Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Computer-assisted detection of typologically relevant semantic shifts in world languages*

Ilya Gruntov
Institute of Linguistics
Bol.Kislovsky per. 1/12,
Moscow, 125009, Russia
altaica@narod.ru

Elisei Rykov
HSE University
20 Myasnitskaya ulitsa,
Moscow, 101000, Russia
esrykov@edu.hse.ru

Abstract

The paper contains the description of a semi-authomatic method for the detection of typologically relevant semantic shifts in the world's languages. The algorithm extracts colexified pairs of meanings from polysemous words in digitised bilingual dictionaries. A machine learning classifier helps to separate those semantic shifts that are relevant to the lexical typology. Clustering is applied to group similar pairs of meanings into semantic shifts.

Keywords: Lexical typology, semantic shifts, NLP, computational semantics, polysemy
DOI: 10.28995/2075-7182-2023-22-161-171

Автоматический поиск типологически релевантных семантических переходов в языках мира

Илья Грунтов
Институт языкознания РАН
Россия, г. Москва, 125009
Бол. Кисловский пер. 1/12
altaica@narod.ru

Елисей Рыков
Национальный исследовательский
университет «Высшая
школа экономики»
Россия, г. Москва, 101000
ул. Мясницкая, д. 20
esrykov@edu.hse.ru

Аннотация

Статья описывает полуавтоматический метод выявления типологически важных семантических переходов в языках мира. Алгоритм извлекает пары значений многозначных слов, встречающихся в двуязычных словарях. Машинно обученный классификатор позволяет определить степень релевантности перехода для лексической типологии. Пары значений объединяются в семантические переходы посредством кластеризации.

Ключевые слова: Лексическая типология, семантические переходы, АОТ, полисемия, вычислительная семантика

1 Introduction

The typology of semantic changes in the languages of the world is one of the areas of linguistics that can greatly benefit from computational methods. The current work is based on the datasets collected within the framework of the Database of Semantic Shifts in the Languages of the World (DSS, https://datsemshift.ru) which was founded in 2002. The main concept of the project has been described in (Zalizniak et al., 2012). The semantic shift is defined as a cognitive proximity between two meanings which are represented in different ways in the languages of the world, e.g. a relation between two meanings of a polysemic word, etymological cognates in related languages, semantic evolution at

[*] The paper is supported by the grant of Russian Science Foundation №22-18-00586, https://rscf.ru/project/22-18-00586/

different historical stages of a language, relations between the meaning of a loanword and a source word in two languages, the meanings of two morphological derivatives from a single root etc. See a more complete list in (Zalizniak, 2018). In the present paper, however, we restrict our subject to a specific subset of these possible types of semantic shifts, namely with a relation between two different meanings that can be represented by a colexification within a polysemic word of a given language. Such pairs of meanings are here called **realisations** of a semantic shift. For example, in at least 275 languages there are different realisations of the semantic shift "moon" - "month" [1]. Cf. also Table 1

Language	**Entry**	**Meaning 1**	**Meaning 2**
Ancient Hebrew	ḥōdeš	new moon ḥōdeš māḥār ’tomorrow is the new moon’	month ḥōdeš bəšānā ‘one month in the year’
Swahili	mwezi	moon mwezi uliliwa na joka ’lunar eclipse’	month kila mwezi ’every month’
Adyghe	maze	moon мазэр къыкъокIыгъ ’the moon rose’	month январь мазэм ’in the month of January’
Tibetan	zla-ba	moon	month

Table 1: Several realisations of the semantic shift "moon" - "month" out of 275 present in the DSS

The purpose of this paper is to describe our attempt to automatically detect those semantic shifts that can be reproduced in different languages independently or through borrowing or loan translation. This may deepen our understanding of the cognitive mechanisms that give rise to polysemy. We have created a dataset of pairs of meanings representing a semantic shift extracted from the polysemous words of 75 languages, developed a machine learning classifier that defines the degree of typological relevance of a given pair of meanings, and performed clustering to group similar pairs of meanings from different languages. The result of our work is a pool of potential semantic shifts that can be further evaluated by a team of experts who can then incorporate them into the DSS.

