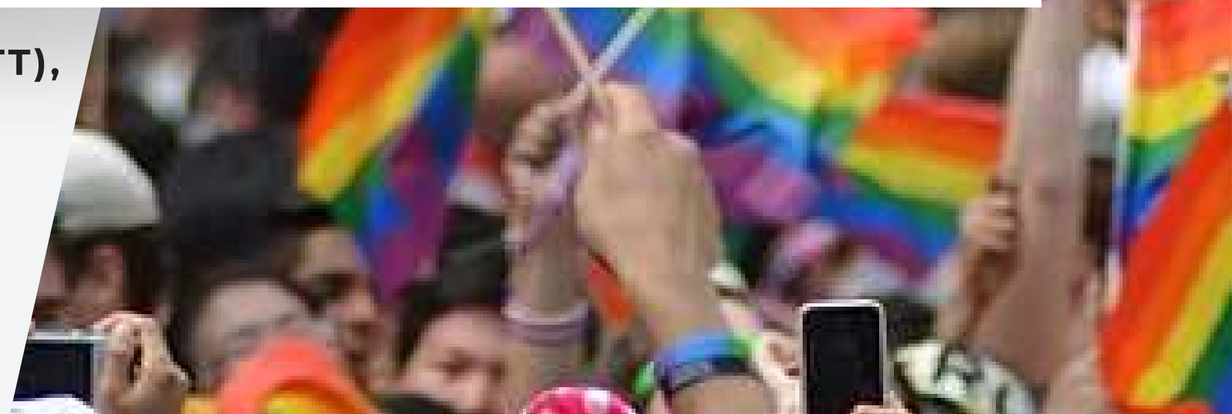
VU VRIJE UNIVERSITEIT AMSTERDAM

TOWARDS SEMI-AUTOMATIC CONSTRUCTION OF MULTILINGUAL LGBTQ+ CONCEPTUAL MODELS

Maria Adamidou, Shuai Wang

Int. Conf. Multilingual Digital Terminology Today (MDTT), Thessaloniki, Greece, June 2025



OVERVIEW

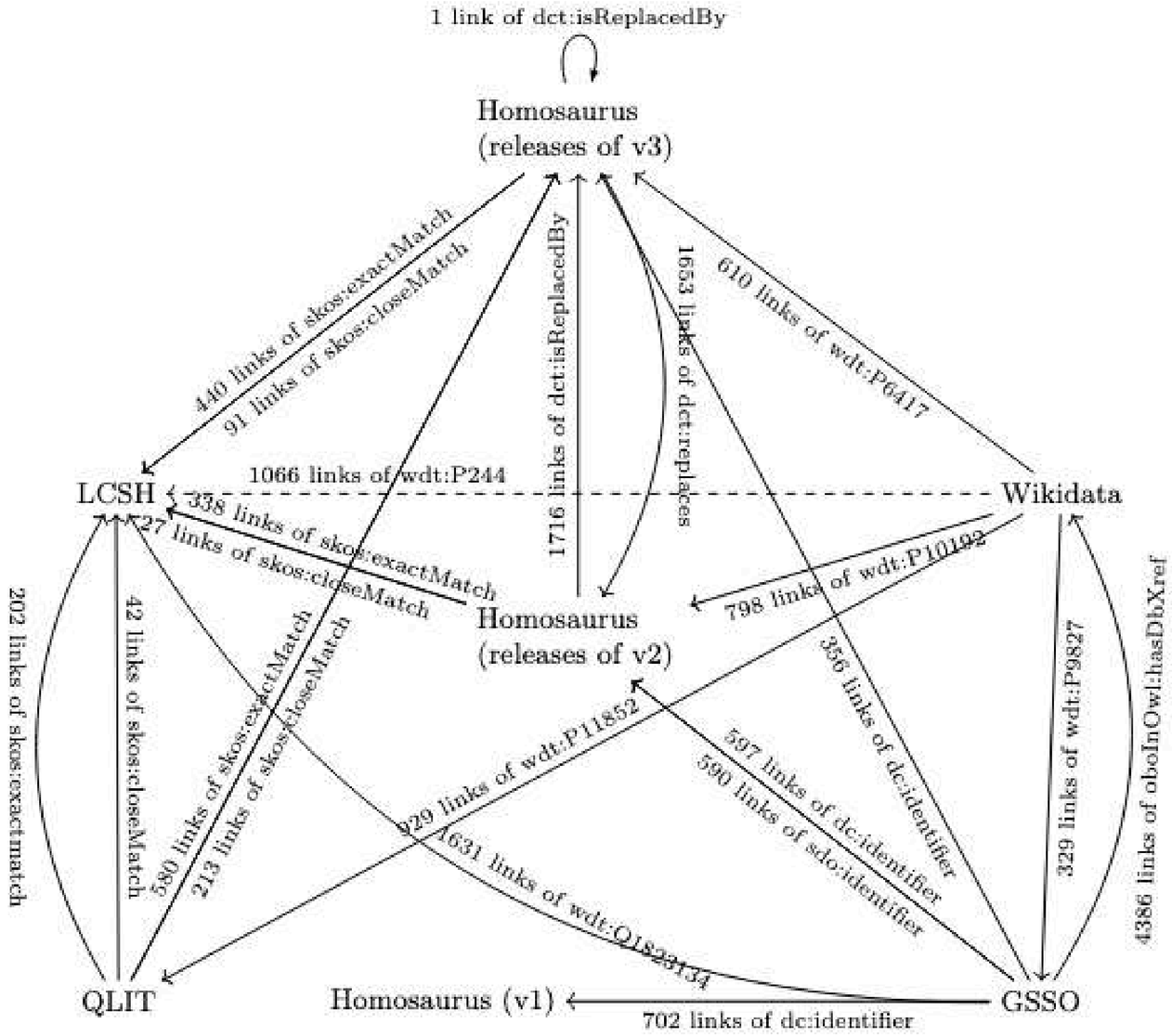


Fig. 1. Conceptual models and their extracted links. The dashed edge indicates that only edges about LCSH entities that appear in the rest of the selected concept models were chosen in this study for further integration and analysis.

