Towards Semi-automatic Construction of Multilingual LGBTQ+ Conceptual Models

Maria Adamidou[1], Shuai Wang[1]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands

Abstract

Recent studies and applications have highlighted the necessity for organized multilingual LGBTQ+ vocabularies. Manual translation presents multiple difficulties, and the accuracy of translated terms heavily relies on the expertise of specialists without standardized evaluation criteria. Some recent research showed the possibility of using machine translation tools and reusing multilingual information from other resources to speed up the process and ensure consistency in translation. This paper evaluates the accuracy of a machine translation tool specifically for LGBTQ+-related terminology. We propose a semi-automated approach with supplementary resources to expedite the translation process accompanied by evaluation criteria.

Keywords

Multilinguality, Homosaurus, machine translation, LGBTQ+, Queer

1. Introduction

In recent years, a significant amount of projects focusing on LGBTQ+ themes have been implemented in various domains, including developing structured vocabularies for the metadata in libraries and archives [1], online LGBTQ+ literature databases [2], hateful speech detection in social media [3] and linguistic analysis in health systems [4]. These projects emphasize the growing focus on LGBTQ+ language and concepts and their applications. Resources used include domain-specific thesauri/structured vocabularies (e.g. Homosaurus, QLIT) that are published as linked data [1, 5, 2], ontologies (e.g. GSSO) [6], and general-purpose knowledge bases (e.g. Wikidata) [7]. For convenience in describing these diverse resources, we use the umbrella term "conceptual models". Although they have been used in some multilingual applications, the need for resources about LGBTQ+ topics across various languages has been increasing. A convenient way to obtain such conceptual models is to translate existing ones.

Multilingual LGBTQ+ labels in conceptual models are crucial in today's globalized and diverse information landscape. These labels can improve searchability and interoperability in resource sharing across different language environments to better serve users from diverse linguistic backgrounds, supporting the equity of access, and thus help ensure inclusivity, respect, and sensitivity to the cultural contexts of their users. Manual editing is indispensable, for example, in cases where the meaning of some translated terms can vary in contexts (e.g. 'queen') and other cases where there is no corresponding term in the target language (e.g. "straight" does not have a Czech counterpart, thus translated to "heterosexuální" [8]). However, this manual approach can be challenging due to the numerous factors that must be considered and exhibit several drawbacks. Some terms have many synonyms, leading to overlooked alternative labels. Furthermore, concepts and their multilingual labels can have cultural barriers and can change (see examples of concept drift and convergence in [9]), resulting in frequent maintenance of their corresponding multilingual labels (e.g. Homosaurus is released twice per year with updates and new terms).

Many terms exhibit a comparable syntactic structure in Homosaurus, such as "African American asexual people", "African American bisexual people", "African American gay men", etc. By standardizing the translation of certain tokens, they can be accurately translated automatically. In addition, manually

4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.

adamimaria96@gmail.com (M. Adamidou); shuai.wang@vu.nl (S. Wang)

0009-0005-9536-5452 (M. Adamidou); 0000-0002-1261-9930 (S. Wang)

translating additional information (e.g. the scope notes and comments) can slow down the development. Considering the scale and intricacy, this manual approach can be both time-intensive and potentially exhausting for experts, leading to uncertainty and inconsistency. As the authors are aware, there is no standard (semi-/automatic) workflow, nor an established evaluation metric.

This study investigates the potential for creating a semi-automated workflow to reduce the burden on experts and accelerate the procedure. We use Homosaurus and QLIT for the evaluation of our approach. The Homosaurus[1] is a linked data vocabulary of LGBTQ+ terms for catalog representation of MORGAI (Marginalized Orientations, Relationships, Gender Identities, and Intersex) [1]. It has its roots in the internal thesaurus of the IHLIA LGBTI Heritage in the Netherlands as the first bilingual LGBTQ+ thesaurus/vocabulary (English and Dutch) and has become well-used in the metadata of libraries and archives. Most recently, Homosaurus is to be enriched with Spanish labels [10]. QLIT (Queer Literature Indexing Thesaurus) is a Swedish thesaurus published in 2023 consisting of 848 entities, including 757 reused from Homosaurus' 2021 release, primarily used for the online Swedish LGBTI literature database, QueerLit [2]. Other related work includes a linked open vocabulary system with LGBTQ+ terms enriched with multilingual labels in Hindi and Bangla by Mukhopadhyay et al. [11]. The demand for a Chinese translation remains [12]. Initial evaluations of MT tools' precision were performed utilizing a subset of Homosaurus terms alongside established benchmarks [13]. Recently, Wang et al. [9] studied the reuse of multilingual LGBTQ+ resources for enrichment.