The main source for our work are bilingual dictionaries "Language X -> Language Y". Usually language Y is one of the major world languages, such as English, Spanish, French, Russian etc. For these high-resource languages there are sophisticated language models, lexical databases of semantic relations and other linguistic tools. Thus, these languages are in fact meta-languages for describing meanings in the source Language X, and we can make cross-linguistic comparison of these data by applying NLP methods to these meta-languages. This approach therefore allows data from both low and high resource languages to be included in the common framework.

2 Methods and related works

There are many methods for detecting semantic shifts. One of them is the study of historical word corpora which contain documents from several hundred years, so that we can track changes in word meanings within a language over an extended time period. See e.g. (Hamilton et al., 2016), (Rodina and Kutuzov, 2020), (Fomin et al., 2019), (Kutuzov and Kuzmenko, 2017)), (Kutuzov et al., 2018) etc. Corpora analysis allows to find out that a given word has different meanings, but it is a highly difficult task to define strictly these meanings.

Another approach implies extraction data from etymological dictionaries and databases which usually contain cognate sets of diachronically related words with often different meanings in the modern languages. The etymological cognates are clear cases of semantic shift, but the main problem here is

[1] https://datsemshift.ru/shift0856

the robustness and reliability of the etymology. The shallower the language family, the more reliable the etymologies, but at deeper levels it becomes much more difficult to prove the validity of cognate relationships. See, e.g. the example of such diachronic studies in (Федотова, И.В., 2020).

Another approach detects the semantic changes arising within morphological derivation. See e.g. such works on Ukrainian (Melymuka et al., 2017) or Czech (Musil et al., 2019).

However, the present paper is concerned only with a particular subclass of semantic shifts, namely those that can be extracted from a dictionary as a separate pair of meanings of a polysemous word. The closest project to this task, apart from the DSS mentioned above, is another large project devoted to the aggregation of colexifications from various languages of the world called CLICKS (The Database of Cross-Linguistic Colexifications), the main principles of which are described in (Rzymski et al., 2020). CLICKS extracts meanings from the word-lists and maps them to concepts from the Concepticon catalogue (List et al., 2019) by semi-automatic fuzzy search.

Although we are aiming for a somewhat similar result, namely a database of semantic shifts, we took rather different approach to selecting and comparing meanings. According to our approach, not all pairs of meanings presented in the dictionaries should be included in the final database, but only those that are "non-trivial", "non-automatic" and thus most relevant for the typology of semantic changes.

There are certain criteria we use to define the degree of relevance of a semantic shift:

- The difference between meanings within a pair of meanings extracted from the dictionary should be "substantial". Of course, when one deals with semantic change, there is often a continuum between completely different and slightly different meanings. The degree of difference can vary greatly depending on the cultural and linguistic context of a given language. If a word ceased to mean ’yellow’ and began to mean ’white’, this would obviously be a major change, but if the meaning of a word changed from ’yellow’ to ’light-yellow, pale’, this would hardly be considered a major change by anyone. So it is not a distinct boundary between two classes of ’major change’ vs. ’minor change’, but a scale.
- The pair of meanings should not be related by regular metonymy. For example, "content" vs. "container", "author" vs. "author’s work", etc.
- Neither meaning should be too vague.
- Differences in syntactic actant structure are not sufficient to turn a pair of meanings into a realisation of a suitable semantic shift. For example, we would not consider the semantic changes between *’to drink’ (transitive) - ’to drink’ (intransitive)* as a valid semantic shift (as in examples like: *He drinks water* vs *he has quit drinking*).
- The meanings within a pair should be connected immediately, i.e. if there is a minimal semantic shift between *foot* ’body part’ and *foot* ’the lower part of the hill’ and on the other hand there is another minimal semantic shift between *foot* ’body part’ and *foot* ’unit of length’, we would not postulate a semantic shift between the meanings *foot* ’lower part of the hill’ and *foot* ’unit of length’.

To make these preliminary criteria more formal and quantifiable we use a number of factors to train the ML classifier to separate suitable semantic shifts from the others. The factors and the classifier itself are described in the "Machine learning classifier" section below.

3 Materials

We used data from 75 digitised dictionaries of the following languages:

- **Altaic:** Azeri, Buriad, Crimean Tatar, Dagur, Japanese, Khamnigan, Khalkha Mongolian, Korean, Nogai, Soyot, Tatar, Turkish
- **Austroasiatic:** Vietnamese
- **Austronesian:** Indonesian, Tagalog
- **Bantu:** Swahili, Zulu
- **Chukotko-Kamchatkan:** Koryak
- **Cushitic:** Somali
- **Indo-European:** Afrikaans, Ancient Greek, Armenian, Belarusian, Bulgarian, Czech, Danish, French, German, Greek, English, Icelandic, Italian, Kurdish, Latin, Latvian, Lithuanian, Norwe-

gian, Old English, Ossetian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Spanish, Swedish, Tajik, Tat, Ukrainian, Welsh
- **North-East Caucasian:** Avar, Bagvalal, Botlikh, Chamalal, Chechen, Lezgin, Rutul, Tabasaran
- **North-West Caucasian:** Adyghe, Abaza
- **Semitic:** Arabic, Hebrew
- **Sino-Tibetan:** Chinese
- **Uralic:** Estonian, Finnish, Hungarian, Komi Zyrian, Mari, Mokshan, Nganasan, Selkup, Yukaghir
- Basque (isolate), Esperanto (artificial).

These dictionaries vary greatly in size and in the percentage of polysemous words they contain, yielding from 500 pairs of meanings in Soyot up to 56000 in Azeri. In total we extracted from all these dictionaries a dataset consisting of 1,300,000 pairs of meanings each of them colexified in a lexical entry.

4 Pipeline

Our pipeline for detecting typologically relevant semantic shifts looks as follows[2].

1. Extraction of the polysemy from the digitised dictionaries by parsing.
2. Vectorization of all meanings of polysemous words using an encoder
3. Processing the data with a machine learning classifier that filters out all trivial or otherwise irrelevant cases.
4. Clustering of the filtered realisations from different languages into groups. These groups correspond to semantic shifts.
5. Postprocessing of the data by linguists.

In the following we will look at each stage in more detail.

5 Parsing

In the first stage we parse digitised dictionaries via Python scripts. The minimal entity we work with is a pair of meanings colexified within a lexical entry. In the simplest case the polysemic word has only two meanings. Often, however, the polysemic word has more than two meanings. It is rather impractical and far from linguistic reality to include all the possible combinations of meanings in our dataset. In our approach, to avoid combinatorial explosion, we assume that each meaning of a polysemic word is derived from its main meaning, which is defined by the dictionary authors as the first meaning. We know that this is an oversimplification derived from the simple model of radial polysemy (all secondary meanings are derived from the main one), while the real polysemy patterns may be different, e.g. chain polysemy (when the third meaning is derived from the second, not from the first). Perhaps, in further studies we will be able to overcome this limitation.

We extract polysemy from the dictionaries relying on the description of the meanings provided by the authors of the corresponding dictionaries. Sometimes, the meanings of a stem diverge so drastically that the authors of a dictionary decide to put them in separate entries as homonyms. In our approach, we did not distinguish between true homonyms which are different words that happen to be identical in form and false homonyms, which are the result of the evolution of the same word. We included such words in our data with the label “homonym”. Another possible decision would be to completely exclude such words from our data, but we supposed that recall was more important here than accuracy, and we do not want to lose the information of these “false” homonyms. The shift in meaning which is so strong that it gives life to two separate words synchronously is of great interest for semantic typology.

The result of this stage is a tsv file in which each line contains a lexical entry, a pair of meanings taken from the dictionary, the language name and dictionary metadata

6 Machine learning classifier

After parsing we obtained over 1,300,000 colexified pairs of meanings from 75 languages. However, not all of them represents a relevant semantic shift. Therefore, pruning is required to filter out irrelevant cases according to the criteria described above in Section 2.

[2] The data and scripts are available at `https://github.com/lmeribal/semantic-shifts`

To compare different classification approaches, we used the F1-measure. To train the classifier, we randomly selected a sample of pairs of meanings extracted from polysemous words and annotated it with the help of a team of linguistic experts[3]. The annotation was performed for two classes: positive class, when the given polysemy represents a semantic shift and vice versa. We considered a pair of meanings as marked for a certain class if at least 3 experts put it into that class. All the experts were experienced in semantic typology and have been working with the Database of semantic shifts for a considerable amount of time. As instruction for the markup they used the criteria for detecting relevance of a semantic shift presented in Section 2. Annotators inter-rater agreement according to Krippendorff's Alpha was 0.46. In total, over 22700 judgements were obtained for 2500 pairs of meanings. 1700 pairs become our train set. Validation set and test set consisted of 375 pairs each. The annotated sample from the dataset is shown in Table 2

language	entry	Meaning 1	Meaning 2	mark
Latvian	jēls	'сырой, сырое мясо'	'разг. непристойный, сальный'	1
Zulu	isi-bindi	'печень'	'смелость, храбрость'	1
Ancient Greek	σκολιός	'кривой, изогнутый'	'лукавый, коварный'	1
Welsh	marchnadaeth	'ware(s), merchandise'	'trade, traffic, commerce, business'	0
Lithuanian	dambra	'дудка, свирель'	'губная гармоника'	0
Azeri	şıltaq	'каприз'	'привередник, привередница'	0

Table 2: Sample of the annotated polysemy from the dictionaries

6.1 Methods

To find out which pairs of meaning represent a valid semantic shift relevant to the lexical typology, we tried several methods.

Cosine measure: As a baseline for comparing different approaches, we chose a method based on cosine distance alone. If the cosine distance between the embeddings of the definitions is greater than 0.5, this pair of meanings qualifies as a valid realisation of a semantic shift, and vice versa. Our dictionaries were usually bilingual translation dictionaries, not explanatory dictionaries. They contained a word in the original language and its translation in English, Russian, Spanish, etc. Since we needed to compare the translations of the original words, we looked for a multilingual embedding model that could compare sentences and noun phrases from different languages in the same embedding space. Therefore, we chose a Multilingual Universal Sentence Encoder (MUSE) [4] to obtain embeddings from dictionary definitions.

Feature-based classifier: The main source for feature engineering and feature selection was the dictionary definitions of the extracted meanings. We extracted features such as the number of words in these definitions and their lengths, the normalised Levenshtein distance between them and the cosine distances between their MUSE embeddings, the presence of common hyperonyms, synonyms, part of speech tags for the syntactic heads of the definitions, and so on. For a more detailed list, see the Table 4. We then trained a gradient boosting classifier on the data for 500 iterations with a depth parameter of 3. The algorithm chosen was Catboost (Dorogush et al., 2018).

Frozen LM fine-tuning: Previous work such as (Radford et al., 2017) shows that Language Models (LMs) perfectly solve downstream tasks related to language understanding, even when learning simple tasks such as next word or character prediction. Therefore, we assumed that Language Models already had the necessary knowledge about languages and their encoder embeddings could be applied to our task. Since we work with different high-resource languages, multilingual LMs such as BERT, RoBERTA, mt0 (Muennighoff et al., 2022) and FLAN-T5 (Chung et al., 2022) were used for training. As the mt0 and FLAN-T5 models are complete transformers, only frozen encoders were extracted from these models. To

[3] Annotation was performed by Ilya Gruntov, Sofia Durneva, Idaliya Fedotova, Viktoria Kaprielova, Veronika Kondratieva, Tatiana Mikhailova, Maria Orlova, Maksim Rousseau, Elisey Rykov, Anna Smirnitskaya, and Anna Zalizniak

[4] `https://tfhub.dev/google/universal-sentence-encoder-multilingual/3`

extract the embedding of the whole meaning, we applied mean pooling to the token embeddings. For all models, output embeddings for two meanings were obtained independently. Later, we concatenated the embeddings of two meanings and applied a classification head with binary output to these concatenated embeddings. In total, our model contained only 2050 trainable parameters. We also explored pre-trained language models of different sizes. During the training, we optimised Cross-Entropy Loss using the AdamW optimizer. Since our data is unbalanced (about 70% of the samples belong to the negative class), we passed class weights into the loss object. Negative class weight was calculated as the ratio of number of positive class samples to total samples, and vice versa. For each model, we trained the classifier for 50 epochs with a learning rate of 1e-3. For validation, we used the version of the classifier from the epoch with the lowest test loss value.

Multilingual Universal Sentence Encoder: In addition, we trained a classifier with the method described above, using embeddings from MUSE.

6.2 Evaluation

The evaluation of the classifiers is shown in the Table 3. For each method we calculated precision, recall and F1. ROC-AUC was only calculated for methods that were able to return probabilities of classes, so we didn't calculate it for our baseline.

The method based on a cosine measure does not achieve a high metric values. Thus, a noticeable difference between the definitions of the meanings of a polysemous word does not necessarily make it a valid semantic shift. BERT, RoBERTa and FLAN-T5 show similar results to the feature-based method. Classifier that accepts embeddings from the mt0-small model shows the best result and outperforms other methods. In addition, the method using embeddings from the MUSE shows ROC-AUC comparable to larger sizes of mt0.

Method		**Precision**	**Recall**	**F1**	**ROC-AUC**
Cosine measure		0.40	0.41	0.40	-
Feature-based		0.59	0.58	0.56	0.67
Multilingual Universal Sentence Encoder		0.65	0.63	0.62	0.71
Frozen LM fine-tuning	bert-base-multilingual-cased	0.62	0.60	0.59	0.64
	xlm-roberta-base	0.65	0.64	0.63	0.69
	xlm-roberta-large	0.63	0.62	0.61	0.66
	flan-t5-small	0.58	0.57	0.56	0.61
	flan-t5-base	0.61	0.61	0.61	0.65
	flan-t5-large	0.61	0.59	0.59	0.65
	mt0-small	**0.68**	**0.67**	**0.67**	**0.74**
	mt0-base	0.67	0.65	0.65	0.71
	mt0-large	0.65	0.64	0.63	0.71

Table 3: Performance of different classification models. For the Frozen LM fine-tuning method, the name of the pre-trained model from the HuggingFace is shown.

For the trained CatBoost classifier we additionally extracted importance of the input features. This data is presented in the Table 4. Cosine distance makes the greatest contribution, as well as the normalised Levenshtein distance.

To infer a mark on unmarked pairs, we used a classifier that accepts embeddings from the mt0-small, which had the highest metrics values. After classification, 800,000 pairs were marked as unsuitable, and 530,000 pairs were valid.

Feature	Importance
Cosine distance between definitions	23.23
Normalised Levenshtein distance between definitions	15.68
Normalised Levenshtein distance between hyperonyms	13.77
Cosine distance between hyperonyms	13.29
Cosine distance between synonyms	10.57
Common hyperonyms between syntactic heads of the definitions	6.12
Part of speech of the syntactic head of the first definition	4.37
Part of speech of the syntactic head of the second definition	4.21
Normalised Levenshtein distance between synonyms	3.77
Common parts of speech	3.14
Common synonyms between syntactic heads of the definitions	1.85

Table 4: Feature importance from the trained CatBoost classifier

7 Clustering

Each semantic shift consists of similar realisations from different world languages. One of our tasks is to group realisations into semantic shifts, which is similar to the clustering task in machine learning. So each realisation from a particular language would be a separate point within a cluster, while the cluster itself would represent a semantic shift as an abstract entity. We can therefore obtain embeddings of the realisations, and then apply any clustering algorithm to these embeddings to obtain clusters of the new semantic shifts.

