The PERTOMed Project: Exploiting and validating terminological resources of comparable Russian-French-English corpora within pharmacovigilance.

Cedric Bousquet, Maria Zimina

The PERTOMed Project:
Exploiting and validating terminological resources of comparable Russian-French-English corpora within Pharmacovigilance

Cedric Bousquet
INSERM UMR_S 872, Eq 20 (Faculté de Médecine - Paris 5)
cedric.bousquet@spim.jussieu.fr

Maria Zimina
CRIM-INaLCO (Paris) / EA2290 SYLED (Paris 3)
zimina@msh-paris.fr

Abstract:

The PERTOMed project is a pluridisciplinary research initiative undertaken by several institutions in France. Applications considered within the part of the project described in this article concern pharmacovigilance and adverse drug reactions. We had several objectives: to create a specialised Russian Internet corpus; to test new tools and methods for term extraction from comparable multilingual texts and to build terminological resources including Russian. Trilingual Russian-French-English lexicon resulting form this work is freely available from the PERTOMed server.

Key-words: adverse drug reactions, alignment, comparable corpora, corpus linguistics, medical texts, natural language processing, parallel corpora, pharmacovigilance, Russian, term extraction.

1. Introduction: the PERTOMed project

Our study presents the results of a part of a collective research project on production and evaluation of terminological and ontological resources in the medical field – PERTOMed (in French: Production et Evaluation de Ressources Terminologiques et Ontologiques dans le domaine de la MEDecine) [Charlet *et al.,* 2006].

Development of terminological resources in medicine is a major issue to allow collecting data and browsing knowledge databases. In this respect, one of the main objectives of the project (officially finished in December 2005) was to explore *Natural Language Processing* (NLP) tools and methodologies for compiling terminological resources from parallel and comparable medical texts in several languages. Terminology extraction from texts was then considered to

design better tools to help specialists coding acts and diagnoses with an ontology-based software [Baneyx *et al*., 2007].

Potential applications considered within PERTOMed covered several fields:

- Pharmacovigilance
- Pneumology
- Drug-drug interactions
- Multilingual terminology management.

The project resulted in cross-field collaborations, critical thinking and exchanges among linguists, specialists of natural language processing and knowledge engineers, information scientists and medical practitioners. Several research institutions contributed to PERTOMed:

- INSERM UMR_S 872, Eq 20, Faculté de Médecine - Paris 5 (France).
- ERSS: Equipe de Recherche en Syntaxe et Sémantique, UMR 5610 CNRS and Toulouse le Mirail University (France).
- CRIM: Centre de Recherche en Ingénierie Multilingue, INaLCO (Paris, France).

Overall project management and coordination were performed by the INSERM research team under the direction of M.-Ch. Jaulent and J. Charlet.

As a part of our contribution to the project, we conducted a series of experiments in order to develop, explore and evaluate terminological resources of comparable Russian-French-English medical text corpora sharing common semantic and pragmatic characteristics within pharmacovigilance. According to World Health Organization (WHO), pharmacovigilance is *"the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems"* [WHO, 2002].

We pursued three objectives for multilingual terminological study using Russian:

- Creation of Russian Internet corpus within Pharmacovigilance;
- Study of methods for building terminologies from comparable corpora;
- Elaboration of French-English-Russian adverse reaction terminological resource.

The trilingual Russian-French-English lexicon resulting from this work is available on the

Web from the PERTOMed server.[1]

Multilingual terminology management within PERTOMed turned out to be quite a challenge from the point of view of research methodology, tools and software. In this article, we focus on some essential results of this work, methodological findings and suggestions for further research, keeping in mind the fact that many new projects faced with disparities across multilingual medical texts are most likely to come up.

2. Multilingual terminology management within Pharmacovigilance

2.1. Pharmacovigilance

In a most general sense, practical pharmacovigilance involves collecting, monitoring, researching, assessing and evaluating information from healthcare providers and patients on the adverse effects of medications. This pharmacological science is particularly concerned with Adverse Drug Reactions (ADRs). According to World Health Organization, ADR means a response to a drug which is noxious and unintended, and which occurs at doses normally used for the prophylaxis, diagnosis or therapy of disease, or for the modification of physiological function [WHO, 1972].

Pharmacovigilance plays an increasingly important role as the number of drug recalls is growing rapidly. Clinical trials usually involve limited study populations and might be insufficient to detect possible side effects and ADRs at the time a drug enters the market. For this reason, research tools and activities within Pharmacovigilance (including data mining and investigation of case reports) are essential to identify possible relationships between drugs and ADRs.

2.2. International terminologies within Pharmacovigilance

Since reports on side effects and ADRs contain a wide variety of domain specific terms, maintaining international terminological standards for pharmacovigilance is of great relevance. However, due to the existence of well-known historic and cultural boundaries between countries, building terminological resources world-wide is not a trivial problem. For the moment, two international terminological resources are widely acknowledged in this field:

[1] The PERTOMed web server is currently moving to a permanent location : http://pertomed.spim.jussieu.fr/pertomed/

World Health Organization – Adverse Reaction Terminology (WHO-ART) and *Medical Dictionary for Drug regulatory Activities* (MedDRA).

