Error Analysis and the Role of Morphology

Marcel Bollmann
Department of Computer Science
University of Copenhagen
Denmark
marcel@di.ku.dk

Anders Søgaard
Department of Computer Science
University of Copenhagen
Denmark
soegaard@di.ku.dk

Abstract

We evaluate two common conjectures in error analysis of NLP models: (i) Morphology is predictive of errors; and (ii) the importance of morphology increases with the morphological complexity of a language. We show across four different tasks and up to 57 languages that of these conjectures, somewhat surprisingly, only (i) is true. Using morphological features *does* improve error prediction across tasks; however, this effect is *less* pronounced with morphologically complex languages. We speculate this is because morphology is more discriminative in morphologically simple languages. Across all four tasks, case and gender are the morphological features most predictive of error.

1 Introduction

In error analysis, we often blame morphology (Nivre, 2007; Bender, 2009), i.e., the productive inflection and derivation of new word forms. Morphology has been argued to be a major source of error in syntactic parsing (Tsarfaty et al., 2020), semantic parsing (Şahin and Steedman, 2018), machine translation (Irvine et al., 2013; Burlot and Yvon, 2017) and a range of other tasks, in particular in morphologically complex languages (Bender, 2009; Søgaard et al., 2018; Tsarfaty et al., 2020). This paper presents a large-scale study showing that morphology *is*, as commonly conjectured, an important source of error across tasks, but somewhat surprisingly, that morphology is *less* predictive of errors in morphologically complex languages.

English is a morphologically *simple* language, showing very limited inflection and expressing most concepts through syntactic structure instead; it is also the most-represented language at major natural language processing (NLP) venues and that with the largest amount of language resources available (Bender, 2011; Joshi et al., 2020). This makes it easy to ignore morphology when designing model architectures. As a consequence, we frequently observe that performance of NLP systems on morphologically more complex languages lags behind that for English (e.g. Czarnowska et al., 2019; Tsarfaty et al., 2020).

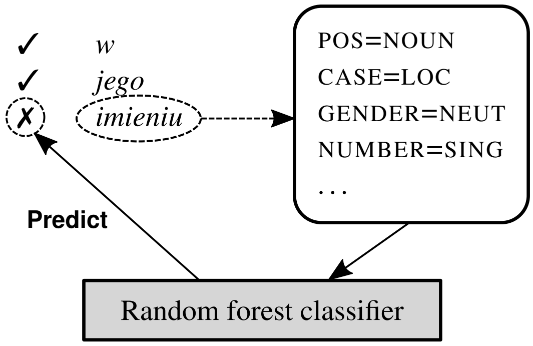

Figure 1: Overview of our methodology: We map each token to a set of morphological features and, based on this representation, predict whether some NLP system (e.g., a dependency parser) was correct (✓) or made an error (✗) on that token.

Complex morphology leads to the occurrence of rare inflected word forms. Polish nouns, for example, can inflect for number and seven different cases; this makes it less likely that all of these inflected word forms appear in the training data for our NLP models. Consequently, a model that correctly handles *imię* 'name' (NOM.SG) might not have seen the less frequent form *imionami* (INST.PL), potentially resulting in errors. If the model has generally seen fewer words in instrumental case, this can lead to systematic errors on this class of inflections.

Nowadays, many NLP systems use statistically learned subword units such as byte-pair encodings (Sennrich et al., 2016) or use characters as input representations, which could allow a system to generalize to individual affixes. However, in practice, these approaches are often found to be insufficient at capturing morphological structure (Vania and Lopez, 2017; Bostrom and Durrett, 2020; Klein and Tsarfaty, 2020).

Contributions In this study, we revisit two common conjectures about the role of morphology that are made in error analysis of NLP systems. Specifically, we ask whether (i) whether morphology is generally predictive of errors across tasks and languages; and (ii) whether the extent to which morphology is predictive depends on the morphological complexity of the language in question. These conjectures are common throughout the literature (Nivre, 2007; Bender, 2009; Manning, 2011).

Looking at data from four shared tasks on semantic role labeling (Hajič et al., 2009), dependency parsing (Zeman et al., 2018), verbal multi-word expression identification (Ramisch et al., 2018), and quality estimation (Fonseca et al., 2019), we map each token in the input data to *a set of morphological features*. Using only this feature set, and *without* using any orthographic or distributional representation of the input, we train random forest classifiers to predict whether a system has made an error on an input token. Figure 1 illustrates this approach.

Using this methodology, we find that, somewhat surprisingly, our results only support the first conjecture. In other words, (i) while morphology *is* helpful in predicting such errors, (ii) the degree to which morphology helps does not increase with the morphological complexity of the language. Moreover, we find and discuss task-specific differences between which morphological features are predictive of error. In general, part of speech, case and gender are most predictive of error.

The code for obtaining the datasets and running the experiments is made publicly available.[1]

2 Background

Morphology is frequently identified as a source of error during qualitative evaluations of NLP systems. Honnibal et al. (2010) observe that inflectional variants cause problems for statistical CCG tagging due to training data sparseness, and explicit morphological analysis helps, even for English. For dependency parsing, Seeker and Kuhn (2013) identify case syncretism as a source of error propagation in data from Czech, German, and Hungarian. Tsarfaty et al. (2020) give a broader overview of the challenges that rich morphological structure presents for dependency parsing, and Şahin and Steedman (2018) discuss the importance of morphology in semantic parsing.