Despite Machine Translation (MT) not being expected to reach human-level accuracy, this paper explores the usability of MT for LGBTQ+-related terms by assessing its accuracy, providing a benchmark of the selected MT tool, and proposing a semi-automatic workflow. More specifically, we study the following three research questions. **RQ1:** How accurate are the terms translated by state-of-the-art MT tools by using customized translated glossaries? **RQ2:** How to construct a semi-automatic workflow for the translation of LGBTQ+ terms and take advantage of multilingual labels from other resources? **RQ3:** What evaluation criteria can we define to evaluate the resulting multilingual conceptual model? The rest of the paper is organized as follows. Section 2 presents the methodology for the benchmarking of translated terms with the evaluation results in Section 3. We propose our workflow in Section 4 with the evaluation criteria introduced in Section 5, followed by the conclusion in Section 6. The supplementary materials can be found on GitHub.[2]

2. Methodology

Incorporating MT into the workflow demands reasonable translation accuracy. We use the DeepL API[3] as a proof-of-concept of our approach for the translation of Homosaurus terms by taking the suggestion by Kazarian et al. [13], whose primitive examination showed that DeepL emerged to be one of the most accurate translation tools for LGBTQ+ terms for translating Homosaurus to Dutch. Additionally, DeepL is free and allows for customized glossaries while translating. In our work, expert insights are requested to further improve the accuracy by providing a translated glossary. This enhancement is crucial, as Kazarian et al. [13] found that some manually constructed rules for refinement can significantly increase the number of terms with a perfect match (from 38.79% to 56.76% with just six simple refinement rules). As the first step, we provide experts with frequent tokens to be manually translated. Our study builds upon earlier research by providing DeepL with extra translated tokens, as we discovered that the translation quality could be enhanced by predetermining the translation of certain tokens. For example, 'LGBTQ+' is often mistakenly translated as 'LGBTQ+' in Dutch instead of 'LHBTQ+'. In cases where a token in the naive DeepL translation did not match the corresponding expert-provided token, it was replaced with the expert version in all cases applicable. For example, 'agender people' was translated by DeepL as 'agender mensen' in Dutch, but since the token 'people' was included in the glossary with

[1] https://homosaurus.org/. We used v.3.4 released on June 2023 with 2,885 entities but only 2,835 are with Dutch labels.

[2] The code, data, and other supplementary materials are available on GitHub: https://github.com/Multilingual-LGBTQIA-Vocabularies/MDTT. The best practices and the evaluation criteria can also be found on Zenodo with the DOI: 10.5281/zenodo.15082538.

[3] https://developers.deepl.com/docs

translation 'personen', the result was adjusted accordingly to 'agender personen'. Given the importance of these tokens, we conduct a comparative analysis with two sets acquired by evaluating the trade-off between translation accuracy and their occurrence frequency in practical scenarios.

Four datasets were taken into consideration when constructing the two sets of tokens. We introduce D_1 and D_2 consisting of tokens extracted from the `skos:prefLabel` of the English terms with their respective frequencies in Homosaurus and QLIT, respectively. D_3 was provided by IHLIA experts[4] with each token associated with frequencies on the occurrence of each Homosaurus token within IHLIA's biggest database of non-fiction books, grey literature, and articles, totaling 116,738 records. Lastly, D_4 was provided by QLIT experts[5] with frequency for each token in Queerlit. The first two concern tokens, while the other two consider the frequency of use. The reason behind incorporating both the frequency of Homosaurus and QLIT tokens and their usage in IHLIA and Queerlit lies in the tradeoff between accuracy in translation of conceptual models as well as the accuracy in most used terms due to the "long-tail distribution" where many tokens are rarely used judged by the frequency obtained. To aggregate D_1 and D_2, we construct D_5 by assigning each token a numerical value that is the sum of its rank of frequencies in D_1 and D_2. For example, the token 'people' has the numbers 1 and 2 in D_1 and D_2 for ranking highest in Homosaurus and second highest in QLIT, resulting in a sum of 3 in D_5.

