Evaluating the Morphosyntactic Well-formedness of Generated Texts

Adithya Pratapa,[1*] Antonios Anastasopoulos,[2*] Shruti Rijhwani,[1] Aditi Chaudhary,[1]
David R. Mortensen,[1] Graham Neubig,[1] Yulia Tsvetkov[3]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Department of Computer Science, George Mason University
[3]Paul G. Allen School of Computer Science & Engineering, University of Washington
{vpratapa,srijhwan,aschaudh,dmortens,gneubig}@cs.cmu.edu
antonis@gmu.edu, yuliats@cs.washington.edu

Abstract

Text generation systems are ubiquitous in natural language processing applications. However, evaluation of these systems remains a challenge, especially in multilingual settings. In this paper, we propose L'AMBRE – a metric to evaluate the morphosyntactic well-formedness of text using its dependency parse and morphosyntactic rules of the language. We present a way to automatically extract various rules governing morphosyntax directly from dependency treebanks. To tackle the noisy outputs from text generation systems, we propose a simple methodology to train robust parsers. We show the effectiveness of our metric on the task of machine translation through a diachronic study of systems translating into morphologically-rich languages.[1]

1 Introduction

A variety of natural language processing (NLP) applications such as machine translation (MT), summarization, and dialogue require natural language generation (NLG). Each of these applications has a different objective and therefore task-specific evaluation metrics are commonly used. For instance, reference-based measures such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and chrF (Popović, 2015) are used to evaluate MT, ROUGE (Lin, 2004) is a metric widely used in summarization, and various task-based metrics are used in dialogue (Liang et al., 2020).

Regardless of the downstream application, an important aspect of evaluating language generation systems is measuring the fluency of the generated text. In this paper, we propose a metric that can be used to evaluate the *grammatical well-formedness* of text produced by NLG systems.[2] Our metric is referenceless and is based on the grammatical rules of the language, thereby enabling fine-grained identification and analysis of which grammatical phenomena the NLG system is struggling with.

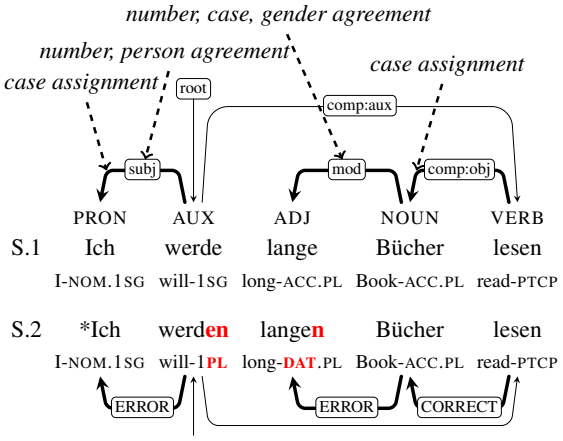


Figure 1: Identifying **grammatical errors** in text using dependency parses and morpho-syntactic rules. Ungrammatical sentence S.2 fails to satisfy subject-verb agreement between PRON and AUX as well as case agreement between ADJ and NOUN. However, it satisfies case assignment rules with the subject in NOM case and the object in ACC case respectively.

Although several referenceless metrics for evaluating NLG models exist, most use features of both the input and output, limiting their applicability to specific tasks like MT or spoken dialogue (Specia et al., 2010; Dušek et al., 2017). With the exception of the grammaticality-based metric of Napoles et al. (GBM; 2016), these metrics are derived from simple linguistic features like misspellings, language model scores or parser scores, and are not indicative of specific grammatical knowledge.

In contrast, there has recently been a burgeoning of evaluation techniques based on grammatical acceptability judgments for both language models (Marvin and Linzen, 2018; Warstadt et al., 2019; Gauthier et al., 2020) and MT systems (Sennrich, 2017; Burlot and Yvon, 2017; Burlot et al., 2018).

*Equal contribution

[1]Code and data are available at https://github.com/adithya7/lambre.

[2]While grammatical well-formedness is often necessary for fluent text, it is not sufficient (Sakaguchi et al., 2016).

However, these methods require an existing model to score two sentences that are carefully crafted to be similar, with one sentence being grammatical and the other not. These techniques are usually tailored towards specific downstream systems. Additionally, they do not consider the interaction between multiple mistakes that may occur in the process of generating text (e.g., an incorrect word early in the sentence may trigger a grammatical error later in the sentence). Most of these methods, with the exception of Mueller et al. (2020), focus only on English or translation to/from English.

In this paper, we propose L'AMBRE, a metric that *both* evaluates the grammatical well-formedness of text in a fine-grained fashion and can be applied to text from multiple languages. We use widely available dependency parsers to tag and parse target text, and then compute our metric by identifying language-specific morphosyntactic errors in text (a schematic overview is outlined in Figure 1). Our measure can be used *directly* on text generated from a black-box NLG system, and allows for *decomposing* the system performance into individual grammar rules that identify specific areas to improve the model's grammaticality.

L'AMBRE relies on a grammatical description of the language, similar to those linguists and language educators have been producing for decades when they document a language or create teaching materials. Specifically, we consider rules describing morphosyntax, including agreement, case assignment, and verb form selection. Following Chaudhary et al. (2020), we describe a procedure to automatically extract these rules from existing dependency treebanks (§3) with high precision.[3]

When evaluating NLG outputs, adherence to these rules can be assessed through dependency parses (Figure 1). However, off-the-shelf dependency parsers are trained on grammatically sound text and are not well-suited for parsing ungrammatical (or noisy) text (Hashemi and Hwa, 2016) such as that generated by NLG systems. We propose a method to train more robust dependency parsers and morphological feature taggers by synthesizing morphosyntactic errors in existing treebanks (§4). Our robust parsers improve by up to 2% over off-the-shelf models on synthetically noised treebanks.

Finally, we field test L'AMBRE on two NLP tasks: grammatical error identification (§5) and machine translation (§6). Our metric is highly correlated with human judgments on MT outputs. We also showcase how the interpretability of our approach can be used to gain additional insights through a diachronic study of MT systems from the Conference on Machine Translation (WMT) shared tasks. The success of our measure depends heavily on the quality of dependency parses: we discuss potential limitations of our approach based on the grammar error identification task.

2 L'AMBRE: Linguistically Aware Morphosyntax-Based Rule Evaluation

In this section, we present L'AMBRE, a metric to gauge the morphosyntactic well-formedness of generated natural language sentences. Our metric assumes a *machine-readable* grammatical description, which we define as a series of language-specific rules $G_l = \{r_1, r_2, \ldots, r_n\}$. We also assume that dependency parses of every grammatical sentence adhere to these rules.[4]

Given a text, we compute a score by verifying the satisfiability of all applicable morphosyntactic rules from the grammatical description. Similar to standard metrics for evaluating NLG, our scoring framework allows for computing scores at both segment-level and corpus-level granularities.

Segment level: Computing L'AMBRE first requires segmentation, tokenization, tagging, and parsing of the corpus.[5] Given the tagged dependency tree for a segment of text and a set of rules in the language, we identify all rules that are applicable to the segment. We then compute the percentage of times that each such rule is satisfied within the segment, based on the parser/tagger annotations. The final score is a weighted average of the scores of individual rules.[6] Our score lies between [0,1], where 1 and 0 represent that rules are perfectly satisfied or not satisfied at all respectively. Consider the example sentence (S.2) from Figure 1. Of the five agreement rules, two rules, number agreement between PRON (Ich) and AUX (werde), and case agreement between ADJ (lange) and NOUN (Bücher) are not satisfied. Both rele-

[3] While such sets of grammar rules could be manually compiled (for example, by linguists), it would require additional centralized effort from a large group of annotators.

[4] Different syntactic formalisms could be applicable, but we work with the (modified) Universal Dependencies formalism (Nivre et al., 2020) due to its simplicity, widespread familiarity and its use in a variety of multilingual resources.

[5] We discuss in §4 how to properly achieve this over potentially malformed sentences.

[6] We assume equal weights among rules, although it would be trivial to extend the metric to use a weighted average.

vant case assignment rules between, PRON (Ich) and AUX (werde), and NOUN (Bücher) and VERB (lesen) are satisfied. Thus, the overall score is 0.71 (5/7). This example showcases how L'AMBRE is inherently interpretable: given a segment (S.2), we can immediately identify that it is grammatically sound with respect to case assignment, but contains two errors in agreement.

Corpus level: To compute L'AMBRE at corpus-level, we accumulate the satisfiability counts for each rule over the entire corpus and report the macro-average of the empirical satisfiability of each applicable rule. This is different from a simple average of segment-level scores and is a more reliable score as it allows the comparison of performance by rule over the entire corpus.

3 Creating a Grammatical Description

In linguistics, grammars of languages are typically presented in (series of) books, describing in detail the rules governing the language through free-form text and examples (see Moravcsik (1978); Corbett (2006) for grammatical agreement).[7] However, to be able to use such descriptions in our metric, we require them to be concise and *machine-readable*.

