ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему Семантические фреймы в архитектуре трансформер: применение в лексико-типологическом исследовании

Тема на английском The Notion of Semantic Frames in Transformer Models: Application to Lexical Typology Study

Студентка 3 курса группы № 182

Аксенова Анна Александровна

Научный руководитель

Сериков Олег Алексеевич

Приглашенный преподаватель

Москва, 2021г.

Table of contents:

1. Introduction

Lexical typology is a relatively young field of linguistics, the main task of which is the comparative analysis of the meanings of words in different languages. It originates from the cognitive linguistics approach and the starting point for the development of lexical typology was the article by Berlin, Kay (1969) on the typology of colour designations. The authors proposed a novel and clear method for comparing the lexicon of different languages, which is still widely used. At present, the interest in the typological analysis of vocabulary only continues to grow.

Although this field of linguistics is quite close to traditional grammatical, phonological or morphological typology, some major methodological differences should be mentioned. First, the sample of languages for the study might be smaller (20 languages compared to 200 languages for grammatical study) (Haspelmath 2003: 217). Second, the interest of lexical typology is the comparison of the systems of related languages and even dialects. For grammatical typology, the key factor is language diversity, and the sample of one language family will not be considered representative. Vocabulary changes are faster than shifts in grammar and phonetics. In addition, lexical typology uses a different method for data collection. For the grammatical typology, the main source of information is the grammars of the studied languages, however, the definitions and examples presented in the dictionaries do not correspond to linguistic reality, therefore they are not used as the main source of data for research in cognitive semantics. The most important tool of a lexical typologist is the typological questionnaire, which contains possible situations of the use of the lexeme of interest. The data of such questionnaires is usually supported by corpus research and dictionary review. In this study, I follow the traditions of the Moscow lexico-typological school (MLexT), which combines the approach of frame semantics and grammatical typology. The most important concept of lexical typology is a frame i.e. prototypical situation in which words of the considered semantic field can be used. As there is a set of arguably universal grammatical concepts (Plungyan 2000) frames are thought to be language independent.

Even though a lot of research has been done in this area (Koch 2008; Rakhilina, Reznikova 2016), some methodological difficulties are still not overcome. First of all, analysis of vocabulary requires extensive and representative material, which in most cases cannot be gleaned from lexicographic sources. The researchers have to create questionnaires and collect the data which also requires the participation of native speakers and experts in the languages under study. The complexity of the entire process does not allow a detailed analysis of large semantic fields in a significant number of languages. Therefore, in most cases, one has to seriously limit either the number of languages in the sample or the degree of detail in the analysis.

Automatisation of the laborious activities, such as collection and processing of lexical material would make it possible to obtain structured data for many languages of the world, prepared for lexicographic processing and direct comparison of the languages. In addition, automation will allow one to get more accurate results, since various methods of data collection are available, for computer research from common crawling to processing the corpus.

However, until recently there were no tools to capture the full spectrum of the semantics of one lexeme to compare it with other words within one language or map it with the words in other languages.

Moreover, clustering of the semantic frames might be an interesting probing objective for modern language models. Today, many language models are trained for a specific task, for example, masked token prediction, translation ranking etc. However, with the development of natural language processing (NLP), tasks become more and more abstract and eventually come to one purpose: learn to understand the request and generate an adequate response in natural language. In modern language models, tokens are represented in the form of embedding vectors, and if at the initial stages of NLP development, embeddings stored only information about the frequent contexts of word use, then modern advanced models create vector representations of words containing syntactic and morphological information (Conneau et al. 2018).

Understanding the nature of embeddings can help us think about why some models succeed and others fail, what they've learned so far, and what improvements have to be made. Modern research aims to interpret the properties of the internal language representations of models such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2018). To check the intuition about whether the properties of word vectors coincide with their linguistic properties (for example, parts of speech), pre-trained embeddings are used as features for the training of another model (for example, a classifier) to determine this property.

The key idea is that achieving a good quality when predicting a linguistic feature correlates with the fact that this feature has already been encoded in the word vector. This means that if a classifier trained on pre-trained embeddings perfectly separates adjectives and participles, then the vectorisation model (ELMo, BERT, LSTM (Hochreiter, Schmidhuber 1997), etc.) managed to capture the difference.

There is quite a lot of research regarding the ability of multilingual models to represent joined vector space for different languages (Choenni, Shutova 2020).

In this paper, I am going to focus on two aspects of the lexical-typological study: the traditional approach and computational modelling of semantic space. I am investigating the verbs of changing in Russian and Swedish. Not being closely related, these languages provide the opportunity to estimate the quality of language representations in multilingual context language models. In section 2 I overview the previous research that has been conducted in this area by theoretical linguists and in section 2.2. I will discuss the computational papers devoted to language modelling and the problem of language-agnosticism in modern architectures. Section 3 is devoted to the semantic field of changing. In section 3.1. I present the field of changing in Swedish (the research was carried out by a group of HSE students including me, Ekaterina Voloshina, Alexandra Trepalenko, Ekaterina Taktasheva), in section 3.2. the field of changing in Russian is described. Section 4 is focused on the attempt to analyse the ability of various models to represent semantic fields of changing in those languages. First, I describe the data and preprocessing (section 4.1.), then I test 5 types of vectorisation models (section 4.2.) and explore the semantic fields in Russian and Swedish separately (section 4.3.). In section 4.4. I describe the translation procedure and discuss the result of the model-based research. In section 5 the traditional and computational approaches are compared.

2. Literature overview

In this section, I am going to overview the previous research that has been conducted in the related field of studies.

2.1. Theoretical papers

The tradition of the Moscow semantic school was founded by Apresyan and the MLexT group is following his framework. Although the research group led by Ekaterina Rakhilina and Tatiana Resnikova is quite young, several studies have already been conducted.

The main ideas of the approach and major studies are summarised in Rakhilina, Reznikova (2013). Authors emphasize the parallels between lexical and grammatical typology showing that the semantic maps of certain fields can predict semantic shifts and represent the connection between various domains of the lexicon. Lexical typology attempts to create a system of patterns in a seemingly chaotic language vocabulary and find interlanguage universals that allow describing the similarities and differences of languages from the point of view of conceptualizing the surrounding reality. First lexical typology studies were devoted to adjective classification (Kjuseva et. al. 2013), nowadays most of the studies are focused on the predicates. Being the root of the sentence predicate denotes the whole situation which makes the study more complicated since more factors should be taken into consideration. An adjective modifies only the noun phrase and the choice of lexeme depends on which noun controls it, while the situation potentially includes both explicitly expressed participants (Agent, Patient, Instrument etc.) and external discursive factors (e.g. background, foreground).

Since typological research includes the analysis of quite a big number of examples, questionnaire survey and consultation with linguists specialized in the language of study, several attempts to create the automatisation of the process have been made. Orexov, Reznikova (2015) suggest the methodology for computational analysis of typological studies. The paper presents the pipeline for comparison of various strategies of conceptualising ‘thickness’ in Russian, English and German. The authors suggest using Google bigrams[1] with target adjectives since the choice of an adjective to a great extent depends on the head noun. The frame identification is based on one-to-one translation of the nouns from Russian bigrams into the target language using online dictionaries. Although this approach allows us to make some assumptions about the frame structure of the field, Orekhov & Reznikova (2015) admit that one of the main limitations of it is the lack of sentence context both for translation and for getting the semantic profile of an adjective.

One of the most novel works in this area was presented by Ryzhova (2020). The author proposes new methods for the automatic collection and analysis of lexical-typological data regarding adjectives. The paper is focused on showing methods that allow you to look at all the stages of research from a different angle. Ryzhova (2020) uses Distributional Semantics Models to represent the word combinations as vectors. Ryzhova (2020) suggests collecting a

[1] http://books.google.com/ngrams

set of words for the questionnaire by multiple translations of the basic set of verbs, through online thesauri.

2.2. Computational background

NLP (Natural Language Processing) is being actively implemented in linguistic research (Orexov, Reznikova 2015; Ryzhova 2020; Rabinovich, Stevenson 2020). Currently, there is a sufficient number of models capable of translating the meaning of a text into a machine-readable format, they are based on methods of distributive semantics. The key idea of distributional semantics is that the meaning of a word can be determined by the context neighbours of the word. (Firth 1957) Each word is represented as a vector, in one form or another, reflecting the words next to which it is present in the training corpus. The models calculate the vector of probabilities of the co-occurrence of the desired word with other words, this allows using the mathematical apparatus to assess the accuracy of the meaning or semantic proximity.

From the point of view of distributive semantics, the meaning of a word can be described using context. Lexical units are represented in the form of vectors, the dimensions of which contain a weighted joint occurrence of a given word with others in a certain corpus. The words with larger frequency get a higher rate, while rare words get lower one, which results in a simpler and more economical vector of "main components" word embedding – this mechanism lies behind the word2vec model. Introduced by Mikolov et. al. (2013) word2vec architecture has become quite successful, however, the efficiency of the model is limited due to the training objective it exploits. It does not consider wide context but tries to predict the word based on the nearest neighbours, therefore it can capture neither syntactic nor discourse information.

More efficient architectures like RNN- (Pineda 1987) and LSTM-based were designed to solve this problem. They process the sentence from the left to the right edge (or vice versa) imitating the process of reading. However, these models also have several disadvantages. The most important one for this paper is that they fail to track long-distance or cross-sentence dependencies. BERT (Bidirectional Encoding Representations from Transformers) model, introduced by Devlin et al. 2018, solves this problem. BERT is based on Transformer architecture which is bidirectional and thus catches wide context. It was pre-trained on two training objectives: Masked Language Modelling and Next Sentence Prediction.

Compared to RNN and LSTM, BERT performs much better due to the self-attention mechanism. The word meaning can be changed based on the new information the model gets while processing the sentence. The attention mechanism is used to weigh the words in the sentence to understand which of them influence the target word the most (Vaswani et. al. 2017). Based on BERT architecture, various modifications with different training objectives have been invented.

