EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 236–248
CILC2016. 8th International Conference on Corpus Linguistics

Supracorpora Databases as Corpus-based Superstructure for Manual Annotation of Parallel Corpora*

Mikhail Kruzhkov
Institute of Informatics Problems, FRC CSC RAS, Russia
magnit75@yandex.ru

Abstract

This paper presents a new type on corpus-based information resource: supracorpora databases (SCDBs). SCDBs are designed to enhance functionality of linguistic corpora by supporting customizable manual annotation of linguistic items, including multi-word items. This is similar to query result categorization functions available in some corpora and to functions provided by some of the standalone corpus annotation tools, although many features supported by SCDBs are more sophisticated (e.g. they allow for detailed annotation of multi-word linguistic items, including specification of main words and immediate context). More importantly still, SCDBs allow researchers to create annotated translation correspondences (TCs) in parallel corpora. Aggregation of searchable TCs in a SCDB represents a unique information resource that facilitates creation of new explicit knowledge about cross-linguistic correspondences and translation models. An overview of four SCDBs developed up to date is also included in this paper.

1 Introduction

This article examines *supracorpora databases* (SCDBs), a new information resource that extends capabilities of researchers working with linguistic corpora, *parallel corpora* in particular.

There are some terminological issues related to the term "parallel corpora". In this paper parallel corpora are understood as corpora containing source texts and their translations, as opposed to comparable corpora which contain two or more monolingual subcorpora designed using to the same sampling techniques. Furthermore, source texts and translations in parallel corpora are *aligned*, usually at the level of sentences†.

* This work was made possible by Russian Foundation for Basic Research (Grant 16-06-00070) and Russian Foundation for Humanities (Grants 15-04-00507 and 16-24-41002).

† For more information on parallel corpora and the alignment process see e.g. McEnery, Xiao & Tono, 2006: 47-51.

A SCDB is essentially a superstructure built on top of the structure of a linguistic corpus that enables researchers to save information about *linguistic items* (LIs[*]) that they discover in the corpus in an orderly manner. More specifically, a SCDB allows researchers to annotate LIs of a certain type with categories from a user-defined classification system that may evolve in the course of a research project. This classification system relies on customizable sets of features that are developed by the researchers themselves taking into account the subject, goals and scope of their research. This is somewhat similar to query result categorization function available in some corpora (e.g. Hoffmann & Evert, 2006: 181) and to some standalone corpus annotation tools, the most prominent – UAM Corpus Tool (O'Donnell, 2008), although some features supported by SCDBs are more sophisticated (such as more detailed annotation of multi-word linguistic items, including specification of main words and immediate context).

While the SCDB concept can be applied to various types of corpora, it has been designed specifically to be applied to parallel corpora. In parallel corpora SCDBs provide features that are quite unique: they allow researchers to save information on *translation correspondences* (TCs) in a consistent way. By aggregating large arrays of TCs researchers create formal descriptions of information on various translation patterns that can be subject to further qualitative and quantitative analysis.

Researchers build TCs in a SCDB by following a certain procedure that is briefly described below. First they search for LIs of a certain type in the source texts (ST) – this search is semi-automated, based on queries that rely on corpus annotation. Each LI is displayed as part of a pair of aligned sentences which makes it easy to manually locate in target text (TT) fragments that correspond to LIs in the ST; such fragments are called *functionally equivalent fragments*[†] (FEFs). At this point researchers annotate both LIs in the ST and their FEFs in the TT using distinct sets of features for different languages. These sets of features are customizable and are developed by researchers according to their research goals. Finally researchers link annotated LIs (in ST) to annotated FEFs (in TT) to produce TCs[‡] which in turn are annotated with features relevant to a TC as a whole. This process will be described in more detail in Section 3.

By processing a parallel corpus (or a subcorpus) in the described fashion researchers produce an array of TCs for LIs of the investigated type. The aggregation of created searchable TCs in a SCDB represents a unique information resource that allows for new possibilities in the field of corpus-based contrastive analysis.

Search functions implemented in SCDBs allow to quickly locate in the information array TCs with certain combinations of features that have been assigned during annotation, which is useful since SCDBs may aggregate thousands of TCs. Furthermore, information in SCDBs may be analyzed statistically to calculate frequencies of various types of translation correspondences or correlations between different context features of the examined LIs.

