A Test Suite and Manual Evaluation of Document-Level NMT at WMT19

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková and **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{rysova, magdalena.rysova, musil, polakova, bojar}@ufal.mff.cuni.cz

Abstract

As the quality of machine translation rises and neural machine translation (NMT) is moving from sentence to document level translations, it is becoming increasingly difficult to evaluate the output of translation systems.

We provide a test suite for WMT19 aimed at assessing discourse phenomena of MT systems participating in the News Translation Task. We have manually checked the outputs and identified types of translation errors that are relevant to document-level translation.

1 Introduction

Currently, the level of machine translation systems can be very good or excellent. For some languages, the systems are on par with humans when evaluated *at the level of individual sentences*, see Hassan et al. (2018) for Chinese-to-English and Bojar et al. (2018) for English-to-Czech translation at WMT18. The main criterion for distinguishing MT systems' quality thus has to shift from evaluating individual sentences to larger units. Ideally, the translated text should be now evaluated as a whole.

We believe that the fundamental criterion of the quality of manual or automatic translation is the extent to which the translation is functional in human communication. These days, the critical basic level in this criterion has been already reached by multiple machine translation systems covering a wide range of language pairs. While the reader of an automatically translated text may be groping at some points in the text, the overall quality of the translation is already so high that the main content of the text and the author's communicative intention is mostly conveyed.

Still, the reader of an MT output takes a higher effort to understand the translated text. For example, morphological errors, shortcomings in the word order, incorrect syntactic relations, failure in translating terminology, or the choice of inappropriate synonyms can hinder the speed and accuracy of text understanding.

In this paper, we first provide a test suite for WMT19 aimed at assessing translation quality of English to Czech NMT systems regarding document-level language phenomena. As qualitative analyses of document-level errors in MT outputs are up-to-date quite rare, this paper further aims at identification, manual annotation and linguistic description of these types of errors relevant to English-Czech NMT and a comparison of performance of the submitted systems in the given areas. We compare NMT systems that translate one sentence at a time with systems that have more than one sentence on input and therefore have potential to translate document-level phenomena better.

After an overview of detected translation errors from various levels of language description, the paper zooms in on three document-level, or coherence-related, phenomena: topic-focus articulation (information structure), discourse connectives and alternative lexicalizations of connectives.[1] We assume that translation systems might have difficulties with these phenomena, as they are related to the previous context and go beyond (or are affected by the phenomena across) the sentence boundary. In this way, they contribute to the overall coherence of the text that should (as a whole) function as an independent unit of human communication.

[1] This work does not address in detail errors in coreference, pronoun and gender translation, as these phenomena have been already widely accounted for, e.g. Guillou et al. (2016); Novák (2016).

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 455–463
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

2 Data

The evaluations in this paper are conducted on a selection of 101 documents from the parallel Prague Czech-English Dependency Treebank (PCEDT, Hajič et al. (2012)), and we also used discourse annotations of the same texts in the Penn Discourse Treebank 3.0 (PDTB, for details see Webber et al. (2019)).

2.1 Prague Czech-English Dependency Treebank

The Prague Czech-English Dependency Treebank is a parallel corpus consisting of English original texts and their Czech translations. The PCEDT contains 1.2 million running words in almost 50,000 sentences in each part.

The English texts come from the Penn Treebank (Wall Street Journal Section; Marcus et al., 1993). They were manually translated into Czech by trained linguists without any support of MT and proofread. The PCEDT is manually annotated on the tectogrammatical (deep-syntactic) layer in both languages. The sentences are represented by dependency structures of content words. The nodes in the tree structures are provided with syntactico-semantic labels as, e.g., predicate, actor, patiens, addressee or locative. Also, the valency frames of verbs (argument structure) are captured, as well as elliptical structures and anaphoric relations.

In addition, the Czech part is automatically tagged and parsed as surface-syntactic dependency trees on the analytical layer. The English part also preserves the original phrase-structure annotation of the Penn Treebank. Also, the annotation of discourse relations, connectives and Altlexes from the Penn Discourse Treebank was extracted and added to our PCEDT dataset.

3 NMT Systems

We evaluated 5 NMT systems from those participating in WMT19 in English-to-Czech translation. In particular, we selected those of the highest quality as estimated by automatic scoring at matrix.statmt.org.[2]

`CUNI-Transf-2018` is last year submission by Popel (2018). It is a neural machine translation model based on the Transformer architecture and trained on parallel and back-translated monolingual data. It translates one sentence at a time.

