HumEval'23 Reproduction Report for Paper 0040: Human Evaluation of Automatically Detected Over- and Undertranslations

Filip Klubička
ADAPT Centre
Technological University Dublin
filip.klubicka@adaptcentre.ie

John D. Kelleher
ADAPT Centre
Maynooth University
john.kelleher@mu.ie

Abstract

This report describes a reproduction of a human evaluation study evaluating automatically detected over- and undertranslations obtained using neural machine translation approaches. While the scope of the original study is much broader, a human evaluation is included as part of its system evaluation. We attempt an exact reproduction of this human evaluation, pertaining to translations on the the English-German language pair. While encountering minor logistical challenges, with all the source material being publicly available and some additional instructions provided by the original authors, we were able to reproduce the original experiment with only minor differences in the results.

1 Introduction

This report presents a reproduction of a human evaluation originally conducted and presented in the paper *As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning* (Vamvas and Sennrich, 2022). The paper proposes an approach for detecting over- and under-translations using *contrastive conditioning* (Vamvas and Sennrich, 2021), a method that relies on hypothetical reasoning over the likelihood of partial sequences and thus has the advantage of not requiring access to the original translation system or to a quality estimation model. The authors evaluate their system based on real machine translations and show that the approach outperforms a supervised baseline in the detection of omissions.

While the scope of their original study is much broader, a human evaluation is included as part of the system evaluation and is described in Section 5.2 of their paper. In this evaluation step, the original authors employ expert annotators to determine whether the spans of text that their system predicts as mistranslated are indeed under- or overtranslations, and do this on the English-German and English-Chinese language pairs. In our reproduction study, we attempt to reproduce the evaluations of the English-German data, by employing expert annotators to evaluate the same data samples.

This reproduction study was conducted as part of the ReproHum project[1] (Belz et al., 2023), the aim of which is to build on existing work on recording properties of human evaluations datasheet-style (Shimorina and Belz, 2022), and assessing how close results from a reproduction study are to the original study (Belz et al., 2022), to systematically investigate what factors make human evaluations more—or less— reproducible. Our choice to reproduce this particular paper is motivated by our previous experience in related fields: both authors have previously worked in the space of machine translation (Popovic et al., 2023; Moslem et al., 2023; Klubička et al., 2022; Bago et al., 2022; Moslem et al., 2022; Toral et al., 2017; Popović et al., 2016; Salton et al., 2014a), have a track record of interest in human evaluation (Klubička et al., 2018b,a; Klubička et al., 2017; Salton et al., 2014b) and reproducibility (Klubička and Fernández, 2018), and are thus well-positioned to conduct this reproduction experiment.

2 Original Study Design

For the English-German language pair, the original study employed two linguistic experts as evaluators. As their annotation interface, the authors opted for Doccano[2] (Nakayama et al., 2018), an open-source text annotation tool which provides annotation features for text classification, sequence labeling, and sequence to sequence tasks. Each expert evaluator was shown 80+720 (dev+test set) randomly sampled positive predictions across both types of coverage errors. Evaluators were shown

[1] https://reprohum.github.io
[2] https://github.com/doccano/doccano

Proceedings of the Third Workshop on Human Evaluation of NLP Systems, pages 153–189, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-088-5_014

the source sequence, the machine translation, and the predicted error span. They were asked whether the highlighted span was indeed translated badly, and were asked to perform a fine-grained analysis based on a list of predefined answer options (see Appendix A). A subset of the samples (100 sentences) was annotated by both raters in order to calculate inter-annotator agreement.

The authors made all predictions, annotations and notebooks used for calculating the precision values available in the GitHub repository[3].

3 Reproduction Study Details

We used the exact same dataset as provided by (Vamvas and Sennrich, 2022) and had each annotator annotate the same set of instances as provided by the original authors. Once we obtained the evaluations, we used the original authors' evaluation script, as provided on their GitHub page. It is worth noting that during the reproduction phase, another team reproducing the same experiment noticed a possible bug in the authors' results processing script. After communication via the ReproHum team, the issues were clarified and corrected, and the authors uploaded a revised script to fix one of the errors that arose. The updated script is also included in their GitHub page and is the one we used for result processing[4].

3.1 Evaluators

Our selection criteria for evaluators required them to be proficient in German and English, with a background in linguistics or (machine) translation, which are all crucial for evaluating a MT-based task on the two languages. The evaluators were recruited via a colleague who teaches a translation studies course and highly recommended them as exceptional students in the course. They are both native German speakers who are fluent in English, currently attending a translation studies course in Ireland.

We sent the evaluators the annotation instructions and had an initial meeting to clear up any questions or uncertainties. We then gave them the smaller development sample to annotate to give them hands-on experience with the task and clear up any confusion that might arise. After this step they were given the full test set for annotation, but were told that they can ask any practical questions should they arise, but should not communicate with each other or ask for opinions on how to annotate questionable instances, but should rely on their own judgement.

We estimated that the annotation would take about 10 hours of work, which turned out to be the case and was consistent with the original authors' experience. Given that participants were paid during the original experiment, we aimed to do the same by following the shared ReproHum procedure for calculating fair pay. However, as the original study was conducted in Switzerland where a minimum wage is not defined, we opted to simply match the rates paid to the evaluators of the original experiment and paid our annotators the equivalent amount in euros, at a rate of €30/hour. This also exceeds the minimum wage in Ireland and would be considered fair pay for an annotation task.

3.2 Differences

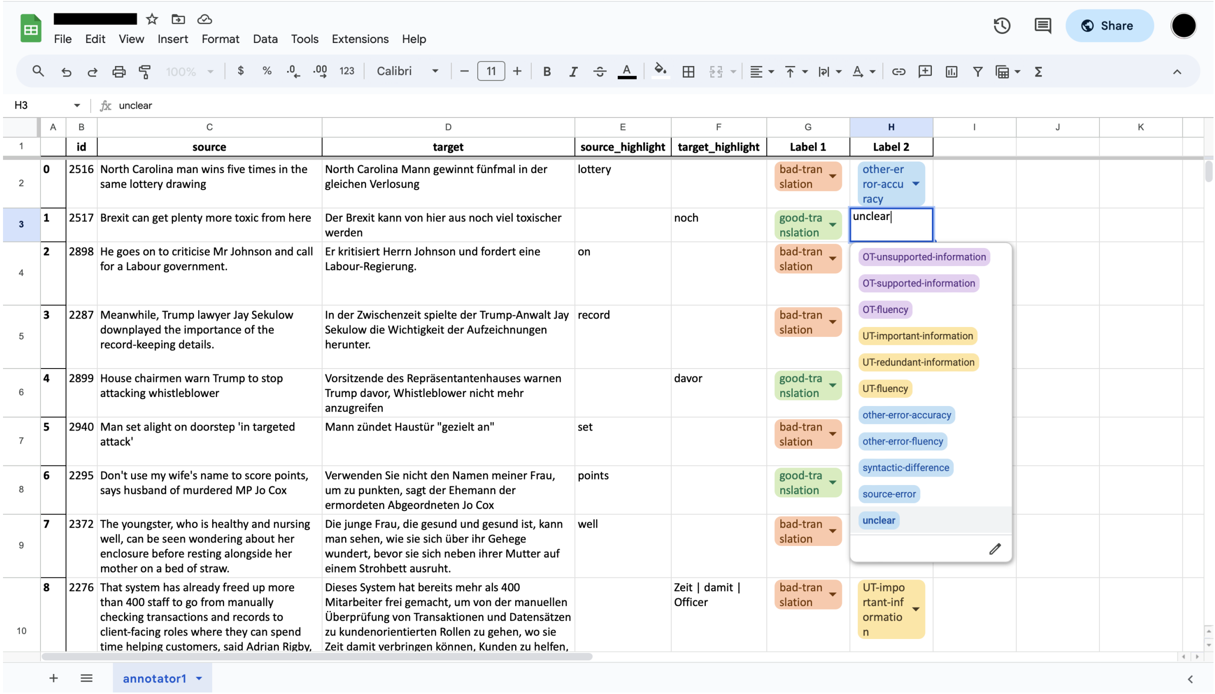
Regarding the choice of annotation interface, we attempted to deploy Doccano to a virtual machine so that the participants could access the application over the web, just as the original authors had. However we faced a number of technical challenges in setting this up, and after a number of attempts had to abandon this direction. The original authors had noted that it is not strictly necessary to use a web application for the annotation, and give liberty to use other methods such as a spreadsheet. Given our difficulties with setting up Doccano, we opted for the spreadsheet option.

Specifically, we used the Google Sheets application and created a separate sheet that contained the data for each annotator individually. This approach made it straightforward to set up and more accessible to the annotators, as it was a familiar interface to them. The annotators were presented with a source sentence, target sentence, the candidate spans in the source and target sentence, and two drop-down menus to select annotation labels, in line with the original study's annotation guidelines. Additionally, we colour-coded the different error categories to reduce the cognitive load of choose from the many possible options. Image 1 shows the annotation interface.

In order to transform the data into the spread-

[3] https://github.com/ZurichNLP/coverage-contrastive-conditioning/tree/master/evaluation/human_evaluation

[4] https://github.com/ZurichNLP/coverage-contrastive-conditioning/blob/master/evaluation/human_evaluation/Human%20Evaluation%20ENDE.v2.ipynb


[redacted]

H3 unclear

	id	source	target	source_highlight	target_highlight	Label 1	Label 2
0	2516	North Carolina man wins five times in the same lottery drawing	North Carolina Mann gewinnt fünfmal in der gleichen Verlosung	lottery		bad-translation	other-error-accuracy
1	2517	Brexit can get plenty more toxic from here	Der Brexit kann von hier aus noch viel toxischer werden		noch	good-translation	unclear
2	2898	He goes on to criticise Mr Johnson and call for a Labour government.	Er kritisiert Herrn Johnson und fordert eine Labour-Regierung.	on		bad-translation	
3	2287	Meanwhile, Trump lawyer Jay Sekulow downplayed the importance of the record-keeping details.	In der Zwischenzeit spielte der Trump-Anwalt Jay Sekulow die Wichtigkeit der Aufzeichnungen herunter.	record		bad-translation	
4	2899	House chairmen warn Trump to stop attacking whistleblower	Vorsitzende des Repräsentantenhauses warnen Trump davor, Whistleblower nicht mehr anzugreifen		davor	good-translation	
5	2940	Man set alight on doorstep 'in targeted attack'	Mann zündet Haustür "gezielt an"	set		bad-translation	
6	2295	Don't use my wife's name to score points, says husband of murdered MP Jo Cox	Verwenden Sie nicht den Namen meiner Frau, um zu punkten, sagt der Ehemann der ermordeten Abgeordneten Jo Cox	points		good-translation	
7	2372	The youngster, who is healthy and nursing well, can be seen wondering about her enclosure before resting alongside her mother on a bed of straw.	Die junge Frau, die gesund und gesund ist, kann man sehen, wie sie sich über ihr Gehege wundert, bevor sie sich neben ihrer Mutter auf einem Strohbett ausruht.	well		bad-translation	
8	2276	That system has already freed up more than 400 staff to go from manually checking transactions and records to client-facing roles where they can spend time helping customers, said Adrian Rigby,	Dieses System hat bereits mehr als 400 Mitarbeiter frei gemacht, um von der manuellen Überprüfung von Transaktionen und Datensätzen zu kundenorientierten Rollen zu gehen, wo sie Zeit damit verbringen können, Kunden zu helfen,		Zeit | damit | Officer	bad-translation	UT-important-information

Figure 1: Screenshot of the annotation interface shown to the evaluators.

sheet annotation interface we had to extract it from the .jsonl format it was provided in. Additionally, given that the original authors' evaluation script relies on the .jsonl data format that is output by Doccano, we also had to convert the annotations from the spreadsheet format back to the required format. It was clear this conversion would be necessary once we opted for the spreadsheet-based approach, and performing the conversion was fairly straightforward, but still made for an added processing step which was not noted anywhere in the reproduction guidelines.

4 Reproduction Results

For the human evaluation aspect, the original paper reports three sets of results: (a) a table containing word-level precision scores of the spans that were highlighted by their automatic approach, based on the human evaluations (Table 2 in the original paper), (b) plots that display the results for the human evaluation of predicted addition and omission errors (Appendix G in the original paper) and (c) Cohen's Kappa scores for inter-annotator agreement (mentioned in the body of Section 5.2 of the original paper).

Above results (a) and (b) fall under **Type I** results as defined in the ReproHum reproduction guidelines, given that they are numerical error counts or precision calculations. Results (c) fall under **Type III**, as they are multi-rater categorical labels attached to text spans.

It should be noted that regarding (a), the original paper does not seem to mention how these precision values are calculated, nor does such a calculation seem to be included in the authors' annotation processing script or reproducibility guidelines, making these results difficult to reproduce without relying on guesswork.

Regarding (b), while the plots presented in the paper are indicative of general trends, precise error counts are difficult to infer from the graphics alone. Fortunately the authors do provide the full annotated data and the exact output of the calculations as part of the notebook on their GitHub page. The same notebook also includes a calculation for (c), making both (b) and (c) straightforward to reproduce. One could argue that the error counts and the Cohen's Kappa are the core reproduction values, as they constitute the raw outputs of the human evaluation. Tables 1 and 2 show the original values provided by (Vamvas and Sennrich, 2022) and our reproduced values side by side. It is worth noting that the original values were not provided in the paper itself, but rather in supplementary material, specifically the notebook on the original author's GitHub page (which is still publicly accessible, but requires some digging to acquire the data).

Type	Label 1	Original	Reproduced
OT	bad translation	54	67
OT	good translation	644	640
UT	bad translation	251	228
UT	good translation	382	418
Type	Label 2	Original	Reproduced
OT	bad+OT-supported-info	10	1
OT	bad+OT-unsupported-info	5	11
OT	bad+UT-important-info	0	19
OT	bad+UT-redundant-info	0	2
OT	bad+other-accuracy	32	28
OT	bad+other-fluency	7	3
OT	good+OT-fluency	117	77
OT	good+OT-supported-info	20	13
OT	good+UT-fluency	0	11
OT	good+UT-redundant-info	0	5
OT	good+syntactic-diff	455	443
OT	good+unclear	52	85
UT	bad+OT-supported-info	0	0
UT	bad+OT-unsupported-info	0	2
UT	bad+UT-important-info	120	109
UT	bad+UT-redundant-info	111	45
UT	bad+other-accuracy	17	61
UT	bad+other-fluency	3	11
UT	good+OT-fluency	0	4
UT	good+OT-supported-info	0	0
UT	good+UT-fluency	72	101
UT	good+UT-redundant-info	25	55
UT	good+syntactic-diff	260	198
UT	good+unclear	25	56

Table 1: Error annotation counts broken down by error type, comparing originally reported values (after the minor bug fix) and our own reproduced values.

Labels	Oκ	Rκ
Question 1	0.56	0.58
Question 1+2	0.33	0.46

Table 2: Cohen's Kappa values for inter-annotator agreement, comparing (O) originally reported values (after the minor bug fix) and (R) our own reproduced values.

4.1 Findings Comparison

The original results presented in the paper by (Vamvas and Sennrich, 2022) find that **(a)** fine-grained answers allow to quantify the word-level precision of the spans highlighted by their approach, both with respect to coverage errors in particular and to translation errors in general; **(b)** precision is higher than expected when detecting omission errors in English–German translations, but is still low for additions; **(c)** the distribution of the detailed answers suggests that syntactical differences between the source and target language contribute to the false positives regarding additions; **(d)** many of the predicted error spans are in fact translation errors, but not coverage errors in a narrow sense–e.g. more than 10% of the spans marked in English–German translations were classified by their raters as a different type of accuracy error, such as mistranslation.

