Quantifying Synthesis and Fusion and their Impact on Machine Translation

Anonymous ACL submission

Abstract

Theoretical work in morphological typology offers the possibility of measuring morphological diversity on a continuous scale. However, literature in NLP typically labels a whole language with a strict type of morphology, e.g. fusional or agglutinative. In this work, we propose to reduce the theoretical rigidity of such claims, by quantifying the morphological typology at the word and segment level. We consider Payne (2017)'s approach to classify morphology using two indices: synthesis (from 1 for analytic to 3 or more for polysynthetic) and fusion (from 0 for agglutinative to 1 for fusional). For computing synthesis, we test unsupervised and supervised morphological segmentation methods for English, German and Turkish, whereas for fusion, we propose a semi-automatic method using Spanish as a case study. Then, we analyse the relationship between machine translation quality and the degree of synthesis and fusion at word (nouns and verbs for English-Turkish, and verbs in English-Spanish) and segment level (previous language pairs plus English-German in both directions). We complement the word-level analysis with human evaluation, and overall, we observe a consistent impact of both indexes on machine translation quality.

1 Introduction

One of the first barriers to develop language technologies is morphology, i.e., how systematically diverse their word formation processes are. For instance, agglutination and fusion are two morphological kind of processes that concatenate morphemes to a root with explicit or non-explicit boundaries, respectively. Processing morphologically-diverse languages and evaluating morphological competence in NLP models is relevant for language generation and understanding tasks, such as machine translation (MT). It is unfeasible to develop models with capacity large enough to encode the full vocabulary of every language, and it is a must to rely on subword segmentation approaches that help to constrain the capacity when generating rare, or even new words (Sennrich et al., 2016). Hence, understanding morphology is essential to develop robust subword-based models and evaluate the quality of their outputs (Vania and Lopez, 2017). Nevertheless, there is a potential gap between the probing of whether an NLP model can handle "morphological richness", and what is a proper measure of "morphological richness" from linguistic typology.

In most of the recent NLP literature, different types of languages (e.g. agglutinative, polysynthetic) are chosen to test a more diverse handling of morphological richness (Ponti et al., 2019). There is, however, a debate as to whether languages can indeed be classified into discrete morphological categories. Payne (2017) provided a morphological typology measurement in a continuous spectrum using the indices of synthesis and fusion. Synthesis measures if a segment is highly analytic or synthetic (from 1 to more), whereas fusion measures whether it is highly agglutinative or fusional (from 0 to 1). And surprisingly, with respect to the NLP literature, it is possible to identify English sentences with a very low fusion index, meaning that they are highly agglutinative[1].

From a more applied perspective, if the references of an evaluation set (in any language generation task) are labelled with the indices, we could perform a stratified analysis (e.g. low fusion and high fusion) to determine how well an NLP model handles morphology for multiple languages. For example, we could assess whether a machine translation model is failing in generating more fusional than agglutinative segments for a specific target language. Knowing and quantifying that problem

[1]For instance, in the following fragment (Payne, 2017), the index of fusion is 1/8 or 0.125 (fusional morpheme joints are marked with a dot and the rest with a hyphen): "The company-'s great break-through came.PAST when they decided to buy trike-s to sell their ice cream around the street-s in the nine-teen twenty-s".

concerning morphology is the first step towards proposing a solution. Our contributions then are listed as follows:

- We present the first computational quantification of synthesis and fusion using standard NLP evaluation sets.
- We analyse the relationship between the two indices and machine translation quality at word-level, and observe that a higher degree of synthesis or fusion usually corresponds to less accurate translations in specific word types (studying nouns and verbs in English-Turkish, and verbs in English-Spanish).
- We complement this evaluation with manual annotation of synthesis and fusion[2].
- We extend the analysis at segment-level, using the aforementioned language pairs plus English-German in both directions, and identify that some synthesis and fusion-based predictors are significant for MT system outputs.

2 Background and related work

2.1 Morphological typology

Early approaches to morphological typology tended to characterise languages in a holistic way, in terms of their word formation strategies, such as agglutination or fusion (Sapir, 1921). First was the idea that languages can be characterised holistically and unambiguously in terms of their word and sentence-building processes, but different studies started to quantify these strategies, such as in Payne (2017), that recently argued about synthesis and fusion, which are defined as follows.

2.1.1 Synthesis

The index of synthesis offers a scale to contrast highly analytic or synthetic languages. This implies whether a word is composed by one (analytic) or several (synthetic) morphemes (Payne, 2017). Synthesis can be computed as the ratio of number of morphemes per words, it is closer to 1 when the language is more analytic (e.g. Mandarin, or English to a less degree), and gets higher the more synthetic the language is (e.g. Turkish, Inuktitut). Polysynthesis can be present when the synthesis degree is higher than 3, although the boundary is arguable. Besides, as we claim in this study, any language can present different levels of synthesis if we evaluate them at a more fine-grained level.

[2]All annotated data will be released upon acceptance.

2.1.2 Fusion

Fusion is the ratio of the fusional morphemes joints[3] per the total number of joints. This index goes from 0 to 1, or from highly agglutinative (e.g. Turkish) to highly fusional (e.g. Spanish) cases. However, we noticed that the computation of fusion is complex to automatiseFor instance, Payne (2017) indicates potential cases to identify fusional joints, such as in prefixes, suffixes, infixes, circumfixes, compounding, non-concatenation processes (reduplication, apophony, substractive morphology) or autosegmental morphemes. Current automatic tools are not designed to identify these cases for most languages.

