Testsuite on Czech–English Grammatical Contrasts

Silvie Cinková **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
<surname>@ufal.mff.cuni.cz

Abstract

We present a pilot study of machine translation of selected grammatical contrasts between Czech and English in WMT18 News Translation Task. For each phenomenon, we run a dedicated test which checks if the candidate translation expresses the phenomenon as expected or not. The proposed type of analysis is not an evaluation in the strict sense because the phenomenon can be correctly translated in various ways and we anticipate only one. What is nevertheless interesting are the differences between various MT systems and the single reference translation in their general tendency in handling the given phenomenon.

1 Introduction

English and Czech are typologically different languages. It goes without saying that some structural phenomena of either lack a direct structural equivalent in the other; for instance, Czech has not grammaticalized noun definiteness, while it boasts a complex system of verb aspect, which is absent in English. Such *1:n* correspondences can pose translation problems in human as well as in machine translation. Intuitively, a translation system that has mastered these *1:n* phenomena ought to be more successful than one that has not. Therefore we investigate whether there is a positive correlation between mastering some of these problematic phenomena and the performance of an En-Cs MT system.

2 Selected Linguistic Phenomena

Based on our experience as Czech learners of English, translators and developers/evaluators of MT systems, we have selected the following phenomena for EN-CS translation evaluation: **English gerundial clause** and English verb control with **controlled infinitive**.

The data comes from a manually-parsed, word-aligned parallel treebank of English news texts and their human Czech translations (see Section 3).

2.1 English gerundial clause (and other *ing*-forms)

Modern Czech has no counterpart of the English gerund. Older Czech (i.e. until approximately 1950), used to have a verb form called present **transgressive**, which would be very handy to translate many cases of English gerundial clauses, but this form is perceived as archaic and hardly ever used. Modern Czech has the following options to render the English gerund:

1. finite clause with a choice of subordinators or conjunctions;

2. non-finite clause (infinitive clause, nominalization, or adjective/present participle).

In this study we tested whether the Czech equivalent in the reference vs. automatic translation was a finite clause or anything else.

2.1.1 Czech finite clause as equivalent to English gerundial clause

Czech is more sensitive to convoluted expressions than English. Therefore non-finite clauses are usually most smoothly translated with finite clauses. To keep the Czech text coherent, though, human translators usually link the gerundial clause to the main clause with an explicit discourse connective – either a conjunction or a subordinator, based on their knowledge of context and their world knowledge. This may pose a challenge for MT systems. The most typical discourse connectives used to translate gerundial clauses would be *-li* (a clitic *if* or *whether*), *což* (*which* referring to a predicate), *protože* (*because*), *když* (*when*), *že* (*that* as subordinator), *jak* (approximately *as* expressing

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 565–575

https://doi.org/10.18653/v1/W18-6434

an event parallel to the main-clause event), and *a* (*and*). Example:

(1) When they arrived at the door, all were afraid to go in, **fearing** that they would be out of place.
Ale když přišli ke dveřím, všichni se báli vstoupit, **protože se báli**, že budou působit trapně.
(But when they arrived at the door, all were afraid to go in, **because they feared** that they would be out of place.)[1]

(2) He said he was surprised by the EC's reaction, **calling** it "vehement, even frenetic."
Řekl, že byl překvapen reakcí ES, **a nazval** ji „prudkou, ba i bouřlivou".
(He said he was surprised by the EC's reaction, **and he called** it "vehement, even frenetic".)

2.1.2 Czech infinitive as equivalent to English gerundial clause

Infinitive clause occurs in our sample to translate gerundial clauses in the subject position and in control in some verbs. Example:

(3) **Avoiding** failure is easy.
Vyhnout se neúspěchu je snadné.
(**To avoid** failure is easy.)

(4) So far no one has suggested **putting** the comptroller back on the board.
Zatím nikdo nenavrhl znovu **dosadit** do Rady také kontrolora.
(So far no one has suggested **to put** the comptroller back on the board.)

2.1.3 Nominalizations as equivalents to gerundial clause

The choice between deverbal noun and event noun is lexically motivated. A deverbal noun is a noun derived from a verb stem by suffixes *-ní, -tí*; e.g. *stát* v. – *stání* n., *proklít* v. – *prokletí* n. This is an almost universal derivational mechanism, but it is stylistically associated with officialese and easily overused.

An event noun is a noun with either no derivative relation to any semantically close verb stem (*restaurace*, n. – NULL v.[2]) or a less productive derivation relation to a verb stem; e.g. *podpořit* v. – *podpora* n., *letět* v. – *let* n.). Also these nominalizations are to be used sparingly to preserve readability.

Example:

(5) Consider **adopting** your spouse's name.
Zvažte **přijetí** příjmení svého partnera.
(Consider **the adopting** of your spouse's name.)

