Translate your Own: a Post-Editing Experiment in the NLP domain

Rachel Bawden[1] **Ziqian Peng**[2] **Maud Bénard**[3] **Éric de la Clergerie**[1]
Raphaël Esamotunu[3] **Mathilde Huguin**[4] **Natalie Kübler**[3] **Alexandra Mestivier**[3]
Mona Michelot[3] **Laurent Romary**[1] **Lichao Zhu**[3] **François Yvon**[2]

[1]Inria, Paris, France
[2]ISIR, CNRS et Sorbonne Université, Paris, France
[3]CLILLAC-ARP, Université Paris Cité, Paris, France
[4]INIST, CNRS, Nancy, France

Abstract

The improvements in neural machine translation make translation and post-editing pipelines ever more effective for a wider range of applications. In this paper, we evaluate the effectiveness of such a pipeline for the translation of scientific documents (limited here to article abstracts). Using a dedicated interface, we collect, then analyse the post-edits of approximately 350 abstracts (English→French) in the Natural Language Processing domain for two groups of post-editors: domain experts (academics encouraged to post-edit their own articles) on the one hand and trained translators on the other. Our results confirm that such pipelines can be effective, at least for high-resource language pairs. They also highlight the difference in the post-editing strategy of the two subgroups. Finally, they suggest that working on term translation is the most pressing issue to improve fully automatic translations, but that in a post-editing setup, other error types can be equally annoying for post-editors.

1 Introduction

In most, if not all scientific domains, academic communication and publication activities take place mostly in English (Gordin, 2015). While sharing a common language can be viewed as a facilitating factor in many cases, it also generates tensions, frictions and inequalities (Amano et al., 2023), and hinders the exposure of science that is not discussed in English. Furthermore, in non-English-speaking countries, it creates a linguistic barrier between the scientific community and the general public that can only amplify misunderstandings and doubts. These issues have motivated calls for changes as expressed in the "Helsinki initiative".[1] Among the Natural Language Processing (NLP) community, this has motivated the ACL 60-60 special initiative,[2] aimed at using automatic tools (speech recognition, machine translation (MT)) and resources (multilingual term lists) to help remove these barriers.

In this paper, we report our attempts to use existing MT technologies to translate English scientific documents in the NLP domain into French. As has been well documented for the biomedical domain in the course of the challenges organised at the Conference on Machine Translation since 2016 (see (Neves et al., 2023) for the latest published edition), academic texts pose specific translation challenges, related notably to term translation and the generation of lexically consistent outputs.

Our main goal in this work is to evaluate the current state-of-the-art in MT for the translation of academic NLP texts with a view to using MT to aid NLP authors in the translation and post-editing of abstracts in non-English languages. We base our evaluation on manually post-edited documents by two populations of post-editors: apprentice and well-trained professional translators on the one hand and NLP experts (academics) who are encouraged to post-edit their own articles on the other. The results of this pilot study will help us design and organise a large-scale experiment that

[1]https://www.helsinki-initiative.org/
[2]https://www.2022.aclweb.org/dispecialinitiative

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 431–443
June 24-27, 2024 ©2024 European Association for Machine Translation

will ultimately cover all scientific domains. The main questions we aim to answer are the following: (a) what is the effort needed for academics to post-edit automatic translations of texts in their domain of expertise? (b) can we measure the quality of the resulting translations? (c) can we see a difference between existing translation tools? and (d) what are the residual errors that still hinder the translation of academic publications?

To answer these questions, we designed a post-editing protocol aimed to facilitate the voluntary participation of academics in our domain and collected a set of more than 350 abstracts in the NLP domain, which were post-edited once or several times. A large subset of them are also associated with post-editor feedback on the types and severity of errors present. We analysed them in terms of the post-editing effort, measured using HTER (Snover et al., 2006) and studied them in terms of differences in post-editing patterns. We release the resulting corpus and the code for the post-editing interface for future use.[3]

2 Related Work

Numerous challenges are faced when developing and adapting NLP models to scientific texts, including how to handle domain-specific terminology (including acronyms), and how to ensure coherence at the document level. In recent years, the development of such tools has been a growing area of interest for NLP researchers, with multiple models being published for the scientific and scholarly domains, e.g. SciBERT (Beltagy et al., 2019), PubmedBERT (Gu et al., 2021), Galactica (Taylor et al., 2022) and ScholarBERT (Hong et al., 2023). Specifically for MT, there have been several initiatives, including the recent IWSLT shared task on translating ACL presentations (Agarwal et al., 2023; Salesky et al., 2023). The project that is closest to our own is the COSMAT project (Lambert et al., 2012), whose aim was to develop a pipeline for integrating the translation of scientific documents into the HAL[4] archiving platform for English–French translation.