We build upon Chaudhary et al. (2020) that constructed first-pass descriptions of grammatical agreement from syntactic structures of text, in particular, dependency parses.[8] In general, rules based on a complete formalized grammar govern several aspects of language generation, including syntax, morphosyntax, morphology, morphophonology, and phonotactics. In this work, we focus on agreement, case assignment, and verb form choice.

3.1 Agreement

We define the *agreement* rules as $r_{agree}(x, y, d) \rightarrow f_x{=}f_y$. Such rules refer to two words with parts-of-speech x (dependent) and y (head/governer) connected through a dependency relation d. These two words must exhibit agreement on some morphological feature f. For instance, the noun *Bücher* ('Book') and its modifying adjective *lange* ('long') in the German example S.1 (Figure 1) agree in number, gender, and case. We denote this general agreement rule as $r_{agree}(\texttt{ADJ}, \texttt{NOUN}, \texttt{mod}) \rightarrow \texttt{Case}, \texttt{Gender}, \texttt{Number}$.

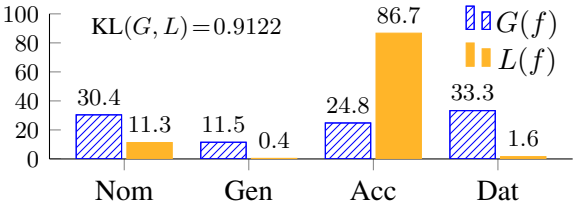


Figure 2: Argument structure rules: the global case distribution $G(f)$ of German NOUN is very different from its local $L^{depd}(f)$ distribution in *comp:obj* dependency with VERB, allowing us to the identify case assignmnt rule $r_{as}(\texttt{NOUN},\texttt{VERB},\texttt{comp:obj}) \rightarrow \texttt{Case}_{\texttt{NOUN}} = \texttt{Acc}; \texttt{Nom}$, i.e the Case can be either Acc or Nom.

For each dependency relation d between a dependent POS x and head POS y, we compute the fraction of times the linked tokens agree on feature f in the treebank. We consider $r_{agree}(x, y, d) \rightarrow f$ as a potential agreement rule if the fraction is higher than 0.9. The resulting set still contains a long tail of less-frequent rules. These are unreliable and could just be because of treebank artifacts. Therefore, we incorporate additional pruning to only select the most frequent rules, covering a cumulative 80% of all agreement instances in the treebank. This is a simplified formulation compared to Chaudhary et al. (2020), but as we show later, this frequency-based approach still results in a high-precision set of rules.

3.2 Case Assignment and Verb Form Choice

We define case assignment and verb form choice rules as $r_{as}(x, y, d) \rightarrow f_x{=}F$. A word with POS x at the tail of a dependency relation d with head POS y must exhibit a certain morphological feature (i.e., f_x must have the value F). Occasionally, a similar rule might be applicable for the head y. For instance, a pronoun that is the child of a `subj` relation (that is, it is the subject of a verb) in most Greek constructions must be in the nominative case, while a direct object (`obj`) should be in the accusative case. In this example, we can write the rules as $r_{as}(\texttt{PRON}, \texttt{VERB}, \texttt{subj}) \rightarrow \texttt{Case}_{\texttt{PRON}} = \texttt{Nom}$ and $r_{as}(\texttt{PRON}, \texttt{VERB}, \texttt{obj}) \rightarrow \texttt{Case}_{\texttt{PRON}} = \texttt{Acc}$.

Our hypothesis is that certain syntactic constructions require specific morphological feature selection from one of their constituents (e.g., pronoun subjects *need* to be in nominative case, but pronoun objects only allow for genitive or accusative case in

[7] Many linguists also produce highly formal accounts of grammatical phenomena. However, many of these formalisms are difficult to implement computationally because they are equivalent (in the most egregious cases) to Turing machines.

[8] We use the Surface-Syntactic Universal Dependencies (SUD) 2.5 (Gerdes et al., 2019). See A.1 for a comparison of UD and SUD.

Greek).[9] This implies that the "local" distribution that a specific construction requires will be different from a "global" distribution of morphological feature values computed over the whole treebank. Figure 2 presents an example for German-GSD.

We can automatically discover these rules by finding such cases of distortion. First, we obtain a global distribution ($G(f_x) = p(f_x)$) that captures the empirical distribution of the values of a morphological feature f on POS x over the whole treebank. Second, we measure two other distributions, local to a relation d, for the dependent ($L^{depd}(f_x \mid d) = p(f_x \mid \langle x, *, d \rangle)$) and head positions ($L^{head}(f_x \mid d) = p(f_x \mid \langle *, x, d \rangle)$)

To identify these morphosyntactic rules with high precision, we measure the KL divergence (Kullback and Leibler, 1951) between global and local distributions and only keep the rules with KL divergence over a predefined threshold of 0.9. Similar to the case of agreement rules, we impose a frequency threshold on the count of dependency relation in the respective treebank. For all the agreement, case assignment and verb form choice rules, we use the largest SUD treebank for the language.

3.3 Human Evaluation

Though our grammatical description incorporates agreement, case assignment and verb-form selection, which are highly indicative of the fluency of natural language text, it is by no means exhaustive. However, these rules are relatively easy to extract from dependency parses with high precision. To measure the quality of our extracted rule sets, we perform a human evaluation task with three linguists.[10] Similar to Chaudhary et al. (2020), for each rule, we present three choices, "*almost always true*", "*sometimes true*" and "*need not be true*", along with 10 positive and negative examples from the original treebank.[11] In Table 1, we show the results for Greek, Italian and Russian.

Our rules are in general quite precise across the three languages, with most rules marked as "almost always true" by linguists. However, we found interesting special cases in Russian, where the annotator stated that dependency relations are "overloaded" to capture several phenomena (explaining the "sometimes" annotations). The SUD schema merges `obj` and `ccomp` into a single `comp:obj` relation, thereby we notice instances where the rule $r_{as}(\texttt{PRON}, \texttt{VERB}, \texttt{comp:obj}) \rightarrow \texttt{Case}_{\texttt{PRON}} = \texttt{Acc}$ (which pertains to direct objects) is incorrectly enforced on a `ccomp` relation. We also notice some issues with cross-clausal dependencies, e.g., the rule $r_{as}(\texttt{VERB}, \texttt{NOUN}, \texttt{subj}) \rightarrow \texttt{VerbForm}_{\texttt{VERB}} = \texttt{Inf}$ is valid in the sentence, "the *goal* is to *win*" but not in "the *question* is why they *came*".

Rules	Greek	Russian	Italian
r_{agree}	11:0:0	17:3:0	-
r_{as}	9:3:0	6:9:3	10:1:0

Table 1: Results on human evaluation of our automatically extracted rules. Numbers denotes # rules labeled as (*always*):(*sometimes*):(*need not*)

It is important to note that these automatically extracted rule sets are approximate descriptions of morpho-syntactic behavior of the language. However, L'AMBRE is flexible enough to utilize any additional rules, and arguably would be even more effective if combined with hand-curated descriptions created by linguists. We leave this as an interesting direction for future work. In our code, we provide detailed instructions for adding new rules.

4 Parsing Noisy Text

Within our evaluation framework, we rely on parsers to generate the dependency trees of potentially malformed or noisy sentences from NLG systems. However, publicly available parsers are typically trained on clean and grammatical text from UD treebanks, and may not generalize to noisy inputs (Daiber and van der Goot, 2016; Sakaguchi et al., 2017; Hashemi and Hwa, 2016, 2018). Therefore, it is necessary to ensure that parsers are robust to any morphology-related errors in the input text. Ideally, the tagger should accurately identify the morphological features of incorrect word forms, while the dependency parser remains robust to such noise. To this end, we present a simple framework for evaluating the robustness of pre-trained parsers to such noise, along with a method to train the robust parsers necessary for our application.

[9] This class of rules are also often lexicalized, depending on the lexeme of either the head or the dependent. In the example S.1 of Figure 1, the object phrase *lange Bücher* ('long Book') is inflected in the accusative case because of the verb *lesen* ('read'). Other constructions might require the object declined in genitive or dative, depending on the verb lexeme.

[10] Disclaimer: One annotator is also an author on this work.

[11] Due to large number of Russian rules, we present only a subset to the linguists.

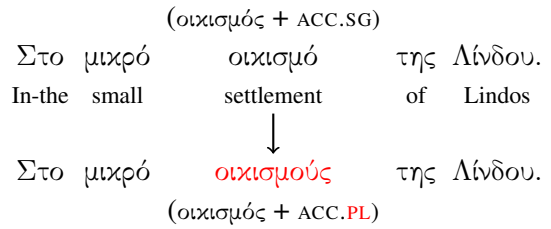


Figure 3: Creating noisy input examples for parsers. In this Greek example, we modify the original word form οικισμό (Singular) to a plural inflection οικισμούς.