(6) The Canadian wound up **writing** a check.
Kanaďan ukončil vysvětlování **vypsáním** šeku.
(The Canadian wound up **with the writing** of a check.)

(7) Fear of AIDS hinders **hiring** at few hospitals.
Strach z AIDS komplikuje **nábor** v několika nemocnicích.
(Fear of AIDS hinders **recruitment** at few hospitals.)

2.1.4 Present participle as equivalent to gerundial clause

The Czech present participle is derived from a verb but behaves like a regular adjective, including inflection; e.g. *spát* v. – *spící* adj.

As an equivalent to the English gerundial clause it requires a syntactic transformation of the source clause, approximately as though the original clause contained a participial clause instead of the gerund. Square brackets in the following example show the syntactic dependencies in English *imagined by the translator* and the corresponding structure in Czech. The main predicate is typeset in bold. Example:

(8) [[Mr. Fukuyama, [peering]] through binoculars at the end of history, **said**] ... [[Francis Fukuyama [nakukující]] skrz brýle na konci historie, ... **uvádí**], že ...

(9) [Other steelmakers **envision** steel [roofs [covering]] suburbia.] [Další výrobci oceli si **představují** ocelové [střechy [pokrývající]] předměstí.]

[1] To make the structure of the target *Czech* reference sentence more accessible to non-Czech speakers, we enhance this paper with their literal English translations. We enclose these—naturally awkward—sentences in parentheses. Where useful, we highlight the contrast in **bold**.

[2] The verb *restaurovat* means *restore*, whereas the noun *restaurace* means *restaurant*.

2.2 English infinitive clause

The English infinitive clause has many functions; e.g. verb control or a convoluted subordinate clause.[3]

Infinitive as controlled verb in verb control is present in both languages, but the sets of infinitive-controlling verbs differ. Other uses of the English infinitive clause, also present in the data, have different structure equivalents in Czech—mostly different types of finite subordinate clauses. A correct parsing would possibly make it easier for an MT system to select a plausible Czech equivalent structure, but the parser was not able to reliably identify the correct syntactic governing node of an infinitive clause in our data sample.

Since we could not rely on the parser to tell infinitive clause as an argument from an adjunct, we did not limit our search to arguments. Our sample contains the following Czech structural equivalents to English infinitive clauses:

1. **infinitive** or **noun phrase**;
2. **finite clause**.

2.2.1 Infinitive as controlled verb

A proportion of verb control cases have a *1:1* translation to Czech.

Example:

(10) Comair said it paid cash but *declined* **to disclose** the price.
Společnost Comair uvedla, že zaplatila hotově, avšak *odmítla* **uvést** cenu.

However, many English controlling verbs have a Czech equivalent verb that cannot act as a controlling verb. To avoid a verbose paraphrase with an expletive pronoun and a subordinate content clause, Czech can resort to a nominalization (deverbal noun or event noun; see Section 2.1.3), e.g.:

(11) Mr. Friend says he agreed **to strike** Mr. Alexander above the belt.
Pan Friend říká, že souhlasil **s udeřením** pana Alexandera nad opaskem.
(Mr. Friend says he agreed **with a striking** of Mr. Alexander above the belt.)

[3]English infinitive clauses occur in the *En-control* task. Despite its name, the *En-control* dataset does not only contain pure instances of grammatical verb control (e.g. *Peter planned to leave*, where *Peter* is the subject of *plan*, *leave* alike), but also other infinitive clauses depending on a finite verb, e.g. consecutive clauses (*The party has gathered enough votes to force the bill through*).

The verbose translation would say the following, being more explicit on the subject of the hitting event that the original Czech translation was:

(12) Pan Friend říká, že souhlasil **s tím, že udeří** pana Alexandera nad opaskem.
(Mr. Friend says he agreed **with it that he would strike** Mr. Alexander above the belt.)

2.2.2 Finite clause as equivalent to English infinitive clause

English has an infinitive structure that resembles a consecutive clause but involves a semantic shift towards temporal sequence of two events. This structure exists in Czech, too, but it is not common. A more natural translation would use a coordination of finite clauses. Example:

(13) The stock gained $2.75 Thursday **to close** at a then-52 week high.
Cenný papír ve čtvrtek navýšil o 2.75 dolaru **a uzavíral** na vrcholu tehdejších 52 týdnů.
(The stock gained $2.75 Thursday **and was closing** at a then-52 week high.)

Purpose and consecutive clauses, as well as content clauses, are typically finite in Czech, using a range of subordinators (cf. Section 2.1.4).

Examples:

(14) It also redesigned Oil of Olay's packaging, stamping the traditional pink boxes with gold lines **to create** a more opulent look.
Společnost rovněž změnila obal krému, na tradiční růžová políčka přidala zlaté linky, **čímž vytvořila** lukrativnější vzhled.
(It also redesigned Oil of Olay's packaging, stamping the traditional pink boxes with gold lines**, by which it created** a more opulent look.)