`CUNI-DocTransf-T2T` is a Transformer model following Popel (2018), but trained on WMT19 document-level parallel and monolingual data. During decoding, each document was split into overlapping multi-sentence segments, where only the "middle" sentences in each segment are used for the final translation. `CUNI-Transf-T2T` is the same system as `CUNI-DocTransf-T2T`, just applied on separate sentences during decoding.

`CUNI-DocTransf-Marian` is document-level trained Transformer in Marian framework following Popel (2018), but finetuned on document-level parallel and monolingual data by translating triples of adjacent sentences at once. If possible, only the middle sentence is considered for the final translation hypothesis, otherwise a double or single sentence context is used.

`Online-B` is an anonymized online system which we know also from several previous years of WMT.

`Reference` is the Czech side of the PCEDT corpus.

4 Annotation Design

The 101 PCEDT documents selected for translation and manual evaluation belong to the "essay" and "letter" genre labels according to the classification of PDTB given in Webber (2009). At the same time, the selected texts have a length of 20–50 sentences. These documents were submitted as an additional test suite for Machine Translation of News shared task at the WMT 2019. Because we are interested in document-level translation and the effect of context on the translation, we only selected documents with cross-sentence discourse relations.

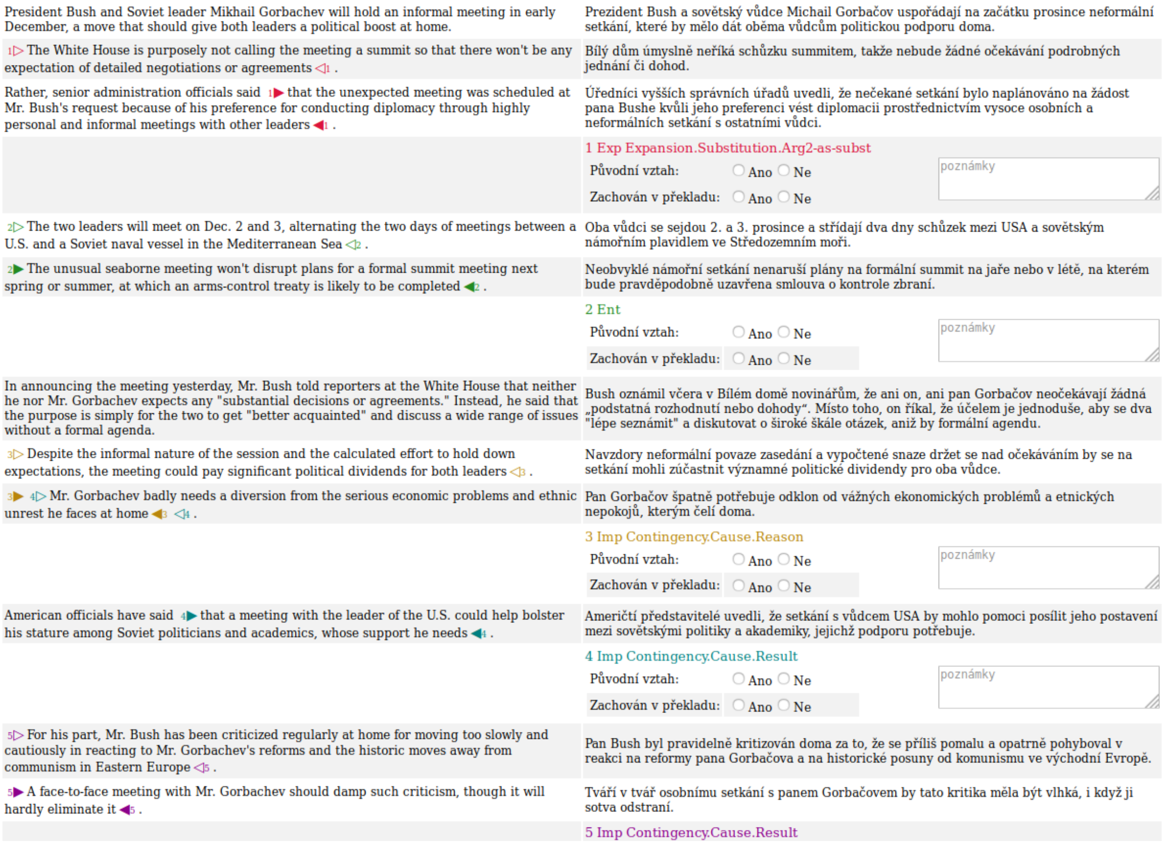
We have created a simple annotation interface (see Figure 1), which allows the annotator to mark the items that were translated correctly.

