Evaluating Machine Translation of Latin Interjections in the *Digital Library of Polish and Poland-related News Pamphlets*[⋆]

Maciej Ogrodniczuk[1][0000−0002−3467−9424]
and Katarzyna Kryńska[2][0000−0003−3446−122X]

[1] Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
maciej.ogrodniczuk@ipipan.waw.pl

[2] Institute of Polish Language, Polish Academy of Sciences
al. Mickiewicza 31, 31-120 Kraków, Poland
katarzyna.krynska@ijp.pan.pl

Abstract. In *The Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries* a small fraction of items contains manually created Latin–Polish dictionaries explaining Latin fragments injected into Polish content. At the same time, rapid development of machine translation creates new opportunities for creating such dictionaries automatically. In this paper, we verify whether existing translation solutions are already capable of generating useful results in this Latin-Polish setting. We investigate two systems available for this language pair: the familiar Google neural engine and the GPT-3 model, then we test the translation of isolated and context-embedded phrases and evaluate its results with both automatic and human metrics: BLEU and White's 5-point scale of adequacy and fluency.

Keywords: digital library, machine translation, evaluation, middle Polish

1 Introduction

The process of injecting Latin fragments into Polish texts was a frequent practice in the 16th to 18th centuries. In *The Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries* (Pol. *Cyfrowa Biblioteka Druków Ulotnych polskich i Polski dotyczących z XVI, XVII i XVIII wieku*, hereafter abbreviated CBDU[3] [5, 9, 10], containing pre-press documents from this period such Latin interjections constitute 7–8% of the textual content of

[⋆] The work was financed by a research grant from the Polish Ministry of Science and Higher Education under the National Programme for the Development of Humanities for the years 2019–2023 (grant 11H 18 0413 86, grant funds received: 1,797,741 PLN).

[3] See https://cbdu.ijp.pan.pl/.

the print[4]. While still understandable for people familiar with Latin and experts, such passages may be difficult to decipher by a non-specialist reader. For this reason, manually created dictionaries of Latin terms and phrases were added to prints. Unfortunately, due to financial constraints of the original project, only 23 prints were annotated in this way with a total of 600 entries.

Today, with new machine translation developments doing reasonably well for modern languages and available also for Latin, we might try to use them in supplementing the previous work. Not only new prints containing Latin can be translated but also existing dictionaries can be extended with missing words and phrases, previously regarded as easy to understand (which is not always the case). But a prerequisite for starting this process should be verification of the quality of existing Latin-to-Polish machine translation solutions. To our best knowledge, no figures have been so far reported for this language pair.

What is important to note is that we do not intend to create a new state-of-the-art Latin-to-Polish translation engine but evaluate existing out-of-the-box solutions on the available manually created dictionaries and to suggest customizations which could improve their usage.

2 Data Preparation

Our translation dataset was based on the manually created Latin-Polish dictionaries attached as metadata to some prints which can already serve as ready-to-use evaluation data. Out of 23 prints with such dictionaries only 21 prints contained Polish equivalents of Latin phrases (with two dictionaries preserving only Latin and no Polish, presumably marked for translation but never completed[5]) and one duplicate, i.e. the same four entries, both Latin originals and translations, in the same context[6]). This filtering step resulted in 566 entries from 20 prints.

The entries were pre-processed to remove line breaks, editorial comments (such as uncertainty signalled with a question mark in brackets or a reference to a particular book in the bible which was the source of the citation). This process also resulted in several corrections in the library, e.g. for one print the dictionary of difficult old Polish words (with explanations in Polish) was found to be swapped with the Latin-Polish dictionary and one Latin entry contained an excessive fragment in Polish. Additionally, seven obvious typos in the Polish texts were corrected.

The dictionary-based dataset could be successfully used for the translation of isolated Latin fragments but we already planned to carry out a context-wise experiment, i.e. test whether the availability of a larger context could improve

[4] Calculated based on a sample of 18 transcribed prints containing Latin-Polish dictionaries; 1895 words out of 24506.