The following example also illustrates the effects of expressing information structure, leading to a different word order:

(15) At least three other factors have encouraged the IMF **to insist** on increased capital.
Nejméně tři další faktory přiměly MMF **k tomu, aby** na zvýšení kapitálu **trval**.
(At least three other factors have encouraged the IMF **to that it should insist** on increased capital.)

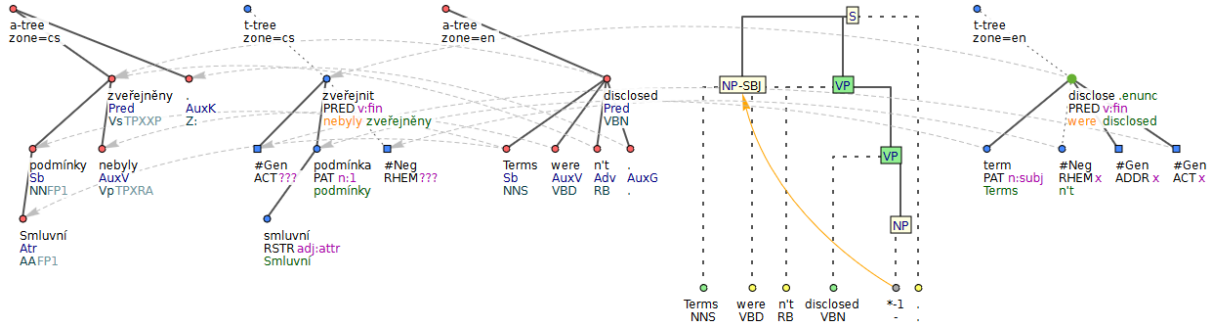

Figure 1: A sentence representation in PCEDT 2.0. The English part also contains the original PennTreebank.

3 Data Set

Our sentences come from the Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) (Hajič et al., 2012). PCEDT 2.0 is a multi-layered parallel treebank with automatic word alignment, manually built upon the Penn Treebank (Marcus et al., 1994) and its translation into Czech. It has two syntactic layers of rooted dependency trees with labeled edges: the analytical (*a*-) layer with surface syntax and the tectogrammatical (*t*-) layer with deep syntax.

In the a-layer, each word token is represented by one node. The inner structure of each node contains the word form, lemma, POS-tag, dependency label (*afun*), and reference to the governing node. The t-layer represents the linguistic meaning of each sentence by a tree that somewhat abstracts from details of morphology and surface syntax, but remains, by and large, a syntactic dependency tree. Each node contains references to the a-layer corresponding a-layer node(s), along with a whole range of other attribute values. Different reference types to content and auxiliary words, respectively. Apart from that, the t-layer provides semantic role labeling (functors), as well as coreference and ellipsis resolution.

Figure 1 illustrates the data structure of PCEDT 2.0 including the alignment links pointing from English to Czech.

3.1 Selected Sentences

We have automatically selected 3235 sentences, using the the PMLTQ search query engine (Štěpánek and Pajas, 2010). The Czech counterpart of the corpus served as the reference translation. A subset of 507 sentences was manually selected that we considered particularly apt for examination of the selected linguistic phenomena. We selected these sentences for each phenomenon separately. Each sentence in the manual selection contains (mostly) only one instance of the phenomenon in question; it is neither syntactically complex nor exceedingly long, and it is comprehensible without context.

We have made the manual selection for future experiments. These are going to involve additional manual annotation, which we do not have yet. In this paper, we therefore just compare the manual selection (called "refined set" in the following) with the 3235 automatically pre-selected sentences (called "pre-selected set") that did not meet the criteria of the manual selection.

All sentences were included in inputs of MT systems participating in the WMT18 News Translation Task. In addition to the "primary" systems CUNI Transformer, UEDIN and the online systems, we also added three baseline (contrastive) systems: CUNI Chimera, CUNI Chimera noDepfix and CUNI Moses.

CUNI Transformer (Popel, 2018) is a carefully trained system (Popel and Bojar, 2018) based on the Transformer architecture (Vaswani et al., 2017) and thus without recurrent connections.

UEDIN is an ensemble of deep RNN systems translating left-to-right and reranked by a deep right-to-left RNN model.

CUNI Moses serves as the ultimate baseline. It is phrase-based (Koehn et al., 2007) and trained on a very large parallel corpus and further adapted for the news text.

CUNI Chimera is the hybrid setup that served very well in 2013–2015 (Bojar et al., 2013). A phrase-based backbone is used to combine translations by a transfer-based system TectoMT (Žabokrtský et al., 2008), by Nematus (Sennrich et al., 2017) and by Neural Monkey (Helcl et al., 2018) with phrase pairs from the large parallel cor-

pus. The final step of Chimera was the application of a dependency-based automatic error correction tool Depfix (Rosa et al., 2012). In this paper we report the performance of both the full CUNI Chimera and a version without a the depfix post-correction, labelled CUNI Chimera noDepFix.

