Findings of the 2022 Conference on Machine Translation (WMT22)

Tom Kocmi
Microsoft

Rachel Bawden
Inria, Paris

Ondřej Bojar
Charles University

Anton Dvorkovich
Neurodub

Christian Federmann
Microsoft

Mark Fishel
University of Tartu

Thamme Gowda
Microsoft

Yvette Graham
Trinity College Dublin

Roman Grundkiewicz
Microsoft

Barry Haddow
University of Edinburgh

Rebecca Knowles
NRC

Philipp Koehn
Johns Hopkins University

Christof Monz
University of Amsterdam

Makoto Morishita
NTT

Masaaki Nagata
NTT

Toshiaki Nakazawa
University of Tokyo

Michal Novák
Charles University

Martin Popel
Charles University

Maja Popović
Dublin City University

Mariya Shmatova
Neurodub

Abstract

This paper presents the results of the General Machine Translation Task organised as part of the Conference on Machine Translation (WMT) 2022. In the general MT task, participants were asked to build machine translation systems for any of 11 language pairs, to be evaluated on test sets consisting of four different domains. We evaluate system outputs with human annotators using two different techniques: reference-based direct assessment and (DA) and a combination of DA and scalar quality metric (DA+SQM).

1 Introduction

The Seventh Conference on Machine Translation (WMT22)[1] was held online with EMNLP 2022 and hosted a number of shared tasks on various aspects of machine translation. This conference built on 15 previous editions of WMT as workshops and conferences (Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021).

For more than a decade, the machine translation (MT) community has focused on the news domain, which has many desirable features for MT evaluation, such as sufficiently long and grammatically correct sentences that are easy for both professionals to translate (to produce references) and for human raters to evaluate without specific in-domain knowledge. However, with recent advances in MT and potential overfitting on the news domain (with methods such as fine-tuning on past WMT testsets), we decided to open a fresh research direction of testing the "General Machine Translation" capabilities.

How to test general MT capabilities is a research question in itself. Countless phenomena could be evaluated, the most important being:

- various domains (news, medicine, IT, patents, legal, social, gaming, etc.)
- style of text (formal or spoken language, fiction, technical reports, etc.)
- noisy or robust user-generated content (grammatical errors, code-switching, abbreviations, etc.)

Evaluating all possible phenomena is near impossible and creates many unforeseen problems. Therefore, we decided to simplify the problem and start with an evaluation of different domains. We select the following four domains: news, e-commerce, social, and conversational, chosen to represent various topics with different content

[1] http://www.statmt.org/wmt22/

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1–45

styles. Additionally, these domains are understandable for humans without special in-domain knowledge, thus not requiring specialized translators or human raters for evaluation.

Another significant change for this year is the redesign of our human evaluation procedure for English→X and non-English language pairs. We introduce SQM-style DA rating, improved sampling of sentences for human judgements, and we opt in for using professional raters.

In addition to language pairs evaluated yearly, we introduce several new language pairs that have never been evaluated at WMT or other venues: Ukrainian↔English, Ukrainian↔Czech, Livonian↔English, Yakut↔Russian and English→Croatian.

Lastly, with multiple different shared tasks run at WMT evaluating different phenomena over the same language pairs, we proposed to aggregate test sets and ask participants of different shared tasks to also translate test sets from other shared tasks (for shared language pairs), allowing cross-task evaluation of systems on various phenomena. More details are in Section 4.2.

General MT task submissions and human judgements are available at `https://github.com/wmt-conference/wmt22-news-systems`. The interactive visualization and comparison of differences between systems is at `http://wmt.ufal.cz` using MT-ComparEval (Sudarikov et al., 2016).

The structure of the findings is as follows. We describe process of collecting, cleaning and translating of test sets in Section 2 followed by summary of allowed training data for constrained track Section 3. We list all submitted systems in Section 4. We use two different techniques for human evaluation. Reference-based DA is used to evaluate languages into English and described in Section 5. DA+SQM technique used for non-English and from English translation directions is described in Section 6. In Section 7, we describe our analysis of English→Croatian, translation direction containing professional and student produced references. We conclude the findings in Section 8.

2 Test Data

In this section, we describe the process of collecting data in Section 2.1, followed by the explanation of preprocessing steps in Section 2.2. Producing human references is summarized in Section 2.3 and test set analysis is conducted in Section 2.5. Lastly, Section 2.4 describes specific language pairs that are prepared differently.

2.1 Collecting test data

As in the news shared tasks in previous years, the test sets consist of unseen translations prepared specially for the task. However, in contrast, we introduce several domains instead of only the news domain. The test sets are publicly released to be used as translation benchmarks. Here we describe the production and composition of the test sets.

With the new direction towards testing general MT capabilities, we redesign the content of the test sets. We decided to collect data from four domains (news, social, e-commerce, and conversation). For all language pairs, we aimed for a test set size of 2000 sentences and to ensure that the test sets were "source-original", namely that the source text is written in the source language, and the target text is the human translation. This is to avoid "translationese" effects on the source language, which can have a detrimental impact on the accuracy of evaluation (Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020). We collected roughly the same number of sentences (around 500 sentences with document context) for each domain. For some languages, we could not locate high-quality data and therefore selected more sentences from other domains.

News domain - This domain contains data prepared in the same way as in previous years (Akhbardeh et al., 2021). We collect news articles from the second half of 2021 extracted from online news sites, keeping document information. The news domain is mainly of the highest quality.

Social domain - For most languages (Czech, English, French, German, and Japanese), we extract data from public Reddit discussions, keeping separate posts as a single document. We target subreddits that come from countries speaking a given language. We remove all posts marked by Reddit as inappropriate.

We use different data source for Chinese and Russian social domain as there is not enough Reddit content. For Chinese, we collected posts from various social media webpages used in China, a list provided by our Chinese colleague. For Russian, we took data from Zen, one of the most popular blog platforms among Russian-speaking users.

E-commerce domain - Contains product descriptions donated by individual companies.

Source / Domain	#segments conversation	ecommerce	news	social	other	total
Chinese	349	518	505	503	-	1875
Czech	-	-	957	491	-	1448
Czech (to Ukrainian)	-	-	-	-	1930	1930
English	484	530	511	512	-	2037
English (to Croatian)	-	1015	656	-	-	1671
French	501	524	504	477	-	2006
German	462	501	506	515	-	1984
Japanese	502	503	505	498	-	2008
Livonian	-	-	-	-	420	420
Russian	-	508	1004	-	504	2016
Russian (to Yakut)	-	-	-	-	1123.0	1123
Ukrainian	-	-	-	-	2018	2018
Ukrainian (to Czech)	-	-	-	-	2812	2812
Yakut	-	-	-	-	1123	1123

Table 1: Number of segments for individual source languages used in the general translation test sets.

For Japanese e-commerce domain, we used search advertising text ads provided by an advertising company with their client's prior consent. Defining documents and sentences in search ads is tricky. Clients define multiple titles and multiple descriptions, called assets. We defined a document as the longest possible combination of assets. We also defined a sentence as either an asset or a unit separated by sentence-ending punctuations within an asset. Since the diversity of Japanese ad sentences is small, we chose the test sentences greedily to minimize the test set's self-BLEU.

Conversational domain - data for English, German, French, and Chinese are provided by the Chat Shared Task organizers (Farinha et al., 2022). These data contain a discussion between an agent, talking in English, and a customer, each of them talking in a different language. To avoid the effects of translationese, we split conversations into individual messages and handled each as a separate document, only using messages written in the original language (therefore, the English side only contains messages from agents) resulting in often short documents.

For Japanese conversational domain, We used question-answer pairs from a community question-answering website, *Oshiete!goo*[2]. The operator provided us with a dump as of March 2022. Topics are diverse, ranging from life advice to entertainment. Since there were usually many answers to a question, we extracted question-answer pairs whose answers were marked as the best answer. We considered a question-answer pair as a document and randomly sampled test data from question-answer pairs with a total length of 180 characters or fewer. We did not indicate the boundary between them.

After collecting all data, we applied several steps to filter out documents of lower-quality, see Section 2.2. Specifically paying attention to short documents. Whenever we had enough data, we removed the shortest documents, usually a single or two sentences. We advised linguists who were checking the data to further remove short documents. This helped us to add document context to the test set.

2.2 Human preprocessing of test data

In the News task of previous years, we asked humans to check collected data and carry out minor corrections (mainly checking sentence splits and discarding similar or repeated content), which was sufficient for the news domain because is often clean and without serious problems. However, with the expansion towards general MT, we run into an issue of source data being noisier and not well formatted that needs to be handled before translation.

Although testing of robustness of MT is an important task, the noisy data introduces problems for human translators and annotators. Therefore, we decided to discard data that are considered too noisy. Furthermore, publicly available data often contains inappropriate content, which can stress either human translators or human annotators, leading to a decrease in the quality (for example, translators refuse to translate political content considered censored in their countries).

Therefore, the source data for test sets[3] goes

[2] https://oshiete.goo.ne.jp

[3] Except for sources from the following translation directions: English→Croatian, Livonian↔English, Yakut↔Russian, Ukrainian→English, Ukrainian↔Czech. Data for these directions have been checked differently and should not contain noisy or inappropriate content.

through human validation checks involving linguists discarding inappropriate content altogether or carrying out minor textual corrections to the data. You can find the linguistic brief for prepossessing in Appendix C.

2.3 Test set translation

The translation of the test sets was performed by professional translation agencies, according to the brief in Appendix D. Different partners sponsored each language pair and various translation agencies were therefore used, which may affect the quality of the translation. The exception is that Chinese↔English, German↔English, Ukrainian↔English and reference-B for Czech↔English were translated by the same agency. These languages also received a special treatment of being translated by one translator and checked by a second different translator.

Several language pairs received special attention. For Chinese↔English, Czech↔English, German↔English, and English→Croatian, we obtained a second reference in each direction from different translators.

For Czech↔English, our partner paid professional agency to provide high-quality translations. However, as it turned out, the quality is rather low. We fixed manually the reference with grammar correction tools, however, that isn't sufficient. We provide this reference as reference–C. There is no issue with reference-B as that was provided by different partner.

Human translations would not be possible without the sponsorship of our partners. We are thankful for the support from: Microsoft, Charles University, LinguaCustodia, NTT, Dublin City University, Google, and Phrase.

2.4 Language pairs prepared differently

English→Croatian The English-Croatian test data is a sub-corpus of the DiHuTra corpus[4] (Lapshinova-Koltunski et al., 2022). The English source texts include Amazon product reviews and news articles. The document information is available for both domains.

The *reviews* were selected from the publicly available **Amazon product reviews**[5,6] containing reviews divided into 24 categories (topics). The selected corpus covers fourteen categories, paying attention to the data balance: an equal number of positive and negative reviews and a balanced distribution of categories (topics). In total, 196 reviews (1015 sentences) were included, fourteen from each of the fourteen selected topics: 'Beauty', 'Books', 'CDs and Vinyl', 'Cell Phones and Accessories', 'Grocery and Gourmet Food', 'Health and Personal Care', 'Home and Kitchen', 'Movies and TV', 'Musical Instruments', 'Patio, Lawn and Garden', 'Pet Supplies', 'Sports and Outdoors', 'Toys and Games' and 'Video Games'.

The *news articles* were selected from the News test corpus of the WMT (2019 and 2020) shared task.[7] In total, 68 news articles (656 sentences) from different sources are included.

These English texts were then translated into Croatian by professional translators and by translation students, thus providing two reference translations. Both professional and student translations were produced in cooperation with the University of Zagreb and the University of Rijeka in Croatia. In total, four professional translators and twenty translation students participated, all native speakers of Croatian and fluent in English. Translation experience of professional translators ranges between five and ten years, while for students the range is from zero to five years, the majority being in the range between two and four years. The two students who indicated no experience (zero years) also indicated that they had no real professional experience yet, only work in the framework of their studies. All students were in their first or second year of master's studies.

The translators were asked too keep the sentence (segment) alignment (not to merge or to split segments so that each English segment corresponds to one translated segment) and not to use any kind of machine translation in the process. No further restrictions were given to the translators.

Yakut↔Russian Source texts for Yakut↔Russian translation were selected from Ulus media, which is Yakutia's official news aggregator. The majority of the data are local news.[8] The professional translators were asked to translate 42 news texts for the test set. Yakut is one of the minor languages spoken by around

[4] https://github.com/katjakaterina/dihutra

[5] https://s3.amazonaws.com/amazon-reviews-pds/readme.html

[6] http://jmcauley.ucsd.edu/data/amazon/

[7] http://www.statmt.org/wmt20/translation-task.html

[8] https://ulus.media/

450,000 native speakers. It is one of the official languages of Sakha (Yakutia), a federal republic in the Russian Federation.

Livonian↔English The source language for Livonian↔English was English, since the amount of Livonian monolingual and parallel data is severely limited. The source texts were selected from various news articles published in 2022; politically neutral topics were selected. One addition to the set was the text describing the addition of WMT'22 Livonian↔English shared task itself. Translations were done by two professionals. Livonian is a critically endangered language spoken in Latvia but belonging to the Finno-ugric language family. Its last native speaker passed in 2013 and currently there are about 20 near-native speakers; however, there is an Institute of the Livonian Language at the University of Latvia that leads efforts on collecting and preserving Livonian texts as well as other materials (audio, video, hand-written, etc).

Ukrainian↔Czech and Ukrainian→English Source texts for Ukrainian↔Czech and Ukrainian→English translation were selected from the inputs collected through the Charles Translator for Ukraine.[9] Charles Translator for Ukraine is an online translation service that has been developed by the team from the Charles University, Prague[10] as a response to the wave of Ukrainian refugees coming to the Czech Republic after the 2022 Russian invasion of Ukraine.[11] The service is powered by a model trained with Block Backtranslation (Popel et al., 2020b). With users' consent, the service can log their inputs for the purpose of creating a dataset of real use cases. The datasets are extracted from the inputs collected in March and April 2022.

After automatic filtering,[12] we asked linguistically-educated annotators to filter and preprocess the source data manually. The filtering aimed at obtaining a data sample with diverse examples. The preprocessing was performed according to the brief in Appendix C with the following modifications. First, as the content is closely related to the war, someone may always find it polarizing or controversial. We did not filter out texts based on this criterion. Second, we asked the annotators not to delete or fix noisy inputs as long as they are comprehensible. This concerns, for instance, errors in casing, punctuation, diacritics, grammar and typos. Furthermore, all emojis are kept. Third, our annotators were instructed to join multiple related sentences to the same line whenever they found them too short compared to the rest of the dataset. The dataset thus does not satisfy the rule that each line contains a single sentence. Finally, any personal data related to people other than well-known people was pseudonymized.

The user inputs cover three broader domains: (1) personal communication, (2) news, and (3) formal communication. Our annotators assigned these categories (often accompanied by a finer subcategory) to every data example. If none of the above categories fit, they labeled the example with the "other" tag.

The source texts were translated by professional translation agencies principally following the brief in Appendix D. A sample of translated sentences were checked by native speakers of the target language. It revealed that post-edited MT had allegedly been used for parts of the Ukrainian→Czech test set, although this was denied by the translator. Therefore, we decided to add additional data to the test set for this direction translated by a different translation agency. This extra data consists of about 600 segments downloaded from the web (news, example CV) and about 200 segments from the Charles Translator inputs logs. It was pre-processed similarly as described above except for the domain annotation (all segments have the "unknown" tag assigned).

2.5 Test set analysis

As described previously, the aim was for the test sets to be composed of approximately 500 sentences per domain, although this depended on the language pair. The number of segments for each domain (including unspecified domain 'other') is given in Table 1 per source language, with the target language being specified where the composition differs. All four domains are available for Chinese, English, French, German and Japanese source texts, whereas only certain domains are available

[9] http://translate.cuni.cz

[10] http://ufal.mff.cuni.cz/u4u

[11] At that time, the most popular online MT services either did not support translation between Czech and Ukrainian (e.g. DeepL) or they seemed to pivot the translation for the language pair via English (e.g. Google Translate, Microsoft Translator).

[12] This includes the removal of intermediate inputs, HTML-tagged inputs, inputs identified as written in a language other than the source language, and backtranslated inputs.

for Czech, Russian and English into Croatian.

Document context Document context is available for most language pairs (the exception being Livonian↔English). The length of documents varies considerably by domain but also by language pair. As can be see in Table 2, e-commerce documents tend to be longest, followed by news and social (together), with conversational documents being shortest, although this does not hold for all languages. For example, the Ukrainian test set has short documents (2.28 segments on average), whereas Yakut↔Russian has very long ones (26.12 segments on average).

Lexical diversity We can compare the type-token ratio (TTR) to get an idea of the relative lexical diversity of (i) domains and (ii) original vs. translated sentences.[13,14] Raw TTRs for each language pair and domain are given in Table 28 in Appendix E. Regarding domains, the TTR is generally lowest for conversations, whereas e-commerce and news are most diverse, followed by social. Translated texts appear to show a lower lexical diversity than original texts. If we look at the ratio between the TTRs of a language A and a language B (i.e. the diversity of A with respect to B), this ratio is higher when A is the source and B the target than when B is the source and A the target. For example, given the language pair Czech↔English, the ratio of the TTRs of Czech and English (i.e. $\frac{\text{TTR}_{\text{cs}}}{\text{TTR}_{\text{en}}}$ is higher when Czech is the original text and lower when it is the translation. This can be seen in Table 3 comparing for individual domains.