Many observations of the effect of morphology come from evaluating machine translation (MT) systems. Federico et al. (2014) show that morphological errors are common for MT into Arabic and Russian and strongly affect human quality judgement. For English–Romanian MT, Peter et al. (2016) find that tense and verb form on the target side are a common source of error. Klubička et al. (2017) find that errors in English–Croatian MT are more common for some morphological categories, such as case. In a similar vein, Burlot and Yvon (2017) evaluate morphological competence of MT systems using contrast pairs and show that systems have different strengths and weaknesses for different morphological phenomena. Beyond parsing and MT, morphology has also been shown to present a challenge for tasks such as Arabic handwriting recognition (Habash and Roth, 2011) or Russian anaphora resolution (Toldova et al., 2016).

Most of the studies cited above predate contextual embedding models such as BERT (Devlin et al., 2019), which are now considered state-of-the-art for many NLP tasks. So far, few studies have explicitly analysed BERT with regard to morphology. Edmiston (2020) analyses morphological content in BERT-style models for five languages and finds that "[morphological] ambiguity is negatively correlated with performance on classification, and to a significant degree in many cases", suggesting that morphology is still a significant source of error in these models. We go significantly beyond this work by studying a much larger set of morphological variables, across several architectures and tasks, and across up to 57 languages.

3 Datasets

We collect datasets from shared tasks that (i) publish system outputs along with their gold annotations, (ii) span a variety of languages, and (iii) cover different NLP tasks. Based on these criteria, we pick datasets from the following shared tasks:

- **SEM**: CoNLL-2009 Shared Task on Semantic Dependencies (Hajič et al., 2009), covering semantic role labeling for seven languages.

[1] https://github.com/coastalcph/eacl2021-morpherror

- **UDP**: CoNLL-2018 Shared Task on Universal Dependencies Parsing (Zeman et al., 2018), covering syntactic parsing for 57 languages.

- **VMWE**: PARSEME 2018 Shared Task on Automatic Identification of Verbal Multiword Expressions (MWE; Ramisch et al., 2018), covering nine languages.

Additionally, we use the following dataset for its *gold* annotations:

- **MT**: WMT 2019 Shared Tasks on Quality Estimation (Fonseca et al., 2019), covering word-level quality estimation for English–German and English–Russian machine translation.

Here, we are not interested in the system outputs from the shared task; instead, we use the gold annotations for the quality estimation, which give us token-level error labels for the underlying machine translation outputs. Section 4.2 describes in detail how we assign error labels to these datasets.

4 Methodology

We train a classifier to predict errors made by NLP systems based on morphological features of the input tokens, in order to then analyze which morphological features (if any) are most predictive of such errors. We first describe how we obtain these features (Sec. 4.1) and how we classify when an NLP system has made an error (Sec. 4.2), then describe the classifier itself (Sec. 4.3).

4.1 Feature Extraction

We represent each token in the input data using a binary feature set. Each individual feature is named using the convention of {CATEGORY}={VALUE}, where the former is a feature category (such as POS for "part of speech") and the latter is a value within that category (e.g. VERB). We encode these features in a binary manner, i.e., for each feature in our inventory, that feature is either present or not present. Importantly, the classifier itself has no notion of "feature categories" as it only sees a single, binary feature vector.

The full feature inventory is summarized in Table 1; what follows is a description of these features and how we derived them.

Morphological features Our *morphological feature inventory* consists of (i) Universal Dependencies (UD) features, (ii) lexical features, and (iii) string-based features.

UD features include the universal part-of-speech (POS) category and the universal feature set as defined by Universal Dependencies; e.g. U:POS=VERB or U:TENSE=PAST.[2] The UDP shared-task gold data already provides this annotation; for the other tasks, we obtain these features by running UDPipe[3] (Straka and Straková, 2017) with the largest pre-trained model for the language in question.[4]

We complement this with the following additional *lexical features*: (i) SYNCRETIC specifies to what extent a token can be representative for several morphological feature sets: e.g., *ask* can be either U:MOOD=IND or U:MOOD=IMP, depending on context; (ii) AMBIG$_{\text{POS}}$ specifies to what extent the universal part-of-speech tag of the token can differ based on context: e.g., *book* could be either U:POS=VERB or U:POS=NOUN; and (iii) AMBIG$_{\text{LEX}}$ specifies whether or not the token belongs to multiple lexemes: e.g., *ruling* is a form of both '(to) rule' and '(the) ruling'. To determine these features for a given token, we use UDLexicons[5] (Sagot, 2018); in case a language is not covered by UDLexicons, we fall back to UniMorph[6] (Kirov et al., 2018).

Finally, we define purely *string-based features* based on comparing the token with its lemma. We perform character-based string alignment using Edlib (Šošić and Šikić, 2017) and derive the following features: (i) EDIT=PRE and EDIT=SUF when there is an edit at the beginning or the end of the sequence, respectively; (ii) EDIT=IN when there is an edit in the middle of the sequence; and (iii) EDIT=FULL when there is no character alignment between the strings. These features are intended to approximate prefixation, suffixation, infixation or other word-internal processes, and suppletion, respectively.

Control features To estimate the relative importance of our morphological features for the error prediction task, we additionally introduce a set of *control features* that are not morphologically motivated (cf. Tab. 1). These are (i) string length fea-

[2] The U: prefix serves to distinguish them from the other features we define. For the full feature set, see: https://universaldependencies.org/u/feat/

[3] http://ufal.mff.cuni.cz/udpipe

[4] The VMWE task only provides POS and morphology annotation for a subset of its languages, so for consistency, we choose to run UDPipe for all languages in this task as well.