To give an overview of the overall quality of the systems, Table 1 presents manual (Bojar et al., 2018) and automatic[4] scores on the WMT18 newstest set of sentences.

One caveat to keep in mind is that this evaluation is based on a different set of sentences than we use in our testsuite. Since our sentences originally come from the WSJ section of the Penn Treebank, they at least belong to the domain of the translation task.

4 Evaluation

For each phenomenon we implemented a small test relying on an automatic analysis of the source English to the surface syntactic tree (a-layer, in the terminology of PCEDT), an automatic analysis of the Czech translation to surface (a-layer) along with a deep (t-layer) syntactic tree, and on automatic word alignments between the English a-layer and Czech a-layer and t-layer. We aligned directly English to each of the Czech layers; a more rigorous approach would have been aligning only the a-layers and follow the links between a-layer and t-layer on the Czech side, but since all our annotations are automatic, we do not expect much difference in these approaches due to random errors in all processing steps. The annotation was provided by the pipeline used in the creation of corpus CzEng (Bojar et al., 2016)[5] as implemented in the Treex toolkit (Popel and Žabokrtský, 2010). For the alignment, we relied on an intersection of GIZA++ (Och and Ney, 2000) alignments.

The test searched for the keyword related to the phenomenon (e.g. the controlled English verb), followed the word-alignment links to the Czech translation and tested some morphological or syntactic properties of the corresponding Czech word or node in t-layer analysis. The result of the test was "Good" if the Czech expression was the best possible translation, "Bad" otherwise, and "Unknown" if the target word or node was not found, e.g. due to errors in word alignment.

It is important to note that "Bad" does not always mean an unacceptable translation. It merely means that the translation is not the most straightforward one.

Table 2 presents the detailed results of these tests on the smaller manually refined set and also on the larger, preselected set which can be somewhat less reliable. In general, the reference seems a little harder to process ("Unk" higher than for MT systems), probably due to a less verbatim translation and thus a less straightforward word alignment. Disregarding the "Unknowns", we plot the results in Figure 2 and Figure 3 by systems and by phenomena, respectively.

5 Discussion

Do the results of the tests on our evaluation suite suggest any association of the overall translation quality with the distribution of Czech finite clauses, subordinative clauses, and non-finite clauses as equivalents of English gerund and infinitive clauses?

Figures 2 and 3 compare the systems and the numbers of Good occurrences in the individual EN-phenomenon / CS-equivalent pairs from two perspectives.

Figure 2 shows for each system how many instances of the given English phenomenon it translated with the given Czech structural equivalent. For instance, the top bar in the UEDIN subgraph, *EN-control-CS-nofinclause*, shows that UEDIN translated approximately 75% of instances of English infinitive clauses with a Czech non-finite clause (detected by automatic parsing and alignment). These detected instances are counted as Good. The proportion has been computed from the sum of Good and Bad, disregarding Unknown. The online-B system, on the other hand, translated only approx. 60% this way.

Each bar in each subgraph is accompanied with an orange diamond. This diamond represents the mean proportion of Good on the given phenomenon-equivalent pair across all systems. We can see that UEDIN was just at the mean translating English infinitives with Czech non-finite clauses, while CUNI Transformer and online-B were evidently below the mean.

Each system has its overall proportion of Good across all phenomenon-equivalent pairs indicated

[4] As calculated by http://matrix.statmt.org/matrix/systems_list/1883

[5] http://ufal.mff.cuni.cz/czeng

System	Ave. z	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
CUNI Transformer	0.594	26.6	26.0	0.638	0.567	0.532
UEDIN	0.384	24.0	23.4	0.666	0.554	0.550
CUNI Chimera noDepFix	–	21.0	19.8	0.703	0.528	0.600
CUNI Chimera	–	20.8	19.2	0.704	0.522	0.605
online-B	0.101	20.0	19.4	0.710	0.523	0.597
CUNI Moses	–	17.5	16.4	0.739	0.509	0.632
online-A	−0.115	16.7	15.7	0.75	0.507	0.619
online-G	−0.246	16.2	15.1	0.770	0.503	0.631

Table 1: Manual and automatic results of WMT18 English-Czech systems. Sorted according to BLEU, which correlates with the manual evaluation "Ave. z" where available, but which may overestimate the quality of especially phrase-based systems like CUNI* ones.