We aim to provide experts with a small and concise set of around 90-100 tokens for manual translation. For a comparative study, we obtain two sets of tokens from different parametric settings considering the trade-offs of frequency in the conceptual models and real-life application scenarios. Following that, we compare their impact on MT accuracy. In the first set, S_1, we include the 60 most frequent tokens from D_5, along with the 30 most frequent tokens from D_3 and D_4 respectively. S_2 has a different parametric setting with more weights on application scenarios featuring 40 from D_1 and 50 from D_3, along with 40 from D_2 and 50 from D_4. Due to some overlap of selected tokens across datasets, there are 83 tokens for S_1 and 101 tokens for S_2. Finally, both were translated by experts for translation in the next step.[6]

Using the glossaries, when a token in the naive DeepL translation did not match the corresponding token provided by the experts, it was replaced with the expert-provided translation. This approach resulted in the creation of two improved versions of the initial naive DeepL translations, one incorporating the modified tokens from S_1 and S_2, respectively. Finally, inspired by [13], we perform some additional semi-automatic rule-based refinement[7] of the results to ensure the syntactic consistency of the terms. For example, replace 'biseksuele mensen' by 'biseksuelen' and replace 'lesbiennes' by 'lesbische vrouwen' in the translated terms in Dutch. In addition, some spaces were removed to concatenate two words or added for splitting into two words. Next, we evaluate their corresponding translation results. We employ two well-known metrics: the Jaccard similarity computes the number of identical matches of translated words disregarding the order of words [14]; the Levenshtein distance for the minimum number of edits required to transform an attempted translation into its accurate translation [15].

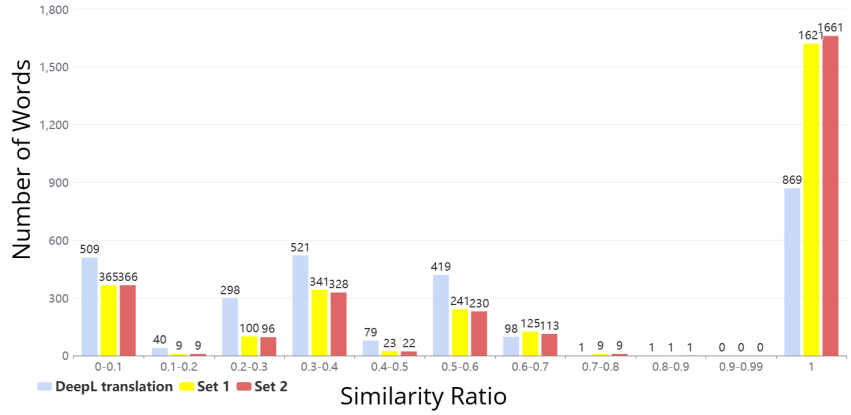
3. Evaluation

Homosaurus provides labels using both `skos:prefLabel` and `skos:altLabel` for Dutch terms, allowing the naive DeepL translations to be compared against the results using S_1 and S_2. Note that not all entities have alternative labels specified by `skos:altLabel`. The improvement in results presented in Table 1 confirms the effectiveness of the use of the customized glossary and the additional refinement. Moreover, using S_2 shows a slight advantage over S_1. However, this marginal difference may be attributed to the greater token count in S_2, preventing a definitive conclusion about the superiority of S_2. Moreover, recognizing the limitations of the translation result, particularly in identifying translated terms with a significant similarity but minor difference in spelling, the Jaccard similarity and the Levenshtein distance are employed to quantify editing efforts by experts. For example, the Homosaurus