Specifically, several types of cross-sentence discourse relations are considered on the source side (reusing the annotations available in the Penn Discourse Treebank 3.0).

The target side was validated by trained linguists. For each of the observed connectives / AltLex, the annotators indicated whether:

(1) the given expression/phrase in the source fulfills the function of a connective – according to the

[2] http://matrix.statmt.org/matrix/systems_list/1896



President Bush and Soviet leader Mikhail Gorbachev will hold an informal meeting in early December, a move that should give both leaders a political boost at home.	Prezident Bush a sovětský vůdce Michail Gorbačov uspořádají na začátku prosince neformální setkání, které by mělo dát oběma vůdcům politickou podporu doma.
1▷ The White House is purposely not calling the meeting a summit so that there won't be any expectation of detailed negotiations or agreements ◁1 .	Bílý dům úmyslně neříká schůzku summitem, takže nebude žádné očekávání podrobných jednání či dohod.
Rather, senior administration officials said 1▶ that the unexpected meeting was scheduled at Mr. Bush's request because of his preference for conducting diplomacy through highly personal and informal meetings with other leaders ◀1 .	Úředníci vyšších správních úřadů uvedli, že nečekané setkání bylo naplánováno na žádost pana Bushe kvůli jeho preferenci vést diplomacii prostřednictvím vysoce osobních a neformálních setkání s ostatními vůdci.
	1 Exp Expansion.Substitution.Arg2-as-subst Původní vztah: ○ Ano ○ Ne Zachován v překladu: ○ Ano ○ Ne poznámky
2▷ The two leaders will meet on Dec. 2 and 3, alternating the two days of meetings between a U.S. and a Soviet naval vessel in the Mediterranean Sea ◁2 .	Oba vůdci se sejdou 2. a 3. prosince a střídají dva dny schůzek mezi USA a sovětským námořním plavidlem ve Středozemním moři.
2▶ The unusual seaborne meeting won't disrupt plans for a formal summit meeting next spring or summer, at which an arms-control treaty is likely to be completed ◀2 .	Neobvyklé námořní setkání nenaruší plány na formální summit na jaře nebo v létě, na kterém bude pravděpodobně uzavřena smlouva o kontrole zbraní.
	2 Ent Původní vztah: ○ Ano ○ Ne Zachován v překladu: ○ Ano ○ Ne poznámky
In announcing the meeting yesterday, Mr. Bush told reporters at the White House that neither he nor Mr. Gorbachev expects any "substantial decisions or agreements." Instead, he said that the purpose is simply for the two to get "better acquainted" and discuss a wide range of issues without a formal agenda.	Bush oznámil včera v Bílém domě novinářům, že ani on, ani pan Gorbačov neočekávají žádná „podstatná rozhodnutí nebo dohody“. Místo toho, on říkal, že účelem je jednoduše, aby se dva "lépe seznámit" a diskutovat o široké škále otázek, aniž by formální agendu.
3▷ Despite the informal nature of the session and the calculated effort to hold down expectations, the meeting could pay significant political dividends for both leaders ◁3 .	Navzdory neformální povaze zasedání a vypočtené snaze držet se nad očekáváním by se na setkání mohli zúčastnit významné politické dividendy pro oba vůdce.
3▶ 4▷ Mr. Gorbachev badly needs a diversion from the serious economic problems and ethnic unrest he faces at home ◀3 ◁4 .	Pan Gorbačov špatně potřebuje odklon od vážných ekonomických problémů a etnických nepokojů, kterým čelí doma.
	3 Imp Contingency.Cause.Reason Původní vztah: ○ Ano ○ Ne Zachován v překladu: ○ Ano ○ Ne poznámky
American officials have said 4▶ that a meeting with the leader of the U.S. could help bolster his stature among Soviet politicians and academics, whose support he needs ◀4 .	Američtí představitelé uvedli, že setkání s vůdcem USA by mohlo pomoci posílit jeho postavení mezi sovětskými politiky a akademiky, jejichž podporu potřebuje.
	4 Imp Contingency.Cause.Result Původní vztah: ○ Ano ○ Ne Zachován v překladu: ○ Ano ○ Ne poznámky
5▷ For his part, Mr. Bush has been criticized regularly at home for moving too slowly and cautiously in reacting to Mr. Gorbachev's reforms and the historic moves away from communism in Eastern Europe ◁5 .	Pan Bush byl pravidelně kritizován doma za to, že se příliš pomalu a opatrně pohyboval v reakci na reformy pana Gorbačova a na historické posuny od komunismu ve východní Evropě.
5▶ A face-to-face meeting with Mr. Gorbachev should damp such criticism, though it will hardly eliminate it ◀5 .	Tváří v tvář osobnímu setkání s panem Gorbačovem by tato kritika měla být vlhká, i když ji sotva odstraní.
	5 Imp Contingency.Cause.Result