		Total	Bad	Good	Unk	Total	Bad	Good	Unk
EN-control-CS-finclause	Reference	76	7.9	72.4	19.7	100	8.0	71.0	21.0
EN-control-CS-finclause	UEDIN	76	34.2	53.9	11.8	100	39.0	50.0	11.0
EN-control-CS-finclause	CUNI Chimera	76	≀ 27.6	51.3	21.1	100	≀ 31.0	47.0	22.0
EN-control-CS-finclause	CUNI Chimera noDepFix	76	27.6	51.3	21.1	100	31.0	47.0	22.0
EN-control-CS-finclause	CUNI Transformer	76	≀ 23.7	≀ 56.6	19.7	100	≀ 23.0	≀ 60.0	17.0
EN-control-CS-finclause	online-B	76	23.7	≀ 59.2	17.1	100	27.0	53.0	20.0
EN-control-CS-finclause	online-A	76	68.4	22.4	9.2	100	71.0	20.0	9.0
EN-control-CS-finclause	online-G	76	71.1	18.4	10.5	100	≀ 69.0	≀ 21.0	10.0
EN-control-CS-finclause	CUNI Moses	76	≀ 67.1	17.1	15.8	100	≀ 67.0	18.0	15.0
EN-control-CS-nofinclause	Reference	104	0.0	70.2	29.8	1819	0.6	74.2	25.2
EN-control-CS-nofinclause	UEDIN	104	20.2	64.4	15.4	1819	18.0	68.4	13.6
EN-control-CS-nofinclause	CUNI Chimera	104	23.1	60.6	16.3	1819	20.2	62.3	17.4
EN-control-CS-nofinclause	CUNI Chimera noDepFix	104	23.1	60.6	16.3	1819	20.2	≀ 62.4	17.4
EN-control-CS-nofinclause	CUNI Transformer	104	26.9	54.8	18.3	1819	≀ 18.2	≀ 64.9	16.9
EN-control-CS-nofinclause	online-B	104	28.8	52.9	18.3	1819	18.6	61.2	20.2
EN-control-CS-nofinclause	online-A	104	≀ 10.6	≀ 77.9	11.5	1819	≀ 8.7	≀ 79.2	12.0
EN-control-CS-nofinclause	online-G	104	11.5	76.0	12.5	1819	≀ 7.3	≀ 81.3	11.4
EN-control-CS-nofinclause	CUNI Moses	104	≀ 5.8	≀ 80.8	13.5	1819	≀ 5.3	79.2	15.6
EN-control-CS-subjunctclause	Reference	90	2.2	65.6	32.2	1130	4.0	70.7	25.3
EN-control-CS-subjunctclause	UEDIN	90	23.3	≀ 66.7	10.0	1130	22.4	60.7	16.9
EN-control-CS-subjunctclause	CUNI Chimera	90	≀ 21.1	58.9	20.0	1130	29.6	47.6	22.7
EN-control-CS-subjunctclause	CUNI Chimera noDepFix	90	21.1	58.9	20.0	1130	29.6	47.6	22.7
EN-control-CS-subjunctclause	CUNI Transformer	90	≀ 18.9	≀ 61.1	20.0	1130	≀ 21.7	≀ 57.5	20.8
EN-control-CS-subjunctclause	online-B	90	20.0	58.9	21.1	1130	25.3	52.7	21.9
EN-control-CS-subjunctclause	online-A	90	50.0	36.7	13.3	1130	52.8	30.4	16.7
EN-control-CS-subjunctclause	online-G	90	57.8	30.0	12.2	1130	64.6	20.4	15.0
EN-control-CS-subjunctclause	CUNI Moses	90	63.3	16.7	20.0	1130	≀ 61.6	17.1	21.3
EN-gerund-CS-finclause	Reference	75	25.3	56.0	18.7	165	21.2	59.4	19.4
EN-gerund-CS-finclause	UEDIN	75	26.7	≀ 58.7	14.7	165	28.5	57.0	14.5
EN-gerund-CS-finclause	CUNI Chimera	75	≀ 24.0	≀ 60.0	16.0	165	≀ 24.2	56.4	19.4
EN-gerund-CS-finclause	CUNI Chimera noDepFix	75	24.0	60.0	16.0	165	24.2	56.4	19.4
EN-gerund-CS-finclause	CUNI Transformer	75	28.0	50.7	21.3	165	26.7	51.5	21.8
EN-gerund-CS-finclause	online-B	75	34.7	48.0	17.3	165	32.7	47.3	20.0
EN-gerund-CS-finclause	online-A	75	60.0	26.7	13.3	165	57.6	32.7	9.7
EN-gerund-CS-finclause	online-G	75	61.3	≀ 28.0	10.7	165	63.6	27.9	8.5
EN-gerund-CS-finclause	CUNI Moses	75	≀ 38.7	≀ 45.3	16.0	165	≀ 36.4	≀ 48.5	15.2
EN-gerund-CS-nofinclause	Reference	218	0.5	72.9	26.6	368	2.2	70.1	27.7
EN-gerund-CS-nofinclause	UEDIN	218	16.1	67.4	16.5	368	19.8	64.9	15.2
EN-gerund-CS-nofinclause	CUNI Chimera	218	29.4	53.7	17.0	368	28.3	53.8	17.9
EN-gerund-CS-nofinclause	CUNI Chimera noDepFix	218	29.4	53.7	17.0	368	28.3	53.8	17.9
EN-gerund-CS-nofinclause	CUNI Transformer	218	≀ 17.0	≀ 62.8	20.2	368	≀ 16.6	≀ 63.0	20.4
EN-gerund-CS-nofinclause	online-B	218	19.3	59.6	21.1	368	20.7	58.2	21.2
EN-gerund-CS-nofinclause	online-A	218	≀ 12.8	≀ 75.7	11.5	368	≀ 14.7	≀ 72.8	12.5
EN-gerund-CS-nofinclause	online-G	218	17.4	71.6	11.0	368	16.3	71.5	12.2
EN-gerund-CS-nofinclause	CUNI Moses	218	33.5	50.5	16.1	368	32.1	51.4	16.6