Note that the authors frame their core findings as pertaining to the precision results, which they did not provide a way to calculate, so we are not able to verify their claims. They also do not go into detail discussing the distribution of human evaluations themselves, and say little about the obtained inter-annotator agreements. This is understandable, as the human annotation was only a small fraction of their work, but consequently there are few findings for us to compare in this regard. Still, we are able to note that based on the distribution of error types our annotators have achieved a similar distribution of errors on the same data, and have achieved comparable agreement on Label 1 (good/bad translation), while also having somewhat higher agreement on Label 1+2 than in the original study.

5 Conclusion

While we were not able to reproduce the core findings on model precision due to lack of information, we did manage to achieve similar Cohen's Kappa scores for our annotator agreement on one question, and a somewhat higher score on the more difficult question. We also reproduced the distribution of labels on Question 1 and on most categories in Question 2.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreements No. 13/RC/2106 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin and Maynooth University. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme, and is co-funded under the European Regional Development Fund.

References

Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, Niels Runar Gislason, Andre Kasen, Filip Klubička, Gauti Kristmannsson, Helen McHugh, Roisin Moran, Orla Ni Loinsigh, Jon Arild Olsen, Carla Parra Escartin, Akshai Ramesh, Natalia Resende, Paraic Sheridan, and Andy Way. 2022. Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced languages. *Revista de Llengua i Dret*, (78):9–34.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Filip Klubička, Lorena Kasunić, Danijel Blazsetin, and Petra Bago. 2022. Challenges of building domain-specific parallel corpora from public administration documents. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 50–55, Marseille, France. European Language Resources Association.

Filip Klubička, Giancarlo D. Salton, and John D. Kelleher. 2018a. Is it worth it? budget-related evaluation metrics for model selection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2018b. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*, 32(3):195–215.

Filip Klubička, Antonio Toral Ruiz, and M. Víctor Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Filip Klubička and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *Proceedings of 4REAL: 1st Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Maja Popović, Mihael Arčan, and Filip Klubička. 2016. Language related issues for machine translation between closely related South Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52, Osaka, Japan. The COLING 2016 Organizing Committee.

Maja Popovic, Vasudevan Nedumpozhimana, Meegan Gower, Sneha Rautmare, Nishtha Jain, and John Kelleher. 2023. Using mt for multilingual covid-19 case load prediction from social media texts. European Association for Machine Translation.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014a. An empirical study of the impact of idioms on phrase based statistical machine translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014b. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Antonio Toral, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way. 2017. Crawl and crowd to bring machine translation to under-resourced languages. *Language resources and evaluation*, 51:1019–1051.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

A Annotator Guidelines

Annotation Guidelines

Thank you for taking part in this annotation project – we appreciate it! In case of questions, feel free to reach out to Jannis Vamvas (vamvas@cl.uzh.ch) at any time.

Task Description

You will be shown a series of source sentences and translations. One or several spans in the text are highlighted and it is claimed that the spans are translated badly. You are asked to determine whether the claim is true.

The highlighted spans can be either in the source sequence or in the translation. If a span is in the source sentence, check whether is has been correctly translated. If a span is in the translation, check whether it correctly conveys the source.
Sometimes, multiple spans are highlighted. In that case, focus your answer on the span that is most problematic for the translation.

In a second step, you are asked to select an explanation. On the one hand, if you agree that the highlighted span is translated badly, please explain your reasoning by selecting your explanation. On the other hand, if you disagree and think that the span is well-translated, please select an explanation why the span might have been marked as badly translated in the first place.
Should multiple explanations be equally plausible, select the first plausible explanation from the top.

Annotation Interface

Please sign in and click on the annotation project named after you, e.g. “Jannis' Annotations”.
Click on the “Start Annotation” button.
You can use the arrow keys to move between samples, or the pagination on the upper right.
A sample is fully annotated if two labels have been selected. The first label is the general assessment (agree/disagree) and the second label is the explanation.
Your annotations are saved automatically.

Examples (English–German)

Examples for bad translations

The span contains information that is missing in the translation.
The government, reeling from low oil prices, says it hopes tourism will contribute up to 10 percent of the gross domestic product.
Die Regierung hofft, dass der Tourismus bis zu 10 Prozent des Bruttoinlandsprodukts ausmachen wird.

Other: The span is badly translated because of an accuracy error.
after millions of people joined a protest in the run-up to a U.N. climate summit.
... nachdem sich im Vorfeld eines Klimagipfels in den Vereinigten Staaten Millionen Menschen einem Protest angeschlossen hatten.

Other: The span is badly translated because of a fluency error.
after millions of people joined a protest in the run-up to a U.N. climate summit.
... nachdem sich im Vorfeld eines Vereinte Nationen Klimagipfels Millionen Menschen einem Protest angeschlossen hatten.

Examples for good translations

The span contains information that is missing in the translation but that can be inferred or is trivial.
... to ensure the country has an adequate supply of medical drugs.
... um sicherzustellen, dass das Land über eine ausreichende Versorgung mit Medikamenten verfügt.

The words in the span are redundant but fluent.
The way it was done ...
Die Art und Weise, wie es gemacht wurde, ...

The translation is syntactically different from the source.
During a conversation with the female tech founders ...
Während eines Gesprächs mit den Tech-Gründerinnen ...

Label explanations

bad-translation
- The span is badly translated.

good-translation
- The span is well translated.

OT-unsupported-information
- OverTranslation: The span adds unsupported information.
- applies only to bad translations

OT-supported-information
- OverTranslation: The span adds information that is supported by the context or trivial.
- applies to band and good translations

OT-fluency
- OverTranslation: The words in the span are redundant but fluent.
- applies only to good translations

- UT-important-information
- UnderTranslation: The span contains information that is missing in the translation.
- applies only to bad translations

UT-redundant-information
- UnderTranslation: The span contains information that is missing in the translation but that can be inferred or is trivial.
- applies to good and bad translations

UT-fluency
- UnderTranslation: The words in the span do not need to be translated.
- applies only to good translations

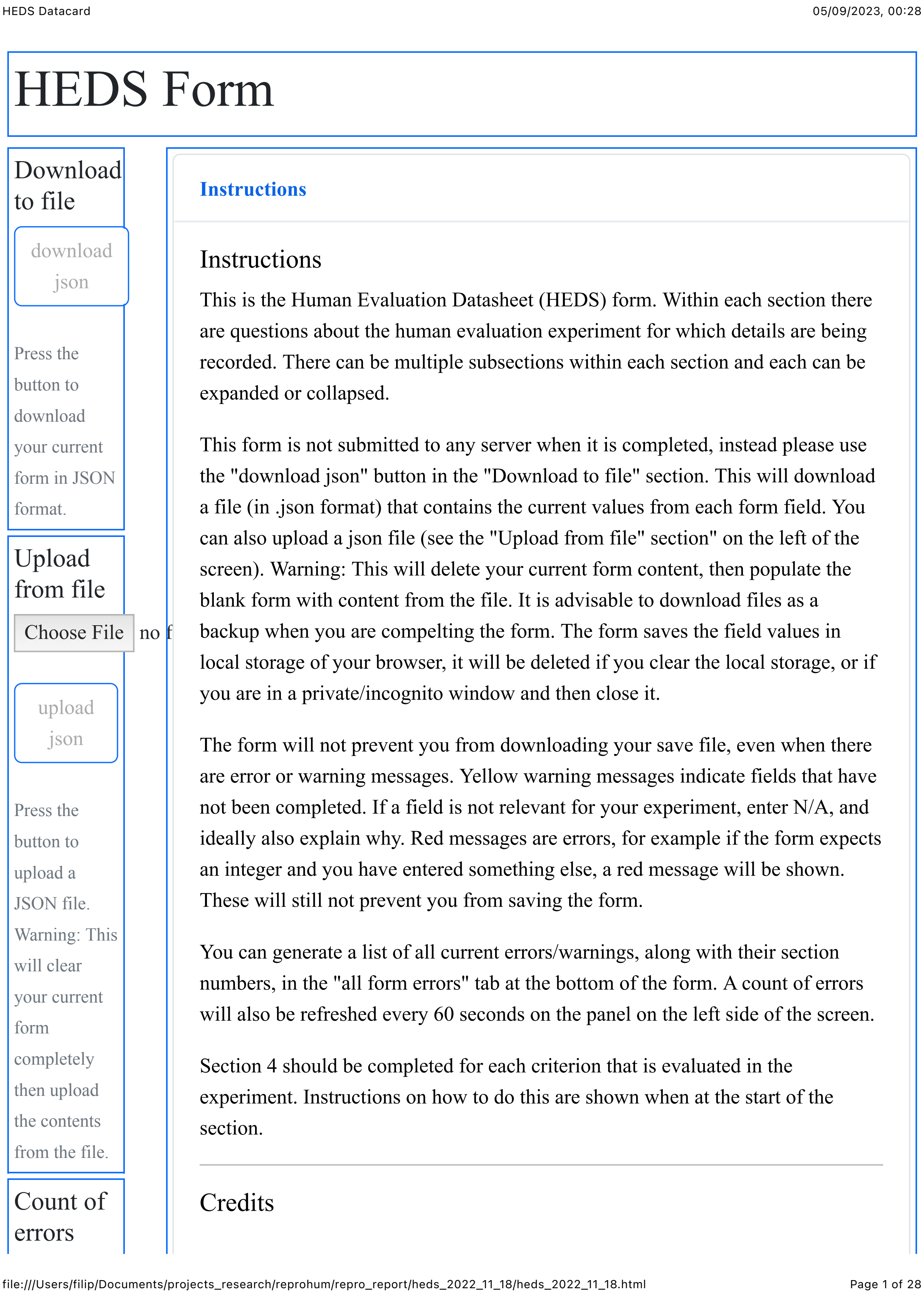
other-error-accuracy
- Other: The span is badly translated because of an accuracy error.
- this can be used both when the text is Over- and Under-Translated

other-error-fluency
- Other: The span is badly translated because of a fluency error.
- this can be used both when the text is Over- and Under-Translated

syntactic-difference
- The translation is syntactically different from the source.
- applies only to good translations, can use when the text is both Over- and Under Translated

source-error
- The translation fixes an error in the source.
- applies only to good translations, can use when the text is both Over- and Under Translated

unclear
- I don’t know.
- applies only to good translations, can use when the text is both Over- and Under Translated



HEDS Form

Download to file

download json

Press the button to download your current form in JSON format.

Upload from file

Choose File no f

upload json

Press the button to upload a JSON file. Warning: This will clear your current form completely then upload the contents from the file.

Count of errors

Instructions

Instructions

This is the Human Evaluation Datasheet (HEDS) form. Within each section there are questions about the human evaluation experiment for which details are being recorded. There can be multiple subsections within each section and each can be expanded or collapsed.

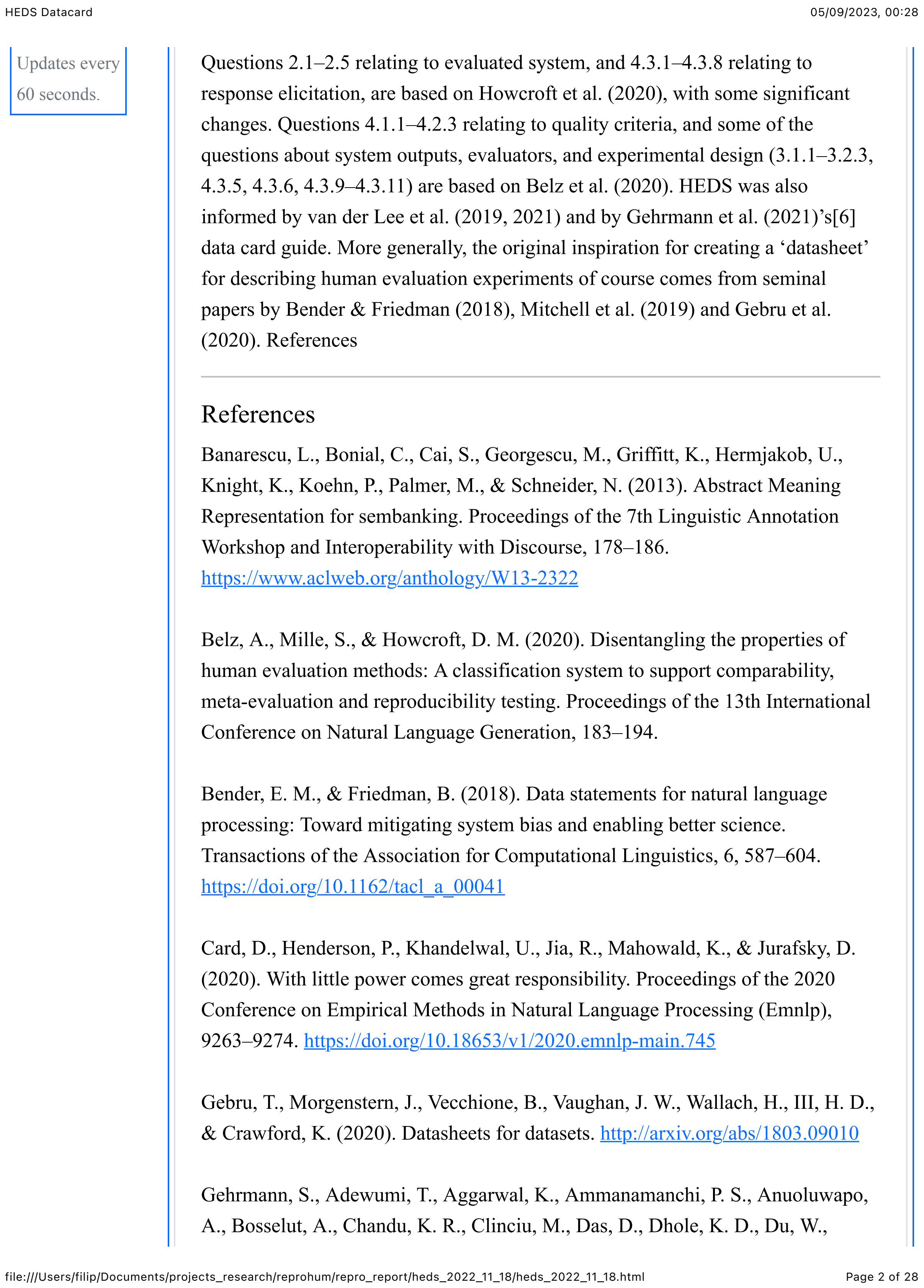
This form is not submitted to any server when it is completed, instead please use the "download json" button in the "Download to file" section. This will download a file (in .json format) that contains the current values from each form field. You can also upload a json file (see the "Upload from file" section" on the left of the screen). Warning: This will delete your current form content, then populate the blank form with content from the file. It is advisable to download files as a backup when you are compelting the form. The form saves the field values in local storage of your browser, it will be deleted if you clear the local storage, or if you are in a private/incognito window and then close it.

The form will not prevent you from downloading your save file, even when there are error or warning messages. Yellow warning messages indicate fields that have not been completed. If a field is not relevant for your experiment, enter N/A, and ideally also explain why. Red messages are errors, for example if the form expects an integer and you have entered something else, a red message will be shown. These will still not prevent you from saving the form.

You can generate a list of all current errors/warnings, along with their section numbers, in the "all form errors" tab at the bottom of the form. A count of errors will also be refreshed every 60 seconds on the panel on the left side of the screen.

Section 4 should be completed for each criterion that is evaluated in the experiment. Instructions on how to do this are shown when at the start of the section.

Credits



Updates every 60 seconds.