Anonymisation One characteristic that stands out is the presence of placeholders for anonymised elements in the conversation and social domains. There are a total of 17 difference placeholders, indicated by the entity type surrounded by #, e.g. #NAME#, #EMAIL#, #Product1#, #Product2#, etc. The entities are identical in the reference translation (where there is a direct translation), rather than the entity being translated (e.g. #NAME# and not #NOM# for French). Manual corrections were carried out to homogenise variants in terms of capitals, space issues and placeholders that were translated rather than copied by the professional translators.

Translation quality As mentioned previously, the quality of the human references differed according to the agency used. A few translations were erroneous due to problems with anonymisation, where some overzealous anonymisation added entity tags within non-entity words, therefore making the source sentence non-sensical. However this affected only one or two sentences, and some minor corrections were introduced. There were some particular problems with the quality of Czech→English translations, including wrong quote marks, grammatical and spelling mistakes and unnatural translations as mentioned in Section 2.3.

3 Training Data

Similar to previous years, we provide a selection of parallel and monolingual corpora for model training. The provenance and statistics of the selected parallel datasets are provided in Appendix in Table 20 and Table 19. Specifically, our parallel data selection include large multilingual corpora such as Europarl-v10 (Koehn, 2005), Paracrawl-v9 (Bañón et al., 2020), CommonCrawl, NewsCommentary-v16, WikiTitles-v3, WikiMatrix (Schwenk et al., 2021), TildeCorpus (Rozis and Skadiņš, 2017), OPUS (Tiedemann, 2012), UN Parallel Corpus (Ziemski et al., 2016), and language specific corpora such as CzEng-v2.0 (Kocmi et al., 2020), YandexCorpus,[15] ELRC EU Acts, YakutCorpus [16], JParaCrawl (Morishita et al., 2020), Japanese-English Subtitle Corpus (Pryzant et al., 2018), Livonian multiparallel corpus Liv4ever (Rikters et al., 2022),[17] KFTT(Neubig, 2011), TED (Cettolo et al., 2012), CCMT, and back-translated news. Similar to previous years, we provided links to these datasets on the task web page.[18] However, new to this year, we automate the data preparation pipeline using a tool named MTDATA (Gowda et al., 2021).[19] MTDATA downloads all available datasets, except

[13]The TTR is the ratio of unique tokens to total tokens, and it is higher the diverse the vocabulary of a text is. It is dependent on the morphological complexity of a language, but can also vary due to other factors.

[14]Texts are tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) where available. For Czech, Livonian and Yakut, for which Spacy models are not available, we took as a rough approximation models for Croatian, Finnish and Russian respectively.

[15]https://github.com/mashashma/WMT2022-data

[16]https://github.com/mashashma/WMT2022-data/tree/main/yakut

[17]https://huggingface.co/datasets/tartuNLP/liv4ever

[18]https://statmt.org/wmt22/translation-task.html

[19]https://statmt.org/wmt22/mtdata

Source / Domain	#segments per doc conversation	ecommerce	news	social	other	all
Chinese	2.13	17.86	13.29	20.12	-	7.32
Czech	-	-	14.07	7.12	-	10.57
Czech (to Ukrainian)	-	-	-	-	1.86	1.86
French	2.61	22.78	14.00	14.03	-	7.04
German	2.87	17.28	11.50	13.92	-	7.32
English	5.20	23.04	16.48	15.06	-	11.25
English (to Croatian)	-	5.18	9.65	-	-	6.33
Japanese	4.40	4.49	15.30	8.03	-	6.26
Livonian	-	-	-	-	1.00	1.00
Russian	-	10.58	12.55	-	5.09	8.88
Russian (to Yakut)	-	-	-	-	26.12	26.12
Ukrainian	-	-	-	-	2.28	2.28
Ukrainian (to Czech)	-	-	-	-	2.98	2.98
Yakut (to Russian)	-	-	-	-	26.12	26.12

Table 2: Average document length (in # segments) for individual source languages used in the general translation test sets.

Lang. pair	conversation →	conversation ←	ecommerce →	ecommerce ←	news →	news ←	social →	social ←	other →	other ←
Czech–English	-	-	1.9	1.58	1.73	1.57	-			
Czech–Ukrainian	-	-	-	-	1.06	0.93				
German–English	1.39	1.00	1.50	1.13	1.35	1.15	1.38	1.13	-	
German–French	1.25	0.95	1.50	1.15	1.35	1.15	1.26	1.08	-	
English–Czech	-	-	0.63	0.52	0.64	0.58	-			
English–German	1.00	0.72	0.89	0.67	0.87	0.74	0.88	0.72	-	
English–Japanese	1.50	1.00	1.41	1.20	1.44	1.13	1.28	1.00	-	
English–Livonian	-	-	-	-	0.74	0.74				
English–Russian	-	0.69	0.59	0.67	0.57	-	-			
English–Chinese	1.15	0.71	1.09	0.70	1.00	0.68	0.92	0.74	-	
French–German	1.06	0.80	0.87	0.67	0.87	0.74	0.93	0.79	-	
Japanese–English	1.00	0.67	0.83	0.71	0.88	0.69	1.00	0.78	-	
Livonian–English	-	-	-	-	1.36	1.36				
Russian–English	-	1.69	1.46	1.75	1.50	-	-			
Russian–Yakut	-	-	-	-	0.89	0.89				
Yakut–Russian	-	-	-	-	1.12	1.12				
Ukrainian–Czech	-	-	-	-	1.08	0.94				
Chinese–English	1.41	0.87	1.43	0.92	1.47	1.00	1.35	1.09	-	

Table 3: For each language pair A–B, the ratio of the TTRs of A and B, for the A→B test set (→; i.e. A is the original text) and for the B→A test set (←, i.e. A is the translated text).

the two which required user authentication: CCMT and CzEng-v2.0.

4 System submissions

In 2022, we received a total of 107 primary submissions[20] and 82 online systems. The participating institutions are listed in Table 4 and detailed in the rest of this section. Each system did not necessarily appear in all translation tasks. We also included online MT systems (originating from 5 services), which we anonymized as ONLINE-A,B,G,W,Y. All submissions, sources and references are made available via github.

For presentation of the results, systems are treated as either constrained or unconstrained. When the system submitters report that they were only trained on our provided data, we class them as constrained. The online systems are treated as unconstrained during the automatic and human evaluations, since we do not know how they were built. In Appendix F, we provide brief details of the submitted systems, for those where the authors provided such details.

4.1 OCELoT

To collect submissions, we used the open-source OCELoT platform[21] again, which provides anonymized public leaderboards for several WMT22 shared tasks.[22] Similarly to the setup from the previous year, only registered and verified teams with correct contact information were allowed to submit their system outputs and each

[20] GTCOM was removed from human evaluation, however, we calculate automatic scores in Appendix G.

[21] https://github.com/AppraiseDev/OCELoT

[22] https://ocelot-wmt22.mteval.org

Team	Language Pairs	System Description
AISP-SJTU	en-ja, en-zh, ja-en, zh-en	Liu et al. (2022)
AIST	ja-en	(no associated paper)
ALMANACH-INRIA	cs-en, cs-uk, ru-en, uk-cs, uk-en	Alabi et al. (2022)
AMU	cs-uk, uk-cs	Nowakowski et al. (2022)
ARC-NKUA	en-uk, uk-en	Roussis and Papavassiliou (2022)
CUNI-BERGAMOT	en-cs	Jon et al. (2022)
CUNI-DOCTRANSFORMER	cs-en, en-cs	Jon et al. (2022)
CUNI-TRANSFORMER	cs-en, cs-uk, en-cs, uk-cs	Jon et al. (2022)
CHARLESTRANSLATOR	cs-uk, uk-cs	Popel et al. (2022)
DLUT	en-ja, en-zh, ja-en, zh-en	(no associated paper)
GTCOM	cs-uk, en-hr, en-uk, en-zh, uk-cs, uk-en	Zong and Bei (2022)
HUAWEITSC	cs-uk, en-hr, en-liv, en-ru, en-uk, en-zh, liv-en, ru-en, uk-cs, uk-en, zh-en	Wei et al. (2022)
JDEXPLOREACADEMY	cs-en, de-en, en-cs, en-de, en-ja, en-ru, en-zh, ja-en, ru-en, zh-en	Zan et al. (2022)
KYB	en-ja, ja-en	Kalkar et al. (2022)
LT22	de-en, de-fr	Malli and Tambouratzis (2022)
LAN-BRIDGE	cs-en, cs-uk, de-en, en-cs, en-de, en-hr, en-ja, en-ru, en-uk, en-zh, fr-de, ja-en, ru-en, ru-sah, sah-ru, uk-cs, uk-en, zh-en	Han et al. (2022)
LANGUAGEX	en-ja, en-zh, ja-en, zh-en	Zeng (2022)
LIV4EVER	en-liv, liv-en	Rikters et al. (2022)
NAIST-NICT-TIT	en-ja, ja-en	Deguchi et al. (2022)
NT5	en-ja, ja-en	Morishita et al. (2022)
NIUTRANS	en-hr, en-liv, liv-en, zh-en	Shan et al. (2022)
OPENNMT	en-de	(no associated paper)
PROMT	de-en, en-de, en-ru, uk-en	Molchanov et al. (2022)
SRPOL	en-hr, en-ru, ru-en	Dobrowolski et al. (2022)
TAL-SJTU	en-liv, liv-en	He et al. (2022)
TARTUNLP	en-liv, liv-en	Tars et al. (2022)
ETRANSLATION	en-ru, en-uk, fr-de	Oravecz et al. (2022)
MANIFOLD	en-zh	Jin et al. (2022)
SHOPLINE-PL	cs-en	(no associated paper)

Table 4: Participants in the shared translation task. The translations from the online systems were not submitted by their respective companies but were obtained by us, and are therefore anonymized in a fashion consistent with previous years of the workshop.

verified team was limited to 7 submissions per test set. Submissions on leaderboards with BLEU and CHRF scores from SacreBLEU (Post, 2018) were displayed anonymously to avoid publishing rankings based on automatic scores during the submission period. Until one week after the submission period, teams could select a single primary submission per test set, specify if the primary submission followed a constrained or unconstrained setting, and submit a system description paper abstract. All entries were mandatory for a system submission to be included in the human evaluation campaign.

OCELoT has helped to simplify the submission process—from collecting submissions to gathering system information—and it supported the multi-domain shift introduced in the general task this year. The platform was also used for the Biomedical Shared Task (Neves et al., 2022). This made it easier to include the biomedical test set as another domain data in the test sets of the general task for languages that overlapped between the two tasks, which made it possible to collect outputs from the general domain systems for the biomedical domain.

4.2 Collaboration across WMT shared tasks

There are various shared tasks at WMT evaluating same language pairs but with different participants. This leads into inability to compare systems specialized for a particular task with participants of other tasks.

Therefore, we decided to open a collaboration across WMT shared tasks by asking participants to translate test sets from other shared tasks as well. This open the possibility to see how general MT systems compete for example in biomedical domain, or what is the general translation quality of specialized systems.

We set up a collaboration with Biomedical Shared Task (Neves et al., 2022) on all shared language pairs (Chinese-English, German-English, Russian-English).

This effort did not increase the number of participants for General MT Task because all participants of Biomedical Shared Task also participated in General MT. However, other participants of General MT have been evaluated on biomedical domain, too. For details, see Neves et al. (2022).

Language Pair	Sys.	Assess.	Assess/Sys
Czech→English	12	20,094	1,674.5
German→English	10	21,006	2,100.6
Japanese→English	14	28,638	2,045.6
Livonian→English	5	4,638	927.6
Russian→English	10	27,651	2,765.1
Ukrainian→English	9	20,305	2,256.1
Chinese→English	13	28,120	2,163.1
Total to-English	73	150,452	2,061

Table 5: Amount of data collected in the WMT22 manual evaluation campaign for evaluation into-English; after removal of quality control items.

	All	(A) Sig. Diff. Bad Ref.	(A) & No Sig. Diff. Exact Rep.
Czech→English	373	91 (24%)	78 (86%)
German→English	365	92 (25%)	84 (91%)
Japanese→English	538	129 (24%)	113 (88%)
Livonian→English	101	15 (15%)	15 (100%)
Russian→English	601	140 (23%)	125 (89%)
Ukrainian→English	395	88 (22%)	83 (94%)
Chinese→English	395	98 (25%)	79 (81%)
Total	**1,422**	**428 (30%)**	**388 (91%)**

Table 6: Number of crowd-sourced workers taking part in the reference-based SR+DC campaign; (A) those whose scores for bad reference items were significantly lower than corresponding MT outputs; those of (A) whose scores also showed no significant difference for exact repeats of the same translation; note: many workers evaluated more than one language pair.

5 Human Evaluation of Translation into English

As in previous years, reference-based Direct Assessment (DA, Graham et al., 2013, 2014, 2016) was employed as the primary method of evaluation for translation into English. DA human evaluation has several important features including accurate quality control of crowd-sourcing and standard methods of significance testing differences in ratings for systems. Human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation or source language input on an analogue scale, which corresponds to an underlying absolute 0–100 rating scale.[23] Direct Assessment is also employed for evaluation of video captioning systems at TRECvid (Graham et al., 2018; Awad et al., 2019) and multilingual surface realisation (Mille et al., 2018, 2019, 2020).

For evaluation of translation into-English, we

[23] No sentence or document length restriction is applied during manual evaluation.

use the monolingual configuration of DA, where the human evaluator reads and rates the system output translation and compares its meaning to an English reference translation, which was manually translated by a human translator. As recommended in Graham et al. (2020), we only employ forward-created test data to avoid potential bias. Since evaluating segments without their context (i.e. the surrounding document) can cause further bias (Läubli et al., 2018; Toral et al., 2018), we evaluate sentences in turn taken from a single document and system (described as "SR+DC" in previous WMT reports).[24] Similarly to last year, for all language pairs for which document context was available, we include it when evaluating translations. Note that the ratings are nevertheless collected on the segment level, motivated by the power analysis described in Graham et al. (2020), as well as better inter-annotator agreement and lower effort described in Castilho (2020).

In terms of the manual evaluation for the translation task for into-English language pairs, a total of 428 Turker accounts were involved.[25] 510,451 translation assessment scores were submitted in total by the crowd, of which 187,922 were provided by workers who passed quality control.[26]

System rankings are produced from a large set of human assessments of translations, each of which indicates the absolute quality of the output of a system. Table 5 shows total numbers of human assessments collected in WMT22 for into-English language pairs contributing to final scores for systems.[27]

Quality control was carried out exactly as described in last year's WMT for crowd-sourcing into-English translation assessments on Amazon Mechanical Turk (see Akhbardeh et al. (2021) for full details). Table 6 shows results of workers who passed quality control (by showing significant differences in scores attributed to translations of known to be of distinct qualities) and numbers of workers who also showed no significant difference for ratings of identical pairs of translations judged separately in repeat tests. Data from the non-reliable workers in all language pairs were removed prior to calculation of results.

Similar to last year, all rankings for to-English translation were reached through segment ratings presented one at a time in their original document order (SR+DC). As is usual with DA assessments, human assessment scores for translations were first standardized according to each individual human assessor's overall mean and standard deviation score. Average standardized scores for individual segments belonging to a given system were then computed, before the final overall DA score for a given system is computed as the average of its segment scores (Ave z in Table 7). Results are also reported for average scores for systems, computed in the same way but without any score standardization applied (Ave % in Table 7).

Human performance estimates calculated through the evaluation of human-produced reference translations are denoted by "HUMAN" in all tables. Translations HUMAN-C in Czech→English are known to be of lower quality than usual for manual translations.

Clusters are identified by grouping systems together according to which systems significantly outperform all others in lower ranking clusters, according to Wilcoxon rank-sum test.

All data collected during the human evaluation is available at `http://www.statmt.org/wmt22/results.html`. Appendix B shows the official results for the underlying head-to-head significance test for all pairs of systems.

6 Human Evaluation of Translation out of English and without English

Human evaluation for out-of-English and non-English translation directions[28] was performed with source-based ("bilingual") direct assessment of individual segments in context similar to the approach described in Akhbardeh et al. (2021). We use open-source framework Appraise for the evaluation (Federmann, 2018).

This year, several changes were made to the annotation procedure, the data sampling, and the interface display. In contrast to the standard DA (sliding scale from 0-100) used in 2021, this year annotators performed DA+SQM (Direct Assessment + Scalar Quality Metric). In DA+SQM, the annotators still provide a raw score between 0 and 100, but also

[24] The implementation still has the limitation that the assessors cannot go back to the previous segment.

[25] Numbers do not include the 988 workers on Mechanical Turk who did not pass quality control.

[26] Both numbers include quality control segments.

[27] Number of systems for WMT22 includes "human" systems comprising human-generated reference translations used to provide human performance estimates.

[28] We decided not to run human evaluation for French↔German due to the small number of system submissions this year.