[5] http://pauillac.inria.fr/~sagot/index.html#udlexicons

[6] https://unimorph.github.io/

Feature	Definition	
Morphological features		

U:POS={VALUE}
U:{FEAT}={VALUE}
	universal part-of-speech tag, e.g. U:POS=VERB
	universal feature according to the UD specification, e.g. U:TENSE=PAST	

AMBIG$_{\text{POS}}$=NO AMBIG$_{\text{POS}}$=YES AMBIG$_{\text{POS}}$=HIGH	$\|P_t\| = 1$ $1 < \|P_t\| < 5$ $\|P_t\| \geq 5$	where P_t is the set of all observed universal POS tags for t
AMBIG$_{\text{LEX}}$=NO AMBIG$_{\text{LEX}}$=YES	$\|L_t\| = 1$ $\|L_t\| > 1$	where L_t is the set of all observed lemmata for t

SYNCRETIC=NO
SYNCRETIC=YES
SYNCRETIC=HIGH
	$\|M_t\| = 1$
	$1 < \|M_t\| < 5$
	$\|M_t\| \geq 5$
		where M_t is the set of all observed morphological feature combinations for t

EDIT=PRE
EDIT=SUF
EDIT=IN
 	
EDIT=FULL
	$x_0 \neq \text{MATCH}$
	$x_n \neq \text{MATCH}$
	$\exists i, j, k : i < j < k$
	$\wedge x_i = \text{MATCH}$
	$\wedge x_j \neq \text{MATCH}$
	$\wedge x_k = \text{MATCH}$
	$\forall i : x_i \neq \text{MATCH}$
		where $[x_0, \ldots, x_n]$ is the sequence of edit alignments between t and l,
		$x_i \in \{\text{MATCH}, \text{MISMATCH}, \text{GAP}\}$

Control features		

LEN=1-3
LEN=4-6
LEN=7-9
LEN=10+
	$1 \leq \|t\| \leq 3$
	$4 \leq \|t\| \leq 6$
	$7 \leq \|t\| \leq 9$
	$\|t\| \geq 10$
		where $\|t\|$ is the string length of t

FREQ=99
FREQ=98
FREQ=95
FREQ=90
FREQ=UNCOMMON
FREQ=RARE
	$P_{99} \leq f(t)$
	$P_{98} \leq f(t) < P_{99}$
	$P_{95} \leq f(t) < P_{98}$
	$P_{90} \leq f(t) < P_{95}$
	$4 \leq f(t) < P_{90}$
	$f(t) < 4$
		where $f(t)$ is the absolute frequency count of t
		and P_n is the n-th percentile of the frequency distribution

Table 1: Inventory of extracted features (cf. Sec. 4.1). t always denotes the token, l its lemma.

tures, where each token is assigned exactly one such feature depending on its length; and (ii) token frequency bins. For the latter, we count token frequencies in the Universal Dependencies treebanks and assign each token a frequency feature. These features are based on frequency bins that we manually curated to provide a roughly balanced distribution of tokens to bins: e.g., FREQ=99 denotes a token that is in the 99$^{\text{th}}$ percentile of the frequency distribution of all types, while FREQ=RARE denotes a token occurring less than four times overall (see Table 1 for all definitions).

Pruning and statistics Since very rare features are not very informative, for any given dataset, we remove features that occur less than 10 times in that dataset. Depending on the task and language, we generate between 17 and 120 unique features this way, with an average of 68.

4.2 Classifying errors in system outputs

The target variable for our classifier is a binary label corresponding to whether or not the shared-task system has made an error on the input token. This requires comparing the outputs of a system to the gold data and classifying each token as either *correct* or *incorrect*. We will also refer to the latter as the *error* class. This classification follows the original evaluation criteria by the shared tasks to the extent possible.

For SEM, a prediction is classified as "correct" *iff* the semantic dependencies and label columns are an exact match with the gold data. For UDP, we do the same with the syntactic head and dependency relation columns; this is the same criterion that underlies the labeled attachment score (LAS) commonly used to evaluate dependency parsing. VMWE is a little more challenging since its prediction involves a set of tokens within a sentence. For each sentence, we match up each gold MWE with the predicted MWE that has the same label and the largest token overlap. We then consider a token "incorrectly" predicted if has a MWE annotation that does not belong to one of these matched MWEs, or if it lacks a MWE annotation that it should have according to the gold data.

As mentioned before, we treat the MT data a little differently: here, the *gold data* provides binary labels in the form of "OK" and "BAD" tags, corresponding to the correctness of some machine translation system. These tags are provided both for tokens and gaps between tokens (to account for the deletion/insertion of words in machine translation). We use the token-level tags from the gold data directly as our error classification labels.

Appendix A gives an example for the error classification approach on VMWE and MT.

4.3 Training classifiers

With the extracted features (from Sec. 4.1), we can now train classifiers to predict the error variable (from Sec. 4.2). Concretely, we train *random forest classifiers* (Breiman, 2001) as implemented by Scikit-learn[7] (Pedregosa et al., 2011) on each output file provided by each shared task. Random forests are ensembles of decision trees and are quick to train: the average training time on our datasets was 14 seconds on CPU, with no single run taking longer than five minutes.

As an alternative to random forests, we also experimented with randomized logistic regression classifiers followed by *stability selection* (Meinshausen and Bühlmann, 2010) to select predictive features. In our trials, this approach showed a worse performance (in terms of F_1-score) compared to random forests, while also taking considerably longer to run (averaging 7 minutes per dataset). We therefore only report results with random forest classifiers.

5 Analysis

For each shared task (Sec. 3), we ran our classification pipeline (Sec. 4) separately for each combination of (i) system submission and (ii) language evaluated on. Since random forests are largely interpretable, our analysis focuses on the important features in our learned models.

