Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

The Pilot Corpus of the English Semantic Sketches

Maria Petrova
ABBYY
Moscow, Russia
m.petrova@abbyy.com

Maria Ponomareva
HSE, ABBYY
Moscow, Russia
maria.ponomareva@abbyy.com

Alexandra Ivoylova
RSUH, MIPT, ABBYY
Moscow, Russia
aleksandra.ivoilova@abbyy.com

Abstract

The paper is devoted to the creation of the semantic sketches for English verbs. The pilot corpus consists of the English-Russian sketch pairs and is aimed to show what kind of contrastive studies the sketches help to conduct. Special attention is paid to the cross-language differences between the sketches with similar semantics. Moreover, we discuss the process of building a semantic sketch, and analyse the mistakes that could give insight to the linguistic nature of sketches.

Keywords: word sketches, semantic sketches, frame semantics, word sense disambiguation, corpus lexicography

DOI: 10.28995/2075-7182-2022-21-436-446

Пилотный корпус английских семантических скетчей

Мария Петрова
ABBYY
Москва, Россия
m.petrova@abbyy.com

Мария Пономарева
ВШЭ, ABBYY
Москва, Россия
maria.ponomareva@abbyy.com

Александра Ивойлова
РГГУ, МФТИ, ABBYY
Москва, Россия
aleksandra.ivoilova@abbyy.com

Аннотация

Работа посвящена созданию семантических скетчей для глаголов английского языка. Пилотный корпус состоит из англо-русских пар скетчей, на примере которых демонстрируется, какие сопоставительные исследования скетчи позволяют проводить. Особое внимание уделяется межъязыковым различиям скетчей одного семантического поля в разных языках. Кроме того, в статье обсуждается процесс построения скетча, возможные ошибки и их лингвистическая природа.

Ключевые слова: скетчи слов, семантические скетчи, семантика фреймов, разрешение лексической многозначности, корпусная лексикография

1 Introduction

In the current paper, we present the pilot corpus of the English semantic sketches and compare the English sketches with their Russian counterparts. The semantic sketch is a lexicographical portrait of a verb, which is built on a large dataset of contexts and includes the most frequent dependencies of the verb. The sketches consist of the semantic roles which, in turn, are filled with the most typical representatives of the roles.

The influence of context on word recognition has been well-known for quite a time. Semantic context allows faster word recognition and the inferring of the skipped words while reading. The research in this area has been conducted in psycholinguistics since the 1970s, with the earliest works by (Tweedy et al., 1977) and (Becker, 1980). Here the focus is on visual word recognition while reading and word recognition by bilingual persons (Assche et al., 2012). Another aspect of the topic is the automatic inferring of the skipped words by context, widely known as a common NLP task today.

The ability to represent the word by its context is the central idea of distributional semantics. It serves as a basis for the bag-of-words task, which is a training objective for static vectors like word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). In the approach, the context has a set length, and the words entering the fixed window are considered equally.

The semantic sketches do not have such disadvantages, as they are based on the result of the semantic parsing and therefore take into account not all the words occurring in the context, but only the words that semantically depend on the given core. That is, we take not the linearly nearest tokens, but the tokens close in the parsing graph, where the type of the links is considered as well.

The BERT (Devlin et al., 2019) contextual embeddings, which followed the static vectors and became a state-of-the-art solution for meaning representation, also rely on the idea of expressing word semantics through its context, using the objective of masked language modeling.

One of the main weaknesses of all vector representations is their interpretation and quality evaluation. The common practice is to consider the vectors as good, if they allow one to get the necessary quality for the down-stream task.

The advantage of the semantic sketches is in their interpretability and clear creation process. The sketches can be regarded as human-interpretable representation of word meanings, which one gets automatically with the help of the statistical methods used on the large text datasets.

The semantic sketches were first demonstrated in (Detkova et al., 2020), where we presented the idea of the semantic sketches itself and analysed the semantic mark-up used for building the sketches. Further, the pilot corpus of the Russian sketches has been created (Ponomareva et al., 2021). Herein, we have continued the work and created the pilot corpus of the English semantic sketches. The corpus is bilingual: each English sketch is accompanied by the Russian analogue with the same semantics, so one can compare the English sketch with the Russian one and analyse the contrastive differences between the sketches. Thereby, the contribution of the current paper is the creation of the English semantic sketches, on the one hand, and the creation of the parallel bilingual sketch corpus – on the other.

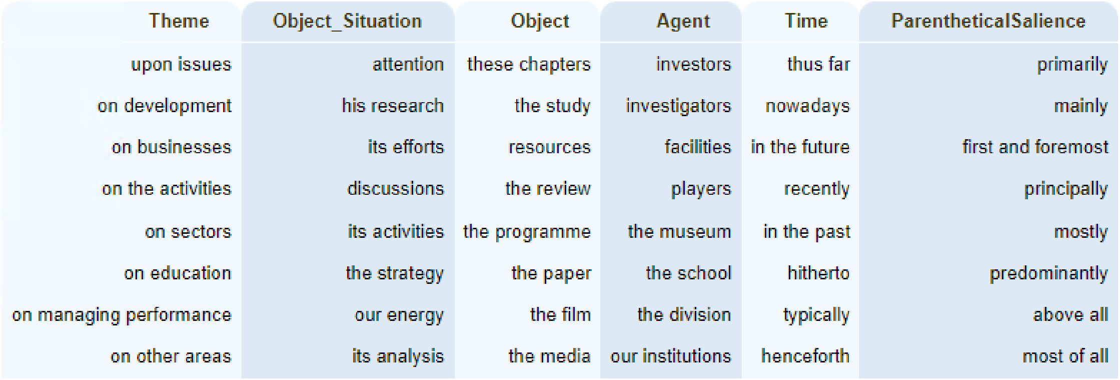
The structure of the paper is as follows. First, we briefly characterise the semantic sketches themselves. Second – give a description of the suggested corpus and explain what kind of verbs it contains. After that, we analyze the mistakes one faces when building the sketches, and focus on the cross-language differences between the sketches with similar semantics. In conclusion, we summarize the results.