Since the research is related to typological comparison, the multilingual model is required. Multilingual BERT represents token and sentence embeddings for more than a hundred languages including Russian and Swedish that we are interested in. However, it has been proven that the cross-linguality of mBERT embeddings is debatable since some of the languages were underrepresented in the training data (Libovický et. al. 2019). The grouping of languages in the vector space is correlated with their genealogical relation and embeddings do not represent similar semantic phenomena in different languages similarly which is crucial for our research, since we use frames as language-neutral items (Libovický et. al. 2019, Kulshreshtha et. al. 2020).

To make the word embeddings language-agnostic (Rabinovich, Stevenson 2020) suggest using the MUSE model. Authors align the fastText monolingual vectors to compare the cross-linguistic similarity of semantic concepts. This study proposes a novel approach towards lexical-semantic typology showing that language models are capable of capturing typological differences between lexicons of various languages. I believe that multilingual transformer embeddings that generally capture semantic information better than fastText, can be used in typological research.

3. Theoretical basis: ‘Change’ in Russian and Swedish

Traditionally typological study consists of 4 stages (Ryzhova 2020):

1. Drawing up a questionnaire, i.e. the preliminary definition of a set of frames and selection of diagnostic contexts for each frame.

2. Collecting data from other languages in the sample.

3. Drawing up a semantic map that allows one to describe and visualize the system characteristic of each language, as well as reflect the main typologically relevant principles of the organization of the field under consideration.

4. Analysis of the types of systems implemented in different languages.

This paper is devoted to the first two stages of research.

The questionnaire plays a crucial role in the typological studies, however, there is no conventional way to create the list of necessary and sufficient stimuli, i.e. for each study it depends on the background and assumptions of a linguist. Several studies allow scholars to create questionnaires in a clear way since they consist of a limited set of predefined features, such as colors (Berlin, Kay 1969). In lexical typology, the study is based on the linguistic behaviour of lexemes and not on the physical properties of the referent, given the fact that the metaphorical use of lexemes is also taken into account. These kinds of studies require time-consuming and thorough work with corpora. Moreover, some proficiency in the language is required. I would like to emphasize that the final set of frames and stimuli is not universal and is based on a linguist's personal opinion to a certain extent.

Several computational approaches have been introduced during the last decade. (Dahl 2007; Wälchli, Cysouw 2012) works should be mentioned. The proposed method is based on the word correspondence in the parallel corpora and the questionnaire is formed based on the information about several languages. There is no doubt that this approach is quite limited since for many languages there exists neither enough data nor parallel corpora.

The other approach was presented by Orexov, Reznikova (2015) for adjectives. It exploits the Google n-gram Corpora. For the basic list of lexemes denoting features in one language, the bigram collocation list is collected and sorted by frequency. Then all the nouns from the most frequent bigrams are translated to the target language and clustered by their collocational properties. There are several major drawbacks of this approach. Firstly, it is limited to the list of languages presented in the Google n-grams. Secondly, it is arguably applicable only for adjectives since the only argument they modify is a noun. If one wants to study verbs, bigrams will not be representative enough. Moreover, considering only bigrams adds some limitations to the questionnaire. There are many cases when it is not the noun that defines the choice of adjective but the pragmatics of the whole sentence. Finally, as Ryzhova (2020) points out, questionnaires compiled automatically based on a bigram corpus are too cumbersome and in this form cannot be used for fieldwork with informants.

Within the MLexT group approach, questionnaires are collected in the following way: the first draft is made in Russian, based on the linguistic intuition of the scholars since it is their native language. The intuition is always supported by the analysis of corpus data and lexicographic sources i.e. dictionaries, thesauri etc. Potential frames are formed and controversial points are clarified by the intuition of naive native speakers. After that, the questionnaire is tested in other languages and corrected regarding oppositions and transfers that were not found in Russian. This methodology was used for the traditional study described later.

In this paper, I am going to focus on the verbs and frames related to change.

The semantic field of changing is represented mostly by the transitive verbs and includes the Agent-like participant and the object of changing. The latter can be both concrete and abstract. Some other additional factors might influence the choice of the predicate, for instance, the reason for changing e.g. malfunction of the object. The languages can also differentiate between the situations with controlled change, change caused by the external influencer. I conducted a traditional study for verbs of changing in Russian and Swedish and introduced semantic maps for the field in these languages.

The questionnaire[2] consisted of 34 stimuli and the respondents were asked to translate the sentences from English to their native languages. The list of stimuli included the following possible oppositions:

1. Partial VS complete change of the object

[2] The questionnaire can be found here: https://docs.google.com/spreadsheets/d/1WoHV1bv3csVzHUutqzrw01lTxhHqgYmIWKYrVrQiHjg/edit?usp=sharing

She has noticeably [changed] since getting married.

The chef was angry with the cook because he [had modified] the recipe even though the dish turned out to be excellent.

2. Type of the possessor (animate/inanimate)

Leaves [change] their color in autumn.

Have you [changed] your haircut? You look great!

3. Type of object (abstract/concrete)

If you work hard, you can [change] your life for the better.

I [exchanged] a sheep for a goat.

4. Includes change of the spacial position

She noticed that he added something into her wine and discreetly [swapped] their wine glasses when he turned away.

5. Change of the object VS Change of the quality

We [changed] the color of the walls. We were tired of grey walls, so we decided to paint them green.

She asked her son [to change] his shirt, but he refused to do it and kept the dirty one.

6. The reason for changing

He [replaced] a burned-out lightbulb.

She asked her son [to change] his shirt, but he refused to do it and kept the dirty one.

7. Equal VS Not equal exchange

I [exchanged] a sheep for a goat.

They say that Rubens original painting [has been replaced/swapped] during the transportation.

8. Cyclic change VS Linear change

I [exchanged] a sheep for a goat.

Autumn [turns to] winter.

3.1. Swedish

Based on the data collected from 6 Swedish native speakers[3] I got the following results.

The field of changing includes 4 predicates: *förändra, ändra, byta, växla*. The main opposition is based on the manner of changing.

The verb *förändra* is used in contexts of global and comlex change.

(1) *Alla barn drömmer väl om att kunna förändra världen.*

All children dream probably of being able to change the world.

Förändra can be used in the same contexts as *ändra*, however, the meaning is different.

(2) *Han ändrade sin åsikt med åren om en del av reformerna, även om han fortsatte att försvara vissa av dem.*

He changed his mind over the years about some of the reforms, although he continued to defend some of them.

(3) *Att förändra någons åsikter med fakta är i stort sett omöjligt.*

Changing someone's opinions with facts is virtually impossible.

In example (2) one can see the partial change, while for (3) the change is complete and requires complex actions, although the verbs govern the same object. This type of context does not present any problems for the traditional theoretic approach, since native speakers can explain the difference, whilst for the model which is based only on the context and does not represent linguistic intuition it might be problematic.

Ändra is one of the most commonly used verbs. It is employed to express modification like in (4).

[3] For Swedish verbs I use the data collected by me and my research group in HSE for the lexical typology course 2020.

(4) Regeringen ändrade regler för beskattning.

The government changed the taxation rules.

The change is usually controlled and applies to the object itself.

The most common word from the field of change in Swedish is *byta*. The basic semantics behind the verb *byta* is replacing one object or feature with another. The last frame might be hard to distinguish as the contexts are quite similar to the modification expressed by *ändra***.** One of the sentences that presented problems for the interpretation of the responses of the native speakers was (5).

(5) Löv byter/ändrar färg på hösten.

Leaves change their color in autumn.

The background for the difference is the metaphorical conceptualisation of modification as a change of the object: “the leaves changed color so much as if they had turned into different leaves”. This description of the difference was confirmed by our respondents.

Moreover, *byta* can be used in contexts of changing clothes, changing position or exchange. It can be used with exchanging both two homogenous objects (objects of one class) and two completely different ones (6).

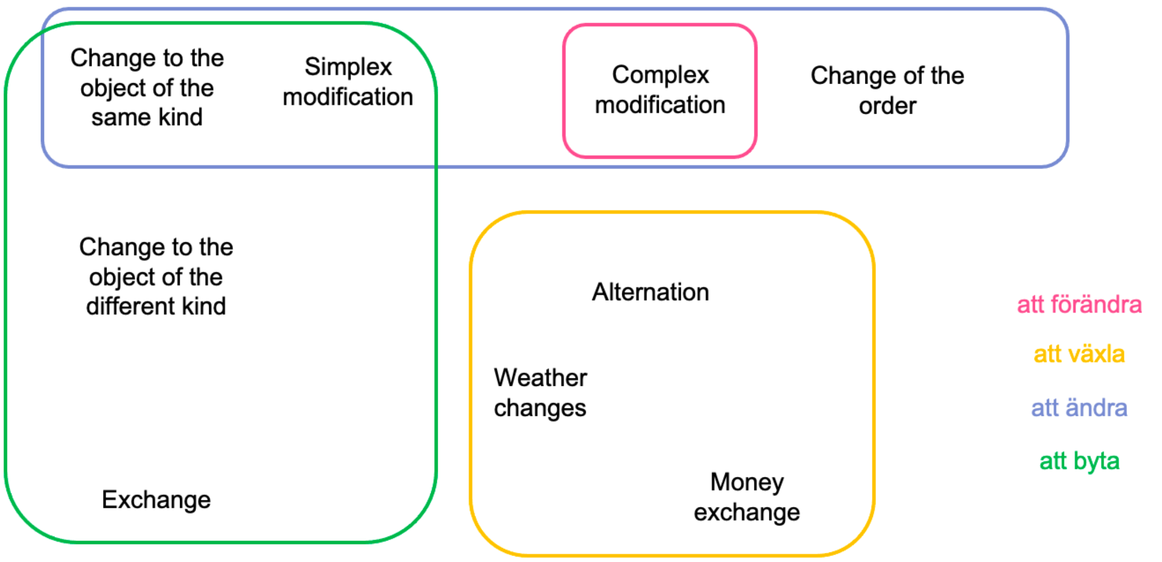
(6) Harald har bytt sin gamla bil mot en ny, fin cykel.