Czech→English

Rank	Ave.	Ave. z	System
1	74.0	0.133	Online-W
2	75.3	0.055	CUNI-DocTformer
2	69.8	0.050	Lan-Bridge
2	70.7	0.037	Online-B
2	72.5	−0.004	JDExploreAcad
2	70.5	−0.014	Online-A
2	71.2	−0.015	CUNI-Transformer
2	71.4	−0.028	Online-G
2	71.9	−0.086	SHOPLINE-PL
10	67.7	−0.145	Online-Y
11	61.2	−0.290	HUMAN-C
11	64.0	−0.301	ALMAnaCH-Inria

Japanese→English

Rank	Ave.	Ave. z	System
1	66.7	0.069	DLUT
1	66.1	0.068	NT5
1	66.3	0.059	JDExploreAcademy
1	67.0	0.054	LanguageX
1	68.2	0.049	Online-B
1	66.1	0.046	Online-W
1	68.5	0.016	Lan-Bridge
1	67.1	0.006	Online-G
1	64.8	0.006	Online-A
1	63.8	−0.018	AISP-SJTU
1	66.5	−0.021	NAIST-NICT-TIT
1	66.6	−0.035	Online-Y
1	62.5	−0.056	KYB
14	26.2	−1.285	AIST

Russian→English

Rank	Ave.	Ave. z	System
1	77.5	0.055	JDExploreAcademy
1	77.5	0.040	HuaweiTSC
1	75.0	0.033	Online-G
1	76.7	0.008	Lan-Bridge
1	75.2	0.005	Online-Y
1	74.6	−0.003	SRPOL
1	74.3	−0.011	Online-B
1	74.7	−0.021	Online-A
1	76.1	−0.039	Online-W
10	69.8	−0.238	ALMAnaCH-Inria

German→English

Rank	Ave.	Ave. z	System
1	68.8	0.004	Lan-Bridge
2	70.8	−0.023	Online-W
2	68.1	−0.038	JDExploreAcademy
2	64.1	−0.057	Online-G
2	67.3	−0.070	Online-A
2	68.3	−0.086	HUMAN-B
2	66.5	−0.089	Online-Y
2	66.3	−0.092	Online-B
2	64.8	−0.126	LT22
2	66.2	−0.127	PROMT

Ukrainian→English

Rank	Ave.	Ave. z	System
1	73.5	0.048	Lan-Bridge
1	74.8	0.047	Online-B
3	69.8	0.039	HuaweiTSC
3	69.8	0.007	Online-A
3	73.6	−0.010	PROMT
3	73.4	−0.023	Online-G
7	71.0	−0.071	Online-Y
7	70.2	−0.082	ARC-NKUA
9	68.8	−0.246	ALMAnaCH-Inria

Livonian→English

Rank	Ave.	Ave. z	System
1	67.7	0.024	TartuNLP
1	66.0	−0.014	TAL-SJTU
1	64.0	−0.035	HuaweiTSC
1	63.5	−0.079	Liv4ever
5	60.4	−0.346	NiuTrans

Chinese→English

Rank	Ave.	Ave. z	System
−	73.4	0.134	HUMAN-B
1	69.8	−0.026	JDExploreAcademy
1	69.0	−0.034	HuaweiTSC
1	69.1	−0.063	AISP-SJTU
1	69.2	−0.079	LanguageX
1	69.7	−0.083	Online-A
1	68.6	−0.083	DLUT
1	67.4	−0.089	Online-B
1	69.9	−0.098	Online-G
1	66.5	−0.109	Online-W
1	65.3	−0.117	Lan-Bridge
1	66.5	−0.122	Online-Y
1	66.3	−0.164	NiuTrans

Table 7: Official results of WMT22 General Translation Task for translation into-English (SR+DC). Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges are based on the same test; grayed entry indicates resources that fall outside the constraints provided.

are presented with seven labeled tick marks, as visible in Figure 1. Discrete SQM (0-6) was found to correlate well with MQM (Multidimensional Quality Metrics) annotations by Freitag et al. (2021), while internal preliminary experiments suggested that DA+SQM helps to stabilize scores across annotators (as compared to DA). Annotators performing DA+SQM annotations at IWSLT 2022 human evaluation campaign (Anastasopoulos et al., 2022) also provided positive feedback about the annotation format. In previous years, full documents were sampled for annotation. This year we sampled a maximum of 10 consecutive segments from a document (a document "snippet") for annotation. This provides the potential to annotate segments from a more diverse range of documents while still maintaining a similar number of total annotations. Up to 10 source segments preceding and following the snippet being evaluated are displayed as static extra context for the annotator in the interface, as presented in Figure 1. As in past years, annotators provide both segment-level scores and document-level scores (in this case it is more accurate to call them snippet-level scores), however only the segment-level scores were used to compute the official rankings. As the English–Livonian data was not document-level, those annotations are run with segment-level-only DA+SQM. HITs (using the Amazon terminology of "Human Intelligence Task" to describe an annotation task) contained quality control segments, as described in Section 6.2. Rankings are computed as described in Section 6.4 based on segment-level scores.

6.1 Human Annotators

All annotations in the bilingual human evaluation campaign were carried out by hired professional annotators. This year, for the first time, we did not ask participants of the general task to contribute to human evaluation, but instead made it voluntary. The main motivations for this change were the attempt to increase the reliability and consistency of the judgements and the immense amount of time that was needed to be devoted to the process of collecting annotations from participating teams. Annotations for different language pairs were provided by different parties with their pool of annotators of distinct profiles as summarized in Table 8.

Charles University provided annotators for language pairs involving the Czech language, i.e. English→Czech and Ukrainian↔Czech. Their annotators were linguists, translators, researchers and students who are native speakers of the target language[29] with high proficiency in the source language.

University of Tartu provided the annotations for Livonian↔English, with 15% of the Livonian-speaking population participating in the annotation efforts. All three participants were near-native speakers of Livonian and participated in source-based Livonian-English and English-Livonian annotations, as well as reference-based Livonian annotation.

The second annotator group was provided by Toloka AI,[30] who collected annotations for English→Russian and Russian↔Yakut. Toloka AI is a global data labeling company that helps its customers to generate machine learning data at scale by harnessing the wisdom of the crowd from around the world. It relies on a geographically diverse crowd of several million registered users[31] (Pavlichenko et al., 2021). Toloka tests proficiency of their annotator crowd and excludes from future annotations anyone who does not pass quality control in the Appraise tool.

The last part of annotations was sponsored by Microsoft, who contributed with their pool of qualified paid bilingual speakers experienced in the MT evaluation process. Microsoft provided annotations for English into Chinese, Croatian, German, and Japanese, as well as Chinese→English as a comparison for reference-based evaluation described above and MQM evaluated in Metrics shared task (Freitag et al., 2022). For this pool of annotators, their performance is tracked over time, and those who fail quality control are permanently removed from the pool. This process increases the overall quality of the human assessment.

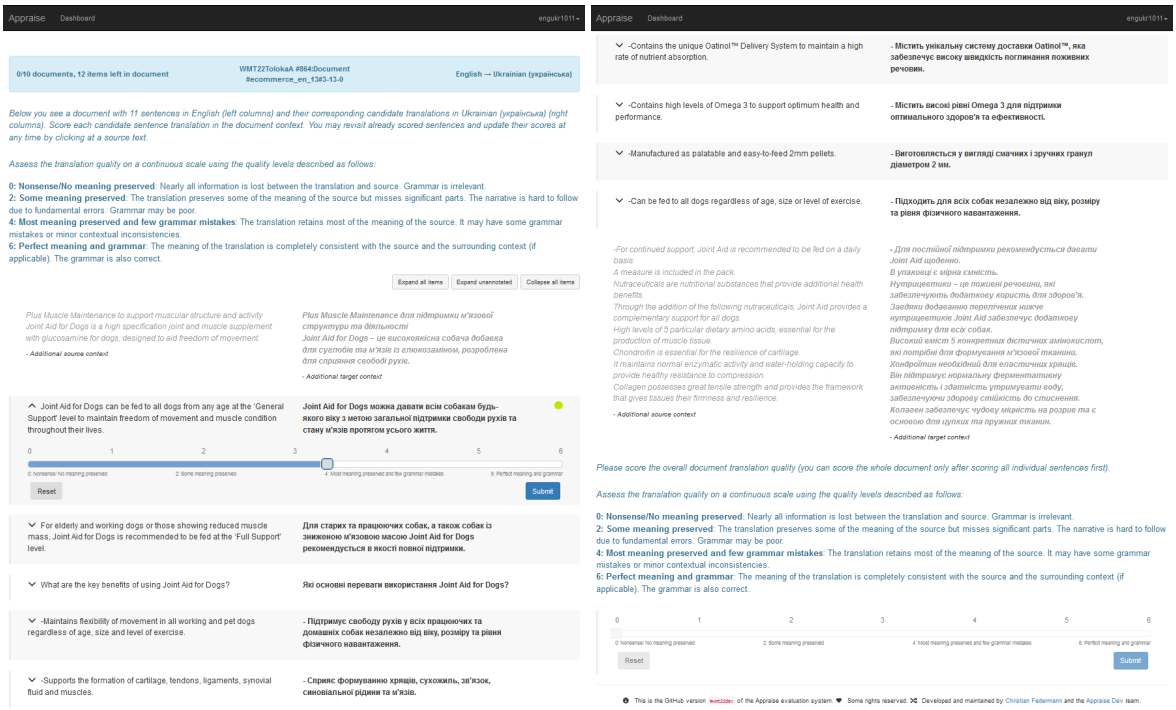
6.2 Sampling and Quality Control

In past WMT annotations, document-system pairs were sampled randomly for annotation, resulting in different subsets of the test set being annotated for each system. This year we first randomly sample a subset of document snippets from each of the domains for annotations, sampling the domains

[29] Some of Ukrainian→Czech annotators were not native Czechs, but native Ukrainians with near-native knowledge of Czech.

[30] https://toloka.ai

[31] https://hackernoon.com/evolution-of-the-data-production-paradigm-in-ai



(a) Top part of the screen with segment-level scoring.

(b) Bottom part of the screen with document-level scoring.

Figure 1: Screen shot of the document-level DA+SQM configuration in the Appraise interface for an example assessment from the human evaluation campaign for out of English language pairs. The annotator is presented with the entire translated document snippet randomly selected from competing systems (anonymized) with preceding and following contexts and is asked to rate the translation of individual segments and then the entire document on sliding scales.

Language pairs	Annotators' profile
English→Chinese/Croatian/German/Japanese	Microsoft annotators: bilingual target-language native speakers, professional translators or linguists, experienced in MT evaluation
English→Czech	Czech paid linguists, annotators, researchers, students with high proficiency in English
English→Livonian	Livonian speakers
English→Russian/Ukrainian, Russian↔Yakut	Toloka paid crowd: bilingual target-language native speakers
Ukrainian↔Czech	Paid translators and target-language native speakers

Table 8: Human annotator types for each language pair in bilingual human evaluation.

with approximately the same number of segments per domain. We use document snippets with 10 consecutive segments, or fewer in the case of short documents. In this way, all systems are annotated over almost exactly the same subset of document snippets.[32] All HITs consists of exactly 100 segments and are generated as in the past: (1) first snippet-system pairs are randomly sampled (from the restricted set of pre-sampled snippets) with up to 80 segments; (2) then random snippets with the remaining 20 (or more) segments are duplicated to serve as quality control items; (3) BAD references are introduced to the random segments in the duplicated snippets to have about 12-14% of quality control segments per HIT.[33] BAD references consist of segments in which an embedded sequences of tokens is replaced from a randomly placed phrase of the same length, sampled from a different reference segment.

We perform quality control by measuring an annotator's ability to reliably score BAD translations significantly lower than corresponding original system outputs using a paired significance test with $p < 0.05$. We pair two HITs into a single annota-

[32]For English→{Czech, German, Japanese, Russian, Ukrainian, Chinese} and the additional Chinese→English collection, all systems received annotations for all the sampled snippets. For Czech→Ukrainian, Ukrainian→Czech, English→Croatian, Yakut→Russian, and pairs including Livonian, annotation coverage of sampled snippets was incomplete; not all systems were scored over exactly the same set of segments.

[33]For full details, see the HIT and batch generation code: https://github.com/wmt-conference/wmt22-news-systems

Language Pair	Sys.	Assess.	Assess/Sys
Chinese→English	14	26,800	1,914.3
Czech→Ukrainian	12	21,285	1,773.8
English→Czech	12	24,000	2,000.0
English→German	11	21,800	1,981.8
English→Croatian	10	19,046	1,904.6
English→Japanese	14	27,600	1,971.4
English→Livonian	6	3,903	650.5
English→Russian	12	46,675	3,889.6
English→Ukrainian	9	35,048	3,894.2
English→Chinese	14	27,800	1,985.7
Yakut→Russian	3	4,200	1,400.0
Ukrainian→Czech	12	14,622	1,218.5

Table 9: Amount of data collected in the WMT22 manual evaluation campaign for evaluation out-of-English; including human references as systems; after removal of quality control items.

Language Pair	Ann.	HITs	HITs/Ann.
Chinese→English	12	134	11.2
English→Czech	16	120	7.5
English→German	14	109	7.8
English→Croatian	13	96	7.4
English→Japanese	17	138	8.1
English→Chinese	8	139	17.4

Table 10: Numbers of individual annotators taking part in the WMT22 human evaluation campaign and the average number of HITs collected per annotator.

tion task with about 24-28 quality control segments to ensure a sufficient sample size for the statistical test. If an annotator is not able to demonstrate reliability on BAD references, they are excluded from further annotations, the HITs are reset and annotated from scratch by another annotator.[34]

In addition to the quality control items, because this annotation is performed bilingually, reference translations are also evaluated as though they were submitted systems.

For language pairs where there was a concern about having sufficient annotations, two smaller batches of HITs were generated (such that at least all segments in the first batch could be covered for all systems, with the second campaign completed if possible; in the case of translation between Czech and Ukrainian, due to a large number of single-sentence documents, larger documents were sampled first).

6.3 Calibration HITs

For several language pairs (English→{Chinese, Croatian, Czech, German, Japanese} and Chinese→English), we collect calibration HITs in the DA+SQM interface: one identical HIT with 100 randomly selected segments completed by all annotators, in addition to their regular annotation HITs. By providing a small set of sentences annotated by all annotators, we are better able to examine questions about inter-annotator consistency. We release these alongside the other annotations and the anonymized mapping between annotators and HITs in order to enable additional analysis.

Table 10 shows the number of unique annotators for these languages, along with the total number of HITs and average number of HITs per annotator. For all pairs of annotators who completed both a calibration HIT and additional HIT(s) within a given language pair, we compute the Spearman's rank correlation coefficient between the two an-

[34]The quality control in bilingual human evaluation excluded 17 HITs in total: 1 Yakut→Russian, 2 English→Russian, 3 English→Ukrainian, 7 English→Livonian, 4 Czech↔Ukrainian.

Language Pair	Min.	Max.	Med.
Chinese→English	0.03	0.77	0.40
English→Czech	0.15	0.81	0.49
English→German	-0.18	0.47	0.21
English→Croatian	0.23	0.65	0.41
English→Japanese	-0.11	0.68	0.24
English→Chinese	-0.13	0.56	0.16

Table 11: Minimum, maximum, and median Spearman's rank correlation coefficients between pairs of annotators on calibration HIT segments.

Source-Based English→Livonian (Official WMT22 ranking)

Rank	Ave.	Ave. z	System
1	74.4	1.255	HUMAN-A
2	46.2	0.215	TAL-SJTU
3-4	36.9	-0.147	HuaweiTSC
3-4	36.3	-0.175	TartuNLP
5	33.8	-0.262	Liv4ever
6	17.9	-0.853	NiuTrans

Ref.-Based English→Livonian

Rank	Ave.	Ave. z	System
1	39.5	0.499	TAL-SJTU
2-4	31.8	0.077	TartuNLP
2-4	31.5	0.051	Liv4ever
2-4	31.0	0.037	HuaweiTSC
5	18.3	-0.656	NiuTrans

Source-Based Livonian→English

Rank	Ave.	Ave. z	System
1	81.7	1.009	HUMAN-A
2-3	60.3	0.257	TartuNLP
2-3	60.2	0.252	TAL-SJTU
4	50.4	-0.084	HuaweiTSC
5	41.3	-0.406	Liv4ever
6	23.1	-1.052	NiuTrans

Table 12: Three rankings for systems translating between English and Livonian.

notators' scores of the segments in the calibration HIT. Table 11 shows the minimum, maximum, and median correlations obtained by pairs of annotators for each language. These vary quite widely between languages, and we also note that across the calibration HITs, annotators vary widely in their use of the scoring space and the shape of their score distributions. Even within the same language pair (i.e., scoring the exact same set of segments in the calibration HIT), some annotators' scores are distributed across most of the 0-100 scoring space, some only produce scores above a certain threshold, and some treat the scale as though it were discretized according to the numerical scale shown in the interface (clustering most of their scores at the numerical marks the one can see in Figure 1).

6.4 Human Ranking Computation

The official rankings shown in Table 13 are generated on the basis of the segment-level DA+SQM scores that are collected within document snippet context for all language pairs.[35] The quality control (BAD) segments and any HITs that failed to pass quality control are removed prior to computing the rankings. Means and standard deviations for computing z-scores are computed at the HIT level. To compute system-level averages (both raw and z-score), any instances of multiple scores for the same segment are first averaged together, then all segment-level scores are averaged per system to compute the system-level scores. The clusters are computed using the Wilcoxon rank-sum test with $p < 0.05$. Rank ranges indicate the number of systems a particular system underperforms or outperforms (i.e., the top end of the rank range is $l + 1$ where l is the number of losses, while the bottom is $n - w$ where n is the total number of systems and w is the number of systems that the system in questions significantly wins against).

The rankings for translation between Livonian and English shown in Table 12 are computed in the same manner described above, but because the test set does not include document boundaries the data was collected without document context and some of the data collection was source-based while other portions were reference-based. As the official ranking for English→Livonian we consider the ranking computed from source-based human evaluation.

[35]The code used to generate the rankings in Table 13 can be found here: https://github.com/AppraiseDev/Appraise/blob/main/Campaign/management/commands/ComputeWMT21Results.py

6.5 Comparison of Human Evaluation Methods

In collaboration with the metrics shared task (Freitag et al., 2022), human annotation data for the Chinese→English direction was collected using three different approaches: the official monolingual reference-based SR+DC DA (Section 5, Table 7), the source-based fully document-level DA+SQM approach used for out-of-English and non-English directions (Section 6), and the Multidimensional Quality Metrics (MQM) framework (Freitag et al., 2021, 2022). We present the rankings produced by the three approaches in Table 14.