Questions 2.1–2.5 relating to evaluated system, and 4.3.1–4.3.8 relating to response elicitation, are based on Howcroft et al. (2020), with some significant changes. Questions 4.1.1–4.2.3 relating to quality criteria, and some of the questions about system outputs, evaluators, and experimental design (3.1.1–3.2.3, 4.3.5, 4.3.6, 4.3.9–4.3.11) are based on Belz et al. (2020). HEDS was also informed by van der Lee et al. (2019, 2021) and by Gehrmann et al. (2021)'s[6] data card guide. More generally, the original inspiration for creating a 'datasheet' for describing human evaluation experiments of course comes from seminal papers by Bender & Friedman (2018), Mitchell et al. (2019) and Gebru et al. (2020). References

References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract Meaning Representation for sembanking. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, 178–186. https://www.aclweb.org/anthology/W13-2322

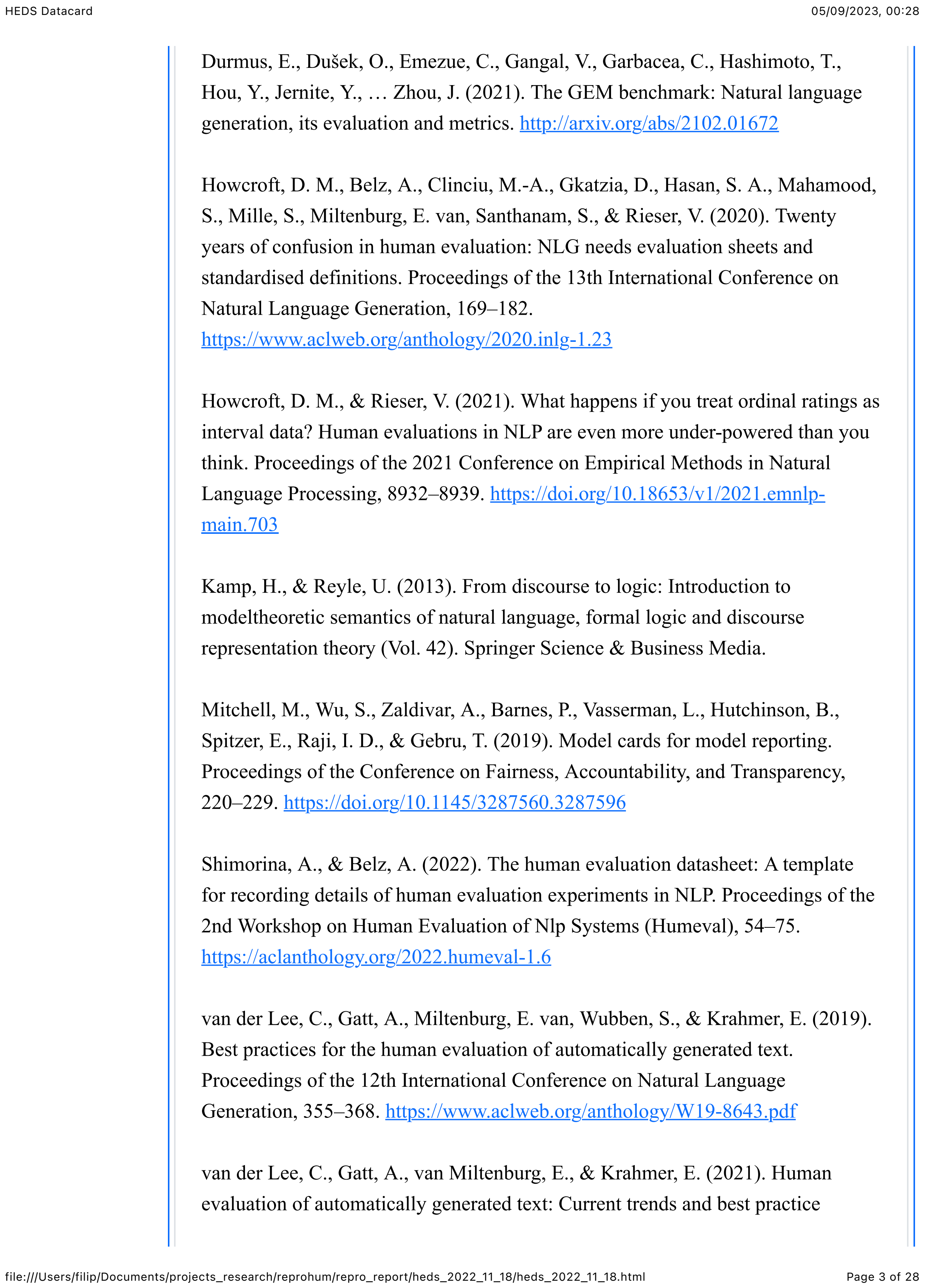
Belz, A., Mille, S., & Howcroft, D. M. (2020). Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. Proceedings of the 13th International Conference on Natural Language Generation, 183–194.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020). With little power comes great responsibility. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Emnlp), 9263–9274. https://doi.org/10.18653/v1/2020.emnlp-main.745

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2020). Datasheets for datasets. http://arxiv.org/abs/1803.09010

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Anuoluwapo, A., Bosselut, A., Chandu, K. R., Clinciu, M., Das, D., Dhole, K. D., Du, W.,



Durmus, E., Dušek, O., Emezue, C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., … Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics. http://arxiv.org/abs/2102.01672

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Miltenburg, E. van, Santhanam, S., & Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. Proceedings of the 13th International Conference on Natural Language Generation, 169–182. https://www.aclweb.org/anthology/2020.inlg-1.23

Howcroft, D. M., & Rieser, V. (2021). What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 8932–8939. https://doi.org/10.18653/v1/2021.emnlp-main.703

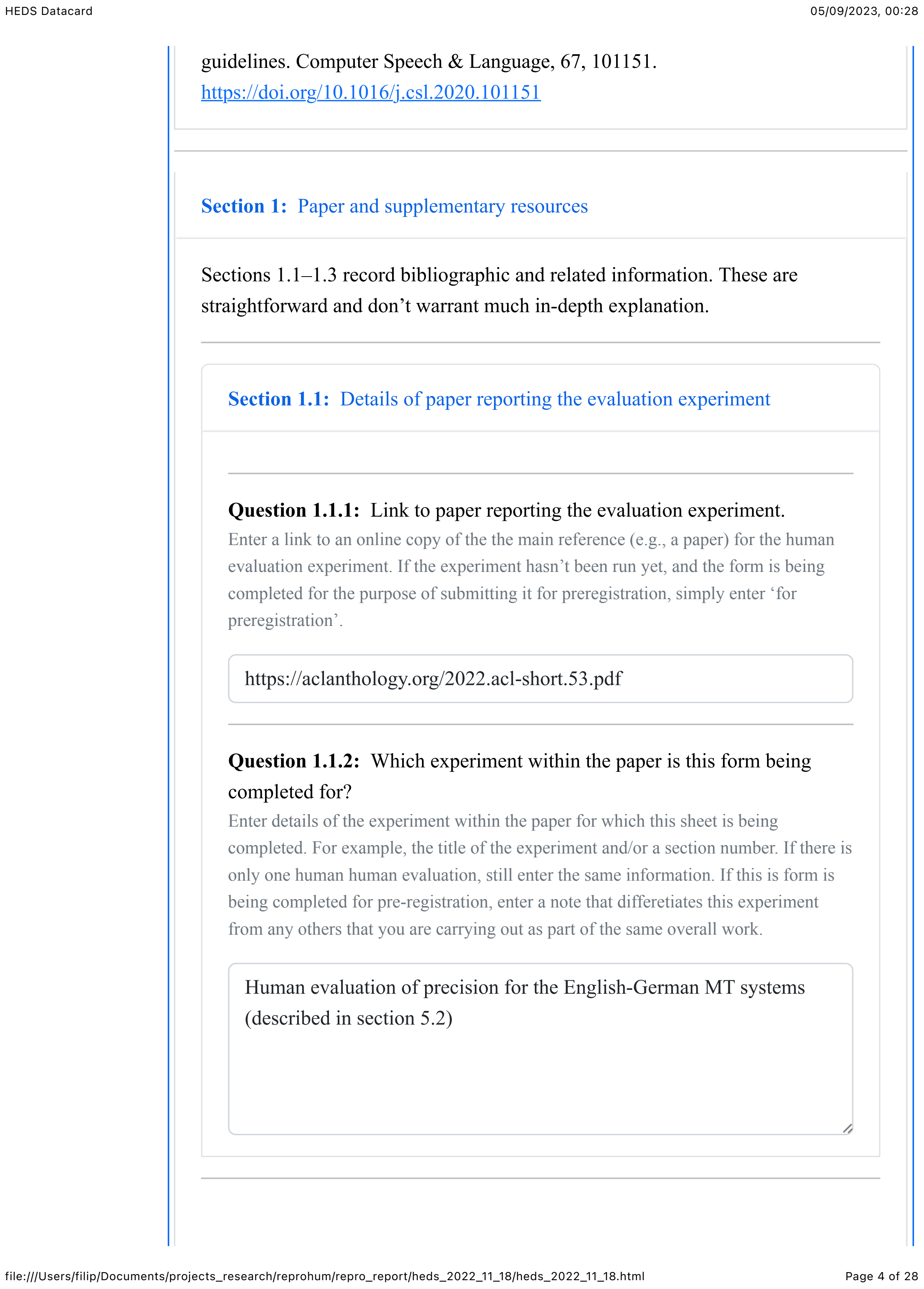
Kamp, H., & Reyle, U. (2013). From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory (Vol. 42). Springer Science & Business Media.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Shimorina, A., & Belz, A. (2022). The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. Proceedings of the 2nd Workshop on Human Evaluation of Nlp Systems (Humeval), 54–75. https://aclanthology.org/2022.humeval-1.6

van der Lee, C., Gatt, A., Miltenburg, E. van, Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. Proceedings of the 12th International Conference on Natural Language Generation, 355–368. https://www.aclweb.org/anthology/W19-8643.pdf

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice



guidelines. Computer Speech & Language, 67, 101151. https://doi.org/10.1016/j.csl.2020.101151

Section 1: Paper and supplementary resources

Sections 1.1–1.3 record bibliographic and related information. These are straightforward and don’t warrant much in-depth explanation.

Section 1.1: Details of paper reporting the evaluation experiment

Question 1.1.1: Link to paper reporting the evaluation experiment.

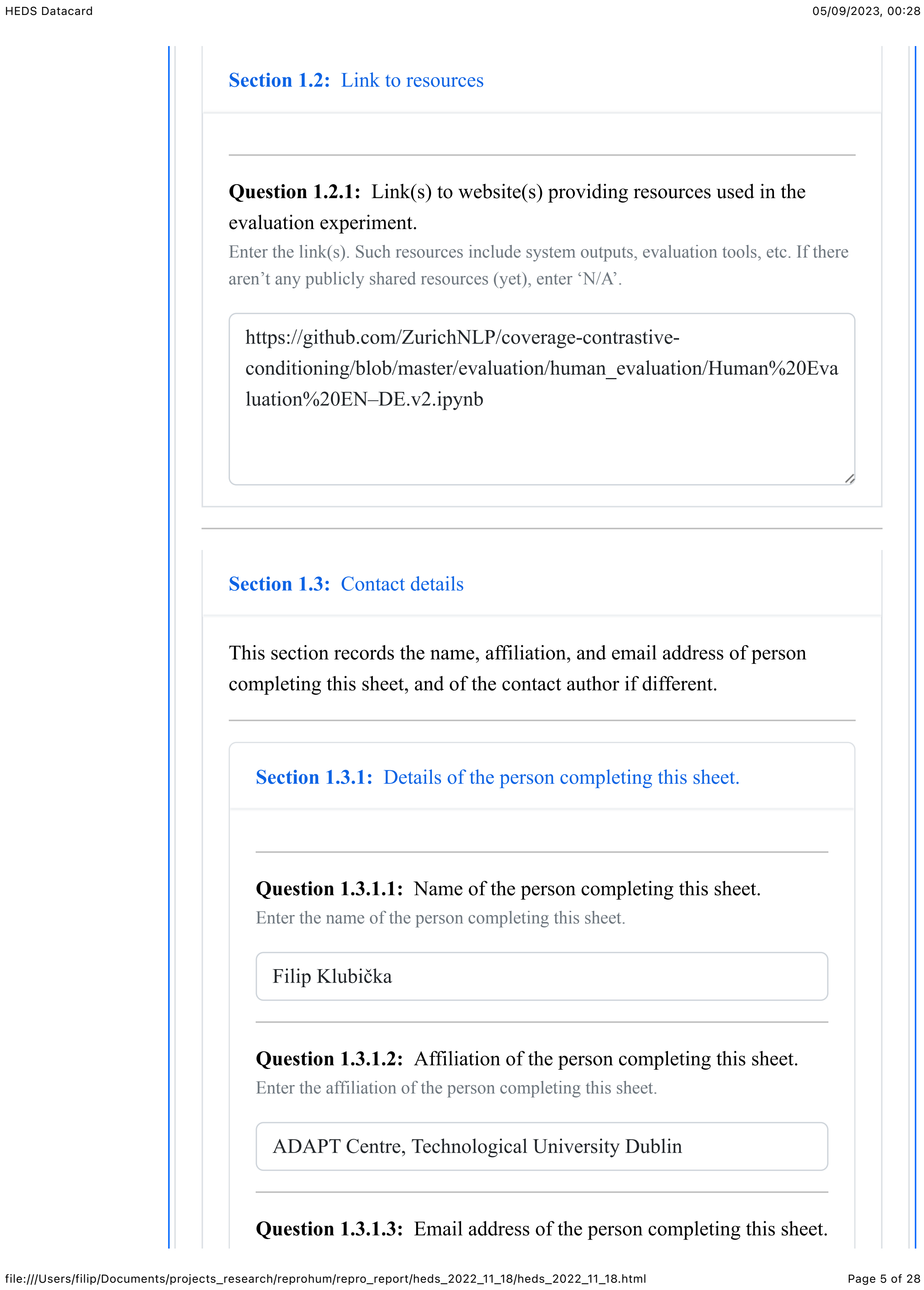
Enter a link to an online copy of the the main reference (e.g., a paper) for the human evaluation experiment. If the experiment hasn’t been run yet, and the form is being completed for the purpose of submitting it for preregistration, simply enter ‘for preregistration’.

https://aclanthology.org/2022.acl-short.53.pdf

Question 1.1.2: Which experiment within the paper is this form being completed for?

Enter details of the experiment within the paper for which this sheet is being completed. For example, the title of the experiment and/or a section number. If there is only one human human evaluation, still enter the same information. If this is form is being completed for pre-registration, enter a note that differetiates this experiment from any others that you are carrying out as part of the same overall work.

Human evaluation of precision for the English-German MT systems (described in section 5.2)



Section 1.2: Link to resources

Question 1.2.1: Link(s) to website(s) providing resources used in the evaluation experiment.

Enter the link(s). Such resources include system outputs, evaluation tools, etc. If there aren’t any publicly shared resources (yet), enter ‘N/A’.

https://github.com/ZurichNLP/coverage-contrastive-conditioning/blob/master/evaluation/human_evaluation/Human%20Evaluation%20EN–DE.v2.ipynb

Section 1.3: Contact details

This section records the name, affiliation, and email address of person completing this sheet, and of the contact author if different.

Section 1.3.1: Details of the person completing this sheet.

Question 1.3.1.1: Name of the person completing this sheet.

Enter the name of the person completing this sheet.