Figure 1: Screenshot of the annotation interface.

annotator, or the function of AltLex – according to the original English annotation displayed. If yes, then whether its Czech translation is (2):

- adequate and correctly placed,[3]
- adequate but incorrectly placed,
- omitted and it does not harm the output
- omitted and it harms the output
- not adequate.

The questionnaire for word order annotation is analogous, compare the description of tables with results below in Section 7. The original translation into Czech from PCEDT could serve as a reference translation but similarly to Bojar et al. (2018), we opted for a bilingual evaluation, showing the annotators always the source and the candidate translation. The benefit is that the human translation can be evaluated using the same criteria as the MT system outputs.

There were 6 annotators, all of them students of linguistics. Each annotator evaluated 8 documents in the first round. For each document, they evaluated the output of one MT system (without knowing which MT system produced the output). To measure the inter-annotator agreement, we organized a second round of evaluation, where each annotator was given documents and systems combination that was in the first round evaluated by another annotator. Details on the IAA are given in Section 7.

5 Linguistic Analysis of Translations Errors across Language Levels

We carried out a complex linguistic analysis of a sample of the translated texts and revealed that even the best translations contained cca 15–20 linguistic issues (per text of 35 sentences). This means that although the content reliability and linguistic level of (the best) MT systems is very high, they still do not reach communication skills of humans. This fact may be challenging for their authors, as there are still possibilities for improvement. However, a systematic improvement of MT systems is rather difficult due to non-systematic nature of language errors found in the analysis – e.g. if there appeared an untypical word order in a

[3] for Altlexes: and preserves the original discourse meaning

sentence, it does not mean that word order errors are also present in the rest of the translated text. It turned out, on the contrary, that the errors / problematic issues appear individually, as singularities.

In the following part, we discuss the problematic places in a sample of translated texts. We tried to select the best or (at least) good MT systems to demonstrate that even in such an advanced translation, there are still issues requiring improvement.

5.1 Morphology

We were able to detect errors from various levels of language description. Some problematic issues concerned even such basic phenomena as e.g. the use of a verbal mood or other morphological issues (*It's as if France decided to give only French history questions to students in a European history class, and when everybody aces the test, they say their kids are good in European history – Je to, jako by se Francie rozhodla dávat studentům evropských hodin dějepisu jen otázky z francouzštiny,* ***a když všichni v testu excelují, říkají****, že jejich děti jsou v evropských dějinách dobré*; the Czech translation is not consistent in maintaining potentiality: the intended content should be translated into Czech as: *jako* ***kdyby*** *se Francie..., a* ***až by*** *všichni v textu excelovali, řekli* ***by****...*) with the obligatory conditional morpheme *by*, also as a part of the conjunction *kdyby*, used in past (unreal) conditions.

5.2 Lexicon

Other issues concerned the choice of vocabulary. The individual translations included e.g. inappropriate repetition of the same word (ie. the MT systems produced a non-natural output by not attempting to use a synonym, cf. *in test-coaching workbooks and worksheets — v pracovních sešitech a pracovních sešitech* "in test-coaching workbooks and in test-coaching workbooks"). In some of them, there also appeared incorrect literal translations of terms (cf. *a joint venture of McGraw-Hill Inc. and Macmillan's* ***parentt****, Britain's Maxwell Communication Corp – společným podnikem McGraw-Hill Inc. a Macmillanovým* ***rodičem****, britskou společností Maxwell Communication Corp*).