2 Semantic Sketches

The idea to represent word compatibility in the form of the sketch belongs to Adam Kilgarriff (Kilgarriff et al., 2014) and is currently realized in the Sketch Engine project[1]. Verbal dependencies are classified according to their syntactic roles and statistically ranged, which allows one to see all of the most frequent syntactic dependencies of the verb at the same time.

The problem is that the syntactic sketches do not differentiate between various meanings of the verbs and combine all possible meanings in one sketch. To overcome this problem, we suggested the semantic sketches, which take the semantic models into account and classify the dependencies by their semantic relations with the core instead of their surface realizations (Detkova et al., 2020). For instance, see fig. 1 with the sketch of the verb ‘to focus’ in the meaning ’to concentrate on smth., to pay special attention to smth.’:

[1] www.sketchengine.eu



Theme	Object_Situation	Object	Agent	Time	ParentheticalSalience
upon issues	attention	these chapters	investors	thus far	primarily
on development	his research	the study	investigators	nowadays	mainly
on businesses	its efforts	resources	facilities	in the future	first and foremost
on the activities	discussions	the review	players	recently	principally
on sectors	its activities	the programme	the museum	in the past	mostly
on education	the strategy	the paper	the school	hitherto	predominantly
on managing performance	our energy	the film	the division	typically	above all
on other areas	its analysis	the media	our institutions	henceforth	most of all

Figure 1: The sketch for the verb ‘to focus:TO_FOCUS’

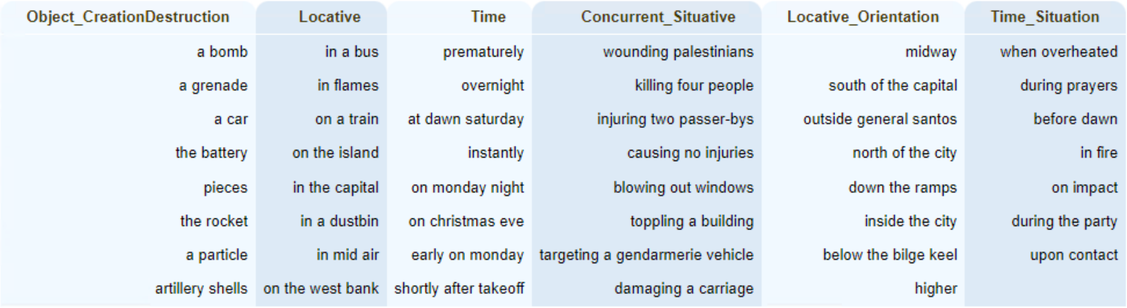
Such sketches are built for each meaning separately, however, it demands a significant text corpus with full semantic mark-up. The authors settled on the Compreno mark-up built by the Compreno parser, which includes not only actant dependencies, but all possible links.

In the Compreno model, all words are presented in the form of a thesaurus-like semantic hierarchy, which consists of the semantic classes (semantic fields), and a set of the semantic roles for the classes (for detail, see (Anisimovich et al., 2012), (Petrova, 2014)). If a verb has several meanings, it enters several semantic classes with its own semantic model each. The semantic class is specified for each sketch.

3 English SemSketches Corpus

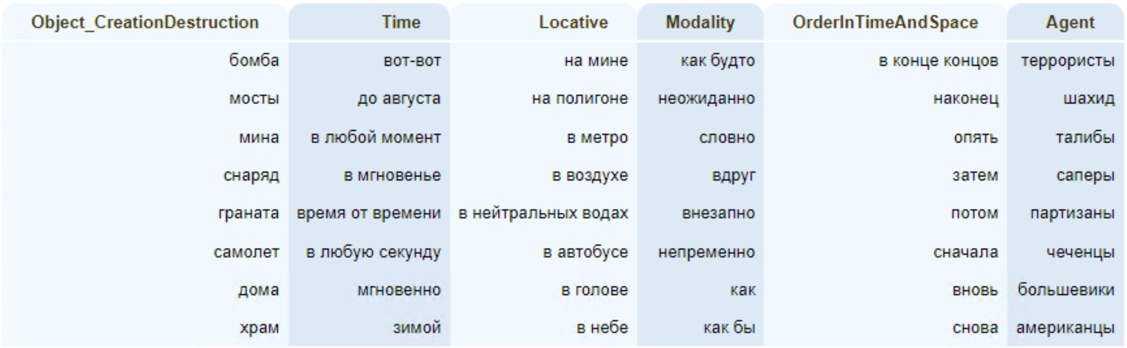
The SemSketches pilot corpus consists of 100 English sketches which are manually checked. It means that we have chosen the sketches manually according to their quality. The sketches are built on the corpus of the English texts comprising different genres, such as technical texts, news, fiction, and containing 14 million syntactic verbal links, that is, links which depend on the verbal cores.

Each English sketch is provided with the parallel Russian sketch from the same semantic class, as shown in fig. 2 and 3:



Object_CreationDestruction	Locative	Time	Concurrent_Situative	Locative_Orientation	Time_Situation
a bomb	in a bus	prematurely	wounding palestinians	midway	when overheated
a grenade	in flames	overnight	killing four people	south of the capital	during prayers
a car	on a train	at dawn saturday	injuring two passer-bys	outside general santos	before dawn
the battery	on the island	instantly	causing no injuries	north of the city	in fire
pieces	in the capital	on monday night	blowing out windows	down the ramps	on impact
the rocket	in a dustbin	on christmas eve	toppling a building	inside the city	during the party
a particle	in mid air	early on monday	targeting a gendarmerie vehicle	below the bilge keel	upon contact
artillery shells	on the west bank	shortly after takeoff	damaging a carriage	higher	

Figure 2: The sketch for the verb ‘to explode:TO_BLOW_UP’



Object_CreationDestruction	Time	Locative	Modality	OrderInTimeAndSpace	Agent
бомба	вот-вот	на мине	как будто	в конце концов	террористы
мосты	до августа	на полигоне	неожиданно	наконец	шахид
мина	в любой момент	в метро	словно	опять	талибы
снаряд	в мгновенье	в воздухе	вдруг	затем	саперы
граната	время от времени	в нейтральных водах	внезапно	потом	партизаны
самолет	в любую секунду	в автобусе	непременно	сначала	чеченцы
дома	мгновенно	в голове	как	вновь	большевики
храм	зимой	в небе	как бы	снова	американцы

Figure 3: The sketch for the verb ‘взорвать’:TO_BLOW_UP’

For 100 English sketches, 84 Russian sketches are used: it means that some Russian sketches correspond to more than one English sketch. Totally, the corpus includes 113 English-Russian sketch pairs.