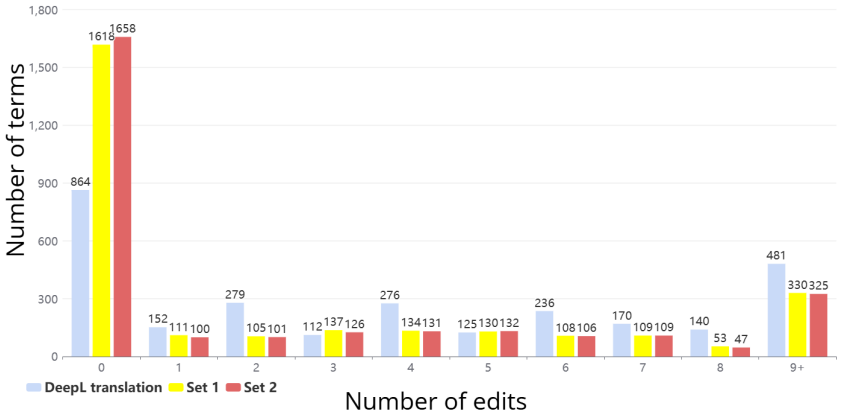
[4]Received with authorization for use on June 28, 2023.
[5]Received with authorization for use on July 3, 2023.
[6]We combined them into a unified set ($S_1 \cup S_2$), totaling 115 tokens for manual translation by a Dutch-speaking expert from IHLIA and Swedish-speaking experts from QLIT.
[7]More details are in the supplementary material.

		Without Refinement		With Refinement		Score
		prefLabel	altLabel	prefLabel	altLabel	
Homosaurus	Baseline (naive DeepL translations)	864 (30.5%)	48 (1.7%)	864 (30.5%)	48 (1.7%)	6.7M
	Translation using S_1	1064 (37.5%)	50 (1.8%)	1618 (57.1%)	49 (1.7%)	23.1M
	Translation using S_2	1076 (38%)	55 (1.9%)	1658 (58.5%)	48 (1.7%)	27.1M
QLIT	Baseline (naive DeepL translations)	268 (30.1%)	93 (10.5%)	268 (30.1%)	93 (10.5%)	166.9K
	Translation using S_1	238 (26.8%)	80 (9%)	511 (57.5%)	74 (8.3%)	180.3K
	Translation using S_2	243 (27.4%)	71 (8%)	518 (58.3%)	72 (8.1%)	229.2K

Table 1
The accuracy of the identical match for Homosaurus and QLIT and their scores of correct translation in use.



(a) Jaccard Similarity



(b) Levenshtein Distance

Figure 1: A comparison of Jaccard Similarity and Levenshtein distance for translated terms using the naive DeepL translations, and the results using Sets 1 and 2 for Homosaurus.

term "5-alpha reductase deficiency" has a Dutch translation as "5-alpha-reductasedeficiëntie". However, its Dutch label is "5-alpha-reductase deficiëntie". While an exact match algorithm does not recognize this as a match, an expert would see the need for a minor change.

Figure 1a shows that, using Jaccard similarity, in both cases, the results show that over 70% of the translation with at least 50% similarity to the Dutch labels by Homosaurus. Furthermore, as in Figure 1b, with S_2, the Levenshtein distance indicates the superior performance compared to the naive DeepL translations, accounting for just over 70% (i.e. 1,985 terms) of the Dutch translations requiring at most 3 edits to match exactly with Homosaurus' translation. Due to page limit, similar evaluation results for QLIT were included in the supplementary material.

Further evaluation of these translations considers the frequency of tokens in the database of IHLIA and