First, though, we look at the overall F_1-score of the individual classifiers, which we evaluate via stratified 5-fold cross-validation on each data point (Sec. 5.1). Additionally, to better estimate the importance of morphology, we run our cross-validation pipeline a second time *without* the morphological features, i.e., only providing the classifiers with the "control features" shown in Tab. 1. We refer to these two feature sets as "full" and "control" settings, respectively, and analyze their differences in F_1-score (Sec. 5.2).[8] Finally, we analyse the importance of individual morphological features (Sec. 5.3).

5.1 How well do the classifiers predict errors?

To evaluate how well the full classifiers learned the task, we consider their F_1-score for predicting the "error" class. Across all of our datasets, we observe a mean F_1 of 0.43 with a standard deviation of ± 0.18. Note that our setup is not comparable to most other NLP classification tasks: we evaluate a classifier trained to detect the errors of state-of-the-art systems, which means that (i) the task is inherently hard, as those systems are optimized to fix easily detectable errors, and (ii) there is no reason to assume a priori that this task is well learnable from morphological input features alone. Therefore, we believe an F_1 score of 0.43—albeit with considerable variance in performance across tasks and languages—is a strong result.

Error rate There is one important aspect to consider: the frequency of the "error" class depends on the system performance of the data point we look at, and as such our class distribution can be highly imbalanced and varied. Indeed, F_1-score and frequency of the error class correlate very strongly with Pearson's $r = 0.93$. Figure 2 plots this relationship.[9] This suggests that the errors introduced by state-of-the-art NLP systems, unsurprisingly, become harder and harder to predict the better the underlying systems perform.

Note that data imbalance is in the nature of the error prediction task, as we expect errors in state-of-the-art systems to be rare. Additionally, different

[7] We use the default parameters in Scikit-learn 0.23, with the exception of setting class weights to be "balanced" according to their frequencies in the input data.

[8] To complement the results and analyses presented here, we also provide a detailed table with the results for all task/language pairs in Appendix B.

[9] It might look surprising that many data points have very high error rates, with some even going above 0.95; i.e., more than 95% of all predictions in the respective file are deemed to be "incorrect" according to the criteria in Sec. 4.2. Spot-checking reveals that this is, however, plausible: for example, in UDP, the average labeled attachment score (LAS) on the Thai TH_PUD treebank was only 1.38 (Zeman et al., 2018, Table 15), with 23 systems achieving a LAS of only 0.77 or lower (out of 100; cf. http://universaldependencies.org/conll18/results-las.html), which is reflected by an error rate of ≥99.23% in our data.

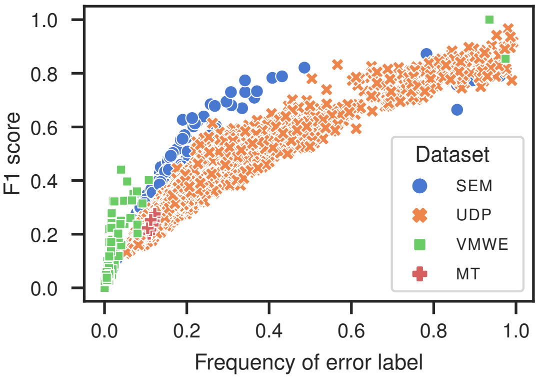

Figure 2: F_1-scores of trained error classifiers in relation to the frequency of error, i.e. the error rate of the original model (cf. Sec. 4.2).

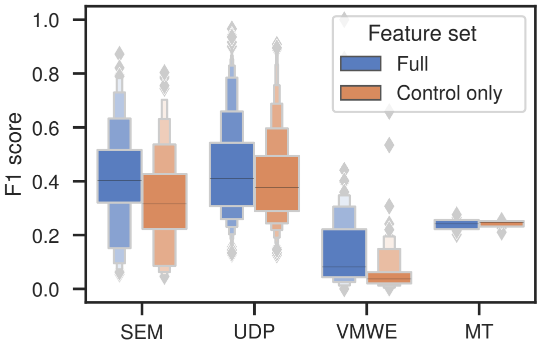

Figure 3: Quantiles of the F_1 distribution by dataset, and whether the classifiers were trained using the full feature set from Tab. 1 (blue) or only the control features (orange).

languages have differently-sized morphological tag inventories, affecting the total number of input features for the classifier. We do not attempt to apply data balancing techniques to counteract this, since this would make the task artificially easy and our results overly optimistic.

5.2 How important is morphology for predicting errors?

Figure 3 provides an alternative view of the F_1-scores presented in Fig. 2, this time as a letter-value plot (Hofmann et al., 2017) showing quantiles of the F_1 distribution. Additionally, we compare the classifier with the full feature set to the control set where morphological features were not included.

We observe that the classifiers learn best on UDP followed by SEM, while classifier F_1 is relatively poor on VMWE data. A probable explanation for this is the generally low error rate in VMWE (cf. Fig. 2). The other important observation is that classifiers in the "control" setting score consistently lower than the classifiers that have access to morphological features.

Importance by language For looking at individual languages, we restrict ourselves to the UDP data. Firstly, UDP covers 57 languages—more than any other task in our comparison—and there are no languages in the other tasks that are not also contained in UDP. Secondly, our classifier performance is generally highest on UDP (cf. Fig. 3), allowing for a more meaningful interpretation of results, particularly of selected features.

Furthermore, to factor out the effect of a data point's error rate (as discussed in Sec. 5.1), we look at the *difference* between the F_1-score of the full classifier and the control classifier trained on the same data point. In other words, we define

$$\Delta F_1 = F_1(g_f) - F_1(g_c) \qquad (1)$$

where g_f and g_c are the classifiers with the full and the control feature set, respectively. This gives us a way to judge the importance of morphological features relative to the non-morphological ones while minimizing the effect of the error rate on the results, since ΔF_1 no longer shows a strong correlation with the error rate ($r = 0.29$).