So far the SCDB concept has been tested in four separate studies that are still in progress, though on different stages (e.g., see http://a179.ipi.ac.ru/corpora_dynasty/main.aspx). All of the studies use the Russian-French parallel corpus of the Russian National Corpus (RNC, http://ruscorpora.ru/search-para-fr.html) to contrast different types of Russian LIs with their French correspondences. This corpus contains mostly Russian literary works and their translations into French made by professional translators. Both grammatical and lexical LIs have been examined:

1. Russian personal verbal forms

[*] Term LI here refers to both lexical and grammatical linguistic items that may include one or more words that may be either adjacent (cf. French connector ***et avec ça***, *il était gentil*), or separated by other words (cf. French connector ***non seulement*** *le visage* ***mais aussi*** *l'âme*).

[†] The term 'functionally equivalent fragment' was proposed by Dobrovolsky D.O. (Dobrovolsky, Kretov & Sharoff 2005).

[‡] A coupled pair of annotated LI (ST) and FEF (TT) is also often referred to as a *monoequivalence* (ME), the term proposed by Anna A. Zalizniak (Loiseau, Sitchinava, Zalizniak & Zatsman, 2013: 102-103).

2. Russian language-specific items
3. Russian connectors
4. Russian impersonal forms

These studies and their associated SCDBs will be described in Section 4.

Use of SCDBs facilitates creation of new explicit knowledge about cross-linguistic correspondences and translation models. For example, during the study of Russian personal verbal forms regular variants of translation of Russian grammatical tenses into French were extracted from the SCDB data, some of them not previously described in contrastive Russian-French grammars (Zatsman & Buntman, 2015).

2 Motivation

The concept of SCDB was created as a result of bringing together two groups of specialists. On one hand, there was a group of linguistic experts who where conducting contrastive analysis of Russian linguistic items based on data contained in the Russian-French parallel corpus of the RNC, for simplicity here they will be called "the linguists". On the other hand, there was a group of software developers who were ready to provide their resources and expertise in order to facilitate the linguists' efforts.

The linguists found out that the tools available in the RNC were not efficient enough for conducting their analysis. One problem was that while the corpus was able to provide annotation at the level of a word, there was no obvious way to annotate the actual linguistic items (LIs) found in the corpus. Another problem was that even though the parallel corpus presented Russian-French texts aligned at the level of sentences, there was no obvious way to produce any formal representation of actual translation correspondences (TCs) that would be able to match annotated LIs from ST to their functionally equivalent fragments (FEFs) in TT.

These problems made it difficult to conduct the contrastive analysis the linguists were aiming at. That's why they called on the developers to provide them tools that would be able to address those problems. The developers decided that the best way to do it would be to create a database that would be able to integrate all the information contained in the parallel corpus and to add a new layer of information on top of it that would allow the linguists to save data about LIs, FEFs and TCs. Working closely together, the linguists and the developers eventually managed to create such a database that was the first in a series of similar databases that eventually came to be known as supracorpora databases, or SCDB[*]. At the same time the underlying concept of SCDB was created.

3 The SCDB Annotation Concept

The SCDB concept implies that during each study the linguists are investigating one class of LIs (e.g. personal verbal forms, connectors, phrasal subordinating conjunctions, etc.) in the source language "in the mirror" of a target language[†]. During the study the linguists aim to manually annotate all the LIs of the investigated class in the source texts of the parallel corpus and to create a TC for each one (this includes location and manual annotation of the appropriate FEFs in the target texts). This approach allows to apply qualitative analysis to the array of TCs in SCDB. When dealing with highly frequent LIs, such as personal verbal forms, it may be not possible to manually annotate all the

[*] The term was first proposed in Kruzhkov, 2015.
[†] Eventually we plan to support several target languages.

LIs in the corpus – in such cases the linguists may decide to limit their research to a certain subcorpus of the available corpus.

Furthermore, the SCDB concept implies that the investigated LIs may be divided into a finite number of non-overlapping subclasses. Each LI in the ST is annotated with a number of properties. Two types of properties are distinguished:

- *Basic type* is the principle property of a LI, one that allows to attribute it to one of the distinct subclasses mentioned above. It may specify either lexical or grammatical form of LI, or both. Each LI must always be assigned exactly 1 basic type (see Figure 1). For example, in the SCDB of personal verbal forms most basic types for Russian LIs are combinations of tense and aspect of the verb, such as *future-perfective*, *future-imperfective*, *past-perfective*, *past-imperfective*, *imperative-perfective*, *imperative-imperfective*[*], etc.
- *Supplementary features* specify additional properties of the investigated LIs. They usually refer to the context of a LI. Each LI may be assigned zero, 1 or more supplementary features (see Figure 1). For example, in the SCDB of personal verbal forms there are such supplementary features for Russian language as *presence of subordinate perfective/imperfective infinitive* (SubInfPF/SubInfIPF), *presence of a modal determiner* (ModDet), *presence of negation* (Neg), etc.

