# The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, François Yvon

# The WMT18 Morpheval test suites for English–Czech, English–German, English–Finnish and Turkish–English

**Franck Burlot**
Lingua Custodia
1, Place Charles de Gaulle
78180 Montigny-le-Bretonneux
franck.burlot@linguacustodia.com

**Yves Scherrer**
Department of Digital Humanities
University of Helsinki
Helsinki, Finland
yves.scherrer@helsinki.fi

**Vinit Ravishankar**
Charles University
Faculty of Mathematics and Physics
ÚFAL; Prague, Czech Republic
vinit.ravishankar@gmail.com

**Ondřej Bojar**
Charles University
Faculty of Mathematics and Physics
ÚFAL; Prague, Czech Republic
bojar@ufal.mff.cuni.cz

**Stig-Arne Grönroos**
Aalto University
Department of Signal
Processing and Acoustics
Espoo, Finland
stig-arne.gronroos@aalto.fi

**Maarit Koponen**
School of Languages and
Translation Studies
University of Turku
Turku, Finland
maarit.koponen@utu.fi

**Tommi Nieminen**
Department of Digital Humanities
University of Helsinki
Helsinki, Finland
tommi.nieminen@helsinki.fi

**François Yvon**
LIMSI, CNRS, Université Paris Saclay
Campus Universitaire d'Orsay
F-91 403 Orsay Cédex
francois.yvon@limsi.fr

## Abstract

Progress in the quality of machine translation output calls for new automatic evaluation procedures and metrics. In this paper, we extend the Morpheval protocol introduced by Burlot and Yvon (2017) for the English-to-Czech and English-to-Latvian translation directions to three additional language pairs, and report its use to analyze the results of WMT 2018's participants for these language pairs. Considering additional, typologically varied source and target languages also enables us to draw some generalizations regarding this morphology-oriented evaluation procedure.

## 1 Introduction

The success of rather opaque neural machine translation systems has called for more fine-grained types of evaluation than traditional automatic evaluation metrics offer. In particular, we would like to obtain more detailed information about systems performance than just one overall number (even if it correlates well with human judgement). Evaluation metrics that focus on various aspects of the translation, such as syntax or morphology, rather than on general translation quality, have thus seen renewed interest. This interest has spurred the inclusion of additional test suites into the WMT 2018 news translation task.

Burlot and Yvon (2017, B&Y in the following) present a test suite for evaluating the morphological competence of machine translation systems. They provide a set of sentence pairs in the source language that differ by one morphological contrast. A sentence pair is considered correct if the morphological contrast is also conveyed in the target language translations of the two sentences of the pair. B&Y developed their test suite for English–Czech and English–Latvian and applied it to a selection of MT systems that participated in WMT 2017. For WMT 2018, we have



*Proceedings of the Third Conference on Machine Translation (WMT)*, Volume 2: Shared Task Papers, pages 550–564

https://doi.org/10.18653/v1/W18-64060

extended the English–Czech test suite[1] and created similar Morpheval test suites for three additional translation directions: English–German,[2] English–Finnish,[3] and Turkish–English.[4] All primary WMT submissions of these translation directions were evaluated.[5]

We start by summarizing the components of the Morpheval test suites and their language-specific implementations.

## 2 The Morpheval test suites

A Morpheval test suite according to B&Y consists of three aspects:

- the definition of a set of contrasts that can be triggered in the source language and evaluated in the target language;
- a procedure to generate contrast pairs from a monolingual source language corpus;
- and a procedure to score the target language translations of the contrast pairs.

B&Y describe three types of contrasts. Type A contrasts resemble paradigm completion tasks, in which one single morphological feature (number, gender, tense, etc.) is evaluated. The two sentences of a contrast pair only differ in one word (or phrase) and across one feature at a time. Type B contrasts contain somewhat more complicated substitutions that are mainly evaluated in terms of agreement. For example, a contrast pair contains a pronoun or an adjective-noun noun phrase, and its evaluation is correct if the adjective and noun agree. Type C contrasts concern lexical replacements of the same category, testing whether the morphological agreement still holds if an adjective is replaced by a hyponym. Table 1 summarizes the set of contrasts implemented for the different language pairs, according to this typology. The contrasts that are not described in B&Y will be presented in detail in the following sections.

Before that, we discuss some language-specific implementation differences of the generation and scoring procedures.

### 2.1 Sentence selection and contrast generation

We follow the algorithm provided by B&Y for sentence selection and contrast generation:

1. Collect a large number of short sentences (length $< 15$ words) containing a source feature of interest.

   As source corpora, we use the English News-2007 and 2008 corpora (for EN-CS and EN-DE), the English News-2007 corpus (for EN-FI), and SETIMES2 (for TR-EN). In order to detect the source features, the corpora are annotated using TreeTagger (Schmid, 1994) and/or CoreNLP (Manning et al., 2014) (for English), or an Apertium (Forcada et al., 2011) morphological analyser (for Turkish). For the named entities feature used in EN-FI, we additionally annotate the source corpora with the Stanford NER tagger (Finkel et al., 2005).

2. Generate a variant as prescribed by the contrast feature.

   For English corpora, we follow B&Y and use the Pymorphy morphological generator[6] to create the variants. For the Turkish corpus, we use Apertium.

3. Compute an average language model (LM) score for the base/variant pair, and remove the 33% worst pairs based on the LM score.

   We use a 5-gram language model trained on all English monolingual data available at WMT 2015. No language model filtering is applied to the Turkish data.

4. Randomly select 500 pairs per feature (400 for Turkish) for inclusion.

B&Y identify one of the sentences of a contrast pair as the "base" and the other one as the "variant". We keep this terminology for the sake of simplicity, but do not intend to imply (1) that the base is in any way "easier" to translate than the

[1] Contributors: Franck Burlot and François Yvon; test suite and evaluation scripts are available at https://github.com/franckbrl/morpheval_v2

[2] Contributors: Franck Burlot and François Yvon; test suite and evaluation scripts are available at https://github.com/franckbrl/morpheval_v2

[3] Contributors: Yves Scherrer, Maarit Koponen, Tommi Nieminen, Stig-Arne Grönroos; test suite, evaluation scripts and logs are available at https://github.com/Helsinki-NLP/en-fi-testsuite

[4] Contributors: Vinit Ravishankar and Ondřej Bojar

[5] The same method has also been adapted to English-to-French: significance tests, as well as concrete examples, are provided for this language pair in Burlot and Yvon (2018).