Figure 4 (bottom half) shows the quartiles of ΔF_1 scores by language in the UDP dataset. They span a wide range of values, with the median ΔF_1 varying gradually between -0.03 (for Turkish, TUR) and 0.24 (for Nigerian Pidgin, PCM). Morphological features appear to be important for some languages while being unhelpful, and sometimes even detrimental, for others.

Morphological complexity Are the differences in ΔF_1 scores (in Fig. 4) somehow related to the morphological complexity of the languages? To analyze this relationship more systematically, we use the measure of *morphological feature entropy* (MFE) introduced by Çöltekin and Rama (2018). MFE is sensitive to both the size of a language's morphological feature inventory as well as its distribution, with a more uniform distribution of features resulting in a higher MFE. Since MFE is a treebank measure that relies on the association between tokens and morphological tags, it is affected by tokenization and annotation choices of the treebank used to calculate it; therefore, it can only be considered a rough approximation of the underlying language's complexity. Like Çöltekin

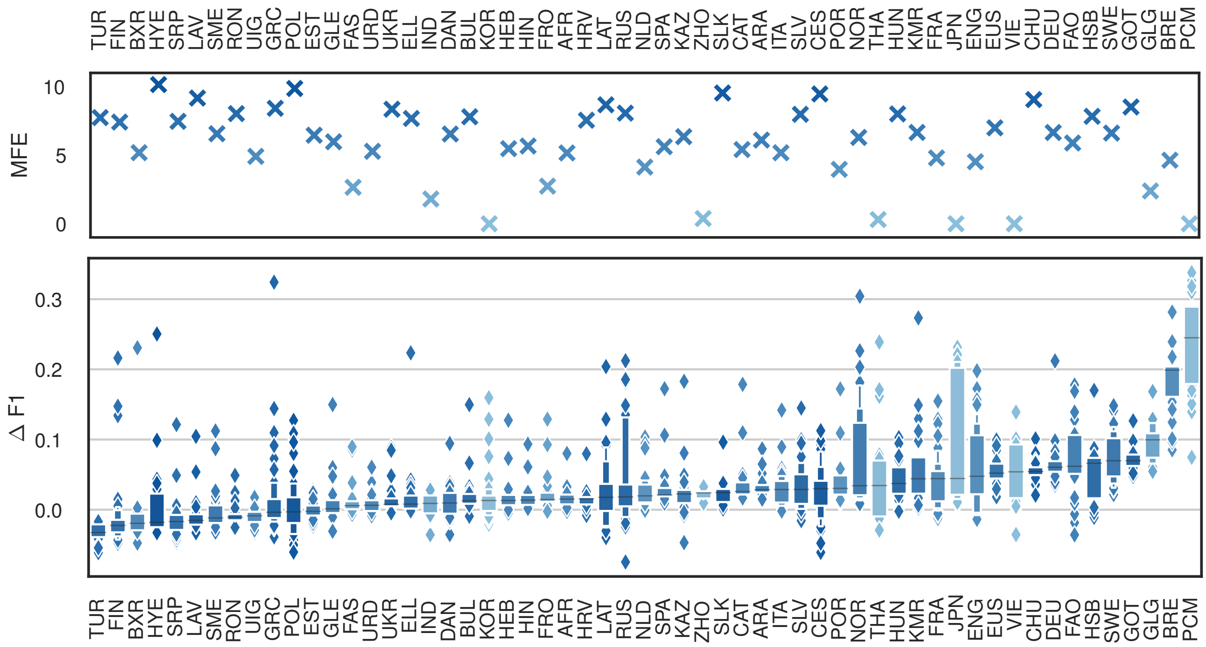

Figure 4: Classifier performance on UDP by language, sorted by median ΔF_1, where ΔF_1 is the difference in F_1-scores between training with the full and the control feature set (cf. Eq. 1). Bottom half shows the quartiles of the ΔF_1 distribution, top half shows the morphological feature entropy (MFE) for the given language; color shading is also based on MFE (with darker shade = higher MFE). Full names for all language codes as well as exact numeric values can be found in Appendix B.

and Rama (2018), we calculate the MFE score for each language on the UD treebanks.[10]

The MFE score for each language is shown in the top half of Fig. 4. Surprisingly, we find a slight, negative correlation between MFE and ΔF_1 (Pearson's $r = -0.24$). While languages with high MFE appear across the whole range of the ΔF_1 distribution, a number of languages with low MFE—and thus deemed to be more morphologically simple, such as Thai (THA), Japanese (JPN), or Nigerian Pidgin (PCM)—are found to profit *more* from the inclusion of morphological features. One possible explanation is that the control features are already very strong, which we will look at more closely in Sec. 5.3. Another possible factor is that morphologically complex languages introduce a much larger set of morphological features; if, for a given language, most of them are not relevant for predicting errors in the UDP task, they might hurt the overall classifier performance.

5.3 What morphological features are most predictive of errors?

Morphological features provide a helpful signal to the classifiers, though its overall magnitude differs by language (cf. Sec. 5.2). Now, we ask *which* of the morphological features are particularly relevant for error prediction. Since plain feature importances of trained random forest classifiers can be misleading (Strobl et al., 2007; Parr et al., 2018), we follow the approach of explicitly removing features and retraining (Parr et al., 2018; Hooker and Mentch, 2019). Unlike the analyses above, we are not concerned with generalization here, but with identifying features that are especially predictive for the error variable on each dataset as a whole. Therefore, we do not use a cross-validation strategy, but rely on the full dataset for both training and obtaining feature importances.