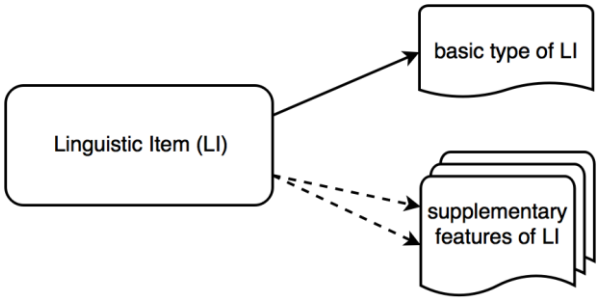

Figure 1: Annotation of LIs in SCDB

FEFs in TT are annotated according to the same principle, except that basic types of FEFs must not be restricted to any particular class of LIs in the target language – they may refer to any means translators use to reproduce functions of the appropriate LIs from the source text in the target language. The assortment of basic types of FEFs often has to be expanded in the course of the study as new translation patterns are being discovered.

To provide an example, Table 1 below presents the list of French basic types (types of FEFs) used in the current version of the SCDB of Russian personal verbal forms. As we can see, 19 of these basic types are French tenses, which means that they represent congruent translations. The rest of basic types in the table represent divergent translations[†] revealed in the corpus, in Table 1 they are given in italic. Basic types 20-25 represent several types of non-finite verb forms, basic type 26 (*Substantif*) represents cases when Russian verbal forms are translated into French by noun phrases. There are also two special values included in this list: "*Non determiné*" (27) is assigned to a FEF when a linguist cannot assign an appropriate basic type to a FEF (but the corresponding LI has been translated by some non-obvious means), and "*Omission*" (28) is assigned to a FEF when meaning of the corresponding LI is not present in the translation at all (the translator has chosen to skip it for some reason).

[*] In the SCDB interface and in this article names of properties are often abbreviated, e.g. Fut-PF, Fut-IPF, Past-PF, Past-IPF, etc.

[†] For more information on congruent and divergent translations see Johansson, 2007: 24-25.

No	basic type	No	basic type
1	present	15	subjonctif imparfait
2	passé composé	16	subjonctif plus-que-parfait
3	passé simple	17	conditionnel present
4	imparfait	18	conditionnel passé
5	plus-que-parfait	19	conditionnel passé 2
6	passé antérieur	20	*participe present*
7	passé immédiat	21	*participe passé*
8	futur simple	22	*participe passé compose*
9	futur antérieur	23	*gérondif*
10	futur immédiat	24	*infinitive*
11	futur immédiat dans le passé	25	*infinitif passé*
12	impératif	26	*Substantif*
13	subjonctif présent	27	*Non determine*
14	subjonctif passé	28	*Omission*

Table 1: List of basic types of French FEFs in the SCDB of Russian personal verbal forms. Basic types 20-28 (*in italic*) represent divergent translations.

Below we provide two examples of the annotation scheme used in SCDB. The first one (Figure 2) is an example of an annotated French FEF taken from the SCDB of Russian personal verbal forms[*]. The basic type of the FEF is specified in the topmost dark-grey box, while the supplementary features are specified in light-grey boxes below, in angle brackets[†].

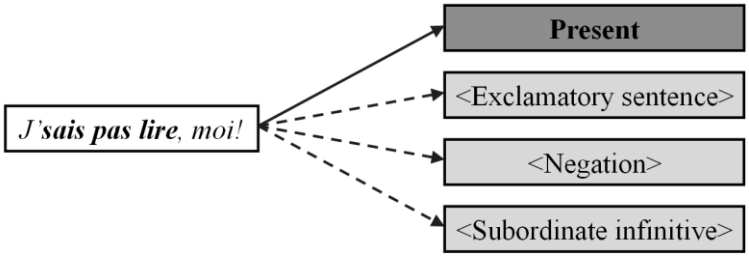

Figure 2: Example of annotation of a FEF in the SCDB of Russian personal verbal forms ('*Me, I can't read!*').

The second example (Figure 3) is taken from the SCDB of Russian connectors. As one can expect, in this database basic types are mostly specified by lexical form of LIs.