[6] http://pymorphy.readthedocs.io/

| Feature | B&Y | EN-CS | EN-DE | EN-FI | TR-EN |
|---|---|---|---|---|---|
| **Paradigm contrast features:** | | | | | |
| Singular vs. plural noun | A-1 | ✓ | ✓ | ✓ | |
| Singular vs. plural pronoun | A-2 | ✓ | ✓ | ✓ | |
| Masculine vs. feminine pronoun | A-3 | ✓ | ✓ | S | |
| Present vs. future tense | A-4 | ✓ | ✓ | S | ✓ |
| Present vs. past tense | A-5 | ✓ | ✓ | ✓ | ✓ |
| Indicative vs. conditional mode | | ✓ | ✓ | | |
| Positive vs. comparative adjective | A-6 | ✓ | ✓ | ✓ | |
| Positive vs. superlative adjective | | ✓ | ✓ | | |
| Affirmative vs. negative verb form | A-7 | ✓ | ✓ | ✓ | ✓ |
| Compound generation | | | ✓ | | |
| Human vs. non-human pronoun | | | | ✓ | |
| Definite vs. possessive determiner | | | | ✓ | |
| Definite vs. indefinite determiner | | | | S | |
| Reported speech subordinate clauses | | | | ✓ | |
| First vs. second person verb form | | | | | ✓ |
| Present vs. future subject participle | | | | | ✓ |
| Present vs. future object participle | | | | | ✓ |
| **Agreement features:** | | | | | |
| Pronoun vs. Adj+Nouns | B-1 | ✓ | ✓ | ✓ | |
| Pronoun vs. coordinated nouns | B-2 | ✓ | | | |
| Simple vs. coordinated verbs | B-3 | ✓ | ✓ | | |
| Adposition case (+ position) | B-4 | ✓ | ✓ | ✓ | |
| Coreference link | | ✓ | ✓ | | |
| Strong/weak adjective | | | ✓ | | |
| Local postposition/adverb case | | | | ✓ | |
| **Rare word features:** | | | | | |
| Named entities | | | | ✓ | |
| Numbers | | | | ✓ | |
| **Consistency features:** | | | | | |
| Adjective hyponyms | C-1 | ✓ | ✓ | | |
| Noun hyponyms | C-2 | ✓ | ✓ | | |
| Verb hyponyms | C-3 | ✓ | ✓ | | |

Table 1: List of contrast features implemented in the Morpheval test suites. The features already proposed by B&Y are marked by their corresponding code in the second column. *S* indicates features used to measure stability (see Section 2.5).

variant, (2) that the base always is the unmodified sentence extracted from the corpus and the variant the automatically modified one, or (3) that the evaluation of the base would be more lenient than the evaluation of the variant.

For consistency features (see Table 1), we select a noun, an adjective or a verb and replace it with a random hyponym, producing an arbitrary number of sentences. Sentence selection slightly differs from the description above: during step 2, we generate as many variants as possible. Each variant is then scored with a language model and only the top four variants are kept, leading to buckets of five sentences. For hyponym generation, we use WordNet (Miller, 1995).

### 2.2 Scoring procedures

The automatic scoring procedure for a given contrast pair receives two target language sentences (the MT output of the two source language sentences forming the contrast pair) as input and returns a binary correct/incorrect judgement. A contrast pair is judged correct if the two target sentences differ and the differences encode the contrast that is expressed in the source sentences. A contrast pair is judged incorrect if the two sentences are identical or if they differ in a way that is irrelevant to the examined contrast.

For consistency features, we wish to assess the MT system consistency with respect to lexical variation in a fixed context; accordingly, we measure the success based on the average normalized entropy of morphological features in the set of target sentences.

The target language sentences of all participating systems are morphologically analyzed to facilitate scoring. The following tools are used:

- Czech: MorphoDiTa (Straková et al., 2014)
- German: SMOR (Schmid et al., 2004)
- Finnish: The *finnish-analyze-words* script[7] provided by the Language Bank of Finland[8] and based on the Omorfi morphology (Pirinen, 2015) and the HFST toolkit (Lindén et al., 2011)
- English: MorphoDiTa (Straková et al., 2014)

[7] http://urn.fi/urn:nbn:fi:lb-2018041701

[8] https://www.kielipankki.fi/

As shown by B&Y, there is no need to perform a full morphological disambiguation in the target side, as we merely need to check whether some morphological features are present or absent. In fact, full automatic disambiguation could be harmful due to error propagation.

### 2.3 Additional English–Czech contrasts

The English–Czech evaluation procedure follows B&Y, to which we added a handful of new tests.

#### Conditional

Paradigm contrast features introduce a new verbal test. In the test suite, a verb in future tense is turned into its conditional form: *I will write* → *I would write*. In the Czech variant, we check whether the verb translation is in conditional mode.

#### Superlative

The superlative task is comparable to the comparative task introduced in B&Y. The base sentence contains an adjective and the variant contains its superlative form. In the output, we look for the adjective translation and check whether is has a superlative form.

#### Coreference

Agreement features introduce a new coreference task. The test suite for this task was produced using English coreference annotations obtained using CoreNLP (Manning et al., 2014). We collected sentences containing a coreference link involving a personal pronoun (*it*) or a relative pronoun (*that, which, who, whom, whose*). The base sentence remains unchanged. In order to generate the variant, the antecedent noun of the pronoun is then changed to a synonym using WordNet (Miller, 1995):

- Personal pronoun: *This **cat** is cute and I love **it**.* → *This **dog** is cute and I love **it**.*
- Relative pronoun: *The **woman who** left was angry.* → *The **man who** left was angry.*

In the output of the MT system, we are then able to locate the antecedent of the pronoun by looking for the only noun that differs between the base and variant translations (namely, the translation of *cat/woman* in the base and *dog/man* in the variant). Finally, we check whether the noun and personal

pronoun bear the same gender.[9] We also check number agreement for the relative pronoun. Note that for this specific task, we can compute accuracy scores on both base and variant.

### 2.4 Additional English–German contrasts

English–German is a new language pair we introduce in the current paper. It takes most of the previous tasks introduced in B&Y for English into Czech and Latvian. Conditional, superlative and coreference tasks are also adapted to German (see Section 2.3).

#### Compounds

This task consists in assessing the ability of the MT system to generate correct compounds that actually exist in German. For this purpose, the base sentence in the English test suite contains a multi-word expression that is *most likely* translated by a compound in German. To generate the variant, we modify one single English word in the multi-word expression, such that the new German translation should result in a compound that has at least one morpheme in common with the one seen in the base translation. For instance, the English expression *apple juice* in the base translates into the German compound *Apfelsaft*. We modify the word *apple* and obtain *orange juice*, which translates into *Orangensaft*. In the MT output we finally compare both compounds *Apfelsaft* and *Orangensaft* and report a success if they have at least one morpheme in common. Here, the common morpheme is *-saft*.

For the test suite generation, we needed a translation dictionary containing compounds on the German side and multi-word expressions on the English side. We gathered all the English-German parallel data we could find on OPUS (Tiedemann, 2012) and removed the data available at the WMT18 News Translation shared task. This resulted in nearly 40M parallel sentences. We obtained a phrase table out of this data using the Moses toolkit (Koehn et al., 2007). We finally extracted from this phrase table a dictionary containing a compound on the German side and several multi-word expressions on the English side (removing punctuation and other noisy tokens).

[9] We do not control whether the chosen synonym translates to a noun of a different gender. In some cases, the translation of the pronoun in the base and variant should thus remain unchanged, sharing the same gender.

The test suite generation starts with the identification in the base sentence of an English multi-word expression that is present in our dictionary. We then look for a new English multi-word expression that has at least one common word with the previous one (we have *apple juice*, we get the expression *orange juice*, since both have *juice* in common). Finally, if both expressions translate into German compounds that have at least one morpheme in common (relying on SMOR analysis), the new English expression is inserted into the sentence, which produces the variant sentence.

At evaluation step, we look for the word in the base sentence that is not in the variant sentence and vice-versa. We report a success when both words are known compounds and when they contain at least one common morpheme (using SMOR analysis).