WHO-ART has been developed over more than thirty years for coding adverse reaction terms in relation to drug therapy. It is widely used by drug regulatory agencies and pharmaceutical manufacturers in many countries. WHO-ART is initially developed in English with more or less complete translations into French, German, Spanish, Portuguese and Italian. The system is maintained by the WHO Collaborating Centre for International Drug Monitoring, Uppsala Monitoring Centre (UMC).

MedDRA defines fully equivalent medical terms in different languages, including English, French, German, Japanese and Spanish. This international terminology applies to almost all stages of drug development, as well as health effects and malfunction of devices. MedDRA is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA).

2.3. Pharmacy-related issues in Russia

Russian pharmaceutical industry and distribution network has undergone radical transformations after the break-up of the USSR. Rapid political changes and transition to a market economy resulted in massive disruption to pharmaceutical production and distortion of cooperative links between former partners. Today, many efforts are still required to establish countrywide activities as pharmaceutical distribution in Russia remains fragmented with thousands of pharmaceutical wholesalers. Only few of them offer nationwide coverage.

In Russia, all drugs and biological products must be registered with the Ministry of Health at the federal level. The federal and regional governments are then entitled to develop and use lists of essential drugs and provide recommendations for supply and use in public health system. Enterprises and insurance companies develop their own versions. Thus, many lists of essential drugs have been developed in parallel based on several criteria, such as efficacy, safety, price (as well as origin and production characteristics).

In 1997, Russian Federation established a federal centre for monitoring adverse reactions which has joined adverse drug reaction monitoring programme within WHO. For the moment, WHO-ART is not translated into Russian. Terminological resources in Russian are available within *International Classification of Diseases* (ICD), currently being adopted in Russia to classify diseases and health problems on many types of medical records.

Progressive integration of Russian Federation and Russian language into World Health system is an active process. At the international level, there exist medical terminologies including Russian, such as Russian translation of *Medical Subject Headings* (MeSH) in at least two versions developed by different institutions.[2] On a national level, translation of terminological resources is in process for healthcare disciplines and medical sciences in general.[3]

3. Parallel *vs.* comparable text processing: state of the art

In the past few years, many efficient tools and methods have been developed for bilingual lexicon extraction from parallel corpora (source texts accompanied by their translations in one or several languages) [Véronis, 2000]. It is common knowledge already that parallel corpora are not always available for specific domains/language pairs. Nowadays, the objective is to exploit the richness of non-parallel yet comparable corpora existing in almost any field of knowledge.

According to Fung and McKeown [1997], a rather problematic task of bilingual lexicon extraction from non-parallel corpora can be sorted out by statistical study of context word similarity between candidate terms, representing mutual translation pairs. Following this direction, most of the work in this field is done comparing the distributional contexts of source and target words, testing several weighting factors and similarity measures [Chiao and Zweigenbaum, 2002]. This approach relies upon existing bilingual resources ('base lexicon' or 'pivot translations') used to calculate translation similarities between source and target words' context vectors [Sadat *et al*., 2002]; [Morin and Daille, 2004].

Developments within comparable text processing motivate new experiments and research initiatives aiming at terminology extraction from multilingual texts coming from several cultural and linguistic sources.

---

[2] MeSH offers controlled vocabulary and thesaurus features used to index, catalogue and retrieve the world's medical literature. It has been translated into Russian by the US *National Library of Medicine* (NLM), Russian National Public Library for Science and Technology (CYRMESH integrated within automated library system IRBIS).

[3] English translation is provided on the following Russian web sites: *Antibiotics and Antimicrobial Therapy*: http://antibiotic.ru/index.php?newlang=eng and *Interregional Association for Clinical Microbiology and Antimicrobial Chemotherapy*: http://www.iacmac.ru/iacmac/en/

4. The trilingual PERTOMed corpus on adverse drug reactions

For the PERTOMed project, parallel English/French translation texts of *Summaries of Product Characteristics* (SPC) were a good starting point for ADR terminological study. As for Russian, only comparable texts are currently available in the field of Pharmacovigilance. Moreover, Russian pharmaceutical documentation is of heterogeneous nature. Lack of visibility and decentralisation creates difficulties in terms of localisation of high-quality authentic medical data in Russia (comparable to SPCs).

4.1. Parallel English/French sub-corpus

The English/French part of the PERTOMed corpus was composed out of 156 SPC downloaded in PDF format from the EMEA web site by the INSERM team in February 2004.

A SPC is a special type of medical text with a description of a certain medicinal product's properties and the conditions attached to its use. This document is intended for product certification performed by European Medicines Agency (EMEA)[4] and for medical professionals. The SPCs are provided in all EU languages (without official Russian translation).

Figure 1 shows an extract of the English/French sub-corpus from the PERTOMed project. Only two SPC sections (*Section 4.5* and *Section 4.8*) containing information on drug-drug interactions and undesirable effects, were used for the corpus.