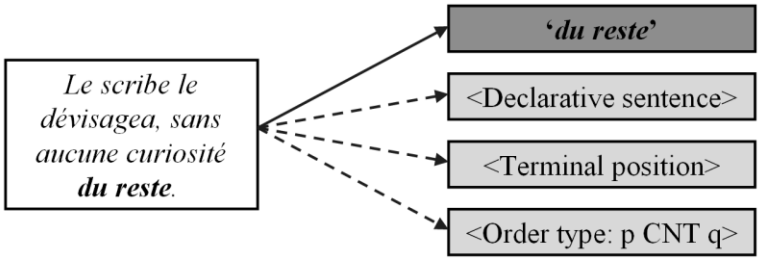

Figure 3: Example of annotation of a FEF in the SCDB of Russian connectors ('*The scribe looked at him, without any interest however*'). Supplementary feature "order type" specifies the order of the elements that are linked by a connector.

[*] Since most of the audience of this paper is probably more familiar with French than Russian, where possible the preference is given to French examples over Russian ones.

[†] In the examples we aim to give a general idea of how annotation works, that's why in most cases we present only some of the supplementary features assigned to LIs and their FEFs.

After matching an annotated LI in the ST to its appropriate annotated FEF in the TT a researcher produces a TC. As a distinct object, a TC can also be annotated with features from a separate list of features that are relevant to the TC as a whole but not to either LI or FEF. Such features are called TC-relevant features. One example of a TC-relevant feature in the SCDB of Russian personal verbal forms is the case of *subject change* between the ST and TT which will be illustrated in the following example.

Here we will revise the steps a researcher must take in order to produce a TC based on aligned texts in a parallel corpus:

1. Researcher locates a fragment that contains a LI of interest in the source texts of the parallel corpus – see the grayed text in the left column of Figure 4 that contains a Russian personal verbal form (lit.: *'I go to the theater occasionally'*). This step is semi-automated, assisted by custom-built queries.
2. Researcher manually locates appropriate FEF for this LI in the translation segment of the aligned pair – see the grayed text in the right column of Figure 4 (lit.: *'It happens to me to go to the theater'*).

Вы все дома сидите.	Vous restez à la maison tout le temps.
– Я иногда в театр хожу да в гости: вот бы…	— Il m'arrive d'aller au théâtre, en visite: à ce moment-là...
– Что за уборка ночью!	— Joli nettoyage, la nuit!

Figure 4: A fragment of an aligned text contained in the parallel corpus.

3. Researcher annotates both the LI and the FEF: he specifies main words and immediate context and assigns the basic type and supplementary features (Figure 5).
4. Researcher creates a TC by matching the annotated LI to the annotated FEF. If necessary, he also assigns TC-relevant features to the TC as a whole (see Figure 5 – here the researcher assigned *subject change* feature to the TC).

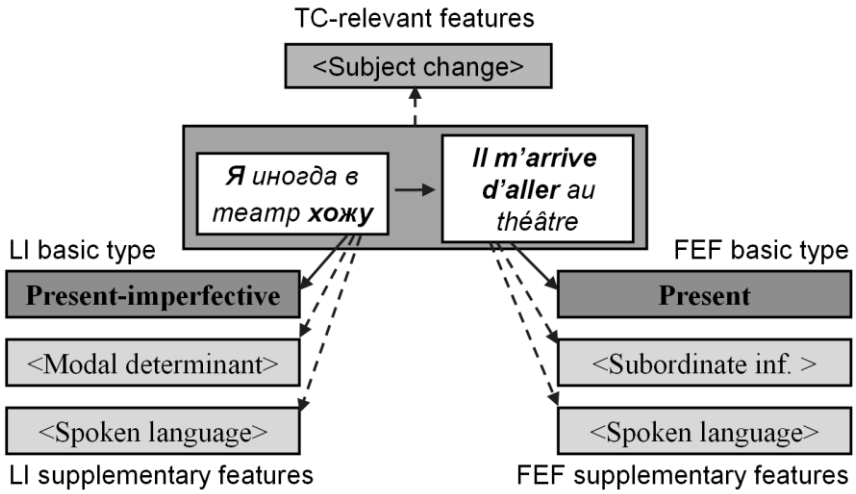

Figure 5: Overall structure of a TC created as a result of manual annotation based on data from a parallel corpus.

Normally this process is repeated until all LIs of the investigated class in the parallel corpus (or a chosen subcorpus) have been processed in this way.

4 SCDBs Created Up To Date

SCDBs are built around specific linguistic studies conducted as part of concrete research projects. The most important distinctive feature of a SCDB is the object of the respective study, e.g. the class of LIs investigated as part of that study. Another distinctive feature of a SCDB is the principle of partitioning of the investigated class of LIs into subclasses (which is defined by the list of basic types of LIs, see Section 3). In this section we present an overview of SCDBs created up to date, in chronological order. During this overview we will emphasize how the SCDB structure and concept have been adopted in order to assimilate nature of particular classes and subclasses of LIs that we had to deal with.