#### Verb position

The test suite is generated by locating complex sentences where (a) the principal clause can be omitted and (b) the subordinate clause leads to a German translation where the verb should be located at the end of the clause. Using CoreNLP annotations, we focus on specific English conjunctions that lead to a verb shift in German, like *that* → *dass*, *because* → *weil*, etc. In order to generate the variant sentence, we simply omit all words from the beginning of the sentence up to the conjunction: *I think **that** life is hard.* → *Life is hard.*

Once both sentences are translated into German, we simply check that the conjugated verb is closer to the end of the sentence in the base than it is in the variant: *Ich denke, dass das Leben hart **ist**.* (last position) → *Das Leben **ist** hart.* (second to last).

#### Strong adjective

This task focuses on the contrast between weak and strong forms of the German adjective. We rely on a quite simple rule of German, stating that an adjective following a definite article does not contain any gender marker in its ending, whereas it does contain it when following, e.g. a possessive determiner.

We therefore identified English sentences with a subject noun phrase containing a definite article, an adjective and a noun (according to CoreNLP analysis). To generate the variant, we simply replace the article by a possessive determiner: *The small dog is gone.* → *Our small dog is gone.*

In the MT output, we check whether the variant contains a strong form of the adjective (using SMOR analysis): *Der kleine Hund ist weg.* → *Unser **kleiner** Hund ist weg.*

### 2.5 Additional English–Finnish contrasts

For English–Finnish, we reuse most of B&Y's paradigm contrast features, but repurpose some of them as stability features (see Table 1 and below). We reuse a limited subset of agreement features. After initial experiments, we decided against using consistency features, as they yielded a high percentage of unnatural and sometimes even unintelligible sentences. We provide additional features tailored to Finnish in both categories and provide an additional class of language-independent rare word features. In the following sections, we describe these features in more detail.

#### Human vs. non-human pronoun

Both English and Finnish distinguish between pronouns whose antecedents are human (English *I, he, she, ...*, Finnish *minä, hän, ...*) and pronouns whose antecedents are non-human (English *it*, Finnish *se*).

The conversion procedure identifies base sentences with instances of *me, us, him*, or *her*, and generates the variants by replacing the pronouns with *it*. We discard subject contexts and make sure that no other pronoun is present in the sentence. We also discard prepositional phrase contexts which would command the use of possessive suffixes in Finnish. Note that no treatment is applied to the antecedent of the pronoun. This is generally not an issue because we do not need to preserve the meaning between the base and variant sentence, we only need to check if human vs. non-human aspect of the pronoun is preserved.

The scoring procedure checks if the correct Finnish pronoun lemma (*se*) is used in the variant.

#### Definite vs. possessive determiner

In contrast to English, Finnish uses suffixation to indicate possession, e.g. *-ni* for the 1st person singular and *-si* for 2nd person singular as in *kirja+ni* 'my book', *kirja+si* 'your book'. We wanted to test how well current MT systems are able to generate these suffixes.

The conversion procedure selects variant sentences with noun phrases containing a possessive determiner and generates the base by replacing the possessive determiner with *the*.

The scoring script checks whether the possessive suffix (or alternatively, the possessive determiner) of the correct person is generated.

#### Reported speech subordinate clauses

In English, the structure of affirmative and interrogative subordinate clauses is rather similar: *X says that A* vs. *X asks if A*, without any structural differences in X or A. In Finnish, various types of expressions A are possible for *say+that*, but none of them is structurally identical to the *ask+if* subordinate clause, which corresponds to a (direct) question with the question particle *-ko/kö*.

The conversion procedure is bidirectional: it selects sentences containing *say+that* and transforms them to *ask+if* and vice-versa. Idiomatic constructions like *having said that* or *when asked if* are discarded.

The scoring procedure reports success if one of the correct constructions is identified in the affirmative sentence, and if the *-ko/kö*-construction is identified in the interrogative sentence.

#### Stability features

Two of the paradigm contrast features reported by B&Y do not apply to Finnish. Feature A-3 tests whether the masculine/feminine contrast between the pronouns *he* and *she* is conveyed in the target language, but Finnish uses the same pronoun *hän* regardless of the gender of the antecedent. Feature A-4 tests whether the present tense/future tense contrast is conveyed in the target language, but Finnish does not have a future tense and generally uses present tense in such cases.

Instead of measuring contrast, we can use these two features to measure *stability*: an MT system can be considered stable if two source sentences differing only in one word according to the contrasts presented above yield completely identical translations. Note that stability is not necessarily a good measure of overall translation quality: text can be translated in various ways, and two completely different translations can still be both correct, adequate and natural. However, stability may be an important criterion for particular applications of machine translation. For instance, for purposes of manual post-editing, stability may be preferable as it leads to easier predictability of the output. Our findings concerning the relation between stability and general translation quality will be discussed below.

We introduce a third stability feature that relies

on the absence of determiners in Finnish: we select sentences with noun phrases containing the indefinite determiner *a* and replace it with the definite determiner *the*. We try to avoid noun phrases in object positions, where determinacy can be expressed through case in Finnish.

The scoring procedure for stability features is simple: a contrast pair is considered stable if the strings of both translations are identical.

These stability features can be compared to the consistency features used for Czech and German. For both feature types, the variants are created through some type of transformation that is supposed to be invariant with respect to target morphology. For the consistency features, this transformation is semantic (based on the hyponymy relation), whereas it is morphological for the stability features.

**Adposition case**

B&Y introduce a feature where an English preposition is replaced by another one such that their counterparts in the target language govern two different cases. In Finnish, case government is closely tied to word order: most adpositions are postpositions and require genitive case, but some adpositions are prepositions and require partitive case. There are only two frequent prepositions, namely *ennen* 'before' and *ilman* 'without'. We restrict this feature to the former, as the latter often appears in idiomatic expressions from which variants are difficult to generate.

The sentence selection script produces the contrast pairs *before* → *after* and *before* → *during* (base followed by variant). Idiomatic constructions such as *named after, looking after, come before* are discarded, as well as particle readings of these words.

The scoring procedure verifies if a preposition with a noun or pronoun in partitive case to its right is present in the base, and if a postposition with a noun or pronoun in genitive case to its left is present in the variant. It also accepts the postpositional use of *ennen* in conjunction with pronouns (*sitä ennen*), as well as the use of bare (ad-/in-)essive case instead of the postposition *aikana* 'during'.

**Local postposition case**

Finnish local postpositions (the equivalents of *over, under, next to, between*, etc.) can be inflected themselves using the Finnish local cases, e.g. *sisällä/sisältä/sisälle* 'inside/from inside/towards inside', *edessä/edestä/eteen* 'in front of/from in front of/towards in front of'.

The conversion procedure yields the following contrast pairs: *in front of* → *behind*, *underneath* → *next to*, *outside* → *inside*, *inside* → *outside*, *above* → *below*, *below* → *above*. Non-prepositional and idiomatic readings are discarded as far as they could be discovered during development.

The scoring procedure checks that the English prepositions are translated correctly and that the case type (locative/separative/lative, as in the examples above) matches between the two sentences of the contrast pair.