The DA rankings produced large clusters only for this language pair; that is, it was not possible to separate the performance into many system clusters with statistical significance. It is also important to note that the set of data over which each of these rankings was produced may have differed (e.g., the distribution over topic domains or the amount of coverage of the full test set), making it difficult to determine whether these differences in rankings represent differences due to data or due to different annotation methods.

7 Manual Error Analysis of English→Croatian translations

In addition to the official human evaluation by assigning DA scores, an analysis of errors in English→Croatian translations was carried out by an MT researcher with experience in human translation. The evaluation was carried out bilingually, while looking at the original English segment and all of its translations, both machine and human, all mixed together in a random order. The segments were presented in the natural order in the document, and the entire document (news article or review) was available by scrolling down or up.

The analysis was performed on the first 100 documents (80 reviews and 20 news articles), containing 603 segments (416 in reviews and 187 in news). All 14 review topics mentioned in Section 2.4 are included, although not uniformly distributed. The annotations are publicly available at https://github.com/wmt-conference/wmt22-news-systems/humaneval/en-hr/.

The errors were not coupled to any quality criterion (adequacy, fluency, readability) – all problematic words found in the translations were tagged as errors, no matter whether they are related to the source language, or are specific to the target lan-

English→Czech

Range	Ave.	Ave. z	System
1	91.2	0.335	HUMAN–C
2	90.9	0.279	Online-W
3	88.6	0.158	JDExploreAcad.
4-6	85.3	0.045	Online-B
4-6	87.1	0.041	Lan-Bridge
4-6	85.1	0.029	HUMAN-B
7-10	84.2	−0.059	CUNI-Bergamot
7-10	83.7	−0.074	CUNI-DocTransf.
7-10	84.0	−0.087	Online-A
7-10	83.2	−0.128	CUNI-Transf.
11-12	83.3	−0.258	Online-G
11-12	80.8	−0.310	Online-Y

Czech→Ukrainian

Range	Ave.	Ave. z	System
1	85.6	0.295	HUMAN-A
2-5	84.6	0.225	Online-B
2-3	84.1	0.151	AMU
3-6	82.5	0.125	Lan-Bridge
3-6	81.1	0.065	HuaweiTSC
4-8	81.9	0.062	CharlesTranslator
6-8	80.2	0.026	CUNI-JL-JH
6-8	80.2	−0.002	CUNI-Transf.
9-10	79.8	−0.008	Online-G
9-10	79.2	−0.075	Online-A
11	76.0	−0.257	Online-Y
12	68.4	−0.669	ALMAnaCH-Inria

Ukrainian→Czech

Range	Ave.	Ave. z	System
1	89.6	0.417	HUMAN-A
2-3	85.6	0.182	AMU
2-4	83.5	0.148	HuaweiTSC
4-8	83.5	0.127	Lan-Bridge
3-8	82.0	0.110	CUNI-Transf.
4-8	82.5	0.082	CharlesTranslator
4-8	81.4	0.052	CUNI-JL-JH
4-8	81.9	0.042	Online-B
9-10	80.0	-0.101	Online-A
9-10	77.5	-0.138	Online-G
11	73.9	-0.351	Online-Y
12	69.2	-0.617	ALMAnaCH-Inria

Yakut→Russian

Range	Ave.	Ave. z	System
1	71.3	0.708	HUMAN-A
2	54.6	0.178	Online-G
3	16.0	−0.873	Lan-Bridge

English→Chinese

Range	Ave.	Ave. z	System
1	81.7	0.154	HUMAN-A
2-5	81.9	0.099	Online-W
2-5	80.9	0.074	HUMAN-B
2-9	80.3	0.073	JDExploreAcad.
2-7	79.7	0.026	Online-Y
4-11	80.0	0.020	Lan-Bridge
4-11	78.5	0.019	Manifold
5-12	79.4	−0.012	LanguageX
5-12	79.4	−0.019	Online-B
6-12	78.7	−0.020	Online-A
8-12	79.6	−0.043	HuaweiTSC
6-12	79.0	−0.045	AISP-SJTU
13-14	77.5	−0.150	DLUT
13-14	77.2	−0.153	Online-G

English→German

Range	Ave.	Ave. z	System
1-6	93.9	0.116	HUMAN-A
1-4	93.6	0.106	Online-B
1-4	93.4	0.106	Online-W
1-5	92.4	0.071	JDExploreAcad.
3-7	93.8	0.051	HUMAN-B
5-9	93.6	0.015	Lan-Bridge
4-9	91.1	−0.019	Online-A
6-11	92.2	−0.054	Online-Y
6-11	93.2	−0.066	Online-G
8-11	90.8	−0.110	PROMT
8-11	89.9	−0.189	OpenNMT

English→Japanese

Range	Ave.	Ave. z	System
1	86.3	0.218	HUMAN-A
2-11	84.1	0.103	NT5
2-9	83.6	0.099	LanguageX
2-9	84.3	0.093	JDExploreAcad.
2-8	84.3	0.087	Online-B
2-9	83.9	0.078	DLUT
2-11	83.2	0.058	Online-Y
3-11	82.9	0.022	Lan-Bridge
6-11	82.9	0.018	Online-A
2-11	83.3	0.004	NAIST-NICT-TIT
11-12	81.9	−0.027	AISP-SJTU
6-12	83.0	−0.029	Online-W
13	79.5	−0.311	Online-G
14	76.9	−0.434	KYB

English→Russian

Range	Ave.	Ave. z	System
1-2	87.3	0.222	Online-W
1-2	86.6	0.194	HUMAN-A
3-5	86.0	0.136	Online-G
3-5	84.4	0.131	Online-B
3-5	84.2	0.096	JDExploreAcad.
6-7	84.3	0.046	Lan-Bridge
6-7	82.5	0.005	Online-Y
8-10	80.7	−0.086	Online-A
8-11	81.0	−0.123	PROMT
8-11	79.5	−0.159	SRPOL
9-12	79.6	−0.203	HuaweiTSC
11-12	79.4	−0.220	eTranslation

English→Croatian

Range	Ave.	Ave. z	System
1	93.7	0.327	HUMAN-A
2-3	92.6	0.264	HUMAN-st.
2-3	92.0	0.232	Online-B
4	91.2	0.155	Lan-Bridge
5-8	88.5	−0.018	Online-A
5-8	87.3	−0.057	HuaweiTSC
5-8	88.5	−0.068	SRPOL
5-8	87.0	−0.094	NiuTrans
9	84.5	−0.333	Online-G
10	82.3	−0.414	Online-Y

English→Ukrainian

Range	Ave.	Ave. z	System
1	87.1	0.319	HUMAN-A
2-4	84.0	0.124	Online-B
2-4	84.3	0.118	Lan-Bridge
2-4	83.5	0.092	Online-G
5-6	82.8	−0.018	Online-A
5-7	82.0	−0.037	HuaweiTSC
6-7	80.5	−0.105	eTranslation
8-9	79.6	−0.185	Online-Y
8-9	79.8	−0.233	ARC-NKUA

English→Livonian

Range	Ave.	Ave. z	System
1	74.4	1.255	HUMAN-A
2	46.2	0.215	TAL-SJTU
3-4	36.9	-0.147	HuaweiTSC
3-4	36.3	-0.175	TartuNLP
5	33.8	-0.262	Liv4ever
6	17.9	-0.853	NiuTrans

Table 13: Official results of WMT22 General Translation Task for translation out of English or without English. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges indicate the number of systems a system significantly underperforms or outperforms; grayed entry indicates resources that fall outside the constraints provided. All language pairs except English→Livonian used document-level evaluation.

guage, or both. There was no distinction of error severity ("major", "minor" or similar).

All identified errors (issues) were tagged by their possible causes and/or plausible explanations of their origin, as in (Popovic, 2021). Some of the identified "issue types" are equivalent to the typical error classes that can be found in MQM or similar schemes (such as "mistranslation", "gender", etc.), while some go beyond that, often including several different intertwining types of errors. Some of them involve single words, while others might involve a large group of words. The main difference between such tags in comparison to MQM or similar tags is that they are related to (linguistically motivated) causes of errors, also taking into account differences between source and target language as well as the translation process, and not only to the "symptoms" manifested in the MT output.

For example, the most frequent issue is related to "rephrasing", and refers to a sequence of words that is not translated properly for some of the following reasons: 1) the translation of the source words follows the structure of the source language although it should be expressed differently in the target language (rephrasing is needed); 2) rephrasing is needed but incorrectly applied; 3) rephrasing is not needed but is applied, and/or 4) the choice of target words is related to source words but seems random, both in lexical as well as grammatical terms. The issue is manifested by several consecutive different but intertwined types of errors such as case, gender, verb form, mistranslation, function word, omission, addition, word order, etc. Incorrect translation of multi-word expressions and collocations falls under this type.

Overall error rates Table 15 presents the aggregated error rates for each translation, calculated as the number of words which were tagged as any type of error divided by the total number of words in the text. Thus, the interpretation of, for example, the overall error rate of 12.76% for the MT system ONLINE-B is that about 12-13 incorrect words were found in each group of 100 words. The error rates are presented for the entire analysed text, as well as separately for the two domains. The translations are ranked from the lowest to the highest overall error rate.

The ranking is similar to the official direct assessment results presented in Table 13, however there are some different tendencies. The main difference is the preference for human translations – error rates exhibit a clear preference for human translations over MT outputs. While both scores agree on the four best translations (two human and two MT outputs), error rates clearly distinguish the two human translations with about 10% less errors than in the best MT output. Direct assessment scores, however, are all close, ranging from 93.7 to 91.2, and even put student translations at the same rank as the best MT output. The same tendency has been reported in Freitag et al. (2021), where the MQM error annotation on English→German and Chinese→English translations clearly distinguished human translations from MT outputs, contrary to direct assessment scores. These findings indicate that for evaluating human translations in any context (comparing different human translations, comparing with MT outputs), some kind of error annotation should be performed.

Another potentially interesting difference is the system ONLINE-G, which is clearly ranked as second worst by direct assessment, but less clearly as third worst by error annotation. A potential reason is the different nature of errors in different MT systems discussed below. Other differences between the two rankings affect only the mid-range systems which have very close scores in both set-ups.

It can be seen that errors were detected both in human and in machine translations, although the error rates are notably lower in human translations. Overall error rate is lower than 1% for professional translations and lower than 3% for students' translations, while in MT outputs, the overall error rates range from 12 to 22%.

In human translations, error rates are similar for both domains. In MT outputs, however, the error rates are notably higher for reviews than for news, which is not surprising given that there are much less training resources for reviews. Furthermore, it can be noted that the rankings would be slightly different if only one of the domains were used: NIUTRANS would be ranked higher on news while ONLINE-G would be ranked higher on reviews and HUAWEITSC would be ranked lower. Nevertheless, those variations in rankings can be observed only for the mid-ranged systems where differences in error rates are small anyway.

Comparing machine and human translations Table 16 presents issue types identified in machine and in human translations and their corresponding error rates. In addition, the distribution between the two domains is presented for each, meaning

	SR+DC DA				DA+SQM				MQM	
	Rank	Ave.	Ave. z	Order	Range	Ave.	Ave. z	Order	MQM score	Order
HUMAN-A	-	-	-	-	1-3	82.4	0.137	1	1.223	1
HUMAN-B	1	73.4	0.134	1	8-12	80	-0.029	9	1.997	2
JDExploreAcademy	2	69.8	-0.026	2	3-7	81.5	0.048	6	2.827	6
HuaweiTSC	2	69	-0.034	3	3-7	80.7	0.056	5	3.089	8
AISP-SJTU	2	69.1	-0.063	4	8-10	80.8	-0.013	8	3.187	9
LanguageX	2	69.2	-0.079	5	1-6	82	0.109	2	2.738	5
Online-A	2	69.7	-0.083	6	9-14	79.1	-0.078	10	3.731	11
DLUT	2	68.6	-0.083	7	11-14	79	-0.181	14	-	-
Online-B	2	67.4	-0.089	8	1-3	81.9	0.1	3	2.714	4
Online-G	2	69.9	-0.098	9	3-7	81.4	0.065	4	2.933	7
Online-W	2	66.5	-0.109	10	9-14	78.4	-0.098	12	3.953	12
Lan-Bridge	2	65.3	-0.117	11	4-7	81	0.041	7	2.471	3
Online-Y	2	66.5	-0.122	12	8-12	79.6	-0.086	11	3.281	10
NiuTrans	2	66.3	-0.164	13	11-14	79	-0.107	13	-	-

Table 14: Comparison of three methods of generating human annotations and rankings. Note that each method used different subsets of the test data, and the DA approaches only produced weak clusterings.

en→hr translation		error rate (%) ↓ overall	news	reviews
HT	professionals	0.71	0.86	0.60
	students	2.43	2.23	2.59
MT	online-B	12.76	11.19	13.98
	Lan-Bridge	13.42	11.46	14.95
	HuaweiTSC	17.39	12.87	**20.83**
	online-A	17.69	14.30	20.29
	SRPOL	17.96	14.55	20.56
	online-G	18.43	16.60	**19.80**
	NiuTrans	18.99	**13.51**	23.15
	online-Y	21.51	18.48	23.82

Table 15: Percentage of words marked as errors (error rate) in all translations: two human translations (by professional translators and by students) and eight machine translation hypotheses. The percentages are presented for the entire text (overall) and separately for news and for reviews. The translations are ranked from best to worst according to the overall error rate. Bold values indicate domain-specific ranks which are different from the overall rank.

that, for example, 32.2% of all rephrasing errors are found in news and 67.8% in reviews. Issue types are ranked according to their percentage in MT outputs.

The most prominent issues in MT outputs are similar to those reported in in (Popovic, 2021): rephrasing (described at the beginning of the section), ambiguity (different meanings of a word in different contexts), noun phrases (sequences of nouns and possibly adjectives) and omissions (either a part of the source text is omitted or something is missing in the target language), with the error rates ranging from 1% to 5%. Interestingly, the same issue types are the most frequent issues in human translations, too, although with much smaller error rates (less than 0.4%).

The majority of issue types in MT outputs is found more frequently in reviews than in news, although the differences vary. From the most prominent issues, only noun phrase errors are slightly more frequent in news. In human translations, the distribution of issue types between the two domains is more even, although the most prominent four are more frequent in reviews.

Somewhat surprisingly, hallucination errors were identified in the human translation of news. Further manual inspection revealed that in one sentence, a phrase not related to any part of the source text indeed appears in the professional translation. The probable reason is a somewhat specific financial term “like-for-like" meaning “financial growth". The source sentence “Drink-led pubs and bars performed by far the strongest with *like-for-likes* up more than restaurants were down." ended up translated as “Drink-led pubs and bars performed by far the strongest, *while pubs and bars selling both drinks and food had* more up than restaurants were down". The translator probably did not recognise the term and assumed that it refers to something similar to the previously mentioned “drink-led pubs and bars", so they added the phrase about ‘pubs and bars selling both drinks and food’ which were not mentioned whatsoever in the source. Without this hallucination, all error rates (overall, news and reviews) for professional translations presented in Table 15 would be 0.60%.

Comparing MT systems Table 17 presents the most frequent issue types (with error rate greater than 1%, or, in other words, which were found at least once in each 100 words) in each of the eight MT outputs. The outputs are ranked from best to worst according to the overall error rate (Table 15). For each issue type, its overall error rate together with the separated error rates in news and reviews

en→hr	MT outputs			human translations		
	error	% of the issue type		error	% of the issue type	
issue type	rate %	in news	in reviews	rate %	in news	in reviews
rephrasing	5.12	32.2	**67.8**	0.27	47.9	**52.1**
ambiguity	3.38	32.8	**67.2**	0.21	27.0	**73.0**
noun phrase	2.55	**53.6**	46.4	0.14	20.8	**79.2**
omission	1.22	48.0	**52.0**	0.37	46.9	**53.1**
named entity	0.86	47.4	**52.6**	0.05	50.0	50.0
verb form	0.86	31.3	**68.7**	0.06	**80.0**	20.0
gender	0.85	27.3	**72.7**	0.05	0	**100**
pron/det	0.64	12.7	**87.3**	0.02	0	**100**
preposition	0.54	42.0	**58.0**	0.07	**69.2**	30.8
untranslated	0.52	17.0	**83.0**	0.07	15.4	**84.6**
case	0.50	37.9	**62.1**	0.11	**73.7**	26.3
mistranslation	0.48	38.1	**61.9**	0.07	**61.5**	38.5
addition	0.43	14.8	**85.2**	0.01	0	**100**
source	0.34	2.6	**97.4**	0.02	0	**100**
order	0.28	33.7	**66.3**	0.03	**66.7**	33.3
non-existing	0.25	35.6	**64.4**	0.04	0	**100**
passive	0.19	**53.4**	46.6	0.01	0	**100**
number	0.17	24.1	**75.9**	0.01	0	**100**
-ing	0.16	**59.5**	40.5	0.01	0	**100**
rel. phrase	0.09	**66.7**	33.3	0	0	0
POS ambiguity	0.08	3.4	**96.6**	0	0	0
hallucination	0.07	30.8	**69.2**	0.06	**100**	0
negation	0.06	0	**100**	0	0	0
repetition	0.02	43.8	**56.2**	0.01	**100**	0

Table 16: Identified issues in all MT hypotheses and in both HT references: error rate together with the distribution between news and reviews. The issue types are ordered by their percentage in MT hypotheses. Bold values indicate the domain with the higher amount of a particular issue type.

is shown.