4.1 The SCDB of Russian Personal Verbal Forms

This SCDB was created in 2012, it was the first SCDB of this kind. The study of personal verbal forms ran parallel to development of the concept and structure of the underlying database that only later came to be known as SCDB. The project head and main inspirer behind the idea of SCDB was Anna A. Zalizniak, but in total there were a dozen people involved in the development process, both linguists and software developers. This work was reported in Zalizniak, Sitchinava, Loiseau, Kruzhkov & Zatsman, 2013 and Loiseau et al., 2013.

The object of the underlying study was Russian personal verbal forms (with some restrictions, e.g. the verb *byt'* [to be] was excluded from this study). These verbal forms were investigated "in the mirror" of their French translation correspondences, which allowed to clarify peculiarities of their function and meaning in Russian and, at the same time, to reveal various cross-language correspondences and translation patterns.

Because verbal forms are highly frequent, only a subcorpus of the available Russian-French corpus has been processed. 10061 TCs were created in this SCDB. Table 2 below presents all basic types of Russian personal verbal forms that were examined during this study with total number of TCs created for each basic type. All basic types in this project are defined by a combination of tense and aspect (grammatical features) but some also specify specific lexical items that substantially modify function and meaning of verbal forms (lexical features).

No	**basic type**	**Total TCs**	**No**	**basic type**	**Total TCs**
1	past-perfective	3365	9	past-imperfective +*by*	73
2	past-imperfective	2664	10	past-perfective +*chtoby*	36
3	present-imperfective	2356	11	past-perfective +*bylo*	35
4	future-perfective	628	12	past-perfective +*chtoby*	20
5	imperative-perfective	389	13	past-perfective +*esli by*	16
6	imperative-imperfective	212	14	past-imperfective +*esli by*	7
7	past-perfective +*by*	173	15	past-imperfective +*bylo*	7
8	future-imperfective	80			

Table 2: List of Russian basic types in the SCDB of Russian personal verbal forms with total numbers of TCs created for each basic type.

The prevalence of the past tense in the analyzed corpus is due to the fact that it is composed of Russian classical literature pieces for which the past tense is the default register.

Most of the processed texts included multiple translations made by different translators. As a result, in most cases multiple TCs were created for the same LI in the source text. Such TCs were automatically combined into 4135 TCs with multiple translation variants. An example of such a multivariant TC is given in Figure 6 below.

LI (ST)	LI Features	FEFs (Target Texts)		
			FEF	FEF Features
Чего я там **не видал**? –	**Past-IPF** < Neg > < Interrog > < DialRepl >	TT1	Qu'est-ce que j'**en ai à faire**?	**Pres** < SubInf > < Interrog > < DialRepl >
		TT2	Qu'est-ce que j'y **trouverai**?	**Fut** < Interrog > < DialRepl >

Figure 6: An automatically generated TC with multiple translation variants
(lit., ST: *'What haven't I seen there?'*; TT1: *'What am I to do [there]?'*; TT2: *'What will I find there?'*).

A part of this database is publicly available at: http://a179.ipi.ac.ru/corpora_dynasty/main.aspx.

4.2 The SCDB of Russian Language-Specific Items

This SCDB was created as part of another project headed by Anna A. Zalizniak. This work was reported in Zalizniak, 2015.

The object of the study underlying this SCDB was the class of Russian lexical LIs that are generally considered as hard to translate into other languages, into French in particular. The principle of partitioning this class into subclasses was not grammatical but lexical – the basic form of a LI in question was assumed as the basic type of this LI. Therefore, the list of basic types in this project was much longer than in the project described above (114 basic types currently) and the basic types themselves were much more diverse. This also caused higher diversity of supplementary features that may be relevant for various basic types of LIs. To facilitate navigation through numerous basic types and supplementary features in this project we had to introduce clusters of basic types and supplementary features.

This SCDB also introduced creation of French-Russian TCs along with Russian-French TCs. While Russian-French TCs allowed to investigate translation models (ways how a certain LI may be translated into another language), French-Russian TCs allowed to investigate translation stimuli (which items in another language may have caused a certain LI to appear in a Russian translation, see Zalizniak et al., 2013: 102). The latter is an especially interesting question in respect to Russian language-specific items which are assumed not to have any close equivalents in another language.

4198 Russian-French TCs and 1000 French-Russian TCs have been created in the SCDB.

In many cases the Russian-French corpus of the RNC was not sufficiently large for this study – there were too few examples for many of the investigated LIs.