**Rare word features**

In the early days of NMT, translation of out-of-vocabulary words was virtually impossible and hampered the performance when compared with SMT. In recent years however, most systems have adopted an approach in which rare words are split into "subwords" during preprocessing (see e.g. Sennrich et al., 2016), such that any unknown word can be composed of various subword chunks during test time. Several subword chunking algorithms with various parameter settings can be used, but their respective performance differences are hard to assess as they typically concern low-frequency words with low impact on general translation quality. Therefore, we introduce two features that specifically deal with low-frequency items. These features are language-independent and do not require the use of a morphological analyzer.

For the first feature, we identify large numbers (at least 3 digits) in the English source text and modify them by subtracting a constant number. For example, the number *27,801* would be transformed into *27,628*. The scoring procedure verifies if the original and modified numbers are found in the respective sentences.

For the second feature, we use the Stanford Named entity recognizer to identify named entities in the English source text. We then consider two subsets of named entities, frequent ones (occurring more than 1000 times) and rare ones (occurring between 20 and 100 times). Contrast pairs are generated by identifying sentences with a frequent named entity, and replacing it by a rare one. We restrict the replacement to single-word named entities of the same class and make sure that the replacement candidate contains at least two differing

| System | BLEU | Ave. z |
|---|---|---|
| CUNI-Transformer | 26.6 | 0.594 |
| uedin | 24.0 | 0.384 |
| online-B | — | 0.101 |
| online-A | — | -0.115 |
| online-G | — | -0.246 |

Table 2: BLEU scores and human evaluation scores computed on newstest-2018 for English–Czech.

| System | BLEU | Ave. z |
|---|---|---|
| online-Z | — | 0.653 |
| online-B | — | 0.561 |
| Microsoft-Marian | 48.9 | 0.551 |
| MMT-production-system | 46.7 | 0.539 |
| UCAM | 47.1 | 0.537 |
| NTT | 47.0 | 0.491 |
| KIT | 46.9 | 0.454 |
| online-Y | — | 0.396 |
| JHU | 43.9 | 0.377 |
| uedin | 44.9 | 0.352 |
| LMU-nmt | 40.6 | 0.213 |
| online-A | — | 0.060 |
| online-F | — | -0.385 |
| online-G | — | -0.416 |
| RWTH-UNSUPER | 15.9 | -0.966 |
| LMU-unsup | 15.8 | -1.122 |

Table 3: BLEU scores and human evaluation scores computed on newstest-2018 for English–German.

characters, as in the following example: *Extensive damage was reported in Cuba.* → *Extensive damage was reported in Tuzla.*

The scoring procedure checks that both named entity strings are found in the respective sentences. The frequent named entities are likely to be translated (e.g., English *Africa* would become Finnish *Afrikka*, in oblique cases *Afrika-*). Therefore, we add a small hand-crafted dictionary containing the most frequent entities, and compare these entries with the base forms obtained by the morphological analyzer. We currently do not verify case consistency, as many rare entities are not recognized by the morphological analyzer.

### 2.6 Additional Turkish–English contrasts

We introduce Turkish–English as another new language pair in the paper. Note that the translation direction is opposite to the other pairs, with English acting as the target language. We include the B&Y tests for verb tense and polarity and add several tests for Turkish-specific features.

| System | BLEU | Ave. z |
|---|---|---|
| NICT | 18.2 | 0.521 |
| HY-NMT | 17.8 | 0.466 |
| uedin | 16.7 | 0.324 |
| Aalto | 16.2 | 0.271 |
| HY-NMT2step | 14.5 | 0.258 |
| talp-upc | 14.3 | 0.238 |
| CUNI-Kocmi | 14.7 | 0.184 |
| online-B | — | 0.183 |
| online-A | — | -0.212 |
| online-G | — | -0.233 |
| HY-SMT | 10.5 | -0.334 |
| HY-AH | 6.4 | -0.369 |

Table 4: BLEU scores and human evaluation scores computed on newstest-2018 for English–Finnish.

| | SETIMES2 | newstest-2018 | |
|---|---|---|---|
| System | BLEU | BLEU | Ave.z |
| online-G | 25.86 | — | 0.101 |
| online-A | 27.03 | — | 0.077 |
| Alibaba-Ensemble | — | — | 0.030 |
| online-B | 24.84 | — | 0.027 |
| uedin | 48.42 | 26.9 | -0.008 |
| NICT | 40.64 | 26.7 | -0.040 |

Table 5: BLEU scores computed on SETIMES2 and newstest-2018 and human evaluation scores on newstest-2018 for Turkish–English.

**Verb person**

Turkish models verbal agreement with number and person agglutinatively, often making pronouns superfluous. We modify first person verbal agreement to second person, keeping the number intact: *kitap okuyorum* → *kitap okuyorsun*. We check the MT output for the presence of the pronoun *you*: *I am reading a book* → ***you*** *are reading a book*.

**Participles**

Turkish features several participles that form relative clauses. These include, relevant to our tests, present-tense subject and object participles, and future tense participles. We introduce two tests. One transforms present tense subject participles to future tense ones: *Bu gelen adam* → *Bu gelecek adam*. For the English translations, our (fairly simple) test involves searching through the translation output for the tense-imparting strings, *will, shall, would* and *going* (as a simple test for the presence of *'going to'*): *The man who is coming* → *The man who* ***will*** *come*.

Object participles function similarly, however, they use transitive verbs that take an argument: *Okuduğum kitap* → *okuyacağım kitap*. Our tests for the MT output are similar: *the book that I read*

| | Verbs | | | | Pronouns | | Nouns | Adjectives | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| System | Past | Future | Cond. | Neg. | Fem. | Plur. | Plur. | Compar. | Superl. | |
| CUNI-Transformer | 84.2 | **88.0** | 59.0 | **97.4** | 94.2 | 92.2 | 76.4 | 74.0 | **89.8** | 83.9 |
| uedin | **92.0** | 83.0 | **73.4** | 96.6 | 94.2 | **92.8** | 78.8 | **78.8** | 88.8 | **86.5** |
| online-B | 87.8 | 77.6 | 57.4 | 94.2 | 92.8 | 92.0 | 80.0 | 75.8 | 69.8 | 80.8 |
| online-A | 86.8 | 86.8 | 71.2 | 94.4 | 94.0 | 89.6 | **81.2** | 74.6 | 61.0 | 82.2 |
| online-G | 81.4 | 84.0 | 70.8 | 78.4 | **98.0** | 89.4 | 79.2 | 73.0 | 50.4 | 78.3 |

Table 6: Accuracy values for the English-Czech test suite (paradigm contrast features).

| | Coordinated verbs | | | Co. N. | Adj+Nouns | | | Coref. rel. | | Prep. | Cor. per. | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System | Nbr | Pers | Tense | Case | Gdr | Nbr | Case | Gdr | Nbr | Case | Gdr | |
| CUNI-Transformer | **88.0** | **88.4** | **84.8** | 99.6 | **96.4** | **97.0** | **97.0** | 76.5 | 77.3 | 95.4 | **66.4** | **87.9** |
| uedin | 84.8 | 84.8 | 81.8 | 99.8 | 94.6 | 94.8 | 94.8 | 81.0 | 82.2 | **96.9** | 64.3 | 87.3 |
| online-B | 82.2 | 83.2 | 80.0 | 99.8 | 91.2 | 91.8 | 91.6 | 79.5 | 80.1 | 91.6 | 64.8 | 85.1 |
| online-A | 68.8 | 67.6 | 65.0 | 99.2 | 91.0 | 89.4 | 91.2 | **81.9** | **82.9** | 96.1 | 55.0 | 80.7 |
| online-G | 62.6 | 61.2 | 58.8 | **100.0** | 83.2 | 80.4 | 82.6 | 76.7 | 77.5 | 84.6 | 42.0 | 73.6 |