The choice of the English verbs is based on the Russian corpus which was built in (Ponomareva et al., 2021). The Russian corpus, in turn, includes only polysemantic verbs as an important point is to investigate how good the sketches can differentiate between various meanings of the verbs.

To form the English sample, we have taken the verbs from the same semantic classes and set the threshold of 200 semantic links for each English verb: it means, the verb must have at least 200 links in the English texts corpus. (For comparison, the threshold for Russian verbs was 2000 links, but the Russian sketches were collected on the bigger dataset which includes more than 36 million links.)

After it, 100 English sketches were chosen, which met the above mentioned criteria and seemed to be enough representative to show the ability of the sketches to deal with polysemy, word sense disambiguation (WSD) problem, and asymmetrical compatibility of the verbs with similar semantics in different languages. Of course, the pilot corpus of 100 sketches is not enough for conducting representative contrastive research, however, certain observations seem to be of interest for comparative studies even on the small sample, as it is demonstrated below.

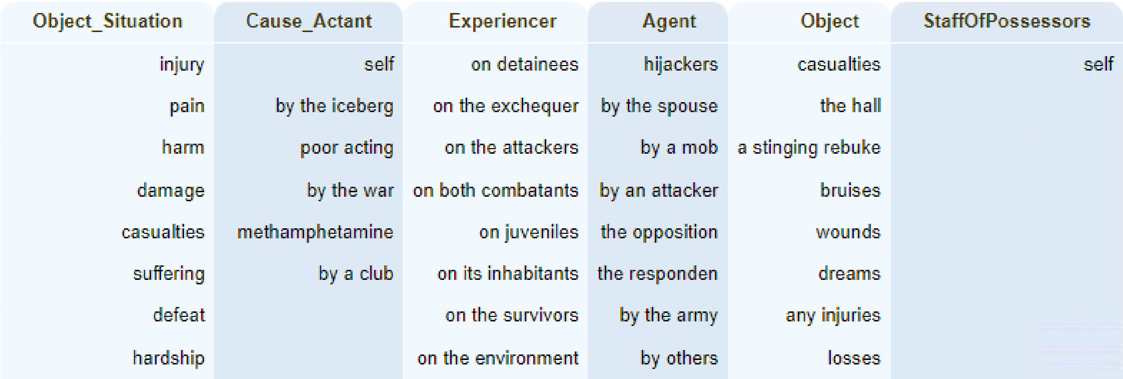
4 What the mistakes in the sketches demonstrate

The sketches are based on (1) the semantic relations the verb has in the text collection; (2) the work of the parser which classifies the relations according to their semantic roles and defines the meanings of the verbs. Therefore, the view of the sketch depends on the number of links the verb has in the corpus, and on the correctness of the parser’s work. Herein the following mistakes are possible, which concern the automatic generation of the sketches.

4.1 ‘Empty’ sketches

The insufficient number of links leads to partly ‘empty’ sketches, where the semantic roles contain very few fillers, up to only one. So when the semantic role column is partly empty, it can mean that the number of the role’s links in the corpus turned out to be poor (for instance, see the [Cause_Actant] slot in the sketch for ‘inflict’ on fig. 4]). As the number of texts grows, this problem occurs rarer.

Another reason for the lack of fillers comes from the narrowness of the semantic role filling. That is, slots like [Object] or [Cause] have rather wide filling, while [Locative] and [Time] are more restricted in this respect. In turn, the Compreno parser has a large set of characteristic slots (for size, colour, speed, modality, and so on), so some slots possess rather narrow semantics and include a small set of fillers (like the [StaffOfPossessors] slot in the same sketch on fig. 4).



Object_Situation	Cause_Actant	Experiencer	Agent	Object	StaffOfPossessors
injury	self	on detainees	hijackers	casualties	self
pain	by the iceberg	on the exchequer	by the spouse	the hall	
harm	poor acting	on the attackers	by a mob	a stinging rebuke	
damage	by the war	on both combatants	by an attacker	bruises	
casualties	methamphetamine	on juveniles	the opposition	wounds	
suffering	by a club	on its inhabitants	the responden	dreams	
defeat		on the survivors	by the army	any injuries	
hardship		on the environment	by others	losses	

Figure 4: The sketch for the verb ‘to inflict:TO_BRING_STATE_TO_SMB’

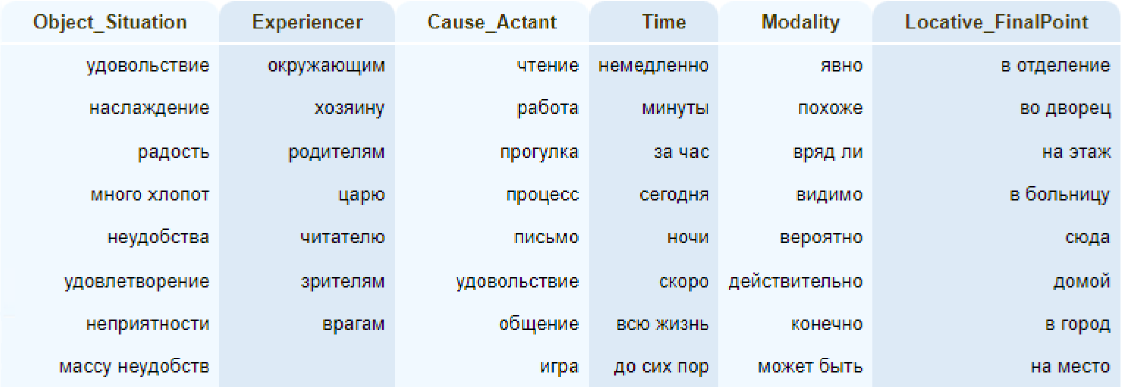
Moreover, there are verbs with narrow compatibility, such as lexical functions. For instance, see the [Object] slot in the sketch of ‘играть:TO_COMMIT’ (fig. 13).