A few corpora are available for scientific document translation, covering different types of publications. The biomedical task at WMT, for instance, has produced parallel test sets for a number of years extracted from article abstracts from PubMed that are available in several languages (Jimeno Yepes et al., 2017; Neves et al., 2018; Bawden et al., 2019; Bawden et al., 2020; Yeganova et al., 2021; Neves et al., 2022; Neves et al., 2023). The SciPar parallel corpus of scientific texts (Roussis et al., 2022) is composed of master's and doctoral theses across several domains and in multiple languages. S2ORC (Lo et al., 2020) is also multi-discipline and contains monolingual English articles from Semantic Scholar. In the NLP domain, Mariani et al. (2019) compiled and explored a large-scale comparable corpus of about 65k NLP papers from multiple sources, while Tanguy et al. (2020) focus on French, providing a monolingual corpus from the TALN conferences.

Evaluating MT for scientific documents is challenging, particularly as standard metrics may well underestimate the impact of mistranslating scientific terminology if they are considered equal to other words. This is particularly the case for simple surface-based metrics such as BLEU (Papineni et al., 2002), but is also a currently unknown factor for other automatic metrics such as COMET (Rei et al., 2020). According to the human evaluations of the WMT biomedical shared tasks, e.g. (Neves et al., 2023), term translation was one of the factors most impacting judgments of quality over other factors such as style and naturalness. Another problem is the scarcity of parallel texts that can be used for reference-based evaluation. Moreover, those that exist may not be perfect translations, either because there is no guarantee that two abstracts for the same paper in multiple languages were intended to be perfect translations or because the authors are non-native speakers of at least one of the languages. This therefore motivates alternative approaches to evaluation, including reference-less evaluation (for automatic evaluation, this would refer to quality estimation (Specia et al., 2010)) and human evaluation, through post-editing or error annotation for example.

Post-editing has previously been used as a means of evaluating MT quality, either through the time taken to render a text to an acceptable standard or (largely related) through the number of changes that were made, which is the basis for the HTER metric ("Human-targeted Translation Edit Rate") (Snover et al., 2006; Dorr et al., 2011). This task-based evaluation strategy is less costly both financially and in terms of effort on the part of translators, and can provide clues as to what types of er-

[3] https://github.com/ANR-MaTOS/Resources
[4] https://hal.science

rors are being produced by MT systems. It is also a realistic setting in many cases, including ours, where MT systems can be used to provide an initial translation of a text that the author can then modify. For example, the previously mentioned COSMAT project aimed to integrate such software into the publishing platform to facilitate the production of texts in multiple languages by the authors.

3 Data Collection

We collect a corpus of over 20k English NLP titles and abstracts that we translate automatically into French and of which a selection is then post-edited. Basic statistics on the most common types of publications included are in Table 1. As shown in Figure 1, the corpus contains titles and abstracts from various publication types, the most common being conference papers, journal articles, book sections, preprints, reports and books. Once the initial corpus extracted (Section 3.1), each of the titles and abstracts is automatically translated into French using three MT systems (Section 3.3), and finally, we collect post-edits of the translations by translators and members of the NLP community (Section 3.4). The research protocol received a positive evaluation from our university institutional review board. All code and the resulting corpora will be made publicly available. The abstracts belong to the metadata of the articles and therefore can be freely distributed.[5]

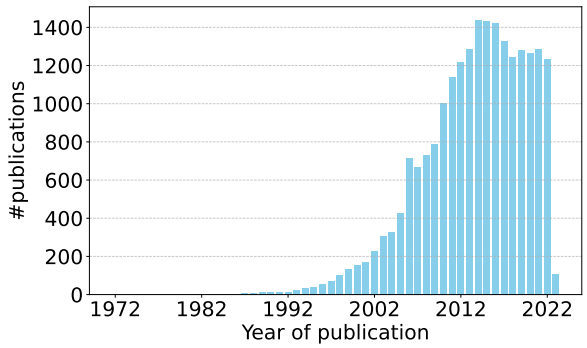


Figure 1: Distribution of publications in the corpus by year.

[5]The metadata of articles published on the HAL platform is under a CC0 licence. It is specified that "HAL's metadata can be consulted in whole or in part by harvesting in compliance with the intellectual property code.", pursuant to the so-called French Law for a Digital Republic [Loi n°2016-1321 du 7 octobre 2016 (art. 30)] see, `https://doc.archives-ouvertes.fr/en/legal-aspects/`

Publication by type	#	Avg. #toks title	Avg. #toks abstract
Total	21,748	9.86	148.81
Conference paper	14,312	9.70	138.16
Journal article	4,362	10.54	178.68
Book section	1,047	9.38	138.91
Preprint	585	9.87	164.23
Report	506	8.80	150.29
Book	271	9.98	152.15
...			

Table 1: Statistics of the initial NLP corpus overall and for the six most common publication types. Tokens here are defined simply as white-spaced delimited sequences of characters.

3.1 Extracting Scientific Abstracts

Our source texts are English titles and abstracts from scientific publications in the NLP domain from the HAL open archive,[6] extracted using the dedicated API. In order to select a maximum number of publications with as few non-NLP publications as possible, we carried out the following steps to extract and filter the data: (i) download data from several domains, included the wide domain of "informatics", (ii) filter to retain only NLP publications, (iii) further filter to remove abstracts that already have a French translation.