First, it can be noted that in the two best-ranked systems, there are three clearly predominant issue types for both domains: rephrasing, ambiguity and noun phrase. These three issue types are predominant in other systems, too, however with higher error rates.

Furthermore, for all systems, rephrasing errors and ambiguity problems are more frequent in reviews, whereas noun phrase errors are more frequent in news. Also in all systems, there are slightly more omissions in news than in reviews.

When looking at lower ranked systems, it can be noted that not only the error rates for the generally most prominent issue types increase, but also more error types emerge: incorrect verb forms, incorrect gender and problems with pronouns or determiners in reviews.

The most interesting system is ONLINE-G: while the rephrasing error rate is only slightly worse than the two best-ranked systems, and ambiguity and noun phrase errors are also not much worse than some of the higher-ranked systems, it is the only system with notable problems with named entities (more than 2%) and mistranslations (more than 1%) in both domains, as well as generating non-existing words in reviews (more than 1%). This specific distribution of error types could be the reason that this system was clearly ranked as the second worst by direct assessment, although it has similar error rate as some other systems.

In the lowest-ranked systems, apart from the higher error rates for all common issue types, the appearance of untranslated words in reviews can be noted in NIUTRANS, and problems with named entities in news in ONLINE-Y.

Apart from the described quantitative analysis, a qualitative inspection of the translation showed, as can be expected, that the MT outputs generally are close to the source language, without divergences. Nevertheless, some very creative and very nice machine translations were found, too.

Comparing human translations Table 18 presents the most frequent issue types (with error rate greater than 0.1%, or in other words, that were found at least once in each 1000 words) in each of the two human translations. The translations are ranked from best to worst according to the overall error rate (Table 15). For each issue type, its overall error rate together with the separated error rates in news and reviews is shown.

First, it can be noted that the most frequent error in both human translations in omission, being more frequent in student translations. The second issue

en→hr: MT hypotheses				
	most frequent	error rate ↓		
MT system	issue types	overall	news	reviews
online-B	rephrasing	4.14	2.87	5.13
	ambiguity	2.52	1.95	2.96
	noun phrase	1.86	2.77	1.15
Lan-Bridge	rephrasing	4.33	2.98	5.38
	ambiguity	2.62	2.03	3.08
	noun phrase	2.01	2.90	1.31
HuaweiTSC	rephrasing	5.49	3.97	6.65
	ambiguity	3.43	2.43	4.19
	noun phrase	2.64	2.77	2.54
	omission	1.04	1.23	<1
	verb form	<1	<1	1.04
	gender	<1	<1	1.10
online-A	rephrasing	5.06	3.68	6.12
	ambiguity	3.85	2.99	4.50
	noun phrase	2.88	3.20	2.63
	omission	1.11	1.38	<1
	gender	<1	<1	1.18
SRPOL	rephrasing	5.33	4.12	6.25
	ambiguity	3.82	2.69	4.68
	noun phrase	2.69	3.25	2.26
	omission	1.44	1.52	1.38
	verb form	<1	<1	1.00
	pron/det	<1	<1	1.08
online-G	rephrasing	4.59	3.54	5.38
	ambiguity	3.06	2.33	3.61
	noun phrase	2.17	3.28	1.33
	named entity	2.11	2.59	1.75
	omission	1.41	1.61	1.26
	mistranslation	1.37	1.11	1.57
	non-existing	1.06	<1	1.32
	verb form	<1	<1	1.22
	gender	<1	<1	1.04
	pron/det	<1	<1	1.16
NiuTrans	rephrasing	5.76	4.14	6.99
	ambiguity	3.30	2.29	4.08
	noun phrase	2.84	2.93	2.77
	omission	1.69	1.78	1.63
	gender	1.03	<1	1.45
	verb form	<1	<1	1.24
	untranslated	<1	<1	1.14
	pron/det	<1	<1	1.18
online-Y	rephrasing	6.26	5.08	7.16
	ambiguity	4.43	3.75	4.95
	noun phrase	3.29	4.07	2.70
	omission	1.32	1.49	1.20
	verb form	1.15	<1	1.32
	named entity	1.14	1.38	<1
	gender	1.13	<1	1.42
	pron/det	<1	<1	1.18

Table 17: The most frequent issue types (error rate $\geq 1\%$) in each of the eight MT hypotheses separately, overall as well as separately for news and reviews. The hypotheses are ranked from best to worst according to the overall error rate (Table 15).

en→hr: human translations				
	most frequent	error rate ↓		
	issue types	overall	news	reviews
prof.	omission	0.20	0.13	0.25
	rephrasing	0.14	0.18	0.10
	hallucination	0.11	0.26	0
stud.	omission	0.54	0.64	0.45
	rephrasing	0.41	0.41	0.41
	ambiguity	0.37	0.23	0.47
	noun phrase	0.25	0.13	0.35
	case	0.14	0.18	0.10
	untranslated	0.14	<0.1	0.21
	mistranslation	0.13	0.15	0.10
	preposition	0.11	0.15	<0.1
	verb form	<0.1	0.18	<0.1
	order	<0.1	0.10	<0.1
	named entity	<0.1	0.10	<0.1
	non-existing	<0.1	0	0.14
	gender	<0.1	<0.1	0.14

Table 18: The most frequent issue types (error rate $\geq 0.1\%$) in each of the two human reference translations separately, overall as well as separately for news and reviews. The translations are ranked from best to worst according to the overall error rate (Table 15).

type is rephrasing, also more frequent in student translations. The third ranked issue in professional translations are hallucinations, which is discussed in one of the previous paragraphs. For students, the third ranked issue are ambiguous words, apparently more problematic in reviews.

Furthermore, a number of issue types with error rate larger than 0.1% in student translations are less frequent or even not appearing at all in professional translations.

Apart from the described quantitative analysis, a qualitative inspection of the translation showed that students generally diverged more from the source language than professionals, which is the opposite of what could be intuitively expected. This is the probable reason that for all MT outputs, both automatic metrics, COMET and CHRF, are lower when calculated using student references.

8 Conclusions

The General Machine Translation Task at WMT 2022 covered 21 translation pairs, 15 of which had English on the source or target side and 6 were without English. Direct assessment (DA) was the main golden truth, although the style varied across language pairs. Into-English translation was evaluated against human reference translation, preserving the order of sentences in a document but not presenting the whole document at once (SR+DC). Out-of-English and non-English pairs offered the context to the annotators and allowed them to re-

visit the scores assigned to individual segments (DA+SQM), evaluating against the source.

9 Limitations

We opened a research question of testing general capabilities of MT systems. However, we have simplified this approach. Firstly, we only used four domains that are not specialized. Secondly, we used only cleaner sentences avoiding noisy in the source sentences.

Although we accept human judgement as a gold standard, giving us more reliable signal than automatic metrics, we should mention that human annotations are noisy (Wei and Jia, 2021) and their performance is affected by quality of other evaluated systems (Mathur et al., 2020). Moreover, reference-based human judgements are biased by the quality of references.

The error analysis of Croatian translations was carried out by one evaluator. Also, the selected sample is different than the one used for direct assessment.

10 Ethical consideration

Several of the domains contained texts that included personal data, for example the conversational data (See Section 2.5 for more details). Entities were replaced by anonymisation tags (e.g. #NAME#, #EMAIL#) to preserve the anonymity of the users behind the content.

The sentences in Ukrainian datasets (as described in Section 2.4) were collected with users' opt-in consent and any personal data related to people other than well-known people was pseudonymized (using random first names and surnames). Sentences where such pseudonymization would not be enough to preserve reasonable anonymity of the users (e.g. describing events uniquely identifying the persons involved) were not included in the test set.

As described in Section 2.2 and in the linguistic brief (Appendix Section C), inappropriate, controversial and/or explicit content was filtered out prior to translation, particularly keeping in mind the translators and not exposing them to such content or obliging them to translate it. A few sentences containing explicit content managed to escape the filter, and we removed these sentences from the test sets without translation.

Human evaluation using Appraise for collecting human judgements was fully anonymous. Automatically generated accounts associated with annotation tasks with single-sign-on URLs were distributed randomly among pools of annotators and did not allow for storing personal information. For language pairs for which we used calibration HITs, we received lists of tasks completed by an individual anonymous annotator.

Acknowledgments

This task would not have been possible without the sponsorship of monolingual data, test sets translation and evaluation from our partners. Namely Microsoft, Charles University, Toloka, NTT Resonant, LinguaCustodia, Webinterpret, Google, CyberAgent, and Phrase. This work was supported by the European Commission via its H2020 Program (project WELCOME, contract no. 870930) and by 20-16819X (LUSyD). We greatly appreciate the help of Jaroslava Hlaváčová, Lucie Poláková, Pavel Pecina, and Mariia Anisimova with preparation of the Ukrainian↔Czech testset. Additionally, we would like to thank Loïc Barrault, Markus Freitag, Jesús Gonzáles Rubio, Marcin Junczys-Dowmunt, Raheel Qader, Matiss Rikters and many others.

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

Maja Popović's participation was funded by the ADAPT SFI Centre for Digital Media Technology, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

Michal Novák's and Martin Popel's participation was funded by LM2018101 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic.

Yvette Graham's participation was supported by the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Trinity College Dublin funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa,

Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Jesujoba Alabi, Lydia Nishimwe, Benjamin Muller, Camille Rey, Benoît Sagot, and Rachel Bawden. 2022. Inria-almanach at wmt 2022: Does transcription help cross-script machine translation? In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Luca Diduch, Alan F. Smeaton, Yvette Graham, and Wessel Kraaij. 2019. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID*, volume 2019.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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

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

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference*

on Machine Translation: Volume 2, Shared Task Papers, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. 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: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Colin Cherry and George Foster. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*.

Hiroyuki Deguchi, Kenji Imamura, Masahiro Kaneko, Yuto Nishida, Yusuke Sakai, Justin Vasselli, Huy-Hien Vu, and Taro Watanabe. 2022. Naist-nict-tit wmt22 general mt task submission. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3278–3284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Dobrowolski, Mateusz Klimaszewski, Adam Myśliwy, Marcin Szymański, Jakub Kowalski, Kornelia Szypuła, Paweł Przewłocki, and Paweł Przybysz. 2022. Samsung r&d institute poland participation in wmt 2022. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021a. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021b. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Ana C. Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the wmt 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, George Foster, Chi kiu Lo, Craig Stewart, Tom Kocmi, Eleftherios Avramidis, Alon Lavie, and André F. T. Martins. 2022. Results of the wmt22 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, pages 1–28.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Bing Han, Yangjian Wu, Gang Hu, and Qiulin Chen. 2022. Lan-bridge mt's participation in the wmt 2022 general translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Zhiwei He, Xing Wang, Zhaopeng Tu, Shuming Shi, and Rui Wang. 2022. Tencent ai lab - shanghai jiao tong university low-resource translation system for the wmt22 translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Chang Jin, Tingxun Shi, Zhengshan Xue, and Xiaodong Lin. 2022. Manifold's english-chinese system at wmt22 general mt task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2022. Cunibergamot submission at wmt22 general translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shivam Kalkar, Yoko Matsuzaki, and Ben LI. 2022. Kyb general machine translation systems for wmt22. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations (ICLR)*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *CoRR*, abs/2007.03006.

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

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 337–338, Ghent, Belgium. European Association for Machine Translation.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Guangfeng Liu, Qinpei Zhu, Xingyu Chen, Renjie Feng, Jianxin Ren, Renshou Wu, Qingliang Miao, Rui Wang, and Kai Yu. 2022. The aisp-sjtu translation system for wmt 2022. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–Machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.

Marilena Malli and George Tambouratzis. 2022. Evaluating corpus cleanup methods in the wmt'22 news translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (SR'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The second multilingual surface realisation shared task (SR'19): Overview and evaluation results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. Association for Computational Linguistics.

Simon Mille, Anya Belz, Bernd Bohnet, Thiago Castro Ferreira, Yvette Graham, and Leo Wanner. 2020. The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.

Alexander Molchanov, Vladislav Kovalenko, and Natalia Makhamalkina. 2022. Promt systems for wmt22 general translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Makoto Morishita, Keito Kudo, Yui Oka, Katsuki Chousa, Shun Kiyono, Sho Takase, and Jun Suzuki. 2022. Nt5 at wmt 2022 general translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-Lopez, Eulalia Farre-Maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the wmt 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam mickiewicz university at wmt 2022: Ner-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Csaba Oravecz, Katina Bontcheva, David Kolovratnìk, Bogomil Kovachev, and Christopher Scott. 2022.

etranslation's submissions to the wmt22 general machine translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. CrowdSpeech and Vox DIY: Benchmark Dataset for Crowdsourced Audio Transcription. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Martin Popel. 2020. CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Martin Popel, Jindřich Libovický, and Jindřich Helcl. 2022. Cuni systems for the wmt 22 czech-ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020a. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020b. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

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

Maja Popovic. 2021. On nature and causes of observed MT errors. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 163–175, Virtual. Association for Machine Translation in the Americas.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Matīss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits, and Mark Fishel. 2022. Machine translation for Livonian: Catering to 20 speakers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 508–514, Dublin, Ireland. Association for Computational Linguistics.

Dimitrios Roussis and Vassilis Papavassiliou. 2022. The arc-nkua submission for the english-ukrainian general machine translation shared task at wmt22. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Weiqiao Shan, Zhiquan Cao, Yuchen Han, Siming Wu, Yimin Hu, Jie Wang, Yi Zhang, Hou Baoyu, Hang Cao, Chenghao Gao, Xiaowen Liu, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022. The niutrans machine translation systems for wmt22. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, and Ondřej Klejch. 2016. Using MT-ComparEval. In *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 76–82.

Maali Tars, Taido Purason, and Andre Tättar. 2022. Teaching unseen low-resource languages to large translation models. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang, and Ying Qin. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Johnny Wei and Robin Jia. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6840–6854, Online. Association for Computational Linguistics.

Changtong Zan, Keqin Peng, Liang Ding, Baopu Qiu, Boan Liu, Shwai He, Qingyu Lu, Zheng Zhang, Chuang Liu, Weifeng Liu, Yibing Zhan, and Dacheng Tao. 2022. Vega-mt: The jd explore academy machine translation system for wmt22. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Hui Zeng. 2022. No domain left behind. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Hao Zong and Chao Bei. 2022. Gtcom neural machine translation systems for wmt22. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

A Statistics of training data

This section describes statistics of the training corpora.

ja-en	Segments	en Toks	en Types
JParacrawl-v3	25.74M	682.78M	2.84M
NewsComm-v16	1.84k	45.28k	6.28k
WikiTitles-v3	757.04k	2.02M	281.88k
WikiMatrix	3.90M	72.32M	1.11M
JESC	2.80M	23.90M	161.38k
KFTT	440.29k	11.54M	190.88k
TED	241.74k	4.95M	64.04k
Total	33.88M	797.55M	3.75M
zh-en	**Segments**	**en Toks**	**en Types**
ParaCrawl(bonus)	14.17M	253.78M	1.87M
NewsComm-v16	313.67k	7.98M	76.36k
WikiTitles-v3	921.96k	2.55M	380.23k
UNPC	17.45M	479.54M	939.62k
CCMT			
WikiMatrix	2.60M	58.62M	1.06M
BackTrans News	19.76M	416.57M	1.19M
Total	55.22M	1.22B	4.01M

Table 19: Training data statistics for ja-en and zh-en. Only the English side statistics are reported, which are obtained after running MosesDecoder's `tokenizer.perl`, similar to Table 20.