In sketches like these, empty lines in the semantic slots are correct.

4.2 Incorrect semantic roles or incorrect fillers

Other errors concern either the incorrect choice of the semantic slot for the given verb meaning, or the wrong fillers of the slot. As one of the key points is to examine how well the sketches solve the WSD problem, this type of mistakes is important for us.

An illustration for the incorrect semantic slot is the Russian sketch for ‘доставлять:TO_BRING_STATE_TO_SMB’ (fig. 5), parallel for the above shown ‘inflict:TO_BRING_STATE_TO_SMB’. It contains the [Locative_FinalPoint] slot, which must definitely belong to another meaning of the verb – ‘bring to some place’.



Object_Situation	Experiencer	Cause_Actant	Time	Modality	Locative_FinalPoint
удовольствие	окружающим	чтение	немедленно	явно	в отделение
наслаждение	хозяину	работа	минуты	похоже	во дворец
радость	родителям	прогулка	за час	вряд ли	на этаж
много хлопот	царю	процесс	сегодня	видимо	в больницу
неудобства	читателю	письмо	ночи	вероятно	сюда
удовлетворение	зрителям	удовольствие	скоро	действительно	домой
неприятности	врагам	общение	всю жизнь	конечно	в город
массу неудобств		игра	до сих пор	может быть	на место

Figure 5: The sketch for the verb ’доставлять:TO_BRING_STATE_TO_SMB’

Examples of the wrong fillers have already been shown in (Ponomareva et al., 2021). The reasons are usually bound either with the statistics, or with the work of the parser. At the analysis stage, all possible hypotheses are built for the sentence – with all possible homonyms that can fit. The final structure turns out to be the one with the highest scores. In some cases, hypotheses with more frequent homonyms win due to their higher frequency, in spite of the fact that the whole structure with the wrong homonym gets lower evaluations.

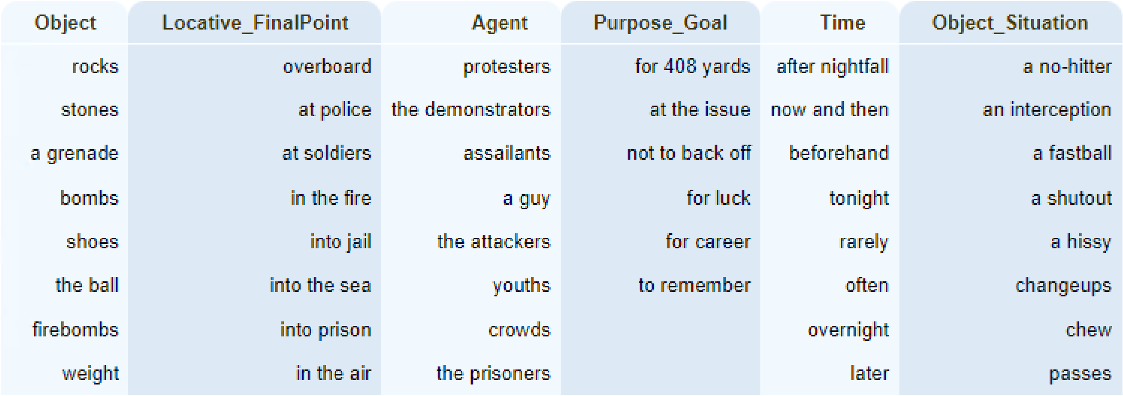
As the text collections for building the sketches grow, the statistics of the proper analysis improves, therefore, we expect that most part of the errors will be corrected with enlarging the corpora. Nevertheless, in case of the improper work of the parser, the opportunity to correct the semantic models that the parser uses exists as well.

4.3 The syntactic homonymy

Key difference between the semantic and the syntactic sketches is that in the former 1 surface realisation can correspond to various semantic roles. For instance, ‘for’-dependency can introduce Time, Purpose, Distance, Motive and a number of other relations.

Usually, the proper semantic role is chosen according to the semantic model of the given verb in Compreno – namely, the set of the semantic slots with the necessary surface realisation, the fillers of the semantic slots, and their status (which marks the role as more or less preferable).

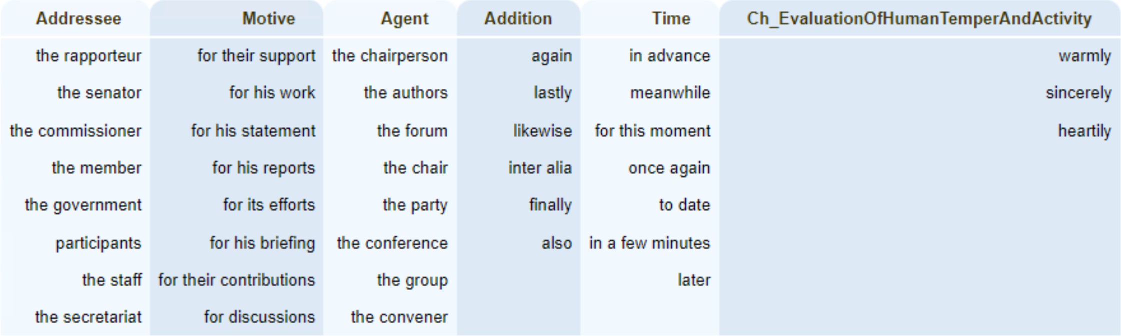
When the model or the statistics give improper results, the semantic role of the dependency can be defined incorrectly. For instance, see the [Purpose_Goal] slot of the verb ‘throw:TO_THROW’: the first line contains the nominal group ‘for 408 yards’, which must evidently belong to the [Locative_Distance] slot (fig. 6).