Concretely, for each feature *category* (as introduced in Sec. 4.1), we retrain the model *without* features from that category and note the drop in error-class F_1-score compared to the model with the full feature set. Formally, let Φ be the full feature set and $\phi_c \subset \Phi$ the subset of features belonging to category c (e.g., $c =$ U:TENSE). The *importance* of category c is then defined as

$$f(c) = F_1(C_\Phi) - F_1(C_{\Phi\setminus\phi_c}) \qquad (2)$$

where C_X is a random forest classifier trained using feature set X. Higher values for $f(c)$ mean a higher

[10] We use UD version 2.5 (Zeman et al., 2019).

Category	FI
U:POS	32.65
FREQ	15.51
LEN	11.38
U:CASE	7.25
U:GENDER	6.96
EDIT	6.75
U:NUMBER	3.78
U:NAMETYPE	2.77
U:ANIMACY	2.73
U:ADPTYPE	1.96

(a) SEM

Category	FI
U:POS	34.74
FREQ	31.99
LEN	21.79
U:CASE	16.51
U:GENDER	10.98
EDIT	9.01
U:NUMBER	6.47
U:ANIMACY	3.78
U:ASPECT	2.28
SYNCRETIC	2.08

(b) UDP

Category	FI
FREQ	38.24
LEN	28.97
U:CASE	12.56
U:POS	12.47
U:GENDER	10.15
EDIT	9.78
U:ANIMACY	7.24
U:NUMBER	7.15
U:ASPECT	5.81
U:TENSE	2.68

(c) VMWE

Category	FI
FREQ	29.50
LEN	19.63
U:CASE	12.39
U:POS	10.93
U:GENDER	9.69
EDIT	5.91
U:NUMBER	3.94
U:ASPECT	3.25
SYNCRETIC	2.84
U:ANIMACY	2.30

(d) MT

Table 2: Top 10 feature categories by *average* feature importance (FI) for each task. All FI scores given $\cdot 10^{-3}$

importance of category c, while negative values mean that including c is actually detrimental to the F_1-score.

Average feature importances Table 2 shows the top 10 feature categories for each task, averaged over all languages and datasets. The two control features, FREQ and LEN, always appear among the three most important categories, only trumped by U:POS for the UDP and SEM tasks. Notably, these three are the only feature categories that are *guaranteed* to appear with every token. It is no surprise that token frequency is strongly related to the likelihood of errors, while Zipf's law tells us that token length is strongly correlated with frequency.

Figure 5 shows the *distribution* of feature importances for the top 10 categories of UDP (cf. Tab. 2b). U:POS spans a much wider range of FI values than any of the other categories, although the outliers at the upper end all come from Nigerian Pidgin (PCM). Moreover, categories with a low average FI (e.g., U:ASPECT or SYNCRETIC) do not show outliers, i.e., are of low importance across languages. This is also true for the remaining feature categories.

Individual part-of-speech tags Since U:POS is an important feature category across tasks (cf. Tab. 2), we also look at feature importances for *individual* POS tags. For this, we use the same approach as for the feature categories (cf. Eq. 2), except that we now only remove a single U:POS feature from Φ at a time.

Table 3 shows the average feature importances for individual U:POS features, though this time we restrict ourselves to the subset of languages in UDP that are also covered in SEM.[11] This way, we can better isolate the *task-specific* differences in FI scores, without conflating them with the different *language-specific* distributions of part-of-speech tags that may affect these results. We find that adverbs (ADV) are the most important part-of-speech category for both tasks, while INTJ and PART are found to be important for predicting errors in UDP, but not in SEM. This aligns with our intuitions about what is hard in syntactic and semantic parsing, further supporting the validity of our approach.

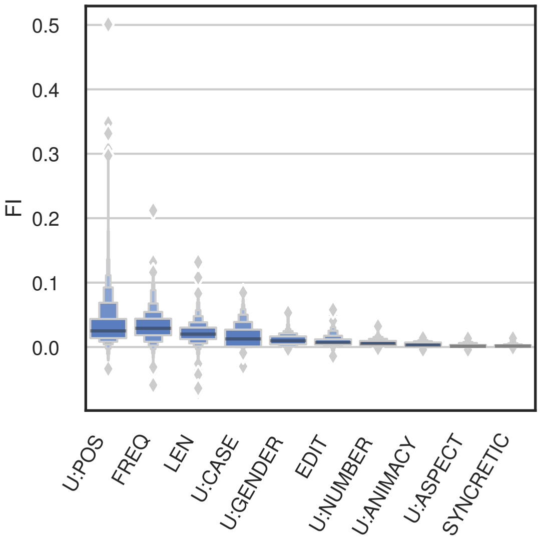

Figure 5: Quartiles of the top 10 feature categories on UDP by average feature importance (FI).

6 Conclusion

We presented a large-scale error analysis focusing on the role of morphology. Our analysis spans a range of morphological variables, four NLP tasks, and up to 57 languages. We confirm the common conjecture that morphological variables—especially case and gender—are predictive of errors

[11] These are Catalan, Czech, German, English, Japanese, Spanish, and Chinese; cf. Appendix B.