Table 7: Accuracy values for the English-Czech test suite (agreement features).

| | Nouns | Adjectives | | | Verbs | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| System | Case | Gender | Number | Case | Number | Person | Tense | Negation | |
| CUNI-Transformer | 0.109 | 0.191 | 0.193 | 0.203 | 0.110 | 0.077 | **0.096** | 0.069 | 0.131 |
| uedin | **0.095** | **0.185** | **0.184** | **0.189** | **0.099** | 0.081 | 0.097 | 0.072 | **0.125** |
| online-B | 0.105 | 0.186 | 0.186 | 0.195 | 0.108 | **0.071** | 0.099 | **0.067** | 0.127 |
| online-A | 0.096 | 0.202 | 0.201 | 0.207 | 0.182 | 0.129 | 0.154 | 0.105 | 0.159 |
| online-G | 0.153 | 0.229 | 0.229 | 0.237 | 0.242 | 0.161 | 0.190 | 0.119 | 0.195 |

Table 8: Entropy values for the English-Czech test suite (consistency features).

→ *the book that I **will** read.*

## 3 Results

In Tables 2–5, we summarize the WMT18 submissions of the four language directions in terms of BLEU scores[10] and human evaluation scores on the official test set (Bojar et al., 2018).[11]

In the following, we present the results of all our tests across languages in an as uniform way as possible. Bolding in the tables means simply the best result in that category. We do not use any significance tests here. All tables are sorted according to the human evaluation scores.

### 3.1 English–Czech

Results for the paradigm contrast features in English–Czech are shown in Table 6. Not taking into account *online* systems whose architectures are unknown, the table shows a contrast between a Recurrent Neural Network model (*uedin*) and a Transformer model (*CUNI-Transformer*). The former obtains slightly higher accuracies than the latter. This is especially obvious in verb tasks (past and conditional), as well as for noun number. This might suggest that Transformer models have more difficulty in conveying a morphological feature from source to target.[12]

However, we observe no such difference for agreement features (Table 7), where *uedin* obtains an average accuracy of 87.3 and *CUNI-Transformer* obtains 87.9. The latter is slightly better for coordinated verbs and noun phrase inner agreement (see the Adj+Nouns columns), but the former is significantly better in terms of coreference with a relative pronoun (Coref. rel.).

Both systems obtain similar average entropy values in Table 8. These results can be compared to the ones shown in Table 7 of B&Y, although they were computed on another version of the test suite containing different sentences. Whereas the

[10] http://matrix.statmt.org

[11] With the exception of Turkish–English, we are not able to compute BLEU scores on the test suite data, as no reference translations are available.

[12] It is however important to note that *not* preserving past of conditional form of the verb needs not lead to a lower translation quality in general because in many situations, less precise wording does not really affect the overall meaning. The reader may subconsciously correct smaller discrepancies among sentences while enjoying the more fluent or more common wording.

| | Verbs | | | | Pronouns | Nouns | | Adjectives | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| System | Past | Future | Cond. | Neg. | Plur. | Compd. | Nbr. | Compar. | Superl. | |
| online-Z | 85.0 | 42.2 | 79.2 | 95.8 | 97.7 | 63.1 | 58.2 | 87.6 | 93.9 | 78.1 |
| online-B | 91.3 | 86.8 | 92.3 | 98.4 | 99.2 | 63.7 | 67.3 | 92.8 | 98.7 | 87.8 |
| Microsoft-Marian | 90.4 | 71.3 | 97.6 | **99.4** | 98.6 | 63.6 | 65.2 | 94.9 | 99.6 | 86.7 |
| MMT-production-system | 92.3 | 79.7 | 86.4 | 98.4 | 97.3 | **67.2** | 63.1 | 93.1 | 98.9 | 86.3 |
| UCAM | 94.7 | 84.6 | 98.0 | 99.2 | 99.0 | 64.0 | 68.0 | **97.5** | **100.0** | **89.5** |
| NTT | 93.9 | 89.2 | 97.2 | 99.2 | 99.6 | 61.2 | 68.5 | 96.5 | **100.0** | **89.5** |
| KIT | 89.6 | 74.4 | 96.6 | 98.8 | 98.8 | 61.9 | 64.9 | 93.4 | 99.6 | 86.4 |
| online-Y | 91.5 | 81.4 | 91.3 | 98.8 | 98.8 | 66.1 | 67.3 | 94.0 | 99.1 | 87.6 |
| JHU | 92.6 | **90.4** | 94.6 | 97.4 | 99.6 | **67.2** | 69.0 | 93.6 | 98.9 | 89.3 |
| uedin | 93.1 | 79.4 | 97.4 | **99.4** | 97.2 | 66.4 | 65.4 | 94.4 | 99.1 | 88.0 |
| LMU-nmt | 93.6 | 80.2 | 98.4 | **99.4** | 97.3 | 66.8 | 70.9 | 94.9 | 99.8 | 89.0 |
| online-A | 93.5 | 87.5 | 95.4 | 99.2 | 99.4 | 62.6 | **71.9** | 95.5 | 99.1 | 89.3 |
| online-F | **98.7** | 2.1 | 98.4 | **99.4** | **100.0** | 63.4 | 70.5 | 95.1 | 99.3 | 80.8 |
| online-G | 90.2 | 52.8 | 92.8 | 98.8 | 95.8 | 54.2 | 63.1 | 90.7 | 97.5 | 81.8 |
| RWTH-UNSUPER | 92.3 | 52.3 | **99.2** | 98.6 | 95.7 | 18.9 | 71.1 | 88.3 | 98.1 | 79.4 |
| LMU-unsup | 74.7 | 45.4 | 97.2 | 88.6 | 93.8 | 58.0 | 67.2 | 84.7 | 99.7 | 78.8 |

Table 9: Accuracy values for the English-German test suite (paradigm contrast features).