Another lexical issue was the use of an inaccurate synonym in a given context (cf. *but he doesn't deny that some items are similar – ale nepopírá, že některé předměty jsou podobné*; the word *předměty* may be a synonym to the original *items* but not in this context, the Czech word here means rather tangible *objects*).

Generally, the MT systems succeed in translating basic words or phrases but sometimes they fail in translating terms or technical words and in lexical variety (often resulting in word repetition and failure to use an appropriate synonym).

5.3 Syntax

The translations also exhibit signs of incorrect syntactic relations, e.g. excessive genitive accumulation, which is untypical for Czech (cf. *About 20,000 sets of Learning Materials teachers' binders have also been sold in the past four years. – Asi 20 000 souborů (Noun in Gen) učebních materiálů (NP in Gen) učitelských pořadačů (NP in Gen) bylo také prodáno v posledních čtyřech letech.*). Another typical syntactic error appears in translation of syntactically potentially homonymous phrases, as in the example above in 5.1 the phrase *European history class*, translated wrongly as *evropských hodin dějepisu* (European classes of history).

Also, a large problematic area was revealed in word order configurations. Some translations contained the word order adopted from English, where it is untypical or even incorrect in Czech. This issue is related to sentence information structure or topic-focus articulation, as the word order is connected with contextual boundness (cf. *. . . says "well over 10 million" of its Scoring High test-preparation books have been sold since their introduction 10 years ago – uvádí, že "více než 10 milionů" jeho testovacích knih Scoring High se prodalo od jejich zavedení před 10 lety*; the expression *"více než 10 milionů"* is the focus of the sentence and therefore it should be placed in the final position in Czech). Similar issue (concerning topic-focus articulation) may be observed in the sentence *Scoring High and Learning Materials are the best-selling preparation tests. – Scoring High and Learning Materials jsou nejprodávanější přípravné testy*. Again, the expression *Scoring High and Learning Materials* should be (as focus proper of the sentence) placed in the final sentence position in Czech.

5.4 Semantics

Semantic issues (to a certain extent) are already partly included in the incorrect translations of terms as discussed above. Other are related especially to factual inaccuracy, e.g. the expression

French history questions was incorrectly translated as *otázky z francouzštiny* "questions from French".

In some cases, even a whole part of the original text was completely omitted in the translation – the meaning of the sentence was thus negatively affected (*... and Harcourt Brace Jovanovich Inc.'s Metropolitan Achievement Test and Stanford Achievement Test – ... a Harcourt Brace Jovanovich*).

5.5 Discourse

Further issues in translations also appeared on higher levels of language description, crossing the sentence boundary and mostly affecting text understanding as a whole. These discourse-related phenomena include especially coreference and discourse (semantico-pragmatic) relations, largely expressed by discourse connectives or their paraphrases (AltLexes). A detailed analysis of discourse-related translation errors is given below in Section 6.1.

6 Linguistic Analysis of Selected Document-Level Errors

6.1 Selected coherence phenomena

A comprehensive linguistic analysis of a sample of translated texts showed that even the best translations are not completely error-free (the best ones contained about 15–20 errors per text). These errors were further analyzed – they appear across individual levels of language description. Unfortunately, the main common feature of the errors seems to be the fact that they are not systematic. The key to a good distinction of translation quality is thus their complex linguistic analysis. For the annotation, we have chosen three document-level types of the errors discovered in the output analysis, namely those concerning **topic-focus articulation, discourse connectives** and the meanings they convey and **alternative lexicalizations of connectives** (AltLexes). The annotators then assessed them on a larger sample of translated data from all the systems and the reference translation. The finding are analyzed linguistically in the rest of this Section and quantitatively below in Section 7.

6.1.1 Topic-focus articulation and word order

First, we observed the phenomenon of topic-focus articulation (we follow this phenomenon as presented within the Functional Generative Description, see Sgall (1967) or Sgall et al. (1986)). In our experiment, we took advantage of the fact that English and Czech have a different word order system in combination with topic-focus articulation and contextual boundness.[4] While English has a fixed word order, strongly influenced by grammar, Czech has a free word order mainly influenced by the contextual boundness of individual sentence constituents. It is thus necessary to harmonize the word order in a Czech sentence always with respect to the previous (con)text.