4.3 The SCDB of Russian Connectors

This SCDB was created as part of a project headed by Olga Inkova. This work was reported in Zatsman, Inkova, Kruzhkov & Popkova, 2016.

The study behind this SCDB investigates translation of Russian connectors into French and besides it pays special attention to investigation of internal structure of Russian connectors. A connector is a linguistic item whose function is to establish a certain type of relation between two (and

sometimes more) predicative elements (for more detail see Inkova, 2016). Furthermore, complex connectors may be viewed as combinations of components that can in turn be simple or complex connectors or parts of connectors (compare, for example, various Russian connectors that contain *ne tol'ko* [*'not only'*]: *ne tol'ko… no* [*'not only… but'*]; *ne tol'ko… no i* [*'not only… but also'*]; *ne tol'ko… no dazhe* [*'not only… but even'*]; *ne tol'ko… no prosto* [*'not only… but simply'*]; *ne tol'ko… no eshche* [*'not only… but in addition'*]; *ne tol'ko… no i voobshche* [*'not only… but also quite*]; *ne tol'ko… no, naprotiv* [*'not only… but, to the contrary'*]; etc.).

To make it possible to account for internal structure of connectors during the annotation process, this SCDB first introduced second-level classification of basic types (which in this SCDB are called *discourse realizations*). Now it is possible to simultaneously assign discourse realizations to multiple clusters. For example, in this way discourse realization *'not only… but, to the contrary'* can be simultaneously assigned to clusters <*not only*>, <*but*> and <*to the contrary*>. This allows researchers to later trace translation correspondences not only at the level of actual discourse realization, but also at the level of clusters they represent.

In addition, in this SCDB TCs of different types were created. TCs of Type 1 recorded information on translations of complex connectors taken as a whole; TCs of Type 2 recorded information on translations of each of the two main elements of two-part correlative connectors; TCs of Type 3 recorded information on translations of other parts of connectors.

Finally, this SCDB first introduced creation of TCs for machine translation tools that are available online. By analyzing translations of connectors made by various translation engines and contrasting them to translations made by professional human translators linguists reveal systematic errors of translation engines related to translation of connectors.

As of the moment of writing this paper, 1565 TCs have been created in this SCDB.

Once again, in some cases the current size of the Russian-French corpus of the RNC was not sufficient for this study – some of the connectors that the linguists aim to investigate are not very frequent. To that end, we actively contribute to expansion of the Russian-French parallel corpus of the RNC.

4.4 The SCDB of Russian Impersonal Verbal Forms

This SCDB was created as part of the project on impersonal constructions headed by Anna A. Zalizniak. The underlying study aims to investigate Russian impersonal verbal forms (those without a canonical subject) and also impersonal forms in the broad sense, including indefinite-personal verbal forms and generalized-personal verbal forms.

One of the issues related to creation of this SCDB is the problem of finding a right balance between acceptable rates of noise and losses while searching for impersonal forms in the corpus. To locate a potential impersonal form we have to find a verb in second or third person without a noun or a pronoun in the immediate surrounding that agrees with that verb. However the problem is that in the RNC homonymy is not automatically resolved – each word form is assigned multiple sets of morphological properties that may be associated with it, which leads to both noise in search queries (false hits for verbs) and losses in search queries (false hits for nouns and pronouns that we have to exclude). Furthermore, there is no obvious way to tell what should be considered an immediate surrounding of a verb. These problems make creation of such search queries a very intricate task. At the same time, the apparent complexity of location of impersonal forms in the corpus emphasizes value of SCDBs since they allow to preserve results of manual linguistic analysis.

As of the moment of writing this paper, only 118 TCs have been created in this SCDB, which means that this project is still in a very early stage and the lists of basic types of LIs and supplementary features are still being developed. This preliminary analytical stage is a very important task as it lays ground for future work and eventually affects the value of the SCDB to the linguists who are to analyze the data array to be created. Once this preliminary stage is finished and guidelines

for creation of TCs are finalized, the process of creation of new TC in the SCDB will significantly accelerate.

5 Search and Statistics

Annotation of LIs and TCs in SCDBs makes it possible to quickly find LIs and TCs that match specific criteria and appropriate search functions have been developed in SCDBs. The developed search engine allows users to run search queries by specifying such properties as:

- basic types of LIs;
- basic types of FEFs;
- basic type clusters of LIs and FEFs;
- supplementary features of LIs;
- supplementary features of FEFs;
- TC-relevant features;
- title of the original text and name of translator (for texts with multiple translations);
- lemmas of words in LIs or FEFs, etc.