4.1 Adding Morphology-related Noise

To simulate noisy input conditions for parsers, we add morphology-related errors into the standard UD treebanks using UniMorph dictionaries (McCarthy et al., 2020). UniMorph provides a schema for inflectional morphology by listing paradigms with relevant morphological features from an universal schema (Sylak-Glassman, 2016). Given an input sentence, we search for alternate inflections for the constituent tokens, based on their lemmata.[12] For simplicity, we only replace a single token in each sentence and for this token, we substitute with a form differing in exactly one morphological feature (e.g., Case, Number, etc.). For each sentence in the original treebank, we sample a maximum of one altered sentence. Figure 3 illustrates the construction of a noisy (or altered) version of an example sentence from Greek-GDT treebank.[13]

In general, we were able to add noise to more than 80% of the treebanks' sentences, but in a few cases we were constrained by the number of available paradigms in UniMorph (see A.2 for more details). A potential solution could utilize a inflection model like the `unimorph_inflect` package of Anastasopoulos and Neubig (2019), but we leave this for future work.

For evaluation, we induce noise into the *dev* portions of the treebanks and test the robustness of off-the-shelf taggers and parsers from Stanza (Qi et al., 2020) (indicative results on Czech, Greek, and Turkish are shown in Figure 4). Along with the overall scores on the dev set, we also report the results only on the altered word forms ("Altered Forms"). Across the three languages, we notice a significant drop in tagger performance, with a more than 30% drop in feature tagging accuracy of the altered word forms. The parsing accuracy is also affected, in some cases significantly. This reinforces observations in prior work and illustrates the need to build more robust parsers and taggers.

4.2 Training Robust Parsers

To adapt to the noisy input conditions in practical NLP settings like ours, our proposed solution is to re-train the parsers/taggers directly on noisy UD treebanks. With the procedure described above (§4.1) we also add noise to the *train* splits of the UD v2.5 treebanks and re-train the lemmatizer, tagger, and dependency parser from scratch.[14] To retain the performance on clean inputs, we concatenate the original clean train splits with our noisy ones. We experimented with commonly used multilingual parsers like UDPipe (Straka and Straková, 2017), UDify (Kondratyuk and Straka, 2019), and Stanza (Qi et al., 2020), settling on Stanza for its superior performance in preliminary experiments. We use the standard training procedure that yields state-of-the-art results on most UD languages with the default hyperparameters for each treebank. Given that we are inherently tokenizing the text to add morphology-related noise, we reuse the pre-trained tokenizers instead of retraining them on noisy data.

Figure 4 compares the performance of the original and our robust parsers on three treebanks. Overall, we notice significant improvements on both LAS (with similar gains on UAS) and UFeat accuracy on the altered treebank as well as the altered forms. Importantly, our robust parsers retain the state-of-the-art performance on clean text. In all the analyses reported henceforth (unless explicitly mentioned), we use our *robust* Stanza parsers trained with the above-described procedure.

5 Does L'AMBRE Capture Grammaticality?

Before deploying L'AMBRE on automatically generated text, we need to ensure that our approach is indeed able to identify syntactic ill-formedness. Grammar error correction (GEC) datasets are an ideal test bed. In its original formulation, the GEC

[12] For each token, we first map the morphological feature annotations in the original UD schema to the UniMorph schema (McCarthy et al., 2018).

[13] Tan et al. (2020) follows similar methodology using English-only `LemmInflect` tool, but our approach is scalable to the large number of languages in UniMorph.

[14] We added errors into UD, and not SUD, as it allows for reusing the original Stanza hyperparameters, and also facilitates for application of robust parsers outside of L'AMBRE. Note that, conversion between UD and SUD can be done with minimal loss of information.

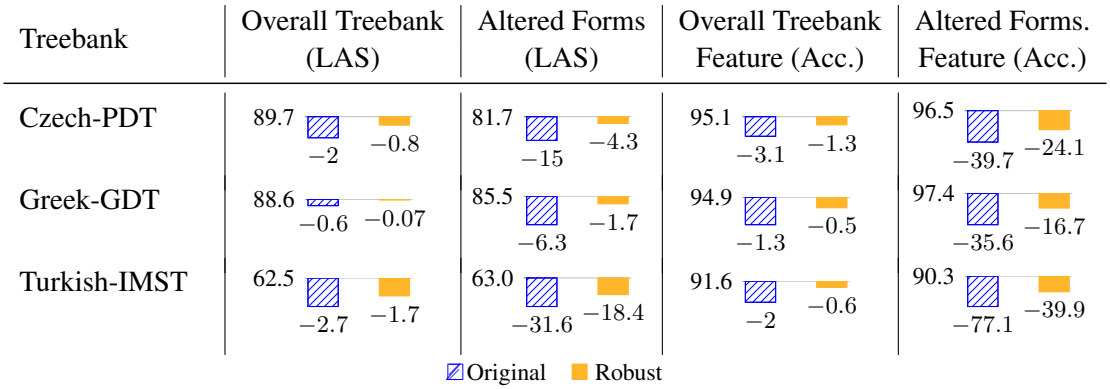


Figure 4: Our robust parsers reduce the errors on the noisy evaluation set (example over three treebanks) compared to the original pre-trained ones. The baseline axis in each plot corresponds to the performance on the clean evaluation set. Our models are more robust on both parsing (LAS) and morphological feature prediction. We report results both over the whole treebank and over only the erroneous tokens.

task involves identifying and correcting errors relating to spelling, morphosyntax and word choice. For evaluating L'AMBRE, we only focus on *grammar error identification* (GEI) and specifically on identification of morphosyntactic errors.

We experiment with two morphologically rich languages, Russian and German. We use the Falko-MERLIN GEC corpus (Boyd, 2018) for German and the RULEC-GEC dataset (Rozovskaya and Roth, 2019) for Russian. We focus on error types related to morphology (see A.3).

Evaluation: To evaluate the effectiveness of L'AMBRE, we run it on the *training*[15] splits of the German and Russian GEC datasets. GEC corpora typically annotate single words or phrases as errors (and provide a correction); in contrast, we only identify errors over a dependency link, which can then be mapped over to either the dependent or head token. This difference is not trivial: a subject-verb agreement error, for instance, could be fixed by modifying either the subject or the verb to agree with the other constituent. To account for this discrepancy, we devise a schema to ensure the proper computation of precision and recall scores. First, we detect any errors at a given token by evaluating all the valid L'AMBRE rules between the curren token and its dependency neighbors (head, dependents). If there is a gold error at the current token, we consider it a true positive or false negative depending on whether or not we detect the error. For false positive cases, we divide the score between the current token and the neighbor via the erroneous dependency link (see algorithm 1 in A.3).

[15]We use the train portion due to its large size, therefore gives a better estimate of our L'AMBRE performance. Note that, in this experiment, we do not aim to compare against state-of-the-art GEI tools.

Lang.	Parser	$r_{agree} \cup r_{as}$ P	$r_{agree} \cup r_{as}$ R	r_{agree} P	r_{agree} R	r_{as} P	r_{as} R
German	Original	32.6	29.6	34.2	28.9	**14.9**	**1.0**
	Robust	33.6	**34.5**	35.2	**33.8**	12.2	0.8
	Robust++	**40.0**	34.1	**42.5**	33.4	12.2	0.8
Russian	Original	**18.5**	20.9	**22.6**	18.7	9.9	3.7
	Robust	**18.5**	**25.7**	22.1	**20.6**	**14.6**	**8.1**

Table 2: Precision and Recall of morphosyntactic errors on train splits of German and Russian GEC. Robust++ indicates results after additional manual post-correction of rules.

Table 2 presents the results using both agreement (r_{agree}) and argument structure rules (case assignment and verb form choice, r_{as}).

Analysis: In both languages, we find agreement rules to be of higher quality than case and verb form assignment ones. This phenomenon is more pronounced in German where many case assignment rules are lexeme-dependent, as discussed in §3.

Importantly, our proposed **robust parsers lead to clear gains in error identification recall**, compared to the pre-trained ones ("Original" vs. "Robust" in Table 2). Given the complexity of the errors present in text from non-native learners and the well-known incompleteness of GEC corpora in listing all possible corrections (Napoles et al., 2016), combined with the prevalence of typos and the dataset's domain difference compared to the parser's training data, our error identification module performs quite well.

To understand where L'AMBRE fails, we man-

ually inspected a sample of false positives. First, we notice that tokens with typos are often erroneously tagged and parsed. Our augmentation is only equipped to handle (correctly spelled) morphological variants. Additionally applying a spell checker might be beneficial in future work.

Second, we find that German interrogative sentences and sentences with more rare word order (e.g., object-verb-subject) are often incorrectly parsed, leading to misidentifications by L'AMBRE. In the training portion of the German HDT treebank, 76% of the instances present the subject before the verb and the object appears after the verb in 62% of the sentences. Questions and subordinate clauses that follow the reverse pattern (OVS order) make a significant portion of the false positives.