[5] See `https://cbdu.ijp.pan.pl/id/eprint/700/` and `https://cbdu.ijp.pan.pl/id/eprint/2250/`

[6] See `https://cbdu.ijp.pan.pl/id/eprint/4210/` and `https://cbdu.ijp.pan.pl/id/eprint/4220/`.

translation quality. To be able to do that, we needed larger contextual fragments of dictionary phrases which were not available in the CBDU digital library. At the same time, some full texts from the library were transcribed in another project, intended to create *The Electronic Corpus of the 17th and 18th Century Polish Texts (until 1772)* (Pol. *Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do roku 1772)*) [4], also referred to as the *Baroque Corpus* (Pol. *Korpus Barokowy* – hence its acronym KORBA[7], Out of 20 prints with Latin dictionaries available in the digital library, only 18 have been transcribed for the corpus so we limited our further experiments to these prints to be able to perform the evaluation on a fixed dataset. As a result, the final number of samples in the set dropped to 553. For completeness, it is worth noting that for some prints their content was retrieved from duplicates, i.e. different editions of the same content, registered in CBDU as separate prints[8].

The texts were available in TEI P5 XML, following the format of the National Corpus of Polish [13], with Latin parts marked with `<foreign>` tags which were removed before running the translation. Similarly, when two variants of the same fragment have been available (in `<choice>` elements), their regularized versions have been used (cf. `<orig>corporisþ</orig>` vs. `<reg>corporisque</reg>`).

After stripping XML tags from the corpus files, the Latin dictionary fragments were located using fuzzy matching[9] which was very effective in neglecting minor variations between the transcribed content of the print and the dictionary entry, cf. *O Passi graviora* vs. *O Passi grauiora* or parts of content removed from the Latin entry, usually with a Polish interjection, marked with an ellipsis character, cf. *prosequi … Hostem* corresponding to *prosequi nie mogli Hostem*. The method, apart from producing offsets of the closest match found, also returns Levenshtein edit distance between both matching parts so errors could be easily tracked. In some cases, however, the difference between the dictionary entry and the corpus text made even the fuzzy match impossible, e.g. when acronyms were used in the text but were already expanded in the dictionary (e.g. *jurium coaequationis M. D. L. cum Regno* shown as *jurium coaequationis Magni Ducatus Lituaniae cum Regno*). During the pre-processing step, such abbreviations were inserted into corpus texts in their expanded form.

Another set of decisions had to be made on the size of the context. The most straightforward scope would be a complete syntactically valid unit containing the phrase being translated. This is not easy to obtain for an old Polish text so initially we decided to extract the left and right context surrounding the phrase or sentence of arbitrary 50-character length and then cut it at potentially syntactically valid places, such as commas, widening the window appropriately. Since

[7] See also `https://korba.edu.pl/overview?lang=en`.

[8] See e.g. `https://cbdu.ijp.pan.pl/id/eprint/3760/`, `https://cbdu.ijp.pan.pl/id/eprint/3770/` and `https://cbdu.ijp.pan.pl/id/eprint/3780/` with the content available for the first one and the dictionary present only for the last one or a similar case with `https://cbdu.ijp.pan.pl/id/eprint/13880/` and `https://cbdu.ijp.pan.pl/id/eprint/13890/`.

[9] With the `fuzzy_index` function from the `Text::Fuzzy` Perl module.

this procedure became overly complicated, we got back to a simpler solution of passing full paragraphs to the translation module. A positive side-effect of this step was that paragraphs containing many interjections could only be translated once.

Still, extracting full paragraphs resulted in lengthy passages, sometimes exceeding 5 000 characters (while e.g. the GPT-3 Davinci model is capable of processing up to 4 000 tokens per request) which introduced more complications without any observable gains (the translation did not look better as compared to shorter passages). So we reverted to the idea of extracting full sentences containing Latin phrases being translated and one sentence before and after the full translated sentences. In some cases, the contexts were manually fine-tuned (when e.g. a full stop did not finish a sentence but an abbreviation). The length of the largest context extracted using this method was 700 characters.

3 Translation Setting

Our translation experiments have been concentrated along two axes: the approach used to machine translation and the size of context needed for using the model successfully. The most popular engine offering Latin to Polish translation is Google Translate, also powering numerous “independent” interfaces available on the Web. The Google API[10] is used e.g. by TranslateKing[11], TranslateHub[12] or Translatiz[13]. A similar model is used e.g. by Yandex API[14], adapted by e.g. ContDict[15], Latin Online[16] or Lingvanex[17] but after preliminary tests we decided to use only Google because of its much better quality than the competitive systems. The experiment used standard Google Translate API calls.