Downloading the Data We downloaded the metadata (of which the abstract is one type of information) corresponding to all publications associated with the "computational linguistics" (cs.CL) but also the wider "informatics" domains.

Retaining Only NLP Publications We filter the publications, only keeping those that (i) contain a known keyword in their title, abstract or keyword list or (ii) are published at a known NLP venue.[7]

We check each publication for NLP-specific keywords (in the list of keywords, the title or the abstract). The list was created by taking the set of user-entered keywords for all publications associated with the cs.CL domain, manually filtering it to remove words that could also be relevant to other domains and adding any missing terms based on domain knowledge. This process required manually verifying publications matched with different keywords and removing those that matched with non-NLP publications.

We identify NLP venues by taking the list of conferences, workshops and journals from the

[6]`https://hal.science`

[7]Both the keyword list and venue list can be found at `anonymised-link`.

ACL anthology corpus (Rohatgi, 2022), adding other known venues and augmenting the list by automatically generating variants of the names (in order to match the various ways authors enter venues), e.g. *13th Nordic Conference of Computational Linguistics (NODALIDA 2001)* also results in *13th Nordic Conference of Computational Linguistics* and *NODALIDA*. We match publications based on the presence of one of the identified venues somewhere in their venue names.

Further Filtering Since the aim is to translate the abstracts into French, we target abstracts that were not originally written in French by filtering out those for which a French abstract exists. This follows the approximation that the presence of a French abstract is likely to indicate that the original language was French.

3.2 Available Metadata

For each article, we collect the following information: title, abstract, list of authors, publication type, venue, date of publication, keywords, language of the text, URL to the paper, licence and the reason for the publication being accepted (out of the filters described above).

3.3 Automatic Translation

We translated the titles and abstracts into French using three commercial neural MT systems: DeepL (professional edition, version 7.5),[8] Systran Translate (professional edition)[9] and e-translation (version 12.3).[10] In practice, we concatenated all titles and abstracts into a single file to be translated, separating each article with a token indicating the ID number of the article. We then retrieved the individual translations. Research in contextual MT has shown that when trained properly (this is the case of commercial systems), models have no issue translating multiple sentences at once, especially for short documents such as abstracts (Maruf et al., 2019; Fernandes et al., 2023).

3.4 Manual Post-editing

We developed an online interface to collect post-edits and to provide feedback on MT quality. Users created an account, filling in basic information that could be useful for future research. They then selected articles to post-edit via the interface and finally gave feedback about the experience. We collected post-edits from two types of post-editors: (i) translators and students[11] in translation studies, and (ii) members of the NLP community, who were encouraged, although not forced, to post-edit their own articles.

Post-editor Metadata A condition for participating in the post-editing experiment was fluency in French. Post-editors remained anonymous, but we collected information about their profile that is important for future research, namely their native language(s), other language(s) spoken, the number of years of experience in NLP (<3, 3-10 or 10+) and whether they have previously written an abstract in English and written an abstract in French. We also ask for their general appreciation of MT tools by asking (i) whether they have previously used MT tools to help write scientific articles and (ii) whether they would consider it useful to integrate MT for abstracts into HAL. They can also leave free comments if they wish. Any other information is not available due to anonymity reasons.

Post-editing via the Interface Given that NLP community members were encouraged to post-edit their own publications as experts in the content to be translated, we made sure that they could search the database of publications by ID, keyword (in the title or abstract) and by author name. Otherwise, they could also choose a random publication. To ensure that the same publications were not post-edited too often, publications were presented in a random order in the interface, with a random seed dependent on the ID of the user. Each title and abstract could be post-edited a maximum of three times (once for each MT system). A screenshot of the interface is displayed in Figure 2.

An automatic translation was randomly selected out of the three (the post-editor is unaware of which MT was used). Guidelines were provided on the post-editing page: to modify the text (title and abstract) so that it is clear, understandable and acceptable, as they would do for a journal article written in French. The post-editors could then edit the MT output without a time limit and provide basic feedback on its quality (Figure 3). We also log the time taken to finish post-editing.

[8] https://deepl.com

[9] https://www.systran.net/en/translate

[10] https://webgate.ec.europa.eu/etranslation.

[11] The students worked under the close supervision of their teachers.