Last, we find that morphological taggers exhibit very poor handling of syncretism (i.e., forms that have several possible analyses), often producing the most common analysis regardless of context. For example, nominative-accusative syncretism is well documented in modern German feminine nouns (Krifka, 2003). German auxiliary verbs like *werden* ('will') that share the same form for 1st and 3rd person plurals, are almost always tagged with the 3rd person. As a result, our method mistakenly identifies correct pronoun-auxiliary verb subject dependency constructions as violations of the rule $r_{agree}(\texttt{PRON}, \texttt{AUX}, \texttt{subj}) \rightarrow \texttt{Person}$, as the PRON and AUX are tagged with disagreeing person features (1st and 3rd respectively). By manually correcting for this issue over our German rules (by specifically discounting such cases) we improve L'AMBRE's precision by almost 7 percentage points ("Robust++" in Table 2).

Comparison with Other Metrics: We also compare L'AMBRE to other metrics that capture fluency and/or grammatical well-formedness, namely perplexity as computed by large language models and the grammaticality-based metric (GBM) of Napoles et al. (2016). To provide a fair comparison of L'AMBRE, perplexity and GBM, we reformulate the GEI task into an acceptability judgment task. Specifically, we check if the metrics' score the grammatical target sentence higher than the ungrammatical source sentence in the GEI test split for Russian and German. To compute the perplexity scores, we use transformer-based LMs (Ng et al., 2019). GBM relies on the open-source Language Tool (Miłkowski, 2010), which is a widely-used rule-based proofreading software to detect sentence-level errors.[16] GBM measures error count rate as $1 - \frac{\#\text{errors}}{\#\text{tokens}}$. For additional details on the task setup, we refer the readers to A.4 in Appendix.

Our findings are two-fold. First, perplexity performs better than the other metrics. However, perplexity cannot provide any error diagnosis, so it by itself is not useful for providing feedback to a user. Second, L'AMBRE is better at capturing morpho-syntactic rules necessary for grammatical correctness (especially in Russian), while GBM is better at other fluency-related aspects. While both L'AMBRE and GBM are interpretable, L'AMBRE's UD-based rule construction makes it easier to extend it to new languages.[17] For complete results, refer to Table 4 in Appendix A.4.

In our GEI analysis, we utilized the Russian and German GEC corpora for evaluating the quality of L'AMBRE. In future work, it would be interesting to expand the analysis to datasets from other languages, Czech (Náplava and Straka, 2019) and Ukrainian (Syvokon and Nahorna, 2021).

6 Evaluating NLG: A Machine Translation Case Study

Grammaticality measures, including L'AMBRE, can be useful across NLG tasks. Here, we chose MT due to the wide-spread availability of (human-evaluated) system outputs in many languages.

In addition to BLEU, chrF and t-BLEU[18] are commonly used to evaluate translation into morphologically-rich languages (Goldwater and McClosky, 2005; Toutanova et al., 2008; Chahuneau et al., 2013; Sennrich et al., 2016). Evaluating the well-formedness of MT outputs has previously been studied (Popović et al., 2006). Recent WMT shared tasks included special test suites to inspect linguistic properties of systems (Sennrich, 2017; Burlot and Yvon, 2017; Burlot et al., 2018), which construct an evaluation set of contrastive source sentence pairs (typically English). While such contrastive pairs are very valuable, they only *implicitly* evaluate well-formedness and require access to underlying MT models to score the contrastive sentences. In contrast, L'AMBRE *explicitly* measures well-formedness, without requiring access to trained MT models.

[16] https://languagetool.org/dev

[17] In L'AMBRE, we reuse the expertise of UD annotators, but in GBM, expanding to a new language requires native speakers to craft new regexes.

[18] t-BLEU (Ataman et al., 2020) measures BLEU on outputs tagged using a morphological analyzer.

en→:⁘	cs	de	et	fi	ru	tr
WMT'18						
all	0.84	-0.06	0.68	0.86	0.86	0.58
r_{agree}	0.91	0.07	0.83	0.96	0.71	0.64
r_{as}	0.78	-0.10	0.62	0.77	0.89	-0.31
WMT'19						
all	0.80	0.16	-	0.85	0.57	-
r_{agree}	0.89	0.14	-	0.87	0.70	-
r_{as}	0.70	0.13	-	0.82	0.45	-

Table 3: With a few exceptions, our grammar-based metrics correlate well with human evaluations of WMT18 and WMT19 systems (Pearson's r against Z-scores). Results using robust Stanza parsers.

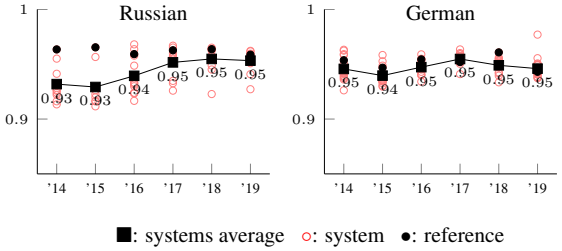


Figure 5: A diachronic study of grammatical well-formedness of WMT English→X systems' outputs. The systems in general are becoming more fluent. In the last two years the best systems produce as well-formed outputs as the reference translations.

For evaluating MT systems, we use the data from the Metrics Shared Task in WMT 2018 and 2019 (Ma et al., 2018, 2019). This corpus includes outputs from all participating systems on the test sets from the News Translation Shared Task (Bojar et al., 2018; Barrault et al., 2019). Our study focuses on systems that translate from English to morphologically-rich target languages: Czech, Estonian, Finnish, German, Russian, and Turkish. We used all relevant languages from the WMT shared task except for Lithuanian and Kazakh, which lack reasonable quality parsers.

Correlation Analysis The MT system outputs are accompanied with human judgment scores, both at the segment and system level. In contrast to the reference-free nature of human judgments, our scorer is both reference-free *and* source-free.

Following the standard WMT procedure for evaluating MT metrics, we measure the Pearson's r correlations between L'AMBRE and human z-scores for systems from WMT18 and WMT19. We follow Mathur et al. (2020) to remove outlier systems, since they tend to significantly boost the correlation scores, making the correlations unreliable, especially for the best performing systems (Ma et al., 2019). Table 3 presents the correlation results for WMT18 and WMT19.[19]

We generally observe moderate to high correlation with human judgments using both sets of rules across all languages, apart from German (WMT18,19). This confirms that grammatically sound output is an important factor in human evaluation of NLG outputs. The correlation is lower with case assignment and verb form choice rules, with notable negative correlations for German, and Turkish (WMT18). In the case of German, a significant number of case assignment rules are dependent on the lexeme (as noted in §3) and we expect future work on lexicalized rules to partially address this drawback. In Turkish, the low parser quality plays a significant role and highlights the need for further work on parsing morphologically-rich languages (Tsarfaty et al., 2020). Last, we note that human judgments, unlike L'AMBRE, incorporate both well-formedness *and* adequacy (with respect to the source). Therefore, we recommend using L'AMBRE in tandem with standard MT metrics to obtain a good indication of overall performance, both during model training and evaluation.

We additionally perform a correlation analysis of L'AMBRE with perplexity, BLEU and chrF on the WMT system outputs (A.5 in Appendix). As expected, we see a strong negative correlation with perplexity (low perplexity and high L'AMBRE). For BLEU and chrF, the results are quite similar to the correlations with human z-scores.

Diachronic Analysis We present an additional application of L'AMBRE through a diachronic study of translation systems submitted to the WMT news translation tasks. We run our scorer on system outputs from WMT14 (Bojar et al., 2014) to WMT19 (Barrault et al., 2019) for translation models from English to German and Russian.[20] Figure 5 shows the scores of all systems and highlights the average trend of system scores. We also present the scores on the reference translations for comparison. We observe that systems have gotten more fluent over the years, often as good as the reference translations in the most recent shared tasks.[21]

[19] See A.5 for the corresponding scatter plots.

[20] Similar analysis on Czech, Finnish and Turkish in A.6.

[21] Such comparison of well-formedness scores is reasonable, to an extent, even though test sets differ year-to-year, as we measure the grammatical acceptability but not adequacy.

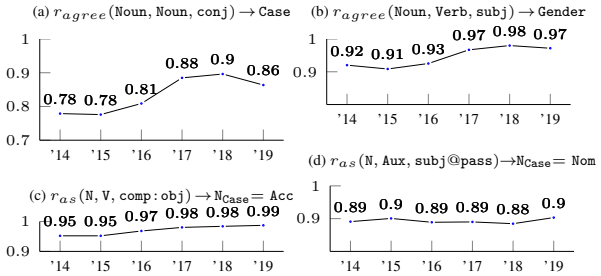


Figure 6: Diachronic analysis of select agreement (r_{agree}) and argument structure (r_{as}) rules in Russian. We report *median* well-formedness score per year. WMT systems have consistently improved on their well-formedness, but some phenomena are still challenging, such as handling agreement across conjuncted nouns (a) or casing in passive constructions (d).