Examining LGBTQ+-related Concepts in the Semantic Web: Link Discovery, Concept Drift, Ambiguity, and Multilingual Information Reuse

International Conference on Knowledge Engineering and Knowledge Management (EKAW), Amsterdam, the Netherlands, Nov 2024

RESEACH QUESTIONS

Community-driven, addressing urgent needs by the community, aiming for easy maintenance and easy of use

01

Accuracy of MT

How accurate are the terms translated by state-of-the-art MT tools by using customized translated glossaries?

REUSE

Other resources

How much multilingual information can be reused?

02

Workflow

How to construct a semi-automatic workflow for the translation of LGBTQ+ terms and take advantage of multilingual labels from other resources?

03

Evaluation

What evaluation criteria can we define to evaluate the resulting multilingual conceptual model?

RQ1: THE ACCURACY OF MT WITH CUSTOMIZED TRANSLATED GLOSSARY

select sets of tokens and provide them with their translations to the MT tools. Compare their results.

IHLIA LGBTI HERITAGE — Frequency of tokens in use

Queerlit — Frequency of tokens in use

HOMOSAURUS — Frequency of tokens in vocabulary

QLIT — Frequency of tokens in vocabulary

Long-tail distribution

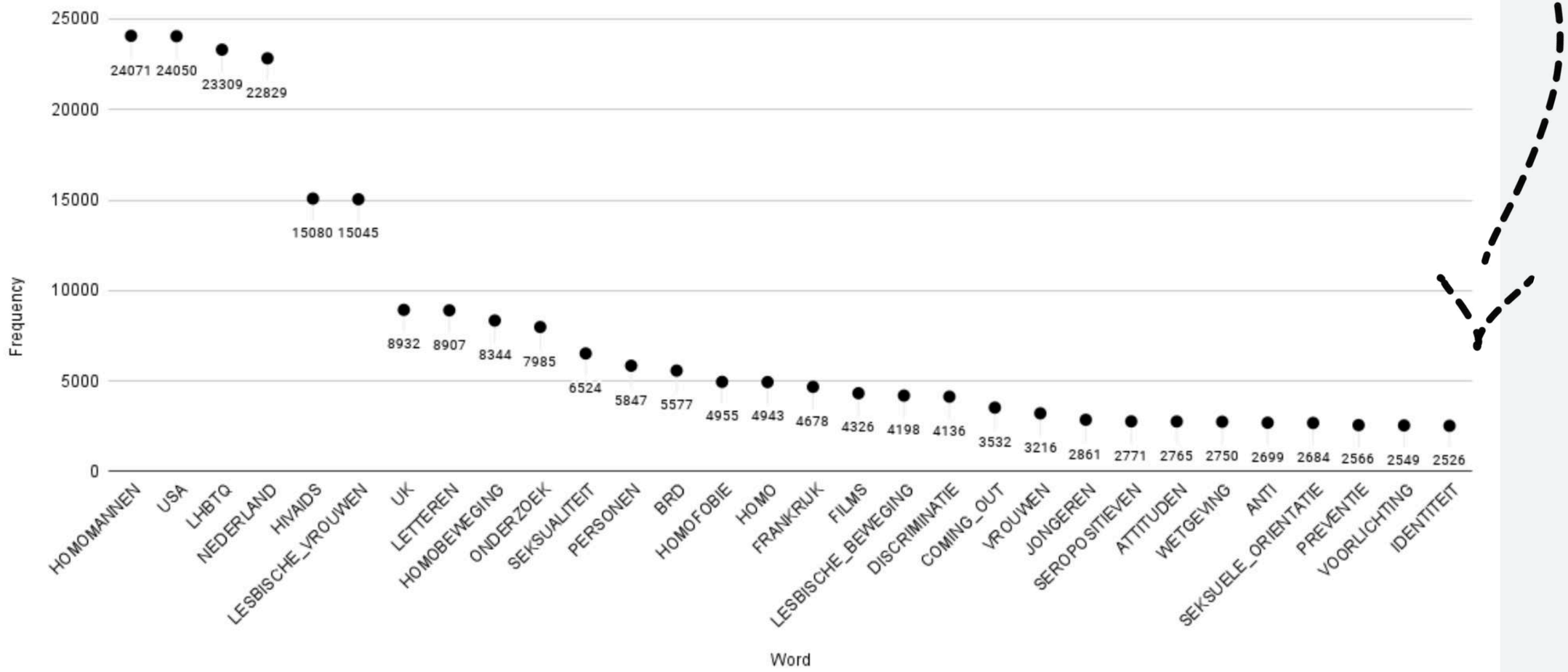


Figure 4.2: *Frequency of Dutch words in IHLIA.*

Long-tail distribution

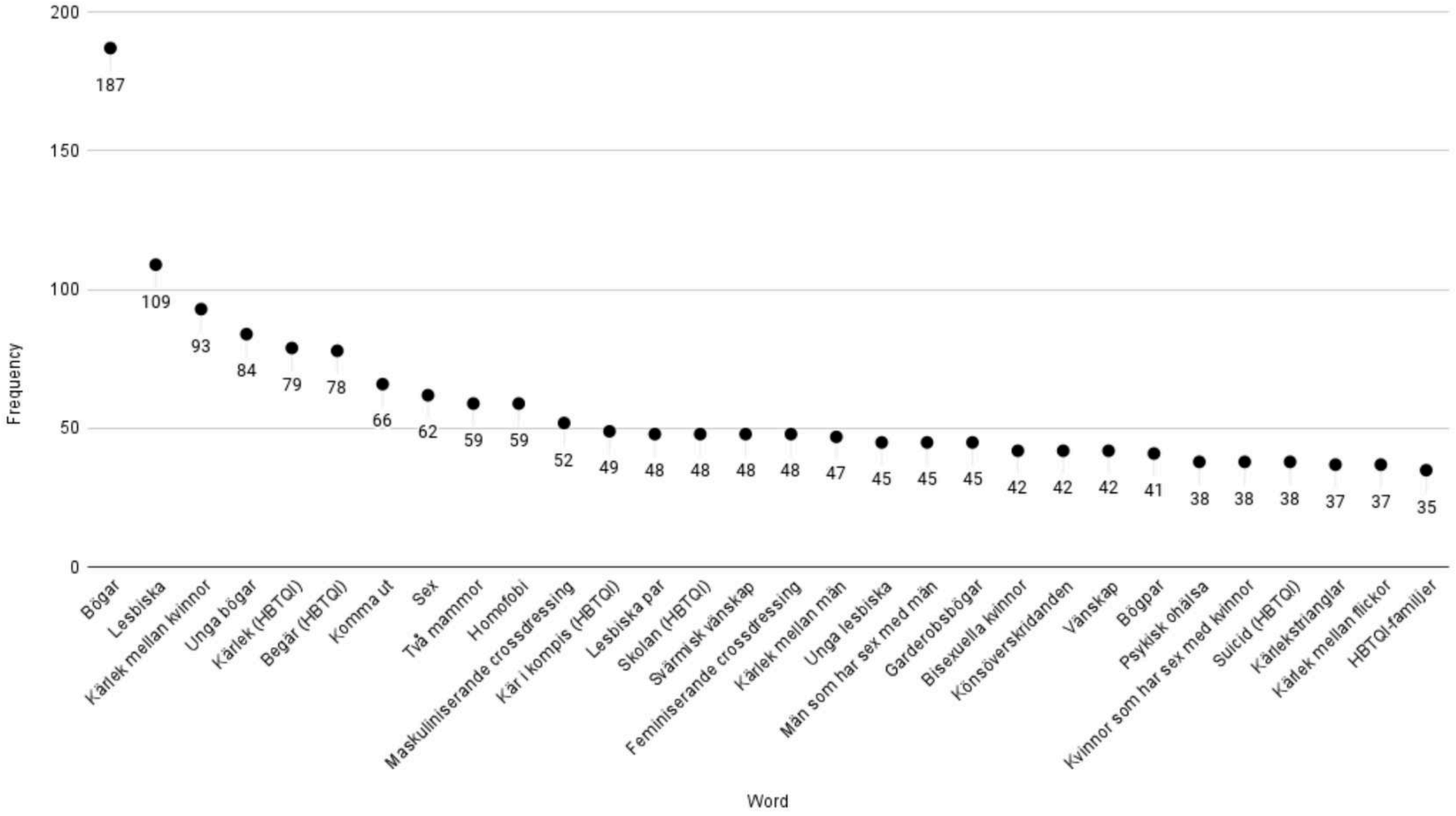


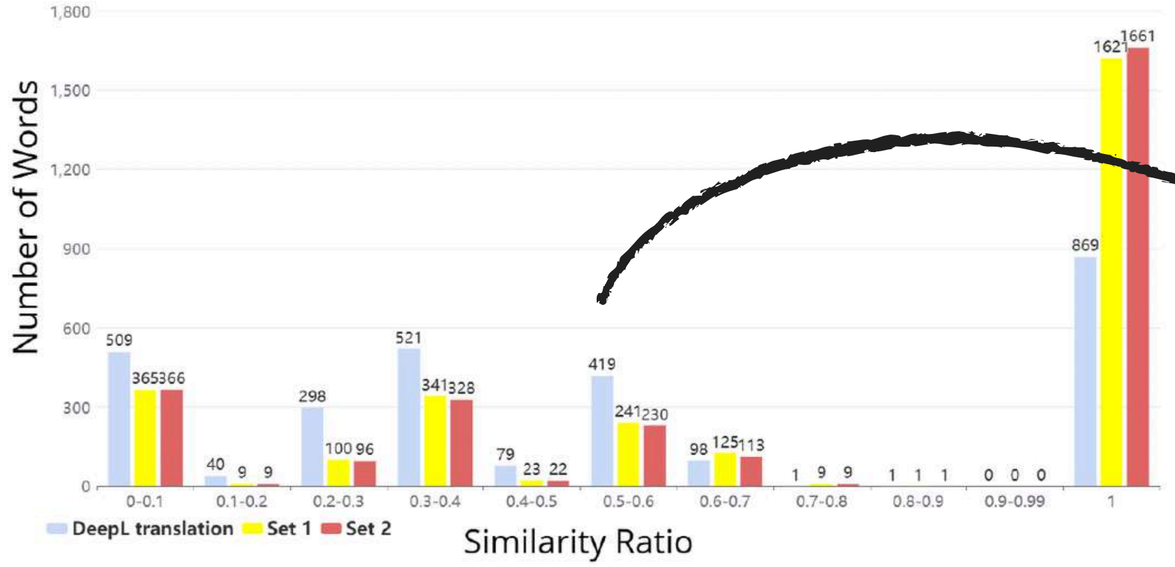
Figure 4.3: *Frequency of Swedish words in the Queerlit Database.*

