34	11.28	9.07	11.21
	German	9.28	9.28	9.19	14.02	13.81	14.18
	Russian	11.73	11.54	11.92	8.8	8.62	8.97
	Spanish	4.34	5.13	3.89	13.63	13.88	13.14
	Turkish	13.77	13.69	13.6	12.24	12.17	11.96
UD	Finnish	15.86	15.17	15.08	10.68	10.32	10.02
	German	7.86	7.05	7.61	11.74	11.55	12.36
	Russian	7.03	7.26	7.88	3.41	3.69	4.26
	Spanish	2.71	2.79	2.98	12.08	11.85	12.32
	Turkish	10.66	10.53	10.36	7.79	7.98	7.87
XNLI	German	3.22	5.21	5.47	9.4	10.58	10.62
	Russian	4.25	7.19	7.15	2.81	3.95	4.02
	Spanish	1.95	2.79	3.01	10.98	11.89	11.93
	Turkish	3.98	8.13	8.39	2.99	5.77	5.89
SRL	Finnish	15.92	14.82	15.69	10.75	9.92	10.9
	German	9.89	10.83	11.12	14.68	15.39	15.48
	Spanish	5.86	5.96	6.16	14.72	14.56	15.01
	Turkish	10.25	9.9	9.92	7.58	7.28	7.26

Appendix B. Correlation

In order to provide more insight into the relation between the downstream tasks and the probing tests, we show only the significant Spearman correlations with $p = 0.2$ in Figure B.1.

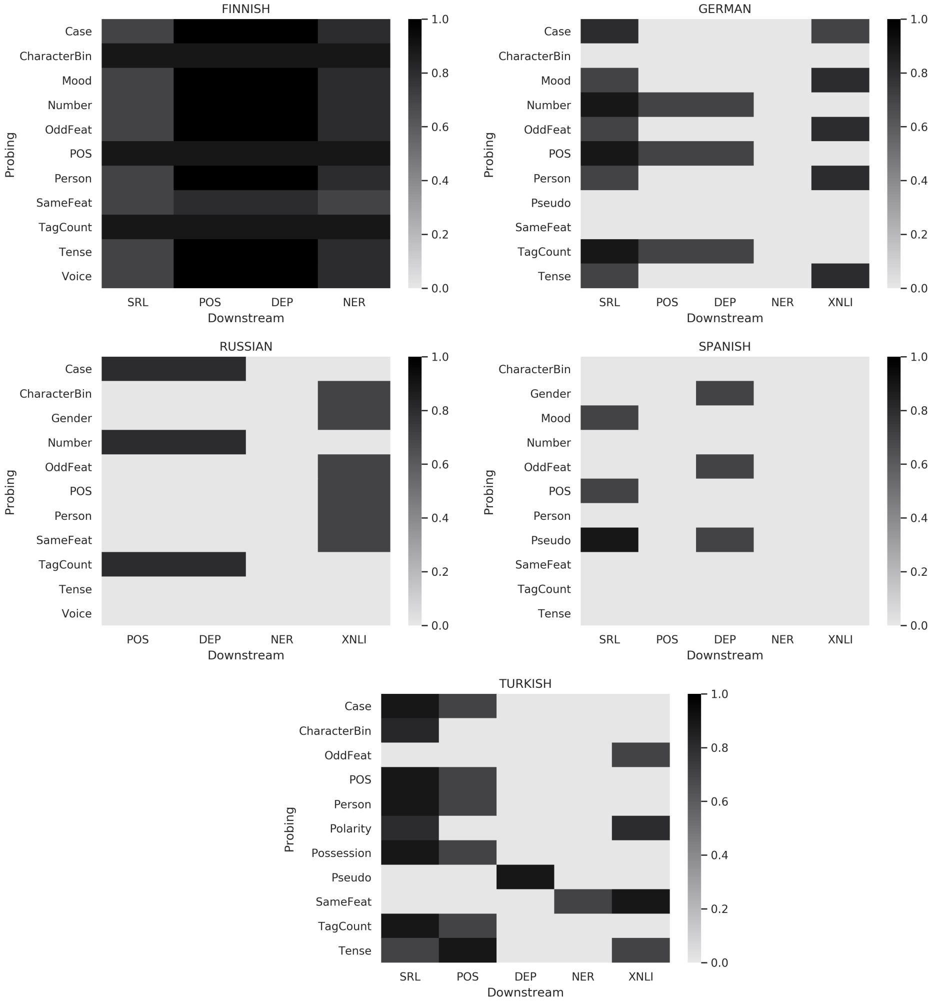

Figure B.1
Spearman correlation between probing and downstream tasks for each language. Weak correlations are not shown.