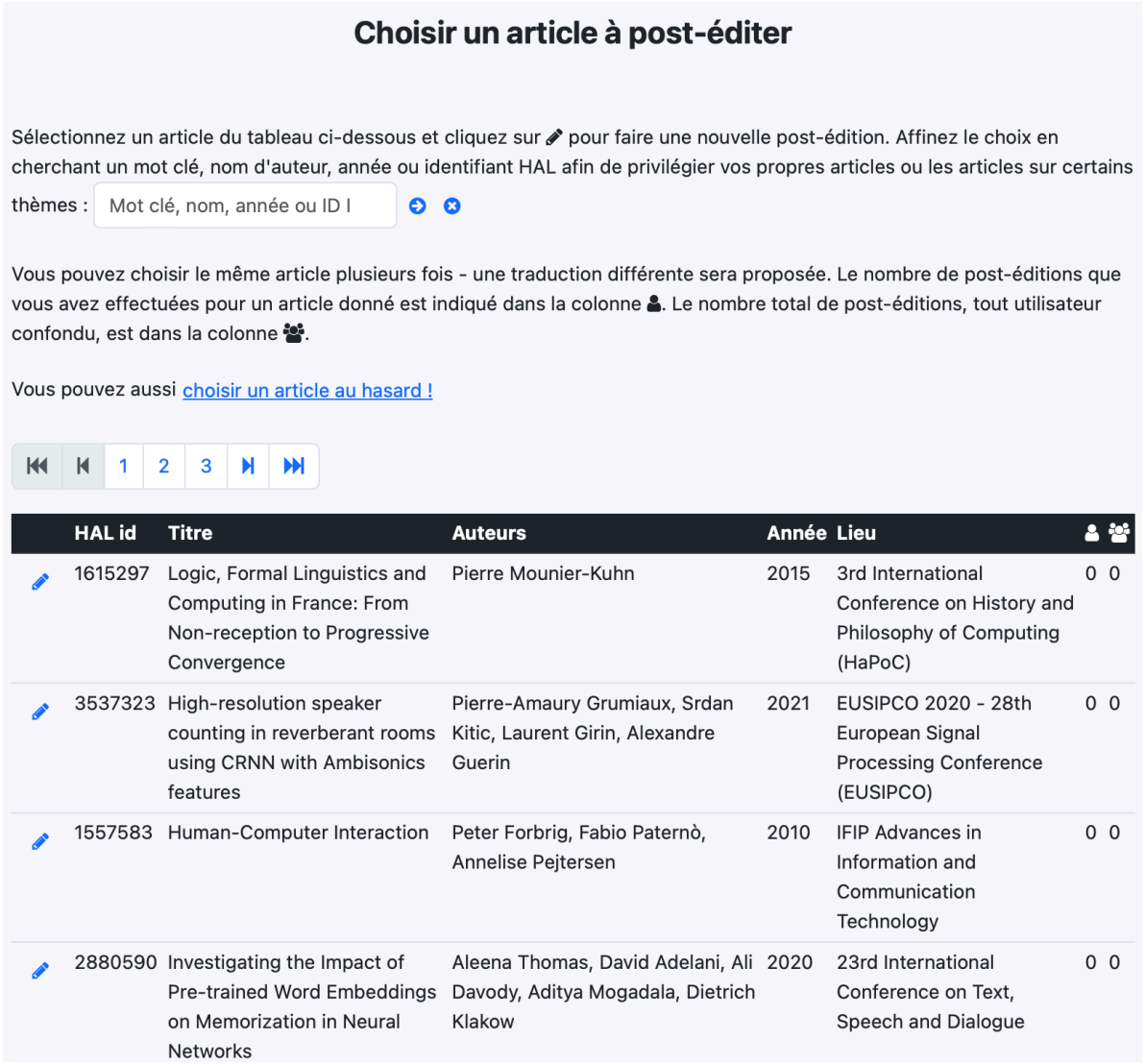


Figure 2: Interface for article selection. The instructions read "Select an article from the table below and click on [the pen emoji] to start a new post-edit. Filter your selection by searching for a keyword, author name, year or HAL ID in order to prioritise your own articles or articles from certain themes. [...] You can choose the same article several times - a different translation will be given. The number of times a given article has been post-edited by you is indicated in the column [with the person emoji]. The total number of times it has been post-edited by all users is given in the column [with the people emoji]."

Post-editor Feedback The type of feedback differs depending on the profile of the post-editor. For members of the NLP community, they indicate a single feedback score corresponding to the question "What importance do you give to the MT problems seen?" (as shown in Figure 3), with possible responses "No problem", "Not very serious (spelling, punctuation, etc.)", "moderately serious (not interfering with comprehension but not linguistically or stylistically acceptable)" and "serious (interfering with understanding, not faithful to the source)". As NLP experts are not specialists in manual error annotation, they were presented with four easy-to-use categories. For translators, the question is more detailed, asking for each error type (faithfulness, grammar, terminology, spelling and punctuation, style, document coherence) whether the problems seen correspond to the same four degrees of quality ("No problem", "not serious", moderately serious" or "serious"). In both cases, post-editors can leave a free form comment. The error categories were defined based on the MeLLANGE error typology (Kübler, 2008).

4 NLP Post-edit Corpus

In Table 2, we report basic statistics concerning the post-editing corpus, for documents post-edited by the community (by NLP researchers), for documents post-edited by translator and for two categories combined (all). Given that a single abstract

Figure 3: Example of the post-editing interface (NLP community member view). The instructions read "Modify the text (title and abstract) so that it is clear, understandable and acceptable, as you would do for a publication in a French journal (e.g. the TAL journal). While post-editing, please do not use machine translation tools. If possible, please complete your post-edition without interruptions so that the registered duration corresponds to the actual time to post-edit... Warning: if you leave this page (by closing the window or going back to the previous page), you will lose your modifications.". Post-editing is performed without prior sentence segmentation and does not assume that the source and target texts have matching number of sentences.

can be translated multiple times (using different MT systems), we distinguish the statistics concerning the number of abstracts that have been post-edited and the number of translations that have been post-edited (a translation being specific to a particular abstract).