2.2 Morphological typology on NLP

A survey by Ponti et al. (2019), on computational typology for NLP, pointed out that morphological knowledge is potentially helpful for analysing the difficulty in generation tasks such as language modelling and neural MT for both unsupervised and supervised settings. More specifically, they suggested that the degree of fusion (related to the index of fusion proposed by Payne (2017)) impacts in the rate of less frequent words, which is a relevant parameter for generation tasks.

Besides, the studies that address morphological typology are related to either the development of morphological analysis systems or the evaluation of typologically diverse languages in terms of morphology (Vania and Lopez, 2017; Xu et al., 2020). However, the typology used to distinguish languages varies across different studies. For instance, Vania and Lopez (2017) considers four phenomena to label languages: fusionality, agglutination, reduplication and root-pattern; whereas Xu et al. (2020) considers more fine-grained elements such as affixation (prefixation, infixation and suffixation) or partial reduplication. It is important to note that none of the previous studies have addressed the phenomena as an index but rather as a discrete label for a language.

Furthermore, other studies refer only to morphological typological features as part of the task of typological feature prediction from linguistic databases (Bjerva and Augenstein, 2018; Bjerva et al., 2020).

[3]Or how many grammatical, syntactic and semantic features are joint. More than one feature can be fused in a single morpheme.

2.3 Morphological segmentation and analysis

Morphological segmentation was first introduced by Harris (1951). Unsupervised methods are popular with the morfessor (Creutz and Lagus, 2002, 2007; Poon and Domingos, 2009) family of methods, including semi-supervised versions (Kohonen et al., 2010; Grönroos et al., 2014). Also, Adaptor Grammars have been applied with great success to the task (Eskander et al., 2019). Besides, supervised methods have achieved the best results, such as pointer generator networks (Mager et al., 2020).

Besides, the most widespread unsupervised segmentation methods (Byte-Pair-Encoding (BPE; Sennrich et al., 2016) and a method based on unigram language modelling (Kudo, 2018)) are not linked at all to morphological segmentation, but they are used to constrain the vocabulary size for neural generation tasks.

Finally, it is important to note that the index synthesis can be computed with a robust morphological analyser or segmentation model (to count the number of morphemes), but neither of them are built to compute the index of fusion directly.

3 How to compute Synthesis and Fusion?

3.1 Synthesis: automatic computation

To automatically compute the index of synthesis, we require to perform a robust morphological segmentation. A rule-based morphological analyser and disambiguator might be the best option if available (which we use later for Turkish in §4.2), but for the purpose of the study, we compare well-known supervised and unsupervised methods:

- Byte-Pair-Encoding (BPE) and Unigram Language Model (uniLM)[4] from SentencePiece (Kudo and Richardson, 2018).
- Morfessor (Poon and Domingos, 2009).
- Pointer Generator Network (PtrNet) from the implementation of Mager et al. (2020).

3.1.1 Datasets and evaluation

We used the CELEX dataset of segmented words for English and German (Steiner, 2016, 2017), where we randomly split training and evaluation data (80-10-10). Besides, for the unsupervised methods, we use the newscommentary-v15 (Barrault et al., 2019) and EuroParl-v10 (Koehn, 2005) corpora[5]. Furthermore, we define two metrics to assess the performance on computing synthesis:

- Accuracy count: Evaluates if the number of obtained morphemes in the hypothesis segmentation is the same as in the reference.
- Exact segmentation precision: Analyses if the split morphemes are the same. We first perform an automatic alignment between the hypothesis and reference segments with the parallel Needleman-Wunsch algorithm for sequences (Naveed et al., 2005), and then compute the exact match at morpheme level.

	English				German			
#morphs.	1	2	3	4	1	2	3	4
	16,914	28,900	1,798	73	13,061	32,007	5,808	360
	Accuracy Count							
uniLM$_{64k}$	0.54	0.52	0.49	0.59	0.35	0.27	0.21	0.18
BPE$_{64k}$	0.5	0.53	0.5	0.52	0.29	0.33	0.28	0.26
Morfessor	0.22	0.47	0.55	0.48	0.17	0.26	0.28	0.25
PtrNet	**0.82**	**0.84**	**0.56**	**0.81**	**0.74**	**0.86**	**0.7**	**0.42**
	Exact Segmentation Precision							
uniLM$_{64k}$	0.54	0.52	0.6	0.8	0.29	0.38	0.32	0.22
BPE$_{64k}$	0.5	0.44	0.56	0.76	0.24	0.33	0.23	0.08
Morfessor	0.21	0.58	**0.7**	0.78	0.17	0.45	0.44	0.36
PtrNet	**0.76**	0.67	**0.81**	**0.8**	**0.67**	**0.73**	**0.72**	**0.62**

Table 1: Accuracy count and segmentation precision for English and German using unsupervised and supervised segmentation methods. Results are grouped by the expected number of morphemes (e.g. "1" means that the word should not be split).

3.1.2 Results and discussion

Table 1 shows the scores on morphological segmentation for both English and German. For uniLM and BPE. we observe that they under-perform when it is not expected to split the word (column "1"). This is a pattern observed by Bostrom and Durrett (2020), where they noted that unsupervised segmentation methods tend to over split the roots of words. They both improve their accuracy and precision when the number of morphemes expected is larger. Unexpectedly, Morfessor also under-performs in case "1" for both languages, and only surpasses the other unsupervised methods when we measure precision for many morphemes. Furthermore, The PtrNet supervised method outperforms the rest in almost all scenarios.