4.2. Comparable Russian sub-corpus

Localisation and data retrieval for Russian were carried out by the CRIM-INaLCO team [Ivanova et Nuk, 2005]. It was first necessary to establish a list of available pharmaceutical resources on the Russian web and to develop a set of criteria for corpus construction.

Sampling frame

In corpus linguistics studies, a comparable multilingual corpus is usually defined as a corpus containing components that are collected using the same sampling frame, similar balance and representativeness. The components representing the languages involved must match with each other in terms of proportion, genre, domain and sampling period [McEnery and Xiao,

[4] *European Medicines Agency* (EMEA) is a decentralised EU body with headquarters in London. The EMEA issues certificates of a medicinal product in conformity with the arrangements laid down by the World Health Organisation.

2005].

In order to collect comparable medical data for the PERTOMed corpus, special attention was paid to the following characteristics of Russian pharmaceutical texts:

- Degree of specialisation
- Recognition by domain experts in Russia
- Style (summarization)
- Clarity and precision
- Information granularity
- Availability of active component and product name indexation.[5]

LAMIVUDINE

English	French
Lamivudine may inhibit the intracellular phosphorylation of zalcitabine when the two medicinal products are used concurrently.	La Lamivudine peut inhiber la phosphorylation intracellulaire de la zalcitabine lorsque ces deux produits sont administrés de manière concomitante.
Zeffix is therefore not recommended to be used in combination with zalcitabine.	Par conséquent, il n'est pas recommandé d'utiliser Zeffix en association avec la zalcitabine.
In clinical studies of patients with chronic hepatitis B, Lamivudine was well tolerated.	La Lamivudine a été bien tolérée au cours des essais cliniques réalisés chez des patients atteints d'hépatite B chronique.
The most common adverse events reported were malaise and fatigue, respiratory tract infections, throat and tonsil discomfort, headache, abdominal discomfort and pain, nausea, vomiting and diarrhoea./…/	Les effets indésirables le plus souvent rapportés étaient : malaise et fatigue, infections respiratoires, gêne au niveau de la gorge et des amygdales, céphalées, douleur ou gêne abdominale, nausées, vomissements et diarrhée./…/

Figure 1. Parallel English/French sub-corpus: extract.

Following preliminary research described in [Ivanova et Nuk, 2005], three Russian web sites were selected for the project:

[5] Following these principles, the choice of Russian pharmaceutical texts was made by selecting active components identical to those mentioned in English/French SPC sub-corpus.

- RECIPE: http://www.recipe.ru
- РЛС: http://www.rlsnet.ru
- Russian Vidal: http://www.vidal.ru

The **RECIPE** site is an on-line information catalogue devoted to legal pharmacological documentation of Russian Federation. It provides Medline user manual, detailed index of Russian bio-medical web sites and offers possibilities to search for medical products using several criteria (including ICD-10).

The **РЛС** site (Russian acronym of *Регистр Лекарственных Средств России*, in English *Register of Medical Substances of Russia*) is an on-line encyclopaedia of medical products with precise product description.

The **Russian Vidal** web site is maintained and regularly updated by the private company AstraPharmService in accordance with the Industrial Standard of Russian Federation.

These three web sites provide standard medical data acknowledged by legal authorities of Russian Federation and referenced by international pharmaceutical partners present on the Russian market. In order to compile a balanced Russian sub-corpus with a size similar to that of the English/French one, text-to-text alignment was used to achieve best results. For each drug product listed within the English/French SPC part, localisation of corresponding texts in Russian was performed through the following steps:

- Cross-check if the product is commercialised in Russia.
- Search *RECIPE*, *РЛС* and *Russian Vidal* using product name or active component.
- Localise product description pages.
- Extract textual data present within adverse reaction section only.
- Keep all variants in case of several descriptions possible for the same product.

Figure 2 shows the extract of the Russian sub-corpus resulting from this study described in [Ivanova and Nuk, 2005].

<Recipe LAMIVUDINE>

Побочные действия :

Лейкопения, тромбоцитопения, анемия; недомогание, утомляемость, головная боль, периферическая нейропатия, парестезии; гастралгия, тошнота, рвота, диарея, панкреатит; повышение активности «печеночных» трансаминаз, гиперкреатининемия, панкреатит, повышение активности сывороточной амилазы; лихорадка, развитие вторичной инфекции. Передозировка: усиление проявлений описанных побочных действий. Лечение: промывание желудка, активированный уголь, симптоматическая терапия, непрерывный гемодиализ.

<РЛС LAMIVUDINE>

Побочные действия:

Головная боль, головокружение, слабость, нарушение сна, бессонница, гипотимия, нейропатия, кашель, гриппоподобный синдром, анорексия, тошнота, диарея, рвота, боль в эпигастральной области, некротический панкреатит (возможен летальный исход), миалгия, артралгия, лейкопения, анемия, лихорадка, потливость, аллергические реакции.