Apart from that, there is another task when we already have semantic shifts, and our goal is to find new realisations for a given shift from our collection of realisations. For this we can use special algorithms that can be initialised with centroids, such as K-Means (Lloyd, 1982), and pass embedded shifts as cluster centroids. Alternatively, it is possible to cluster all realisations, and further match clusters with shift embeddings. If the distance between a cluster of realisations and a shift embedding is less than a certain threshold, we can say that this cluster corresponds to that shift. If there is no such shift for a given threshold, this may be the cluster of a new semantic shift. To speed up the matching process, efficient similarity search algorithms such as FAISS (Johnson et al., 2019) can be used.

Since each realisation is a pair of meanings, different methods can be used to obtain an embedding for the whole realisation:

- Sum of the embedding of meanings
- Average embedding of meanings
- Concatenation of embeddings of meanings
- Embedding of concatenation of meanings.

As a benchmark to test the quality of different clustering approaches we use the manually selected dataset of semantic shifts and their realisations from the DSS. Sample from the obtained dataset is shown in the Table 5. The dataset consisted of 7441 semantic shifts and 25407 realisations of these shifts from 1300 world languages.

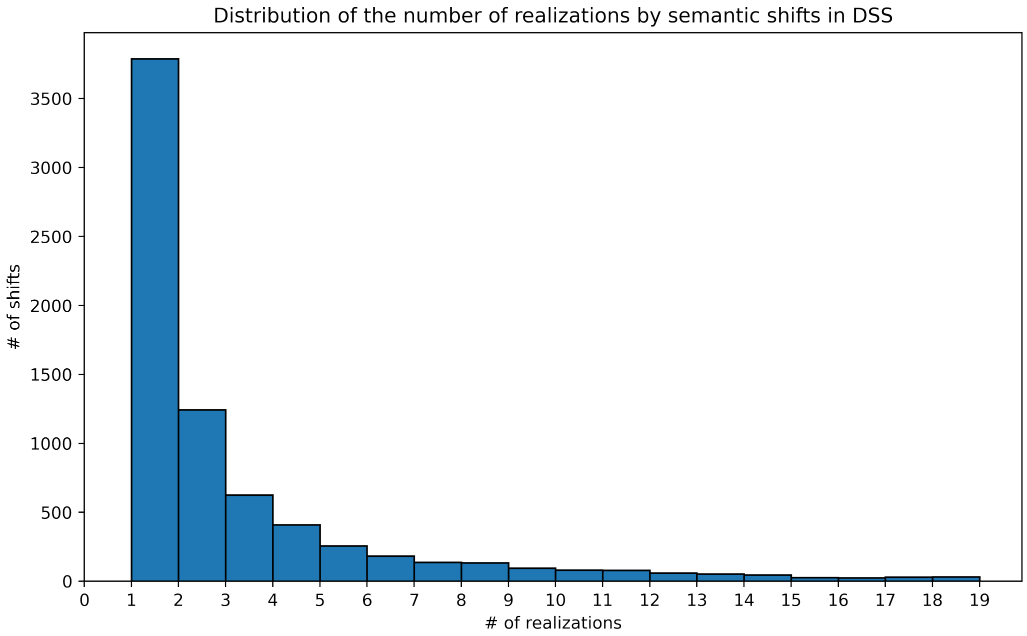

Figure 1: Distribution of the number of realisations by semantic shifts in DSS

ID	Shift ID	Language	Meaning 1	Meaning 2
19659	5218	Hungarian	head	chapter (of a book)
10706	3253	Thai	belly button	whirlpool
7336	1151	Swahili	to mix, to stir, to shake	to mix, to derange (plans etc)
1985	57	Latin	child, baby; boy	young servant, slave
8412	57	Fula	young boy	servant
11977	2798	Meadow Mari	interpreter	talkative person