POS	**FI ($\cdot 10^{-6}$)**
ADV	39.5
INTJ	38.4
PART	26.5
AUX	15.0
PROPN	11.2
CCONJ	5.2
ADP	4.9
SCONJ	3.4
PRON	3.2
SYM	3.1
X	2.0
DET	0.6
VERB	-0.3
NUM	-1.7
PUNCT	-13.8
ADJ	-14.7
NOUN	-15.9

(a) UDP

POS	**FI ($\cdot 10^{-6}$)**
ADV	9.9
AUX	6.8
X	4.3
ADP	3.7
SCONJ	2.9
SYM	1.7
NOUN	-1.2
PRON	-1.2
PART	-2.6
INTJ	-4.3
NUM	-4.3
PROPN	-5.0
VERB	-5.0
DET	-6.2
CCONJ	-6.8
ADJ	-8.1
PUNCT	-9.3

(b) SEM

Table 3: Average feature importance (FI) for U:POS features on the subset of languages that are both in UDP and SEM.

across NLP tasks and languages. Somewhat surprisingly, we found that the usefulness of morphological variables is *negatively* correlated with the morphological complexity of the language in question. We speculate this is because morphological information is more discriminative in morphologically simple languages.

Acknowledgments

Marcel Bollmann was funded from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 845995. Anders Søgaard was funded by a Google Focused Research Award.

References

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Çağrı Çöltekin and Taraka Rama. 2018. Exploiting Universal Dependencies treebanks for measuring morphosyntactic complexity. In *Proceedings of the First Shared Task on Measuring Language Complexity*, Torun, Poland.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 974–983, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. arXiv:2004.03032.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653, Doha, Qatar. Association for Computational Linguistics.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Nizar Habash and Ryan Roth. 2011. Using deep morphology to improve automatic error detection in Arabic handwriting recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 875–884, Portland, Oregon, USA. Association for Computational Linguistics.

Jan Hajič, 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 (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Heike Hofmann, Hadley Wickham, and Karen Kafadar. 2017. Letter-value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, 26(3):469–477.

Matthew Honnibal, Jonathan K. Kummerfeld, and James R. Curran. 2010. Morphological analysis can improve a CCG parser for English. In *Coling 2010: Posters*, pages 445–453, Beijing, China. Coling 2010 Organizing Committee.

Giles Hooker and Lucas Mentch. 2019. Please stop permuting features: An explanation and alternatives. arXiv:1905.03151.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing'11, pages 171–189, Berlin, Heidelberg. Springer-Verlag.

Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Joakim Nivre. 2007. Data-driven dependency parsing across languages and domains: Perspectives from the CoNLL-2007 shared task. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 168–170, Prague, Czech Republic. Association for Computational Linguistics.

Terence Parr, Kerem Turgutlu, Christopher Csiszar, and Jeremy Howard. 2018. Beware default random forest importances.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Jan-Thorsten Peter, Tamer Alkhouli, Hermann Ney, Matthias Huck, Fabienne Braune, Alexander Fraser, Aleš Tamchyna, Ondřej Bojar, Barry Haddow, Rico Sennrich, Frédéric Blain, Lucia Specia, Jan Niehues, Alex Waibel, Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, François Yvon, Mārcis Pinnis, and Stella Frank. 2016. The QT21/HimL combined machine translation system. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 344–355, Berlin, Germany. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Benoît Sagot. 2018. A multilingual collection of CoNLL-u-compatible morphological lexicons. In

Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Gözde Gül Şahin 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 (Volume 1: Long Papers)*, pages 386–396, Melbourne, Australia. Association for Computational Linguistics.

Wolfgang Seeker and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55.

Rico Sennrich, 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, Berlin, Germany. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).

Svetlana Toldova, Ilya Azerkovich, Alina Ladygina, Anna Roitberg, and Maria Vasilyeva. 2016. Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 74–83, San Diego, California. Association for Computational Linguistics.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

Clara Vania 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, Vancouver, Canada. Association for Computational Linguistics.

Martin Šošić and Mile Šikić. 2017. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé

Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Lisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

	✗	✓	✓	✓	✗	✓	✓	✓	✓
	To	było	uczciwe	postawienie	sprawy	–	utrzymuje	Kurski	.
GOLD:				1	1				
SYS:	1			1					

(a) VMWE example from Polish

		✓		✓		✗		✓		✓		✗		✓	
	␣	ersetzt	␣	die	␣	Standardglyphen	␣	durch	␣	die	␣	Glyphenglyphenglyphen	␣	.	␣
GOLD:	OK	OK	OK	OK	OK	BAD	OK	OK	OK	OK	OK	BAD	OK	OK	OK

(b) MT example from German

Table 4: Examples for how tokens are classified as correct (✓) or incorrect (✗) in our experiments. GOLD shows gold annotations, SYS shows output from an NLP system participating in the respective shared task.

A Examples for error classification

Table 4a shows an example for how we classify errors (cf. Sec. 4.2) in the VMWE dataset on verbal multi-word expression (MWE) identification. In the gold data, a single MWE (*'postawienie sprawy'*) is annotated, while the NLP system has incorrectly identified the MWE as being *'to … postawienie'*. The annotation "1" here is an ID in case there are multiple MWEs within the same sentence. We annotate both *'to'*, which was mistakenly identified as part of the MWE, as well as *'sprawy'*, which was mistakenly left out, as an error (✗). All remaining tokens are marked as correct (✓).

Table 4b shows an example from the MT dataset on quality estimation for machine translation (MT). Here, the gold data provides us with "OK" and "BAD" labels for the individual tokens of the machine-generated translation as well as for the *gaps* between the tokens. The latter is done to be able to annotate missing passages in the machine translation output; i.e., a gap between tokens would be labelled "BAD" if the MT system should have produced more output at a given position in a sentence than it did. Since it is unclear to which (existing) tokens these "gap annotations" should be ascribed to, we do not consider them for the error classification, and only consider "OK/BAD" labels for the tokens that *do* appear in the data.