Appendix C. Intrinsic Experiments on Additional Languages

Our framework allows us to compare the type-level probing task performance between related languages, and we report additional results on Czech from Slavic; Hungarian from Uralic; and French from Romance language families. We first compare the majority baseline scores for each language pair shown in Figure C.1, then apply the same intrinsic experimental setup for the additional languages given in Table C.1.

Hungarian vs. Finnish. As it can be seen in Figure C.1 Hungarian and Finnish follow similar performance trends, in line with the similar MCC scores as given by Cotterell et al. (2018). The slightly bigger gap between POS tasks is due to Finnish data containing words from the "Adjective" class, which the Hungarian data doesn't. When compared, the ranking of different embedding spaces for intrinsic experiments were identical for both languages: D-ELMO, followed by fastText, which was generally true for all languages experimented with. In addition, we observe that while all embeddings achieved relatively high scores for "Number," "POS," and "TagCount" probing tests, "Tense," "CharacterBin," and "OddFeat" had lower scores for both languages. Apart from similarities, when we compare the scores of the best performing embedding on intrinsic experiments, we observe that it achieved lower scores for the majority of the tasks for Finnish. This can be explained with high number of paradigms (57,642) in Finnish UniMorph data, compared with Hungarian data that only has 13,989 paradigms.

Czech vs. Russian. Although the majority of baseline scores follow the same trajectory for both languages, we observe the repeating pattern of lower baseline scores for probing tasks in Russian with respect to Czech. The only exception is the "Case" probing task, which can be explained by the different numbers of case markers: Czech has 7, whereas Russian has 6 distinct case marking features in the Unimorph data. These results suggest that Russian probing tasks had more heterogeneous instances (especially for the POS test). This is particularly due to the original Czech Unimorph data covering predominantly nominal inflections whereas the Russian data is more balanced in terms of lexical variety. This is hinted by the large gap between the Unimorph datasizes, where Russian had 28,068 paradigms and 473,481 inflections, whereas Czech data was less than a quarter of it (5,125 and 134,527, respectively). Finally, we find that the accuracy intrinsic

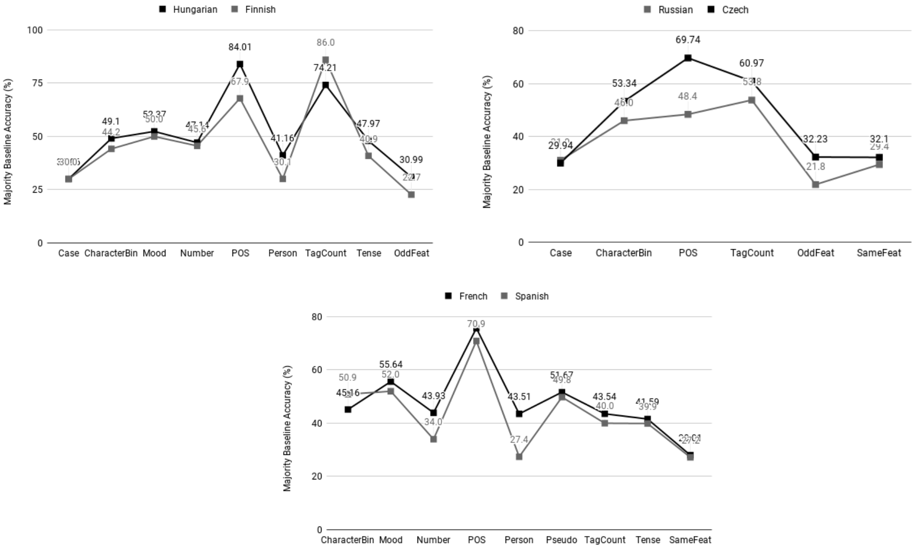

Figure C.1
Majority baseline scores comparison for related languages for the common tasks.

Table C.1
Probing Task Results for Additional Languages from the Same Language Family. **Bold** represents the best score, and *italics* represents the second best.