| | Coordinated verbs | | | Verb | Adj+Nouns | | Coref. rel. | | Cor. per. | Adj. | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| System | Nbr | Pers | Tense | Pos | Gdr | Nbr | Gdr | Nbr | Gdr | Strong | |
| online-Z | 80.5 | 80.7 | 80.3 | 90.8 | 99.8 | 99.8 | 65.5 | 65.5 | 93.2 | 79.8 | 83.6 |
| online-B | 98.7 | 98.7 | 98.7 | 96.0 | **100.0** | **100.0** | 69.6 | 69.6 | 88.7 | 95.7 | 91.6 |
| Microsoft-Marian | 96.3 | 96.3 | 96.3 | 93.8 | **100.0** | **100.0** | 70.1 | 70.1 | 93.8 | 96.7 | 91.3 |
| MMT-production-system | 92.3 | 92.6 | 92.3 | 94.8 | 99.3 | 99.6 | 68.7 | 68.7 | 94.4 | 94.1 | 89.7 |
| UCAM | 97.4 | 97.4 | 97.4 | 93.8 | **100.0** | **100.0** | 68.4 | 68.4 | 94.8 | 99.1 | 91.6 |
| NTT | 98.9 | 98.9 | 98.4 | 94.8 | **100.0** | **100.0** | 70.0 | 70.0 | 93.3 | 96.8 | 92.1 |
| KIT | 97.0 | 97.0 | 96.7 | 93.4 | **100.0** | **100.0** | 69.6 | 69.6 | 93.4 | 96.1 | 91.3 |
| online-Y | 97.1 | 97.1 | 97.1 | **97.2** | **100.0** | **100.0** | 70.8 | 70.8 | 93.0 | 97.9 | 92.1 |
| JHU | 99.6 | 99.8 | **99.8** | 94.2 | **100.0** | **100.0** | 70.1 | 70.1 | 91.6 | 97.4 | **92.2** |
| uedin | 99.1 | 99.1 | 99.1 | 94.0 | 99.5 | **100.0** | 68.5 | 68.5 | 94.1 | 99.6 | **92.2** |
| LMU-nmt | 98.5 | 99.0 | 99.0 | 96.2 | **100.0** | **100.0** | **72.2** | **72.2** | 84.7 | 96.9 | 91.9 |
| online-A | 99.7 | **100.0** | 99.0 | 88.5 | **100.0** | **100.0** | 70.1 | 70.1 | 85.3 | 98.2 | 91.1 |
| online-F | **99.8** | **100.0** | 99.3 | 92.6 | **100.0** | **100.0** | 68.6 | 68.6 | 90.8 | **100.0** | 92.0 |
| online-G | 99.0 | **100.0** | 99.5 | 56.6 | **100.0** | **100.0** | 65.2 | 65.2 | 78.6 | 92.5 | 85.6 |
| RWTH-UNSUPER | 98.7 | 99.7 | 98.4 | 95.4 | 98.0 | **100.0** | 65.7 | 65.7 | 72.6 | 98.9 | 89.3 |
| LMU-unsup | 97.1 | **100.0** | 96.5 | 88.3 | 99.4 | 99.7 | 45.6 | 45.6 | **95.1** | 99.3 | 86.7 |

Table 10: Accuracy values for the English-German test suite (agreement features).

| | Nouns | Adjectives | | Verbs | | | Average |
|---|---|---|---|---|---|---|---|
| System | Case | Gender | Number | Number | Person | Tense | |
| online-Z | 0.038 | 0.034 | 0.030 | 0.069 | 0.055 | 0.091 | 0.053 |
| online-B | 0.015 | 0.010 | 0.008 | 0.025 | 0.013 | 0.055 | 0.021 |
| Microsoft-Marian | 0.014 | 0.006 | 0.004 | 0.022 | 0.015 | 0.051 | 0.019 |
| MMT-production-system | 0.015 | 0.017 | 0.015 | 0.031 | 0.020 | 0.071 | 0.028 |
| UCAM | 0.015 | 0.007 | 0.005 | 0.014 | 0.006 | 0.048 | 0.016 |
| NTT | 0.015 | **0.001** | **0.000** | 0.019 | 0.012 | 0.049 | 0.016 |
| KIT | 0.017 | 0.009 | 0.008 | 0.024 | 0.017 | 0.059 | 0.022 |
| online-Y | 0.019 | 0.013 | 0.011 | 0.033 | 0.019 | 0.073 | 0.028 |
| JHU | 0.009 | 0.007 | 0.006 | 0.027 | 0.013 | 0.063 | 0.021 |
| uedin | 0.011 | 0.005 | 0.003 | 0.024 | 0.017 | 0.051 | 0.019 |
| LMU-nmt | 0.020 | 0.003 | 0.003 | 0.023 | 0.008 | 0.067 | 0.021 |
| online-A | 0.015 | 0.003 | 0.001 | 0.037 | 0.011 | 0.070 | 0.023 |
| online-F | **0.005** | 0.002 | 0.001 | **0.011** | **0.004** | **0.034** | **0.010** |
| online-G | 0.030 | 0.006 | 0.001 | 0.068 | 0.014 | 0.087 | 0.034 |
| RWTH-UNSUPER | 0.031 | 0.015 | 0.010 | 0.060 | 0.009 | 0.115 | 0.040 |
| LMU-unsup | 0.019 | 0.015 | **0.000** | 0.098 | 0.015 | 0.137 | 0.047 |

Table 11: Entropy values for the English-German test suite (consistency features).

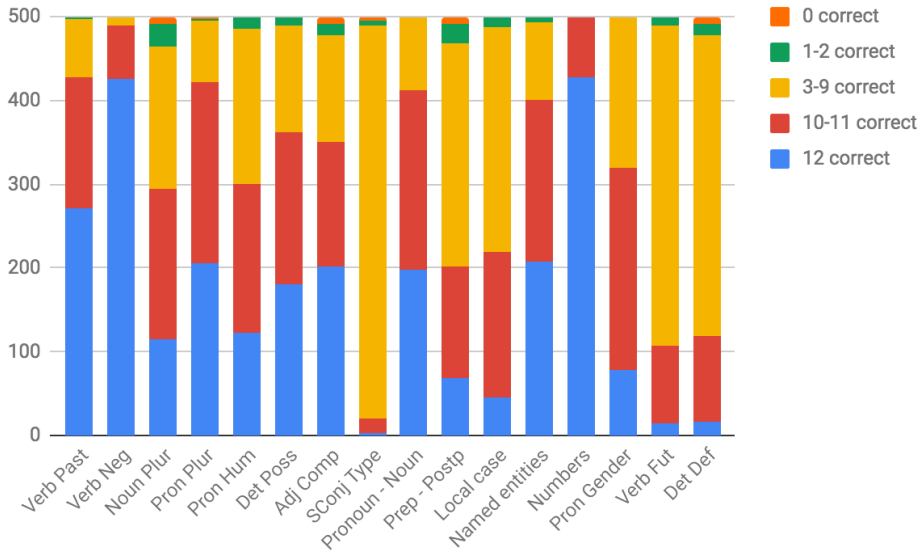


Figure 1: Distribution of correct labels across examples for English–Finnish. *n correct* represents the number of examples (out of the total 500 per contrast) for which $n$ systems (out of a total of 12) were able to generate the contrast correctly.

best system listed there (LIMSI FNMT) obtained an average entropy of 0.168, the WMT 2018 systems *uedin* and *CUNI-Transformer* turn out to be significantly lower (0.125 and 0.131, respectively).

### 3.2 English–German

Results for the paradigm contrast features in English–German are shown in Table 9. It is clear from the table that certain tasks are now too easy for the current state-of-the-art: verb negation, pronoun plural and superlative are very close to a perfect accuracy across nearly all systems. The hardest task seems to be the one involving compound generation (Nouns Compd. in Table 9), where accuracies range from 18.9 to 66.4. Verb future tense also causes considerable difficulties to several systems, including the top-scoring online-Z. As with English–Czech, we see that the systems best ranked according to manual evaluation (closer to the top of the list) do not necessarily score well in this detailed evaluation and vice versa. One example is the anonymous *Online-Z* system, which is rather bad at preserving verb attributes, noun number or comparative adjectives.