Harald has exchanged his old car for a new, nice bike.

There is one more specific verb – *växla* – which is used for highly collocational contexts like currency exchange and weather changes. *Växla i*s also used for alternations that happen periodically.

As one can see in Swedish there are different verbs to express the situations of changing and none of them is dominant. I present the semantic map for the Swedish verbs below.

Figure 1. Semantic map of changing in Swedish



I introduced 9 frames to differentiate between the verbs (Figure 1), namely

1. Change to the objects of the same kind (change the shirt)
2. Change to the object of the different kind (change shirt to a blouse)
3. Exchange (two participants give something for each other)
4. Simplex modification (change of features of the object)
5. Complex change (usually applies for abstract objects and requires complex actions)
6. Change of the order
7. Alternation
8. Weather changes
9. Money exchange

Notice that the last three frames are quite narrow and specialized, but the decision to use them in the map is supported by the strategies that exist in other languages. [4]

[4] Frames were discussed in one of the seminars within the Lexical Typology course with Tatjana Reznikova.

The main type of the opposition in Swedish is replacement vs. modification (*byta*/*ändra*). Interestingly, *byta* is more frequent and neutral because modification can be conceptualised as replacement: we change the property of the object as if we replace the object with another one.

3.2. Russian

For Russian we collected the answers of 5 native speakers. They were asked to translate sentences from English using the verb of changing, all the conclusions were based on their answers and my language intuition. According to the questionnaire there is only one basic root denoting changes in Russian – *men* (‘change’). I did not take into the account simplex imperfective verb *menjat'* since it can be used in all contexts. Variability is created by the prefix combinations for perfective forms. The field is represented by 5 major verbs and several highly specified ones. Let us start with the verbs that are used more often, i.e. *izmenjat', smenjat', pomenjat', zamenjat', obmenjat'*.[5]

Izmenjat' is a prototypical verb for modification in Russian.

(7) Stanislavskij vse vremja perepisyval svoj trud, čto-to izmenjal, dopolnjal.

Stanislavsky was constantly rewriting his work, changing something, supplementing it.

There is no specific word used in the contexts related to weather, which means that almost all the major verbs can be used including *izmenjat'*. One might notice that the frames covered by *izmenjat'* are quite similar to the Swedish *ändra***.** Interestingly, this verb has one more metaphorical meaning: it is used to describe cheating like in (8). This meaning is not taken into account in my paper since it has lost its connection with the field of changing and the verb exhibits different government patterns.

(8) Anna Karenina ne izmenjaet mužu.

Anna Karenina is not cheating on her husband.

[5] Notice that in this study I do not differentiate between perfective and imperfective verb forms. Their status in Russian is highly debatable (Janda et al. 2013) but I assume that the frame semantics of the verb is defined by the prefix and not by the grammatical form.

The second quite frequent verb is *pomenjat'*, which represents the change between two objects. The distribution of *pomenjat'* is quite similar to the one of *byta*, with the only exception: Russian has a specialized verb for the contexts of exchange.

At first glance the verb *smenit'* is very similar to the previous one and is used in the same contexts, however there is an opposition. While *pomenjat'* is quite neutral and generally can be used by native speakers in almost any context in oral speech, *smenit'* is more focused on the object that is changed. It is important that the object is replaced, thus it is usually not used in the contexts when we change an object to one of a different kind: in this case the second object is pragmatically important.

(9) *Mne nužno lobovoe steklo pomenjat'.*

I need to change my windshield.

At the same time there is another verb being focused on the object that replaces the previous one – *zamenit'*. The contexts usually describe the change to another type of object or change to the better object of the same kind when the previous one malfunctions.

(10) *Zameni im zavtrak ili lëgkuju zakusku — i lišnie kilogrammy načnut isčezat'!*

Replace it with breakfast or a light snack - and the extra pounds will start to disappear!

As I have already mentioned, there is a special verb for situations including the second active participant, even though it might not be overtly expressed in the sentence. *Obmenjat'* is used only in such contexts and is also used for the situations related to money exchange. I would like to note that some of the respondents used *pomenjat'* in the sentence like (10), which supports my assumption about *pomenjat'* being a neutral verb for the field of changing in Russian.

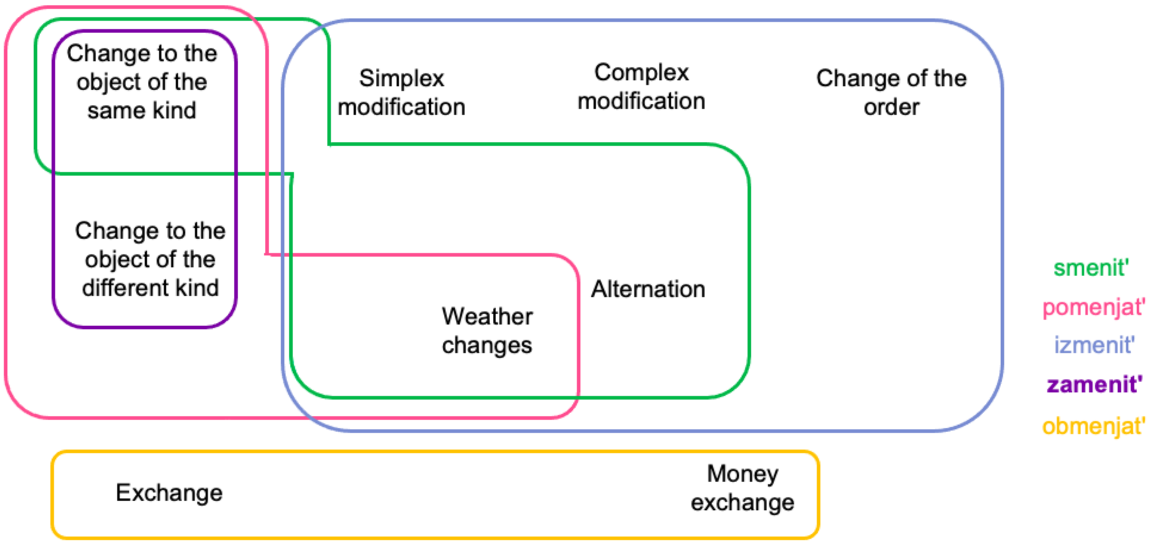
(11) Esli by ja ne znal o soderžimom ètogo čemodančika, ja by ego na vaš ne zadumyvajas' obmenjal.

If I did not know about the contents of this suitcase, I would exchange it for yours without hesitation.

Several verbs are used in a closed set of contexts. For the change with the negative connotation, the verb *promenjat'* is used. The situations 'replace one object with another illegally' and 'take over for someone' are covered by the verb *podmenit'*. I decided not to include them in the map to keep it more simple and comparable with the Swedish one.

The semantic map based on the same frames as in Swedish is presented below.

Figure 2. Semantic map of changing in Russian



The opposition in Russian partly resembles the Swedish one. The verb *izmenit'* corresponds to Swedish *ändra* whilst the group *pomenjat', smenit', zamenit'* behaves similarly to Swedish *byta*. However, the field of change in Russian is more complex. There is a triplet opposition that profiles different stages of the prototypical situation. The verb *smenit'* highlights the initial part of the situation, i.e. the final state is not important, that is why *smenit'* is not used in the contexts where the object is replaced with another one. The final stage of the action is profiled by the verb *zamenit'*. The idea is supported by the fact that *zamenit'* also takes the object that substitutes the previous one as a semantic argument (*zamenit' X na Y*). In case of *zamenit',* the object that is being changed should disappear, that is why it is not typically used in contexts with alternations and weather changes, as these changes are reversible. *Pomenjat'* is a neutral verb in this group but it is important that the change took place only once. As can be seen from the map, it is not used to describe alterations.

To summarise, Swedish and Russian exhibit different strategies to describe the situations of changing. There are more widely used lexemes in Russian and the main type of opposition between them is the modification of the same object VS substitution with another one and the profiled phase of the prototypical situation (the initial point of changing or the final one). As for Swedish, the split is based on the distinction between replacement and modification (the process or the result is profiled) with an extra verb for special collocational contexts like money exchange and weather changes.

4. Experiments

As I have already mentioned it was decided to focus on the first two stages of the LexTyp research pipeline. The goal of this part of my research is to estimate the performance of computational tools in linguistic typology study[6]. As far as I know, there has been no discussion about the contextual language modelling methods that can be used in this kind of investigation. In this section, I am introducing the pipeline for processing the Russian data and extracting Swedish verbs of the semantic field of changing. In section 4.1. I describe the dataset and data preprocessing, In section 4.2. I compare the performance of different encoders and explain why LaBSE shows the best estimation of semantic relations. Afterwards, I apply different kinds of clustering to the data of Russian and Swedish semantic fields of change. In section 4.4. I suggest the way of transferring the information about Russian verbs to Swedish vector space, interpret the results and discuss some limitations of the suggested approach. Finally, in section 4.5. I compare the results of traditional and computational studies.

4.1. Dataset and preprocessing

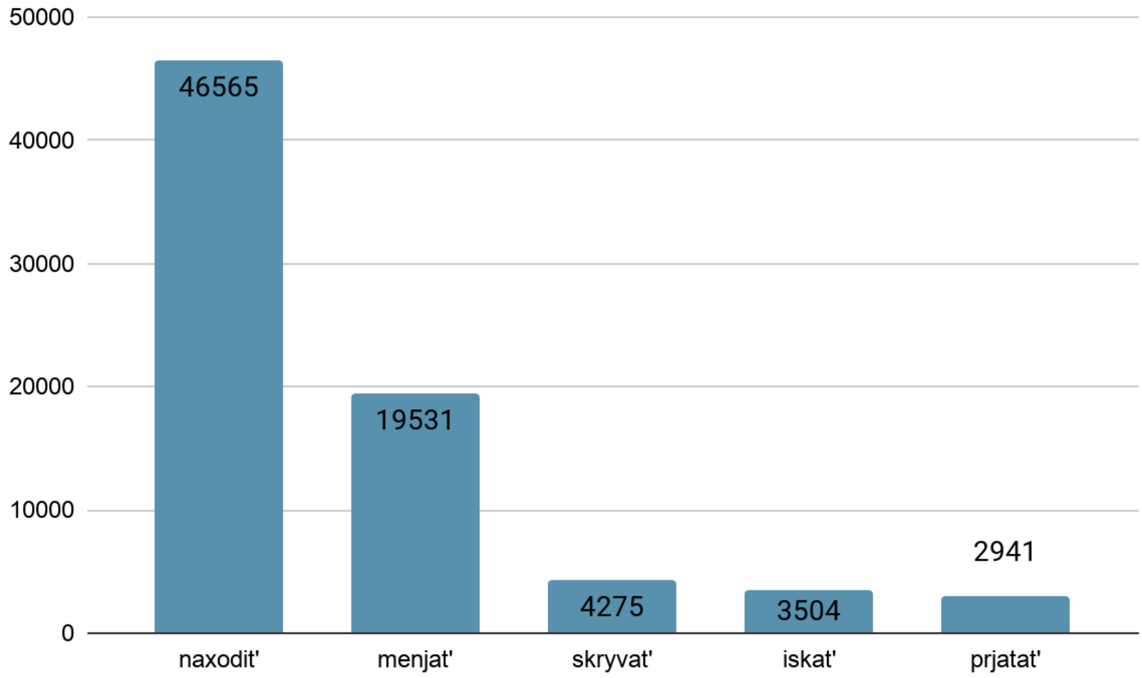
For all of the experiments, I have created a dataset based on a 1 billion token subsample of Taiga corpora[7], which was annotated in UD format (de Marneffe et al. 2006). The corpus consists of texts from literary magazines, naive poetry and news articles. I used regular expressions to extract the sentences with the forms of lexemes *prjatat', iskat', naxodit', menjat', skryvat'*, including various combinations with prefixes e.g. *priprjatat', poiskat'* , for simplicity, in the future, I will call a group of cognate verbs by a simplex verb. I decided to focus on these verbs since all of these semantic fields have already been studied and they have the same transitive model of the government.

Some limitations of the data have to be discussed. Firstly, the markup of the corpus contained quite many mistakes, e.g. punctuation marks were not separated from the lexical tokens, thus I had to check all the unique tokens that matched the regular expression. Moreover, there was no punctuation in some of the examples which means that a small text could have been treated like a sentence. I decided not to filter this kind of examples assuming that the wider context should not change the situational semantics of a frame.

[6] All the code and data can be found here: https://github.com/aaaksenova/term_paper2021
[7] https://tatianashavrina.github.io/taiga_site/segments

The number of sentences representing each type of the basic verb is presented in Figure 3.

Figure 3. The distribution of verbs in the full dataset

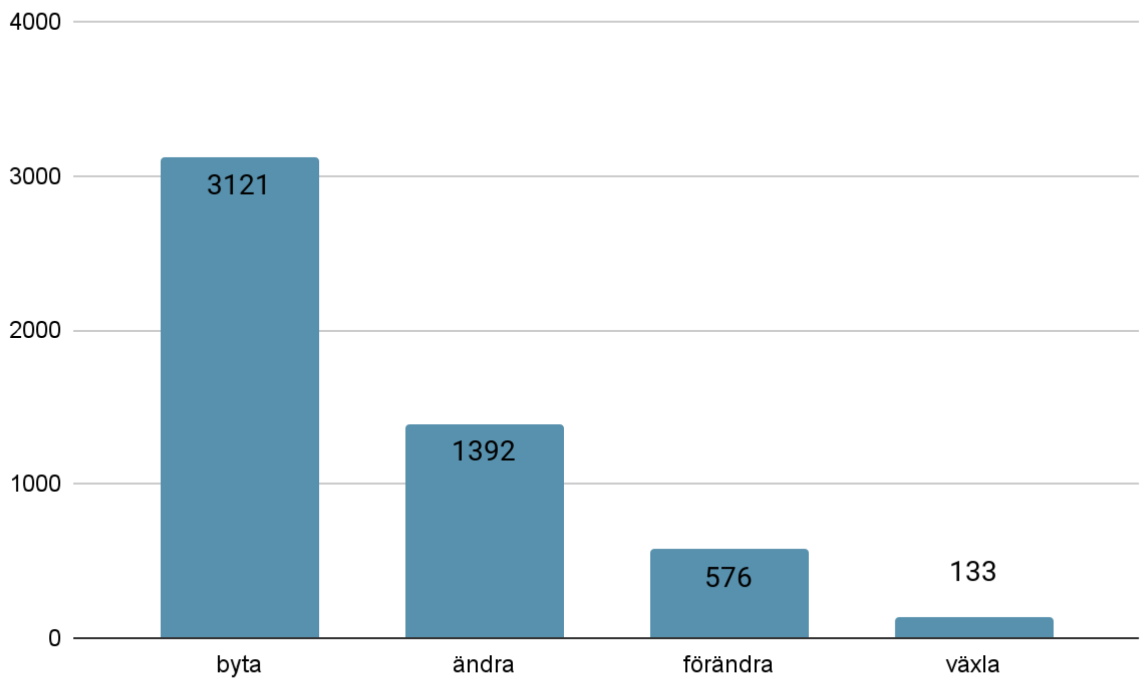


I extracted all the examples of the target field of change in a separate dataset. Afterwards, the sentences were automatically translated into Swedish and for each sentence, the verb corresponding to the Russian verb of change was extracted (for the more detailed discussion see section 4).

For the preliminary clustering experiments with Swedish verbs, I had to collect and annotate a dataset since there were not enough examples with UD-markup presented in Treebank[8]. I used three small corpora from Leipzig Corpora Collection[9]: news, Wikipedia and web. I extracted all the sentences that contained verbs *förändra, ändra, byta, växla* knowing that they make up the field of changing in Swedish (Section 3). All the examples were preprocessed with the same pipeline as the main corpus and converted into a dataset with 5221 examples. Figure 4 presents the distribution of examples.

[8] https://github.com/UniversalDependencies/UD_Swedish-Talbanken
[9] https://wortschatz.uni-leipzig.de/en/download/Swedish

Figure 4. The distribution of verbs in the ‘change’ dataset for Swedish



I also used the small set of 34 stimuli from the questionnaire that includes all the oppositions that had been checked in the traditional field study. I manually translated them in Russian and automatically translated in Swedish using the same tools as for the first dataset.

4.2. Classification task for model selection

The dataset was split into the train (70%) and test (30%) subsets. I considered various ways of vectorisation of target sentences.

1) TF-IDF

I vectorised all the sentences utilizing a built-in sklearn TF-IDF vectoriser (Pedregosa 2011) as a baseline. I must emphasize that this type of vectorisation model overtly encodes the verb that is used in the sentence.

2) BERT [CLS] token embeddings

Given the fact that BERT is one of the most studied models in terms of context-based embeddings, I also vectorised all the sentences from the dataset using a pre-trained multilingual cased BERT base. The idea behind this type of vectorisation is related to the definition of the frame. While differentiating between frames, one should focus not on the verbs but on the contexts since only situations and not lexical entities are universal cross-linguistically. Thus if the model represents the difference between the sentences I may assume that they belong to different types of situations. Although the model receives information about target verbs it will not be clearly presented in the final vector.

3) BERT [MASK] token embeddings

The second basic task for BERT is [MASK]-token prediction. For all of the sentences from the dataset, target verbs were replaced by [MASK] token and then predicted embeddings were extracted. For this experiment, I also used a pre-trained multilingual cased BERT base. The idea of creating a vector prediction based on a particular context also corresponds with the LexTyp approach. I assume that predictions of verbs for similar contexts will be similar. In this experiment, the model does not have any information about the verb but creates embedding vectors based on the context.

4) BERT SVO and SO vocabulary vectors

While collecting the data I also extracted subjects, target verbs and objects for each sentence. The assumption is that the linear combination of non-contextual embeddings of verbs and their core arguments can be a good representation of frames. As I have already discussed in section 3 distinction between frames is often based on the features of the verb's dependents, while the rest of the context may not be as important. Using extracted lexemes I combined them in two different ways: I calculated a sum of subject, verb and object embeddings and only subject plus object embedding. The first combination includes overt information about the predicate, whilst the second one does not.

5) LaBSE

Working with multilingual embeddings is problematic in terms of their representation within one model. The model shows good performance on English embeddings provided with good training data and fine-tuning, whilst for other languages, the quality is quite poor (Libovický 2019). Since I am working with languages that are not closely related, embeddings must be well-aligned and based on a significant number of training sentences. Therefore, I decided to use a novel model presented by Google AI (Feng et al. 2020). LaBSE (Language-Agnostic BERT Sentence Embedding) combines the masked language model (MLM) and the translation language model (TLM) to improve the quality of language representation. As I have already mentioned, sentence-embeddings seem to be a good representation of frames.

6) XLM

XLM (Lample, Conneau 2019) uses three training objectives:

Masked Language Modelling – the same objective sa multilingual BERT exploits, but XLM uses the sequences of the constant length and not the pairs of sentences for masking;

Causal Language Modelling – prediction of the next word based on the given one;

Translation Language Modelling.

According to Choenni, Shutova (2020) XLM is quite good at capturing typological information, which is useful for our research. I used mean pooling to calculate the sentence embedding of this model.

