`GramTypix`: a Python pipeline for subtoken-based language typology and probabilistic semantic mapping from parallel corpora

Nilo Pedrazzini

Background

Token-based typological approaches (Levshina 2019, 2022) provide language typologists with ways of investigating systematic cross-linguistic variation in a data-driven way, by making generalizations and classifying languages based on correspondences between tokens observed in massively parallel corpora. Probabilistic semantic maps (Croft & Poole 2008; Wälchli 2010) have now long been the predominant token-based tool to induce cross-linguistically salient dimensions from massively parallel corpora in the study of various constructions and semantic concepts, such as semantic roles (Hartmann et al. 2014), person-marking (Cysouw 2007), causatives (Levshina 2015), word order entropy (Levshina 2019), motion verbs (Wälchli & Cysouw 2012), and temporal adverbial clauses (Haug & Pedrazzini 2023), to name a few. A major limitation of a purely token-based approach is that it does not allow for the capture of variation within the semantic space of constructions encoded morphologically, instead of lexically. This is particularly an issue for target languages where *several* morphological strategies co-exist in addition to or instead of lexified means, and especially those lacking detailed descriptions of morphological markers that could be leveraged in typological studies.

Purpose of the tool

This contribution presents a 'subtoken-based' approach to building probabilistic semantic maps from parallel corpora.[1] The method integrates associations between lexical items in a source language and character n-grams in target languages, enabling the capture of variation without assuming the exclusive use of either lexical or morphological strategies for encoding a particular construction. The focus of the presentation is on clarifying technical aspects of the pipeline

[1] The tool can currently be found at `https://github.com/linguistanonymous/gramtypix`. Note that this is under an anonymous account, as a research paper using this tool is undergoing anonymous review. I will provide the de-anonymized link should this abstract be accepted for a tool presentation.

(named `GramTypix` as a nod to the concept of GRAM TYPE; Bybee & Dahl 1989; Dahl & Wälchli 2016), from the creation of a research-ready parallel dataset to the application of statistical techniques for the analysis of semantic maps, with particular attention to the customizable components of the pipeline. All parts of the pipeline are performed in Python, rather than in R as with previous tools for probabilistic semantic mapping (e.g. Cysouw 2015), because of Python's greater flexibility, better documentation, ease of integration into diverse research workflows, and potential expandability with emerging AI technologies.

Details of the tool presentation

The following components, summarized in the flowchart in Figure 2, are covered:

1. **Dataset creation**, including a closer look at a typical output of two off-the-shelf automatic word alignment tools, i.e. FastAlign (Dyer et al. 2013), one of the most user-friendly and fast tools, and SyMGIZA++ (Junczys-Dowmunt & Szał 2012), a two-directed, one-to-one model, which is more resource-intense, but more suitable for capturing the translations of certain constructions (e.g. subordinate clauses).

2. ***n*-gram search**. A breakdown of the methods used to look for morphological markers in the target languages, building on Asgari & Schütze's (2017) SuperPivot method, but heavily adapted for probabilistic semantic mapping. In particular, we look at:

   – automatically extracting stopwords in the target languages by leveraging significant associations between character n-grams and bespoke lists of lexical items in the source language;

   – looking for significant associations (by χ^2) between source words and target character n-grams. This step also shows the benefits of using off-the-shelf dependency parsers (e.g. from SpaCy; Honnibal & Montani 2017) to identify the headword (when relevant) of the source word(s) under investigation to look for potential morphological markers in the target languages (e.g. given the source words *when*, *while*, and *after*, we look for morphological markers on the target token aligned to the syntactic head of these subordinators, rather than only the subordinator itself);

   – grouping character n-grams into clusters of likely allomorphs of the same marker;

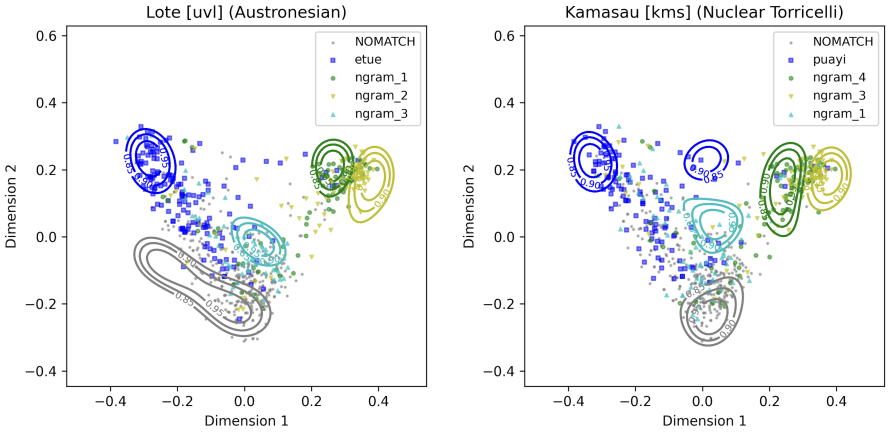
   – refining the token-based parallel dataset, extracted as part the previous step, with morphological markers.

3. **Distance matrix and multidimensional scaling (MDS)**, explaining how Hamming distance is applied as a measure of dissimilarity between pairs of contexts, and how MDS is used to project dissimilarity measures into a lower-dimensional space for visualization.