Object	Locative_FinalPoint	Agent	Purpose_Goal	Time	Object_Situation
rocks	overboard	protesters	for 408 yards	after nightfall	a no-hitter
stones	at police	the demonstrators	at the issue	now and then	an interception
a grenade	at soldiers	assailants	not to back off	beforehand	a fastball
bombs	in the fire	a guy	for luck	tonight	a shutout
shoes	into jail	the attackers	for career	rarely	a hissy
the ball	into the sea	youths	to remember	often	changeups
firebombs	into prison	crowds		overnight	chew
weight	in the air	the prisoners		later	passes

Figure 6: The sketch for the verb ‘to throw:TO_THROW’

Another example is the group ‘for this moment’ in the [Time] slot instead of [Motive] in the sketch of ‘to thank’ (fig. 7). Here, on the contrary, [Motive] is definitely more frequent, but ‘moment’ is a very typical [Time] filler, therefore, high statistical evaluation of the correlation ‘moment’-[Time] made the incorrect structure win.



Addressee	Motive	Agent	Addition	Time	Ch_EvaluationOfHumanTemperAndActivity
the rapporteur	for their support	the chairperson	again	in advance	warmly
the senator	for his work	the authors	lastly	meanwhile	sincerely
the commissioner	for his statement	the forum	likewise	for this moment	heartily
the member	for his reports	the chair	inter alia	once again	
the government	for its efforts	the party	finally	to date	
participants	for his briefing	the conference	also	in a few minutes	
the staff	for their contributions	the group		later	
the secretariat	for discussions	the convener			

Figure 7: The sketch for the verb ‘to thank:TO_THANK’

All the mistakes deal with different aspects of the WSD and homonymy problems. Their number does not seem significant, nevertheless, their statistical estimation must be made when creating a larger sketch corpus.

5 Cross-language differences between the sketches with similar semantics

The one-language sketch corpora suggest good lexicographic portraits of the verbs, showing their most frequent semantic links sorted according to the semantic roles of the dependencies. Moreover, apart from purely lexicographic tasks, the sketches allow one to solve various problems bound both with the context usage of the verbs and with their polysemy.

Another purpose of the sketches deals with contrastive studies. Parallel sketches from different languages give perfect representation of the correlation between similar verbs, therefore, parallel sketch corpora would be helpful in this respect.

Evidently, each sketch can correspond to more than one sketch in another language. To get a full set of all possible counterparts, one should take the necessary sketch in one language and the sketches for all the semantic equivalents in the same semantic class in another language. After it, one can range the counterparts according to their affinity with the primary verb. We have not made such full sets in the pilot corpus, however, adding this option is included in further plans.

At the current stage of the project, the correlations between the English and the Russian sketches do not include all possible correlations for each verb, so the sketch pairs are just a subset of the possible variants.

Some pairs look similar: both English and Russian sketches include the same set of semantic roles, and the semantic roles contain either fillers with close semantics, or just a wide range of fillers with no special semantic restrictions on them.

At the same time, many sketches demonstrate significant differences between the English and Russian equivalents. Most of them concern the following situations:

(a) some semantic slot is present in the sketch of one language and is absent – in the corresponding sketch in another language;

(b) equivalent sketches contain the same sets of the roles in both languages, but the fillers of some role differ significantly;

(c) the semantic field where the considered verbs belong is structured differently in different languages.

5.1 Different semantic roles in the equivalent sketches from different languages

Frequently, the semantic role sets in the parallel sketches do not coincide completely. It concerns both the actant roles and the circumstantial ones. The reasons can be different. First, the semantics of one verb may be wider than the semantics of the other, therefore, the model of the former can include additional roles which are absent in the model of the latter. Second, the model of both verbs can include the same sets of roles, however, the frequency of some roles may differ for various verbs, which can be motivated both by the verb's semantics and by the representativeness and contents of the corpora for building the sketches.

An example of the first case is the correlation between the semantic derivates in different languages. For instance, Russian verb 'трясти' 'to shake' does not attach the initial point dependency in contexts like (1) and (2), while the English 'shake' does:

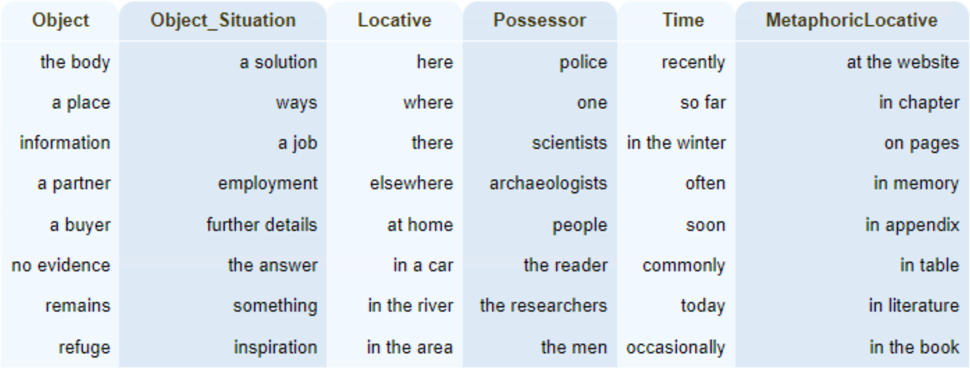
(1) A sound they couldn't shake [from their Locative_InitialPoint: heads] – Звук, который им никак не удавалось вытряхнуть [из Locative_InitialPoint: головы];

(2) I saw immediately that my few belongings had been disturbed–collars not refolded, one of my chemises balled up and pushed into a corner, the tortoiseshell comb shaken [from its Locative_InitialPoint: handkerchief]. – И сразу увидела, что в моих вещах кто-то рылся — воротники были сложены неаккуратно, одна из моих рубашек скомкана и засунута в угол, черепаховый гребень вытряхнут [из носового Locative_InitialPoint: платка].