DeepL

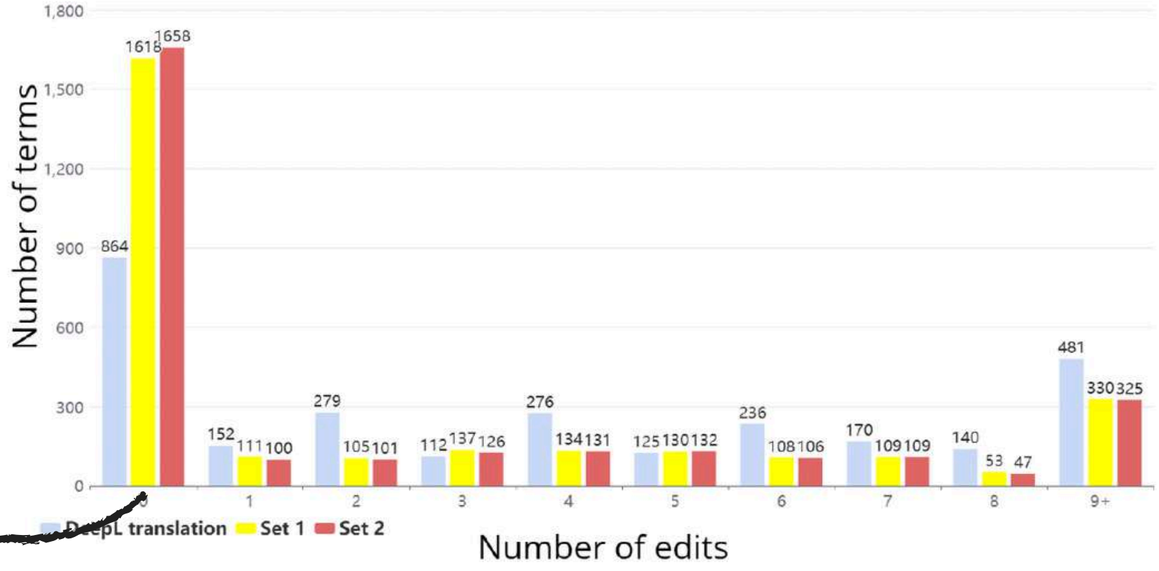
Source	Token Set A	Token Set B
#tokens from Homosaurus terms	-	40 terms
#tokens from QLIT terms	-	40 terms
#tokens from combined Homosaurus and QLIT dataset	60 terms	-
#tokens of high frequency in IHLIA	30 terms	50 terms
#tokens of high frequency in QueerLit	30 terms	50 terms
Total #tokens	83 terms	101 terms

See our previous attempt:
DOI: 10.5281/zenodo.10523283
(a subset)

		Without Refinement		With Refinement	
		prefLabel	**altLabel**	**prefLabel**	**altLabel**
Homosaurus	Baseline (naive DeepL translations)	864 (30.5%)	48 (1.7%)	864 (30.5%)	48 (1.7%)
	Translation using Token Set A	1064 (37.5%)	50 (1.8%)	1618 (57.1%)	49 (1.7%)
	Translation using Token Set B	1076 (38%)	55 (1.9%)	1658 (58.5%)	48 (1.7%)
QLIT	Baseline (naive DeepL translations)	268 (30.1%)	93 (10.5%)	268 (30.1%)	93 (10.5%)
	Translation using Token Set A	238 (26.8%)	80 (9%)	511 (57.5%)	74 (8.3%)
	Translation using Token Set B	243 (27.4%)	71 (8%)	518 (58.3%)	72 (8.1%)



(a) Jaccard Similarity



(b) Levenshtein Distance

Calling for help from the experts in this room

Lessons learned:

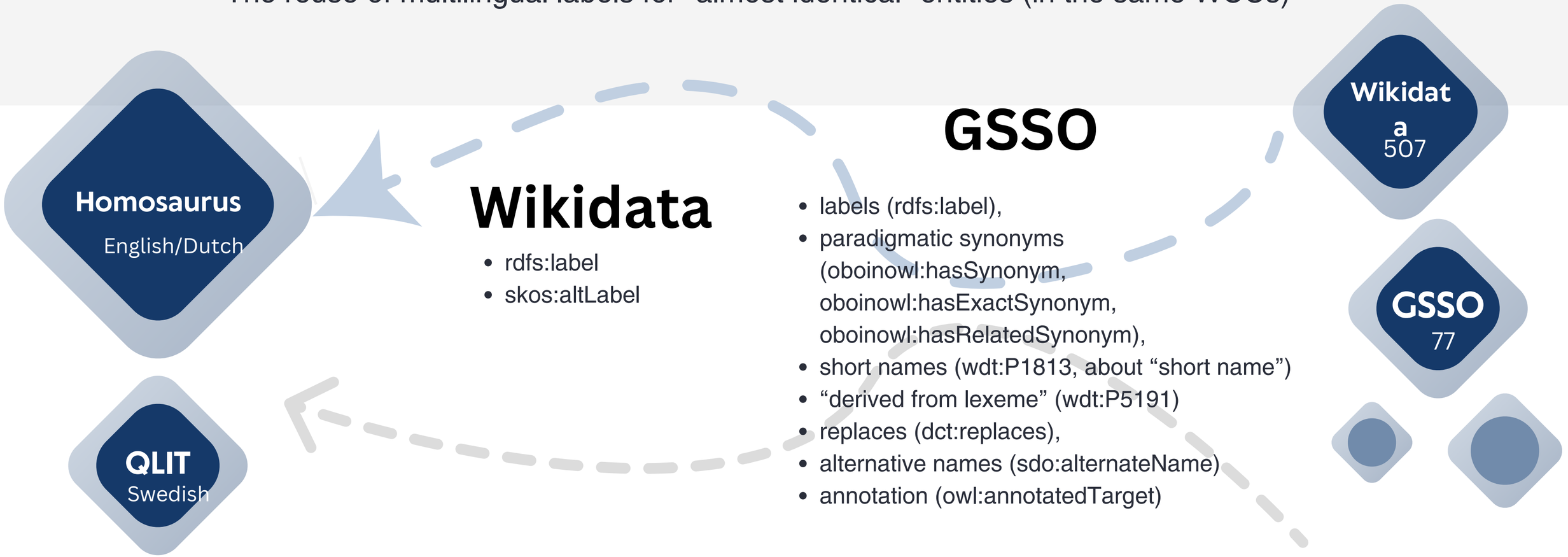
1) manual revision is unavoidable
2) limited accuracy even with customized glossary and refinement rules
3) focus on the prefLabels.
4) a benchmark that still needs to be improved.
5) mistaken translation due to lack of context.

The context was not taken into account

altLabel, scopeNotes, etc.

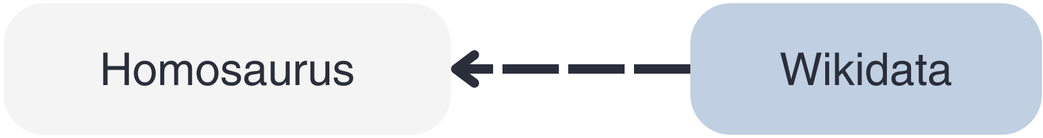
Multilingual info reuse

Community's needs for multilingual labels.
Manually searching for alternative labels.
The reuse of multilingual labels for "almost identical" entities (in the same WCCs)



Reuse!
But how much?

Only 48 entities can be enriched
top 3: English, Danish, and French



429 entities can be enriched
top 4 languages:

- English (1,692 labels for 429 entities)
- Spanish (951 labels for 333 entities)
- Chinese (893 labels for 287 entities)
- Portuguese (881 labels for 299 entities)

QLIT:

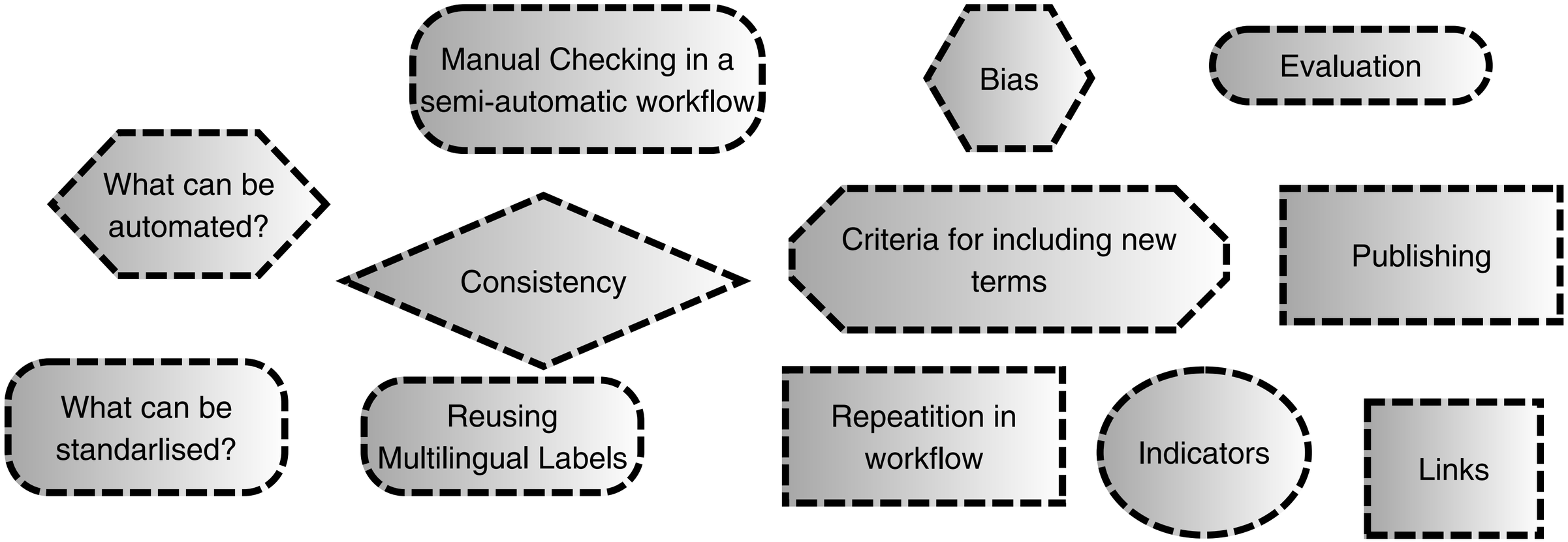
- 914 entities with one prefLabel each
- only 480 altLabels