B Statistics and classifier results

Table 5 presents statistics and classifier results, corresponding to the analyses in Secs. 5.1 and 5.2, for each task/language pair. The column "Avg. error rate" corresponds to the error rates plotted in Fig. 2, while the "MFE" column shows the morphological feature entropy (cf. Sec. 5.2) for the respective language. "Avg. F_1" shows the average F_1-score after stratified 5-fold cross-validation (cf. Sec. 5.1), while "Avg. ΔF_1" corresponds to the ΔF_1-measure defined in Eq. (1).

Dataset	Language (ISO 639-3 + Name)		Avg. Error Ratio	MFE	Avg. F_1	Avg. ΔF_1
SEM	CAT	Catalan	0.15	5.41	0.39	0.13
	CES	Czech	0.31	9.47	0.63	0.08
	DEU	German	0.03	6.65	0.11	0.03
	ENG	English	0.19	4.52	0.42	0.08
	JPN	Japanese	0.23	0.00	0.57	0.09
	SPA	Spanish	0.16	5.62	0.40	0.12
	ZHO	Chinese	0.21	0.38	0.39	0.05
UDP	AFR	Afrikaans	0.20	5.16	0.40	0.02
	ARA	Arabic	0.26	6.10	0.44	0.03
	BRE	Breton	0.84	4.63	0.83	0.18
	BUL	Bulgarian	0.14	7.80	0.31	0.02
	BXR	Russia Buriat	0.87	5.18	0.85	-0.01
	CAT	Catalan	0.13	5.41	0.31	0.04
	CES	Czech	0.16	9.47	0.29	0.03
	CHU	Church Slavic	0.32	9.06	0.50	0.06
	DAN	Danish	0.22	6.55	0.35	0.01
	DEU	German	0.27	6.65	0.44	0.07
	ELL	Modern Greek (1453-)	0.16	7.68	0.33	0.02
	ENG	English	0.22	4.52	0.41	0.06
	EST	Estonian	0.22	6.46	0.34	-0.00
	EUS	Basque	0.25	6.99	0.41	0.06
	FAO	Faroese	0.68	5.89	0.71	0.07
	FAS	Persian	0.18	2.66	0.35	0.01
	FIN	Finnish	0.21	7.41	0.32	-0.01
	FRA	French	0.22	4.80	0.40	0.04
	FRO	Old French (842-ca. 1400)	0.23	2.74	0.40	0.02
	GLE	Irish	0.36	5.98	0.55	0.01
	GLG	Galician	0.26	2.39	0.50	0.09
	GOT	Gothic	0.37	8.51	0.55	0.07
	GRC	Ancient Greek (to 1453)	0.35	8.42	0.49	0.01
	HEB	Hebrew	0.29	5.48	0.48	0.02
	HIN	Hindi	0.11	5.67	0.29	0.02
	HRV	Croatian	0.18	7.55	0.32	0.02
	HSB	Upper Sorbian	0.72	7.83	0.72	0.05
	HUN	Hungarian	0.30	8.01	0.46	0.04
	HYE	Armenian	0.76	10.15	0.73	0.01
	IND	Indonesian	0.24	1.81	0.38	0.01
	ITA	Italian	0.20	5.18	0.39	0.03
	JPN	Japanese	0.42	0.00	0.55	0.11
	KAZ	Kazakh	0.78	6.35	0.80	0.03
	KMR	Northern Kurdish	0.77	6.66	0.75	0.06
	KOR	Korean	0.22	0.00	0.39	0.02
	LAT	Latin	0.35	8.67	0.46	0.02
	LAV	Latvian	0.26	9.17	0.36	-0.01
	NLD	Dutch	0.20	4.13	0.38	0.03
	NOR	Norwegian	0.25	6.29	0.38	0.07
	PCM	Nigerian Pidgin	0.84	0.00	0.79	0.24
	POL	Polish	0.13	9.86	0.25	0.00
	POR	Portuguese	0.16	3.97	0.32	0.04
	RON	Romanian	0.18	8.03	0.32	-0.01
	RUS	Russian	0.27	8.07	0.39	0.04
	SLK	Slovak	0.20	9.54	0.37	0.02
	SLV	Slovenian	0.35	7.98	0.47	0.03
	SME	Northern Sami	0.44	6.56	0.53	0.00
	SPA	Spanish	0.14	5.62	0.31	0.03
	SRP	Serbian	0.17	7.47	0.30	-0.01
	SWE	Swedish	0.23	6.61	0.41	0.07
	THA	Thai	0.93	0.31	0.83	0.04
	TUR	Turkish	0.41	7.74	0.48	-0.03
	UIG	Uighur	0.41	4.91	0.60	-0.01
	UKR	Ukrainian	0.21	8.35	0.36	0.02
	URD	Urdu	0.21	5.28	0.42	0.01
	VIE	Vietnamese	0.51	0.00	0.58	0.05
	ZHO	Chinese	0.31	0.38	0.47	0.02
VMWE	ELL	Modern Greek (1453-)	0.01	7.68	0.11	0.06
	ENG	English	0.06	4.52	0.26	0.10
	EUS	Basque	0.03	6.99	0.17	0.07
	FAS	Persian	0.03	2.66	0.11	0.03
	HEB	Hebrew	0.01	5.48	0.05	0.01
	HUN	Hungarian	0.10	8.01	0.15	0.07
	POL	Polish	0.02	9.86	0.16	0.11
	POR	Portuguese	0.07	3.97	0.15	0.06
	SPA	Spanish	0.02	5.62	0.24	0.17
MT	DEU	German	0.12	6.65	0.26	0.00
	RUS	Russian	0.11	8.07	0.22	-0.00

Table 5: Statistics and classifier results averaged over each task/language pair.