In Russian, the semantic derivate 'вытряхнуть' 'shake out' is used when the initial point role is expressed in a sentence. Therefore, the sketches can show that 'shake' usually corresponds to the Russian 'трясти' (which does not mark the 'direction of shaking'), but can also correspond to 'вытрясти' (which denotes the 'from' direction) with the dependency of the initial point.

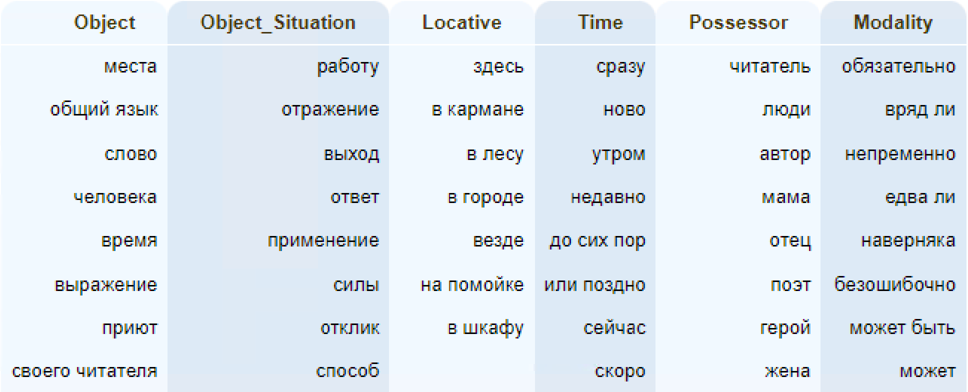
Nevertheless, there can be occasional variations depending on the contents of the corpora, especially as far as less frequent verbs are considered. The more the corpora are, the more stable are the results. Thus we permanently enlarge the size of the dataset for building the sketches.

As an instance of such statistical oscillations, see the sketches for “find:TO_SEEK_FIND” and “найти:TO_SEEK_FIND”. The first five roles coincide, but the sixth one is different – it is [Metaphoric_Locative] for the English ‘find’ and [Modality] for the Russian ‘найти’ (fig. 8, 9):



Object	Object_Situation	Locative	Possessor	Time	MetaphoricLocative
the body	a solution	here	police	recently	at the website
a place	ways	where	one	so far	in chapter
information	a job	there	scientists	in the winter	on pages
a partner	employment	elsewhere	archaeologists	often	in memory
a buyer	further details	at home	people	soon	in appendix
no evidence	the answer	in a car	the reader	commonly	in table
remains	something	in the river	the researchers	today	in literature
refuge	inspiration	in the area	the men	occasionally	in the book

Figure 8: The sketch for the verb ‘to find:TO_SEEK_FIND’



Object	Object_Situation	Locative	Time	Possessor	Modality
места	работу	здесь	сразу	читатель	обязательно
общий язык	отражение	в кармане	ново	люди	вряд ли
слово	выход	в лесу	утром	автор	непременно
человека	ответ	в городе	недавно	мама	едва ли
время	применение	везде	до сих пор	отец	наверняка
выражение	силы	на помойке	или поздно	поэт	безошибочно
приют	отклик	в шкафу	сейчас	герой	может быть
своего читателя	способ		скоро	жена	может

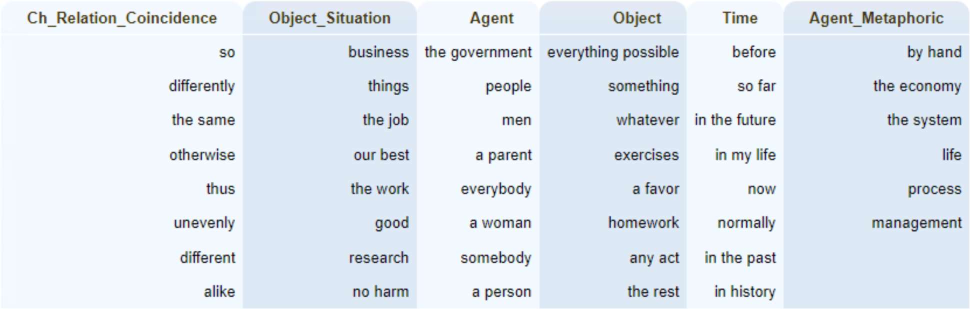
Figure 9: The sketch for the verb ‘найти:TO_SEEK_FIND’

Both roles – [Metaphoric_Locative] and [Modality] – can be frequently used with both verbs. In this case, the difference does not seem meaningful.

5.2 Different fillers of the semantic roles

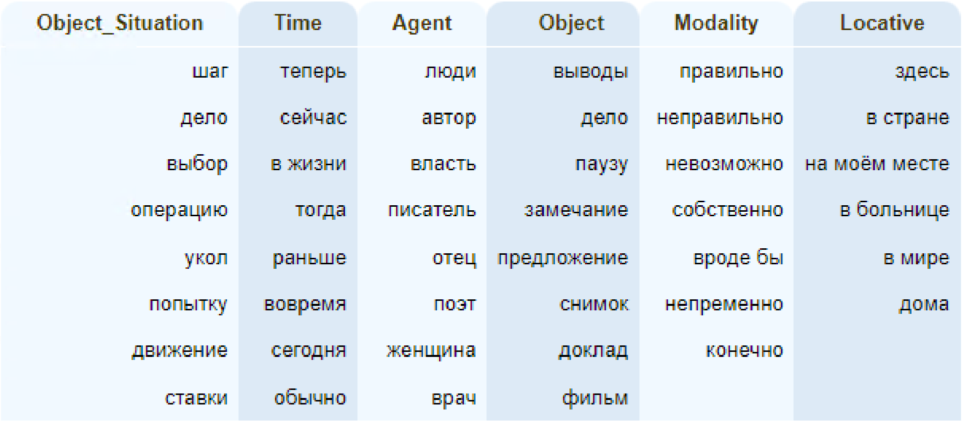
Let us consider some sketches for the descendants of the semantic class “TO_COMMIT”: the English verbs ‘do’, ‘play’ and the Russian verbs ‘делать’,‘играть’. “TO_COMMIT” is a kind of lexical function, where the verbs have rather narrow compatibility in the [Object] role (place trust/hope vs pay a visit vs play a joke/trick vs take a look/try/walk/etc., and so on).