Hungarian (Uralic)

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Case	30.06	55.7	43.2	86.8	*95*	**98.2**
Mood	52.37	64.3	68.7	83.9	*94.3*	**97.6**
Number	47.14	63.3	61	85.9	*94.8*	**97.3**
POS	84.01	89	85	92.2	*97.8*	**99.6**
Person	41.16	58.6	58.1	82.9	*93.8*	**97**
Tense	47.97	67.1	65.9	81.3	*93.2*	**95.1**
CharacterBin	49.1	53.9	47.6	52	*55.4*	**65**
TagCount	74.21	83.7	77.1	90.3	*96.4*	**98.2**
OddFeat	30.99	37.5	32.5	45.8	*49.2*	**52.2**

Czech (Slavic)

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Case	29.94	69.6	64.4	85.8	*94.6*	**97.1**
POS	69.74	80.3	72.2	87.6	*96.1*	**98.5**
CharacterBin	53.34	57.5	54.2	61.0	*65.8*	**71.5**
TagCount	60.97	78.2	68.3	78.2	*92.2*	**95.8**
OddFeat	32.23	46	42.4	41.7	*49.7*	**52.2**
SameFeat	32.1	65.9	69.6	75.7	*84*	**89.7**

French (Romance)

Task	baseline	MUSE	word2vec	GloVe-BPE	fastText	D-ELMo
Mood	55.64	70.6	72.2	79.2	*93.4*	**93.4**
Number	43.93	62.9	72.1	88	*98.3*	**99**
POS	75.7	88.7	84.6	91.4	*98.8*	**99.3**
Person	43.51	58	66.4	85.7	*97.7*	**98**
Pseudo	51.67	*96.9*	75	90.8	85.5	**97.6**
Tense	41.59	62.7	67.6	73.5	**95.3**	*94*
CharacterBin	45.16	57.6	54.6	59.5	*62.1*	**73.9**
TagCount	43.54	66.9	67	67.9	**91.1**	*90.2*
SameFeat	28.01	43.2	67.3	55.3	*72.2*	**72.6**

task ranking of the best performing embedding, D-ELMO, is identical across both languages, that is, "POS" having the highest scores and "OddFeat" having the lowest.

French vs. Spanish. Unlike previous language pairs, Unimorph 2.0 only contains verbal inflection paradigms for both languages, which slightly limits our analysis. Similar to previous language pairs, we observe a comparable baseline trajectory as shown in Figure C.1, with a slightly bigger gap for Number and Person tasks. We attribute it to the removal of more ambiguous forms or infrequent words that contain a diverse set of Number and Person tags in French than Spanish. When comparing the results of the intrinsic experiments, we notice a repeating pattern where all embeddings score their highest for "Number," "POS" and "Person"; while achieving the lowest for "CharacterBin" and "SameFeat" tasks.

Acknowledgments

We first thank the anonymous reviewers who helped us improve the paper. We would like to thank Marvin Kaster for his help on contextualizing the probing tasks and to Max Eichler for his contribution on acquiring experimental results for additional languages in Appendix C. We are sincerely grateful to Adam Lopez, Ida Szubert, Ji-Ung Lee, Edwin Simpson, and members of AGORA Lab for providing feedback on early drafts of this work. This research has been supported by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1), and also by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR) and as part of the Software Campus program under the promotional reference 01IS17050. We gratefully acknowledge the support of NVIDIA Corp. with the donation of the Tesla K40 GPU used for this research. Clara Vania is supported by the Indonesian Endowment Fund for Education (LPDP), the Centre for Doctoral Training in Data Science, funded by the UK EPSRC (grant EP/L016427/1), and the University of Edinburgh. We finally thank UKP system administrators for their system maintenance efforts and Celal Şahin for helping with system configuration errors.

References

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*, pages 1–13.

Aggarwal, Rakesh and Priya Ranganathan. 2016. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspectives in Clinical Research*, 7(4):187–190.

Ataman, Duygu and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Belinkov, Yonatan and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bengtson, Eric and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference*, pages 294–303, Honolulu, HI.

Benikova, Darina, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. Germeval 2014 named entity recognition shared task: Companion paper. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 104–112, Hildesheim.

Bisazza, Arianna and Clara Tump. 2018. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876.

Blake, Barry J. 2001. Case. Cambridge University Press. 119–120.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642, Lisbon.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In the *50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012*, pages 136–145, Jeju Island.

Camacho-Collados, Jose, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver.

Che, Wanxiang, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels.

Chen, Qian, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics, ACL 2017, pages 1657–1668, Vancouver.

Comrie, Bernard and Maria Polinsky. 1998. The great Daghestanian case hoax. In A. Siewierska and J. Jung Song, eds. *Case, Typology and Grammar*, pages 95–114.

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, ACL 2018, Volume 1: Long Papers*, pages 2126–2136, Melbourne.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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, Brussels.

Corbett, Greville G. 2013. Number of genders. In Dryer, Matthew S. and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig. Available at `https://wals.info/chapter/30`.

Cotterell, Ryan, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Cotterell, Ryan, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 2 (Short Papers)*, pages 536–541, New Orleans, LA.