L'AMBRE also allows for fine-grained analysis of NLG systems by identifying specific grammatical issues. We illustrate this through a diachronic comparison of WMT systems for English→Russian on a subset of L'AMBRE's morphosyntactic rules (Figure 6), presenting the median score per rule and year. Such fine-grained analysis reveals interesting trends. For example, while systems have been performing well on some rules over the years (Figure 6 (c)), there are rules that improved only in recent years (Figure 6 (a)). We also identify rules for constructions that remain challenging even for the best systems from WMT19 (Figure 6 (d)).

7 Conclusion and Future Work

In this paper, we introduce L'AMBRE, a framework to evaluate grammatical acceptability of text by verifying morphosyntactic rules over dependency parse trees. We present a method to automatically extract such rules for many languages along with a method to train robust parsing models which facilitate better verification of these rules on natural language text. We demonstrate the practical application of L'AMBRE on the popular generation task of machine translation, focusing on translation into morphologically-rich languages. Directions for future work include (1) incorporating additional morphosyntactic rules (e.g., word order), automatically extracted or hand-crafted ones such as those in Mueller et al. (2020) and (2) building more robust parsers and morphological taggers that are aware of the dependency structure of the sentence.

Acknowledgments

The authors would like to thank Maria Ryskina for help with human evaluation of extracted rules, and Alla Rozovskaya for sharing the Russian GEC corpus with us. This work was supported in part by the National Science Foundation under grants 1761548, 2007960, and 2125201. Shruti Rijhwani was supported by a Bloomberg Data Science Ph.D. Fellowship. This material is partially based on research sponsored by the Air Force Research Laboratory under agreement number FA8750-19-2-0200. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

Duygu Ataman, Wilker Aziz, and Alexandra Birch. 2020. A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. In *International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling,

Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium. Association for Computational Linguistics.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.

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.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687, Seattle, Washington, USA. Association for Computational Linguistics.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

G.G. Corbett. 2006. *Agreement*. Agreement. Cambridge University Press.

Joachim Daiber and Rob van der Goot. 2016. The denoised web treebank: Evaluating dependency parsing under noisy input conditions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 649–653, Portorož, Slovenia. European Language Resources Association (ELRA).

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless quality estimation for natural language generation. In *Proceedings of the 1st Workshop on Learning to Generate Natural Language*, Sydney, Australia.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2019. Improving surface-syntactic Universal Dependencies (SUD): MWEs and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 126–132, Paris, France. Association for Computational Linguistics.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Homa B. Hashemi and Rebecca Hwa. 2016. An evaluation of parser robustness for ungrammatical sentences. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1774, Austin, Texas. Association for Computational Linguistics.

Homa B Hashemi and Rebecca Hwa. 2018. Jointly parse and fragment ungrammatical sentences. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Manfred Krifka. 2003. Case syncretism in German feminines: Typological, functional, and structural aspects. *Ms., ZAS Berlin*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond user self-reported Likert scale ratings: A comparison model for automatic dialog evaluation. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1363–1374, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, 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 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, 40(7):543–566.

Edith A Moravcsik. 1978. *Agreement. Universals of Human Language. Vol 4., ed. by Joseph H. Greenberg, Charles A. Ferguson and Edith Moravcsik*. Stanford: Stanford University Press.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356, Hong Kong, China. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović, Adrià de Gispert, Deepa Gupta, Patrik Lambert, Hermann Ney, José B. Mariño, Marcello Federico, and Rafael Banchs. 2006. Morpho-syntactic information for automatic error analysis of

statistical machine translation output. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 1–6, New York City. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Error-repair dependency parsing for ungrammatical texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

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.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

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.

John Sylak-Glassman. 2016. The Composition and Use of the Universal Morphological Feature Schema (Unimorph Schema).

Oleksiy Syvokon and Olena Nahorna. 2021. UA-GEC: Grammatical error correction and fluency corpus for the ukrainian language.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio. 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.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

A Appendix

A.1 Comparison of UD and SUD

A comparison of the UD and SUD trees for the German sentence from Figure 1 is presented in Figure 7. Unlike the UD parse, the SUD parse directly links the PRON and AUX, allowing for an easy inference of relevant morphosyntactic rules.

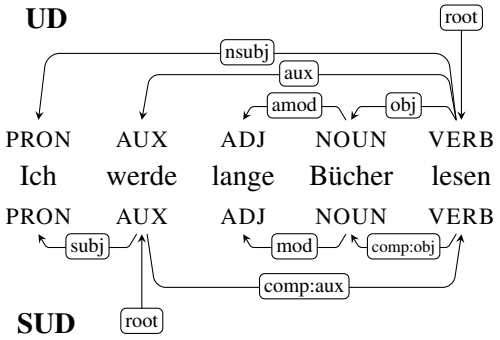


Figure 7: The SUD tree (below) for the sentence "Ich werde lange Bücher lesen" links the auxiliary verb "werde" with its subject "Ich" capturing an agreement rule not present in the UD tree (above).

A.2 Robust parsing

We proposed a methodology to utilize UniMorph dictionaries to add morphology-related noise into UD treebanks. Sometimes, the amount of noise we can add is limited by the number of available paradigms in UniMorph. For example, the Turkish dictionary contains just 3.5k paradigms as compared to 28k in Russian, and we could only corrupt about 55% of the Turkish sentences.

A.3 GEC datasets

In our evaluation on GEC, we only select morphology-related errors in German and Russian GEC datasets. Specifically, we use all errors of the type ⟨POS⟩:form from German Falko-MERLIN GEC corpus. In the Russian RULEC-GEC dataset, we select errors of types Case (Noun, Adj), Number (Noun, Verb, Adj), Gender (Noun, Adj), Person (Verb), Aspect (Verb), Voice (Verb), Tense (Verb), Other (Noun, Verb, Adj) and word form.

The methodology for computing the precision and recall in our GEC evaluation (from Table 2) is presented in algorithm 1.

Algorithm 1: GEC using extracted morphosyntactic rules.
$H(t), G(t)$: estimated and gold error in token t
$h^{(t)}, D^{(t)}$: head and dependents of token t
$f_r(h^{(t)}, t)$: rule r is satisfied in the dependency link between t and $h^{(t)}$

Result: P = $\frac{tp}{tp+fp}$, R = $\frac{tp}{tp+fn}$

$tp, fp, fn = 0, 0, 0$;
for *sent in doc* **do**
  for t *in sent* **do**
    $E(t) = \left\{ t^* : \neg f_r(t^*, t), \forall t^* \in h^{(t)} \cup D^{(t)} \right\}$;
    $H(t) = len(E(t)) > 0$;
    if $H(t)$ **then**
      if $G(t)$ **then**
        tp += 1 ;
      else
        for t^* *in* $E(t)$ **do**
          if $\neg G(t^*)$ **then**
            fp += 0.5 ;
          end
        end
      end
    end
    if $G(t) \wedge \neg H(t)$ **then**
      fn += 1 ;
    end
  end
end

A.4 GEC evaluation

Comparison with Other Metrics We additionally present a comparison of L'AMBRE to other metrics that capture fluency and/or grammatical well-formedness. One metric is perplexity, computed by large language models (LM). Specifically, we use transformer-based LMs (Ng et al., 2019). Second, we use a grammaticality-based metric (GBM) (Napoles et al., 2016) that relies on the open-source Language Tool (Miłkowski, 2010). Language Tool is a widely-used rule-based proofreading software used to detect sentence-level errors. GBM measures error count rate as $1 - \frac{\#\text{errors}}{\#\text{tokens}}$.

To provide a fair comparison of these three methods, we first reframe the GEI task into an acceptability judgment task. Given a source sentence from GEC corpus, we prepare four variants using the annotations provided with the corpus, 1. source sentence itself (no corrections made), 2. morph-corrected sentence (only morphology related corrections are made), 3. rest-corrected sentence (only non-morphology related corrections are made), and

Contrast	L'AMBRE	GBM	PPL
German			
(src, tgt)	0.30 (0.28)	0.63	**0.95**
(src, morph-corrected)	0.31 (0.29)	0.32	**0.70**
(src, rest-corrected)	0.21 (0.19)	0.61	**0.92**
(morph-corrected, tgt)	0.20 (0.18)	0.62	**0.96**
(rest-corrected, tgt)	0.41 (0.39)	0.47	**0.97**
Russian			
(src, tgt)	0.21 (0.20)	0.40	**0.94**
(src, morph-corrected)	0.24 (0.22)	0.14	**0.74**
(src, rest-corrected)	0.12 (0.12)	0.43	**0.88**
(morph-corrected, tgt)	0.18 (0.16)	0.46	**0.95**
(rest-corrected, tgt)	0.35 (0.33)	0.18	**0.94**

Table 4: Accuracy results for L'AMBRE, grammaticality-based metric (GBM), and perplexity (PPL) on various contrastive acceptability judgments on German and Russian GEC (*test* splits). Best score is in **bold**, and the second best score is underlined. Numbers in parentheses are obtained by using original Stanza parsers instead of the proposed robust parsers.