We conclude that, to compute synthesis, we should prioritise, besides a rule-based morpholog-

[4] We analysed several vocabulary sizes (4k, 8k, 16k, 32k, 64k) but report only the best one, which is 64k for all cases.

[5] Other languages like Danish are also available and was tested, but we did not report the results here as there is not complementary machine translation evaluation sets.

ical analyser, a supervised segmentation method like PtrNet if data is available. We take advantage of this for the segment-level analysis in §5.

3.2 Fusion: Semi-automatic computation

Calculating fusion should be approached in a case by case scenario, as there are different considerations provided by Payne (2017). Therefore, there is not an automatic tool that can obtain the fusion score directly. We decided to focus on Spanish[6] as a case study, where verbs and auxiliary verbs contains the highest degree fusion of all the parts-of-speech (POS).

Procedure We observed that we could perform an annotation per paradigm and the termination of the verb (-ar, -er, -ir), as the fusion degree will remain the same regardless of the lemma[7]. Then, on a chosen Spanish corpus:

1. Perform an automatic annotation of POS and morphological features[8].
2. Review the automatic annotation of special cases. For instance, there are specific verb forms that are missed as adjectives. We corrected the POS and morphological annotation of those cases in a manual step.
3. Obtain a set of all unique verb paradigms and morphological features in the corpus, considering the three different types of verb terminations in Spanish as different elements[9].

Now there is a list of unique verb paradigms and terminationsthat can be annotated both in synthesis and fusion. The steps are as follows:

1. For each unique verb paradigm and termination, segment a verb sample into its morphemes. E.g. the verb *habló* ('talked'), is split in *habl-ó*, and *habláramos* ('we were to speak') in *habl-ára-mos*.
2. Analyse how many morphological features are fused in each morpheme: if you change a value of a feature, will the surface form or morpheme will change? E.g. in *habl-ó*, *-ó* participates in 5 features (mode (indicative), subject person (third person), subject number (singular), tense (past) and aspect (perfective)). For *habl-ára-mos*, *-ára* includes the past and subjunctive, whereas *-mos* denotes the person and number. If any of aforementioned feature changes its value, the surface will change too.
3. Count and aggregate the results per morphemes and obtain the fusion for each verb paradigm. E.g. the fusion for *habl-ó* is 4/5 = 0.8, and for *habl-ára-mos* is 2/4 = 0.5.

Finally, with the annotation in the unique list of verb inflections and terminations, we can extend the degree of fusion to all the verbs in the original Spanish corpus.

4 Word-level analysis of Synthesis and Fusion in Machine Translation

In this analysis, we ask the following question: how difficult is translating a word concerning its index of synthesis or fusion? For evaluating synthesis, we work with Turkish[10] nouns and verbs, and for fusion, we keep working on Spanish verbs. For both cases, English is the source language in the translation task.

4.1 Experimental design

The experiment consists of comparing a gold standard reference with machine translation system outputs at the word level:

1. For both the reference and system output, we automatically **tag all the words with a morphological analyser** (the Boun morphological analyser and disambiguator (Sak et al., 2008) for Turkish and an spaCy model trained on the Ancora Universal Dependency parser (Taulé et al., 2008) for Spanish). The POS is needed to filter the target words. For synthesis in Turkish, the number of morphemes works as a proxy, as we are working at the word level. For fusion in Spanish, we need the inflection to obtain the degree of fusion from the annotated unique list (see 3.2).

[6] We chose this language because of the ease of finding annotators and MT training and evaluation data.

[7] Except for irregular ones, which presents a limitation and potential noise. To reduce the risk of a biased assessment, we also performed a human evaluation.

[8] We use the spaCy model `es_dep_news_trf`, available at `https://spacy.io/models/es#es_dep_news_trf`. It has an accuracy of 0.99 in POS and morphological tagging in the UD Spanish AnCora dataset (Taulé et al., 2008), which contains news texts mostly.

[9] Using the Unimorph database (McCarthy et al., 2020) is another alternative for extracting all the possible unique inflections (at least the ones that are annotated), but would have required an extra aligning step of the Unimorph and spaCy tag sets.

[10] Turkish presents high synthesis and agglutination (Zingler, 2018), meaning that there are words composed with several morphemes and the morpheme boundaries are explicit, respectively. We focus on verbs and nouns, which usually contain more morphemes than other parts-of-speech. We chose this language due to the availability of an open-source rule-based morphological analyser and an expert annotator.