A competitive approach is offered by a recently popular paradigm of prompting large general-purpose generative language models for various tasks, including translation, so we also decided to test the OpenAI’s Davinci GPT-3 large-scale language-generation approach with its default settings[18]. GPT-3 prompts were constructed in the form of the static request: *Przetłumacz łacinę na polski:* (*Translate Latin to Polish:*) *”entry”:*.

The scope-based analysis is exploiting the expectation that taking into account contextual information from the text in which the phrase or sentence is embedded might bring more successful results than isolated translation (i.e. just a given phrase or sentence). This might be particularly true for the multilingual generative models which are known for the ability to reuse foreign-language

[10] https://translate.google.pl/?sl=la&tl=pl

[11] https://translateking.com/

[12] https://livetranslatehub.com/

[13] https://translatiz.com/

[14] https://translate.yandex.com/?lang=la-pl

[15] https://www.contdict.com

[16] https://www.latin-online-translation.com/

[17] https://lingvanex.com/demo/

[18] See e.g. its openly accessible ”playground” https://beta.openai.com/playground

hints, well represented in the main text of our prints. In our experiments, we will try to validate this anticipation by comparing the results of isolated and contextual translation.

In the context-less setting the entries have been fed one by one to the translation engine and results collected.

4 Evaluation

Two basic types of evaluation of machine translation results are usually being carried out: automatic and human evaluation. Below we present both, realizing that each of them has its advantages and flaws: automatic evaluation being low-cost and consistent but not necessarily reliable while human evaluation being more meaningful but also time-consuming and subjective.

For evaluation, we randomly selected 100 entries, following the proportion of phrases to full sentences from the complete dataset (95 entries being phrases). When several translation variants were given, they were treated equally during evaluation.

4.1 Automatic Evaluation

A commonly used method to evaluate machine translation automatically is BLEU (BiLingual Evaluation Understudy) [11], an n-gram precision-based metric. Even though numerous other metrics (cf. NIST [2], METEOR [7], TER [14] or chrF [12]) are also used, many new ones are proposed almost every year[19] and despite the well-known problems with BLEU, the machine translation community still uses it as the primary measure of translation quality so we decided to adopt it for our evaluation.

Table 1 presents the results of the automatic evaluation.

Table 1. Automatic evaluation results

Setting	**Engine**	**Cumulative BLEU score**			
		1-gram	**2-gram**	**3-gram**	**4-gram**
Isolated	Google	26.22	18.77	**14.74**	**12.58**
	GPT-3	26.27	16.45	11.35	8.33
Context-wise	Google	23.95	15.46	11.63	8.85
	GPT-3	**31.83**	**20.18**	13.45	9.54

Even though the higher presence of unigrams and bigrams in the translation puts the context-wise GPT-based solution in the best position, the most

[19] See e.g. chapter 6 of [8] for more examples and [6] for evaluation of the correlation of various metrics with human judgements.

frequently used 4-gram BLEU score is the highest for isolated Google solution. Still, what is slightly discouraging, it does not reach 20 points regarded as the minimum value required for getting the gist of the text.

However, it is worth noting that BLEU only measures direct word-to-word similarity and good translations using e.g. synonyms get poor scores because they cannot be matched in the reference text. This is where human evaluation can show its usefulness.

4.2 Human Evaluation

Apart from automatic evaluation, we decided to carry out the human evaluation of translation results to compensate for BLEU deficiencies (e.g. in dealing with synonyms), compare how both sets of results correlate and to better assess various properties of the translation. Again, numerous methods of human evaluation have been proposed[20], with the most straightforward one using a 5-point scale of adequacy and fluency [16] (see Table 2). Adequacy is related to the adherence of the translation to the source text while fluency grades the quality of the target text only. Contrary to more detailed evaluation schemes, such as iSTS (Interpretable Semantic Textual Similarity, see [1]) or detailed error typologies (see. e.g. [15]) it requires a moderate evaluation effort while still measuring the two most important properties of the translation.

Table 2. 5-point scale of adequacy and fluency, based on [16] (Table 3)

Value	Adequacy	Fluency
5	all meaning	flawless Polish
4	most meaning	good Polish
3	much meaning	non-native Polish
2	little meaning	disfluent Polish
1	none	incomprehensible

The evaluation has been carried out by a single translator proficient in Latin and Polish. Table 3 presents its overall results for both translation engines and translation scopes which show relatively high fluency (good to flawless) while retaining much meaning. What is surprising, adding context does not always help with achieving better translation adequacy and always impairs fluency. The adequacy values seem to correlate with automatic results.