In the annotation of the translated texts, we focused on the word order of the subject. While the subject is typically at the beginning of the sentence in English, it can occupy various positions in Czech, depending on whether it is contextually bound or not. We were wondering how individual MT-systems reflect this word order issue.

We automatically selected English original sentences from the PCEDT that contained a noun used with an indefinite article in the subject position and its Czech counterparts in evaluated translated texts. It is assumed that this subject is contextually non-bound (not deductible from the previous context, it is "new" information) and is thus expected elsewhere than at the beginning of the sentence, most likely to follow the predicate in Czech. Moreover, this subject (or the constituent corresponding to it in Czech) could be also so-called *focus proper* standing at the very end of the Czech sentence in written texts.

For Czech translations, it was necessary to check whether the Czech equivalent of the English subject was retained as a contextually non-bound sentence constituent and whether it was appropriately located in the Czech sentence, see the following example.

English text: *What is the best-selling preparation test? A NEW LANGUAGE TEST is the best-selling preparation test.*

Expected Czech translation: *Co je nejprodávanějším přípravným testem? Ne-*

[4] For definitions of terms related to topic-focus articulation and contextual boundness see Hajičová et al. (1998).

jprodávanějším přípravným testem je NEW LANGUAGE TEST.

6.1.2 Discourse connectives and their sentence positions

The second phenomenon assessed in the annotation were discourse connectives. Discourse connectives are rather short function words (e.g. *but, therefore, nevertheless, because,* or *and*) that connect two text units while expressing a discourse (semantico-pragmatic) relation between them, thus ensuring text to a large extent text coherence and cohesion. Here, the problematic issues included the use of a wrong Czech equivalent – both from the semantic and grammatical point of view (e.g. the positions of connectives in a sentence etc.). An example of a wrong connective translation is as follows. ***Since** chalk first touched slate, schoolchildren have wanted to know: What's on the test? – ***Protože** se křída poprvé dotkla břidlice, žáci chtěli vědět: Co je na testu?*

The English connective *since* is homonymous and its meaning may be causal or temporal. In the example, it was translated as causal (by the Czech connective *protože – because*) in a temporal context (the correct Czech translation here would be *od okamžiku, kdy* (from the moment when...). Such an incorrect translation of a discourse connective demonstrates nicely the potential huge impact on overall comprehensibility.

From the word order perspective, even these cohesive devices have their typical positions in a clause – according to their part-of-speech classification. Coordinating conjunctions typically stand between two discourse units (*I play the flute **and** I dance. / Hraju na flétnu **a** tančím.*) both in English and Czech. Subordinating conjunctions typically occur at the beginning of the discourse unit to which they belong syntactically (*Because it rains, I'm not going out. I won't go out because it rains. / **Protože** prší, nepůjdu ven. Nepůjdu ven, **protože** prší.*). Connectives of adverbial origin have looser positions in some cases;[5] they can occur e.g. in the first and second position in the sentence (*For me it is easier to not lose a game than to win it, **thus** I produce better results in stronger tournaments. Both umpires claimed that they were unsighted, and were **thus** forced to give Somny the benefit of the doubt. / Pro mě je snazší neztratit hru, než ji vyhrát, **proto** dosahuji lepších výsledků v silnějších turnajích. Oba rozhodčí tvrdili, že neviděli, byli **proto** nuceni dát Somnymu výhodu pochybovat.*).

In some word-order positions of discourse connectives, English and Czech differ. In other words, a Czech translation should not copy the connective ordering from an English original. In English, some discourse connectives can occur e.g. at the very end of the sentence (cf. *too, as well, instead, nevertheless* etc.), which is not typical for Czech.

To better compare the quality of the individual translations, we observed especially the translation equivalents of multi-word connectives like *as long as* or *as much as* that could be problematic due to their idiomatic character.

6.1.3 Alternative lexicalizations of discourse connectives (AltLexes)

In addition to discourse connectives, discourse relations can also be expressed by their alternatives called AltLexes, see Prasad et al. (2010). Alternative lexicalizations of connectives are often multi-word phrases such as *for this reason*. Since these cohesive structures often have an idiomatic character and they generally do not achieve such degree of grammaticalization as connectives, their forms in languages may vary to a large extent.