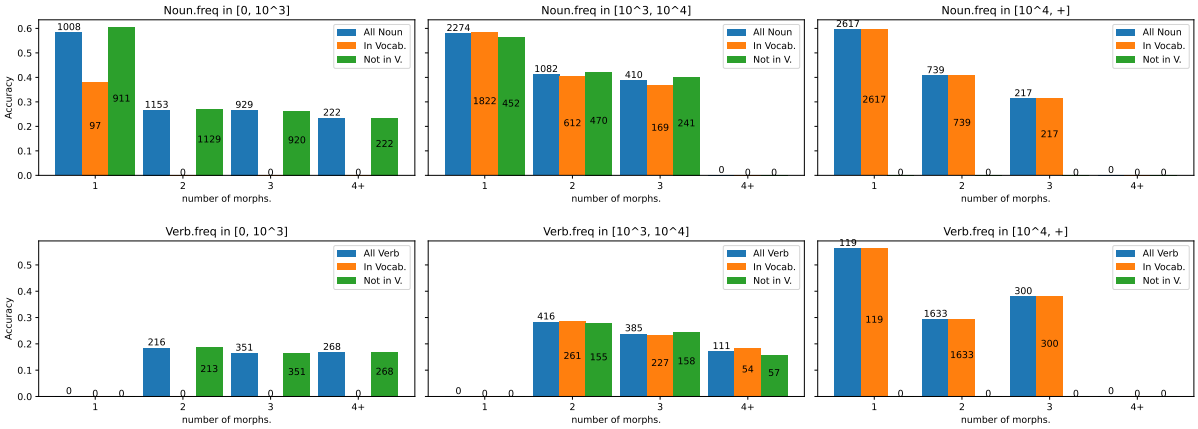

Figure 1: Accuracy (exact translation) for Nouns (top) and Verbs (bottom) in the English→Turkish translations. Results are grouped by the training frequency of the words (less to more frequent from left to right), and each subplot presents the scores for all the words, and whether they belong or not to the vocabulary input of the model. The number of samples are stacked in each bar, and we do not show entries with less than 30 samples.

2. **Align the words** between the reference and system output. We use the awesome-align (Dou and Neubig, 2021) tool by fine-tuning the multilingual BERT (Devlin et al., 2019) model for word-alignment, using the reference and system output as parallel corpora.
3. Calculate the translation accuracy (exact match of the word, 0 or 1) for the target POS.

We then fine-grain the results concerning the degree of synthesis (number of morphemes) or fusion. Additionally, we control different confounds: frequency of the word in the training set, and whether the full word is part of the vocabulary input of the model or not. Finally, we complement the analysis with a human evaluation (see §4.4).

4.2 Synthesis analysis: English→Turkish

Data We use the NEWSTEST2018.EN-TR evaluation set from WMT (Bojar et al., 2018), with 3,000 samples. In the Turkish side there are 45,944 tokens, and Table 2 shows the distribution of the number of morphemes obtained with Sak et al. (2008).

Model We use an English-Turkish system trained with the TIL corpus of 39.9M parallel sentences (Mirzakhalov et al., 2021). On the NEWSTEST2018.EN-TR set, the performance is 13.06 and 49.54 in BLEU and chrF, respectively.

Results and discussion Figure 1 shows the average accuracy (exact translation, 0 or 1) of nouns and verbs in NEWSTEST2018.EN-TR, where the number of morphemes is a proxy for the index of synthesis. In most cases, especially with a higher training frequency, we observe that the average accuracy drops as the number of morphemes increases from 1 to more. This is clearer in nouns than in verbs, which have fewer cases to analyse overall. Between 2, 3 or more than 4 morphemes the differences are not significant, and sometimes is not consistent (e.g. verbs with the highest frequency). However, we can argue that analytic nouns (synthesis=1) are easier to translate than synthetic nouns (synthesis>1) for the English→Turkish direction. The pattern holds for whether the word is part of the vocabulary of the model or not, although rare words (frequency in $[0, 10^3]$ have generally lower translation accuracy than more frequent words (frequency > 100).

	Total	#1	#2	#3	#4	#5+
Verbs	3,834	133	2,265	1,036	308	92
Nouns	10,680	5,899	2,974	1,556	244	7

Table 2: Number of nouns and verbs in the Turskish reference set, and their respective number of morphemes.

4.3 Fusion analysis: English→Spanish

Data We use the NEWSTEST2013.EN-ES evaluation set from WMT (Bojar et al., 2013) with 3,000 samples. In the Spanish side there are 62,055 tokens, with 6,317 verbs, and where 1,411 of them are more agglutinative (fusion=0) and 4,822 more fusional (fusion>0).

Model For training, we use the MarianNMT toolkit (Junczys-Dowmunt et al., 2018), a Transfomer-base model (Vaswani et al., 2017) with default parameters, and four NVIDIA V100 GPUs. We obtained different English-Spanish mod-

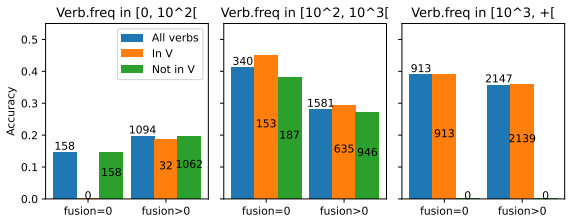

Figure 2: Accuracy (exact translation) for Verbs in the English→Spanish translations. Results are grouped by the training frequency and whether the word belongs to the vocabulary of the model (In V) or not (Not in V).

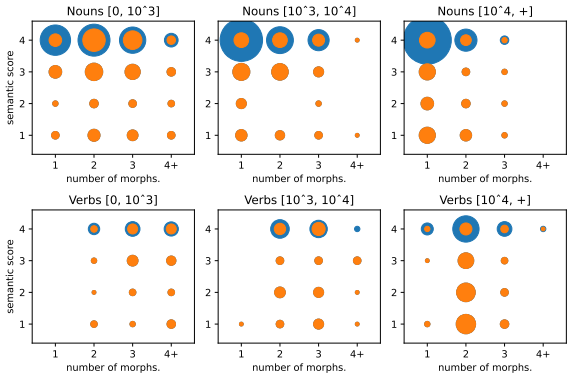

Figure 3: Semantic score annotation for Turkish. Bubbles represent the amount of annotations per score and their respective group. The orange inner bubble represents the amount of samples with 'zero' accuracy (in the automatic analysis) in each category.