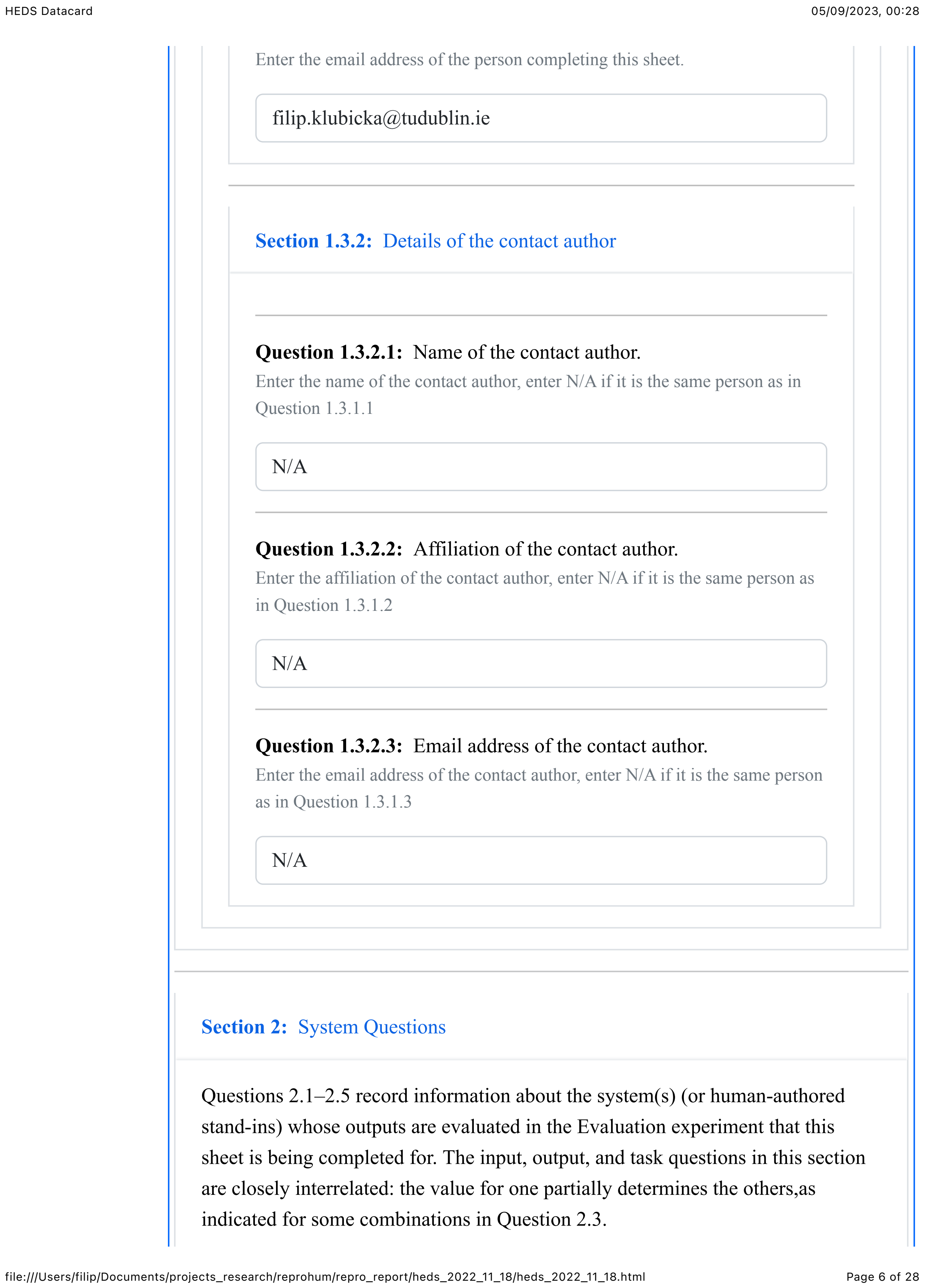
Filip Klubička

Question 1.3.1.2: Affiliation of the person completing this sheet.

Enter the affiliation of the person completing this sheet.

ADAPT Centre, Technological University Dublin

Question 1.3.1.3: Email address of the person completing this sheet.



Enter the email address of the person completing this sheet.

filip.klubicka@tudublin.ie

Section 1.3.2: Details of the contact author

Question 1.3.2.1: Name of the contact author.

Enter the name of the contact author, enter N/A if it is the same person as in Question 1.3.1.1

N/A

Question 1.3.2.2: Affiliation of the contact author.

Enter the affiliation of the contact author, enter N/A if it is the same person as in Question 1.3.1.2

N/A

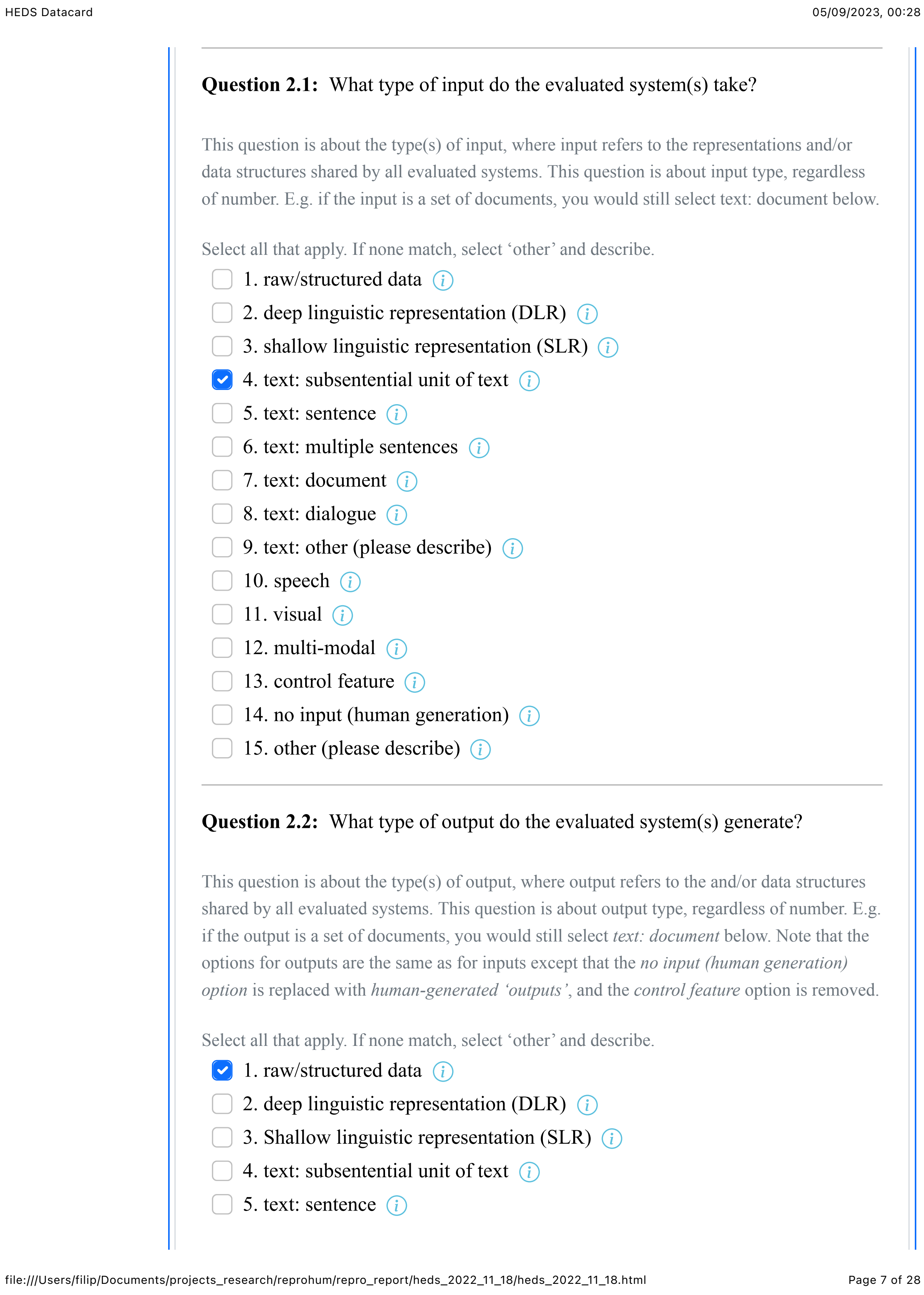
Question 1.3.2.3: Email address of the contact author.

Enter the email address of the contact author, enter N/A if it is the same person as in Question 1.3.1.3

N/A

Section 2: System Questions

Questions 2.1–2.5 record information about the system(s) (or human-authored stand-ins) whose outputs are evaluated in the Evaluation experiment that this sheet is being completed for. The input, output, and task questions in this section are closely interrelated: the value for one partially determines the others,as indicated for some combinations in Question 2.3.



Question 2.1: What type of input do the evaluated system(s) take?

This question is about the type(s) of input, where input refers to the representations and/or data structures shared by all evaluated systems. This question is about input type, regardless of number. E.g. if the input is a set of documents, you would still select text: document below.

Select all that apply. If none match, select ‘other’ and describe.

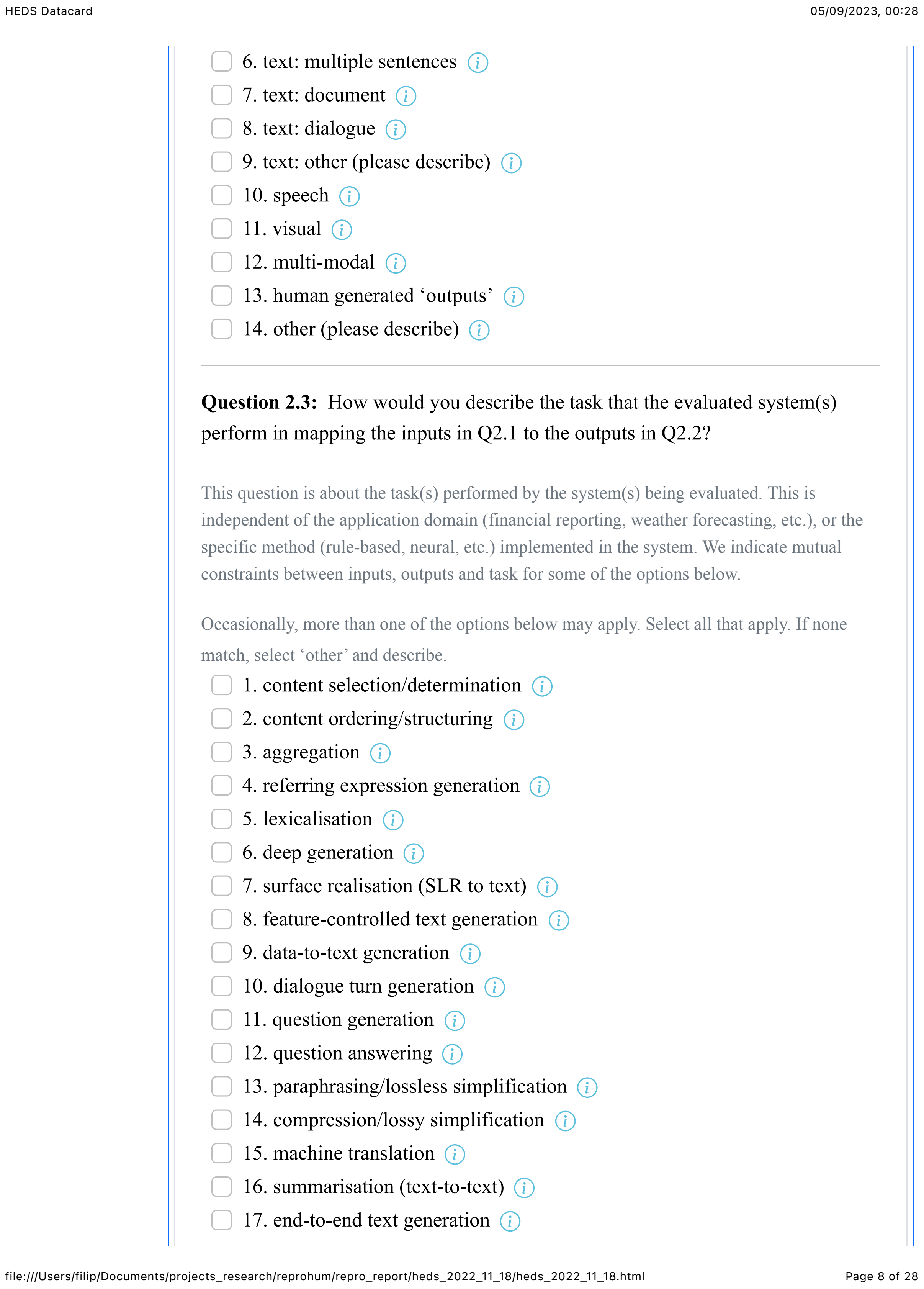
☐ 1. raw/structured data
☐ 2. deep linguistic representation (DLR)
☐ 3. shallow linguistic representation (SLR)
☑ 4. text: subsentential unit of text
☐ 5. text: sentence
☐ 6. text: multiple sentences
☐ 7. text: document
☐ 8. text: dialogue
☐ 9. text: other (please describe)
☐ 10. speech
☐ 11. visual
☐ 12. multi-modal
☐ 13. control feature
☐ 14. no input (human generation)
☐ 15. other (please describe)

Question 2.2: What type of output do the evaluated system(s) generate?

This question is about the type(s) of output, where output refers to the and/or data structures shared by all evaluated systems. This question is about output type, regardless of number. E.g. if the output is a set of documents, you would still select *text: document* below. Note that the options for outputs are the same as for inputs except that the *no input (human generation) option* is replaced with *human-generated ‘outputs’*, and the *control feature* option is removed.

Select all that apply. If none match, select ‘other’ and describe.

☑ 1. raw/structured data
☐ 2. deep linguistic representation (DLR)
☐ 3. Shallow linguistic representation (SLR)
☐ 4. text: subsentential unit of text
☐ 5. text: sentence



- □ 6. text: multiple sentences ⓘ
- □ 7. text: document ⓘ
- □ 8. text: dialogue ⓘ
- □ 9. text: other (please describe) ⓘ
- □ 10. speech ⓘ
- □ 11. visual ⓘ
- □ 12. multi-modal ⓘ
- □ 13. human generated ‘outputs’ ⓘ
- □ 14. other (please describe) ⓘ

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2?

This question is about the task(s) performed by the system(s) being evaluated. This is independent of the application domain (financial reporting, weather forecasting, etc.), or the specific method (rule-based, neural, etc.) implemented in the system. We indicate mutual constraints between inputs, outputs and task for some of the options below.

Occasionally, more than one of the options below may apply. Select all that apply. If none match, select ‘other’ and describe.

- □ 1. content selection/determination ⓘ
- □ 2. content ordering/structuring ⓘ
- □ 3. aggregation ⓘ
- □ 4. referring expression generation ⓘ
- □ 5. lexicalisation ⓘ
- □ 6. deep generation ⓘ
- □ 7. surface realisation (SLR to text) ⓘ
- □ 8. feature-controlled text generation ⓘ
- □ 9. data-to-text generation ⓘ
- □ 10. dialogue turn generation ⓘ
- □ 11. question generation ⓘ
- □ 12. question answering ⓘ
- □ 13. paraphrasing/lossless simplification ⓘ
- □ 14. compression/lossy simplification ⓘ
- □ 15. machine translation ⓘ
- □ 16. summarisation (text-to-text) ⓘ
- □ 17. end-to-end text generation ⓘ

□ 18. image/video description ⓘ
□ 19. post-editing/correction ⓘ
☑ 20. other (please describe) ⓘ

Please describe:

It's binary classification in a sense, predicting 0 or 1, mapped to 'Undertranslation' or 'Overtranslation' labels

Please provide further details for your above selection(s)

Question 2.4: What are the input languages that are used by the system?