Table 2: Detailed results of our automatic tests. Left: Manually refined set, Right: larger, pre-selected set. Systems sorted by average performance in our testsuite. "≀" indicates lines out of sequence in the "Bad" or "Good" columns.

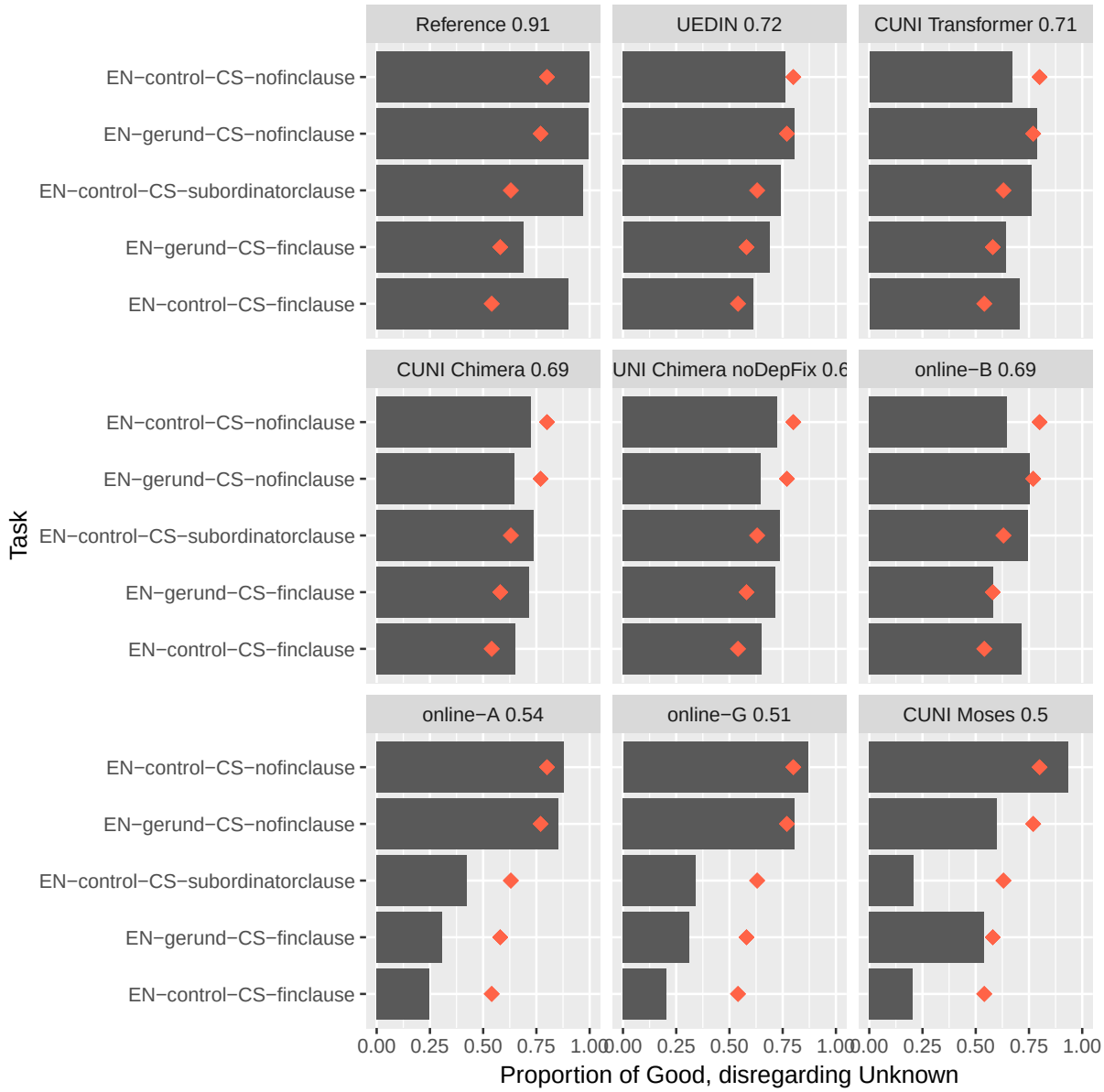

Figure 2: Proportion of Good in different linguistic phenomena broken down by systems, measured on the refined set. Each subgraph represents one MT system. They are ordered top-down according to the average proportion of sentences that each of them judged as Good for the given *En-phenomenon / Cz-equivalent* pair. The proportion is computed from Good + Bad, disregarding Unknown. The orange diamonds show the average occurrence of the Good instances in the given phenomenon-equivalent pair across all systems; that is, the proportion of Good in this phenomenon-equivalent pair captured by a fictitious average system.