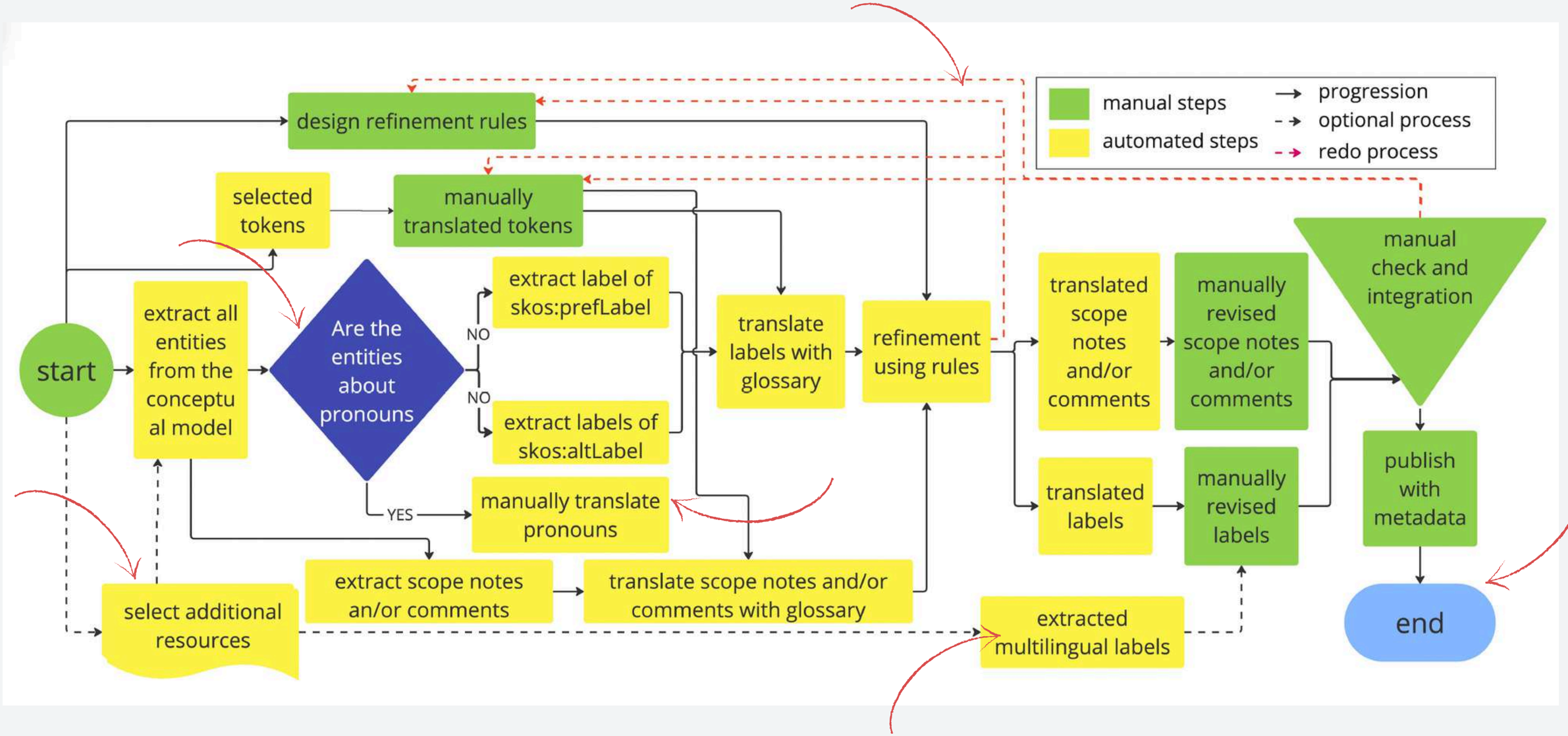
775 Swedish labels in Wikidata (524 prefLabels and 251 altLabels) for 524 entities.

