UNIVERSITÄT
DES
SAARLANDES

Through the looking glass of NMT: Topology of embedding spaces and language typology

Maria Kunilovskaya

1st Roundtable on NLP for Luxembourg(ish)

University of Luxemburg
23 February, 2024

Goal and motivation

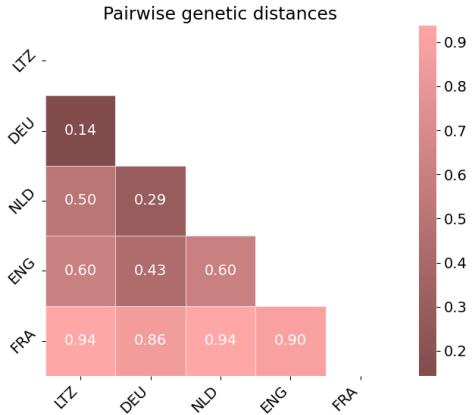
Compare translated dialects from pairs of source languages and see whether typological differences persist through translation into a third language (English)

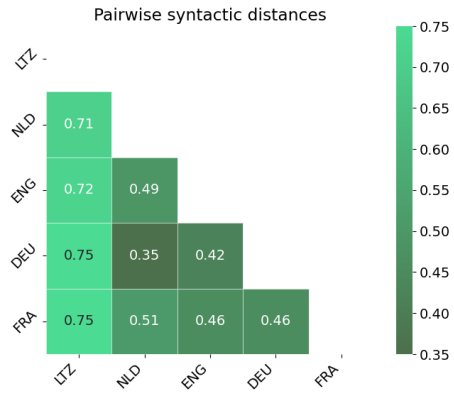
Theoretical underpinnings:

- Translated language retains the footprints of the source language (SL) (Bjerva et al., 2019; Zou et al., 2022)
- The typological similarity between languages/varieties can be approximated by quantifying isometry between their embedding spaces (Chowdhury et al., 2021).
- Embedding spaces can be compared using graph metrics (Patra et al., 2019).

Our hypotheses

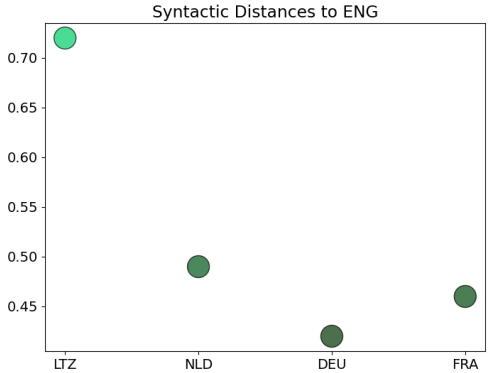
(1) English translations from LTZ, DEU, NLD, FRA should demonstrate the same distances as between non-translated LTZ, DEU, NLD, FRA.
Based on URIEL Typological Compendium (Littell et al., 2017) linguistic feature sets:

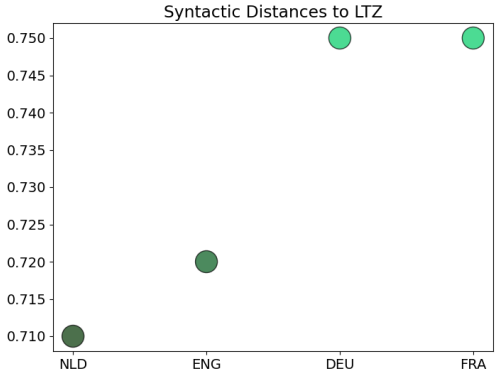




Our hypotheses

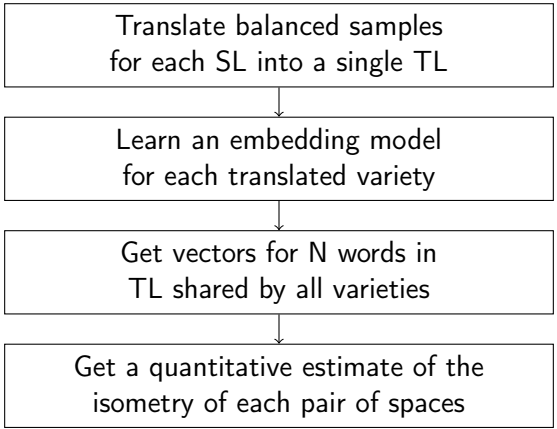
(2) The smaller the structural distance between SL and TL(ENG), the smaller the SL interference in translation (i.e. translations are more similar to 'normal' TL)





Approach

Calculate (dis)similarity between translated varieties as isometry of their embedding spaces based on vectors of N shared words.