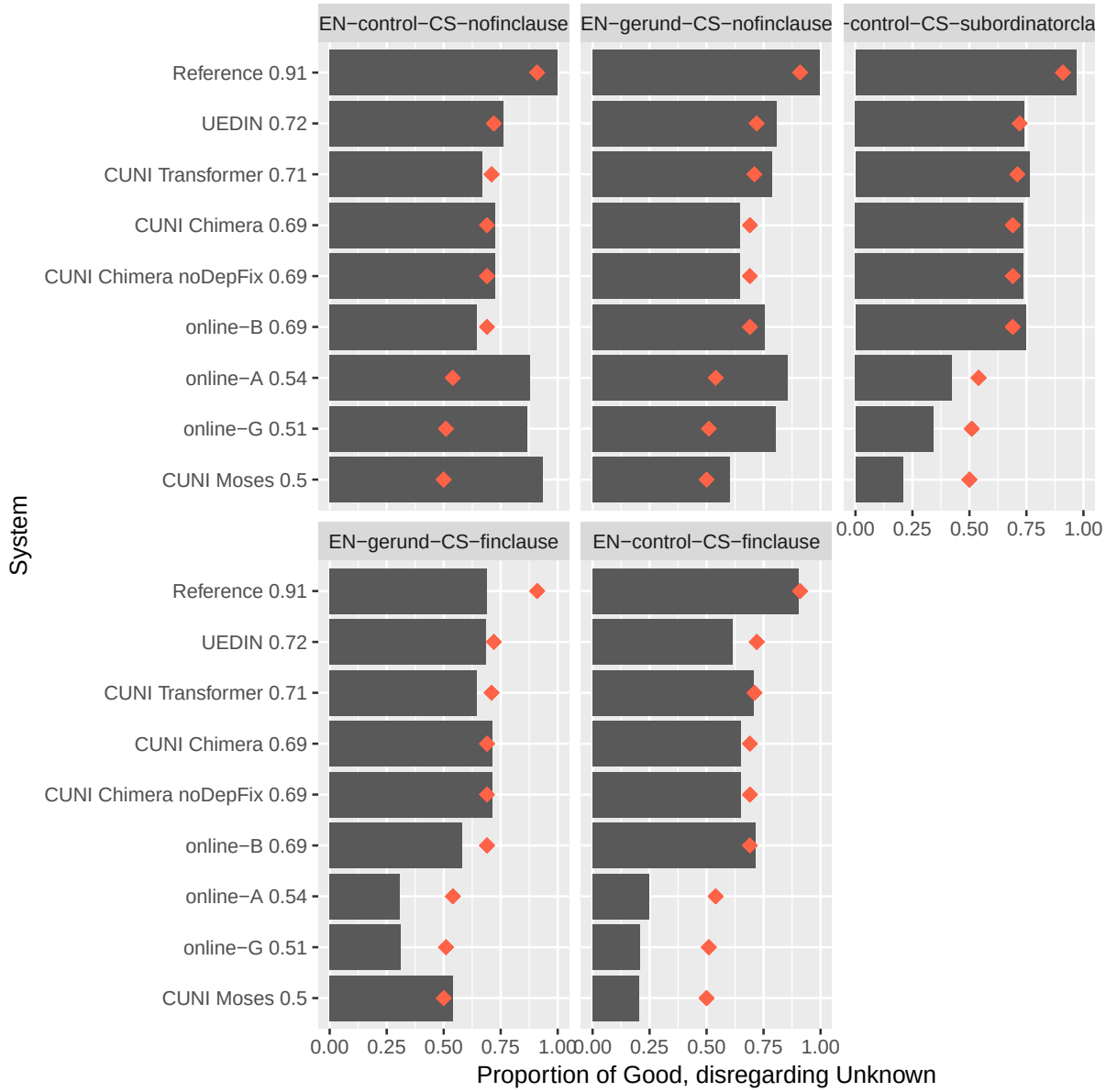

Figure 3: Proportion of Good in the systems broken down by linguistic phenomena, measured on the refined set. Each subgraph represents one pair of *En phenomenon – Cs translation* option. Each bar represents one MT system. The orange diamonds show the average performance of the given system across all phenomenon-equivalent pairs.

in the stripe with its name (e.g. UEDIN 0.72 means that UEDIN has 72% of Good across the five phenomenon-equivalent pairs). The systems are ordered top-down according to this proportion.