<Russian Vidal LAMIVUDINE>

Побочное действие :

Со стороны пищеварительной системы: возможны боли и дискомфорт в эпигастральной области, тошнота, рвота, диарея, снижение аппетита, повышение активности печеночных трансаминаз.

Со стороны ЦНС: возможны повышенная утомляемость, головная боль.

Со стороны дыхательной системы: возможны инфекции дыхательных путей.

Прочие: возможно общее недомогание.

Figure 2. Comparable Russian sub-corpus: extract.

Evaluation by Correspondence Analysis

We used correspondence analysis (CA) available within *Lexico3*[6] textometric toolbox to evaluate whether Russian ADR descriptions collected on three different web sites, using a common sampling frame, could form a homogeneous corpus for a terminological study.

CA is a multidimensional descriptive data analysis method used for describing contingency tables (or cross-tabulations) [Lebart *et al*., 1997]. In our case, correspondence analysis was used to describe a lexical table, cross-tabulating wordforms and medical texts on same drug products collected on the *RECIPE*, *РЛС* and *Russian Vidal* web sites. In other words, our

[6] *Lexico3* textometric toolbox presents a wide range of functions (segmentation, concordances, measurements and counts based on graphical forms, computation of characteristic elements and correspondence analyses of forms and repeated segments):
http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/

table had as many rows as there were wordforms in collected Russian texts and as many columns as there were texts describing adverse drug reactions of each drug product. To analyse the information contained in this table, the row-profile and column-profile tables were calculated and the distances among the words on the one hand and drug product adverse reaction descriptions on the other hand were displayed (see *Figure 3*).

Figure 3 shows the *first plane* of the correspondence analysis (that is, the plane of the first two principal axes) showed a high density of groupings of columns-points (texts) in the origin of the axes, displaying important lexical affinities between texts. Only some incomplete or shortened texts departed from the *average profile*, showing common lexical characteristics of the Russian sub-corpus. This dominance of core vocabulary confirmed global terminological homogeneity of texts coming from three different sources (*Recipe*, *РЛС*, *Vidal*). It was a positive result regarding the initial objectives of our terminological study.

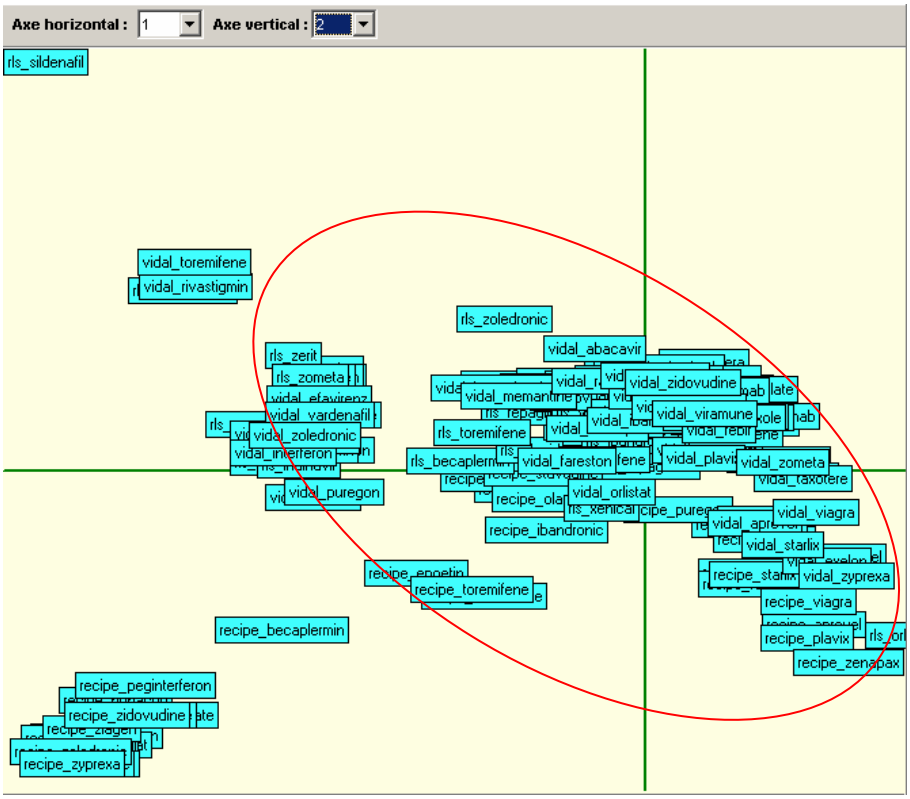

Figure 3. Comparable Russian sub-corpus: first plane of the Correspondence Analysis.

4.3. Main lexicometric characteristics of the trilingual PERTOMed corpus

Figure 4 shows main lexicometric characteristics of the PERTOMed corpus. Each sub-corpus

was subdivided into graphical forms using *Lexico3* textometric toolbox. The results of automatic segmentation are represented in separate columns, showing respectively the total number of occurrences of wordforms (tokens) within each corpus part, the number of different wordforms (types) used, most frequent wordform (maximum frequency) and the total number of hapaxes (*hapax legomenon*, i.e. one-off occurrences of wordforms) in a given sub-corpus.