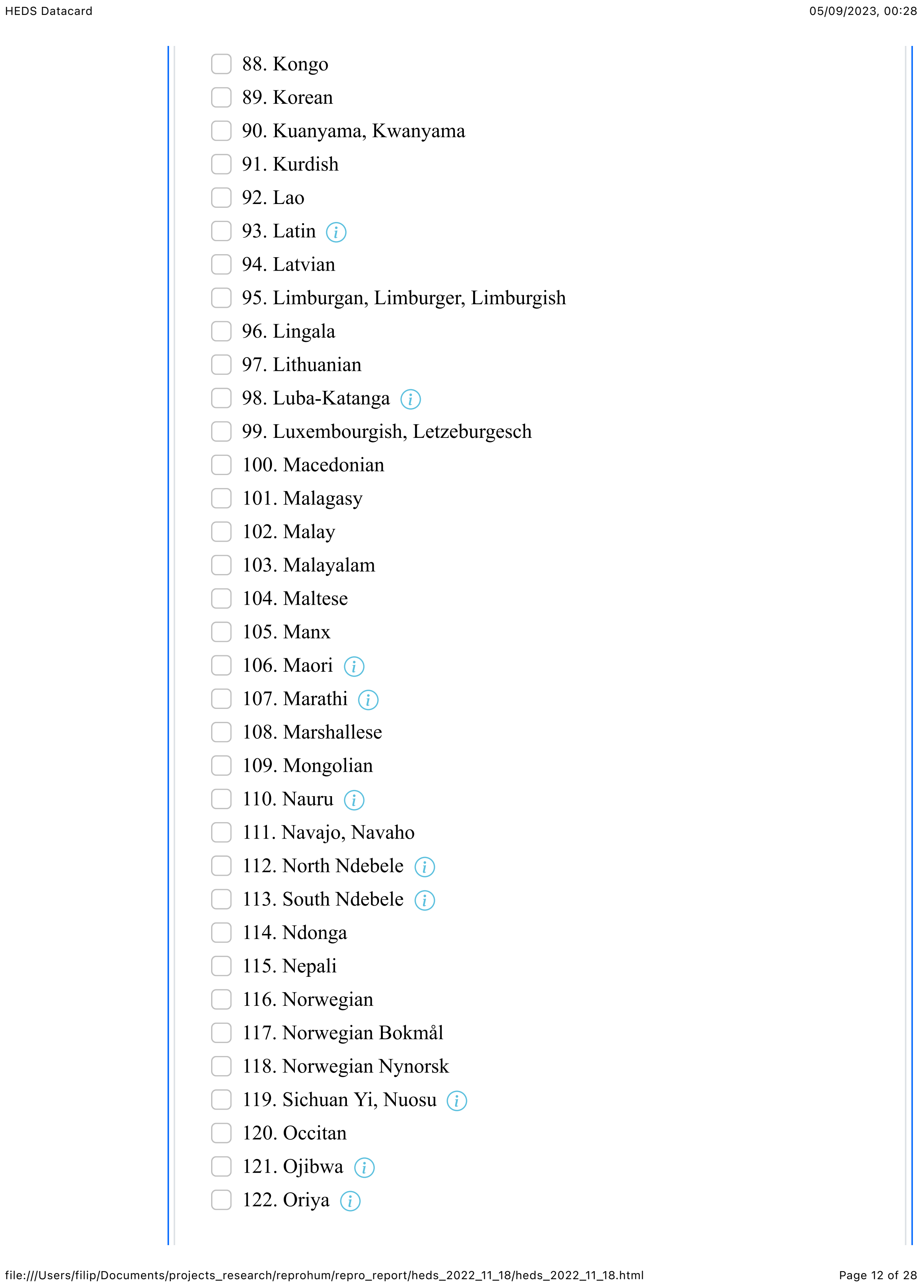
This question is about the language(s) of the inputs accepted by the system(s) being evaluated. Select any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019). E.g. English, Herero, Hindi. If no language is accepted as (part of) the input, select 'N/A'.

Select all that apply. If any languages you are using are not covered by this list, select 'other' and describe.

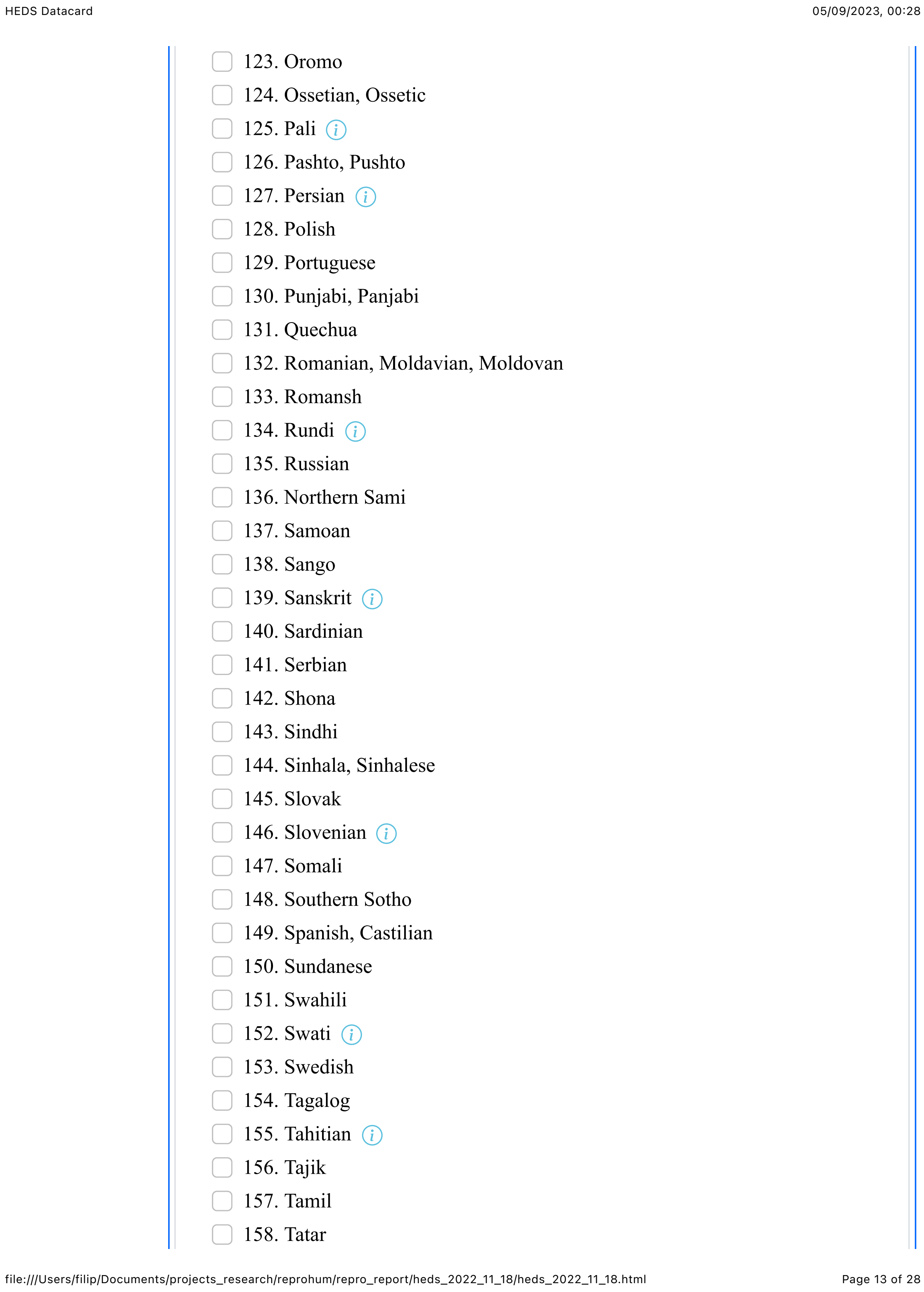
□ 1. Abkhazian ⓘ
□ 2. Afar
□ 3. Afrikaans
□ 4. Akan
□ 5. Albanian
□ 6. Amharic
□ 7. Arabic
□ 8. Aragonese
□ 9. Armenian
□ 10. Assamese
□ 11. Avaric ⓘ
□ 12. Avestan ⓘ
□ 13. Aymara
□ 14. Azerbaijani ⓘ
□ 15. Bambara
□ 16. Bashkir

□ 17. Basque
□ 18. Belarusian
□ 19. Bengali ⓘ
□ 20. Bislama ⓘ
□ 21. Bosnian
□ 22. Breton
□ 23. Bulgarian
□ 24. Burmese ⓘ
□ 25. Catalan, Valencian
□ 26. Chamorro
□ 27. Chechen
□ 28. Chichewa, Chewa, Nyanja
□ 29. Chinese
□ 30. Church Slavic, Old Slavonic, Church Slavonic, Old Bulgarian, Old Church Slavonic ⓘ
□ 31. Chuvash
□ 32. Cornish
□ 33. Corsican
□ 34. Cree
□ 35. Croatian
□ 36. Czech
□ 37. Danish
□ 38. Divehi, Dhivehi, Maldivian
□ 39. Dutch, Flemish ⓘ
□ 40. Dzongkha
☑ 41. English
□ 42. Esperanto ⓘ
□ 43. Estonian
□ 44. Ewe
□ 45. Faroese
□ 46. Fijian
□ 47. Finnish
□ 48. French
□ 49. Western Frisian ⓘ
□ 50. Fulah ⓘ
□ 51. Gaelic, Scottish Gaelic

□ 52. Galician
□ 53. Ganda
□ 54. Georgian
☑ 55. German
□ 56. Greek, Modern (1453–)
□ 57. Kalaallisut, Greenlandic
□ 58. Guarani
□ 59. Gujarati
□ 60. Haitian, Haitian Creole
□ 61. Hausa
□ 62. Hebrew ⓘ
□ 63. Herero
□ 64. Hindi
□ 65. Hiri Motu
□ 66. Hungarian
□ 67. Icelandic
□ 68. Ido ⓘ
□ 69. Igbo
□ 70. Indonesian
□ 71. Interlingua (International Auxiliary Language Association) ⓘ
□ 72. Interlingue, Occidental ⓘ
□ 73. Inuktitut
□ 74. Inupiaq
□ 75. Irish
□ 76. Italian
□ 77. Japanese
□ 78. Javanese
□ 79. Kannada
□ 80. Kanuri
□ 81. Kashmiri
□ 82. Kazakh
□ 83. Central Khmer ⓘ
□ 84. Kikuyu, Gikuyu
□ 85. Kinyarwanda
□ 86. Kirghiz, Kyrgyz
□ 87. Komi



□ 88. Kongo
□ 89. Korean
□ 90. Kuanyama, Kwanyama
□ 91. Kurdish
□ 92. Lao
□ 93. Latin ⓘ
□ 94. Latvian
□ 95. Limburgan, Limburger, Limburgish
□ 96. Lingala
□ 97. Lithuanian
□ 98. Luba-Katanga ⓘ
□ 99. Luxembourgish, Letzeburgesch
□ 100. Macedonian
□ 101. Malagasy
□ 102. Malay
□ 103. Malayalam
□ 104. Maltese
□ 105. Manx
□ 106. Maori ⓘ
□ 107. Marathi ⓘ
□ 108. Marshallese
□ 109. Mongolian
□ 110. Nauru ⓘ
□ 111. Navajo, Navaho
□ 112. North Ndebele ⓘ
□ 113. South Ndebele ⓘ
□ 114. Ndonga
□ 115. Nepali
□ 116. Norwegian
□ 117. Norwegian Bokmål
□ 118. Norwegian Nynorsk
□ 119. Sichuan Yi, Nuosu ⓘ
□ 120. Occitan
□ 121. Ojibwa ⓘ
□ 122. Oriya ⓘ