The Reference subgraph suggests how well the automatic equivalent detection worked. If the phenomenon-equivalent pairs had been mapped manually, all bars would have been at 100%, since the sentences had been selected manually, but the automatic detection of phenomenon-equivalent pairs had not worked perfectly. Therefore we should not compare the proportion of Good in a system with 1 (or 100%) but with the proportion reached by Reference. On the other hand, most of the missed instances were classified as Unknown rather than Bad (hence the outstanding proportion of "Unk" associated with Reference in Table 2), so we can estimate the performance of the automatic detection of phenomenon-equivalent pairs. However, this estimation is approximate, as we do not know whether the detection of the phenomenon-equivalent pairs worked equally well for all systems.

Caution is advised considering En-gerund-CS-finclause, of which only about 70% had been captured even in Reference, and, to somewhat lesser extent considering En-control-CS-finclause (about 87%).

Figure 3 presents the matter from the opposite perspective: each subgraph renders one phenomenon-equivalent pair and each bar one system. The orange diamonds indicate for each system how well it had captured the given phenomenon-equivalent pair compared to its average performance in capturing these pairs. For instance, we can see that, in En-gerund-CS-finclause, Reference really had a problem, lying approx. 20% below its mean proportion of Good across all phenomenon-equivalent pairs.

Having analyzed the performance of the individual systems in translating specific English phenomena with their expected Czech equivalents, we have to consider the overall translation quality of the systems (Table 1). The most important observation is that the order of MT systems in Figures 2 and 3 does roughly match their overall performance for the CUNI Transformer and UEDIN, as well as the Chimeras and online-B. We only have a difference in the order of CUNI Moses vs. online-A and online-G. CUNI Moses turned out the worst detecting the phenomenon-equivalent pairs, but had a higher BLEU score than online-A as well as online-G.

We have measured Spearman's rank correlation ρ between the overall translation performance estimated by BLEU and the phenomenon-equivalent detection and observed a statistically significant correlation (p-value 0.007) at 0.88. We have used the RVaideMemoire R package (Hervé, 2018)) to obtain the 95% confidence interval, which is 0.34-1.00. Given the coarseness of the comparison (overall performance instead of performance on the individual sentences and proportions of Good instead of agreement of the equivalent structure with the Reference on individual sentences), we consider this a promising result that encourages further research.

As a next step, we plan to obtain and compare manual evaluations of individual sentences. The annotators will rate the automatic and reference translations alike, without knowing which is which. We will investigate whether there is a relationship between the equivalent choice regarding the selected linguistic phenomena and the translation quality rating. We will observe the agreement of the structure of Czech equivalents with Reference on individual sentences rather than merely comparing their proportions of Good.

6 Conclusion

We have presented a testsuite of about 3000 automatically pre-selected sentences focused on English–Czech translation of a small set of extremely frequent verb-related phenomena.

Targeted automatic checks on whether the given English grammatical phenomenon is translated as expected generally correlate with whole-sentence measures like BLEU. At the same time, interesting differences between individual MT systems are observed in their handling of the phenomena. For instance, CUNI Transformer's overall performance is deemed better but it departs more often from the most canonic translation of the examined phenomena compared to both the reference translation and UEDIN.

Further investigation and targeted manual evaluation are needed to validate whether the less expected translations for our selected phenomena indeed constitute a better translation quality.

The dataset is publicly accessible via the LINDAT-CLARIN repository:

http://hdl.handle.net/11234/1-2856

Acknowledgments

This work has been supported by the Czech Science Foundation grant 18-24210S and by the Czech Ministry of Education, Youth, and Sports grant LTC18020. We have used the data, tools, and infrastructure of the LINDAT-CLARIN repository (https://lindat.mff.cuni.cz/en/), Research Infrastructure CZ.02.1.01/0.0/0.0/16_013/0001781.

References

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia, Bulgaria. Association for Computational Linguistics.

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*, Brussels, Belgium. Association for Computational Linguistics.

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 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Jindřich Helcl, Jindřich Libovický, Tom Kocmi, Tomáš Musil, Ondřej Cífka, Dušan Variš, and Ondřej Bojar. 2018. Neural monkey: The current state and beyond. In *The 13th Conference of The Association for Machine Translation in the Americas, Vol. 1: MT Researchers' Track*, pages 168–176, Stroudsburg, PA, USA. The Association for Machine Translation in the Americas, The Association for Machine Translation in the Americas.

Maxime Hervé. 2018. *RVAideMemoire: Testing and Plotting Procedures for Biostatistics*. R package version 0.9-69-3.

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 *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics.

Martin Popel. 2018. CUNI Transformer Neural MT System for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural mt systems for wmt17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Jan Štěpánek and Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*,

pages 1828–1835, Valletta, Malta. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA.