Landscape of linguistic research

## Abstract

In this paper we analyze abstracts from linguistic journals in order to create a 2D representation of the entire linguistic field. We extracted and annotated a list of linguistic journals, gathered the abstracts and metadata for all scientific papers from those journals, analyzed them using a language model and visualized them using *t*-SNE .

**Keywords**: landscape of linguistic research, bibliometrics. linguistics, language model

## Introduction

Linguistics is a vast, complex and interdisciplinary field that deals with the study of language, its structure and spheres of its functioning and formation. The understanding of linguistic areas and their interrelationships is crucial for developing a comprehensive understanding of how language science is evolving. However, the amount of linguistic papers that appear each year is overwhelmingly huge, so it is hard to keep track of changes in different subfields of linguistics and their interaction. Specialists in certain subfields know what is going on in the field of their speciality and beyond. However, teaching linguistics to new generations lacks the overall view of the field that was earlier the subject of linguistic theory. In our paper we try to reproduce the division of linguistics into subfields using the bottom-up method rather than predefined theoretical views.

In this paper we present an approach for creating a linguistic literature ‘landscape’ of the subfields of linguistics. There is a similar work on biomedical research [González-Márquez et al. 2023]. The map we create will provide a visual representation of the intersection between different linguistic areas, including more traditional ones, such as phonetics, morphology, syntax, semantics and also branches of applied linguistics. The semantic map will be based on abstracts collected from linguistic journals and other metadata. The methodology we present involves several steps, including journals markup, data collection, preprocessing, model vectorization, and dimensionality reduction.

…

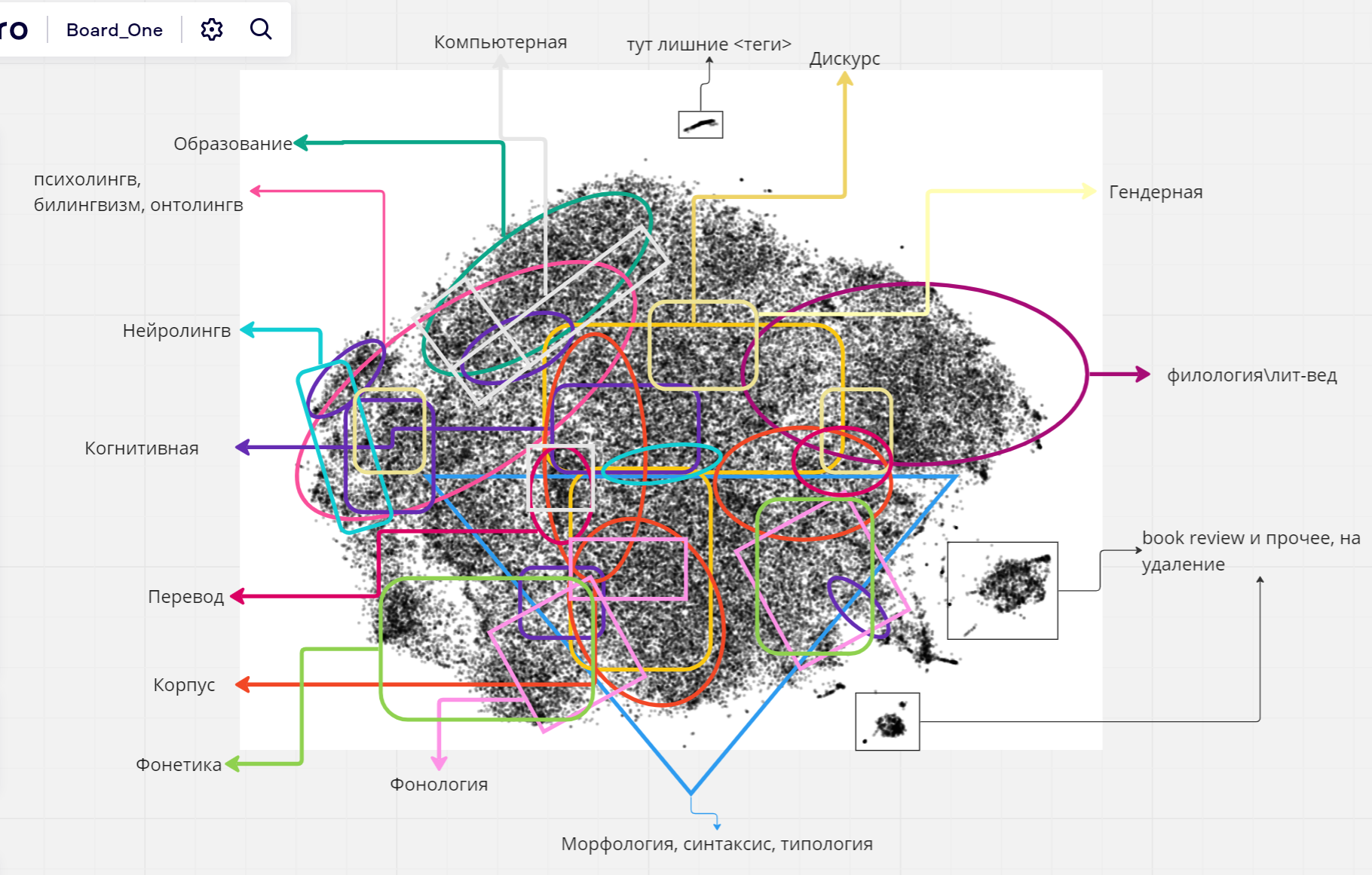
The semantic map we create will provide a valuable contribution to the field of linguistics by visually representing the relationships between different linguistic spheres. The semantic map will be useful for researchers, educators, and students who seek to understand "where they are" in linguistics. Additionally, the methodology we present in this paper can be applied to other fields to create semantic maps that visually represent complex relationships between different directions of subject.