For example, the AltLex *for this reason* is not translated into Czech literary as *pro tento důvod* 'lit. for this reason', but as *z tohoto důvodu* 'lit. from this reason'. Other examples of English AltLexes are *that's all, that's largely due to, attributed that to, it will cause* etc. A list of AltLexes in English is given in Prasad et al. (2007), multi-word connective expressions in Czech are described and presented in Rysová (2018). Due to their high lexical variety and lower degree of grammaticalization, AltLexes were selected for the annotation as potentially interesting expressions for translation.

7 Results

In this section, we present the results of the evaluation.

7.1 Inter-annotator agreement

The inter-annotator agreement was measured pairwise, it ranges from 66 % to 93 % with an average of 80 %. The agreement was on average 69 % for AltLexes, 87 % for connectives and 79 % for questions concerning word order.

[5] For more information see Rysová and Rysová (2018).

7.2 AltLexes

The annotation interface for alternative lexicalizations contained identical questions to those for connective assessment (described above in Section 4), with the exception of their (in)correct placement, as this question is irrelevant for such non-grammaticalized phrases. There were 23 queries in average for each of the evaluated translations. The results for adequacy of AltLex translations in each system output AND the reference are summed up in Table 1.

A source AltLex was assessed as an appropriate connecting device in accordance with the original discourse annotation in 130 cases (Yes), and inappropriate in 42 cases (No). The proportion of negative answers is surprisingly high, but a closer look on the data reveals that the annotators, quite in unity (but in contrast to the PDTB notion of AltLex), resist treating **verbs** as a specific form of connecting devices. This mostly concerns causative verbs like *to explain, to strengthen* or *to blame*. They might be in fact right, these verbs are mostly translated well and their role in discourse coherence is a rather supplementary one. Apart from this issue, Table 1 demonstrates that once an AltLex is approved as a connecting device, it is in vast majority of cases translated correctly (rarely incorrectly), the original discourse meaning is preserved and it is not omitted in the translation. This applies quite equally across all systems, with a small decrease for CUNI-DocTransf-Marian system and the reference (!). A potential explanation is the typically looser human translation (and possibly the context-aware Marian system).

7.3 Connectives

As for connectives, there were 52 queries in average for each of the evaluated translations. The results for adequacy of connective translations in each system output and the reference are summed up in Table 2. A source connective candidate was assessed as an factual connecting device in 303 cases (Yes), and not a connective in 30 cases (No). This proportion seems to be correct, the non-connective readings of some expressions are relevant, e.g. several times for *as much as* in the function (and position) of a quantifier. Once a connective candidate is approved as an actual connective, it translated always correctly (compare column "n" in Table 2), but it is possibly incorrectly placed in the translation (column "ax"). The result figures indicate that there are no significant differences across the systems in translating the traced connectives.

7.4 Word order

The word order evaluation focused the translation of contextually non-bound subjects (representing a new information in the sentence). The annotators first determined, which of the automatically preselected sentences from the English source indeed contain a contextually non-bound subject (85 Yes, 10 No). If yes, they traced whether the subject in the Czech translation also contextually non-bound. The results of manual annotation demonstrate that MT systems in general preserve the contextual non-boundness of the subjects. The figures are comparable across the systems, only the Marian system and the reference achieved a slightly worse scores:

	yes	no
CUNI-Transf-2018	11	1
CUNI-DocTransf-Marian	17	3
online-B	6	1
CUNI-DocTransf-T2T	17	1
CUNI-Transf-2019	6	1
reference	19	4

In a second task, we observed whether the subject in the English original the focus proper of the given sentence. Again, the annotators first filtered out relevant sentences (10 Yes, 36 No). Then they looked at whether the subject in the Czech translation is also the focus proper of the sentence. Similarly as in the previous task, the Marian system's performance is worse, and the performance of CUNI-DocTransf-T2T drops. However, the results here are less significant, as there were only few occurrences of the annotated tokens:

	yes	no
CUNI-Transf-2018	2	0
CUNI-DocTransf-Marian	5	3
online-B	0	1
CUNI-DocTransf-T2T	1	2
CUNI-Transf-2019	0	1
reference	1	6