Before focusing on one semantic field I decided to estimate the ability of word and sentence embeddings to distinguish different semantic fields based on various predicates. According to our assumptions, contextual word and sentence embeddings should have been able to represent the differences between the predicates from different semantic fields (like *prjatat'* and *menjat'*) better than between the predicates of the same field (like *prjatat'* и *skryvat'*). However, both kinds of distinctions should be good enough for us to continue performing experiments on predicates within one field.

I trained a logistic regression classifier on the extracted vectors that were created to estimate the performance of the model in the distinction between semantic fields. The model had to differentiate between 5 classes, namely *prjatat', iskat', naxodit', menjat', skryvat'.* Due to imbalanced data, I chose to report the micro f-score as a metric of quality.

Table 1. Performance of the logistic regression multi-class classifier on different vectors

Vectorizer	micro f-score
TF-IDF Baseline	0.96
BERT CLS	0.78
BERT MASK	0.68
BERT SVO-vocab	1

BERT SO-vocab	0.69
LaBSE	**0.93**
XLM	**0.97**

As can be seen from the table, TF-IDF and SVO vocabulary-based embeddings showed the best performance on the test set. I can assume that for the distinction of one field only lexical information is required, therefore simple models based on word co-occurrence are more efficient. However, if one focuses only on the distinctions within one semantic field the lexical information will not be helpful. For instance, in languages with a dominant system, where only one verb covers all the frames, the SVO approach will fail.

Since the research is devoted to the difference between the frames and not the lexical entries, I suggest running the clustering on the examples within one field on more complex representation. I decided to focus on the predicates of the field of change and use two of the most successful models: LaBSE and XLM.

Figure 5. LaBSE embeddings of verbs *prjatat', iskat', naxodit', menjat', skryvat'*

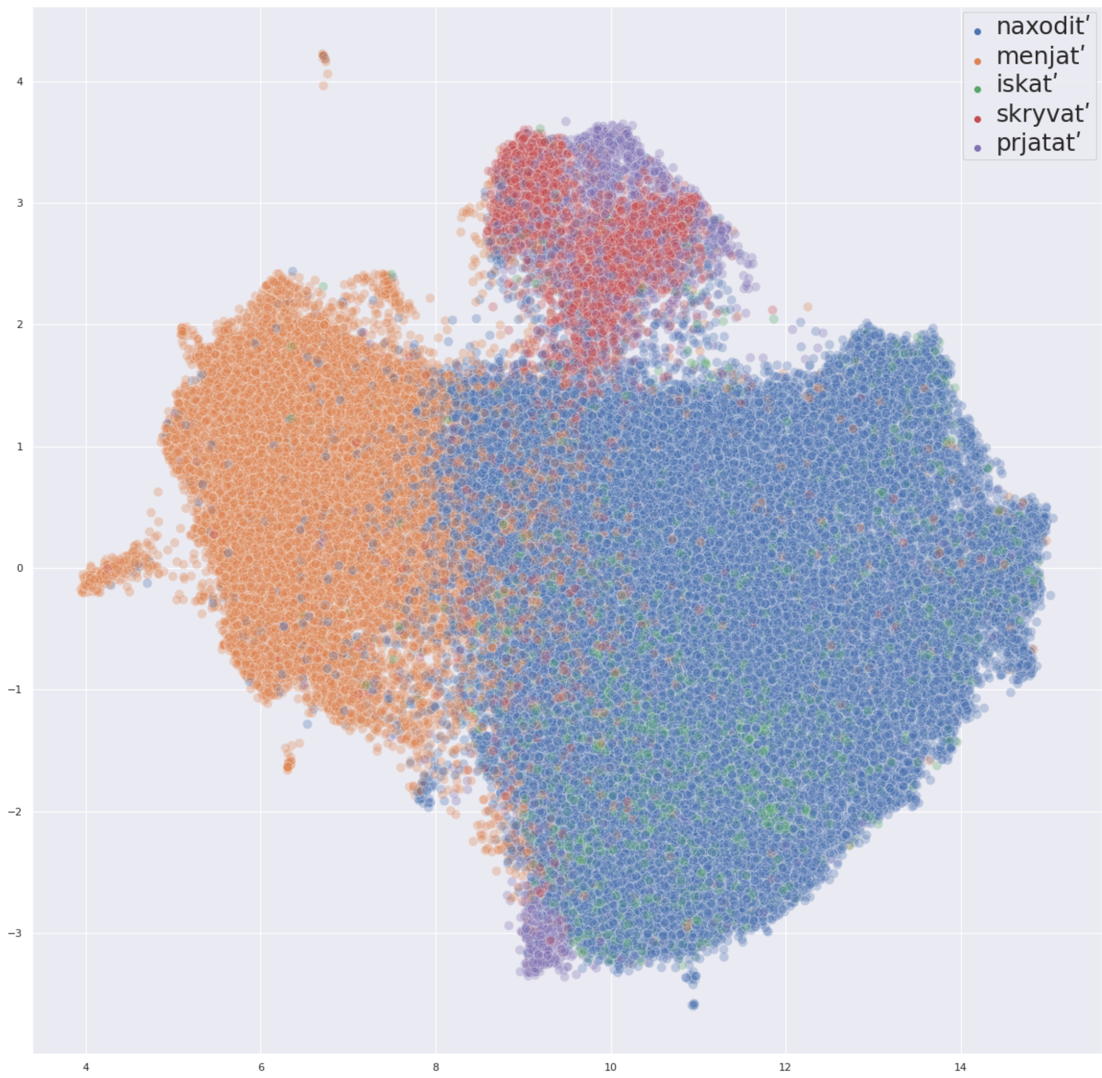
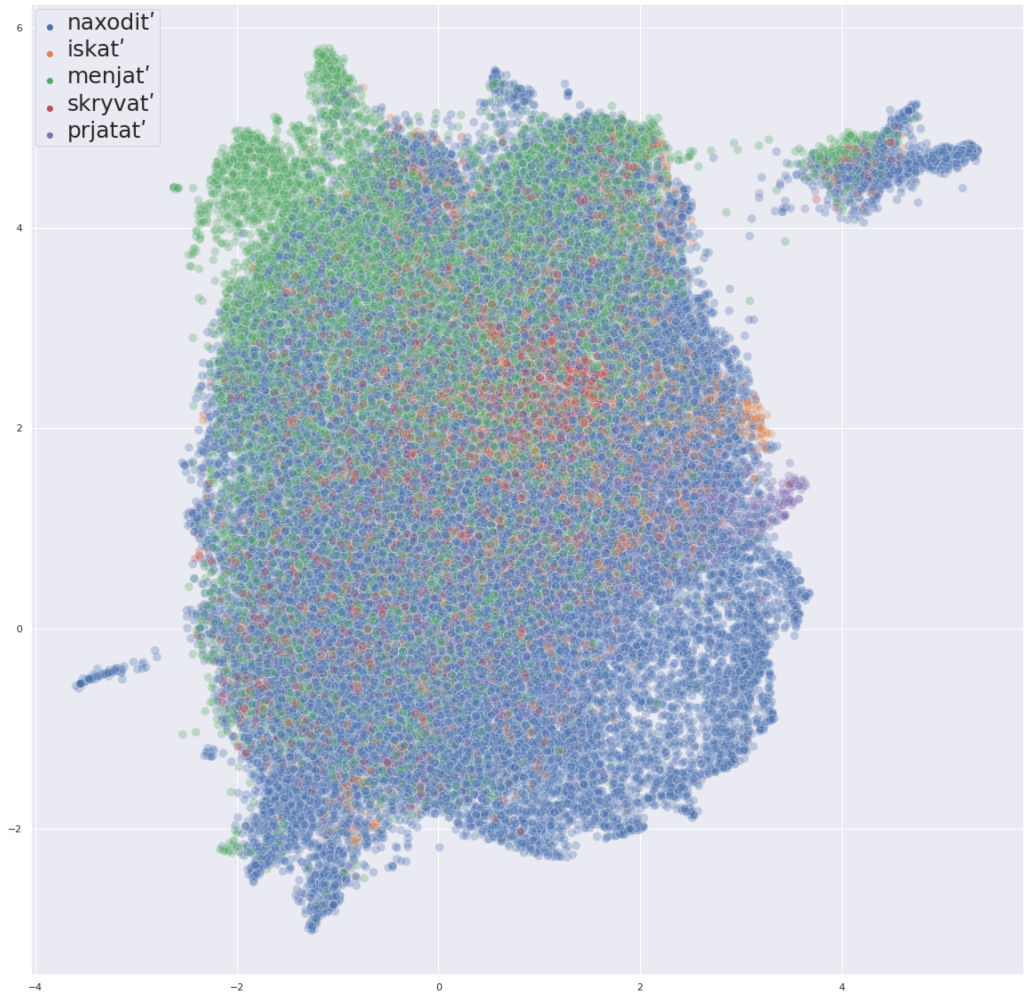


Figure 6. XLM embeddings of verbs *prjatat', iskat', naxodit', menjat', skryvat'*



Scatterplots depict the sentence embeddings built after the dimension reduction by the UMAP algorithm.[10] As can be seen from the visualisation, LaBSE is better at dealing with verb separation, henceforth I used LaBSE embeddings for the further investigation.