4. target sentence (all corrections made).

To evaluate the effectiveness of the three metrics, we make 5 contrastive comparisons as shown in Table 4. For instance, in the comparison (src, tgt), $\forall$ src $\neq$ tgt, we check if L'AMBRE(src) $<$ L'AMBRE(tgt), GBM(src) $<$ GBM(tgt) and PPL(src) $>$ PPL(tgt).[22] Table 4 presents the accuracy results across the 5 contrastive pairs on the *test* splits of German and Russian GEC corpora. Overall, perplexity performs much better than the other metrics across all pairs. However, unlike GBM and L'AMBRE, perplexity doesn't provide error diagnosis, with no feedback on incorrect grammatical rules. Between L'AMBRE and GBM, former is competitive or better at two pairs, (src, morph-corrected) and (rest-corrected, tgt), whereas the latter does better at two other pairs, (src, rest-corrected), (morph-corrected, tgt). These results indicate that the proposed metric, L'AMBRE, is good at capturing morpho-syntactic rules necessary for grammatical correctness (especially in Russian), and the more complex GBM does better at other fluency related rules. Additionally, we observed clear improvements by using our proposed robust parsers (§4) over the original stanza parsers.

In our re-implementation of the GBM, we follow the prior work (Napoles et al., 2016) and utilize Language Tool (LT) for error detection. We use LT for two languages, German (de-DE: Germany) and Russian (ru-RU). The source sentences in both the GEC corpora are pre-tokenized, therefore, we skip whitespace-based rules while using LT. For Russian, we remove whitespace-based rules corresponding to comma, punctuation and hypen. For German, we remove whitespace-based rules corresponding to quotation mark, exclamation mark, unit spaces, comma and parentheses. In the German GEC test split, the total counts of each contrastive pairs (x, y) with sent(x) $\neq$ sent(y), (src, tgt): 1791, (src, morph-corrected): 1169, (src, rest-corrected): 1646, (morph-corrected, tgt): 1582, and (rest-corrected, tgt): 831. In the Russian GEC split, the total counts of each contrastive pairs (x, y) with sent(x) $\neq$ sent(y), (src, tgt): 2381, (src, morph-corrected): 1405, (src, rest-corrected): 2005, (morph-corrected, tgt): 1913, and (rest-corrected, tgt): 1148.

[22] These strict inequalities allow us to capture limitations of rule-based methods. An error might be undetectable if the corresponding rule is absent in the method's rule set.

A.5 WMT Correlation Studies

English→◌	#sys†	r_{agree}	r_{as}	$r_{agree} \cup r_{as}$
WMT'18				
Czech	5	0.91	0.78	0.84
German	12	0.07	-0.10	-0.06
Estonian	12	0.83	0.62	0.68
Finnish	12	0.96	0.77	0.86
Russian	7	0.71	0.89	0.86
Turkish	7	0.64	-0.31	0.58
WMT'19				
Czech	11	0.89	0.70	0.80
German	20	0.14	0.13	0.16
Finnish	12	0.87	0.82	0.85
Russian	11	0.70	0.45	0.57

Table 5: With a few exceptions, our grammar-based metrics correlate well with human evaluations of WMT18 and WMT19 systems. (Pearson's r against Z-scores). †:we remove outlier systems following Mathur et al. (2020). Results using **robust** Stanza parsers.

Correlation with Human z-scores: In Table 5 we present a detailed account of the Pearson's r correlations between human z-scores and L'AMBRE for systems in WMT'18 and WMT'19. We also present the correlations with original Stanza parsers in Table 6. In Figure 11 and Figure 12, we present the scatter plots comparing human z-scores and L'AMBRE for WMT'18 and WMT'19 respectively.

English→◌	#sys†	r_{agree}	r_{as}	$r_{agree} \cup r_{as}$
WMT'18				
Czech	5	0.88	0.85	0.87
German	12	-0.08	0.02	0.04
Estonian	12	0.84	0.68	0.74
Finnish	12	0.96	0.73	0.85
Russian	7	0.66	0.84	0.90
Turkish	7	0.64	-0.86	0.51
WMT'19				
Czech	11	0.86	0.46	0.63
German	20	0.44	0.10	0.17
Finnish	12	0.85	0.81	0.84
Russian	11	0.74	0.58	0.69

Table 6: Correlations with human evaluations of WMT18 and WMT19 systems. (Pearson's r against Z-scores). †:we remove outlier systems following Mathur et al. (2020). Results using **original** Stanza parsers.

English→◌	#sys†	BLEU	chrF
WMT'18			
Czech	5	0.83	0.81
German	12	0.18	0.14
Estonian	12	0.75	0.66
Finnish	12	0.89	0.86
Russian	7	0.85	0.84
Turkish	7	0.61	0.66
WMT'19			
Czech	11	0.85	0.83
German	20	0.01	0.01
Finnish	12	0.86	0.85
Russian	11	0.65	0.53

Table 7: Correlations with BLEU, chrF for WMT'18 and WMT'19 systems using all the rules. (Pearson's r against L'AMBRE). †:we remove outlier systems following Mathur et al. (2020). Results using **robust** Stanza parsers.

Correlation with other metrics In Table 7 we present a comparison of L'AMBRE with BLEU (Papineni et al., 2002) and chrF (Popović, 2015). In Figure 9 and Figure 10, we present scatter plots comparing perplexity and L'AMBRE for WMT systems from WMT'14 to WMT'19. To use perplexity as a corpus fluency measure, we first compute perplexity of each output translation and then take an average over all sentences in the target test set to obtain a corpus perplexity score for each WMT system. On most occasions, as expected, we see a negative correlation between perplexity and L'AMBRE, more strongly in Russian than in German.

A.6 Diachronic analysis of WMT systems

Figure 8a presents a diachronic study of WMT systems for Czech, Finnish, and Turkish using L'AMBRE. Figure 8b shows the morpho-syntactic rule specific trends for Russian WMT.

A.7 Rule Extraction Statistics

For extracting agreement (r_{agree}), case assignment and verb form choice (r_{as}) rules, we use the largest available treebank for the language from SUD. Table 8 presents the rule counts for the languages discussed in this paper.

Treebank	# r_{agree}	# r_{as}
Czech-PDT	28	33
German-HDT	30	23
Greek-GDT	11	12
Estonian-EDT	22	31
Finnish-TDT	19	35
Russian-SynTagRus	25	35
Turkish-IMST	32	6

Table 8: Statistics of rules extracted from SUD treebanks.

A.8 Reproducibility Checklist

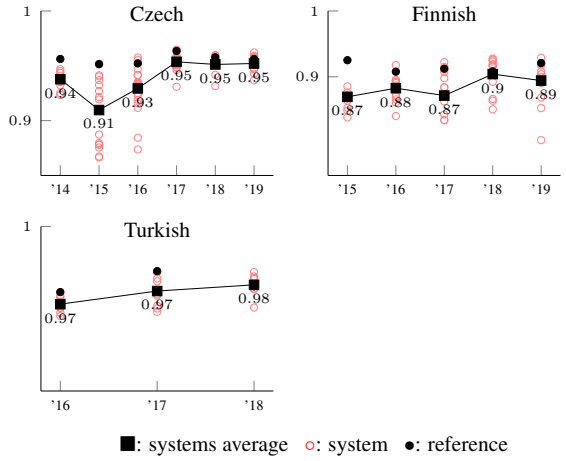
A.8.1 Model Training

For training robust dependency parsers, we use the training infrastructure provided by Stanza authors.[23] We use the same set of language-specific hyperparameters as the original Stanza parsers and taggers. All our training is performed on a single GeForce RTX 2080 GPU.