Şahin, Gözde Gül and Eşref Adalı. 2018a. Annotation of semantic roles for the Turkish proposition bank. *Language Resources and Evaluation*, 52(3):673–706.

Şahin, Gözde Gül and Mark Steedman. 2018. Character-level models versus morphology in semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 386–396, Melbourne.

Dozat, Timothy and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, pages 1–8, Toulon.

Eichler, Max, Gözde Gül Sahin, and Iryna Gurevych. 2019. LINSPECTOR WEB: A multilingual probing suite for word representations. 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 2019, System Demonstrations*, page 127–132, Hongkong.

Erten, Begum, Cem Bozsahin, and Deniz Zeyrek. 2014. Turkish resources for visual word recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2106–2110, Reykjavic.

Eryigit, Gülsen, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Fares, Murhaf, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10*, pages 406–414, Hong Kong.

Gage, Philip. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the*

Association for Computational Linguistics, 6:451–465.

Ghaddar, Abbas and Phillippe Langlais. 2017. Winer: A Wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.

Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA.

Haverinen, Katri, Jenna Kanerva, Samuel Kohonen, Anna Missila, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. The Finnish Proposition Bank. *Language Resources and Evaluation*, 49(4):907–926.

Heinzerling, Benjamin and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2989–2993, Miyazaki.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hohensee, Matt and Emily M. Bender. 2012. Getting more from morphology in multilingual dependency parsing. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 315–326, Montréal.

Huang, Eric, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882.

Huang, Zhiheng, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR, arXiv preprint arXiv: 1508.01991*.

Iggesen, Oliver A. 2013. Number of cases. In Dryer, Matthew S. and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Isgüder, Gözde Gül and Esref Adali. 2014. Using morphosemantic information in construction of a pilot lexical semantic resource for Turkish. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing, LG-LP at COLING 2014*, pages 46–54, Dublin.

Keuleers, Emmanuel and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633.

Kim, Yoon, Yacine Jernite, David A. Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2741–2749, Phoenix, AZ.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, pages 1868–1873, Miyazaki.

Köhn, Arne. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon.

Köhn, Arne. 2016. Evaluating embeddings using syntax-based classification tasks as a proxy for parser performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 67–71.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, pages 1–14, Vancouver.

Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Volume 2: Short Papers*, pages 302–308, Baltimore, MD.

Ling, Wang, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon.

Linzen, Tal. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Luong, Thang, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013*, pages 104–113, Sophia.

McCarthy, Arya D., Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies, UDW@EMNLP 2018*, pages 91–101, Brussels.

McDonald, Ryan, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98, Ann Arbor, MI.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, pages 1–12. Scottsdale, AZ.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Nangia, Nikita, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 1–10.

Nayak, Neha, Gabor Angeli, and Christopher D. Manning. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, RepEval at ACL 2016*, pages 19–23, Berlin.

Nivre, Joakim, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, and Maria Jesus Aranzabe. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543, Doha.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, LA.

Pilehvar, Mohammad Taher, José Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream NLP applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*, pages 1857–1869, Vancouver.

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

Rogers, Anna, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703.

Rogers, Anna, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on*

*Lexical and Computational Semantics (*SEM 2017)*, pages 135–148.

Rubenstein, Herbert and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Ruder, Sebastian, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Sahin, H. Bahadir, Caglar Tirkaz, Eray Yildiz, Mustafa Tolga Eren, and Ozan Sonmez. 2017. Automatically annotated Turkish corpus for named entity recognition and text categorization using large-scale gazetteers. Available at `http://arxiv.org/abs/1702.02363`.

Sang, Erik Tjong Kim and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Schnabel, Tobias, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 298–307, Lisbon.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Shi, Xing, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT Learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Sylak-Glassman, John. 2016. The composition and use of the universal morphological feature schema (UniMorph schema). Technical report, Center for Language and Speech Processing, Johns Hopkins University.

Sylak-Glassman, John, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, Volume 2: Short Papers*, pages 674–680, Beijing.

Tal Linzen, Tal, Grzegorz Chrupała, and Afra Alishahi. 2018. Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–18.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, pages 1–17.

Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054.

Vania, Clara, Andreas Grivas, and Adam Lopez. 2018. What do character-level models learn about morphology? The case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2573–2583.

Vania, Clara and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027.

Veldhoen, Sara, Dieuwke Hupkes, and Willem H. Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@NIPS*, pages 69–77.

Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Yang, Wei, Wei Lu, and Vincent Zheng. 2017. A simple regularization-based algorithm for learning cross-domain word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 2898–2904, Copenhagen.