Table 10 shows even more clearly how easy certain tasks are. Indeed, noun phrase internal agreement (gender and number) seems to be perfectly modeled by every system (accuracies range from 98.0 to 100, see the columns Adj+Nouns). Coordinated verbs and strong/weak adjectives seem rather easy as well, with all accuracies over 90%. Coreference with relative pronouns (Coref. rel.) seems to be the most difficult task. Note that we observe exactly the same results for gender and number: this is due to the fact that the SMOR analysis of relative pronouns is highly ambiguous. E.g. the pronoun *die* is both singular and plural, and has no specific gender in plural form, therefore it may agree with any noun. Strictly all the errors for this task are due to the fact that we could not find the right noun or pronoun in the sentence, which leads to no difference between gender and number. Hence the task does not measure agreement as much as the ability of a system to output a relative pronoun.

Consistency tasks are shown in Table 11. Strikingly, the *online-Z* system, ranked best on human judgement, shows the worst entropy score. Overall, the consistency task figures do not seem to correlate well with general translation quality measures. Compared to to the Czech values in Table 8, we notice that the German average entropy values are quite low. This could be explained by the fact that Czech has a richer nominal, adjectival and verbal morphology than German. For instance, whereas German has four cases, Czech has seven, which impacts the entropy values computed for this task.

| System | Verbs Past | Verbs Neg | Nouns Plur | Pronouns Plur | Pronouns Hum | Det Poss | Adj Compar | SConj Type | Average |
|---|---|---|---|---|---|---|---|---|---|
| NICT | 94.4 | 98.6 | 79.2 | 94.6 | 90.4 | 88.4 | 88.0 | **96.2** | **91.2** |
| HY-NMT | 93.8 | **99.0** | 74.8 | 82.6 | 67.4 | 83.6 | 78.6 | 96.0 | 84.5 |
| uedin | 94.0 | 98.8 | 75.0 | 93.6 | 82.6 | 85.0 | 87.8 | 90.2 | 88.4 |
| Aalto | 93.4 | 98.8 | 72.0 | 88.4 | 77.4 | 90.8 | 81.6 | 87.6 | 86.3 |
| HY-NMT2step | **95.2** | 99.0 | 69.8 | 91.4 | 83.8 | **94.0** | 81.6 | 87.4 | 87.8 |
| talp-upc | 91.0 | 98.4 | 72.2 | 94.0 | 80.2 | 83.0 | 79.8 | 84.4 | 85.4 |
| CUNI-Kocmi | 89.0 | 98.0 | 73.8 | 91.4 | 80.4 | 86.2 | 78.6 | 76.6 | 84.3 |
| online-B | 92.0 | 98.6 | 76.4 | 91.0 | 78.0 | 77.0 | 82.2 | 84.8 | 85.0 |
| online-A | 87.6 | **99.0** | 78.6 | 94.2 | 82.0 | 84.6 | 86.0 | 23.4 | 79.4 |
| online-G | 82.8 | 92.6 | 76.8 | 86.8 | 66.2 | 83.0 | **88.2** | 3.2 | 72.5 |
| HY-SMT | 79.0 | 96.2 | 53.2 | 62.8 | 59.4 | 80.6 | 68.8 | 6.8 | 63.4 |
| HY-AH | 93.4 | 98.2 | **88.8** | **99.0** | **94.8** | 76.0 | 87.2 | 1.0 | 79.8 |

Table 12: Accuracy values for the English–Finnish test suite (paradigm completion features).

| System | Adj+ Noun | Prep / Postp | Local case | Average |
|---|---|---|---|---|
| NICT | **96.2** | **88.2** | 81.2 | **88.5** |
| HY-NMT | 87.8 | 81.8 | 68.6 | 79.4 |
| uedin | 92.0 | 83.0 | 80.4 | 85.1 |
| Aalto | 93.2 | 81.4 | 69.8 | 81.5 |
| HY-NMT2step | 90.0 | 86.8 | 70.4 | 82.4 |
| talp-upc | 91.8 | 70.4 | 77.4 | 79.9 |
| CUNI-Kocmi | 91.4 | 63.8 | 71.2 | 75.5 |
| online-B | 90.2 | 72.8 | 66.0 | 76.3 |
| online-A | 81.2 | 41.4 | 78.4 | 67.0 |
| online-G | 84.6 | 33.8 | 80.2 | 66.2 |
| HY-SMT | 78.6 | 48.0 | 41.4 | 56.0 |
| HY-AH | 89.8 | 74.0 | **81.8** | 81.9 |

| System | Named entities | Numbers | Average |
|---|---|---|---|
| NICT | 90.4 | 99.4 | 94.9 |
| HY-NMT | 91.6 | 98.4 | 95.0 |
| uedin | 92.4 | 99.8 | 96.1 |
| Aalto | 82.4 | 96.0 | 89.2 |
| HY-NMT2step | 81.8 | 97.0 | 89.4 |
| talp-upc | 79.8 | 98.8 | 89.3 |
| CUNI-Kocmi | 86.6 | 99.8 | 93.2 |
| online-B | **94.8** | 99.0 | **96.9** |
| online-A | 90.2 | 99.8 | 95.0 |
| online-G | 86.2 | **100.0** | 93.1 |
| HY-SMT | 81.6 | 93.8 | 87.7 |
| HY-AH | 85.0 | 99.8 | 92.4 |

Table 13: Accuracy values for the English–Finnish test suite (left: agreement features, right: rare word features).

| System | Verbs Fut | Pronouns Gender | Det Def | Average |
|---|---|---|---|---|
| NICT | 68.4 | 87.0 | 70.6 | 75.3 |
| HY-NMT | 65.0 | 84.2 | 58.8 | 69.3 |
| uedin | **73.0** | 84.6 | 65.4 | 74.3 |
| Aalto | 71.2 | 74.8 | 63.6 | 69.9 |
| HY-NMT2step | 64.4 | 75.4 | 57.2 | 65.7 |
| talp-upc | 61.0 | 75.0 | 53.2 | 63.1 |
| CUNI-Kocmi | 54.0 | 65.6 | 48.8 | 56.1 |
| online-B | 68.8 | 88.6 | 55.2 | 70.9 |
| online-A | 59.6 | 84.8 | 70.2 | 71.5 |
| online-G | 62.2 | 91.0 | 73.6 | 75.6 |
| HY-SMT | 33.8 | 79.6 | 42.2 | 51.9 |
| HY-AH | 71.4 | **95.0** | **89.0** | **85.1** |

Table 14: Accuracy values for the English–Finnish test suite (stability features).

| | **Verbs** | | | | **Obj. Part.** | **Subj. Part.** | **Average** |
|---|---|---|---|---|---|---|---|
| System | Person | Future | Past | Neg. | Future | Future | |
| online-G | 60.0 | 67.3 | 75.5 | 68.3 | 41.0 | 21.8 | 55.65 |
| online-A | **71.3** | **72.3** | **77.3** | **72.0** | **49.5** | **30.5** | **62.15** |
| online-B | 46.8 | 66.8 | 76.3 | 66.5 | 40.3 | 26.8 | 53.92 |
| uedin | 53.5 | 65.0 | 66.5 | 64.5 | 39.0 | 17.0 | 50.92 |
| NICT | 57.8 | 69.0 | 73.3 | 67.8 | 45.5 | 22.3 | 55.95 |

Table 15: Accuracy values for the Turkish–English test suite.