Table 5: Sample of the clustering benchmark. If two realisations belong to one semantic shift their Shift Id would be the same

Figure 1 shows the distribution of the number of realisations by semantic shifts in DSS. It is noteworthy (and important for clustering) that most of the shifts have only one realisation. The Adjusted Rand Index (ARI) (Steinley, 2004) was chosen as a quality metric for clustering.

We tried different clustering approaches: K-Means, BIRCH (Zhang et al., 1996), DBSCAN (Ester et al., 1996). Since K-Means requires a number of clusters, a unique number of shifts was passed as a parameter, while we did not pass the number of clusters to other algorithms. If the number of clusters hasn't been passed to the BIRCH algorithm, it returns the subclusters, skipping the last clustering step. The DBSCAN algorithm itself determines the number of clusters using the threshold parameter, which we have left by default. Table 6 shows the performance of different clustering algorithms on the benchmark. The BIRCH algorithm outperforms K-Means and DBSCAN. In addition, we found that the sum of embeddings is the best of the methods observed to obtain the embedding of the whole realisation.

	K-Means	BIRCH	DBSCAN
Sum	0.48	**0.79**	0.73
Average	0.48	0.43	0.78
Embedding concat	0.37	0.76	0.59
String concat	0.44	0.62	0.75

Table 6: Performance of the different clustering approaches on benchmark

Below is the example of the cluster of realisations which corresponds to the semantic shift "fox" - "cunning person". Not all the clustered pairs of meanings are relevant to this semantic shift.

Language	Entry	Meaning 1	Meaning 2
Ancient Greek	ἀλώπηξ	'лиса лисица'	'лиса, разновидность акулы'
Avar	цер	'лиса, лисица'	'хитрец, пройдоха, лиса'
Bulgarian	лисица	'лисица зверь'	'перен. лиса, хитрец '
Chinese	红狐 hónghú	'рыжая лиса'	'лиса обыкновенная'
Czech	liška	'лисица, лиса'	'старая лиса, плут'
Czech	lišák	'лиса '	'старая лиса, хитрец, плут'
English	fox	'лиса, лисица'	'лиса, пронырa, хитрец'
French	renard	'лиса, лисица'	'лиса, хитрец'
German	fuchs	'лисица'	'лиса, хитрец, пройдоха'
Italian	volpe	'лиса, лисица'	'лисий мех, лиса'
Korean	여우	'лисица, лиса'	'лиса'
Lezgian	сик1	'лиса, лисица'	'лиса, хитрец'
Mari	рывыж	'лиса, лисица '	'лиса'
Polish	lis	'лисица, лиса'	'лис'
Slovak	lišiak	'лисица, лиса, самец лис'	'хитрец, плут, лиса'
Slovak	liška	'лиса, лисица'	'лиса'
Spanish	raposo	'лисица'	'хитрец, льстец, лиса'
Spanish	raposa	'лисица, лиса'	'хитрец, льстец, лиса'
Spanish	zorra	'лиса, лисица'	'лиса, шельма'
Spanish	zorrero	'лисий'	'королевский зверолов'
Spanish	zorro	'лис, лиса'	'лиса, лисий мех'
Turkish	tilki	'лиса, лисица'	'лиса, хитрец'
Ukrainian	лис	'лисица, лиса, кобель, диал. лис'	'перен. лиса'

Table 7: Cluster of realisations for the "fox" - "cunning person" semantic shift.