## Data

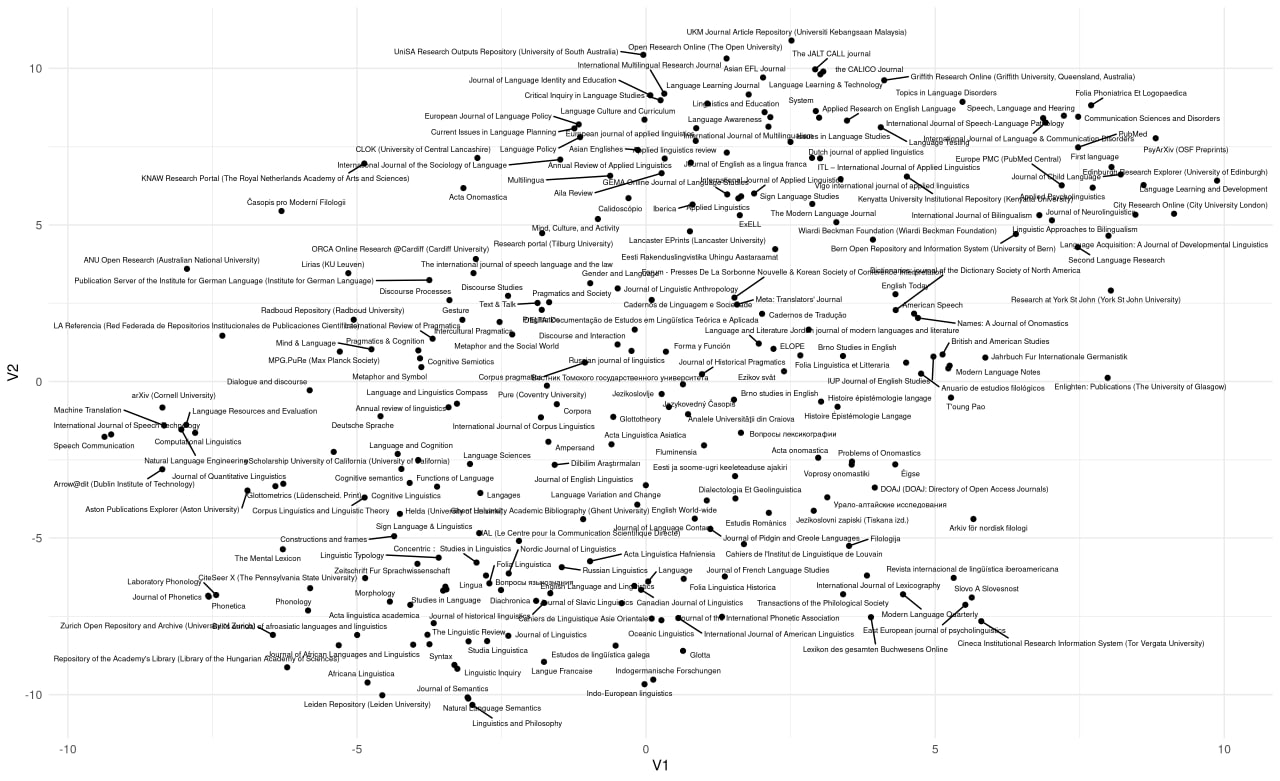
## Methods

## **Results**

**/**тут предварительная карта, данные почти собраны, но еще есть что чистить и необходима разметка. Здесь пока области отмечены исходя из упоминания соответствующего направления в аннотации/



\Попытка сделать journal2vec\



**Methodology**

**Journals markup**

In order to accomplish our objective of constructing a semantic map of linguistics through the vectorization of article abstracts, it was necessary to compile a comprehensive inventory of linguistic journals from which we could gather the requisite abstracts.

Initially, the study involved two data sets comprising 2755 journals in total dedicated to the fields related to “language studies”, obtained from the HSE and the University of Helsinki. In addition, the list of journals also included information about the prestige of each journal according to each source.

However, it was discovered that not all the journals were entirely relevant to the field of linguistics, as some of them were oriented towards other complex areas such as philology, literary studies, media communication, and teaching, among others. As a result, there was a requirement for the manual classification of these journals to determine their appropriateness for inclusion in the study.

To accomplish this, six employees of the Language Convergence Laboratory and several students undergoing summer practice were enlisted to participate in the markup. The decision to classify a journal as belonging to the field of linguistics was based on the aims and goals of each journal (and also according to personal beliefs on what is linguistics, and what is not), which were also collected and added to a common dataset with the journals. This process ensured that only journals that were directly related to the field of linguistics were included in the study, thereby enhancing the accuracy and validity of the data.