As fig. 10, 11, 12 and 13 demonstrate, the compatibility of the verbs ‘do’ and ‘делать’ is rather wide, while ‘играть’ combines with only four Object fillers.



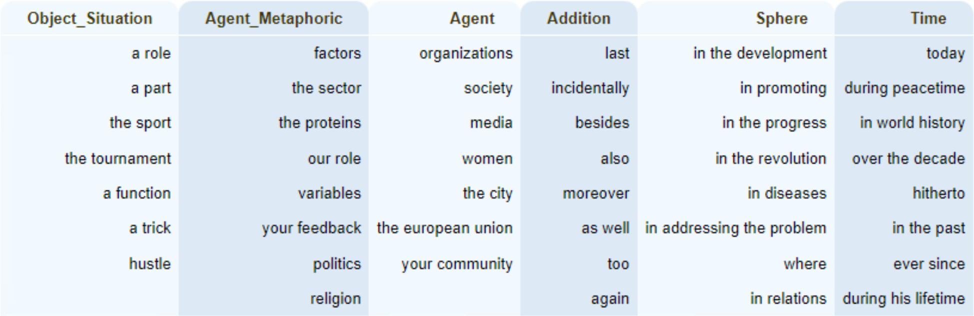
Ch_Relation_Coincidence	Object_Situation	Agent	Object	Time	Agent_Metaphoric
so	business	the government	everything possible	before	by hand
differently	things	people	something	so far	the economy
the same	the job	men	whatever	in the future	the system
otherwise	our best	a parent	exercises	in my life	life
thus	the work	everybody	a favor	now	process
unevenly	good	a woman	homework	normally	management
different	research	somebody	any act	in the past	
alike	no harm	a person	the rest	in history	

Figure 10: The sketch for the verb ‘to do:TO_COMMIT’



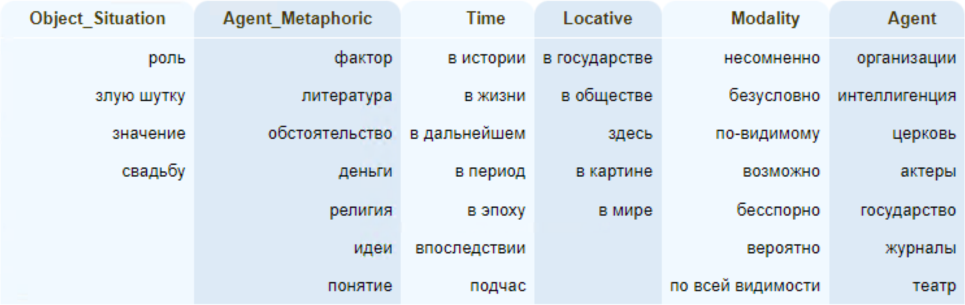
Object_Situation	Time	Agent	Object	Modality	Locative
шаг	теперь	люди	выводы	правильно	здесь
дело	сейчас	автор	дело	неправильно	в стране
выбор	в жизни	власть	паузу	невозможно	на моём месте
операцию	тогда	писатель	замечание	собственно	в больнице
укол	раньше	отец	предложение	вроде бы	в мире
попытку	вовремя	поэт	снимок	непременно	дома
движение	сегодня	женщина	доклад	конечно	
ставки	обычно	врач	фильм		

Figure 11: The sketch for the verb ‘делать:TO_COMMIT’



Object_Situation	Agent_Metaphoric	Agent	Addition	Sphere	Time
a role	factors	organizations	last	in the development	today
a part	the sector	society	incidentally	in promoting	during peacetime
the sport	the proteins	media	besides	in the progress	in world history
the tournament	our role	women	also	in the revolution	over the decade
a function	variables	the city	moreover	in diseases	hitherto
a trick	your feedback	the european union	as well	in addressing the problem	in the past
hustle	politics	your community	too	where	ever since
	religion		again	in relations	during his lifetime

Figure 12: The sketch for the verb ‘to play:TO_COMMIT’



Object_Situation	Agent_Metaphoric	Time	Locative	Modality	Agent
роль	фактор	в истории	в государстве	несомненно	организации
злую шутку	литература	в жизни	в обществе	безусловно	интеллигенция
значение	обстоятельство	в дальнейшем	здесь	по-видимому	церковь
свадьбу	деньги	в период	в картине	возможно	актеры
	религия	в эпоху	в мире	бесспорно	государство
	идеи	впоследствии		вероятно	журналы
	понятие	подчас		по всей видимости	театр

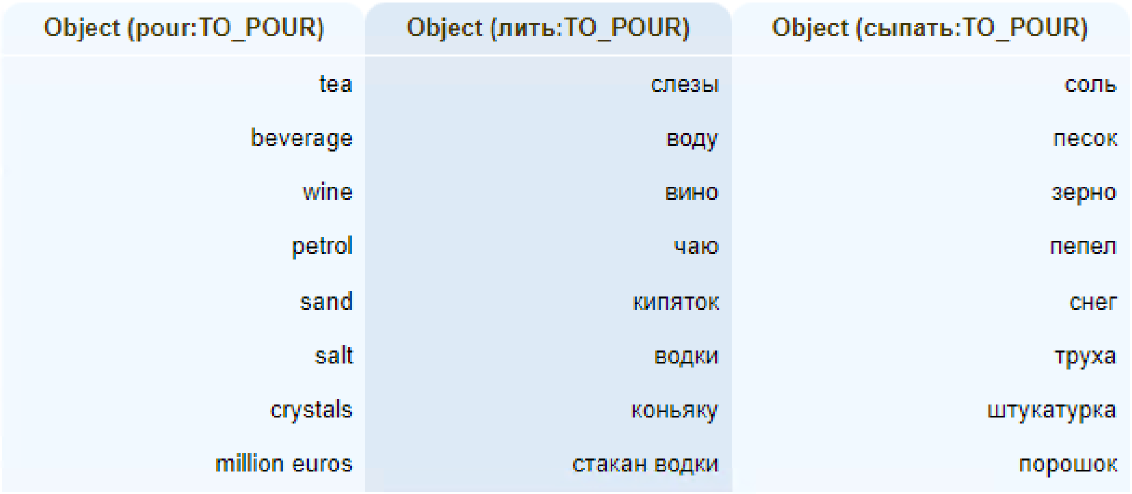
Figure 13: The sketch for the verb ‘играть:TO_COMMIT’