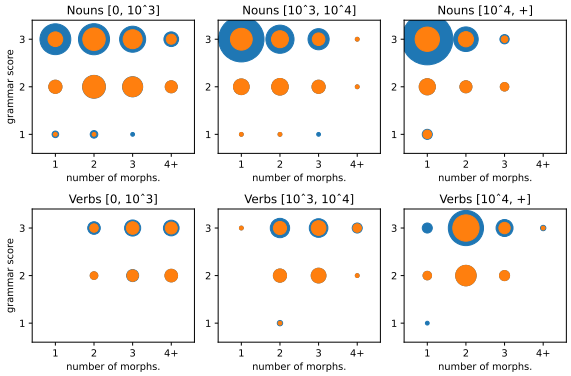

Figure 4: Grammar score annotation for Turkish.

els using the newscommentary-v8 (Bojar et al., 2013) and EuroParl (Koehn, 2005) datasets with joint vocabulary sizes of 8k, 16k and 32k (using unigram-LM from SentencePiece (Kudo and Richardson, 2018)). For this analysis, we chose the best performing system: combining both datasets (2.2M sentences) with 16k pieces. On NEWSTEST2013.EN-ES, the performance is 31.6 BLEU points.

Results and discussion Figure 2 shows the average accuracy of verbs in NEWSTEST2013.EN-ES for verbs without and with some degree of fusion. In the two higher frequency subplots (middle and right), we can observe that the average accuracy of the non-fusional verbs is higher than the fusional ones, and the pattern holds whether the verb is present in the vocabulary input of the model or not. The exception is for the least frequent verbs, although this is explained as the model do not have enough information to learn from, regardless of their degree of fusion.

4.4 Human evaluation

Exact translation accuracy has limitations, as there are potential translations that could be acceptable given a specific context (e.g. a synonym). For that reason, we performed a human evaluation of a sample of sentences on (10%) of each evaluation set, focusing on two scores[11]:

1. Semantic score: evaluates the meaning of the word used in the automatic translation (system output) and how it compares with the gold standard translation. Scale goes from 1 (no relationship at all) to 4 (it is the same lemma).
2. Grammar score: evaluates the grammatical form and how it compares with the gold standard translation. Scale goes from 1 (different inflection) to 3 (same inflection).

Synthesis In Figures 3 and 4 we show the annotation scores for the semantic and grammar metrics, respectively, for both Nouns (top) and Verbs (bottom). We also divide the analysis w.r.t. the frequency of the word in the training data. For Nouns, we observe similar patterns as in the automatic analysis, where the amount of words with one morpheme (synthesis=1) has a higher semantic or grammar score than the rest, suggesting they are easier to generate for the model, except in the least frequent block, which still cannot be well translated. The Verbs tend to have more distributed scores suggesting the difficulty of generating inflected forms may remain equally high even when the words are more frequent. Single morpheme Verbs are very rare in Turkish and generally contain exceptional forms which reflects in the low translation accuracy in Figures 3 and 4. We also observe that a good proportion of translated words with 'zero' accu-

[11] Details of the annotation protocol are in the Appendix

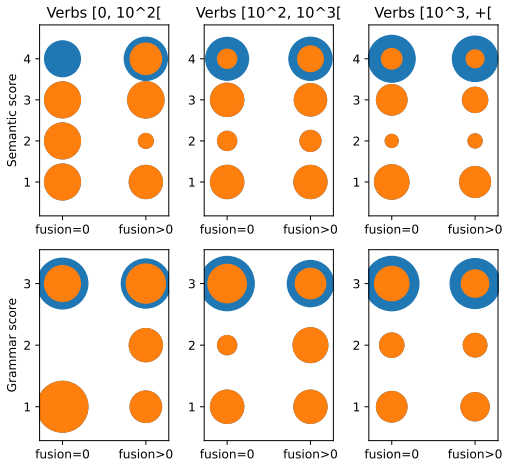

Figure 5: Semantic (top) and Grammar (bottom) annotation for Spanish.

racy (not the exact translation) has been annotated with highest semantic (same lemma) or grammar (same inflection) score, suggesting in some cases the model is successful in generalization, although we see this case when the words are relatively short (1 to 3 morphemes at most).

Fusion Figure 5 shows the semantic and grammar annotation scores for Spanish verbs. For the semantic scores (top), in all levels, the gap between the non-fusional and fusional verbs is reduced, for all the frequency groups. This means that the model is indeed able to generalise and offer alternative translations (not the exact verb), which is more complex to measure with automatic metrics. In the grammar scale (bottom), however, we still note a slight advantage in the maximum score (3) of the non-fusional verbs against the fusional ones for the two highest frequency subplots (middle and right).

5 Segment-level Analysis of Synthesis and Fusion in Machine Translation

To analyse the relationship between machine translation difficulty and the degree of synthesis or fusion at the segment level, we process a selection of systems for the language pair we want to evaluate. We obtain an automatic metric score of the output with respect to the reference (BLEU (Papineni et al., 2002), chrF (Popović, 2015), COMET (Rei et al., 2020)) per sentence, and also compute potential predictor variables for each sentence, such as the degree of synthesis or fusion. We complement the predictor variable list with other heuristics, such as the length of the sentence in characters (*char.count*) or words (*word.count*). The full list of all the predictors per language pair is in the Appendix. For simplification purposes, in the following analysis and plots, we only show the predictors that show a significant effect on the system outputs.