During the journal analysis we discovered a number of journals that exhibit a fundamentally interdisciplinary nature, wherein articles from both linguistic and non-linguistic disciplines are present. These journals were duly identified as such and were not included in this current study. It is plausible that in future research endeavors, these interdisciplinary journals may prove to be valuable resources in the investigation of larger scientific domains beyond the scope of linguistics. Alternatively, the abstracts of articles featured in these journals may be subjected to markup individually.

As a result of the markup, we identified 232 linguistic journals and 203 interdisciplinary journals.

With the journal list compiled, there needed to be an effective parsing algorithm that would browse journal pages and extract the needed data. Several approaches were compared in terms of data mining, and ultimately OpenAlex had been chosen as the most optimal instrument for data analysis.

OpenAlex is …

The key variables in the dataset for each article were:

1. OpenAlex identificator
2. DOI
3. Author(s)
4. Title
5. Year of publication
6. Abstract
7. Journal name
8. Volume number
9. Issue number
10. Page numbers
11. Information about paper retraction

Abstract texts contained a lot of information that was used in the subsequent analysis, but it needed to be processed in order to be fit for further research purposes. One of the main requirements was the stopwords cleanup because the total amount of pronouns, specific verbs (like “do” and “be”) and many other common words would be detrimental to the future model vectorizations.

But beforehand, all words of the text corpus were modified to be lowercase in order to bypass the risk of the vectorization system being case-sensitive and as result negatively affecting its own quality of work.

**Model vectorization**

**Я использовал GloVE, но думаю в сторону Bert**

**Dimensionality reduction**

***t*-SNE**

Пока просто ссылки:

1. Списки журналов ВШЭ Наукометрического центра: <https://scientometrics.hse.ru/lists_test>
2. Списки журналов Хельсинки <https://www.tsv.fi/julkaisufoorumi/haku.php>
3. The landscape of biomedical research: <https://static.nomic.ai/pubmed.html>
4. Crossref <https://cran.r-project.org/web/packages/rcrossref/index.html>
5. OpenAlex <https://cran.r-project.org/web/packages/openalexR/index.html>