Corpus Name	Segments	Tokens		Types	
cs-en		**cs**	**en**	**cs**	**en**
Europarl-v10	644.43k	14.95M	17.38M	172.47k	63.27k
ParaCrawl-v9	50.63M	738.33M	805.54M	4.77M	4.53M
CommonCrawl	161.84k	3.53M	3.93M	210.48k	128.39k
NewsCommentary-v16	253.27k	5.67M	6.27M	176.38k	70.77k
WikiTitles-v3	410.94k	985.54k	1.07M	219.38k	186.37k
WikiMatrix	2.09M	34.82M	39.20M	1.07M	798.09k
Tilde Corpus	2.09M	44.03M	47.83M	349.78k	210.28k
CzEng 2.0	60.98M	757.32M	848.02M	3.68M	2.49M
BackTrans News	126.83M	2.35B	2.66B	5.75M	3.84M
Total	244.10M	3.95B	4.42B		
de-en		**de**	**en**	**de**	**en**
Europarl-v10	1.82M	48.10M	50.47M	371.70k	113.91k
ParaCrawl-v9	278.31M	4.63B	4.90B	31.91M	15.99M
NewsCommentary-v16	388.48k	9.92M	9.83M	215.04k	86.50k
CommonCrawl	2.40M	54.68M	58.90M	1.64M	823.89k
WikiTitles-v3	1.47M	3.23M	3.76M	674.95k	573.28k
WikiMatrix	6.23M	114.22M	118.08M	2.86M	1.83M
Tilde Corpus	5.19M	118.11M	120.82M	986.37k	379.92k
Total	295.81M	4.98B	5.26B		
fr-de		**fr**	**de**	**fr**	**de**
Europarl-v10	1.79M	55.33M	47.49M	144.80k	368.53k
ParaCrawl-v9	7.22M	145.20M	123.51M	1.53M	2.37M
CommonCrawl	622.29k	16.59M	14.23M	332.24k	578.30k
WikiTitles-v3	1.01M	2.54M	2.15M	449.70k	503.34k
NewsCommentary-v16	295.65k	9.34M	7.67M	92.30k	185.28k
Tilde Corpus	4.31M	118.15M	96.00M	391.10k	954.49k
WikiMatrix	3.35M	68.26M	59.85M	1.10M	1.85M
Total	18.60M	415.42M	350.90M		
hr-en		**hr**	**en**	**hr**	**en**
ParaCrawl-v9	3.24M	80.75M	90.83M	1.05M	690.15k
Tilde Corpus	745.62k	14.38M	15.49M	196.78k	109.23k
OPUS	85.56M	928.96M	1.06B	5.26M	4.06M
Total	89.55M	1.02B	1.17B		
ru-en		**ru**	**en**	**ru**	**en**
ParaCrawl-(bonus)	5.38M	99.01M	120.02M	1.73M	1.22M
BackTranslation enru	36.77M	799.38M	839.92M	3.78M	1.92M
Yandex Corpus	1.00M	22.26M	24.30M	697.02k	377.83k
CommonCrawl	878.39k	20.61M	21.54M	712.81k	432.62k
UN Parallel Corpus	985.72k	887.11k	893.73k	5.68k	5.54k
WikiTitles-v3	1.19M	3.24M	3.26M	534.43k	457.93k
NewsCommentary-v16	331.51k	8.37M	8.82M	206.54k	82.93k
WikiMatrix	5.20M	94.00M	102.94M	2.24M	1.59M
Tilde Corpus	34.27k	813.70k	855.68k	62.61k	28.93k
Total	51.77M	1.05B	1.12B		
uk-en		**uk**	**en**	**uk**	**en**
ParaCrawl-(bonus)	13.35M	706.98M	721.28M	1.89M	1.26M
WikiMatrix	2.58M	43.76M	49.06M	1.40M	981.85k
Tilde	1.63k	39.93k	41.15k	8.38k	4.70k
ELRC EU Acts	129.94k	3.20M	3.46M	71.46k	33.52k
OPUS Corpus	48.94M	629.35M	704.32M	4.17M	2.89M
Total	65.01M	1.38B	1.48B		
cs-uk		**cs**	**uk**	**cs**	**uk**
WikiMatirx	848.96k	12.30M	12.28M	586.14k	641.72k
OPUS	11.65M	124.21M	125.84M	1.44M	1.68M
ELRC EU Acts	130.00k	2.86M	3.14M	69.58k	71.67k
Total	12.63M	139.38M	141.26M		
liv-en		**liv**	**en**	**liv**	**en**
Total (from OPUS)	0.77k	23.13k	14.21k	2.51k	2.43k
sah-ru		**sah**	**ru**	**sah**	**ru**
Total (from Yakut corpus)	30.15k	199.94k	225.95k	40.60k	40.64k

Table 20: Statistics for parallel training set provided for General/News Translation Task. All numbers are obtained after running MosesDecoder's `tokenizer.perl`. *Tokens* are the total number of words, whereas *Types* are total number of distinct case-insensitive words. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

B Differences in Human Scores

Tables 23–27 show differences in average standardized human scores for all pairs of competing to-English systems for each language pair. The numbers in each of the tables' cells indicate the difference in average standardized human scores for the system in that column and the system in that row.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied Wilcoxon rank-sum test to measure the likelihood that such differences could occur simply by chance. In the following tables $\star$ indicates statistical significance at $p < 0.05$, † indicates statistical significance at $p < 0.01$, and ‡ indicates statistical significance at $p < 0.001$, according to Wilcoxon rank-sum test.

Each table contains final rows showing the average score achieved by that system and the rank range according to Wilcoxon rank-sum test ($p < 0.05$). Gray lines separate clusters based on non-overlapping rank ranges.

	ONLINE-W	CUNI-DOCTRANSFORMER	LAN-BRIDGE	ONLINE-B	JDEXPLOREACADEMY	ONLINE-A	CUNI-TRANSFORMER	ONLINE-G	SHOPLINE-PL	ONLINE-Y	HUMAN-	ALMANACH-INRIA
ONLINE-W	-	0.08⋆	0.08‡	0.10‡	0.14‡	0.15‡	0.15‡	0.16‡	0.22‡	0.28‡	0.42‡	0.43‡
CUNI-DOCTRANSFORMER	-0.08	-	0.01	0.02⋆	0.06	0.07†	0.07†	0.08‡	0.14‡	0.20‡	0.35‡	0.36‡
LAN-BRIDGE	-0.08	-0.01	-	0.01	0.05	0.06	0.07	0.08⋆	0.14†	0.20‡	0.34‡	0.35‡
ONLINE-B	-0.10	-0.02	-0.01	-	0.04	0.05	0.05	0.07	0.12†	0.18‡	0.33‡	0.34‡
JDEXPLOREACADEMY	-0.14	-0.06	-0.05	-0.04	-	0.01	0.01	0.02	0.08†	0.14‡	0.29‡	0.30‡
ONLINE-A	-0.15	-0.07	-0.06	-0.05	-0.01	-	0.00	0.01	0.07	0.13‡	0.28‡	0.29‡
CUNI-TRANSFORMER	-0.15	-0.07	-0.07	-0.05	-0.01	0.00	-	0.01	0.07⋆	0.13‡	0.27‡	0.29‡
ONLINE-G	-0.16	-0.08	-0.08	-0.07	-0.02	-0.01	-0.01	-	0.06	0.12‡	0.26‡	0.27‡
SHOPLINE-PL	-0.22	-0.14	-0.14	-0.12	-0.08	-0.07	-0.07	-0.06	-	0.06†	0.20‡	0.21‡
ONLINE-Y	-0.28	-0.20	-0.20	-0.18	-0.14	-0.13	-0.13	-0.12	-0.06	-	0.14‡	0.16‡
HUMAN-	-0.42	-0.35	-0.34	-0.33	-0.29	-0.28	-0.27	-0.26	-0.20	-0.14	-	0.01
ALMANACH-INRIA	-0.43	-0.36	-0.35	-0.34	-0.30	-0.29	-0.29	-0.27	-0.21	-0.16	-0.01	-
score	0.13	0.06	0.05	0.04	-0.00	-0.01	-0.01	-0.03	-0.09	-0.14	-0.29	-0.30
rank	1	2–3	2–7	3–8	2–8	3–9	3–8	4–9	7–9	10	11–12	11–12

Table 21: Head to head comparison for Czech→English systems

	HUMAN-B	JDEXPLOREACADEMY	HUAWEITSC	AISP-SJTU	LANGUAGEX	ONLINE-A	DLUT	ONLINE-B	ONLINE-G	ONLINE-W	LAN-BRIDGE	ONLINE-Y	NIUTRANS
HUMAN-B	-	0.16‡	0.17‡	0.20‡	0.21‡	0.22‡	0.22‡	0.22‡	0.23‡	0.24‡	0.25‡	0.26‡	0.30‡
JDEXPLOREACADEMY	-0.16	-	0.01	0.04	0.05	0.06⋆	0.06⋆	0.06	0.07†	0.08‡	0.09‡	0.10‡	0.14‡
HUAWEITSC	-0.17	-0.01	-	0.03	0.05	0.05	0.05⋆	0.06	0.06†	0.08‡	0.08‡	0.09‡	0.13‡
AISP-SJTU	-0.20	-0.04	-0.03	-	0.02	0.02	0.02	0.03	0.04	0.05⋆	0.05⋆	0.06†	0.10‡
LANGUAGEX	-0.21	-0.05	-0.05	-0.02	-	0.00	0.00	0.01	0.02	0.03⋆	0.04⋆	0.04†	0.09‡
ONLINE-A	-0.22	-0.06	-0.05	-0.02	0.00	-	0.00	0.01	0.02	0.03⋆	0.03⋆	0.04†	0.08‡
DLUT	-0.22	-0.06	-0.05	-0.02	0.00	0.00	-	0.01	0.02	0.03	0.03	0.04⋆	0.08‡
ONLINE-B	-0.22	-0.06	-0.06	-0.03	-0.01	-0.01	-0.01	-	0.01	0.02⋆	0.03⋆	0.03†	0.08‡
ONLINE-G	-0.23	-0.07	-0.06	-0.04	-0.02	-0.02	-0.02	-0.01	-	0.01	0.02	0.02⋆	0.07‡
ONLINE-W	-0.24	-0.08	-0.08	-0.05	-0.03	-0.03	-0.03	-0.02	-0.01	-	0.01	0.01	0.06⋆
LAN-BRIDGE	-0.25	-0.09	-0.08	-0.05	-0.04	-0.03	-0.03	-0.03	-0.02	-0.01	-	0.00	0.05⋆
ONLINE-Y	-0.26	-0.10	-0.09	-0.06	-0.04	-0.04	-0.04	-0.03	-0.02	-0.01	0.00	-	0.04
NIUTRANS	-0.30	-0.14	-0.13	-0.10	-0.09	-0.08	-0.08	-0.08	-0.07	-0.06	-0.05	-0.04	-
score	0.13	-0.03	-0.03	-0.06	-0.08	-0.08	-0.08	-0.09	-0.10	-0.11	-0.12	-0.12	-0.16
rank	1	2–6	2–7	2–9	2–9	3–9	4–11	2–9	4–11	8–12	8–12	10–13	12–13

Table 22: Head to head comparison for Chinese→English systems

	LAN-BRIDGE	ONLINE-W	JDEXPLOREACADEMY	ONLINE-G	ONLINE-A	HUMAN-	ONLINE-Y	ONLINE-B	LT22	PROMT
LAN-BRIDGE	-	0.03⋆	0.04†	0.06†	0.07‡	0.09‡	0.09‡	0.10‡	0.13‡	0.13‡
ONLINE-W	-0.03	-	0.02	0.03	0.05	0.06	0.07†	0.07⋆	0.10‡	0.10‡
JDEXPLOREACADEMY	-0.04	-0.02	-	0.02	0.03	0.05	0.05†	0.05⋆	0.09‡	0.09†
ONLINE-G	-0.06	-0.03	-0.02	-	0.01	0.03	0.03⋆	0.03	0.07†	0.07⋆
ONLINE-A	-0.07	-0.05	-0.03	-0.01	-	0.02	0.02	0.02	0.06†	0.06
HUMAN-	-0.09	-0.06	-0.05	-0.03	-0.02	-	0.00	0.01	0.04†	0.04⋆
ONLINE-Y	-0.09	-0.07	-0.05	-0.03	-0.02	0.00	-	0.00	0.04	0.04
ONLINE-B	-0.10	-0.07	-0.05	-0.03	-0.02	-0.01	0.00	-	0.03⋆	0.04
LT22	-0.13	-0.10	-0.09	-0.07	-0.06	-0.04	-0.04	-0.03	-	0.00
PROMT	-0.13	-0.10	-0.09	-0.07	-0.06	-0.04	-0.04	-0.04	0.00	-
score	0.00	-0.02	-0.04	-0.06	-0.07	-0.09	-0.09	-0.09	-0.13	-0.13
rank	1	2–6	2–6	2–7	2–9	2–8	5–10	4–9	8–10	6–10

Table 23: Head to head comparison for German→English systems

	JDEXPLOREACADEMY	HUAWEITSC	ONLINE-G	LAN-BRIDGE	ONLINE-Y	SRPOL	ONLINE-B	ONLINE-A	ONLINE-W	ALMANACH-INRIA
JDEXPLOREACADEMY	-	0.01	0.02	0.05⋆	0.05⋆	0.06⋆	0.07†	0.08†	0.09‡	0.29‡
HUAWEITSC	-0.01	-	0.01	0.03⋆	0.04⋆	0.04⋆	0.05†	0.06†	0.08‡	0.28‡
ONLINE-G	-0.02	-0.01	-	0.02	0.03	0.04	0.04	0.05	0.07⋆	0.27‡
LAN-BRIDGE	-0.05	-0.03	-0.02	-	0.00	0.01	0.02	0.03	0.05⋆	0.25‡
ONLINE-Y	-0.05	-0.04	-0.03	0.00	-	0.01	0.02	0.03	0.04	0.24‡
SRPOL	-0.06	-0.04	-0.04	-0.01	-0.01	-	0.01	0.02	0.04	0.23‡
ONLINE-B	-0.07	-0.05	-0.04	-0.02	-0.02	-0.01	-	0.01	0.03	0.23‡
ONLINE-A	-0.08	-0.06	-0.05	-0.03	-0.03	-0.02	-0.01	-	0.02	0.22‡
ONLINE-W	-0.09	-0.08	-0.07	-0.05	-0.04	-0.04	-0.03	-0.02	-	0.20‡
ALMANACH-INRIA	-0.29	-0.28	-0.27	-0.25	-0.24	-0.23	-0.23	-0.22	-0.20	-
score	0.06	0.04	0.03	0.01	0.01	-0.00	-0.01	-0.02	-0.04	-0.24
rank	1–3	1–3	1–8	3–8	3–9	3–9	3–9	3–9	5–9	10

Table 24: Head to head comparison for Russian→English systems

	DLUT	NT5	JDExploreAcademy	LanguageX	Online-B	Online-W	Lan-Bridge	Online-G	Online-A	AISP-SJTU	NAIST-NICT-TIT	Online-Y	KYB	AIST
DLUT	-	0.00	0.01	0.02	0.02	0.02	0.05⋆	0.06†	0.06†	0.09‡	0.09‡	0.10‡	0.13‡	1.35‡
NT5	0.00	-	0.01	0.01	0.02	0.02	0.05⋆	0.06⋆	0.06⋆	0.09‡	0.09†	0.10†	0.12‡	1.35‡
JDExploreAcademy	-0.01	-0.01	-	0.00	0.01	0.01	0.04	0.05	0.05	0.08†	0.08⋆	0.09†	0.11‡	1.34‡
LanguageX	-0.02	-0.01	0.00	-	0.01	0.01	0.04	0.05	0.05	0.07†	0.07⋆	0.09†	0.11‡	1.34‡
Online-B	-0.02	-0.02	-0.01	-0.01	-	0.00	0.03	0.04⋆	0.04⋆	0.07†	0.07⋆	0.08†	0.10‡	1.33‡
Online-W	-0.02	-0.02	-0.01	-0.01	0.00	-	0.03	0.04	0.04	0.06†	0.07⋆	0.08†	0.10‡	1.33‡
Lan-Bridge	-0.05	-0.05	-0.04	-0.04	-0.03	-0.03	-	0.01	0.01	0.03	0.04	0.05	0.07†	1.30‡
Online-G	-0.06	-0.06	-0.05	-0.05	-0.04	-0.04	-0.01	-	0.00	0.02	0.03	0.04	0.06†	1.29‡
Online-A	-0.06	-0.06	-0.05	-0.05	-0.04	-0.04	-0.01	0.00	-	0.02	0.03	0.04	0.06†	1.29‡
AISP-SJTU	-0.09	-0.09	-0.08	-0.07	-0.07	-0.06	-0.03	-0.02	-0.02	-	0.00	0.02	0.04	1.27‡
NAIST-NICT-TIT	-0.09	-0.09	-0.08	-0.07	-0.07	-0.07	-0.04	-0.03	-0.03	0.00	-	0.01	0.04⋆	1.26‡
Online-Y	-0.10	-0.10	-0.09	-0.09	-0.08	-0.08	-0.05	-0.04	-0.04	-0.02	-0.01	-	0.02	1.25‡
KYB	-0.13	-0.12	-0.11	-0.11	-0.10	-0.10	-0.07	-0.06	-0.06	-0.04	-0.04	-0.02	-	1.23‡
AIST	-1.35	-1.35	-1.34	-1.34	-1.33	-1.33	-1.30	-1.29	-1.29	-1.27	-1.26	-1.25	-1.23	-
score	0.07	0.07	0.06	0.05	0.05	0.05	0.02	0.01	0.01	-0.02	-0.02	-0.04	-0.06	-1.28
rank	1–6	1–6	1–9	1–9	1–7	1–9	3–12	4–12	4–12	7–13	7–12	7–13	11–13	14

Table 25: Head to head comparison for Japanese→English systems

	TartuNLP	TAL-SJTU	HuaweiTSC	Liv4Ever	NiuTrans
TartuNLP	-	0.04	0.06⋆	0.10†	0.37‡
TAL-SJTU	-0.04	-	0.02	0.07	0.33‡
HuaweiTSC	-0.06	-0.02	-	0.04	0.31‡
Liv4Ever	-0.10	-0.07	-0.04	-	0.27‡
NiuTrans	-0.37	-0.33	-0.31	-0.27	-
score	0.02	-0.01	-0.04	-0.08	-0.35
rank	1–2	1–4	2–4	2–4	5

Table 26: Head to head comparison for Livonian→English systems

	Lan-Bridge	Online-B	HuaweiTSC	Online-A	PROMT	Online-G	Online-Y	ARC-NKUA	ALMAnaCH-Inria
Lan-Bridge	-	0.00	0.01⋆	0.04⋆	0.06†	0.07†	0.12‡	0.13‡	0.29‡
Online-B	0.00	-	0.01†	0.04⋆	0.06‡	0.07‡	0.12‡	0.13‡	0.29‡
HuaweiTSC	-0.01	-0.01	-	0.03	0.05	0.06	0.11†	0.12†	0.28‡
Online-A	-0.04	-0.04	-0.03	-	0.02	0.03	0.08†	0.09†	0.25‡
PROMT	-0.06	-0.06	-0.05	-0.02	-	0.01	0.06⋆	0.07⋆	0.24‡
Online-G	-0.07	-0.07	-0.06	-0.03	-0.01	-	0.05⋆	0.06⋆	0.22‡
Online-Y	-0.12	-0.12	-0.11	-0.08	-0.06	-0.05	-	0.01	0.17‡
ARC-NKUA	-0.13	-0.13	-0.12	-0.09	-0.07	-0.06	-0.01	-	0.16‡
ALMAnaCH-Inria	-0.29	-0.29	-0.28	-0.25	-0.24	-0.22	-0.17	-0.16	-
score	0.05	0.05	0.04	0.01	-0.01	-0.02	-0.07	-0.08	-0.25
rank	1–2	1–2	3–6	3–6	3–6	3–6	7–8	7–8	9

Table 27: Head to head comparison for Ukrainian→English systems

C Preprocessing cleanup brief for linguists

In this task, we wish to check the data to remove all inappropriate content, remove repetitive content, or correct minor problems with the text.