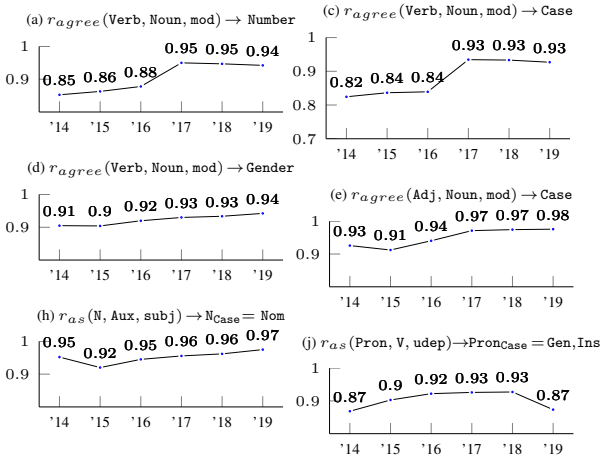
A.8.2 Resources

In this work we use WMT metrics dataset,[24] WMT human evaluation scores,[25] SUD treebanks,[26] UD2SUD converter.[27]

[23] https://stanfordnlp.github.io/stanza/training.html
[24] http://www.statmt.org/wmt19/metrics-task.html
[25] http://www.statmt.org/wmt19/results.html
[26] https://surfacesyntacticud.github.io/data/
[27] https://github.com/surfacesyntacticud/tools

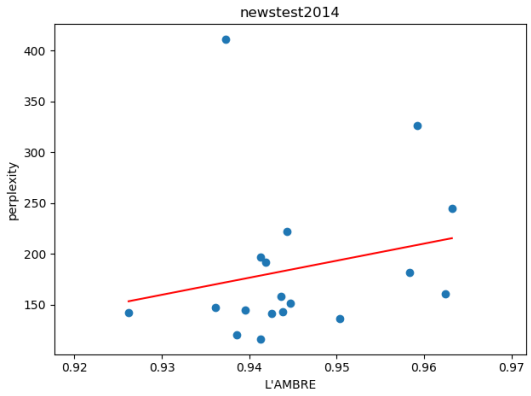


(a) A diachronic study of grammatical well-formedness of WMT English→X systems' outputs. The systems in general are becoming more fluent with very passing year. In the last two years the best systems produce as well-formed outputs as the reference translations.

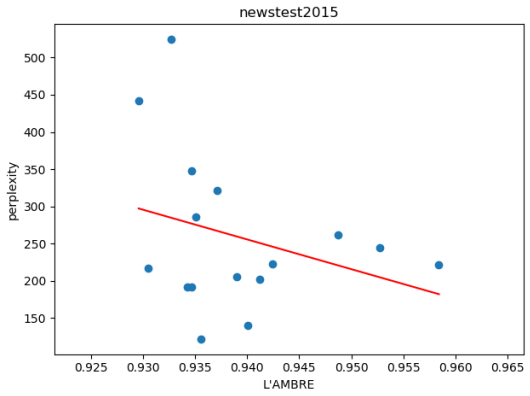


(b) Diachronic analysis of additional agreement (r_{agree}) and argument structure (r_{as}) rules in Russian WMT. We report the *median* well-formedness score for each WMT year.

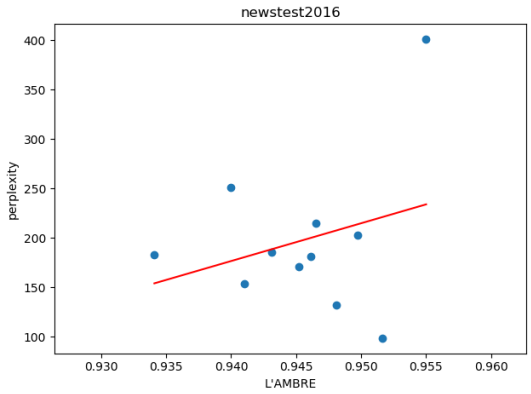
Figure 8: Diachronic study of the grammatical well-formedness of WMT systems.



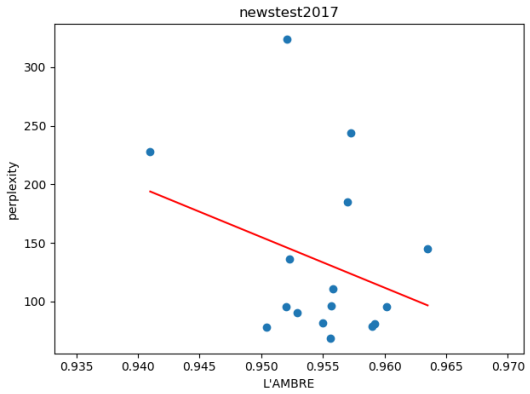
(a) German WMT'14



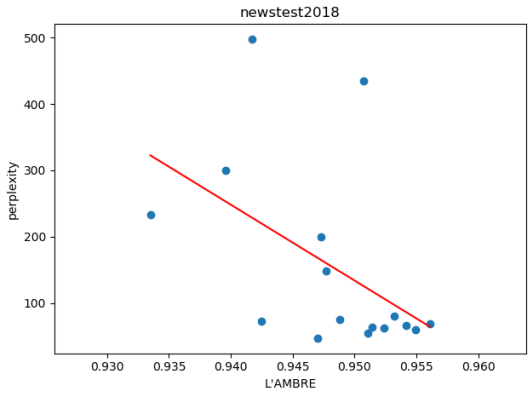
(b) German WMT'15



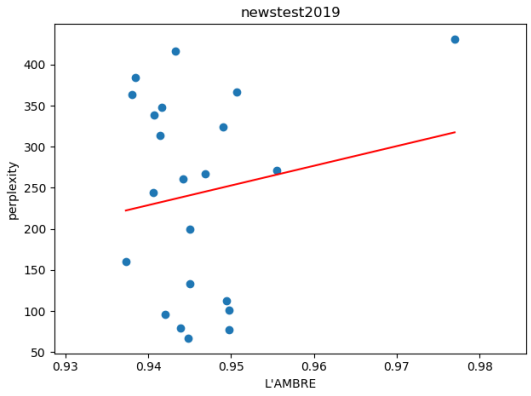
(c) German WMT'16



(d) German WMT'17

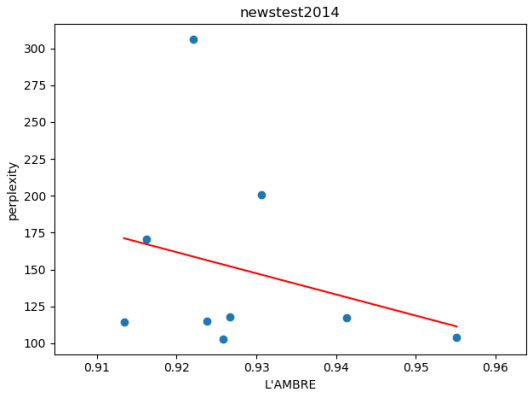


(e) German WMT'18

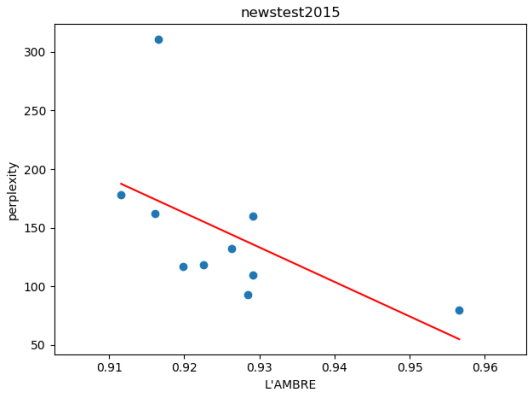


(f) German WMT'19

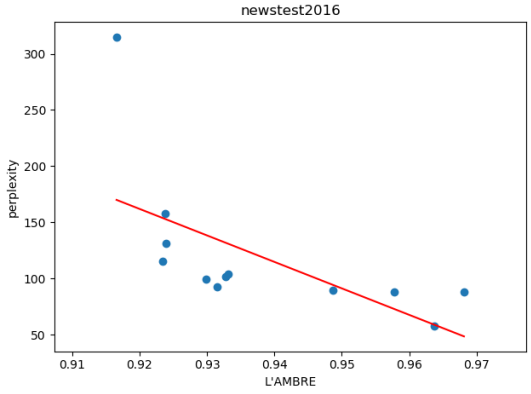
Figure 9: Scatter plot of perplexity and L'AMBRE for German WMT systems from WMT'14→'19. High L'AMBRE and low perplexity indicate better systems. As expected, we see a negative correlation between the two metrics for WMT'15, WMT'17, and WMT'18. But for WMT'14, WMT'16, and WMT'19, we see positive correlations, indicating potential limitations of our metric.