All these properties may be specified (or excluded) in a query at the same time and the returned set of TCs will match all of the specified criteria. For example, in the SCDB of Russian personal verbal forms a user can specify (1) 'past-perfective' as the basic type of LI (2) <subordinate imperfective infinitive> as a supplementary feature of LI (3) 'imparfait' as the basic type of FEF (4) Russian lemma "*stat*"' ('*to become*') in one of the words in a LI in the source text. This query returns 12 TCs (from the total array of 10061 TCs), a fragment of the results is presented in Figure 7 below. Each of the found TCs may be reviewed in more detail in a separate window.

LI & context	LI features	FEF & context	FEF features
стали [...] колени **подгибаться**	**Past-PF** < SubInf-IPF >	ses genoux **se pliaient**	**Imparf**
Он упорно **стал смотреть** налево, в другую сторону	**Past-PF** < SubInf-IPF >	Son **regard**, **qu'il dirigeait** obstinément vers la gauche, de l'autre côté	**Imparf** < SubAttr >
и у него **стал пропадать** свой голос	**Past-PF** < SubInf-IPF >	Ilia Ilitch, [...] lui aussi **perdait** sa voix	**Imparf**
Старые служаки, [...] **стали исчезать**	**Past-PF** < SubInf-IPF >	Les fonctionnaires de la vieille école, [...] **commençaient à disparaître**	**Imparf** < SubInf >

Figure 7: A fragment of the search results returned by the above-specified query from the SCDB of Russian personal verbal forms. 4 of 12 TCs are displayed.

SCDBs can also be used to quickly provide statistics on various translation patterns for the analyzed LIs, based on data about all TCs created in a given SCDB. Of course one should be careful about making inferences from the raw TC statistics because it could be affected by various factors, such as composition of the analyzed corpus, methods of annotation, etc. That's why we provided an

option for qualitative verification of the data. Researchers can trace the numbers right back to the TCs they are based on to analyze whether the observed translation patterns may be considered regular or accidental, general or genre-specific.

Table 3 shows statistics for translation patterns for Russian basic type '*present-imperfective*', commonly known as *simple present tense* (SPT), based on data about 2176 TCs in the SCDB of Russian personal verbal forms. Note that some of the listed patterns have not been previously described in Russian-French contrastive grammars (column 'Status', for more detail see Zatsman & Buntman, 2015: 856-858).

No.	Translation pattern (LI–FEF)	TC count in the SCDB	Status
1.	**SPT– (présent)**	**1587**	*known*
2.	**SPT– (imparfait)**	**328**	*known*
3.	**SPT– (infinitif)**	**71**	*known*
4.	**SPT– (passé composé)**	**30**	*known*
5.	**SPT– (conditionnel présent)**	**23**	*new*
6.	**SPT– (participe passé)**	**22**	*new*
7.	**SPT– (subjonctif présent)**	**19**	*known*
8.	**SPT– (futur simple)**	**19**	*known*
9.	**SPT– (participe présent)**	**19**	*new*
10.	**SPT– (gérondif)**	**15**	*known*
11.	**SPT– (futur immédiat)**	**10**	*known*
12.	**SPT– (passé simple)**	**10**	*new*
13.	**SPT– (plus-que-parfait)**	**8**	*new*
14.	**SPT– (subjonctif imparfait)**	**6**	*new*
15.	**SPT– (impératif)**	**5**	*known*
16.	**SPT– (infinitif passé)**	**3**	*new*
17.	**SPT– (passé immédiat)**	**1**	*new*
	Total for SPT	2176	

Table 3: Translation patterns for Russian simple present tense (SPT) based on data from the SCDB of Russian personal verbal forms. TCs with omitted or unclear translations have been excluded from the statistics.

6 Conclusion

Statistical and search functions provided by SCDB help researchers to accomplish what corpora where originally meant to do – "make visible patterns which were only, if at all, dimly suspected" (Stubbs, 2002: 221). With traditional parallel corpora researchers could trace these patterns at the level of words whereas with SCDBs it is now possible to trace them at the level of specific linguistic items that can be annotated according to the classification system developed by the researchers themselves taking into account the subject, goals and scope of their research.

There are some other corpus tools that also enable manual annotation of linguistic items. The most advanced tool for computer-assisted manual annotation of corpora known to us is UAM Corpus Tool (O'Donnell, 2008). Still, none of such tools can be applied to parallel corpora, which makes SCDBs unique. In the Table 4 below we provide comparison of some of the most important characteristics of UAM Corpus Tool and SCDB.