Next, we followed the systems' ability to place the Czech equivalents of the original English subjects correctly into the Czech output sentence. Here, a correct placement according to the Czech word order rules was mostly achieved by all systems. There was not enough data collected for the online-B system, but the rest is comparable, with both context-aware systems performing slightly

	adequate	missing	wrong
CUNI-Transf-2018	★★★★⯪	☆☆☆☆☆	⯪☆☆☆☆
CUNI-DocTransf-Marian	★★★⯪☆	⯪☆☆☆☆	★☆☆☆☆
online-B	★★★★☆	☆☆☆☆☆	★☆☆☆☆
CUNI-DocTransf-T2T	★★★★☆	☆☆☆☆☆	★☆☆☆☆
CUNI-Transf-2019	★★★★☆	☆☆☆☆☆	★☆☆☆☆
reference	★★★⯪☆	☆☆☆☆☆	★⯪☆☆☆

Table 1: Results for AltLex annotations. Each ★ represents 20 % and the results are rounded to the nearest half-star.

	a	ax	m	n
CUNI-Transf-2018	★★★★⯪	⯪☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
CUNI-DocTransf-Marian	★★★★☆	★☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
online-B	★★★★⯪	⯪☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
CUNI-DocTransf-T2T	★★★★⯪	⯪☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
CUNI-Transf-2019	★★★★⯪	⯪☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆
reference	★★★★☆	☆☆☆☆☆	☆☆☆☆☆	⯪☆☆☆☆

Table 2: Results for connectives annotations. The columns are: (a) adequate and correctly placed, (ax) adequate but incorrectly placed, (m) omitted and it does not harm the output, and (n) not adequate. Each ★ represents 20 % and the results are rounded to the nearest half-star.

worse than others:

	yes	no
CUNI-Transf-2018	14	0
CUNI-DocTransf-Marian	14	5
online-B	3	1
CUNI-DocTransf-T2T	13	3
CUNI-Transf-2019	6	0
reference	19	3

8 Conclusion

In this paper, we have described a test suite of parallel English-Czech texts provided for WMT19 with the aim to assess discourse phenomena in output of MT systems participating in the News Translation Task. We have carried out an extensive manual annotation of the MT outputs and identified types of translation errors relevant to document-level translation. We also compared the systems' performance with respect to the observed phenomena.

In general, the recent NMT systems have achieved such a high level of translation quality that it has become difficult to evaluate their output in a systematic fashion. Most of the errors in the translation cannot be found by a simple comparison with the reference translation, a bilingual evaluation is needed. Moreover, for the observed phenomena, the systems performed with only a minor differences among each other and they reached the quality of the reference. In fact, the reference translation was in some aspects evaluated as worse, which is likely caused by the greater literal adherence of the automatic translations to the original and it does not mean that the reference is incorrect. Contrary to our assumptions, the two context-aware systems did not outperform the others in translating the followed document-level phenomena. This can be attributed to the fact that the systems perform good enough on this task already, and also partly because the evaluation can change a lot using just a slightly different annotation setting, e.g. if we traced also other (ambiguous) connective expressions or anaphoric items. The actual errors are difficult to predict from scratch and they occur randomly. More specifically, while the translations of AltLexes and discourse connectives showed quite satisfactory (at least of those observed here), the most errors (equally across systems) were detected in the area of word order and contextual (non-)boundness of the subjects. The systems prefer to keep the original word also in the translations, not really accounting for the impact of information structure.

Acknowledgement

We acknowledge support from the Czech Science Foundation project no. GA17-06123S (Anaphoricity in Connectives: Lexical Description and Bilingual Corpus Analysis) and the EU project H2020-ICT-2018-2-825460 (ELITR). This study has utilized language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). This research was partially supported by SVV project number 260 453.

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.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 525–542.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3153–3160, Istanbul, Turkey. ELRA, European Language Resources Association.

Eva Hajičová, Barbara H Partee, and Petr Sgall. 1998. *Topic-focus articulation, tripartite structures and semantic content*. Kluwer, Dordrecht.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. https://www.microsoft.com/en-us/research/uploads/prod/2018/03/final-achieving-human.pdf.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19:313–330.

Michal Novák. 2016. Pronoun prediction with linguistic features and example weighing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 602–608.

Martin Popel. 2018. Cuni transformer neural mt system for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1023–1031. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Magdaléna Rysová and Kateřina Rysová. 2018. Primary and secondary discourse connectives: Constraints and preferences. *Journal of Pragmatics*, 130:16–32.

Magdaléna Rysová. 2018. *Diskurzní konektory v češtině: Od centra k periferii*. Institute of Formal and Applied Linguistics, Praha, Czechia.

Petr Sgall. 1967. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(203-225).

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 674–682. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.