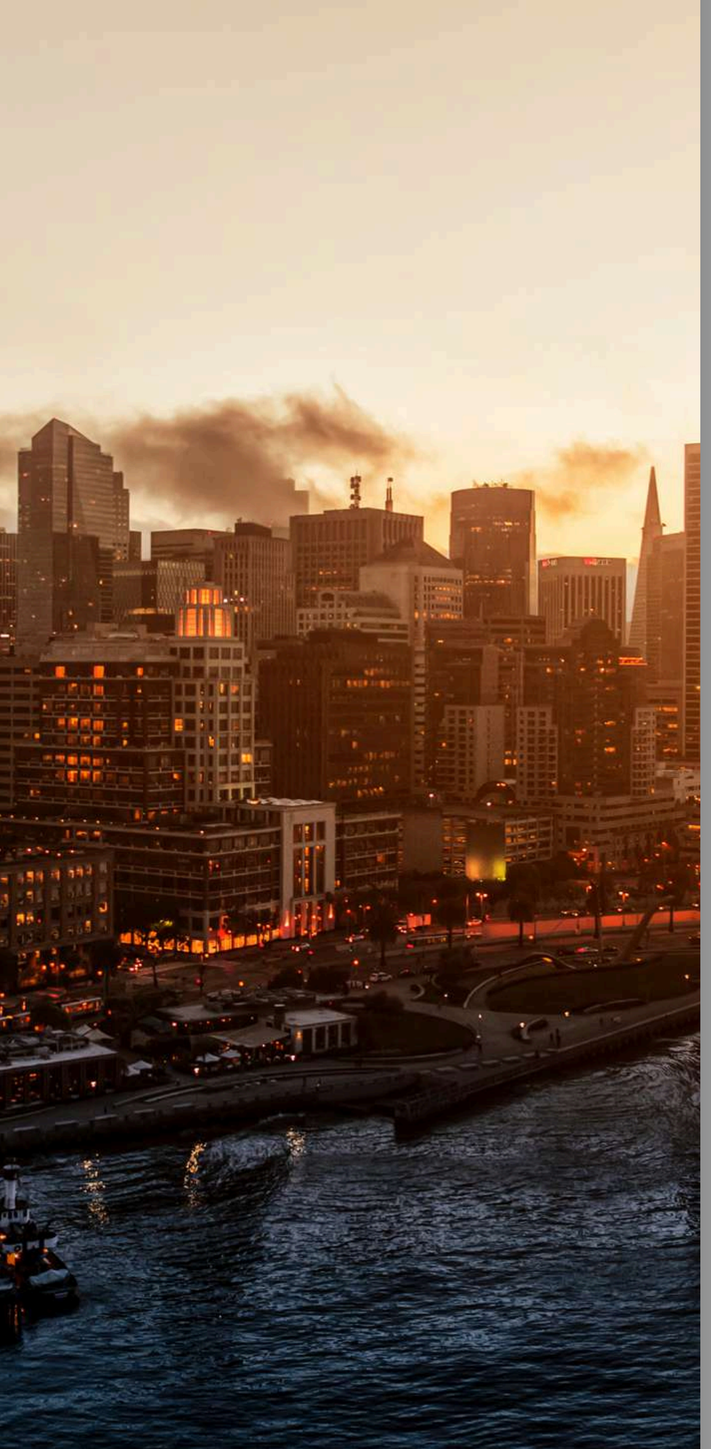
imagine if they had these labels from the beginning!

RQ2: Workflow

How to construct a semi-automatic workflow for
the translation of LGBTQ+ terms and take advantage of multilingual labels from other resources?




manual steps
automated steps
progression
optional process
redo process
design refinement rules
selected tokens
manually translated tokens
start
extract all entities from the conceptu al model
Are the entities about pronouns
NO
NO
YES
extract label of skos:prefLabel
extract labels of skos:altLabel
manually translate pronouns
translate labels with glossary
refinement using rules
translated scope notes and/or comments
manually revised scope notes and/or comments
translated labels
manually revised labels
manual check and integration
publish with metadata
end
select additional resources
extract scope notes an/or comments
translate scope notes and/or comments with glossary
extracted multilingual labels



RQ3: BEST PRACTICES AND EVALUATION CRITERIA

Part 1: Best Practices:

- Clarity and Accuracy
- Consistency
- Cultural and Contextual Sensitivity
- Inclusivity and Ethical Considerations
- Transparency and Community Contribution
- Documentation, Publishing, and Maintenance
- Bias and Transparency
- ...

Part 2: Evaluation Criteria: Indicators

The total number of concepts translated. The overlap with concepts in selected sources.

- The number of preferred labels and alternative labels (for each language/writing system).
- The number of links (e.g. provenance/identification) towards the source of translation as well as other related resources.
- The number of terms about slang and slurs.
- The number of terms about sexual orientations, gender identities, and non-binary experiences.
- The number of pronouns.

.....

Part 3: Evaluation Criteria: Self-assessment Report Template

- Checklist 1: Glossaries and Guidelines of Translation
  - The **writing system** of the target language needs to be specified.
  - A **glossary** is provided to offer standardized translations of selected tokens and terms.
  - There are criteria for the **inclusion and exclusion** of terms for the selection of terms
  - Rules for the translation of **syntactic structures** similar or analogous to maintain the syntactic structures
  - A guideline for translating **historical** terms.
  - .....

Part 3: Evaluation Criteria: Self-assessment Report Template

- Checklist 2: Documentation and Publication
  - The translation should have some documentation
  - Information about all the members and their affiliations, if possible.
  - Changes recorded between different versions of translations and how and which versions.
  - The frequency of maintenance should be specified in the documentation.

https://zenodo.org/records/15082539

Speical thanks to Wilfred van Buuren (IHLIA), Gerrit Wessel (IHLIA), Andrei Nesterov (CWI), and Martin Borissov Mashalov (UvA) for reviewing.