Besides, the four verbs differ in the sets of the semantic roles as well. [Agent], [Object_Situation] and [Time] are present in all four sketches. [Object] is absent in the sketches of ‘играть’ and ‘play’ as their compatibility does not include the corresponding fillers.

‘Do’ and ‘играть/play’ include [Agent_Metaphoric] slot, while ‘делать’ does not include it. The reason seems to be in the semantics of the fillers of the [Object] and [Object_Situation] slots: the most frequent Object_Class fillers are ‘шаг’ ‘step’, ‘выбор’ ‘choice’, ‘операция’ ‘operation’, ‘снимок’ ‘picture’ and so on, which are more often combined with active human-like agent rather than inanimate agents like ‘economy, system, process’ and alike.

As far as the circumstantial dependencies are concerned, both Russian sketches include the semantic roles of [Modality] and [Locative] while the English ‘do’ includes [Ch_Relation_Coincidence] slot (in the Compreno model, it characterizes objects or situations according to their similarity) and ‘play’ – [Addition] and [Sphere]. At first sight, these differences do not seem meaningful, however, it would be interesting to regard the sketches of the whole semantic class TO_COMMIT to examine how regular such correlations are.

Another example concerns verbs with wider compatibility, where the restrictions on the Object role are not purely lexicalized, but concern a wider range of fillers with common semantic features. For instance, let us take the semantic field “TO_POUR” (something liquid or friable). English and Russian structure it differently as far as the core verbs’ compatibility is concerned. Namely, the English verb ‘to pour’ attaches objects which are liquid (water, wine), friable (sand, sugar), or consist of many small pieces (crystals, euros, diced meat, and so on). In Russian, the verb ‘лить’ is used with liquid objects only and the verb ‘сыпать’ – only with friable objects and objects consisting of many small pieces. Therefore, the Object slot fillers differ correspondingly in the sketches (fig. 14).



Object (pour:TO_POUR)	Object (лить:TO_POUR)	Object (сыпать:TO_POUR)
tea	слезы	соль
beverage	воду	песок
wine	вино	зерно
petrol	чаю	пепел
sand	кипяток	снег
salt	водки	труха
crystals	коньяку	штукатурка
million euros	стакан водки	порошок

Figure 14: The fragment of the sketches for the verbs ‘to pour:TO_POUR’, ‘лить:TO_POUR, and ‘сыпать:TO_POUR’

Nonetheless, the amount of eight most frequent fillers which is usually shown in the sketches is not always enough to demonstrate such differences, as the most frequent objects can bear the same semantic features.

As one can see, the sketches suggest a wide range of comparative data in the field of semantics and demonstrate the semantical differences between the verbs of the same semantic class both in different languages and within one language as well.

6 Conclusion

In the given paper, we have presented the pilot corpus of the English semantic sketches.

As the sketches are provided with their semantic parallels in Russian, we have also illustrated what kind of comparative studies the sketches allow to conduct, especially as far as the differences in the semantic roles and their typical fillers are concerned. An important point is the ability of the sketches to deal with polysemy and to differentiate between various homonyms.

We have also discussed common types of mistakes occurring while building the sketches and speculated about their linguistic and technical nature.

Our further plans are to improve the sketches by obtaining them on a bigger dataset, to enlarge the sketch corpus and build the sketches for each verb from the dataset, to provide the corpus with some additional features, such as the opportunity to show more semantic slots and more fillers of the slots when necessary, and to see the correlations between all the verbs of the same semantic class. After it, the work on adding other languages to the sketch corpus will be started.

At the same time, we work on the open corpus of the Compreno semantic mark-up which will include a detailed description of the mark-up principles and the semantic roles used in the mark-up, which will facilitate the understanding of the roles used in the sketches.

The current corpus is available at github[2]. Besides, we continue the work on integrating the semantic sketches in the General Internet-Corpus of Russian (GICR).

We hope the corpus would contribute to different NLP areas, especially to solving the WSD problem.

References

KV Anisimovich, K Yu Druzhkin, KA Zuev, FR Minlos, MA Petrova, and VP Selegei. 2012. Syntactic and semantic parser based on abbyy compreno linguistic technologies. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, P 91–103.

Eva Van Assche, Wouter Duyck, and Robert J Hartsuiker. 2012. Bilingual word recognition in a sentence context. *Frontiers in psychology*, 3:174.

Curtis A Becker. 1980. Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & cognition*, 8(6):493–512.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Julia Detkova, Valeriy Novitskiy, Maria Petrova, and Vladimir Selegey. 2020. Differential semantic sketches for russian internet-corpora. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, Moscow.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. // *Neural and Information Processing System (NIPS)*.

MA Petrova. 2014. The compreno semantic model: the universality problem. *International Journal of Lexicography*, 27(2):105–129.

Maria Ponomareva, Maria Petrova, Julia Detkova, Oleg Serikov, and Maria Yarova. 2021. Semsketches-2021: experimenting with the machine processing of the pilot corpus. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, Moscow.

James R Tweedy, Robert H Lapinski, and Roger W Schvaneveldt. 1977. Semantic-context effects on word recognition: Influence of varying the proportion of items presented in an appropriate context. *Memory & Cognition*, 5(1):84–89.

[2] https://github.com/dialogue-evaluation/SemSketches/tree/main/data/task_2