Type	comm.	trans.	all
PEs	95	242	337
Abstracts w/ PEs	91	241	322
Translations w/ PEs	73	240	313
Abstracts w/ several PEs	17	2	55
Translations w/ several PEs	4	1	30

Table 2: Basic statistics concerning the number of post-editions (PEs) by NLP **comm**unity members, by **trans**lators and by either group (**all**). Among abstracts and translations with several PEs, 46 distinct abstracts are post-edited by both groups, and 28 different translations are post-edited by both.

Concerning the post-editors, there were 4 translators (3 of whom were native French speakers) and 16 NLP experts (13 of whom were native French speakers) and whose experience in NLP ranged from 10+ years (4 users), to 3-10 years (7 users), to under 3 years (5 users).

5 Analysis of the Post-edit Corpus

5.1 Evaluation setup

We primarily base our evaluation of post-editing efforts on the computation of HTER (Human Translation Edit Rate) (Snover et al., 2006), which corresponds to a modified edit distance between the automatic translation and its revised version. We compute HTER with SacreBLEU's implementation[12] (Post, 2018). Scores are computed separately for each whole abstract (viewed as one long line of text) then broken down by post-editor type and averaged over the corresponding documents.

We also report BLEU score differences between the original and modified abstracts, also computed with SacreBLEU,[13] in order to judge how much or little the translations had to be edited to be deemed acceptable. These scores rely on corpus-level statistics, again computed on a per-document basis.[14] Measures of post-editing time were also recorded, but we deem them insufficiently reliable in our experimental setting to perform any fine-grained analysis.

To evaluate the quality of translation without references, we use Comet-QE (Rei et al., 2020), which relies on distances between continuous space representations of source and target texts.[15]

5.2 Results

A first observation is that in our conditions, the automatic translations are mostly of high quality, with an average HTER of 10.7 (BLEU=85.6). Another indication of this high quality is that 13 documents (out of 337) were left entirely unchanged. For the NLP community group, revising an abstract took less than 10 minutes on average.

Comparing the community and translators A more detailed analysis of the post-editing results is illustrated in Figure 4. Two interesting trends can be seen: (a) the distribution of efforts is more concentrated for translators than for the NLP community, (b) the translators also tend to make smaller changes to the translation than the NLP community (HTER=8.0 vs. HTER=18.2), a quite significant difference. This is also obvious when considering the 90% percentile of HTER values (17.2 vs. 32.7). These differences may reveal differences in the way the task was perceived by each population: while translators tend to follow established post-editing guidelines and remain as close as possible to the original MT, field experts are more inclined to rewrite substantial portions of the abstracts.

Without human references, it is difficult to assess the quality of the resulting translations. Computing Comet-QE scores before and after post-editing however reveals a very small improvement (see Table 3). This hints at the lack of sensitivity of QE scores for high-quality translations.

	MT outputs	Post-editions
Translators	76.3	77.0
NLP experts	77.8	78.6

Table 3: Comet-QE(x100) scores of MT outputs and their post-edited versions for each group of post-editors.

Another measure is to take the professional translations as references for the 28 MT outputs post-edited by both groups. For this subset of abstracts, the BLEU score is 76.7 (HTER=18.4).

[12]Version 13.5.

[13]We use the default signature for BLEU: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1.

[14]Sometimes referred to as "document-level" BLEU. Note that BLEU scores cannot be used here to compare translation quality across populations, given that they do not use the same set of references.

[15]The model is `Unbabel/wmt22-cometkiwi-da`. See `https://unbabel.github.io/COMET/`.

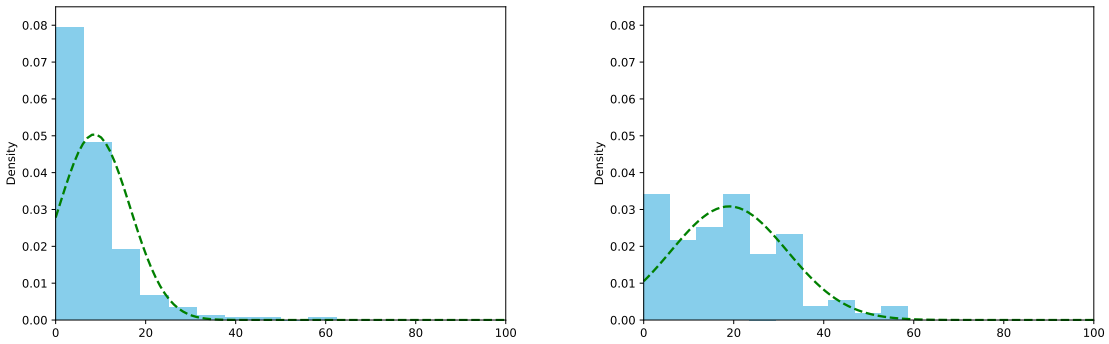


Figure 4: Distribution of HTER scores for translators (left) and NLP experts (right).