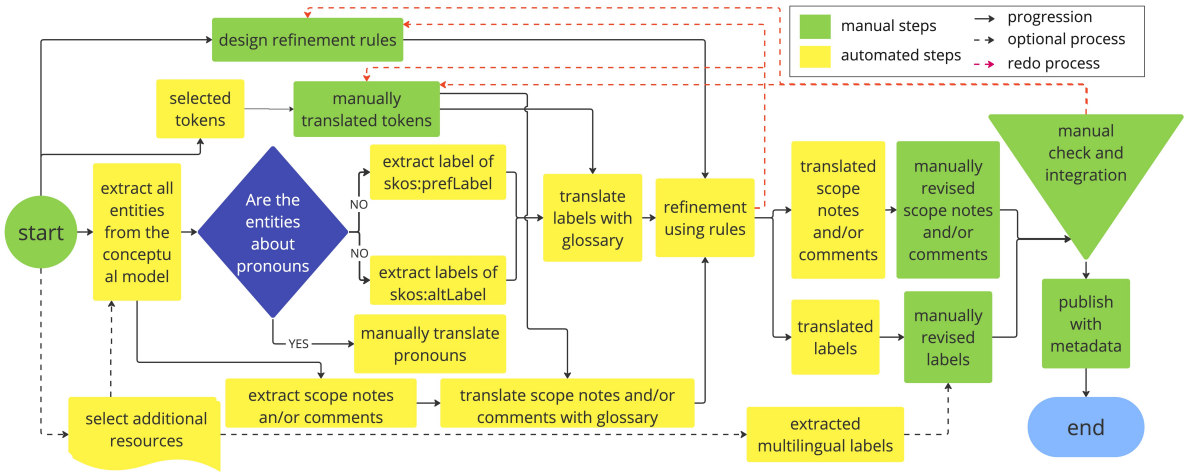


Figure 2: A semi-automatic translation workflow for multilingual LGBTQ+ terms.

Queerlit. This could be achieved by computing the sum of the multiplication of the correct translation by frequency for each of the most frequently used 120 tokens in each database regarding the naive DeepL translation and that using S_1 and S_2. For example, in the IHLIA database, the most frequent token is 'Lesbische' with a frequency of 30,208. If this token is translated correctly 146 out of 150 times in the naive DeepL translation, we accumulate the product of its frequency and the correct occurrence count, 30,208*(146/150). The results in Table 1 further demonstrate how the translations with S_1 and S_2 improve the accuracy in use. Although the accuracy and scores are higher when using S_2, it can be attributed to the fact that S_2 contains more tokens than S_1.

4. A Semi-automatic Translation Workflow

Building upon the findings presented earlier, we design a generic workflow for the fast development of multilingual LGBTQ+ conceptual models. Figure 2 shows our semi-automatic workflow with yellow blocks indicating the use of data processing scripts or MT, and green blocks indicating tasks requiring experts' intervention. Next, we explain this workflow in detail using Homosaurus, S_2, and DeepL. First, we retrieve the latest version of Homosaurus (and other selected multilingual resources). We can then extract labels of the entities to be translated. We exclude the pronouns as they are often not required to translate, but a manual review may be needed to ensure proper spelling. Thus, we focus on the other entities' labels. Meanwhile, we could ask experts to manually translate S_2. The design of refinement rules could start from the beginning (after testing on some examples) and be updated in later iterations. The labels will then go through three steps. First, we translate the labels using DeepL using S_2 and its translation. Then we refine the results with rules to achieve better accuracy. Before integration, the labels should be revised manually. An optional step is to consider the labels extracted from other resources, [9] which could be used as alternative labels. Similarly, the scope notes and comments of a concept or a term, serving to clarify the meaning and use, can be translated in a similar way. After integration, the data could be published together with its (updated) metadata, possibly with multilingual information in the metadata. Translating LGBTQ+ terms often raises numerous concerns, leading to multiple rounds of discussions. Issues such as consistency might emerge later, requiring a review of translated tokens or refinement rule adjustments (the red arrows). Experts may encounter other scenarios, thus making the workflow flexible for extension.

5. Best Practices and Evaluation Criteria

Despite numerous attempts, translations are often not thoroughly assessed, which could be due to the complexity of the task and the lack of comprehensive and systematic evaluation criteria. In the

supplementary material, we propose best practices for 1) clarity and accuracy, 2) consistency, 3) cultural and contextual sensitivity, 4) inclusivity and ethical considerations, 5) transparency and community contribution, and 6) documenting, publishing, and maintenance. For evaluation, we provide some indicators and a checklist for self-assessment on translation, documentation, and publication.

6. Conclusion and Future Work

MT aids experts by suggesting translations and maintaining consistency in the translation workflow, which enhances efficiency. For RQ1, to justify its quality, we presented the first full-scale MT benchmark using Homosaurus and QLIT with two sets of tokens in evaluation. We showed how using MT can help with translation efficiency: when using S_2 with refinement, over 60% can be used as labels (either prefLabel or altLabel) for Homosaurus and over 66% for QLIT. For Homosaurus, about 70% of translations require only at most 3 edits to be accurate. We proposed a workflow for RQ2 for convenient semi-automatic translation. Finally, we established evaluation criteria for quality control (RQ3). Our workflow could benefit from further adjustment and additional hand-picked tokens that exhibit ambiguity. Reproducibility could be improved if the manual refinement is properly documented. While the accuracy of multilingual labels from MT and external sources can be debatable, these labels may be useful, especially for comparative purposes and discussion during manual assessments. Our workflow uses DeepL but other MT tools as well as Large Language Models (LLMs) could be adapted for comparison. In future work, the performance of MT tools can be compared with LLMs, which may outperform MT by taking into account the LGBTQ+ context during translation. Moreover, LLMs may help with the generation of scope notes and comments. It remains to be studied how some bias and ambiguity can be introduced/reduced with MT. Finally, the enhancement of interoperability with multilingual labels and links between conceptual models can be studied.