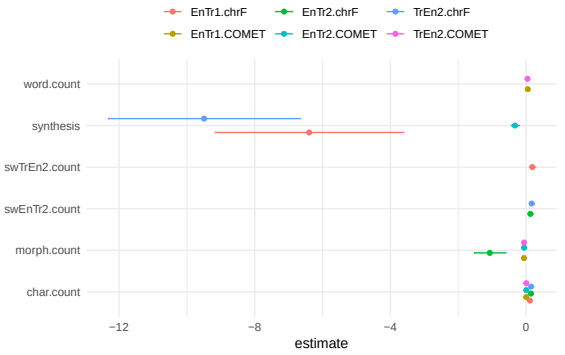

Figure 6: Overview of significant predictors for degree of synthesis across our TR-EN and EN-TR models.

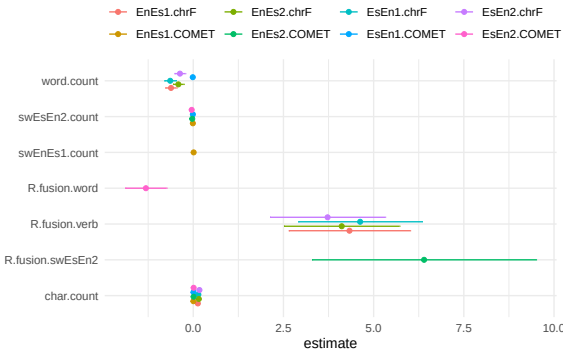

Figure 7: Overview of significant predictors for degree of fusion across our ES-EN and EN-ES models.

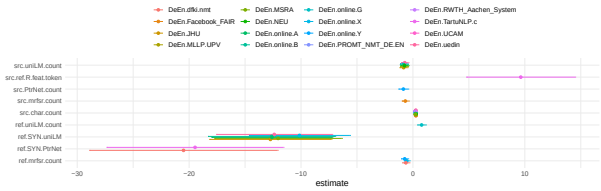

Figure 8: Overview of significant predictors across DE-EN models.

Synthesis on En-Tr and Tr-En We first start evaluating the English-Turkish and Turkish-English language pairs. The evaluated models are EnTr1, EnTr2, and TrEn2 (details in the Appendix).Also, as we are studying synthesis in Turkish, all predictors are computed on the Turkish side, regardless of the translation direction.

We generate a unique model per system output and evaluation metric (we use chrF and COMET),

in which each model's output is set to the degree of synthesis or other heuristics. The goal of this model creation is to identify which predictors (*i.e.*, the aforementioned variables) affect each method's performance. Following model creation, we extract the significant predictors of each model. This provides an indication of which variables can be used to predict the outcome of the model's dependent variable – in our case the degree of synthesis.

In this sense, Figure 6 presents an overview of the significant predictors on En-Tr and Tr-En systems, where we observe a large impact of the *synthesis* variable on the chrF scores of two different systems (EnTr1 and TrEn2). The only other heuristic that achieves a notable impact on system output is *morph.count*, or the length of Turkish sentence in morphemes, split by a morphological analyser. Other predictors have only a minor effect.

Fusion on En-Es and Es-En In a similar way, we evaluate the impact of fusion in English-Spanish (EnEs1, EnEs2) and Spanish-English (EsEn1, EsEn2) models (see Figure 7 and Appendix for details). Again, as we are studying fusion in Spanish, all predictors are computed on the Spanish side, regardless of the translation direction. Following the same procedure as before, Figure 7 presents an overview of the significant predictors, where we can observe that *R.fusion.verb*, or the ratio of the degree of fusion over the number of verbs, is the predictor that has the highest impact in most system outputs (EnEs1, EnEs2 and EsEn2). Additionally, *R.fusion.swEsEn2*, or the ratio of the degree of fusion over the number of subwords input in the EsEn2 model, also has a high impact in one system output (EnEs2, which uses the same segmentation model).

Analysis on En-De and De-En Finally, we extend the analysis to English-German and German-English language pairs, using the respective evaluation sets of the WMT2018 campaign (Bojar et al., 2018), and the system outputs provided for all the participants (measured in BLEU). For computing synthesis, we use the different segmentation methods we compared in §3.1. However, for fusion, we only use a shallow proxy with the number of morphological features that are tagged using a morphological analyser. In this case, the predictors are computed for both the source and target side. We present an overview of these significant predictors for German-English in Figure 8 (and we similarly discuss the English-German results from Figure 9 in the Appendix). We can observe that *ref.SYN.uniLM* and *ref.SYN.PtrNet* are the predictors that impact most of the different system outputs. These variables refer to the synthesis computed on the reference side (English) using uniLM or PtrNet as the morpheme segmentation method, respectively. Furthermore, we observe that *src.ref.R.feat.token* has also some effect over one system output,which is a shallow proxy for the fusion degree in the source w.r.t. to reference segment (using number of tagged morph. features).