Comparison of MT systems We now turn to our third question, which concerns the differences between MT systems. Table 4 reports the average scores for each system and post-editor group and Figure 5 plots the corresponding distributions.

Group	DeepL	eTranslation	Systran
Translators	90.4/7.1 (N=90)	85.9/10.6 (N=79)	87.3/8.5 (N=73)
Experts	81.7/13.8 (N=34)	68.3/24.4 (N=28)	73.1/19.7 (N=33)

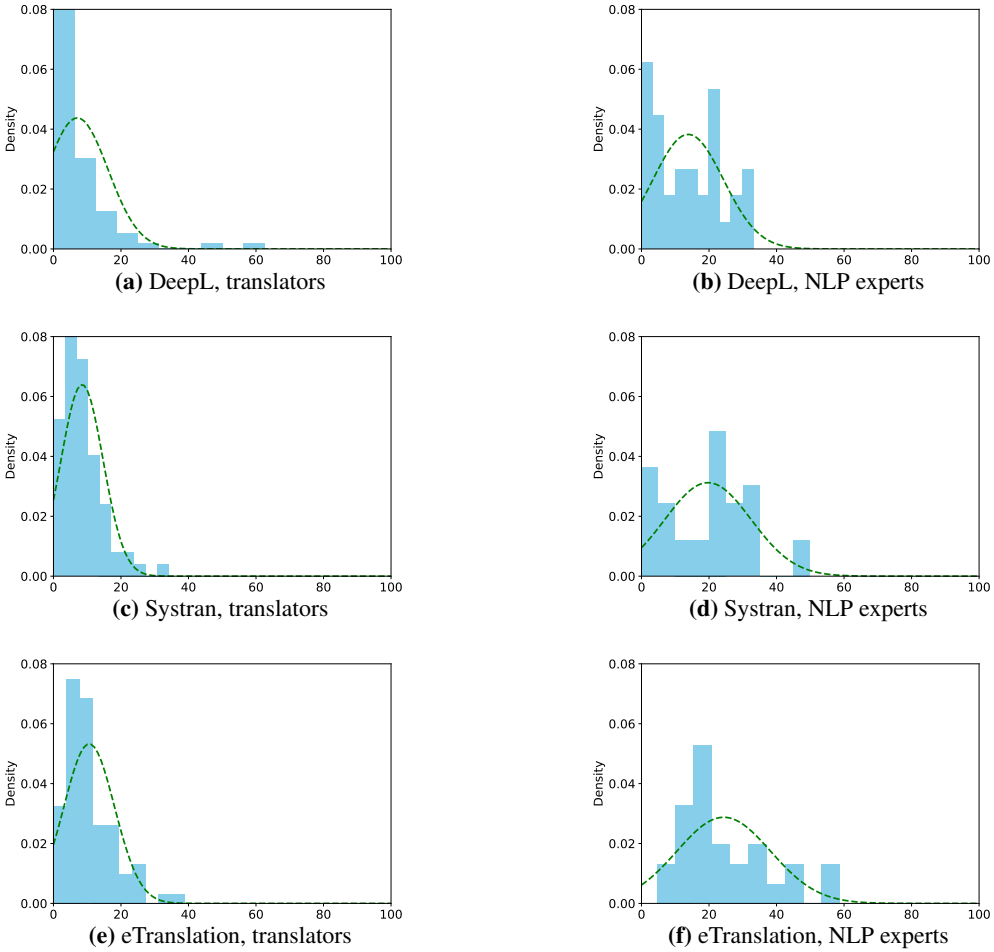
Table 4: BLEU/hTER for each post-editor group and system. The number of abstracts for each category is given in brackets.

The scores in Table 4 show clear preferences, with DeepL yielding the smallest post-editing effort, while e-Translation consistently leads to more corrections. These differences are particularly strong for the NLP experts group. These observations are only partly confirmed by a two-sided student T-test for all pairs of systems: out of 6 comparisons, the only significant differences at p=0.05 are for DeepL vs. eTranslation for both groups, while Systran cannot be viewed as significantly worse than DeepL, nor significantly better than eTranslation.

Qualitative analysis of errors For the translator group, we analyse the post-editor feedback concerning translation errors for the 7 broad error categories introduced in Section 3.4. We report the corresponding statistics in Table 5.

A first observation is the consistency of these judgements: for each error type, more severe errors tend to yield more edits, with some small inconsistencies (e.g. terminology errors with severity 2 and 3). Looking now at error types, we see that that grammar, style, and punctuation errors are mostly associated with the lowest level of severity. This is expected given the very high fluidity of MT outputs. The same trend is observed for faithfulness and coherence errors, which tend to get rarer as the severity level increases. Terminology errors exhibit the reverse trend and are mostly associated with the highest level of severity. However, looking now at the post-editing effort, we observe at all severity levels that fixing term errors always yields the lowest HTER scores, while fixing grammar errors almost always yields the highest ones.