			Nb occ	*Nb forms*	*F max*	*Hapax*
comparable		**Russian**	15 465	3 034	461 (*в*)	1 483
	parallel	**English**	133 984	5 957	4 936 (*of*)	1 836
		French	161 995	7 280	8 022 (*de*)	2 389

Delimiting characters: .,:;!?/_-\'"()[]{}§$

Figure 4. The trilingual PERTOMed corpus: structure and main lexicometric characteristics.

5. Term alignment in parallel English/French texts using SYNTEX

Term alignment in parallel English/French SPC texts was performed by the ERSS team. These bilingual texts were first aligned at the sentence level using *JAPA*[7] tool. Each word in two parts of the corpus was assigned a lemma and a grammatical tag through Part-Of-Speech (POS) tagging performed by *TreeTagger*[8].

Both parts were then analysed with a dependency parser *SYNTEX* [Bourigault and Fabre, 2000]. Identification of syntactic dependencies (for instance subjects, direct and indirect objects of verbs) was performed independently for each language.

Bilingual term extraction was performed using both statistical word similarity measures (such as *Jaccard's Coefficient*) and *alignment by syntactic propagation* [Ozdowska, 2004]; [Ozdowska *et al*., 2005]. The description of the Bilingual French-English lexicon resulting from this work is available on the PERTOMed web site. This hierarchically structured (type head-expansion) lexicon counts 1 278 validated bilingual terminological units (French term ~ suggested equivalent in English).

---

[7] *JAPA* is a program that aligns parallel texts at the sentence level: http://rali.iro.umontreal.ca/Japa/

[8] *TreeTagger* is a tool for annotating text with part-of-speech and lemma information: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

6. Mapping terms in comparable contexts

6.1. Corpus-based approach adopted within the PERTOMed project

Available tools and methods for processing parallel English/French sub-corpora could not be used to process comparable Russian texts. Lack of specific term-extraction software and resources for Russian[9], made us consider quantitative corpus-based tools and methods (such as computation of word *n-gram* repetitions and *multiple co-occurrences*)[10]. This fine-grained context-based approach allowed flexibility in order to work simultaneously on different language combinations in trilingual contexts.

For testing purposes, term extraction for Russian was first done semi-automatically, using fine-grained segmentation into terminological units performed by the CRIM-INaLCO team [Ivanova and Nuk, 2005]. Information chunks on ADRs collected within Russian Internet corpus were automatically splitted into distinct textual units using common punctuation marks: *comma* [,], *colon* [:], *semi-colon* [;], *dot* [.]. This type of segmentation could be applied due to the nature of adverse reaction texts. These concise summary style texts are chiefly composed of enumerations describing particular groups of undesirable effects (cf. *Figure 2*).

Generated lists of term-candidates were then filtered to keep frequent relevant patterns, such as [Ajective + Noun] [*депрессивное состояние*, in English: *depressed state*] or [Noun$_{_Nominative}$ + Noun$_{_Genitive}$] (*потеря аппетита*, in English: *lost appetite*). These terminological units were put into correspondence with validated *SYNTEX* term candidates for French through manual context-based alignment. The results of these experiments are available within the Russian-French lexicon (485 main terms) on the PERTOMed web site.

The experiments of English/French and Russian/French term alignment gave impetus to start working on a trilingual terminological resource of adverse drug reactions. In order to avoid disjoint terminologies with cross-language boundaries, we decided to attempt at creating a unified *Russian-French-English* adverse reaction terminology.

---

[9] Localisation of tools and resources for Russian NLP was unsuccessful during the project. Since then, new tools and resources have been made available. For instance, *TreeTagger* has been adapted for Russian.

[10] *N-gram* is a sub-sequence of *n* items (letters, wordforms, etc.) from a given text sequence. *Co-occurrence* is a simultaneous, but not necessarily contiguous presence of occurrences of two given wordforms in a fragment of text.

6.2. Textometric browsing

Trilingual terminology creation was managed through textometric browsing in comparable contexts. The concept of textometric browsing enables the researcher to move among the results produced by different quantitative methods and the original corpus [Zimina, 2004]. This interactive method of corpus exploration helps to produce an automatic selection of contexts in one of the parts of the multilingual corpus where any textual unit under study (*wordform*, *n-gram*, *collocation*, etc.) is found.

Anchor points

As a rule, initial anchor points (corresponding textual units) for exploring comparable corpora are set using several criteria. Most of the time, cognates (etymologically related words having a common origin in one or more languages, e.g.: English: *syndrome*, French: *syndrome*, Russian: *синдром*), general language translation correspondences (e.g.: English: *pain*, French: *douleur*, Russian: *боль*) as well as existing translation or dictionary resources contribute to anchor points identification.

In our case, *English-French* and *Russian-French* lexicons (sharing common French part) helped to identify simple translation correspondences to start browsing through trilingual texts. Moreover, in order to prospect for lexical anchors, we relied upon frequency counts produced on corpus *vocabulary* (the entire set of wordforms in each corpus part) and on larger units consisting of several wordforms (*segments*) [Lebart *et al*., 1997].