6 Discussion and conclusion

Overall results do not suggest that translating into more analytic languages (e.g. Chinese) or more agglutinative ones (e.g. Turkish) is easier than their counterparts. Highly analytic ones poses the significant issue of word coverage and vocabulary size of the model. Besides, we cannot isolate the fusional degree from synthesis at all. For instance, Turkish is a highly agglutinative language, but also highly synthetic, and there are languages that presents both agglutinative and fusional traits, like Navajo. Therefore, a word level analysis with specific target POS, as in this study, should be fundamental to study the indexes. The language scope is another limitation: is it possible to extend it to further languages in a practical way? Synthesis can be calculated directly only if the morphological analyser splits the word into morphemes. Moreover, the fusion degree poses several issues as mentioned before. However, a less fine-grained analysis in the index (e.g. synthesis=1 vs. synthesis>1 or fusion=0 vs. fusion>0), as in this work, could be beneficial to evaluate more languages.

In conclusion, for the chosen study cases, we observed that higher degrees of synthesis and fusion have an impact in machine translation quality both at word and segment level. Also, we consider that performing our analysis for specific POS and languages could aid NLP systems, like in MT. For instance, as future work, we ask ourselves: how can we make an MT system more aware of fusional joints? And to evaluate the results, we need to fine-grain words with low and high fusion, to observe whether we are achieving improvements.

7 Ethical Considerations

The annotations in this paper were compensated accordingly (see Appendix). Also, for all the datasets

used in the research, we stick to the ethical standards giving credit to the original author in the spirit of *fair scientific usage*. We further strongly encourage future work that use these resources, to cite also the original sources of the data. We also see other ethical risks of this work: for the downstream task of MT, a translation system should not be deployed with low quality translations, as it can mislead the user, and have implicit biases.

References

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.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. 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.

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.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Ramy Eskander, Judith Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195, Florence, Italy. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

Zellig S Harris. 1951. Methods in structural linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, 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.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tahir Naveed, Imitaz Saeed Siddiqui, and Shaftab Ahmed. 2005. Parallel needleman-wunsch algorithm for grid. In *Proceedings of the PAK-US International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET 2005), Islamabad, Pakistan*.

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.

Thomas E Payne. 2017. Morphological typology. In *The Cambridge Handbook of Linguistic Typology*, pages 78–94. Cambridge University Press.

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

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1–10.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *International Conference on Natural Language Processing*, pages 417–427. Springer.

Edward Sapir. 1921. Types of linguistic structure.

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.

Petra Steiner. 2016. Refurbishing a morphological database for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1103–1108, Portorož, Slovenia. European Language Resources Association (ELRA).

Petra Steiner. 2017. Merging the trees - building a morphological treebank for German from two resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 146–160, Prague, Czech Republic.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for

Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Hongzhi Xu, Jordan Kodner, Mitchell Marcus, and Charles Yang. 2020. Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6672–6681, Online. Association for Computational Linguistics.

Tim Zingler. 2018. Reduction without fusion: Grammaticalization and wordhood in turkish. *Folia Linguistica*, 52(2):415–447.

A Human Evaluation

A.1 Annotation Protocol

This study measures the translation quality of translations generated by a translation system. You are given a list of sentences where one column lists each word in the gold standard (correct) translation and the corresponding column the system-generated translations. The evaluation of the translations will rely on the two scores described below. The scores to use in the evaluation are:

Semantic score evaluates the meaning of the word used in the automatic translation (system output) and how it compares with the gold standard translation.

Please assign each word in the output one of the scores you find most appropriate:

1. There is no relationship between the two lemmas
2. The lemmas are different but the translation does not fit well in the context
3. The lemmas are different but it is still an acceptable translation (e.g. synonym)
4. It is the same lemma

Grammar score evaluates the grammatical form and how it compares with the gold standard translation.

Please assign each word in the output one of the scores you find most appropriate:

1. The word is inflected in a different way and it is not necessarily correct
2. The word has different inflection but it is still grammatically correct
3. The words have the same inflection, and it is correct

Please annotate all words in the translations in the file shared with you. In your evaluation try assigning the two scores to each word independently. The inflection of the word measures the morphological feature and should also be evaluated independently from the analyzer output which is automated and may contain errors.

The file contains example annotations for your reference, please ask any questions related to unresolved annotation examples by contacting the project coordinators.

A.2 Annotators

For both Turkish and Spanish, the annotators were contacted directly due to their expertise in morphology (both of them are PhD students in Linguistics and Computational Linguistics, respectively), besides requiring that they are native speakers of the target languages. Also, they were paid more than the minimum wage per hour of annotation of their country of residence, and were told that the annotated data will be released upon acceptance of the study.

B Segment-level Analysis of Synthesis and Fusion

B.1 List of machine translation systems

- EnTr1: the same system used in §4.2
- EnTr2: Transformer-base model (Vaswani et al., 2017) with joint vocabulary size of 8k pieces (unigram language modelling from SentencePiece (Kudo and Richardson, 2018), and trained with a sample (10%) of the corpus of EnTr1.
- EnEs1: the same system used in §4.3
- EsEn1: similar configuration than EnEs1 but in the opposite direction
- EnEs2: same configuration as EnEs1 (model and vocabulary) but with smaller training data.