The data is automatically broken down into individual sentences, which may be wrong sentence splitting. Each document is separated by empty lines. Keep the document-separators intact, split long documents into several by adding empty lines if necessary based on the context (some documents may be merged). In general, documents should be under 30 sentences long.

In the first step, check if a document shouldn't be removed (delete sentences from document) based on the following conditions, be on the save side, rather remove documents where you are uncertain. The conditions for removal of documents are as follows:

- Remove inappropriate content (such as sexually explicit, vulgar, or otherwise inappropriate)
- Remove controversial content (propagandist, controversial political topics, etc.)
- Remove content that is too noisy or doesn't resemble natural text (such as documents badly formatted, hard to understand, containing unusual language, lists or other structured data generated automatically)
- Remove repeated/similar content already part of previous documents

For documents that are not removed, do minor corrections (do not try reformulating the content). The main goal is to make sure each line contains a single sentence (or is empty line which represent document boundaries). The result should be documents that are fluent when reading. Here is a non-complete list of phenomena to pay attention to:

- Each line must be a single sentence, remove anything that dangles around or doesn't fit the context. Also reconnect sentences that have been accidentally split (for example trailing words or punctuation should be appended to the previous line).
- You may do small corrections to make the text cleaner (adding punctuation, correcting small typos, etc.). If text would need more correction, remove whole document. Also, do not polish everything.
- Sentences containing a short phrase or single words that are not necessary for the context (like "Description:" or emoticons like ":)") can be removed.

D Translator Brief for General MT

Translator Brief

In this project we wish to translate online news articles for use in evaluation of Machine Translation (MT). The translations produced by you will be compared against the translations produced by a variety of different MT systems. They will be released to the research community to provide a benchmark, or "gold-standard" measure for translation quality. The translation therefore needs to be a high-quality rendering of the source text into the target language, as if it was news written directly in the target language. However, there are some constraints imposed by the intended usage:

- All translations should be **"from scratch", without post-editing from MT**. Using post-editing would bias the evaluation, so we need to avoid it. We can detect post-editing so will reject translations that are post-edited.
- Translation should **preserve the sentence boundaries.** The source texts are provided with exactly one sentence per line, and the translations should be the same, one sentence per line. Blank lines should be preserved in the translation.
- Translators should **avoid inserting parenthetical explanations** into the translated text and obviously **avoid losing any pieces of information** from the source text. We will check a sample of the translations for quality, and we will check the entire set for evidence of post-editing.
- Please do not translate the anonymization tags (e.g. #NAME#), but use the same form as in the source text. These tags are used to de-identify names and various other sensitive data. In other words, translation must contain given tag #NAME# on a position where it would naturally be placed before anonymization.

The source files will be delivered as text files (sometimes known as "notepad" files), with one sentence per line. We need the translations to be returned in the same format. If you prefer to receive the text in a different format, then please let us know as we may be able to accommodate it.

E Additional statistics of the test sets

Table 28 shows the type-token ratios for the source and target side of each of the test sets, shown for each available domain. As mentioned previously, texts are tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) where available. For Czech, Livonian and Yakut, for which Spacy models are not available, we took as a rough approximation models for Croatian, Finnish and Russian respectively. The type-token ratio is calculated as the number of unique tokens divided by the total number of tokens. The absolute value depends not only on the lexical diversity of the text but also on the morphological complexity of the language in question.

	Type-token ratio (source)					Type-token ratio (target)				
	conversation	ecommerce	news	social	other	conversation	ecommerce	news	social	other
cs-en	-	-	0.40	0.38	-	-	-	0.21	0.22	-
cs-uk	-	-	-	-	0.34	-	-	-	-	0.32
de-en	0.25	0.36	0.35	0.29	-	0.18	0.24	0.26	0.21	-
de-fr	0.25	0.36	0.35	0.29	-	0.20	0.24	0.26	0.23	-
en-cs	0.15	0.24	0.26	0.23	-	0.23	0.36	0.41	0.36	-
en-de	0.15	0.24	0.26	0.23	-	0.15	0.27	0.3	0.26	-
en-hr	-	0.20	0.24	-	-	-	0.31	0.36	-	-
en-ja	0.15	0.24	0.26	0.23	-	0.10	0.17	0.18	0.18	-
en-liv	-	-	-	-	0.25	-	-	-	-	0.34
en-ru	0.15	0.24	0.26	0.23	-	0.21	0.35	0.39	0.33	-
en-uk	0.15	0.24	0.26	0.23	-	0.20	0.34	0.37	0.34	-
en-zh	0.15	0.24	0.26	0.23	-	0.13	0.22	0.26	0.25	-
fr-de	0.19	0.26	0.27	0.26	-	0.18	0.30	0.31	0.28	-
ja-en	0.24	0.20	0.23	0.24	-	0.24	0.24	0.26	0.24	-
liv-en	-	-	-	-	0.34	-	-	-	-	0.25
ru-en	-	0.44	0.35	-	0.43	-	0.26	0.20	-	0.27
ru-sah	-	-	-	-	0.34	-	-	-	-	0.38
sah-ru	-	-	-	-	0.38	-	-	-	-	0.34
uk-cs	-	-	-	-	0.28	-	-	-	-	0.26
uk-en	-	-	-	-	0.28	-	-	-	-	0.13
zh-en	0.24	0.30	0.25	0.27	-	0.17	0.21	0.17	0.20	-

Table 28: Type-token ratio for individual source languages used in the general translation test sets.

F News Task System Submission Summaries

F.1 AISP-SJTU (Liu et al., 2022)

This paper describes AISP-SJTU's participation in WMT 2022 shared general mt task on English->Chinese, Chinese->English, English->Japanese and Japanese->English with constrained training data. Our systems are based on the Transformer architecture with several novel and effective variants, including network depth and internal structure. In our experiments, we employ data filtering, large-scale back-translation, knowledge distillation, forward-translation, iterative in-domain knowledge finetune and model ensemble.

F.2 AIST (no associated paper)

The model was trained similarly to Optimus (Li et al., 2020) with the difference of using BERT (Devlin et al., 2019) for both encoding and decoding instead of BERT for encoding and GPT-2 for decoding as in Optimus, therefore enabling non-autoregressive sequence-to-sequence modeling. We used the pre-trained "bert-base-cased" configuration for English and the "bert-base-japanese" from CL Tohoku for Japanese.

F.3 ALMAnaCH-Inria (Alabi et al., 2022)

ALMAnaCH-Inria's primary submissions are multilingual transformer models between English, Russian, Ukrainian and Russian. The models exploit a dedicated Latin-script transcription convention designed to represent the Slavic languages in a way that maximises character- and word-level correspondences between them as well as with English. For directions where the target language is not English, this involves a final translation step into the original script. Our hypothesis was that bringing the languages

closer together could boost vocabulary sharing and have a positive impact on machine translation results. Initial results indicate that the transcription strategy was not successful, resulting in lower results than baselines. We nevertheless submit these models as our primary systems.

F.4 AMU (Nowakowski et al., 2022)

AMU submission is a weighted ensemble of 4 models based on the transformer-big architecture. Models use source factors to utilize the information about named entities present in the input. Each of the models in the ensemble was trained using only the data provided by the shared task organizers. A noisy back-translation technique was used to augment the training corpora. One of the models in the ensemble is a document-level model, trained on parallel and synthetic longer sequences. During the sentence-level decoding process, the ensemble generated the n-best list (n=200). The n-best list was merged with the n-best list (n=50) generated by a single document-level model which translated multiple sentences at a time. Finally, existing quality estimation models and minimum Bayes risk decoding were used to rerank the n-best list so that the best hypothesis is chosen according to the COMET evaluation metric.

F.5 ARC-NKUA (Roussis and Papavassiliou, 2022)

The ARC-NKUA submission to the WMT22 General Machine Translation shared task concerns the unconstrained tracks of the English-Ukrainian and Ukrainian-English translation directions. The 2 Neural Machine Translation systems are based on Transformer models and our primary submissions were determined through experimentation with (a) checkpoint averaging, (b) ensemble decoding, (c) continued training with a subset of the training data, (d) data augmentation with back-translated monolingual data, and (e) post-processing of the translation outputs. We used various techniques to clean and filter the data provided by the organizers, as well as the additional parallel and monolingual data which we acquired from various sources.

F.6 CUNI-Bergamot (Jon et al., 2022)

CUNI-Bergamot submission is based on block-backtranslation method and MBR decoding using neural metrics. Block-BT is a method which switches between blocks of authentic parallel and backtranslated data during training based on a predefined pattern. The paper compares various parameters of the block-BT method: block size, checkpoint averaging methods, using only BT or also forward translation. The authors also show that MBR decoding can profit from more diverse checkpoints created by this method, as opposed to traditional mixed data training.

F.7 CUNI-DocTransformer (Jon et al., 2022)

Exactly the same as submitted in WMT20 (Popel, 2020), document-level Transformer trained with Block Backtranslation.

F.8 CUNI-Transformer (Jon et al., 2022)

The English↔Czech sentence-level models are exactly the same as submitted in WMT20 (Popel, 2020). The Ukrainian↔Czech models are very similar, also trained with Block Backtranslation. The Czech→Ukrainian system uses in addition special preprocessing (romanization of the Ukrainian side and a novel vocabulary-based inline casing on both sides).

F.9 CharlesTranslator (Popel et al., 2022)

Charles Translator for Ukraine is a free Czech-Ukrainian online translation service available for the public at `https://translator.cuni.cz` and as an Android app. It was developed at Charles University in March 2022 to help refugees from Ukraine by narrowing the communication gap between them and other people in the Czech Republic. It is based on Transformer and Block Backtranslation (Popel et al., 2020a).

F.10 DLUT (no associated paper)

We participate in the WMT 2022 general translation task in 2 language pairs and four language directions, English-Chinese and English-Japanese. Our submission use standard Transformer bilingual models.

We mainly improve performance by data filtering, large-scale data generation (i.e., back-translation, forward-translation, knowledge distillation, R2L training), domain finetuning, model ensemble and post-editing.

F.11 GTCOM (Zong and Bei, 2022)

This submission is based on Transformer architecture and involves data augmentation techniques.

F.12 HuaweiTSC (Wei et al., 2022)

This paper describes the submission of huawei translation services center (HW-TSC) to WMT22 general MT translation task.

F.13 JDExploreAcademy (Zan et al., 2022)

We push the limit of our previous work – bidirectional training (Ding et al., 2021) for machine translation by scaling up two main factors, i.e. language pairs and model sizes, namely the Vega-MT system. As for language pairs, we scale the "bidirectional" up to the "multidirectional" settings, covering all competitive high-resource languages, including en-de, en-cs, en-ru, en-zh, and en-ja, to exploit the common knowledge across languages, and transfer them to the downstream bilingual tasks. As for model size, we scale the transformer-big up to the extremely large model that owns nearly 4.7 Billion parameters, to fully enhance the model capacity for our Vega-MT. Also, we adopt the widely-used data augmentation strategies, e.g. back translation, knowledge distillation, cycle translation, and bidirectional self-training to comprehensively exploit the bilingual and monolingual data. To adapt our Vega-MT to the general domain test set, the noisy channel reranking and generalization tuning are employed.

F.14 KYB (Kalkar et al., 2022)

KYB team participated in the WMT22 general machine translation task on English-to-Japanese and Japanese-to-English directions. Our submissions are based on the transformer model with base setting. We employed several techniques to improve system's performance, such as data cleaning and selection, model ensembling/averaging, beam search, fine-tuning, and post-processing.

F.15 LT22 (Malli and Tambouratzis, 2022)

Our submission consists of translations produced from a series of NMT models of the following two language pairs: german-to-english and german-to-french. All the models are trained using only the parallel training data specified by WMT22. The models follow the transformer architecture employing eight attention heads and six layers in both the encoder and decoder. It is also worth mentioning that, in order to limit the computational resources that we would use during the training process, we decided to train the majority of models by limiting the training to 21 epochs. Moreover, the translations submitted at WMT22 have been produced using the test data released by the WMT22. The aim of our experiments has been to evaluate methods for cleaning-up a parallel corpus to determine if this will lead to a translation model producing more accurate translations. For each language pair, the base NMT models have been trained from raw parallel training corpora, while the additional NMT models have been trained with corpora subjected to a special cleaning process with the following tools: Bifixer and Bicleaner. It should be mentioned that the Bicleaner repository doesn't provide pre-trained classifiers for the above language pairs, consequently we trained probabilistic dictionaries in order to produce new models. The fundamental differences between these NMT models produced are mainly related to the quality and the quantity of the training data, while there are very few differences in the training parameters. To complete this work, we used the following three tools:(i) MARIAN NMT (Version: v1.11.5), which was used for the training of the NMT models and (ii) Bifixer and (iii) Bicleaner, which were used in order to correct and clean the parallel training data. Concerning the Bifixer and Bicleaner tools, we followed all the steps as described meticulously in the relevant article.

F.16 Lan-Bridge (Han et al., 2022)

Team Lan-Bridge's submission are transformer base models. For non-Chinese language pairs, we trained some multilingual models. For Chinese-English and English-Chinese, we train seperated models for each direction.

F.17 LanguageX (Zeng, 2022)

LanguageX submission is an ensemble model equipped with our recent technique of fast domain adaptation and data selection.

F.18 Liv4ever (Rikters et al., 2022)

The submitted translations were generated by an ensemble of three different iterations of multi-lingual transformer models trained on Latvian, Estonian, English and Livonian data from the constrained track. All parallel data were filtered (**?**) before training. After initial training the models were further improved by performing iterative back-translation of batches of 200,000 sentences from each language to the other languages (Livonian monolingual data was upscaled) for four iterations. The ensemble was composed of the single best checkpoint from the last three iterations of the back-translation process.

F.19 NAIST-NICT-TIT (Deguchi et al., 2022)

This paper describes the NAIST-NICT-TIT submission to the WMT22 general machine translation task. We participated in this task in the English-Japanese language pair. Our system is built on an ensemble of Transformer big models, k-nearest-neighbor machine translation (kNN-MT) (Khandelwal et al., 2021), and reranking.

Our base translation system is a combination of kNN-MT and an ensemble of four Transformer big models. Each of the Transformer model instances is trained using a different random seed, and we reuse one of the models for kNN-MT. A notable point of our system is that we construct the datastore for kNN-MT from back-translated monolingual data. We find that using the back-translated data improves translation performance when compared to using a parallel training corpus for the datastore.

We designed a reranking system to select a sentence from among the n-best sentences generated by the base translation system. For each translation hypothesis, the reranker computes a weighted sum of multiple model scores. It then selects the hypothesis with the highest score. We used k-best batch MIRA (Cherry and Foster) to select the weights for the model scores that maximize the BLEU score of the development set. We use context-aware model scores to improve the document-level consistency of the translation.

F.20 NT5 (Morishita et al., 2022)

The NT5 team submission is standard ensemble Transformer models equipped with several extensions, including our recent techniques, followed by a reranking module based on source-to-target, target-to-source, and masked language models. We also applied data augmentation and selection techniques to training data of the Transformer models.

F.21 NiuTrans (Shan et al., 2022)

This paper describes NiuTrans neural machine translation systems of the WMT22 General MT task with constrained data sets. We participated in Chinese to English, English to Croatian, and Livonian-English total of three tasks. We mainly utilized iterative back-translation, iterative knowledge distillation, and iterative fine-tuning. We also use various Transformer variants to improve the model's performance further, e.g., ODE-Transformer, UMST. Moreover, we tried some multi-domain methods, such as multi-domain model structure and multi-domain data clustering method, to adapt to this year's multi-domain test set. We also tried some methods to build a machine translation system using pre-trained language models.

F.22 OpenNMT (no associated paper)

In this paper, we first benchmark the mainstream translators on the English-to-German task by making sure we take into account: - The changes that occurred in the WMT test sets starting 2019 - The post-processing

differences between systems - The recent research in automatic metrics beyond BLEU Over the past 3 years, WMT has shown that both OnlineW and FacebookAI have a clear lead in the human evaluations. When looking at various metrics, we make the assumptions that one reason comes from the very good fluency which exposes a low perplexity when measuring with a GPT-2 language model.

We will therefore try 3 types of experiments: 1) filter various datasets with a GPT-2 model to retain only sentences under a given threshold. 2) Use a noisy channel decoding reranking method (used by FacebookAI) and maybe by OnlineW since their API is way slower then G/M/A. 3) Use a GPT-2 large model distillation during NMT training.

Given the training time of the last experiment we were not able to submit this system, however we will continue and report results in the paper.

F.23 PROMT (Molchanov et al., 2022)

The PROMT systems are trained with the MarianNMT toolkit. All systems use the transformer-big configuration. We use BPE for text encoding, the vocabulary sizes vary from 24k to 32k for different language pairs. All systems are unconstrained. We use all data provided by the WMT organizers, all publicly available data and some private data.

F.24 SRPOL (Dobrowolski et al., 2022)

We present the work of Samsung R&D Institute Poland in WMT 2022 General MT solution for medium to low resource languages: Russian and Croatian. Our approach combines iterative back-translation with noise and iterative distillation. We investigated different monolingual resources and compared their effects on the final translation. We used available BERT-like models to classify texts and to distinguish text domains. We attempted to predict ensemble weight vectors based on BERT-like domain classification for individual sentences. The final models achieved quality comparable to the best online translators using only limited resources during training.