### 3.3 English–Finnish

As a general overview of the English–Finnish features and their difficulty, Figure 1 shows the distribution of correct labels across examples and features. It can be seen that some features (e.g., verb negation or numbers) pose very few problems to current MT systems, whereas others (e.g. subordinate clause type, see SConj Type in the figure) are much more difficult. In contrast to German, the pronoun plural feature seems to be harder for Finnish systems. In particular, the *0 Correct* and *1-2 Correct* categories may indicate potential problems in the example generation or scoring process.

We performed a manual analysis of a small sample of contrast pairs (20-30 examples per feature) regarding the grammaticality of the automatically generated sentences and the recall of the automatic evaluation script. For the features *Noun Plur, Pron Hum, Det Poss, Adj Compar Adj* and *Local case*, more than 20% of the annotated examples showed either problems in the source sentence (incomplete sentences due to splitting errors, ungrammatical or meaningless sentences due to tagging errors, complete meaning changes, etc.), or problems with the evaluation method. Errors of the first class however may not necessarily affect the results of the test suite, as most systems handle incomplete or meaningless sentences rather well. Still, the results of the mentioned features may not be as reliable as those of the remaining ones.

The paradigm completion features (Table 12) show a clear advantage for those two systems that explicitly model target morphology, *HY-NMT2step* and *HY-AH*. On average, these two systems are however outperformed by the *NICT* system, confirming its first rank in the manual evaluation. Most other NMT systems yield comparable accuracies, but it is striking to see that *uedin* repeatedly ranks higher than *HY-NMT* despite its lower BLEU and manual evaluation scores. The only submitted SMT system, *HY-SMT*, clearly underperforms in almost all features. The rule-based *HY-AH* system shows good overall performance, but is penalized by its complete failure on the subordinate clause type task, probably due to some missing or defective rules. We manually checked some examples of the subordinate clause feature, as several systems completely failed on it, and are able to confirm that these systems were indeed unable to correctly generate indirect questions.[13]

The agreement features (left half of Table 13) show a somewhat different picture, with the NICT system clearly leading the board, suggesting that good data selection strategies may be more important for these types of features than explicit modeling of morphology. Still, the *HY-NMT2step* and *HY-AH* yield better scores than their official rankings would suggest.

The rare word features (right half of Table 13) surprise by the exceptional performance of the online systems. It is likely that these systems contain some type of copy mechanism to handle out-of-vocabulary words, whereas such mechanisms are typically not included in research systems. The participating NMT systems use three different subword splitting algorithms: *Aalto* uses Morfessor, *talp-upc* and *CUNI-Kocmi* use wordpieces as implemented in Tensor2Tensor, and *NICT, HY-NMT* and *uedin* use byte-pair encoding. The results suggest that byte-pair encoding performs better than its competitors, but a more careful analysis would be required to confirm this hypothesis. The best performance in rare word features is achieved by online systems B and G, but without knowledge of their internals, we cannot link this performance to training data or dedicated components.

Although a large-scale manual evaluation of the sentence pairs was not within the scope of this paper, a number of English-to-Finnish sentence pairs were extracted for a manual "sanity check". In particular, we focused on cases where only the rule-based system output was evaluated as correct, in order to identify potential false positives/negatives caused by the equally rule-based scoring procedure. One observed weakness of the scoring procedure is that it favors more literal (word-for-word) renderings of the source. This tendency produces false negatives in the cases where the NMT output contained a less literal translation, which may however be both fluent and adequate. False positives can also be observed in some cases where the literal translation in the RBMT output, marked correct, is in fact not a correct translation of the source. These often involved idiomatic expressions (such as *This brings us to X*), which occasionally occur in the sentence pairs even though idioms had been excluded to the ex-

[13] Most failing translations used one of the following constructions: *Hän kysyi, jos se ei tapahtuisi Kaliforniassa.* / *Hän pyyti jos se ei tapahtuisi Kaliforniassa.* 'She asked if it would not happen in California.'

tent possible.

The stability features (Table 14) show lower figures on average. As could be expected, the rule-based system is the most stable one, as it explicitly encodes the mappings between English and Finnish morphological categories. The online systems again performed quite well on these features. Again, the SMT system is worse than the NMT systems, something that was not necessarily expected, as SMT systems tend to produce more literal translations than NMT systems. Similarly to the German consistency features, the Finnish stability features do not seem to correlate strongly with the human judgement scores. In particular, the poor scores of *CUNI-Kocmi* are surprising and not expected from the other features.

As noted above, stability is not necessarily always a reflection of overall quality, and it may not always be most adequate to produce identical translations for sentence pairs differing in only one feature (verb tense, pronoun gender, definiteness). An interesting example of this was observed in the case of indefinite and definite determiners. As Finnish lacks determiners, translations for sentences involving the definiteness contrast were expected to be identical. This was generally the case for the RBMT system, but NMT systems were observed to produce sentences with word order changes that are used in Finnish to indicate distinctions corresponding to the English definite/indefinite articles. The sample extracted for this manual check is insufficient to determine whether these word order differences can be considered something the NMT system has learned from the corpus or simply random variation, but the observation that they occur is interesting. Certainly, NMT systems do have the capacity to learn to express sentence information structure but it is not yet clear if it is sufficiently exemplified in the training data.

An overall point should also be made that the sentence pair evaluation only compares the specific feature being evaluated, or compares whether the sentences are identical in the case of the stability features. The overall correctness, adequacy or fluency is not evaluated, and sentences evaluated as correct for a specific feature may – and indeed often do – contain other errors or problems.

### 3.4 Turkish–English

Finally, we present our evaluation results for Turkish–English in Table 15.[14]

We can observe that none of the systems perform particularly well on either of the participle contrast pairs. Interestingly, performance is worse on the more frequent subject participles. There is also a stark difference in performance across different systems in subject participles, with *Online-A*'s accuracy (30.5%) being almost twice that of *uedin* (17.0%).

Again, the overall translation performance is not quite in line with the performance on our test suite.

## 4 Conclusions

The contrastive evaluation of morphological competence, as introduced by B&Y, has proved to be easy to adapt to additional language pairs and linguistic features. The data collected from the systems participating in WMT18 allows for fine-grained analysis of the impact of system architectures, training parameters and data on the various aspects of morphological competence. In general, the systems that perform well on global quality evaluation also show good morphological competence, but a few striking differences have been found. First, rule-based systems such as *HY-AH* for English–Finnish tend to obtain much higher morphology scores than expected from their overall quality. This is not surprising, as rule-based systems usually contain an explicit morphological generation component, but it requires more research on the factors that influence the correlation between morphological tests and overall translation quality. Second, we found that features focusing on consistency and stability (i.e., those presented in Tables 8, 11 and 14) correlate poorly with human judgement. This suggests that the robustness of current MT system has almost no relation to their quality.

## Acknowledgments

This research has been supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780069 and by the grant 18-24210S of the Czech Science Foundation.

[14] Unfortunately, we did not obtain the output of the *Alibaba-Ensemble* system in time for evaluation.

## References


Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. 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.

Franck Burlot and François Yvon. 2018. Evaluation morphologique pour la traduction automatique: adaptation au français. In *Conférence sur le Traitement Automatique des Langues Naturelles*, TALN, Rennes, France. ATALA.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2011. HFST – framework for compiling and applying morphologies. In *Proceedings of the International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer, Berlin, Heidelberg.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

Tommi A. Pirinen. 2015. Omorfi —free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004*, pages 1263–1266.

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.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).