Predictor	Description
char.count	number of characters
word.count	number of words (no punct. or numbers)
morph.count	number of morphemes.
synthesis	ratio of morph.count / word.count
N+V.word.count	number of Nouns and Verbs
N+V.morph.count	number of morphemes of the Nouns and Verbs
N+V.synthesis	ratio of N+V.morph.count / word.count
swEnTr1.count	number of subwords processed by the EnTr1 model
swEnTr2.count	number of subwords processed by the EnTr2 model
swTrEn2.count	number of subwords processed by the TrEn2 model
syn.swEnTr1	ratio of swEnTr1.count / word.count (synthesis proxy)
syn.swEnTr2	ratio of swEnTr1.count / word.count (synthesis proxy)
syn.swTrEn2	ratio of swEnTr1.count / word.count (synthesis proxy)

Table 3: List of predictors for En-Tr and Tr-En. All variables are computed on the Turkish segment of the evaluation set.

Predictor	Description
char.count	number of characters
word.count	number of words (no punct. or numbers)
verb.count	number of verbs
fusion	sum of the degree of fusion of all the verbs in the segment
R.fusion.verb	ratio of fusion / verb.count
R.fusion.word	ratio of fusion / word.count
swEsEn1.count	number of subwords processed by the EsEn1 model
swEsEn2.count	number of subwords processed by the EsEn2 model
R.fusion.swEsEn1	ratio of fusion / swEsEn1.count
R.fusion.swEsEn2	ratio of fusion / swEsEn2.count
swEnEs1.count	number of subwords processed by the EnEs1 model
swEnEs2.count	number of subwords processed by the EnEs2 model
R.fusion.swEnEs1	ratio of fusion / swEnEs1.count
R.fusion.swEnEs2	ratio of fusion / swEnEs2.count

Table 4: List of predictors for En-Es and Es-En. All variables are computed on the Spanish segment of the evaluation set.

It uses only newscommentary-v8 data, with around 300k sentences).

- EsEn2: similar configuration than EnEs2 but in the opposite direction.

B.2 List of predictors

Tables 3, 4 and 5 describes all the predictors used at the segment level analysis of English-Turkish, English-Spanish and English-German (both directions), respectively.

B.3 Results on English-German

Figure 9 shows the analogous results for English to German, where the synthesis-based variables presents a high impact w.r.t. the other predictors.

Predictor	Description
src.char.count	number of characters in the source side
ref.char.count	number of characters in the target side
src.word.count	number of words in the source side
ref.word.count	number of words in the target side
src.uniLM.count	number of subwords obtained by uniLM in the source
ref.uniLM.count	number of subwords obtained by uniLM in the target
src.SYN.uniLM	synthesis in source = src.uniLM.count / src.word.count
ref.SYN.uniLM	synthesis in target = ref.uniLM.count / ref.word.count
src.mrfsr.count	number of subwords obtained by Morfessor in the source
ref.mrfsr.count	number of subwords obtained by Morfessor in the target
src.SYN.mrfsr	synthesis in source = src.mrfsr.count / src.word.count
ref.SYN.mrfsr	synthesis in target = ref.mrfsr.count / ref.word.count
src.PtrNet.count	number of subwords obtained by PtrNet in the source
ref.PtrNet.count	number of subwords obtained by PtrNet in the target
src.SYN.PtrNet	synthesis in source = src.PtrNet.count / src.word.count
ref.SYN.PtrNet	synthesis in target = ref.PtrNet.count / ref.word.count
src.feat.count	number of morph. features in the source (using spAcy)
src.R.feat.token	ratio of src.feat.count / src.word.count
ref.feat.count	number of morph. features in the target (using spAcy)
ref.R.feat.token	ratio of ref.feat.count / ref.word.count
src-ref.feat.count	src.feat.count minus ref.feat.count
src-ref.R.feat.token	src.R.feat.token minus ref.R.feat.token
ref-src.feat.count	ref.feat.count minus src.feat.count
ref-src.R.feat.token	ref.R.feat.token minus src.R.feat.token

Table 5: List of predictors for En-De and De-En. Variables are computed either on source (src) or target (ref) side.

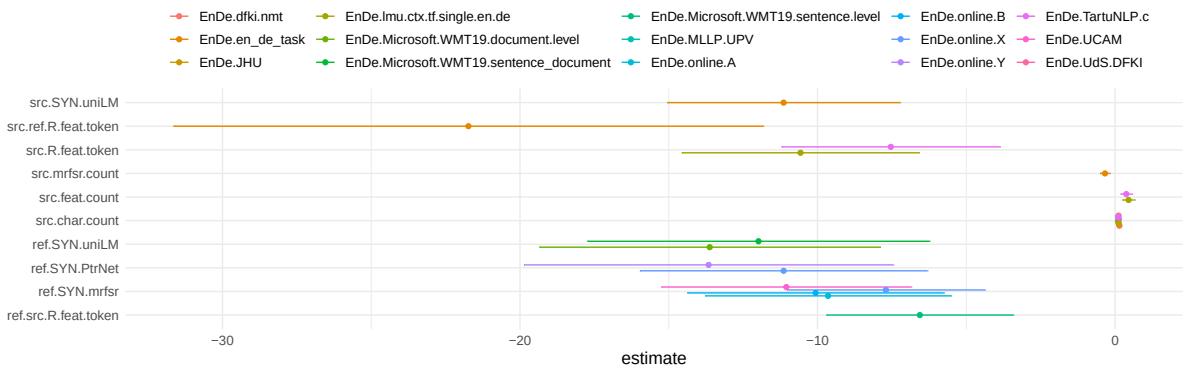

Figure 9: Overview of significant predictors for degree of synthesis across EN-DE models.