Meaning 1	Meaning 2	Number of languages
woman	wife	34
tooth	spike	29
long (size)	long (time)	23
to rip out	to take out	21
voice	sound	21
man	husband	21
head	classifier for round objects	21
grandmother	old woman	21
heel (anatomical)	heel (shoe)	20
Gossypium (plant)	cotton	20

Table 8: Top 10 largest clusters.

8 Validation and Postprocessing

The result is a pool of candidates that might qualify for relevant semantic shifts. In our framework we require a supervision from a linguistic team to exclude possible mistakes, true homonymy and possible inaccuracies of the result. The linguists review the data and include the best semantic shifts into the common database available online at `https://datsemshift.ru`. The manual approach that linguists used to apply previously implied looking through many polysemous words to find a suitable realisation of a semantic shift. Our approach allowed to "enrich" the raw linguistic material via filtering out of unsuitable pairs of meanings by means of a machine learning classifier. In order to quanitfy the value of our method of optimisation of linguistic work we ask the same team of experts who made an initial markup for the classifier to make judgements on a sample of the pairs of meanings which were considered as valid by our ML classifier. It turned out that when the linguists estimate a random sample of pairs of meanings they mark as "valid" only 30% of pairs. However, when they estimate a sample of pairs of meanings that received "valid" mark from the classifier, the approval rate increases up to 69%.

9 Conclusion

The method described above that processes dozens of dictionaries, extracts polysemic words, filters out typologically irrelevant cases and clustering similar pairs of meanings from various languages into semantic shifts is a valuable and powerful tool for detection of typologically relevant semantic shifts. It allows linguists to skip a lot of routine work of manually searching the dictionaries and looking for similar change of semantics in different languages. Thus, linguists can use this tool to make the conclusions about the cognitive mechanisms of the polysemy on the wide typological material.

Acknowledgements

We would like to thank Anna A. Zalizniak who read the preprint version of the paper and gave us her comments. Our colleagues Sofia Durneva, Idaliya Fedotova, Viktoria Kaprielova, Veronika Kondratieva, Tatiana Mikhailova, Maria Orlova, Maksim Rousseau, Anna Smirnitskaya, and Anna Zalizniak contributed substantially to the discussion of the criteria of a valid semantic shift. Alexander Gruntov helped a lot with preprocessing of the digitized dictionaries before parsing.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *CoRR*, abs/1810.11363.

Федотова, И.В. 2020. Полисемия в списках самодийской базисной лексики и языковые контакты. *Урало-алтайские исследования*, 2(37):77–113.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. // *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, P 226–231. AAAI Press.

Vadim Fomin, Daria Bakshandaeva, Julia Rodina, and Andrey Kutuzov. 2019. Tracing cultural diachronic semantic shifts in russian using word embeddings: test sets and baselines. *ArXiv*, abs/1905.06837.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2016:2116–2121.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Andrey Kutuzov and Elizaveta Kuzmenko. 2017. Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. // *Quantitative approaches to the russian language*, P 95–112. Routledge.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *ArXiv*, abs/1806.03537.

JM List, C Rzymski, S Greenhill, N Schweikhard, K Pianykh, and R Forkel. 2019. Concepticon 2.2. *Jena, Germany: Max Planck Institute for the Science of Human History. See https://concepticon. clld. org*.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Mariia Melymuka, Gabriella Lapesa, Max Kisselew, and Sebastian Padó. 2017. Modeling derivational morphology in Ukrainian. // *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

Tomáš Musil, Jonáš Vidra, and David Mareček. 2019. Derivational morphological relations in word embeddings. // *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, P 173–180, Florence, Italy, August. Association for Computational Linguistics.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. // *International Conference on Computational Linguistics*.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Douglas Steinley. 2004. Properties of the Hubert-Arabie Adjusted Rand Index. *Psychological methods*, 9:386–96, 09.

Anna A Zalizniak, Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50(3):633–669.

Anna A. Zalizniak. 2018. The catalogue of semantic shifts: 20 years later. *Russian Journal of Linguistics*, 22/4.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, jun.