[20] See their review e.g. in the Related Work section of [3].

Table 3. Human evaluation results

Setting	**Engine**	**Adequacy**	**Fluency**
Isolated	Google	3.11	4.63
	GPT-3	2.87	4.52
Context-wise	Google	2.61	4.02
	GPT-3	3.00	4.32

5 Qualitative Error Analysis and Translators' Notes

5.1 Analysing Adequacy and Fluency

Passages not translated by the machine translator, whether left in Latin or omitted, were rated lowest on the evaluation scale (1) as *none* or *incomprehensible* in both adequacy and fluency categories.

In many cases, a high degree of fluency can be observed, which we judge following Maučec and Donaj: *"When judging fluency, the source text is not relevant. The evaluators have access to only the translation being judged and not the source data"*[21] while the quality of adequacy in these cases is low. This means that automatic translators mostly do well in generating grammatically and logically correct translations, but often meaningfully distant from the desired version.

The high frequency of adequate translation fluency is influenced by the occurrence of single-word macaronisms in sentences. In the case of a single word translations (e.g. *Februarii*, *discernere*, *protrahere*) or short popular phrases (e.g. *prima die*, *in absentia*), the highest (5) degree of fluency in all four categories is usually recorded, unless the translator leaves a word untranslated in some cases (e.g. *periculum*, *Julii*). However, cases can be noted where the MT solutions translated a stand-alone Latin word correctly, but failed to translate it in context, e.g. *decernemus* (*we decide*) was translated as *współbędźie będźie* (*will co-be*, Google with context) and *dla potrzeb Stanu* (*for the State*, GPT-3 with context).

In some cases, both adequacy and fluency were rated highest in all four categories of translation (e.g. for isolated words *confirmatio* or *notandum*). However, there were also instances of context-less translations which (according to the principle: how well the target text represents the informational content of the source text [8]) were rated highest on the scale and, although correctly translated, they were semantically distant from respective entries in the manually created dictionary. This fact should be noted when analysing the reported percentages. In reality, their actual usefulness will therefore be lower than the statistics show.

[21] See [8], p. 8.

5.2 The Role of Context

In some cases, a correct automatic translation is not possible in the given limited context, while the glossary takes into account a man-made translation adapted to the wider context. E.g. the examined phrase *ultima praeterlapsi* (literally: *last of the past*) requires an addendum specifying the time when the action described in the sentence took place. The time is indicated in a broader, omitted context, and thus in the given limited context, the translator failed to cope with a translation whose meaning (the last day of the past month) should have been guessed from the broader context.

Sometimes the Polish translation present in the dictionary contained excessive explanations, not covered by the original, e.g. *cum omnium applausu & satisfactione* was translated as *zyskując uznanie i zadowolenie wszystkich stanów Królestwa* (En: *gaining recognition and satisfaction from all states of the Kingdom*) which obviously contains more information than the Latin text.

5.3 Transcription Errors

Sometimes the process has been influenced by transcription errors, as in *ludicre* (clearly visible in the original) wrongly transliterated as *ludiere*. At the same time, the neural models are known to deal with such issues successfully.

5.4 Language Model-Induced Consequences

The multilinguality of the GPT model frequently showed when English output was partially produced for translated Latin phrases (e.g. *Salve* rendered as *hello* or *Nobilitatis* as *Nobility*. At the same time, multilinguality is believed to help translate content without the additional step of language identification for individual fragments, previously a necessary step for every mixed-content solution.

The expectation that translation models could make use of syntactic properties of the context to produce properly declined forms was confirmed, cf. *powstáły rożne sensus* (*various meanings have emerged*) was translated as *powstáły*$_{PL}$ *rożne*$_{PL}$ *znaczenie*$_{SING}$ by Google and *powstały*$_{PL}$ *różne*$_{PL}$ *znaczenia*$_{PL}$ by GPT-3 (also note the correction of diacritics, also interpreted properly for misspelt words).

In general, Polish translations retain the proper grammatical form of the tokens, valid in the given context (i.e. making a valid larger whole, e.g. a sentence, when inserted instead of its corresponding Latin interjection). Still, in some cases the translation lacks some essential part, e.g. a preposition, as in *prośił znowu assensum* (*he asked for reconciliation*), when *assensum* was translated as *zgodę* (*reconciliation*$_{ACC}$) while in this context a preposition should be added *o zgodę*. The generative model added this preposition so we corrected such cases in the original translation before running the evaluation.