Acknowledgment

The authors received help from Andrei Nesterov (CWI), Jacco van Ossenbruggen (VU Amsterdam), and experts from the Homosaurus and QLIT/QueerLit project.

Declaration on Generative AI

The authors used TeXGPT (via Writefull) on Overleaf and ChatGPT for paraphrasing.

References

[1] The Homosaurus editorial Board, Homosaurus vocabulary terms, 2021. URL: https://homosaurus.org/.

[2] J. Bergenmar, K. Golub, S. Humelsjö, Queerlit Database: making swedish lgbtqi literature easily accessible, in: DHNB 2022: The 6th Digital Humanities in the Nordic and Baltic Countries Conference 2022, CEUR-WS. org, 2022, pp. 433–437.

[3] E. d. S. Florentino, R. R. Goldschmidt, M. C. R. Cavalcanti, Exploring interactions in youtube to support the identification of crime suspects, in: Proceedings of the XVII Brazilian Symposium on Information Systems, SBSI '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3466933.3466967. doi:10.1145/3466933.3466967.

[4] C. A. Kronk, Gender, Sex, and Sexual Orientation in Medicine: A Linguistic Analysis, Doctoral dissertation, University of Cincinnati, OhioLINK Electronic Theses and Dissertations Center (2021). URL: http://rave.ohiolink.edu/etdc/view?acc_num=ucin1617107411106107.

[5] A. Matsson, O. Kriström, Building and serving the Queerlit thesaurus as linked open data, Digital Humanities in the Nordic and Baltic Countries Publications 5 (2023) 29–39.

[6] C. A. Kronk, J. W. Dexheimer, Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow, Journal of the American Medical Informatics Association: JAMIA 27 (2020) 1110–1115. URL: https://doi.org/10.1093/jamia/ocaa061. doi:10.1093/jamia/ocaa061.
[7] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.
[8] M. ČUDOVÁ, Translating Queer Identities: A Glossary of Terms, Ph.D. thesis, Masarykova Univerzita, 2021.
[9] S. Wang, M. Adamidou, Examining LGBTQ+-related concepts in the semantic web: Link discovery, concept drift, ambiguity, and multilingual information reuse, in: M. Alam, M. Rospocher, M. van Erp, L. Hollink, G. A. Gesese (Eds.), Knowledge Engineering and Knowledge Management, Springer Nature Switzerland, Cham, 2025, pp. 1–17.
[10] Office of Communications and Marketing, An LGTBQ language thesaurus is translated to spanish, 2024. URL: https://www.gc.cuny.edu/news/lgtbq-language-thesaurus-translated-spanish, accessed on May 19, 2024.
[11] P. Mukhopadhyay, R. Mitra, Digital Humanities and Inclusive Librarianship: Designing a Collaborative, Multi-lingual, Skos- compliant Linked Open Vocabulary for LGBTQIA+, Indian Journal of Information, Library & Society 35 (2022) 16–33. URL: https://doi.org/10.5281/zenodo.6814869. doi:10.5281/zenodo.6814869.
[12] D. O. Ihrmark, K. Golub, X. Tan, Subject indexing of lgbtq+ fiction in sweden and china, in: Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities, Ergon-Verlag, 2024, pp. 379–384.
[13] A.-M. Kazarian, S. Wang, Evaluating Automated Machine Translation of LGBTQ+ Terms: Towards Multilingual Homosaurus, 2024. URL: https://doi.org/10.5281/zenodo.10523283. doi:10.5281/zenodo.10523283.
[14] G. Ivchenko, S. Honov, On the jaccard similarity test, Journal of Mathematical Sciences 88 (1998) 789–794.
[15] V. I. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, Soviet Physics Doklady 10 (1966) 707.