4.3. Clustering

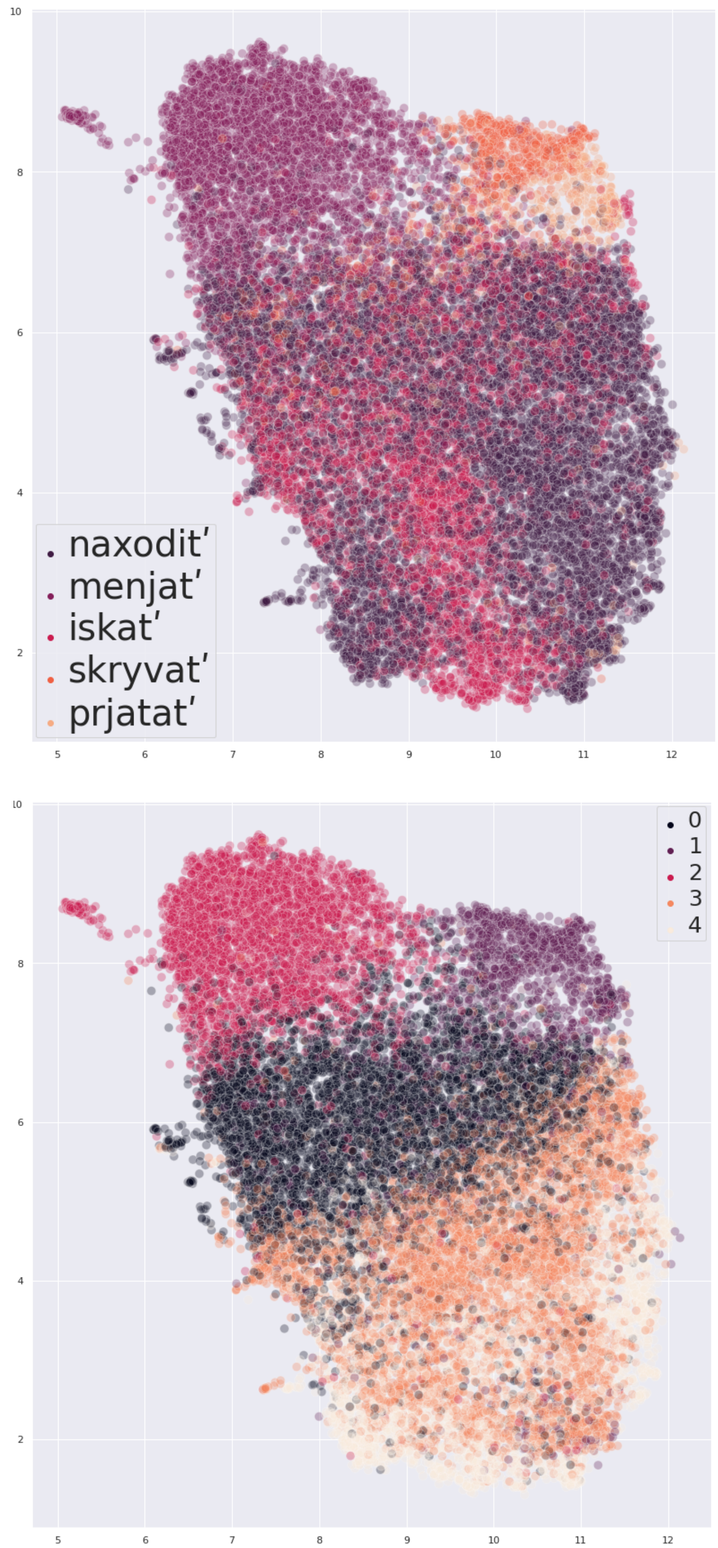
The idea of clustering is to find patterns in the unlabelled data. If a vector representation can capture the semantics then one could expect that the sentences belonging to one family of frames would be located close to each other in vector space, while the ones that describe different situations will be quite far. One important implication is that if the multilingual embeddings are language-agnostic unlike in mBERT (Libovický et. al. 2019) the clusters for Swedish and Russian sentences should be similar. Referring back to our

[10] While applying the tests I also used t-SNE reduction, but since the results did not show the difference, I decided to use UMAP as the least resource consuming algorithm.

theoretical research we expect clustering to distinguish the situational nodes of the field like exchange or alteration (Figure 1), regardless of the verb that is used in the sentence.

As the next stage of the research, I have performed clustering of the LaBSE embeddings that showed good performance on the test classification task and allowed to separate the verbs on the visualisation.

Figure 7. Word embeddings (1) VS K-Means clustering (2) of verbs *prjatat', iskat', naxodit', menjat', skryvat'*



Firstly, I decided to estimate the performance of the clustering algorithm on the whole corpora. In the case of a linguistic investigation, traditional metrics like silhouette score are uninformative since they show the formal quality of grouping, while for me it is more important that they should correspond to linguistic reality. As long as my data is not annotated with frames the estimation will be rather qualitative than quantitative.

As it can be seen from Figure 7 the K-Means algorithm divided the data into 5 clusters which correspond to their semantics. The verbs of hiding (*prjatat', skryvat'*) represent one cluster. The verbs of *change* (*menjat'*) also represent one cluster, whilst for the verbs of *seeking* (*iskat', naxodit'*) the algorithm splitted the examples in three groups, probably because this field was presented by the biggest number of examples (Figure 3).

After that, I focused on the verbs of changing. For now, there is no commonly accepted methodology for frame semantic clustering, therefore I propose to test several clustering algorithms.

I decided to employ three quite often used clustering methods:

1. K-Means

The main idea of the algorithm is to minimize the sum of distances from the points in each cluster to the cluster centroid. The number of clusters is pre-defined.

2. DBSCAN

DBSCAN clustering is based on the density of objects in the vector space. For every iteration, the algorithm checks if there is a sufficient number of points *N* in the environment *e*. Both *N* and *e* are hyperparameters. Hyperparameters were tuned using the k-distance elbow-plot method for silhouette score metric.

3. Agglomerative Clustering

It is the bottom-up algorithm, i.e. it considers every item a separate cluster and then for every iteration merges two the most similar ones in a bigger cluster until all the items become one cluster.

For all of them, cosine distance was chosen as a distance function since it is believed that for NLP tasks it is the most efficient metric.

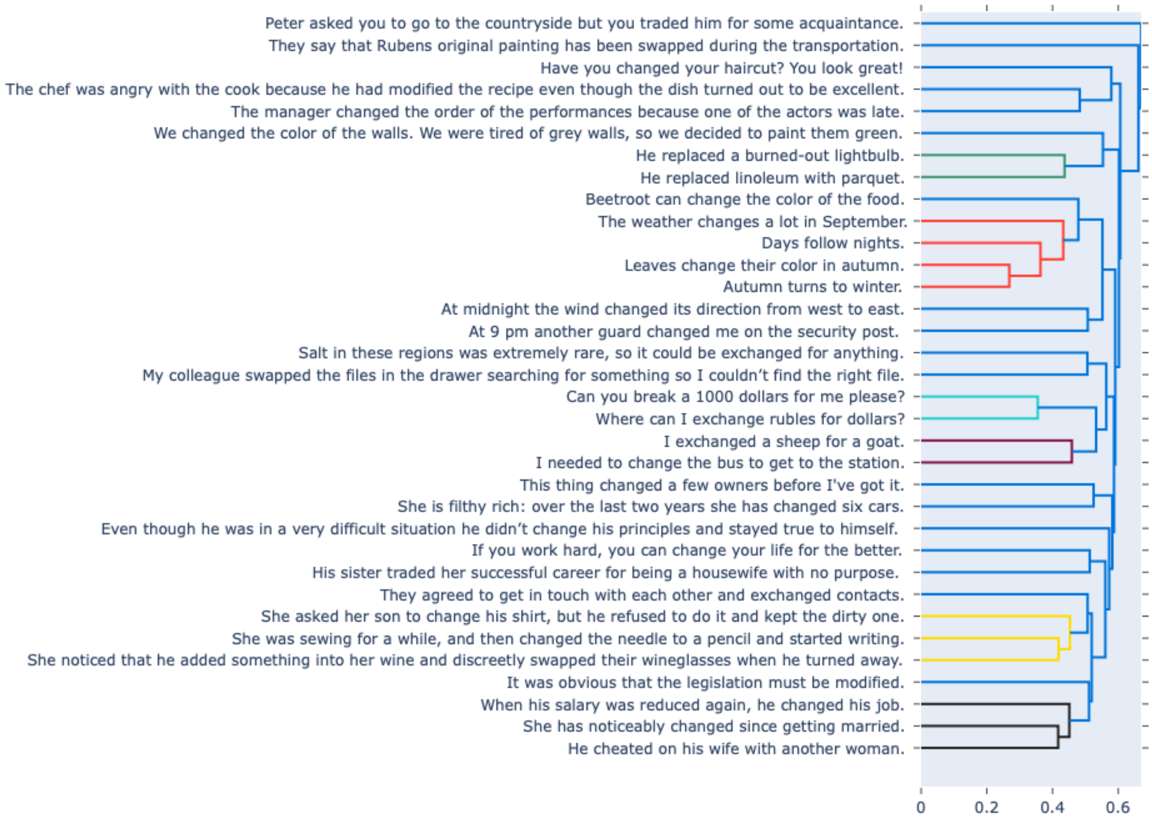
4.3.1. Russian

I ran K-Means clustering on the sentences with changing verbs with k from 2 to 5, but the algorithm was not able to divide the data into interpretable groups. Objects were divided into even uninterpretable sectors (appendix 2), hence the data appeared to be too complex for this method of clustering.

DBSCAN did not appear to be helpful either. With the hyperparameters chosen based on the elbow graph the maximum silhouette score that has been achieved is 0,28. Clusters were assigned almost randomly.

Since Agglomerative Clustering is the most time-consuming algorithm of all that I have used, I decided to run it on the questionnaire examples first. The results appeared to be quite promising. As can be seen from Figure 8 sentences about nature and money are united, moreover, sentences about a change of physical objects are also quite close to each other as one could expect from their semantics. Importantly, the sentences are not clustered by verbs.

Figure 8. Agglomerative Clustering of a questionnaire for the field of change in Russian



However, for the whole 'change' dataset clusters were identified by the model based on the topic of the sentence and not on the frame. For instance, one of the clusters was formed by the sentences related to the law and court. The result is quite expected since I used sentence embeddings that encode topic information better than the type of situation denoted by the verb. The challenges that the algorithms faced also tell us that the frames from the questionnaire cannot be defined by the simple clustering heuristics and require the development of another approach.