After setting aside words with a purely grammatical role as well as the term *patient* for English and French, the most frequent wordforms in the PERTOMed corpus were: *disorders* [Freq=835] in English, *troubles* [Freq=922] in French and *стороны* [Freq=216] in Russian (Genitive of *сторона*, meaning *side* in English). These lexical items were selected as anchors to automatically recognize corresponding contexts in three languages. We used *repeated segments* and *multiple co-occurrences networks* to build context vectors.

	Repeated segments	*Freq*
RU	со ***стороны***	216
	со ***стороны*** сердечно сосудистой системы	28
	со ***стороны*** органов	27
	со ***стороны*** пищеварительной системы	19
	со ***стороны*** органов жкт	18
	побочные действия со ***стороны***	17
	со ***стороны*** цнс	16
	со ***стороны*** нервной системы	15
	со ***стороны*** системы	14
	побочное действие со ***стороны***	13
	со ***стороны*** нервной системы и органов чувств	12
	со ***стороны*** системы кроветворения	12
	со ***стороны*** сердечно сосудистой системы и крови	11
	со ***стороны*** кожных покровов	10
	со ***стороны*** опорно двигательного аппарата	10
	/…/	
FR	***troubles*** du	192
	troubles du système	130
	troubles de	116
	troubles de la	81
	troubles du système nerveux	79
	troubles gastro intestinaux	76
	troubles généraux	76
	troubles cutanés	47
	troubles psychiatriques	46
	troubles du métabolisme et de	42
	troubles généraux et	42
	troubles du métabolisme et de la nutrition	40
	troubles respiratoires	38
	troubles cutanés et	36
	troubles vasculaires	36
	/ …/	
EN	system ***disorders***	178
	disorders common	160
	disorders very	96
	general ***disorders***	84
	nervous system ***disorders***	77
	gastrointestinal ***disorders***	69
	disorders uncommon	68
	disorders very rare	52
	disorders and	50
	tissue ***disorders***	49
	general ***disorders*** and	46
	disorders rare	45
	general ***disorders*** and administration	45
	disorders very common	44
	subcutaneous tissue ***disorders***	44
	/…/	

Figure 5: 15 most frequent repeated segments around the pivotal terms *стороны/troubles/disorders*.

Repeated Segments extraction with *Lexico3*

A *repeated segment* (RS) is a series of consecutive wordforms whose frequency is greater than or equal to 2 in the corpus [Lebart *et al*., 1997]; [Lamalle *et al*., 2005]. Figure 5 shows an excerpt from the repeated segments inventories in Russian, French and English around the pivotal terms *стороны/troubles/disorders*. We noted that even if these terms would never be suggested as translation correspondences in bilingual French/Russian or English/Russian dictionaries, they are positioned preferentially in equivalent contexts (see *Figure 5*).

Multiple co-occurrences networks computation with *COOCS*

Procedures that select repeated segments in a corpus are not yet able to find repetitions that are slightly altered by minor lexical modifications of one of the components. Given a pivotal wordform, several methods can be used to select the set of wordforms that tend to appear often in the neighbourhood of this word. In order to select these words, a unit of context or neighbourhood must be chosen, within which two words are considered to be co-occurring. For example, this unit can be similar to a sentence [Lebart *et al*., 1997].

Martinez [2005] proposed an original method of defining the lexical universe of a given pivotal word based on iterative calculation of lexical attractions: *multiple co-occurrences* networks. In our study, we used the tool *COOCS* resulting from his work in order to discover for each pivotal word (anchor point) a network of words that are positioned preferentially in the same sentences. A detailed presentation of this research method for the identification of translation correspondences can be found in [Martinez and Zimina, 2002].

For comparable texts, the statistical study of the intensity of lexical relations through collocation allowed building context vectors and mapping terminological equivalents within the neighbourhood of corresponding lexical anchors (see *Figure 6*).

Pivot	*Context vectors*
RU ***синдром*** Freq=43	*→ гриппоподобный → озноб → лихорадка*
	→ джонсона → стивенса
	→ токсический
	→ болевой
	→ острый → респираторный → дистресс
	→ лайелла
	→ лизиса
FR ***syndrome*** Freq=145	*→ grippal → pseudo*
	→ fièvre → frissons
	→ johnson → stevens → lyell
	→ johnson → stevens → érythème → multiforme
	→ respiratoire → détresse
EN ***syndrome*** Freq=117	*→ johnson → stevens → multiforme → erythema → epidermal → necrolysis → toxic*
	→ flu → chills
	→ flu → like
	→ respiratory → distress
	→ lupus
	→ fever

Figure 6: Lexical networks around the pivot *синдром/syndrome/syndrome* showing corresponding context vectors (extract).

7. Choosing terms and domains: collaboration domain expert/corpus linguist

Two types of human knowledge were necessary to succeed in creation of our terminological resource: methodological knowledge on text processing from corpus linguist and domain-specific knowledge on adverse drug reactions from domain expert. The details of this collaboration for terminology management process are presented in *Figure 7*.