Qualitative differences between groups Finally, we carry out a small qualitative analysis of the way the two groups post-edit MT. A few interesting examples are given in Table 6, corresponding to cases where a) one group left the MT output unchanged while the other had high HTER and BLEU scores at the sentence level or b) both groups had high but different HTER and BLEU scores at the sentence level. We note that, while there are a few cases in which the translators corrected an MT error that the community seem to overlook ("traduction automatique de neurone" (literally *machine translation of neurons*) in Example 3), in most cases, the community group seems to produce better post-edited texts than the translators. NLP experts seem to better master specialised terminology ("analyse syntaxique en constituants lexicalisés" in Example 2), specialised phraseology (e.g. "les modèles sont entraînés" *models are trained* instead of "les modèles sont formés", literally *models are educated*, in Example 4 (source text: "All our models are trained without the need of cross-modal labeled translation data.")), as well as domain conventions (in Example 1 the acronym "CoMMuTE" is associated with



(a) DeepL, translators **(b)** DeepL, NLP experts

(c) Systran, translators **(d)** Systran, NLP experts

(e) eTranslation, translators **(f)** eTranslation, NLP experts

Figure 5: Distribution of HTER scores across systems and post-editor groups.

	Severity level											
	1			2			3			4		
Problem	N	BLEU	hTER	N	BLEU	hTER	N	BLEU	hTER	N	BLEU	hTER
Grammar	148	90.2	7.0	67	84.8	10.7	14	86.4	10.3	12	79.5	16.4
Spelling & Punct.	130	89.5	7.4	60	86.2	10.0	35	87.4	8.8	16	82.8	13.7
Document	127	91.2	6.0	43	84.9	10.7	39	85.2	10.9	33	82.7	13.5
Style	68	91.1	6.1	84	88.7	8.1	57	85.9	10.4	32	82.7	12.7
Faithfulness	123	92.1	5.4	58	84.9	10.4	35	80.8	14.7	25	84.5	12.4
Terminology	34	95.0	3.2	43	88.4	8.3	73	87.5	8.9	91	85.4	10.7

Table 5: Post-edition efforts evaluated according to the number (N, left), BLEU (middle) and hTER (right) of translations associated with different severity levels (from 1 to 4) for translation problems reported by translators in their feedback.

the full term in brackets, which better conforms to the domain conventions than the solution the translator adopted, i.e. translating the full term). Experts also seem to take more freedom in rearranging constituents and rewriting sentences (Example 5), where translators seem to follow the source sentence structure more closely, a behaviour that is also reflected in the automatic metric scores.

6 Conclusion and Future Work

In this paper, we report the results of a pilot study aimed at evaluating the quality of commercial MT systems for scholarly documents (abstracts) in the NLP domain (for English→French). This study explores a realistic scenario, where domain experts post-edit in their mother tongue their own texts (in English, supposedly their L2). We compare against the use of translators with a partial knowledge of the target domain to perform the same task.

MT	NLP experts post-edits	Translator's post-edits
1- We also release CoMMuTE, a Contrastive Multilingual Multimodal Translation Evaluation dataset, composed of ambiguous sentences and their possible translations, accompanied by disambiguating images corresponding to each translation.		
Nous publions également CoMMuTE, un ensemble de données d'évaluation de la traduction multimodale multilingue contrastive, composé de phrases ambiguës et de leurs traductions possibles, accompagnées d'images désambiguïsantes correspondant à chaque traduction.	Nous publions également le jeu de données CoMMuTE (Contrastive Multilingual Multimodal Translation Evaluation), composé de phrases ambiguës et de leurs traductions possibles, accompagnées d'images visant à leur désambiguïsation et correspondant à chaque traduction.	Nous publions également CoMMuTE, un ensemble de données d'évaluation de la traduction multimodale multilingue contrastive, composé de phrases ambiguës et de leurs traductions possibles, accompagnées d'images désambiguïsantes correspondant à chaque traduction.
2- Multilingual Lexicalized Constituency Parsing with Word-Level Auxiliary Tasks		
Lexicalized Constituency Parsing multilingue avec des tâches auxiliaires de niveau Word	Tâches auxiliaires au niveau des mots pour l'analyse syntaxique en constituants lexicalisés multilingue	Analyse syntaxique de constituants lexicaux multilingues avec tâches auxiliaires au niveau des mots
3- Priming Neural Machine Translation		
Amorçage de la traduction automatique de neurones	Amorçage de la traduction automatique de neurones	Amorçage de la traduction automatique neuronale
4- All our models are trained without the need of cross-modal labeled translation data.		
Tous nos modèles sont formés sans avoir besoin de données de traduction étiquetées intermodales.	Tous nos modèles sont entraînés sans que des données de traduction intermodales annotées soient nécessaires.	Tous nos modèles sont formés sans avoir besoin de données de traduction étiquetées intermodales.
5- On the SPMRL dataset, our parser obtains above state-of-the-art results on constituency parsing without requiring either predicted POS or morphological tags, and outputs labelled dependency trees.		
Sur l'ensemble de données SPMRL, notre analyseur obtient ci-dessus des résultats de pointe sur l'analyse des circonscriptions sans nécessiter une prévision de POS ou d'étiquettes morphologiques, et des sorties marquées d'arbres de dépendance.	Sur l'ensemble de données SPMRL, notre analyseur obtient des résultats supérieurs à l'état de l'art en analyse syntaxique en constituents sans nécessiter de parties du discours prédites ni d'étiquettes morphologiques prédites, et permet de construire des arbres syntaxiques en dépendances étiquetées.	Sur l'ensemble de données SPMRL, notre analyseur obtient des résultats supérieurs à l'état de l'art sur l'analyse des constituants sans nécessiter de prédiction des parties du discours ou des étiquettes morphologiques, ni des sorties marquées d'arbres de dépendance.