6 Conclusions

The presented experiment showed that it is still too early to use Latin-to-Polish machine translation in a completely automatic process. None of the four given categories yielded satisfactory results concerning the adequacy, and the chosen MT solutions should not constitute a reliable end-to-end tool for translating Latin in Polish texts. Moreover, the high percentage of incorrect translations in the Latin survey, both in the wider context (e.g. 45% in Google translator) and stand-alone clearly shows that there is a high risk that automatic translation will generate erroneous results. Still, the evaluator's experiences show that it can prove useful as a translator's aid under human supervision.

The results can be further analysed, taking into account various types of errors produced (wrong syntax, lack of prepositions, inflexion problems, spelling errors etc.)

The fact that the context contained intertwined Latin and Polish text we could also confirm that contemporary translation models can successfully cope with such a multilingual blend. Unfortunately, even though CBDU prints contain interjections in other languages than Latin, no evaluation data is available so no truly multilingual tests could be carried out.

References

1. Agirre, E., Gonzalez-Agirre, A., Lopez-Gazpio, I., Maritxalar, M., Rigau, G., Uria, L.: SemEval-2016 Task 2: Interpretable Semantic Textual Similarity. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 512–524. Association for Computational Linguistics, San Diego, California (2016), `https://aclanthology.org/S16-1082`
2. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. pp. 138–145. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002), `https://aclanthology.org/www.mt-archive.info/HLT-2002-Doddington.pdf`
3. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W.: Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Transactions of the Association for Computational Linguistics **9**, 1460–1474 (2021), `https://doi.org/10.1162/tacl_a_00437`
4. Gruszczyński, W., Adamiec, D., Bronikowska, R., Wieczorek, A.: Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe. Poradnik Językowy (8/2020 (777)), 32–51 (2020), `https://doi.org/10.33896/porj.2020.8.3`
5. Gruszczyński, W., Ogrodniczuk, M.: Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski dotyczących z XVI, XVII i XVIII w. w nauce i dydaktyce (*Digital Library of Poland-related Old Ephemeral Prints in research and teaching*, in Polish). In: Materiały konferencji *Polskie Biblioteki Cyfrowe 2010* (Proceedings of the *Polish Digital Libraries 2010* conference). pp. 23–27. Poznań, Poland (2010)
6. Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., Menezes, A.: To Ship or Not to Ship: An Extensive Evaluation of Automatic

Metrics for Machine Translation. In: Proceedings of the Sixth Conference on Machine Translation. pp. 478–494. Association for Computational Linguistics, Online (2021), https://aclanthology.org/2021.wmt-1.57

7. Lavie, A., Agarwal, A.: METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. pp. 228–231. Association for Computational Linguistics, Prague, Czech Republic (2007), https://aclanthology.org/W07-0734
8. Maučec, M.S., Donaj, G.: Machine Translation and the Evaluation of Its Quality. In: Sadollah, A., Sinha, T.S. (eds.) Recent Trends in Computational Intelligence, chap. 8. IntechOpen, Rijeka (2019), https://doi.org/10.5772/intechopen.89063
9. Ogrodniczuk, M., Gruszczyński, W.: Digital Library of Poland-related Old Ephemeral Prints: Preserving Multilingual Cultural Heritage. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage. pp. 27–33. Hissar, Bulgaria (2011), http://www.aclweb.org/anthology/W11-4105
10. Ogrodniczuk, M., Gruszczyński, W.: Digital Library 2.0 — Source of Knowledge and Research Collaboration Platform. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). pp. 1649–1653. European Language Resources Association, Reykjavík, Iceland (2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/14_Paper.pdf
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002), https://aclanthology.org/P02-1040
12. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (2015), https://aclanthology.org/W15-3049
13. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.): Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw (2012)
14. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: 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. pp. 223–231. Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA (2006), https://aclanthology.org/2006.amta-papers.25
15. Vilar, D., Xu, J., D'Haro, L.F., Ney, H.: Error Analysis of Statistical Machine Translation Output. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). pp. 697–702. European Language Resources Association (ELRA), Genoa, Italy (2006), http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf
16. White, J.S., O'Connell, T.A., O'Mara, F.E.: The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In: Proceedings of the First Conference of the Association for Machine Translation in the Americas. pp. 193–205. Columbia, Maryland, USA (1994), https://aclanthology.org/1994.amta-1.25