Corpus linguist provided methodological knowledge on tools and methods for text exploration as well as quantitative results on corpora. The role of domain expert was to choose and validate relevant terms in case of several variants attested in texts (see *Figure 8*). We also used WHO-ART terminology to check English-French term correspondences.

Task	*Tools, Methods and Resources*
Automatic segmentation into textual units (wordforms, repeated segments)	***Lexico3*** [Lamalle *et al*., 2004]
Identification of trilingual lexical anchors (starting points)	Frequency counts, cognates, general language correspondences, existing bilingual lexicons
Computation of trilingual collocation networks	***COOCS*** [Martinez, 2005]
Identification of similar context vectors	Domain expert / corpus linguist
Semi-automatic segmentation into terminological units	Corpus linguist
Cross-language check	Corpus linguist
Validation of trilingual terminological records	Domain expert
Attribution of domains (organ classes)	Domain expert / corpus linguist
Final resource validation	Domain expert

Figure 7: Management of trilingual ADR terminology creation.

Terminological variants in the PERTOMed corpus		
RU	***гриппоподобный синдром* (13 occurrences)**	☑
	гриппоподобные симптомы (2 occurrences)	
	симптомы гриппоподобного синдрома (1 occurrence)	
	гриппоподобная симптоматика (1 occurrences)	
FR	***syndrome pseudo-grippal* (23 occurrences)**	☑
	syndrome pseudogrippal (2 occurrences)	
	symptômes pseudo-grippaux (7 occurrences)	
EN	***influenza-like symptoms* (9 occurrences)**	☑
	influenza-like illness (6 occurrences)	
	flu-like symptoms (8 occurrences)	
	***flu-like symptom* (4 occurrences)**	☑
	flu-like syndrome (6 occurrences)	
	flu-like illness (5 occurrences)	

Figure 8: Choosing terms through collaboration domain expert/corpus linguist.

The role of domain expert was essential to attribute a particular organ class (domain) for each terminological record (see *Figure 9*).

8. Results: Russian-French-English lexicon of adverse reaction terms

8.1. Qualities

Russian-French-English ADR terminological resource resulting from the project is freely available on the PERTOMed web site: http://pertomed.spim.jussieu.fr/pertomed/

This base lexicon comprises 430 validated trilingual terminological entries in XML format. These entries are structured by 2002 simple terminological records (single word terms) and 1006 complex terminological records (multi-word terms). Accordingly, approximately 50% of the records are complex terms. Each trilingual entry comprises the following fields (see *Figure 9*):

- Simple term (with possible variants).
- Abbreviation (if applicable).
- Related composed term(s).
- Domain(s).
- Drug product(s) concerned.

As shown in *Figure 9*, the structure of the lexicon preserves co-occurrence relations between terms. For example, terminological record *diabetic coma* is listed under two simple main terms *coma* and *diabetes*.

8.2. Limits

For the moment, the trilingual lexicon is a structured list of Russian-French-English terms. It lacks visual aids for navigation and text look-up facilities. This work is still to be done following experiments described in Zimina [2004].

We faced serious difficulties concerning the evaluation of Russian-French-English terminology. The choice of criteria was not straightforward. On the one hand, we had to keep in mind the existence of well-established international terminological standards for English and French, on the other hand, it was important to preserve the specificity of original Russian terms coming from authentic medical texts. Further steps should be taken in this direction in order to think of a balanced approach of this complex methodological issue.

Simple term:	**Simple term:**
диабет *diabète* *diabetes*	*кома* *coma* *coma*
Context:	**Context:**
диабетическая кома *coma diabétique* *diabetic coma*	*диабетическая кома* *coma diabétique* *diabetic coma*
Domain:	**Domain:**
Обмен веществ *Troubles du métabolisme* *Metabolism disorders*	*Обмен веществ* *Troubles du métabolisme* *Metabolism disorders*
Medical product:	**Medical product:**
РЛС: *OLANZAPINE*	РЛС: *OLANZAPINE*

Figure 9: Trilingual terminological entries with complex term co-occurrence.

9. Conclusions and future work

Creating terminological resources from comparable corpora is faced with intrinsic heterogeneity of texts. Following our experiments within the PERTOMed project, we are convinced that the challenge of exploring texts coming from different cultural and linguistic sources should be taken into account in the terminology project feasibility study.

We hope that our exploratory work on creation of Russian Internet corpus in the field of pharmacovigilance will be followed by new research initiatives in order to collect and explore authentic terminological resources in Russian.

The use of textometric browsing for comparable text processing and term extraction gives encouraging results. Suggested methods for Russian corpus exploration should be improved taking into account the availability of new resources for processing Russian medical texts.