UAM Corpus Tool	**SCDB**
Allows to annotate monolingual corpora.	Allows to annotate parallel corpora.
Allows to annotate multi-word linguistic items, but only as uninterrupted segments of text.	Allows to annotate multi-word linguistic items as sets of words that can be either adjacent or separated by other words (e.g. French connector ***non seulement*** *le visage* ***mais aussi*** *l'âme*).
No additional annotation of components of linguistic items is possible.	It is possible to annotate components of linguistic items, including specification of main words, functionally significant words and immediate context.
Open source tool based on open data formats, free for download and use.	Proprietary tool, not publicly available, requires a server database to operate.
Only one user can work with a project at a given time.	Several researchers can work on a project simultaneously allowing for geographic distribution of a project team.
Supports creation of a specialized classification system for annotation of linguistic items.	Also supports creation of a specialized classification system for annotation of linguistic items.

Table 4: Comparison of some of characteristics of UAM Corpus Tool and SCDBs.

The SCDB concept presented in this paper has become the basic framework for solving the following problems:

1. Development of computer-assisted methods for manual annotation of parallel corpora for both local and geographically distributed teams of linguists.
2. Extraction of new cross-linguistic knowledge (both grammatical and lexical) based on data from parallel corpora.
3. Pinpointing significant gaps in existing contrastive grammars and filling these gaps based on data extracted from parallel corpora with help of SCDBs.

SCDBs have proved their effectiveness in dealing with these problems – therefore we can safely state that the SCDB concept provides a solid foundation for extending the line of corpus-based tools for conducting contrastive linguistic analysis.

References

Dobrovolsky D. O., Kretov A. A., Sharoff S. A. (2005) Corpus of parallel texts: Architecture and applications / Korpus parallel'nykh tekstov: Arkhitektura i vozmozhnosti ispol'zovaniya. In: *Russian National Corpus: 2003–2005 / Natsional'nyy korpus russkogo yazyka: 2003-2005*. Moscow: Indrik, 263-296.

Hoffmann, S., Evert, S. (2006). BNCWeb (CQP-Edition): The Marriage of Two Corpus Tools. In: Braun. S., Kohn. K., Mukherjee J. (eds) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods.* (English Corpus Linguistics 3.) Frankfurt am Main: Peter Lang, 177-195.

Inkova, O. Yu. (2016). On the problem of description of multiword connectors of Russian language: ne tol'ko… no i (not only… but also) / K probleme opisaniya mnogokomponentnykh konnektorov russkogo yazyka. In: *Voprosy yazykoznaniya*. Vol.2, 37-60.

Johansson S. (2007). *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. – Amsterdam: John Benjamins.

Kruzhkov, M. G. (2015). Information resources for contrastive linguistic studies: electronic text corpora. / Informatsionnye resursy kontrastivnykh lingvisticheskikh issledovaniy: elektronnye korpusa tekstov. In: *Systems and Means of Informatics*. Vol. 25(2), 140-159.

Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M. (2013). Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus. In: *Informatics and Applications*. Vol. 7(2), 100-109.

McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

O'Donnell, M. (2008). "The UAM CorpusTool: Software for corpus annotation and exploration". In: Bretones Callejas, Carmen M. et al. (eds) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, 1433-1447

Stubbs, M. (2002). *Words and Phrases. Corpus studies of lexical semantics.* Oxford: Blackwell.

Zalizniak Anna A., Sitchinava D. V., Loiseau S., Kruzhkov M., Zatsman I. M. (2013). Database of Equivalent Verbal Forms in a Russian-French Multivariant Parallel Corpus. In: *2013 International Conference on Artificial Intelligence (ICAI'13)*, Vol. 1. – CRSEA Press, Las Vegas, 101-107.

Zalizniak Anna A. (2015). Russian Language-specific Words as an Object of Contrastive Corpus Analysis / Lingvospetsifichnye edinitsy russkogo yazyka v svete kontrastivnogo korpusnogo analiza. In: *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*. Vol. 14(1). – Moscow: RGGU, 683-695.

Zatsman I. M., Inkova O. Yu., Kruzhkov M. G., Popkova N. A. (2016). Representation of Linguistic Knowledge in Suprocorpora Databases / Predstavlenie krossyazykovykh znaniy o konnektorakh v nadkorpusnykh bazakh dannykh. In: *Informatics and Applications*. Vol. 10(1), 106-118.

Zatsman I., Buntman N. (2015). Outlining Goals for Discovering New Knowledge and Computerised Tracing of Emerging Meanings Discovery. In: *Proceedings of the 16th European Conference on Knowledge Management*. – Reading: Academic Publishing International Limited, 851-860.