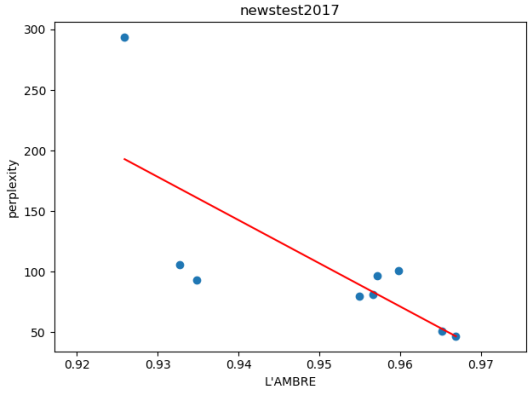
(a) Russian WMT'14



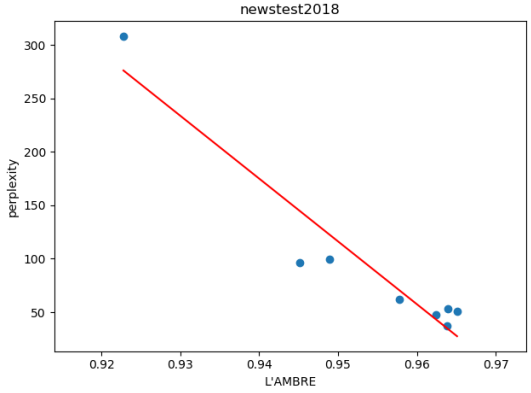
(b) Russian WMT'15



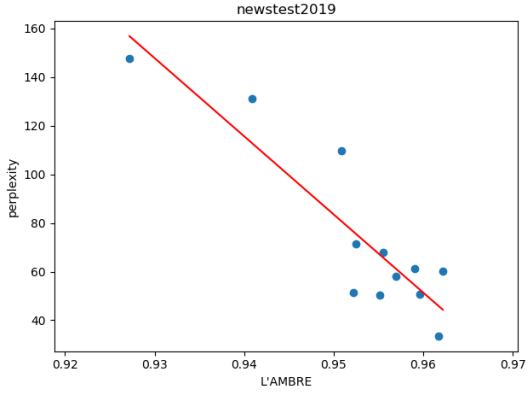
(c) Russian WMT'16



(d) Russian WMT'17

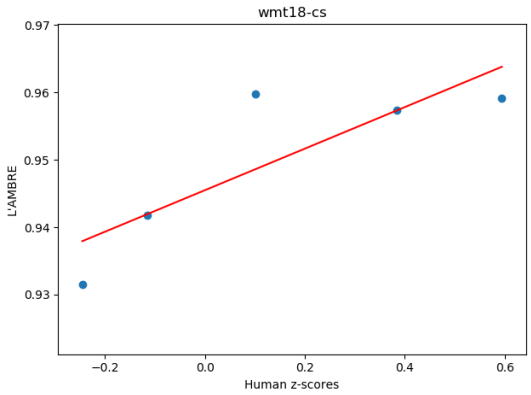


(e) Russian WMT'18

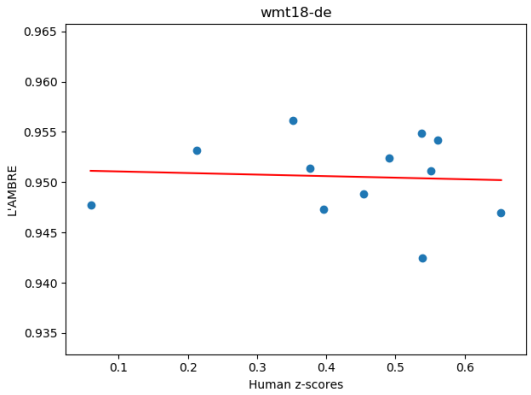


(f) Russian WMT'19

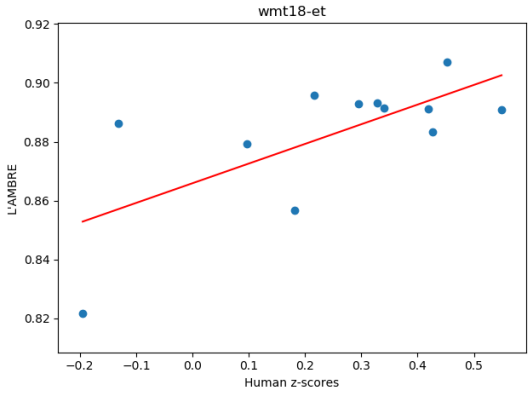
Figure 10: Scatter plot of perplexity and L'AMBRE for Russian WMT systems from WMT'14-'19. High L'AMBRE and low perplexity indicate better systems. As expected, we see negative correlation between the two metrics across the years.



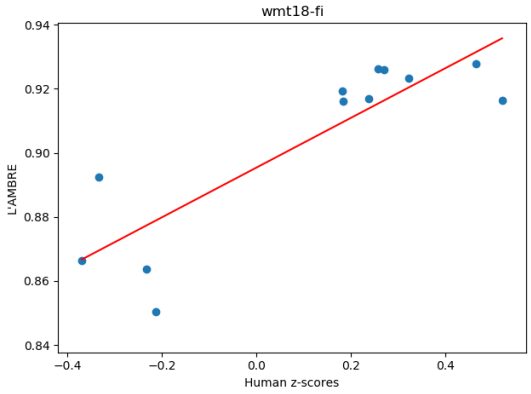
(a) Czech WMT'18



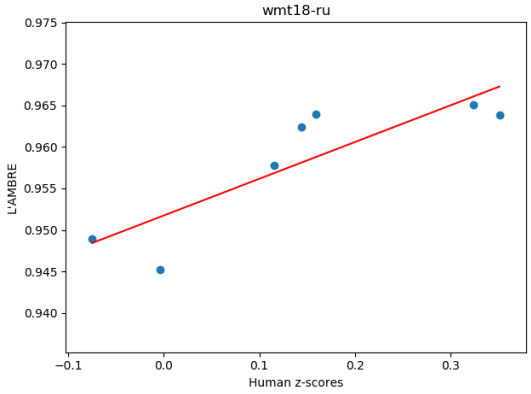
(b) German WMT'18



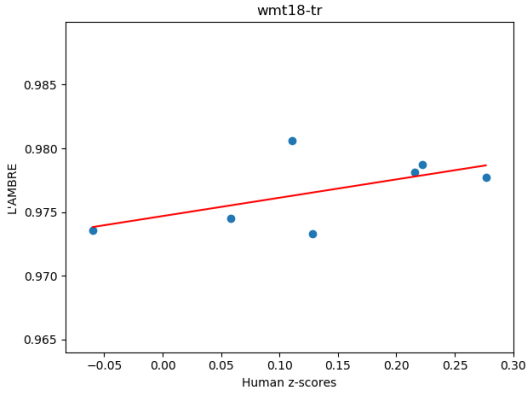
(c) Estonian WMT'18



(d) Finnish WMT'18

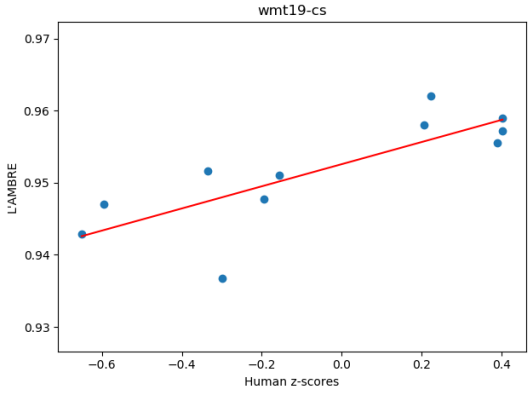


(e) Russian WMT'18

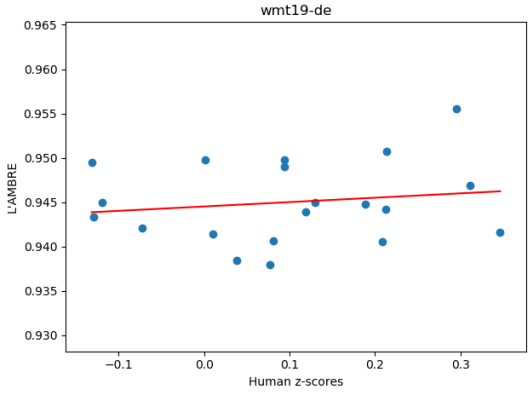


(f) Turkish WMT'18

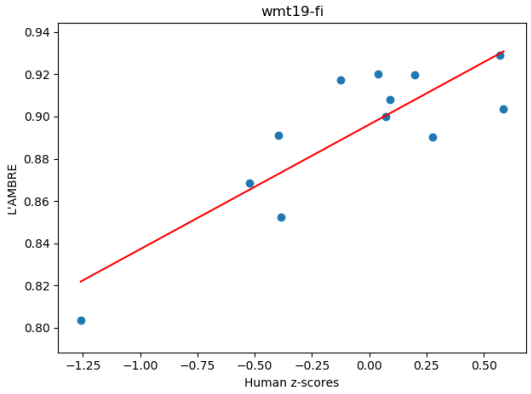
Figure 11: Scatter plot of human z-scores and L'AMBRE for WMT'18 systems.



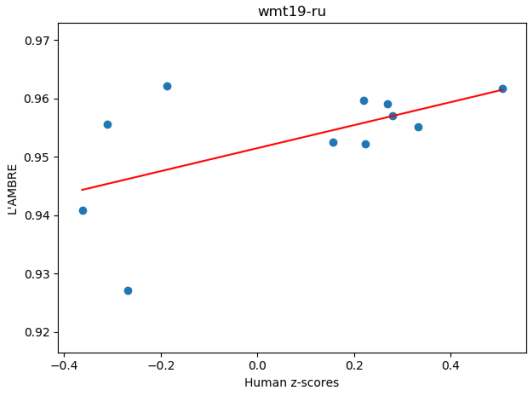
(a) Czech WMT'19



(b) German WMT'19



(c) Finnish WMT'19



(d) Russian WMT'19

Figure 12: Scatter plot of human z-scores and L'AMBRE for WMT'19 systems.