4.3.2. Swedish

Before transferring Russian sentences into Swedish I decided to conduct some preliminary experiments for Swedish. I carried out clustering on the small collected and automatically annotated dataset (section 2).

All the algorithms showed the same performance on the Swedish data: K-Means divided data points into even uninterpretable groups (appendix 3); DBSCAN performed with quite a low silhouette score, separating small pieces of data. Agglomerative clustering divided data into thematic groups.

4.4. Mapping the spaces

The next step for typological research is to expand the analysis from one language to the other. I decided to use a machine translation tool to transfer the data from the vector space of Russian language to the one of Swedish.

4.4.1. Translation

For translation, I used a pre-trained Helsinki NLP opus mt-ru-sv model from the Hugging Face library. The model is based on transformer alignment and is trained on the Tatoeba challenge dataset. Using fine-tuned mBERT alignment (Dou, Neubig 2021), I extracted Swedish verbs that corresponded to Russian verbs with the root *men*.

In many of the contexts, the sentence was paraphrased so that there were no verb but an adjective or participle to describe changing, in some of the sentences the verb without ‘change’ meaning was exploited. In total there were 5971 examples where the verb could not be extracted due to the discussed reasons. Table 2 presents 10 most frequent verbs in Russian and in Swedish that represent the field under study. For Swedish verbs, the percentage was calculated from all the examples where the verb could be extracted.

Table 2. The most frequent verbs representing the field of change in Russian and Swedish

№	Russian verb	Number of occurrences	Percentage of occurrences	Swedish verb	Number of occurrences	Percentage of occurrences
1.	*menjat'*	6629	33.9	*byta*	5185	38.3
2.	*izmenjat'*	5650	28.9	*ändra*	2202	16.3
3.	*smenjat'*	3965	20.3	*förändra*	2038	15.1
4.	*zamenjat'*	2258	11.5	*ersätta*	1662	12.3
5.	*pomenjat'*	293	1.5	*ha*	1410	10.4
6.	*promenjat'*	185	0.9	*komma*	362	2.7
7.	*podmenjat'*	150	0.7	*göra*	100	0.7
8.	*obmenjat'*	126	0.6	*kunna*	90	0.7
9.	*peremenjat'*	122	0.6	*skola*	83	0.6
10.	*razmenjat'*	111	0.6	*ta*	44	0.3

Based on the data in Table 2, I concluded that both for Russian and for Swedish 4 the most frequent verbs are central verbs defining the category of changing. As one can see, there is a drastic change in percentage of occurrences for the lemmas after the 5th. For Swedish the pattern is slightly different because of the auxiliary *ha* (have), which appeared to be quite frequent since for the Present Perfect Tense constructions *ha* was extracted as a verb while lexical verb was parsed as a participle (participle and perfective forms are homonymous in Swedish).

I build up the visualisation for the contexts with four most frequent verbs.

Figure 9. Sentence-embeddings in Russian and Swedish

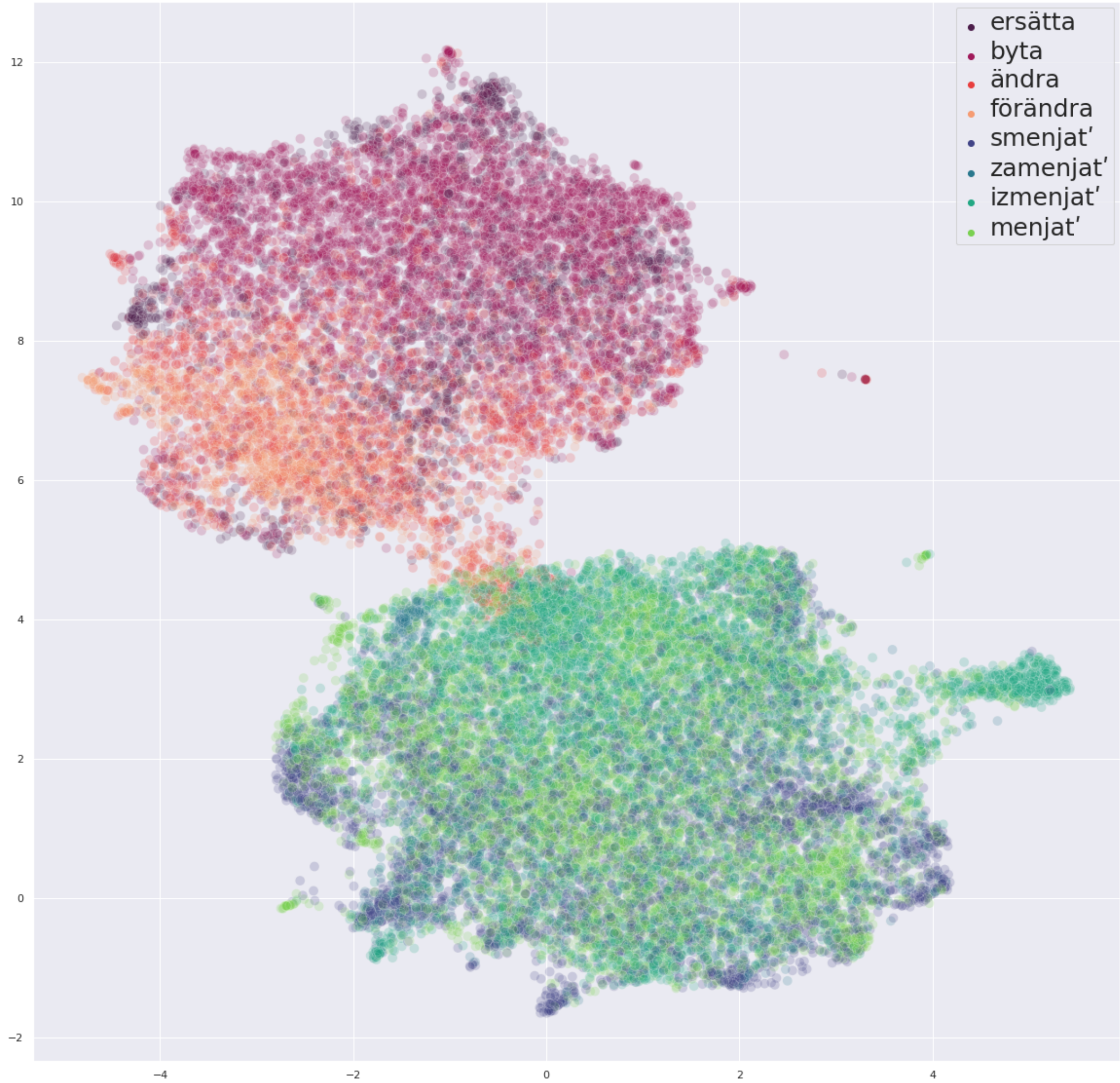


Figure 6 shows a joined vector space for Russian and Swedish examples with different verbs of the field 'change'. The distribution of sentence vectors in space corresponds to linguistic reality.

For Russian the verb *menjat'* (light green) is distributed across the whole space as a neutral imperfective which is what I expected after the theoretical research. *Smenjat'* and *izmenjat'* are also separated. Opposition between them is somewhat similar to the one between *byta* and *ändra* (section 3), however, they are not separated as Swedish verbs are.

In Swedish vector space the picture is also quite promising. Verbs *ändra* (peach) and *förändra* (red) are mixed in the bottom part of the cloud as we expected in section 3, due to the fact that they might be used in the same contexts but add different submeanings. The verb *byta* (purple) is clearly different from *ändra* and *förändra,* since *byta* is used mostly in the contexts with physical objects exchange, whilst the other verbs describe modification of any kind of object. We did not discuss the verb *ersätta* (dark-purple) before because it was only used once in the questionnaire and was not included in the field. *Ersätta* is derived from the verb *sätta* which means to place, which means that it originates from the fields of placing. *Ersätta* can be applied only to the situations of replacement of physical objects and embeddings capture this feature. As we can see, it is closer to the contexts of *byta* than to the contexts of *ändra.*

We can see that although according to Feng et al. (2020) the vectors should be language-agnostic, there is a noticeable consistent shift in the vector space. Sentences are visualized as the clouds of the same shape, with linear transformation. It means that there is language-specific information in the embeddings and LaBSE is not able to liquidate the language bias.

Earlier we assumed that the embeddings are able to represent semantics if they are clustered in the same way in Russian and in Swedish. Thus, I translated and clustered the questionnaire to compare it with the clustering from section 4.3.1., the results can be seen in Figure 10.

Figure 10. Agglomerative Clustering of questionnaire for the field of change in Swedish

In this tree the contexts about the weather change and money exchange form clusters, as it was in Russain. To formally calculate the similarity of clustering I used the Adjusted Rand Index (ARI). The metric considers all pairs of samples and counts the amount of objects that were assigned one cluster in two different clusterings. The metric takes values from -1 to 1, showing the degree of similarity of the cluster assignment.

For the questionnaire the ARI was 0.327, which shows that the clusters were not assigned randomly.

Clustering of the whole translated dataset we got the same results as before. Sentences were divided based on their topics and not the frames the verb describes. For this clustering resulting ARI was 0.05, which means that there is almost no correlation between two clusterings.

I also applied K-Means clustering with 9 clusters as it was the number of frames that I got employing the traditional approach. The resulting ARI was 0.334, which is much higher than for the Agglomerative Clustering. Better performance of K-Means can be explained by the fact that, as we have seen before, it divides data points in similar uninterpretable sectors. Geometrically the vectors of Russian and Swedish sentences form similar clusters, that is why they are splitted similarly. This type of clustering is not applicable for complex linguistic data we have.