□ 123. Oromo

□ 124. Ossetian, Ossetic

□ 125. Pali ⓘ

□ 126. Pashto, Pushto

□ 127. Persian ⓘ

□ 128. Polish

□ 129. Portuguese

□ 130. Punjabi, Panjabi

□ 131. Quechua

□ 132. Romanian, Moldavian, Moldovan

□ 133. Romansh

□ 134. Rundi ⓘ

□ 135. Russian

□ 136. Northern Sami

□ 137. Samoan

□ 138. Sango

□ 139. Sanskrit ⓘ

□ 140. Sardinian

□ 141. Serbian

□ 142. Shona

□ 143. Sindhi

□ 144. Sinhala, Sinhalese

□ 145. Slovak

□ 146. Slovenian ⓘ

□ 147. Somali

□ 148. Southern Sotho

□ 149. Spanish, Castilian

□ 150. Sundanese

□ 151. Swahili

□ 152. Swati ⓘ

□ 153. Swedish

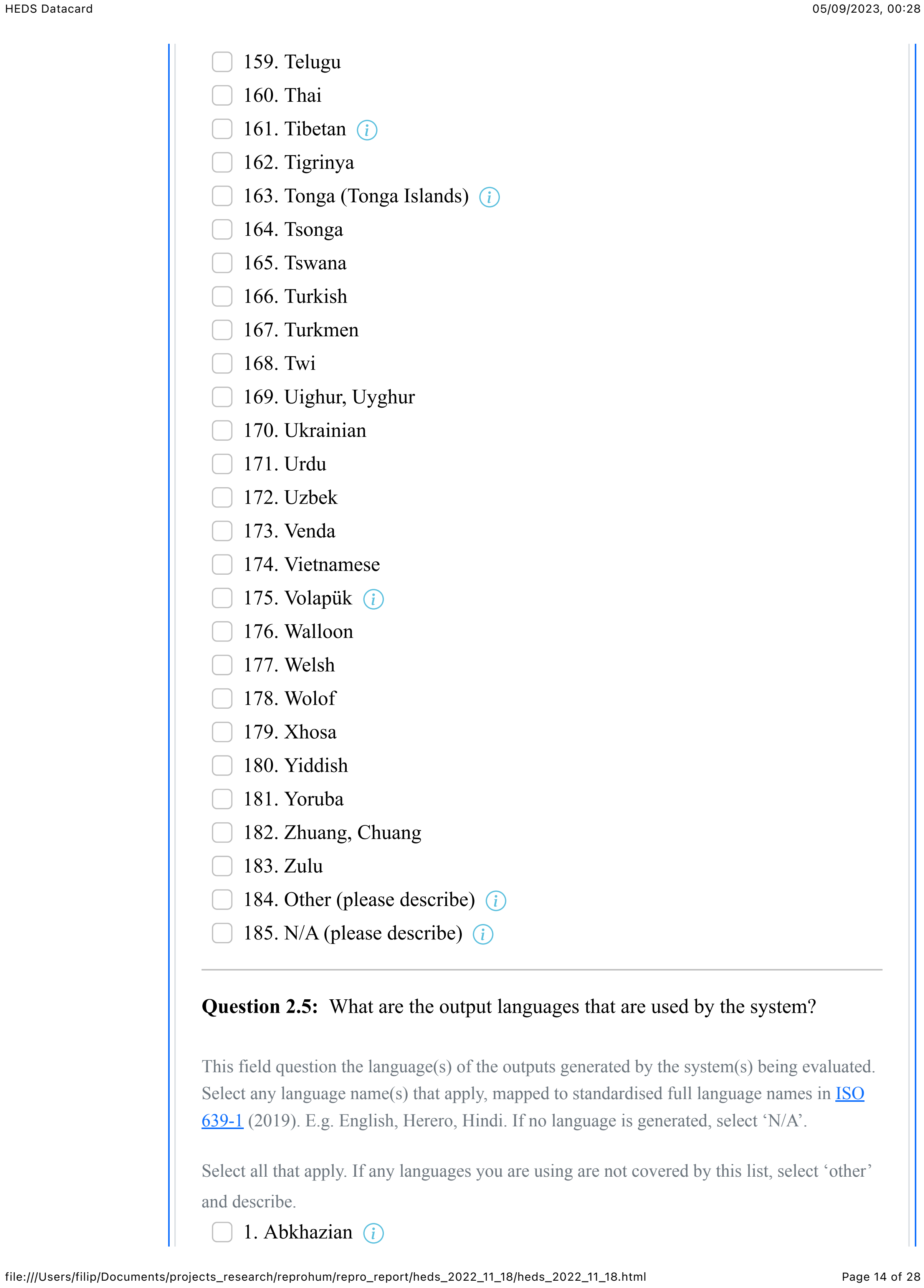
□ 154. Tagalog

□ 155. Tahitian ⓘ

□ 156. Tajik

□ 157. Tamil

□ 158. Tatar



□ 159. Telugu

□ 160. Thai

□ 161. Tibetan ⓘ

□ 162. Tigrinya

□ 163. Tonga (Tonga Islands) ⓘ

□ 164. Tsonga

□ 165. Tswana

□ 166. Turkish

□ 167. Turkmen

□ 168. Twi

□ 169. Uighur, Uyghur

□ 170. Ukrainian

□ 171. Urdu

□ 172. Uzbek

□ 173. Venda

□ 174. Vietnamese

□ 175. Volapük ⓘ

□ 176. Walloon

□ 177. Welsh

□ 178. Wolof

□ 179. Xhosa

□ 180. Yiddish

□ 181. Yoruba

□ 182. Zhuang, Chuang

□ 183. Zulu

□ 184. Other (please describe) ⓘ

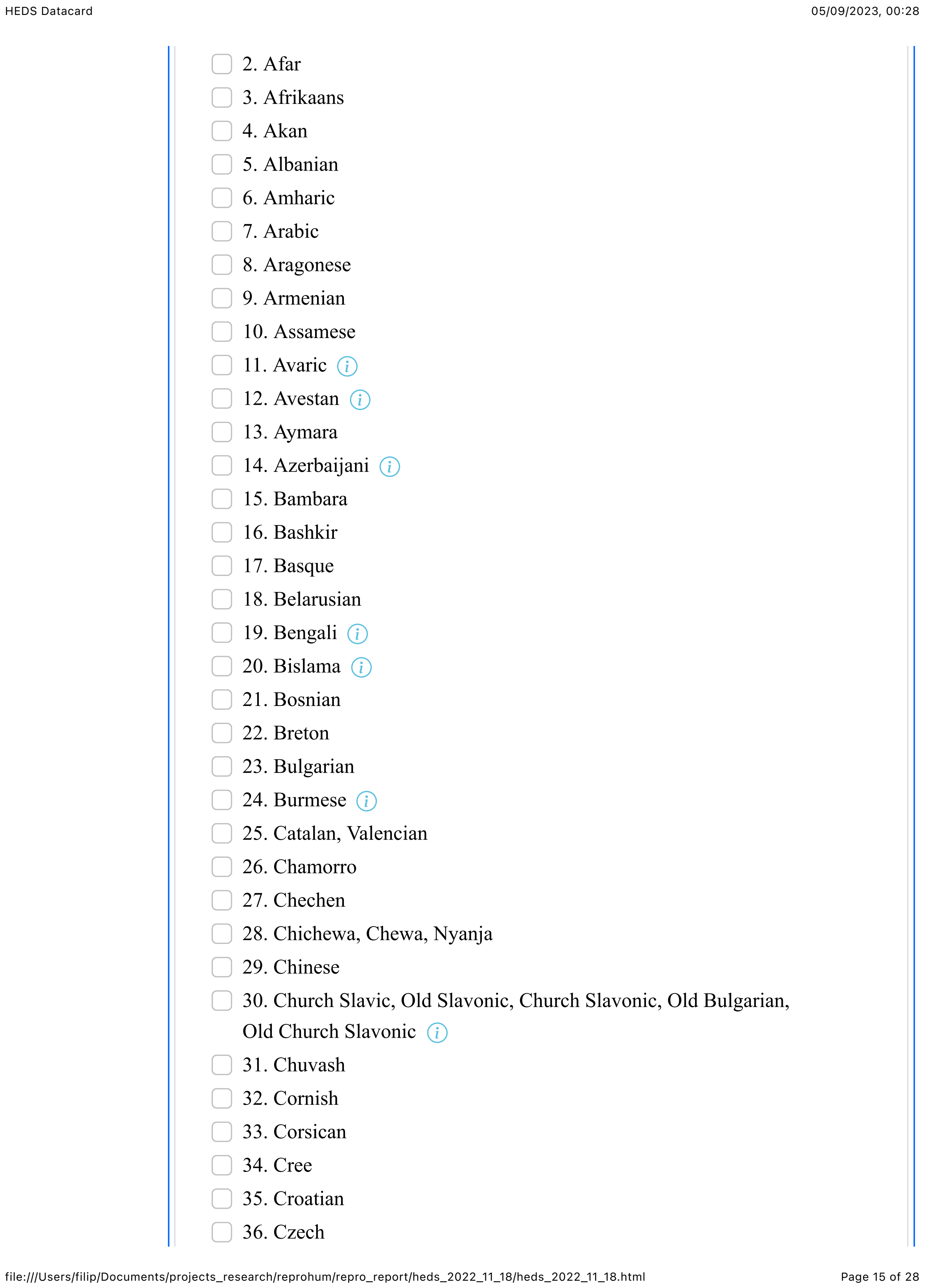
□ 185. N/A (please describe) ⓘ

Question 2.5: What are the output languages that are used by the system?

This field question the language(s) of the outputs generated by the system(s) being evaluated. Select any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019). E.g. English, Herero, Hindi. If no language is generated, select ‘N/A’.

Select all that apply. If any languages you are using are not covered by this list, select ‘other’ and describe.

□ 1. Abkhazian ⓘ



□ 2. Afar

□ 3. Afrikaans

□ 4. Akan

□ 5. Albanian

□ 6. Amharic

□ 7. Arabic

□ 8. Aragonese

□ 9. Armenian

□ 10. Assamese

□ 11. Avaric ⓘ

□ 12. Avestan ⓘ

□ 13. Aymara

□ 14. Azerbaijani ⓘ

□ 15. Bambara

□ 16. Bashkir

□ 17. Basque

□ 18. Belarusian

□ 19. Bengali ⓘ

□ 20. Bislama ⓘ

□ 21. Bosnian

□ 22. Breton

□ 23. Bulgarian

□ 24. Burmese ⓘ

□ 25. Catalan, Valencian

□ 26. Chamorro

□ 27. Chechen

□ 28. Chichewa, Chewa, Nyanja

□ 29. Chinese

□ 30. Church Slavic, Old Slavonic, Church Slavonic, Old Bulgarian, Old Church Slavonic ⓘ

□ 31. Chuvash

□ 32. Cornish

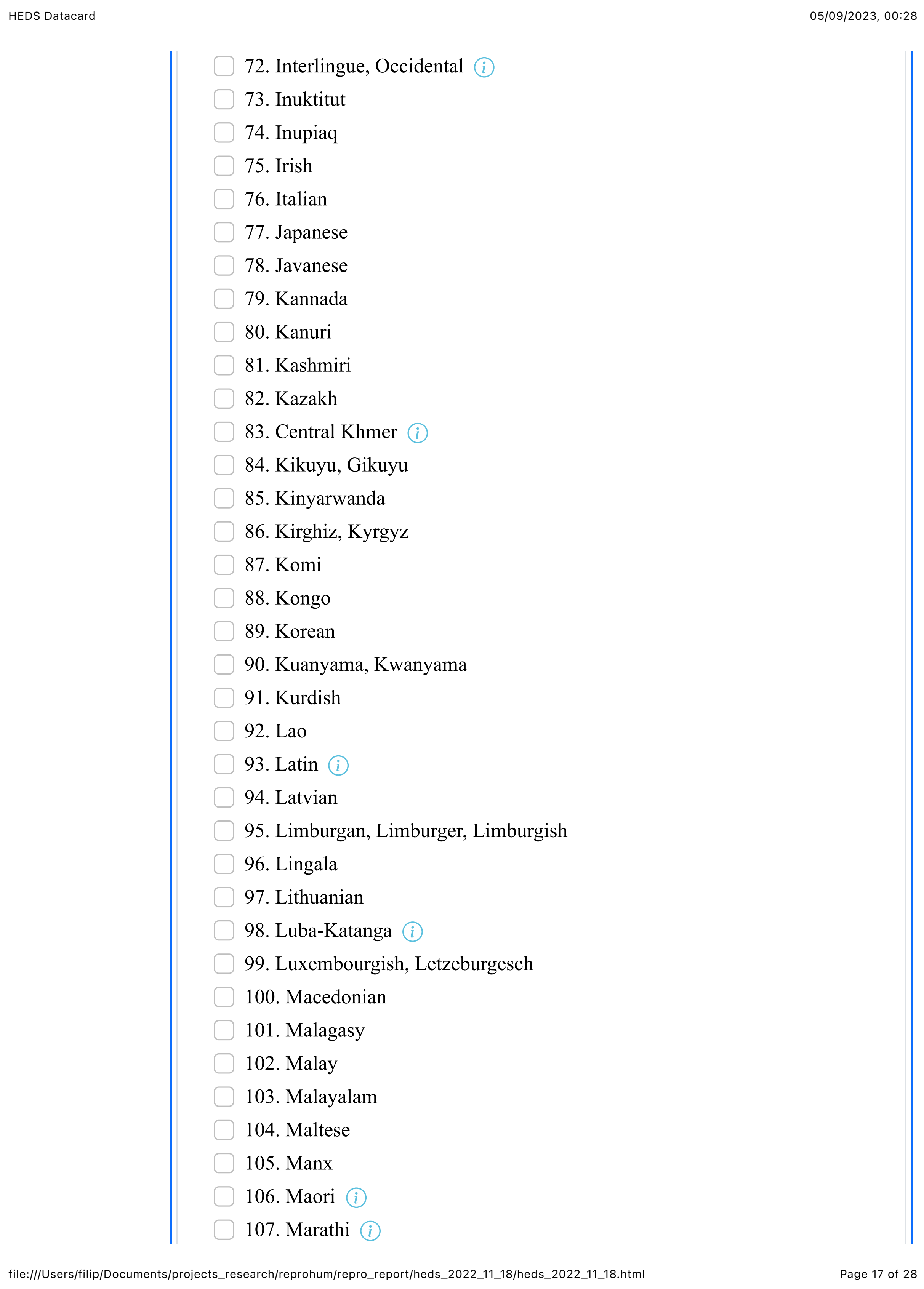
□ 33. Corsican

□ 34. Cree

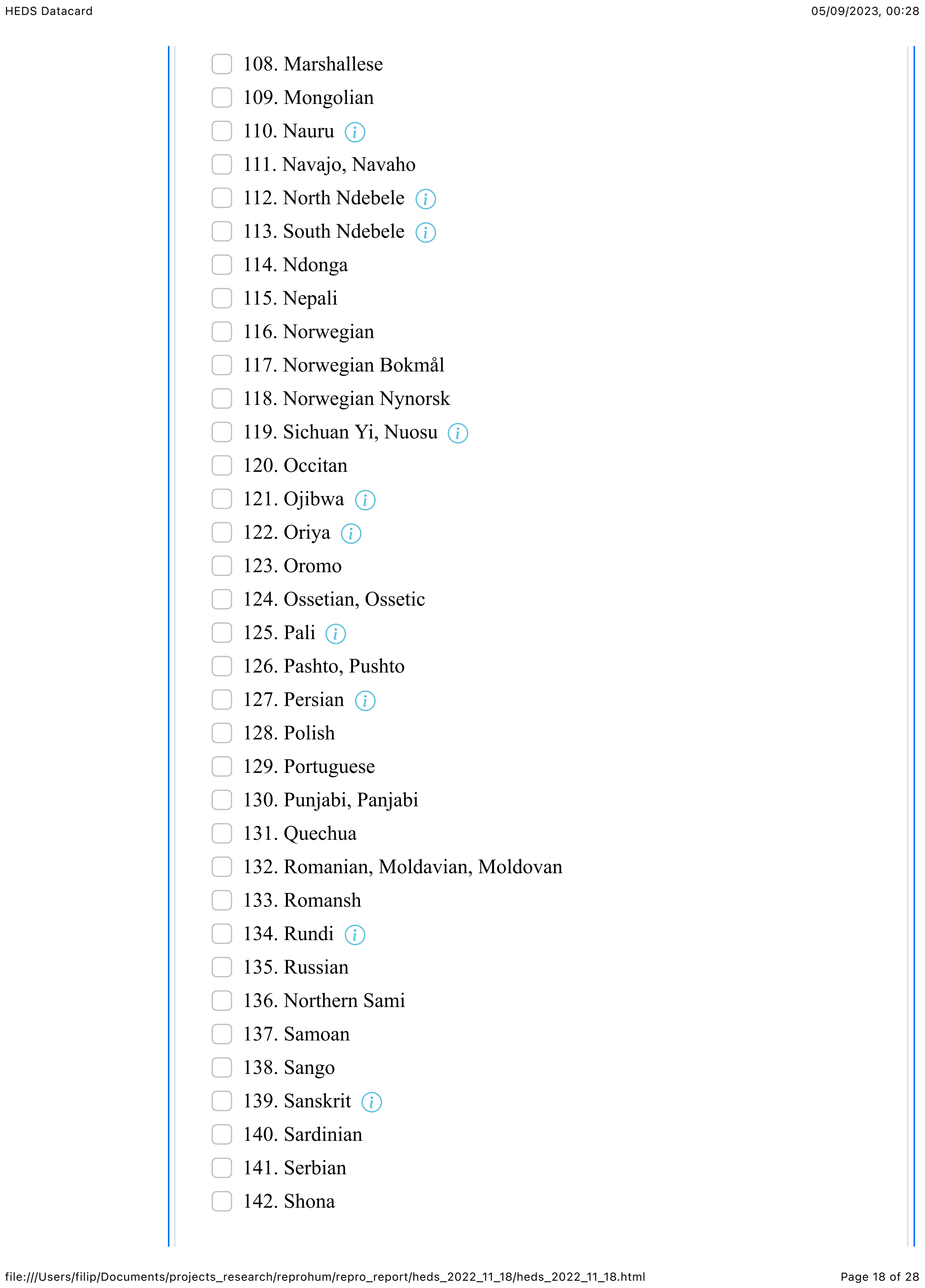
□ 35. Croatian

□ 36. Czech

- ☐ 37. Danish
- ☐ 38. Divehi, Dhivehi, Maldivian
- ☐ 39. Dutch, Flemish ⓘ
- ☐ 40. Dzongkha
- ☑ 41. English
- ☐ 42. Esperanto ⓘ
- ☐ 43. Estonian
- ☐ 44. Ewe
- ☐ 45. Faroese
- ☐ 46. Fijian
- ☐ 47. Finnish
- ☐ 48. French
- ☐ 49. Western Frisian ⓘ
- ☐ 50. Fulah ⓘ
- ☐ 51. Gaelic, Scottish Gaelic
- ☐ 52. Galician
- ☐ 53. Ganda
- ☐ 54. Georgian
- ☑ 55. German
- ☐ 56. Greek, Modern (1453–)
- ☐ 57. Kalaallisut, Greenlandic
- ☐ 58. Guarani
- ☐ 59. Gujarati
- ☐ 60. Haitian, Haitian Creole
- ☐ 61. Hausa
- ☐ 62. Hebrew ⓘ
- ☐ 63. Herero
- ☐ 64. Hindi
- ☐ 65. Hiri Motu
- ☐ 66. Hungarian
- ☐ 67. Icelandic
- ☐ 68. Ido ⓘ
- ☐ 69. Igbo
- ☐ 70. Indonesian
- ☐ 71. Interlingua (International Auxiliary Language Association) ⓘ



□ 72. Interlingue, Occidental ⓘ
□ 73. Inuktitut
□ 74. Inupiaq
□ 75. Irish
□ 76. Italian
□ 77. Japanese
□ 78. Javanese
□ 79. Kannada
□ 80. Kanuri
□ 81. Kashmiri
□ 82. Kazakh
□ 83. Central Khmer ⓘ
□ 84. Kikuyu, Gikuyu
□ 85. Kinyarwanda
□ 86. Kirghiz, Kyrgyz
□ 87. Komi
□ 88. Kongo
□ 89. Korean
□ 90. Kuanyama, Kwanyama
□ 91. Kurdish
□ 92. Lao
□ 93. Latin ⓘ
□ 94. Latvian
□ 95. Limburgan, Limburger, Limburgish
□ 96. Lingala
□ 97. Lithuanian
□ 98. Luba-Katanga ⓘ
□ 99. Luxembourgish, Letzeburgesch
□ 100. Macedonian
□ 101. Malagasy
□ 102. Malay
□ 103. Malayalam
□ 104. Maltese
□ 105. Manx
□ 106. Maori ⓘ
□ 107. Marathi ⓘ