4. **Geostatistical interpolation**. How to implement an Ordinary Kriging model (using the PyKrige library; Müller et al. 2023) to interpolate the linguistic items (i.e. the parallel token or morphological markers) used in each data point by each language in the dataset, to look for patterns of coexpressions. The results from two different sets of hyperparameters are shown, to demonstrate their drastic effect on the experiment.

5. **Probabilistic semantic map generation** from the MDS matrix and the Kriging models (two examples are given in Figure 1).

6. **Statistical analysis**, with a brief demonstration of two approaches for identifying potential GRAM TYPES and for classifying languages typologically, i.e., Gaussian Mixture Modelling and hierarchical clustering, respectively.



(a) (b)

Figure 1: Kriging maps of WHEN-clauses for Lote (Austronesian) and Kamasau (Nuclear Torricelli).

References

Asgari, Ehsaneddin & Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In Martha Palmer, Rebecca Hwa & Sebastian Riedel (eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing*, 113–124. Copenhagen, Denmark: Association for Computational Linguistics. doi: 10.18653/v1/D17-1011. https://aclanthology.org/D17-1011.

Bybee, Joan L. & Östen Dahl. 1989. The Creation of Tense and Aspect Systems in the Languages of the World. *Studies in Language. International Journal*

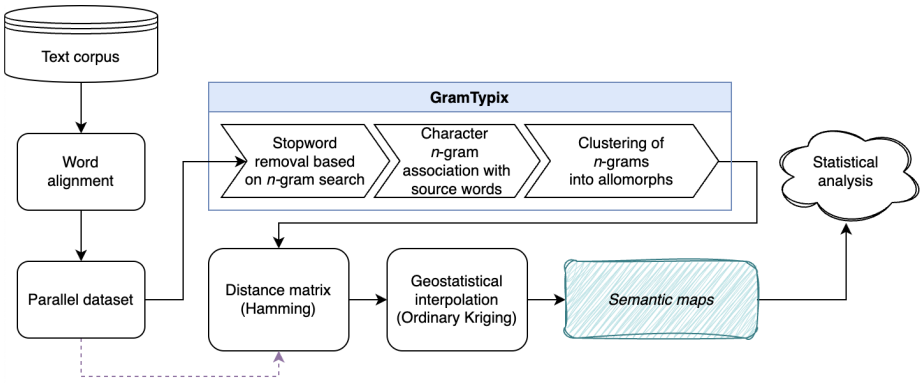


Figure 2: Flowchart of the main tool components.

sponsored by the Foundation "Foundations of Language" 13(1). 51–103. doi: 10.1075/sl.13.1.03byb.

Croft, William & Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics* 34(1). 1–37. https://doi.org/10.1515/THLI.2008.001.

Cysouw, Michael. 2007. *Building semantic maps: The case of person marking* 225–248. Berlin, New York: De Gruyter Mouton. doi:10.1515/9783110198904.4.225.

Cysouw, Michael. 2015. qlcVisualize: zenodo release (v0.1.1). doi:10.5281/zenodo.33174.

Dahl, Östen & Bernhard Wälchli. 2016. Perfects and iamitives: two gram types in one grammatical space. *Letras de Hoje* 325–348. doi:10.15448/1984-7726.2016.3.25454.

Dyer, Chris, Victor Chahuneau & Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies*, 644–648. Atlanta, Georgia: Association for Computational Linguistics. https://aclanthology.org/N13-1073.

Hartmann, Iren, Martin Haspelmath & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38(3). 463–484.

Haug, Dag & Nilo Pedrazzini. 2023. The semantic map of *when* and its typological parallels. *Frontiers in Communication* 8. doi:10.3389/fcomm.2023.1163431.

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Junczys-Dowmunt, Marcin & Arkadiusz Szał. 2012. SyMGiza++: Symmetrized Word Alignment Models for Machine Translation. In Pascal Bouvry, Mieczyslaw A. Klopotek, Franck Leprévost, Malgorzata Marciniak, Agnieszka Mykowiecka & Henryk Rybinski (eds.), *Security and intelligent information systems (siis)*, vol. 7053 Lecture Notes in Computer Science, 379–390. Warsaw, Poland: Springer. http://emjotde.github.io/publications/pdf/mjd2011siis.pdf.

Levshina, Natalia. 2015. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487–520. doi:doi:10.1515/flin-2015-0017.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.

Levshina, Natalia. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology* 26(1). 129–160. https://doi.org/10.1515/lingty-2020-0118.

Müller, Sebastian, Roman Yurchak, Benjamin Murphy, nannau, Malte Ziebarth, Sudipta Basak, Marcelo Albuquerque, Mark Vrijlandt, Matthew Peveler, Daniel Mejía Raigosa, Harry Matchette-Downes, Jordan Porter, Rhilip, Scott Staniewicz, Will Chang & kvanlombeek. 2023. Geostat-framework/pykrige: v1.7.1. doi:10.5281/zenodo.10016909. https://doi.org/10.5281/zenodo.10016909.

Wälchli, Bernhard. 2010. Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery* 8(1). 331–371. doi:10.1349/PS1.1537-0852.A.356.

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50. 671–710. https://doi.org/10.1515/ling-2012-0021.