4.4.3. Comparison with the traditional approach

Using computational methods I obtained different results compared to traditional typological research. For the Russian language, the analysis of the corpora allowed us to get more variability in the verbs for the field of changing. 12 unique lemmas were extracted by regular expressions (*menjat', izmenjat', zamenjat', vzaimozamenjat', namenjat', obmenjat', peremenjat', podmenjat', pomenjat', promenjat', razmenjat', smenjat'*), while the native speakers suggested only 8 of them (*menjat', izmenjat', zamenjat', smenjat', pomenjat', obmenjat', razmenjat', podmenjat'*). On the one hand, the research has better recall for the verbs representing the field. On the other hand, one can always limit the investigation by using only the most frequent verbs as I did in this study.

The word *pomenjat'* appeared to be much less frequent in the corpus data than I expected based on the analysis of the questionnaire. I assumed that *pomenjat'* is quite neutral and can be used in almost any context of changing but more specialized verbs were preferable.

Quite many sentences were translated with a different verb or syntactic structure which made the verb extraction not applicable, however, it was quite expected since while answering the questionnaire native speakers also tend to translate the target verb substituting it with different constructions.

For Swedish, automatization of the analysis also resulted in a different set of verbs. The verb *ersätta* was added, whilst the verb *växla* was used in translation only once. In most of the sentences it was replaced by the verb *byta*, which is quite predictable, given that it is the most common and neutral verb. This case is quite interesting, hence computational methods expanded the field with the lexeme from the joined semantic field, but at the same time did not take low-frequent verbs into account.

5. Conclusions

To sum up, I conducted and compared two parallel studies investigating the semantic field of changing in Russian and Swedish.

For the traditional field linguistics approach I introduced the questionnaire and suggested 9 frames that describe the verb oppositions in languages under study. Based on the analysis I visualised the semantic field as a frame map.

I introduced 2 datasets that can be used for future investigation. The first dataset consists of Russian sentences and subject-verb-object groups extracted for the verbs of changing, hiding and seeking. The second dataset consists of the Russian examples for the verbs with the root *men* (change) and their automatic translations. It also includes verbs of change that were automatically extracted from the Swedish sentences with the help of the alignment tool.

Based on the experiments I carried out, it can be concluded that the language models based on transformer architecture are not as language-agnostic as we want them to be. They manage to differentiate between completely different situations (different semantic fields) and between different verbs but they fail to represent the semantics of similar units similarly for different languages (i.e. there still is a language-specific component in the embeddings). With the methods I used it was not possible to cluster sentences of the one frame in Russian and Swedish similarly. To summarize, the contextual language models manage to capture the opposition between the contexts with different verbs and as expected their distribution is different in two languages but they do not analyse the semantic space in the same way the linguists do since the universal frame structure could not be extracted.

Despite the fact that transformer models do not represent the semantic field as we could expect from the theoretical study, they can still be employed in the initial stages of the research. As I have said in section 4, I managed to extract almost all the lexemes that were used by native speakers in the questionnaire and for both languages even more relevant verbs were extracted. Thus, employing this approach one can improve the recall of the typological study.

This topic requires future development. As a possible direction of the research I suggest annotating the data with the frames and fine-tuning the transformer model on the triplet loss. This will make the model represent the sentences of the same frame similarly and distinguish them from the sentences of the other frames. Moreover, it seems that the study can benefit from the model trained on a special training objective. For instance, it might be the predicate prediction based on its core arguments. Finally, for better results another dataset should be used. As I mentioned before, the data in Taiga corpora appeared to be not annotated properly.

References

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. Univ of California Press.

Choenni, R., & Shutova, E. (2020). What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *arXiv preprint arXiv:2009.12862*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In *Lrec* (Vol. 6, pp. 449-454).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dou, Z. Y., & Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *The new psychology of language*, *2*(1976), 1-30.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Janda, L. A., Endresen, A., Kuznetsova, J., Lyashevskaya, O., Makarova, A., Nesset, T., & Sokolova, S. (2013). *Why Russian aspectual prefixes aren't empty: prefixes as verb classifiers*. Slavica Publishers.

Kjuseva, M. V., Reznikova, T. I., & Ryžova, D. A. (2013). Tipologičeskaja baza dannyx ad″ektivnoj leksiki. In Komp′juternaja lingvistika i intellektual′nye texnologii: po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» (Bekasovo, 29 maja.)Koch, P. (2008). *85. Lexical typology from a cognitive and linguistic point of view* (pp. 1142-1178). De Gruyter Mouton.

Koch, P. (2008). *85. Lexical typology from a cognitive and linguistic point of view* (pp. 1142-1178). De Gruyter Mouton.

Kulshreshtha, S., Redondo-García, J. L., & Chang, C. Y. (2020). Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study. *arXiv preprint arXiv:2009.14304*.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Libovický, J., Rosa, R., & Fraser, A. (2019). How language-neutral is multilingual BERT?. *arXiv preprint arXiv:1911.03310*.

Orexov, B. V., & Reznikova, T. I. (2015). Komp′juternye perspektivy leksiko-tipologičeskix issledovanij. Vestnik Voronežskogo gosudarstvennogo universiteta. Serija: Lingvistika i mežkul′turnaja kommunikacija, (3).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pineda, F. J. (1987). Generalization of back-propagation to recurrent neural networks. *Physical review letters*, *59*(19), 2229.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

RAXILINA, E. V., & REZNIKOVA, T. I. (2013). Frejmovyj podxod k leksičeskoj tipologii. Voprosy jazykoznanija, (2), 3-31.

Rabinovich, E., Xu, Y., & Stevenson, S. (2020). The Typology of Polysemy: A Multilingual Distributional Framework. *arXiv preprint arXiv:2006.01966*.

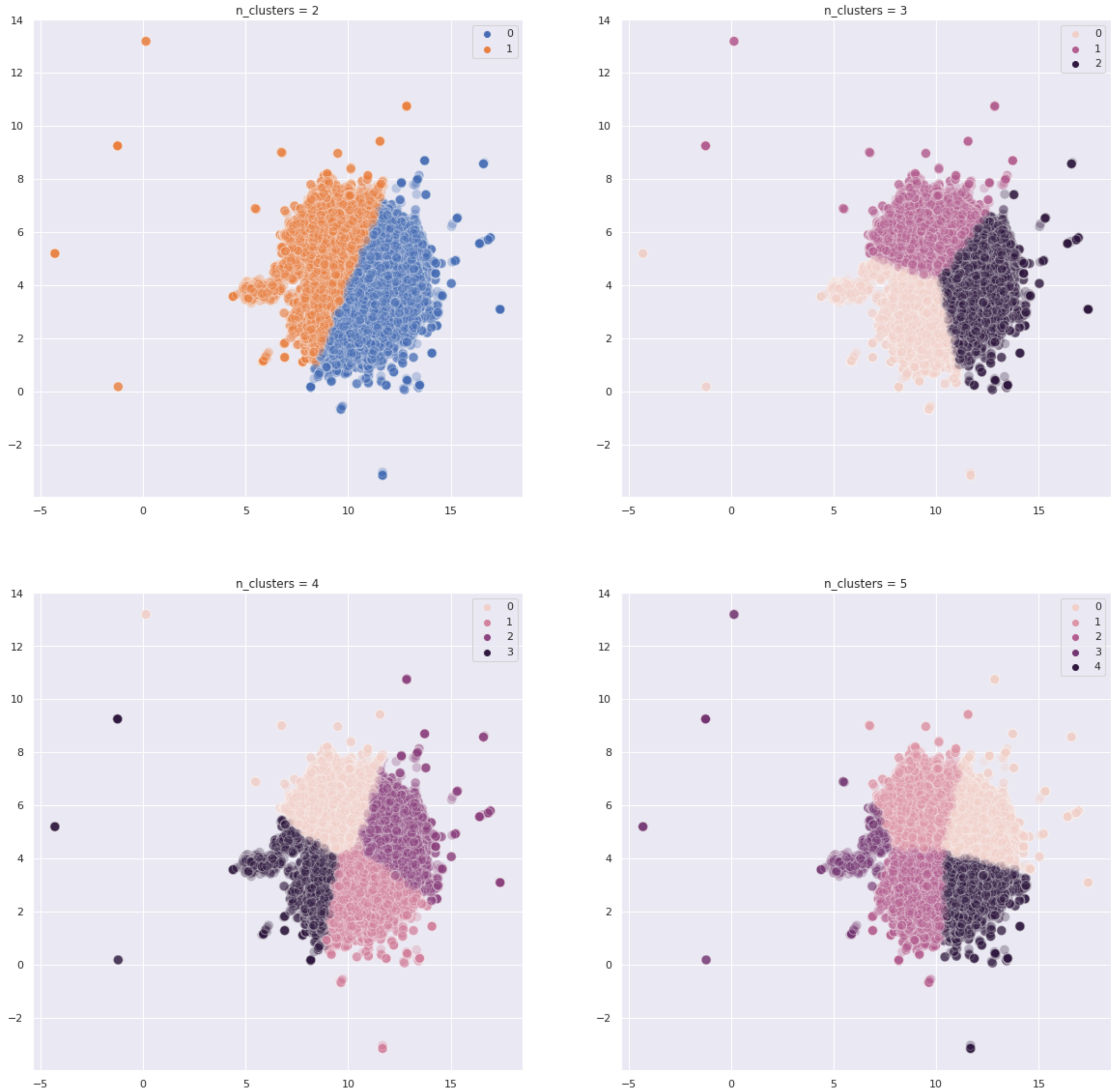
Rakhilina, E., & Reznikova, T. (2016). A frame-based methodology for lexical typology. *The lexical typology of semantic shifts*, *95*, 129.

Ryžova, D. A. (2020) Tipologija leksiki. Komp′juternye metody i instrumenty. SPb.: ALETEJJa, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Appendix

1. Source code: https://github.com/aaaksenova/term_paper2021
2. K-Means in Russian



3. K-Means in Swedish