□ 108. Marshallese

□ 109. Mongolian

□ 110. Nauru ⓘ

□ 111. Navajo, Navaho

□ 112. North Ndebele ⓘ

□ 113. South Ndebele ⓘ

□ 114. Ndonga

□ 115. Nepali

□ 116. Norwegian

□ 117. Norwegian Bokmål

□ 118. Norwegian Nynorsk

□ 119. Sichuan Yi, Nuosu ⓘ

□ 120. Occitan

□ 121. Ojibwa ⓘ

□ 122. Oriya ⓘ

□ 123. Oromo

□ 124. Ossetian, Ossetic

□ 125. Pali ⓘ

□ 126. Pashto, Pushto

□ 127. Persian ⓘ

□ 128. Polish

□ 129. Portuguese

□ 130. Punjabi, Panjabi

□ 131. Quechua

□ 132. Romanian, Moldavian, Moldovan

□ 133. Romansh

□ 134. Rundi ⓘ

□ 135. Russian

□ 136. Northern Sami

□ 137. Samoan

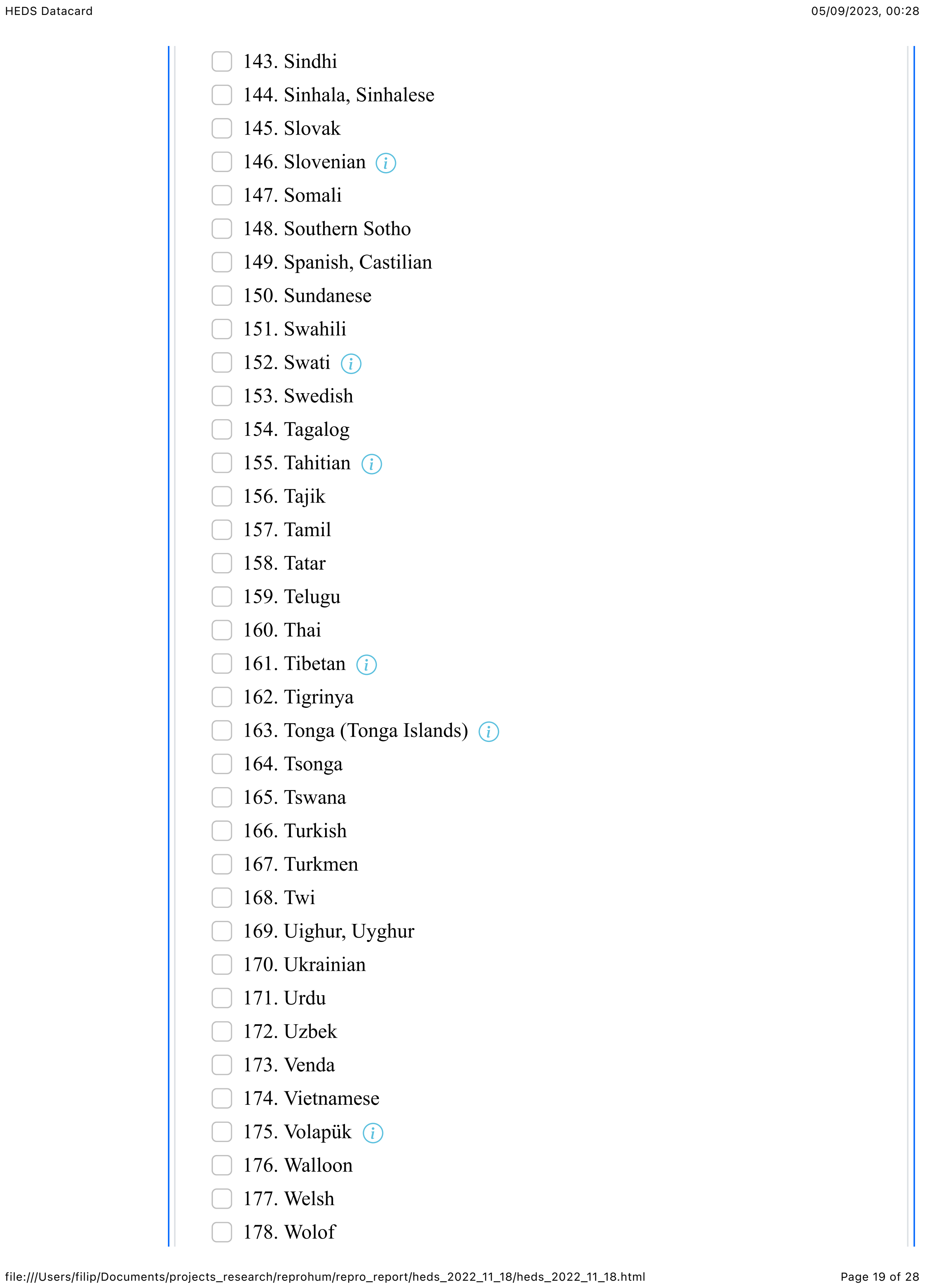
□ 138. Sango

□ 139. Sanskrit ⓘ

□ 140. Sardinian

□ 141. Serbian

□ 142. Shona



□ 143. Sindhi

□ 144. Sinhala, Sinhalese

□ 145. Slovak

□ 146. Slovenian ⓘ

□ 147. Somali

□ 148. Southern Sotho

□ 149. Spanish, Castilian

□ 150. Sundanese

□ 151. Swahili

□ 152. Swati ⓘ

□ 153. Swedish

□ 154. Tagalog

□ 155. Tahitian ⓘ

□ 156. Tajik

□ 157. Tamil

□ 158. Tatar

□ 159. Telugu

□ 160. Thai

□ 161. Tibetan ⓘ

□ 162. Tigrinya

□ 163. Tonga (Tonga Islands) ⓘ

□ 164. Tsonga

□ 165. Tswana

□ 166. Turkish

□ 167. Turkmen

□ 168. Twi

□ 169. Uighur, Uyghur

□ 170. Ukrainian

□ 171. Urdu

□ 172. Uzbek

□ 173. Venda

□ 174. Vietnamese

□ 175. Volapük ⓘ

□ 176. Walloon

□ 177. Welsh

□ 178. Wolof

- ☐ 179. Xhosa
- ☐ 180. Yiddish
- ☐ 181. Yoruba
- ☐ 182. Zhuang, Chuang
- ☐ 183. Zulu
- ☐ 184. Other (please describe) ⓘ
- ☐ 185. N/A (please describe) ⓘ

Section 3: Sample of system outputs, evaluators, and experimental design

Section 3.1: Sample of system outputs

Questions 3.1.1–3.1.3 record information about the size of the sample of outputs (or human-authored stand-ins) evaluated per system, how the sample was selected, and what its statistical power is.

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment?

Enter the number of system outputs (or other evaluation items) that are evaluated per system by at least one evaluator in the experiment. For most experiments this should be an integer, although if the number of outputs varies please provide further details here.

1505

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment?

Select one option. If none match, select ‘other’ and describe:

○ 1. by an automatic random process ⓘ

○ 2. by an automatic random process but using stratified sampling over given properties ⓘ

● 3. by manual, arbitrary selection ⓘ

○ 4. by manual selection aimed at achieving balance or variety relative to given properties ⓘ

○ 5. other (please describe) ⓘ

Section 3.1.3: Statistical power of the sample size.

Section 3.2: Evaluators

Questions 3.2.1–3.2.5 record information about the evaluators participating in the experiment.

Question 3.2.1: How many evaluators are there in this experiment?

Enter the total number of evaluators participating in the experiment, as an integer.

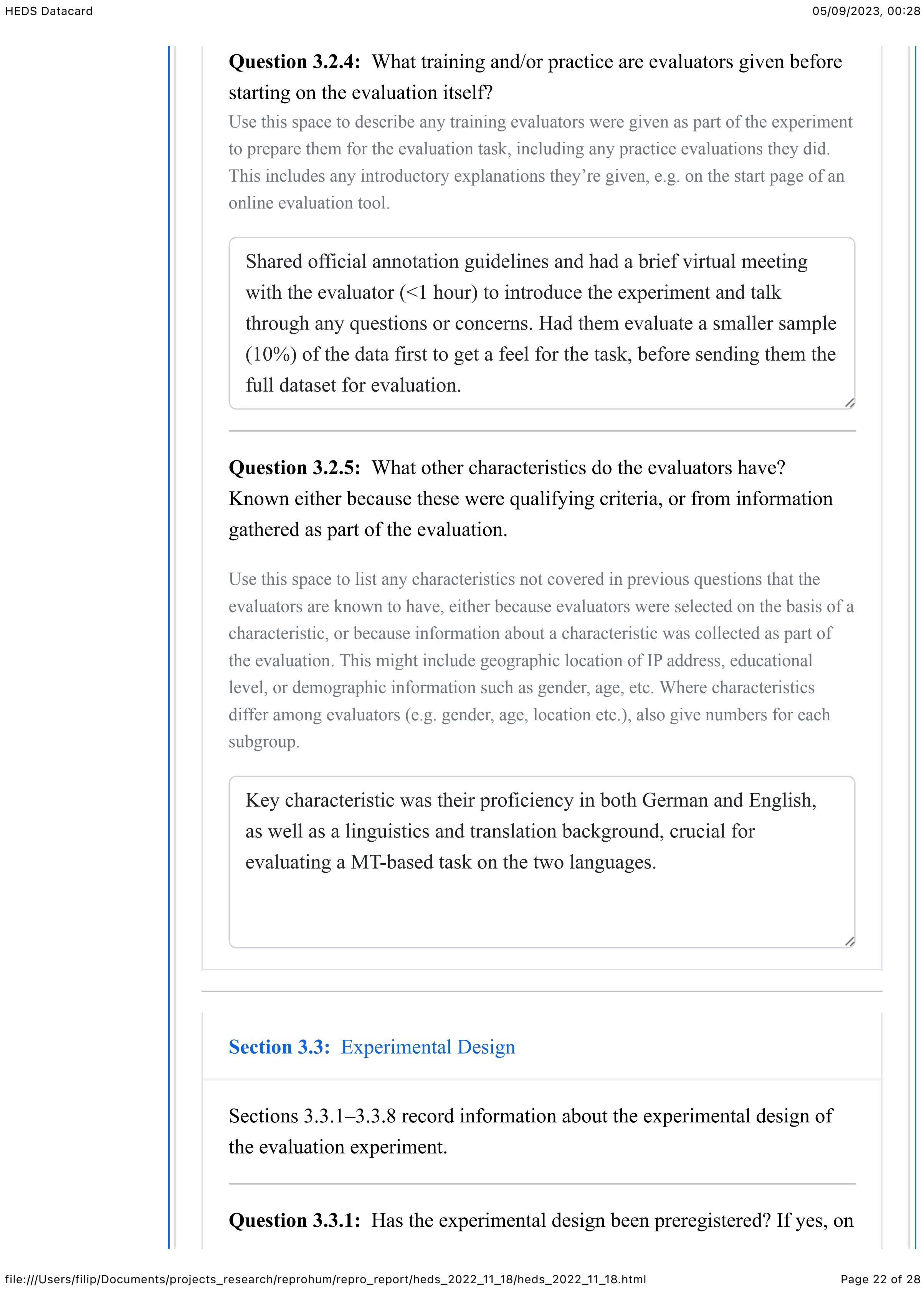
2

Section 3.2.2: Evaluator Type

Question 3.2.3: How are evaluators recruited?

Please explain how your evaluators are recruited. Do you send emails to a given list? Do you post invitations on social media? Posters on university walls? Were there any gatekeepers involved? What are the exclusion/inclusion criteria?

The evaluators came highly recommended by a colleague who teaches the translation studies course.



Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?

Use this space to describe any training evaluators were given as part of the experiment to prepare them for the evaluation task, including any practice evaluations they did. This includes any introductory explanations they're given, e.g. on the start page of an online evaluation tool.

Shared official annotation guidelines and had a brief virtual meeting with the evaluator (<1 hour) to introduce the experiment and talk through any questions or concerns. Had them evaluate a smaller sample (10%) of the data first to get a feel for the task, before sending them the full dataset for evaluation.

Question 3.2.5: What other characteristics do the evaluators have? Known either because these were qualifying criteria, or from information gathered as part of the evaluation.

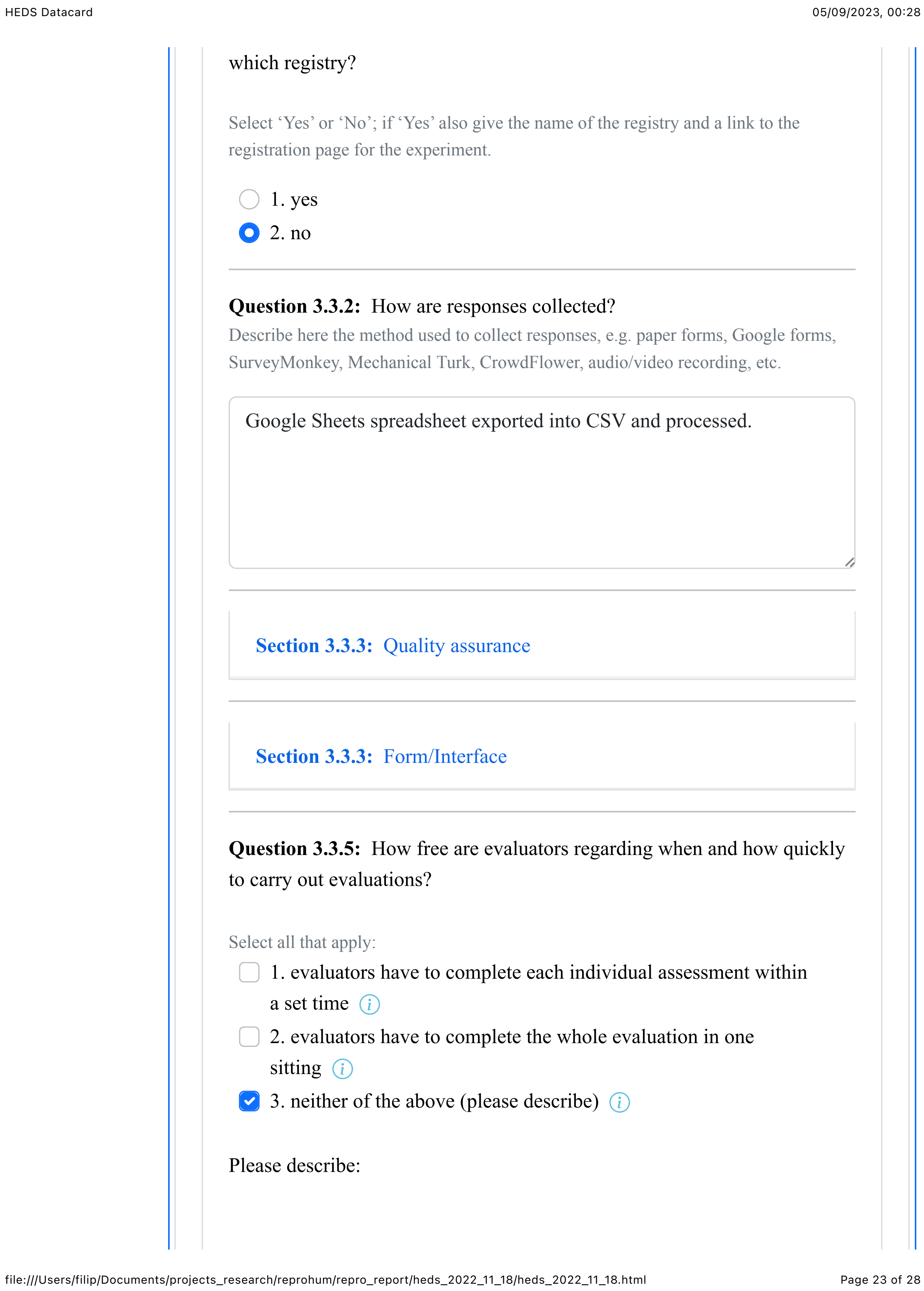
Use this space to list any characteristics not covered in previous questions that the evaluators are known to have, either because evaluators were selected on the basis of a characteristic, or because information about a characteristic was collected as part of the evaluation. This might include geographic location of IP address, educational level, or demographic information such as gender, age, etc. Where characteristics differ among evaluators (e.g. gender, age, location etc.), also give numbers for each subgroup.