Materials and resources

Source of data and register : parliamentary debates

Data parameters :

source language (SL) data

lang	docs	sents	wc
LTZ	500	8353	169,858
DEU	500	8169	146,111
NLD	500	4738	98,056
FRA	500	3530	92,624
ENG	500	7172	154,842

Translations into ENG

WC: 83.6 K (FRA) – 162.6 K (LTZ)

sent_length: 18.79 (DEU) – 23.69 (FRA)

Translation model : `facebook/nllb-200-distilled-600M` (max_length=100)

Vectorisation method : fasttext skipgram (300, -5:5, 1:10, iter=5), lowercase, no punct

Graph-based isometry measure : Gromov-Hausdorff distance (GH), `gudhi` library

The Gromov-Hausdorff distance (GH) distance

In the vector space of each translation variety (=translations from each SL)

1. compute full square distance matrix using pairwise Euclidean distances between vectors of shared words,
2. build a graph from the resulting metric space with N words as vertices, and distances as edges,
3. for the neighbourhood of each word (graph vertice), calculate the stability/robustness of its structure (persistence diagrams),
4. find the shortest distance between the same words from the two spaces for which there exists a perfect match between the points in an orthogonal map of one space into the other (the bottleneck distance).

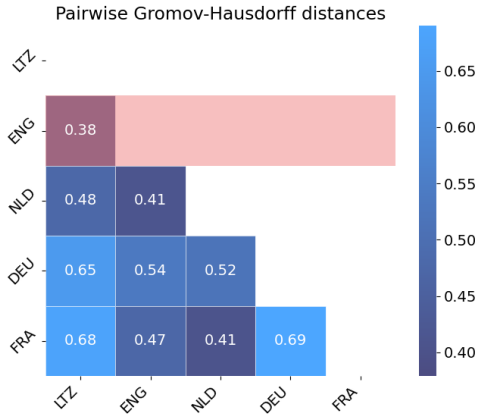
Results

if freq $>$ 10 in any variety, there are only 620 shared words

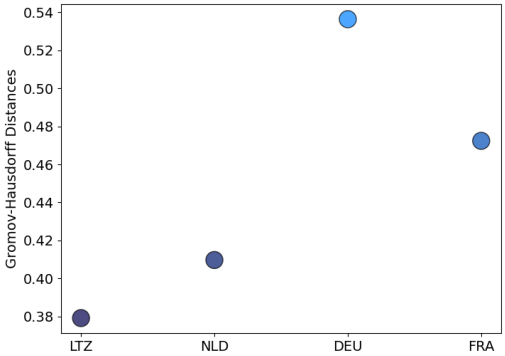
	LTZ	DEU	NLD	ENG	FRA
the	11948	11510	7305	12714	6768
to	4960	5010	2930	5090	2795
of	4533	4562	2946	6032	3441
and	5047	4060	2567	5590	2492
...					
down	45	12	19	22	12
apply	19	11	26	20	13
moment	45	11	24	47	11

Results

Languages: In the looking glass of NMT

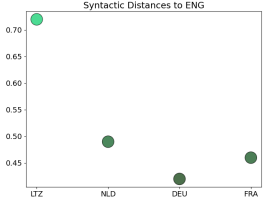


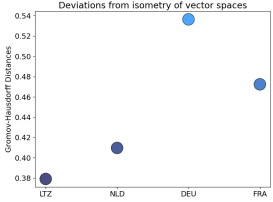
SL-interference effect



Summary and future work

1. Language differences seem to be reflected in translation:
   - (syntactic) LTZ-NLD-ENG-DEU-FRA → (isometry) LTZ-(ENG)-NLD-DEU-FRA
2. Proximity to the SL does not reduce the 'shining-through' effect (interference):





Explore confounding factors:

- NML quality effect (compare to human translations),
- impact of embedding quality (use larger data, other embedding method),
- selection of words that are more or less likely to change their distributional semantics in translation.

Thank you!

Topology of embedding spaces and language typology

SFB 1102 – Information Density and Linguistic Encoding (IDeaL)
funded by the Deutsche Forschungsgemeinschaft, Project ID 232722074

Questions?

Presenter:
Maria Kunilovskaya
maria.kunilovskaya@uni-saarland.de

References I

Bjerva, J., Östling, R., Veiga, M. H., Tiedemann, J., and Augenstein, I. (2019). What do language representations really represent? Computational Linguistics, 45(2):381–389.

Chowdhury, K. D., España-Bonet, C., and van Genabith, J. (2021). Tracing source language interference in translation with graph-isomorphism measures. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 375–385.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

References II

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14.

Patra, B., Moniz, J. R. A., Garg, S., Gormley, M. R., and Neubig, G. (2019). Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. arXiv preprint arXiv:1908.06625.

Zou, L., Saeedi, A., and Carl, M. (2022). Investigating the impact of different pivot languages on translation quality. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research), pages 15–28.