Table 6: Comparison of experts' and translators' post-edits. Source texts are shown in grey.

Using a dedicated interface adapted for the two populations of post-editors, we collected and analysed approximately 350 abstracts and their post-edited versions. Our main result is that the automatic outputs are already quite satisfactory, as acknowledged by a low average post-editing effort (see also (Sebo and de Lucia, 2024)). We also observed that domain experts tend to deviate more from the original text than translators, the two categories displaying different patterns of post-edits. This study also confirmed the prevalence and severity of terminology errors, while other error types are comparatively rarer or less severe. All resources and analyses will be released to the community.

In the future, we plan to both continue analysis of the data, in particular concerning term use and to reproduce this small-scale experiment with another group of academics from a different scientific background. This will however require finding better ways to incentivise researchers to participate in post-editing activities.

Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) under the project MaTOS - "ANR-22-CE23-0033-03". R. Bawden's partcipation was also partly funded by her chair in the PRAIRIE institute funded by the ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Agarwal, Milind, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri,

Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online), July. Association for Computational Linguistics.

Amano, Tatsuya, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, and Diogo Veríssimo. 2023. The manifold costs of being a non-native English speaker in science. *PLoS biology*, 21(7):e3002184, July.

Bawden, Rachel, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy, August. Association for Computational Linguistics.

Bawden, Rachel, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online, November. Association for Computational Linguistics.

Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Dorr, Bonnie, Joseph Olive, John McCary, and Caitlin Christianson, 2011. *Machine Translation Evaluation and Optimization*, pages 745–843. Springer New York, New York, NY.

Fernandes, Patrick, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada, July. Association for Computational Linguistics.

Gordin, Michael D. 2015. *Scientific Babel How Science Was Done Before and After Global English*. University of Chicago Press.

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct.

Hong, Zhi, Aswathy Ajith, James Pauloski, Eamon Duede, Kyle Chard, and Ian Foster. 2023. The diminishing returns of masked language models to science. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1270–1283, Toronto, Canada, July. Association for Computational Linguistics.

Jimeno Yepes, Antonio, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kübler, Natalie. 2008. A Comparable Learner Translator Corpus: creation and use. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora*, pages 73–78, Marrakech, Morocco.

Lambert, Patrik, Holger Schwenk, and Frédéric Blain. 2012. Automatic translation of scientific documents in the HAL archive. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3933–3936, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983, Online, July. Association for Computational Linguistics.

Mariani, Joseph, Gil Francopoulo, and Patrick Paroubek. 2019. The NLP4NLP corpus (i): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3.

Maruf, Sameen, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *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 3092–3102, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Neves, Mariana, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels, October. Association for Computational Linguistics.

Neves, Mariana, 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 (WMT)*, pages 694–723, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Neves, Mariana, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore, December. Association for Computational Linguistics.

Papineni, Kishore, 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, July. Association for Computational Linguistics.

Post, Matt. 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, October. Association for Computational Linguistics.

Rei, Ricardo, 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, November. Association for Computational Linguistics.

Rohatgi, Shaurya. 2022. Acl anthology corpus with full text. Github.

Roussis, Dimitrios, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France, June. European Language Resources Association.

Salesky, Elizabeth, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online), July. Association for Computational Linguistics.

Sebo, Paul and Sylvain de Lucia. 2024. Performance of machine translators in translating French medical research abstracts to English: A comparative study of DeepL, Google Translate, and CUBBITT. *PLOS ONE*, 19(2):1–13, February. Publisher: Public Library of Science.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50, March.

Tanguy, Ludovic, Cécile Fabre, and Yoann Bard. 2020. Impact de la structure logique des documents sur les modèles distributionnels : expérimentations sur le corpus TALN (impact of document structure on distributional semantics models: a case study on NLP research articles). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL,*

22e édition). Volume 2 : Traitement Automatique des Langues Naturelles, pages 122–135, Nancy, France, 6. ATALA et AFCP.

Taylor, Ross, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Yeganova, Lana, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online, November. Association for Computational Linguistics.