F.25 TAL-SJTU (He et al., 2022)

TAL-SJTU submission is based on M2M100 (Fan et al., 2021a) with novel techniques that adapt it to the target language pair: (1) We propose a cross-model word embedding alignment method that transfers a pre-trained word embedding to M2M100, enabling it to support Livonian. (2) We also utilize Estonian and Latvian languages as auxiliary languages for training and pivot languages for data augmentation. (3) Finally, the best result was achieved after fine-tuning the model using the validation set and online back-translation. In model evaluation: (1) We find that previous work (Rikters et al., 2022) underestimated the translation performance of Livonian due to inconsistency in Unicode normalization, which may cause a discrepancy of up to 19 BLEU score. (2) In addition to the standard validation set, we also employ round-trip BLEU to evaluate the models, which we find a more appropriate way for this task.

F.26 TartuNLP (Tars et al., 2022)

TartuNLP's submission is a model based on Transformers. Our main approach was utilizing large pre-trained multilingual neural machine translation models, specifically the M2M-100 model (Fan et al., 2021b). In our systems we used the 1.2 billion parameter model. We fine-tuned the pre-trained model (more specifically we performed cross-lingual transfer learning) to our data, which consisted of WMT22 liv-en, en-liv data and other data from the Finno-Ugric language family for support. The main pipeline was the following: fine-tuning with original parallel data, then two iterations of back-translation and finally fine-tuning on original parallel data again.

F.27 eTranslation (Oravecz et al., 2022)

eTranslations's Fr-De system is an ensemble of 4 big transformers, trained from all available parallel data and with additional tagged, back-translated data generated from a 30M subset of various German monolingual corpora. The monolingual and original parallel data is cleaned up and filtered with heuristic rules. In the model trainings, the original parallel data is upsampled to a 1:1 ratio. Each transformer model is then fine tuned for 3 epochs on the original parallel data. The models use a 32k SentencePiece

vocabulary. The SentencePiece module as built in the Marian toolkit is used for end-to-end text processing, without the standard pre- and postprocessing steps of truecasing, or (de)tokenization.

The En-Uk system is an ensemble of 4 multilingual (En -> Uk, Ru) big transformers, trained from all available parallel data. Each transformer model is then fine tuned only on the En-Uk data for about 50 epochs and the best checkpoint is used in the ensemble. Vocabulary and pre/postprocessing settings are the same as the Fr-De system. The En-Ru system is built with the same setup as the En-Uk, except it is an ensemble of 3 models.

F.28 manifold (Jin et al., 2022)

Manifold's English-Chinese System at WMT22 is an ensemble of 4 models, each trained by one of four different configurations and fine-tuned by applying scheduled-sampling. The four configurations are DeepBig (Xenc), DeepLarger (Xenc), DeepBigTalkingHeads (Xenc) and DeepBig (LaBSE). DeepBig is an extension to TransformerBig, the only difference is the former has 24 encoder layers. DeepLarger has 20 encoder layers and its FFN dimension is 8192. *TalkingHead applies talking-heads trick. For Xenc configs, we selected monolingual and parallel data that is similar to the past newstest datasets using Xenc, and for LaBSE, we cleaned the officially provided parallel data using LaBSE pretrained model.

F.29 shopline-pl

The model we submitted is based on the query results of the transformer and its variants, which includes the integration effect of different models and incorporates the reserved word mechanism.

G Automatic scores

This section contains automatic metric scores. While human judgement is the official ranking of systems and their performance, we share automatic scores to show expected system performance for various testsets.

We use COMET (Rei et al., 2020) as the primary metric and ChrF (Popović, 2015) as the secondary metric, following recommendation by (Kocmi et al., 2021). We present BLEU (Papineni et al., 2002) scores as it is still widely used metric. The COMET scores are calculated with the default model `wmt20-comet-da`. The ChrF and BLEU scores are calculated using SacreBLEU with signature (Post, 2018) is `chrF2|nrefs:all|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0`. Scores are multiplied by 100.

The different suffix represents the name of reference used for calculation (A, B, C, stud), references has been translated by different translators but with the same sponsor. A notable difference is Czech-English, where we are missing reference "A" for it's low quality, which was partly corrected and placed under "C". The second exception is Croatian reference "stud" which was created by students in contrast to "A" prepared by professionals. Lastly, testsets liv-en and ru-sah are reverse testsets to their opposite counterparts (i. e. "en" and "sah" are original sources)

Table 29: Automatic metric scores for en-cs.

System	COMET$_B$ ↑	COMET$_C$	ChrF$_B$	ChrF$_C$	BLEU$_B$	BLEU$_C$
Online-W	97.8	79.3	68.2	51.8	45.8	25.0
Online-B	97.5	76.6	69.0	52.7	48.2	27.0
CUNI-Bergamot	96.0	79.0	63.2	50.3	38.6	24.4
JDExploreAcademy	95.3	77.8	65.1	51.8	41.4	25.5
Lan-Bridge	94.7	73.8	68.2	52.3	45.6	25.9
Online-A	92.2	71.1	65.8	50.8	41.8	24.5
CUNI-DocTransformer	91.7	72.2	63.9	50.8	39.8	25.2
CUNI-Transformer	86.6	68.6	62.1	50.1	37.7	24.5
Online-Y	83.7	62.3	62.9	49.0	37.8	22.8
Online-G	82.3	61.5	62.8	49.0	38.1	22.7

Table 30: Automatic metric scores for en-de.

System	COMET$_A$ ↑	COMET$_B$	ChrF$_A$	ChrF$_B$	BLEU$_A$	BLEU$_B$
Online-W	65.5	64.4	64.1	62.7	36.6	35.3
JDExploreAcademy	63.2	62.5	64.3	63.8	37.8	38.2
Online-B	62.3	61.9	64.6	64.1	38.4	38.3
Online-Y	61.1	60.9	63.7	63.5	37.0	37.2
Online-A	60.6	60.0	63.9	63.6	36.5	37.2
Online-G	60.2	59.3	63.4	63.1	36.4	36.6
Lan-Bridge	58.8	58.3	64.1	63.7	36.1	36.5
OpenNMT	57.2	57.0	62.1	61.5	35.7	35.7
PROMT	55.8	55.3	62.8	62.2	36.1	36.0

Table 31: Automatic metric scores for en-hr.

System	COMET$_A$ ↑	COMET$_{stud}$	ChrF$_A$	ChrF$_{stud}$	BLEU$_A$	BLEU$_{stud}$
Online-B	80.4	77.6	58.5	57.6	31.5	29.8
Lan-Bridge	79.6	76.7	58.5	57.4	31.5	29.7
GTCOM	77.4	74.7	58.1	57.0	30.7	28.6
Online-A	69.5	67.1	56.5	55.9	29.1	28.1
SRPOL	69.4	67.6	56.3	55.6	29.1	27.8
HuaweiTSC	67.6	66.3	56.8	56.1	29.9	28.6
NiuTrans	65.5	63.4	56.3	55.6	29.3	28.1
Online-G	64.2	63.0	53.2	52.5	25.7	24.3
Online-Y	56.7	55.1	54.3	53.6	26.6	25.1

Table 32: Automatic metric scores for en-ja.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
JDExploreAcademy	65.1	36.1	41.5
NT5	64.1	36.8	42.5
LanguageX	62.1	36.1	41.7
Online-B	60.8	35.5	41.2
DLUT	60.5	36.1	41.8
Online-W	59.8	35.2	40.8
Online-Y	56.8	34.4	39.9
Lan-Bridge	56.5	34.1	39.4
Online-A	53.6	34.1	38.8
NAIST-NICT-TIT	53.3	33.8	39.2
AISP-SJTU	52.4	33.9	39.3
KYB	31.8	28.6	33.1
Online-G	24.9	28.0	32.1

Table 33: Automatic metric scores for en-liv.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
TAL-SJTU	-29.5	43.8	17.0
TartuNLP	-36.8	39.2	15.0
HuaweiTSC	-38.9	37.7	12.8
Liv4ever	-39.4	39.6	14.7
NiuTrans	-81.9	30.5	12.3

Table 34: Automatic metric scores for en-ru.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-W	75.1	58.3	32.4
Online-G	73.1	59.5	32.8
Online-B	72.9	59.7	34.9
Online-Y	69.8	58.3	33.2
JDExploreAcademy	69.6	58.4	32.7
Lan-Bridge	67.3	59.0	32.6
Online-A	67.3	58.1	33.1
PROMT	60.3	56.1	30.6
SRPOL	59.7	56.4	30.4
HuaweiTSC	59.2	56.1	30.8
eTranslation	57.9	55.8	29.8

Table 35: Automatic metric scores for en-uk.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-B	73.2	59.3	32.5
GTCOM	72.0	59.0	30.8
Online-G	69.9	57.2	27.2
Lan-Bridge	65.7	58.8	29.5
Online-A	60.9	56.0	28.0
eTranslation	54.5	54.8	26.2
HuaweiTSC	54.4	54.8	26.5
Online-Y	51.9	54.9	26.9
ARC-NKUA	49.2	54.0	25.2

Table 36: Automatic metric scores for en-zh.

System	COMET$_A$ ↑	COMET$_B$	ChrF$_A$	ChrF$_B$	BLEU$_A$	BLEU$_B$
GTCOM	64.7	69.4	44.1	45.7	47.7	50.5
LanguageX	63.8	71.5	49.1	53.1	54.3	59.8
Online-B	61.8	80.4	44.4	68.6	49.1	73.7
JDExploreAcademy	61.7	70.6	44.6	51.1	49.7	57.6
Lan-Bridge	61.4	69.4	42.8	49.2	48.3	56.0
Online-W	61.0	69.5	41.1	47.7	44.8	52.6
Manifold	60.1	71.2	44.2	54.3	48.7	59.6
Online-Y	59.7	71.7	42.3	54.0	46.8	59.9
HuaweiTSC	59.5	73.1	44.5	58.1	49.7	64.4
Online-A	57.3	70.1	42.5	55.5	46.4	60.7
AISP-SJTU	56.5	66.6	43.9	50.9	48.8	57.3
DLUT	52.1	63.0	41.3	50.1	45.2	55.4
Online-G	51.2	62.5	39.4	49.8	43.9	55.2

Table 37: Automatic metric scores for cs-en.

System	COMET$_B$ ↑	COMET$_C$	ChrF$_B$	ChrF$_C$	BLEU$_B$	BLEU$_C$
Online-W	77.5	45.6	79.3	52.0	64.2	23.8
JDExploreAcademy	74.7	49.0	74.4	53.7	54.9	25.1
Lan-Bridge	71.8	47.2	74.0	54.0	54.5	25.5
Online-B	71.8	47.4	73.8	54.0	54.3	25.5
CUNI-DocTransformer	70.6	45.3	72.2	53.0	51.9	24.8
Online-A	69.8	44.3	73.4	53.4	53.3	25.0
CUNI-Transformer	69.2	43.2	71.7	52.0	51.6	23.9
Online-G	63.0	38.8	70.3	52.1	48.5	23.0
SHOPLINE-PL	61.1	39.6	69.2	53.2	46.8	24.6
Online-Y	58.6	35.2	67.9	51.5	44.6	23.1
ALMAnaCH-Inria	19.3	4.9	56.9	48.3	29.9	19.7

Table 38: Automatic metric scores for de-en.

System	COMET$_A$ ↑	COMET$_B$	ChrF$_A$	ChrF$_B$	BLEU$_A$	BLEU$_B$
JDExploreAcademy	58.0	63.5	58.5	61.8	33.7	35.8
Online-B	56.9	63.6	58.3	61.9	33.3	36.6
Lan-Bridge	56.5	63.6	58.5	62.3	33.4	37.0
Online-G	55.2	61.7	58.7	62.5	33.7	36.5
Online-Y	54.6	61.4	58.0	61.9	32.9	36.3
Online-A	54.5	62.2	58.4	62.7	33.3	37.2
Online-W	54.3	61.7	57.7	61.7	32.6	36.0
PROMT	51.8	59.4	57.8	62.1	32.5	36.6
LT22	25.6	33.3	51.3	55.7	26.0	30.9

Table 39: Automatic metric scores for ja-en.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
NT5	42.0	51.3	26.6
Online-W	41.2	51.7	27.8
JDExploreAcademy	40.6	50.1	25.6
Online-B	39.6	49.9	24.7
DLUT	37.2	49.8	24.8
NAIST-NICT-TIT	33.4	48.3	22.7
Online-A	32.9	48.4	22.8
LanguageX	32.9	49.1	22.4
Online-Y	32.3	48.2	21.5
Lan-Bridge	31.9	48.7	22.8
AISP-SJTU	30.1	48.0	22.0
Online-G	22.3	45.7	19.7
KYB	17.3	43.4	18.1
AIST	-152.7	11.4	0.1

Table 40: Automatic metric scores for liv-en.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
TartuNLP	-5.8	53.5	29.9
TAL-SJTU	-8.4	53.2	30.4
HuaweiTSC	-27.3	48.4	23.4
Liv4ever	-44.0	46.7	23.3
NiuTrans	-88.3	35.6	13.0

Table 41: Automatic metric scores for ru-en.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-G	65.1	70.0	46.7
JDExploreAcademy	64.9	68.9	45.1
Online-Y	64.1	68.2	43.8
Lan-Bridge	63.1	68.5	45.2
Online-B	63.1	68.3	45.0
Online-A	62.2	68.3	43.9
Online-W	61.6	66.3	42.6
HuaweiTSC	60.9	68.5	45.1
SRPOL	59.5	67.2	43.6
ALMAnaCH-Inria	26.8	57.9	30.3

Table 42: Automatic metric scores for uk-en.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-B	62.5	67.2	44.4
Lan-Bridge	62.4	67.3	44.6
GTCOM	61.9	67.1	43.9
Online-G	57.4	66.0	43.2
Online-A	52.1	65.2	42.3
HuaweiTSC	50.1	63.9	41.6
Online-Y	49.8	64.6	41.8
PROMT	49.6	64.7	42.1
ARC-NKUA	49.6	64.6	41.9
ALMAnaCH-Inria	21.8	55.6	30.0

Table 43: Automatic metric scores for zh-en.

System	COMET$_A$ ↑	COMET$_B$	ChrF$_A$	ChrF$_B$	BLEU$_A$	BLEU$_B$
Online-G	45.6	36.2	59.7	54.1	29.6	21.7
JDExploreAcademy	45.1	35.2	61.1	54.1	33.5	22.3
LanguageX	44.9	35.3	60.5	54.2	31.9	22.1
Lan-Bridge	43.0	34.0	57.8	52.7	28.1	20.9
HuaweiTSC	42.8	33.5	58.5	52.8	29.8	21.7
Online-B	42.1	32.8	58.2	52.9	28.8	21.1
AISP-SJTU	41.6	32.8	59.2	53.8	29.7	21.4
Online-Y	40.8	31.0	57.6	52.1	27.1	19.8
Online-A	35.2	26.0	57.3	52.1	27.3	19.9
Online-W	31.6	23.1	54.5	49.9	24.0	18.0
NiuTrans	31.3	22.3	56.0	51.2	26.2	19.5
DLUT	30.6	22.0	55.2	50.5	25.0	18.6

Table 44: Automatic metric scores for cs-uk.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
AMU	99.4	61.5	34.7
Online-B	94.3	64.0	38.3
GTCOM	93.4	63.9	36.8
Lan-Bridge	91.8	64.0	38.3
CharlesTranslator	90.8	61.5	34.3
HuaweiTSC	90.7	62.6	36.0
CUNI-JL-JH	90.0	61.6	34.8
Online-G	88.3	60.8	32.5
Online-A	87.8	62.2	35.9
CUNI-Transformer	87.3	61.6	35.0
Online-Y	78.4	59.6	32.1
ALMAnaCH-Inria	61.3	54.5	26.8

Table 45: Automatic metric scores for de-fr.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-B	70.5	74.6	58.4
Online-W	63.6	65.5	43.6
Online-Y	57.8	66.8	46.2
Online-A	52.2	64.5	41.3
Online-G	44.8	62.7	39.0
LT22	10.4	54.4	28.3

Table 46: Automatic metric scores for fr-de.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-W	77.9	81.2	64.8
Online-B	63.7	68.7	46.6
Online-Y	61.6	67.5	45.0
Online-A	59.2	67.2	44.4
eTranslation	55.4	68.4	46.5
Lan-Bridge	51.1	65.0	41.8
Online-G	48.2	66.0	41.1

Table 47: Automatic metric scores for ru-sah.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-G	-17.1	47.0	14.7
Lan-Bridge	-124.3	11.3	0.0

Table 48: Automatic metric scores for sah-ru.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
Online-G	31.1	55.5	29.6
Lan-Bridge	-75.9	28.3	7.1

Table 49: Automatic metric scores for uk-cs.

System	COMET$_A$ ↑	ChrF$_A$	BLEU$_A$
AMU	104.8	60.7	37.0
Online-B	96.5	60.3	36.4
Lan-Bridge	94.5	60.4	36.5
HuaweiTSC	91.4	59.6	36.0
CharlesTranslator	90.2	59.0	35.9
CUNI-JL-JH	89.0	58.7	35.1
CUNI-Transformer	88.5	59.0	35.8
Online-A	85.4	57.5	33.3
Online-G	84.2	56.3	31.5
GTCOM	80.2	55.8	31.3
Online-Y	78.6	55.3	29.6
ALMAnaCH-Inria	62.4	50.7	25.3