Conclusion and Discussion

- We provided a benchmark for machine translation.
- Due to the poor accuracy, we proposed a workflow for semi-automatic construction.
- We proposed best practices and evaluation criteria.
- Licensing of GSSO and Homosaurus → hard to combine it with A.I. & LLM.
- The scopeNotes vary in style and cannot be directly translated (STE for social science?)
- The lack of funding and maintenance remains an issue
- Publishing and data management are still not properly addressed in the community.
- Better engagement with existing communities

Acknowledgement

Andrei Nesterov
CWI

Martin Borissov Mashalov
Uni of Amsterdam (UvA)

Jacco van Ossenbruggen
VU Amsterdam

Jack van der Wel
IHLIA/Homosaurus

Clair Kronk
GSSO

Siska Humlesjö
Olov Kriström
QLIT/QueerLit

Email
shuai.wang@vu.nl

Linkedin
@shuai-ai

Zenodo
10.5281/zenodo.15082538

Github
https://github.com/Multilingual-LGBTQIA-Vocabularies/MDTT

References

1. Trans & Gender Diverse LCSH (2024), https://translcsh.com/, The list was last accessed on 25th May, 2024.
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: international semantic web conference. pp. 722–735. Springer (2007)
3. Bergenmar, J., Golub, K., Humelsj., S.: Queerlit database: Making swedish lgbtqi literature easily accessible. In: DHNB 2022: The 6th Digital Humanities in the Nordic and Baltic Countries Conference 2022. pp. 433–437. CEUR-WS. org (2022)
4. Braquet, D.: Chapter 2 LGBTQ+ Terminology, Scenarios and Strategies, and Relevant Web-based Resources in the 21st Century: A Glimpse, pp. 49–61 (05 2019). https://doi.org/10.1108/S0065-283020190000045009
5. Dobreski, B., Snow, K., Moulaison-Sandy, H.: On overlap and otherness: A comparison of three vocabularies' approaches to lgbtq+ identity. Cataloging & Classification Quarterly 60(6-7), 490–513 (2022). https://doi.org/10.1080/01639374.2022.2090040
6. Ihrmark, D.O., Golub, K., Tan, X.: Subject indexing of lgbtq+ fiction in sweden and china. In: Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities. pp. 379–384. Ergon-Verlag (2024)
7. Jagose, A.: Queer theory: An introduction. NYU Press (1996)
8. Kazarian, A.M., Wang, S.: Evaluating Automated Machine Translation of LGBTQ+ Terms: Towards Multilingual Homosaurus (Mar 2024). https://doi.org/10.5281/zenodo.10523283
9. Kronk, C.A., Dexheimer, J.W.: Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow. Journal of the American Medical Informatics Association 27(7), 1110–1115 (2020)
10. Lynch, K.E., Alba, P.R., Patterson, O.V., Viernes, B., Coronado, G., Du-Vall, S.L.: The utility of clinical notes for sexual minority health research. American Journal of Preventive Medicine 59(5), 755–763 (2020). https://doi.org/https://doi.org/10.1016/j.amepre.2020.05.026, https://www.sciencedirect.com/science/article/pii/S0749379720302774
11. Matsson, A., Kristr.m, O.: Building and serving the queerlit thesaurus
12. Nasim, I., Wang, S., Raad, J., Bloem, P., van Harmelen, F.: What does it mean when your URIs are redirected? Examining identity and redirection in the LOD cloud. In: Proceedings of the 8thWorkshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) (2022)
13. Office of Communications and Marketing: An LGTBQ language thesaurus is translated to spanish (2024), https://www.gc.cuny.edu/news/lgtbq-language-thesaurus-translated-spanish, accessed on May 19, 2024
14. Peterson, R.: Library of congress subject headings for lgbt studies (8 2023), https://guides.libraries.emory.edu/main/queerlcsh
15. Tai, J.: Cultural humility as a framework for anti-oppressive archival description. Reinventing the Museum: Relevance, Inclusion, and Global Responsibilities p. 349 (2023)
16. The Homosaurus editorial Board: Homosaurus vocabulary site (2024), https://homosaurus.org/about, Its documentation was last accessed on 24th May, 2024.
17. Vrandečić, D., Kr.tzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
18. Wang, S., Schlobach, S., Klein, M.: Concept drift and how to identify it. Journal of Web Semantics 9(3), 247–265 (2011). https://doi.org/https://doi.org/10.1016/j.websem.2011.05.003, semantic Web Dynamics Semantic Web Challenge, 2010
19. Wang, S., Maineri, A., Singh, N., Kuhn, T.: FAIR implementation profiles for social science. In: Garoufallou, E., Sartori, F. (eds.) Metadata and Semantic Research. pp. 284–290. Communications in Computer and Information Science, Springer Science and Business Media Deutschland GmbH, Germany (2024). https://doi.org/10.1007/978-3-031-65990-4_26
20. Wang, S., Raad, J., Bloem, P., van Harmelen, F.: Refining large integrated identity graphs using the unique name assumption. In: European SemanticWeb Conference. pp. 55–71. Springer (2023)
21. Watson, B.M.: "there was sex but no sexuality*" critical cataloging and the classification of asexuality in lcsh. Cataloging & Classification Quarterly 58(6), 547–565 (2020)