Bibliography

Web sites:

World Health Organisation Regional Office for Europe: http://www.euro.who.int/

European Medicines Agency: http://www.emea.eu.int/

Uppsala Monitoring Centre: http://www.who-umc.org/

Medical Dictionary for Drug Regulatory Activities: http://www.meddramsso.com/MSSOWeb/index.htm

PERTOMed Project: http://pertomed.spim.jussieu.fr/pertomed/

RECIPE: http://www.recipe.ru

РЛС: http://www.rlsnet.ru

Russian Vidal: http://www.vidal.ru

References:

Baneyx A., Charlet J., Jaulent M.-C. (2007) "Building an ontology of pulmonary diseases with natural language processing tools using textual corpora." *International Journal of Medical Informatics* n°76, pp. 208–215.

Bourigault D., Fabre C. (2000) "Approche linguistique pour l'analyse syntaxique de corpus." *Cahiers de Grammaire* n°25, pp. 131-151, Université Toulouse le Mirail. http://w3.univ-tlse2.fr/erss/textes/pagespersos/cfabre/articles/Bourigault-et-FabreCG00.pdf

Charlet J., Jaulent M.-C., Slodzian M., Bourigault D., Baneyx A., Bousquet C., Mille F., Ozdowska S., Zimina M. (2006) "Pertomed : Production et évaluation de ressources terminologiques et ontologiques dans le domaine de la médecine." *Rapport final*. INSERM U729.

Chiao Y.-Ch., Zweigenbaum P. (2002) "Looking for candidate translational equivalents in specialized, comparable corpora." In Proceedings of *COLING'02*, Taipei, pp. 1208-1212. http://www-test.biomath.jussieu.fr/~pz/FTPapiers/Chiao:COLING2002.pdf

Fung P., McKeown K. (1997) "Finding terminology translations from non-parallel corpora." In *Proceedings of the 5th Annual Workshop on Very Large Copora*, Hong Kong, pp. 192–202. http://www.ece.ust.hk/~pascale/Publications/conference/1997/WVLC-1997.pdf

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A., Lande B., Maisondieu A., Poirot-Zimina M. (2004) *Lexico3 Textometric toolbox User's manual*. SYLED-CLA2T, Université de la Sorbonne nouvelle – Paris 3. http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuelsL3/L3-usermanual.pdf

Lebart L., Salem A., Berry L. (1997) *Exploring Textual Data*. Boston: Kluwer Academic Publishers.

Martinez W. (2005) "COOCS – Outils lexicométriques pour l'analyse des cooccurrences." *Manuel d'utilisation*. SYLED-CLA2T, Université de la Sorbonne nouvelle – Paris 3. http://www.cavi.univ-paris3.fr/ilpga/individus/martinez/

Martinez W., Zimina M. (2002) "Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues." In *Actes des JADT'02*, Saint-Malo, p. 495-506. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/martinez_zimina.pdf

Morin E., Daille B. (2004) "Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé." *Traitement Automatique des Langues* n°45(3), pp. 103-122.

Nuk I., Ivanova J. (2005) "Création d'une terminologie français/russe dans le domaine de la pharmacovigilance." *Mémoire de DESS*. CRIM-INaLCO, Paris.

Ozdowska S. (2004) "Identifying correspondences between words: an approach based on a bilingual analysis of French/English parallel corpora." In *Proceedings of the COLING'04 Workshop on Multilingual Linguistic Resources*, Geneva, pp. 55-62. http://w3.univ-tlse2.fr/erss/textes/pagespersos/ozdowska/publis/ozdowska-coling04.pdf

Ozdowska S. (2005) "Using bilingual dependencies to align words in English/French parallel corpora." In *Proceedings of the ACL Student Research Workshop*, Ann Arbor (USA), pp. 127-132. http://acl.ldc.upenn.edu/P/P05/P05-2022.pdf

Sadat F., Déjean H., Gaussier É. (2002) "A Combination of Models for Bilingual Lexicon Extraction from Comparable Corpora." *Papillon 2002 Seminar*, Tokyo. http://www.papillon-dictionary.org/info_media/1540495.pdf

McEnery A., Xiao R. Z. (2005) "Parallel and comparable corpora: What are they up to?" In G. James and G. Anderman (eds.) *Incorporating Corpora: Translation and the Linguist*. *Translating Europe*. Clevedon: Multilingual Matters.

http://eprints.lancs.ac.uk/59/01/corpora_and_translation.pdf

Véronis J. (Ed.) (2000) *Parallel Text Processing: Alignment and use of translation corpora*. Dordrecht: Kluwer Academic Publishers.

WHO (1972) "International drug monitoring: The Role of National Centres." *Technical Report No 498*. Geneva: World Health Organization.
http://www.who-umc.org/graphics/9277.pdf

WHO (2002) "The importance of pharmacovigilance." *Safety monitoring of medicinal products.* Geneva: World Health Organization.
http://whqlibdoc.who.int/hq/2002/a75646.pdf

Zimina M. (2004) "Bi-text Topography and Quantitative Approaches of Parallel Text Processing." In *Proceedings from the Corpus Linguistics Conference Series*, vol. 1, n°1 Birmingham.
http://www.corpus.bham.ac.uk/PCLC/