Key characteristic was their proficiency in both German and English, as well as a linguistics and translation background, crucial for evaluating a MT-based task on the two languages.

Section 3.3: Experimental Design

Sections 3.3.1–3.3.8 record information about the experimental design of the evaluation experiment.

Question 3.3.1: Has the experimental design been preregistered? If yes, on



HEDS Datacard 05/09/2023, 00:28

which registry?

Select ‘Yes’ or ‘No’; if ‘Yes’ also give the name of the registry and a link to the registration page for the experiment.

○ 1. yes

● 2. no

Question 3.3.2: How are responses collected?

Describe here the method used to collect responses, e.g. paper forms, Google forms, SurveyMonkey, Mechanical Turk, CrowdFlower, audio/video recording, etc.

Google Sheets spreadsheet exported into CSV and processed.

Section 3.3.3: Quality assurance

Section 3.3.3: Form/Interface

Question 3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations?

Select all that apply:

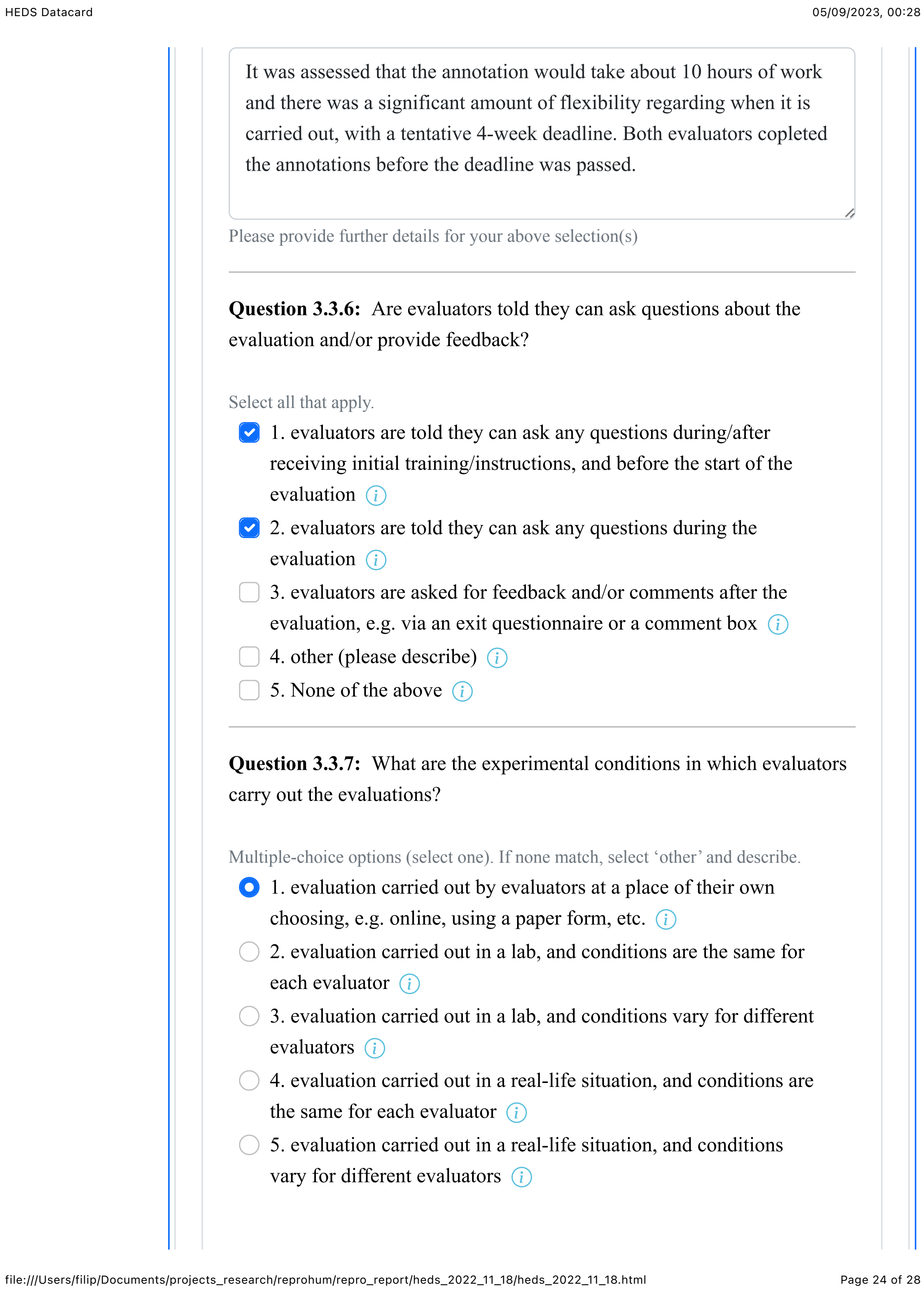
□ 1. evaluators have to complete each individual assessment within a set time ⓘ

□ 2. evaluators have to complete the whole evaluation in one sitting ⓘ

☑ 3. neither of the above (please describe) ⓘ

Please describe:

file:///Users/filip/Documents/projects_research/reprohum/repro_report/heds_2022_11_18/heds_2022_11_18.html Page 23 of 28



HEDS Datacard 05/09/2023, 00:28

It was assessed that the annotation would take about 10 hours of work and there was a significant amount of flexibility regarding when it is carried out, with a tentative 4-week deadline. Both evaluators copleted the annotations before the deadline was passed.

Please provide further details for your above selection(s)

Question 3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback?

Select all that apply.

☑ 1. evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation

☑ 2. evaluators are told they can ask any questions during the evaluation

□ 3. evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box

□ 4. other (please describe)

□ 5. None of the above

Question 3.3.7: What are the experimental conditions in which evaluators carry out the evaluations?

Multiple-choice options (select one). If none match, select ‘other’ and describe.

● 1. evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.

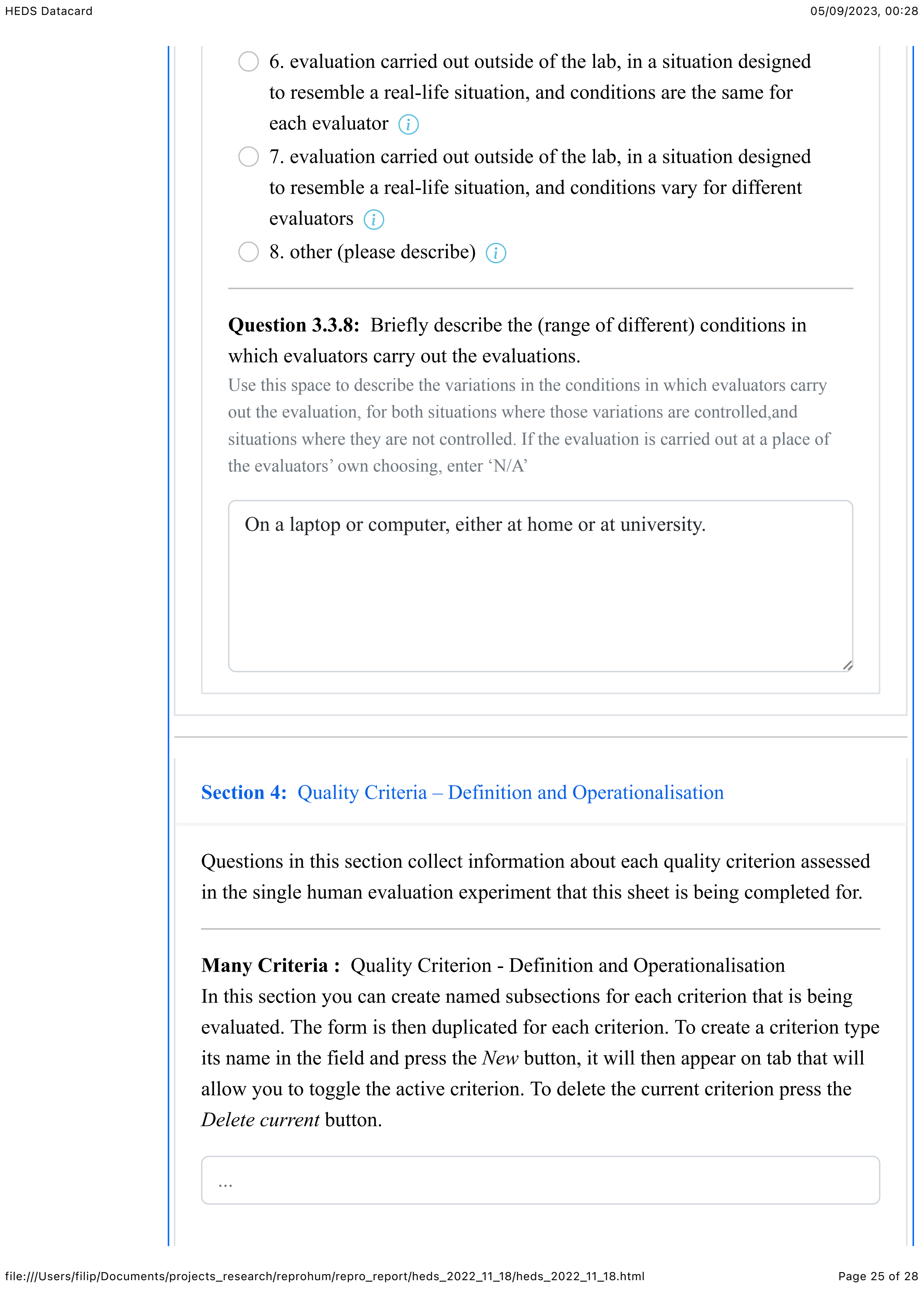
○ 2. evaluation carried out in a lab, and conditions are the same for each evaluator

○ 3. evaluation carried out in a lab, and conditions vary for different evaluators

○ 4. evaluation carried out in a real-life situation, and conditions are the same for each evaluator

○ 5. evaluation carried out in a real-life situation, and conditions vary for different evaluators

file:///Users/filip/Documents/projects_research/reprohum/repro_report/heds_2022_11_18/heds_2022_11_18.html Page 24 of 28



○ 6. evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator ⓘ

○ 7. evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators ⓘ

○ 8. other (please describe) ⓘ

Question 3.3.8: Briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

Use this space to describe the variations in the conditions in which evaluators carry out the evaluation, for both situations where those variations are controlled,and situations where they are not controlled. If the evaluation is carried out at a place of the evaluators' own choosing, enter 'N/A'

On a laptop or computer, either at home or at university.

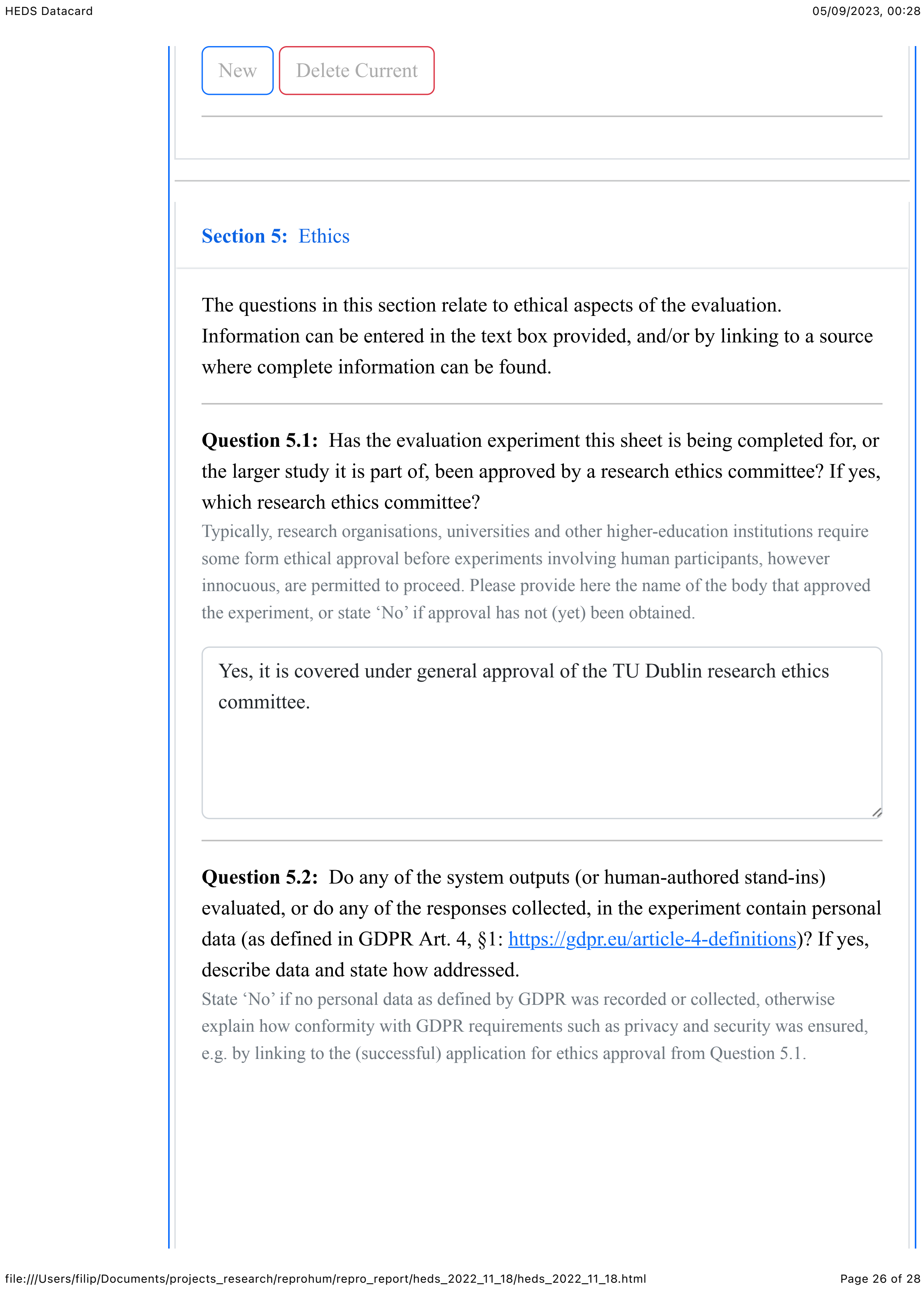
Section 4: Quality Criteria – Definition and Operationalisation

Questions in this section collect information about each quality criterion assessed in the single human evaluation experiment that this sheet is being completed for.

Many Criteria : Quality Criterion - Definition and Operationalisation

In this section you can create named subsections for each criterion that is being evaluated. The form is then duplicated for each criterion. To create a criterion type its name in the field and press the *New* button, it will then appear on tab that will allow you to toggle the active criterion. To delete the current criterion press the *Delete current* button.

...



Section 5: Ethics

The questions in this section relate to ethical aspects of the evaluation. Information can be entered in the text box provided, and/or by linking to a source where complete information can be found.

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

Typically, research organisations, universities and other higher-education institutions require some form ethical approval before experiments involving human participants, however innocuous, are permitted to proceed. Please provide here the name of the body that approved the experiment, or state 'No' if approval has not (yet) been obtained.

Yes, it is covered under general approval of the TU Dublin research ethics committee.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions)? If yes, describe data and state how addressed.

State 'No' if no personal data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements such as privacy and security was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

No.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited)? If yes, describe data and state how addressed.

State ‘No’ if no special-category data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements relating to special-category data was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

No.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

Use this box to describe any *ex ante* or *ex post* impact assessments that have been carried out in relation to the evaluation experiment, such that the assessment plan and process, well as the outcomes, were captured in written form. Link to documents if possible. Types of impact assessment include data protection impact assessments, e.g. under GDPR. Environmental and social impact assessment frameworks are also available.

No.

B Copy of the HEDS sheet

HEDS Datacard 05/09/2023, 00:28

All Form Errors

file:///Users/filip/Documents/projects_research/reprohum/repro_report/heds_2022_11_18/heds_2022_11_18.html Page 28 of 28