Digital editions and corpora of francophone diaries by Alexandre Chicherin and Olga Orlova-Davydova

Alexei Lavrentiev, Michèle Debrenne, Nina Panina, Dmitry Dolgushin, Andrey Borodikhin

DIGITAL EDITIONS AND CORPORA OF FRANCOPHONE DIARIES BY ALEXANDRE CHICHERIN AND OLGA ORLOVA-DAVYDOVA[1]

Alexei Lavrentiev, Michèle Debrenne, Nina Panina, Dmitry Dolgushin, Andrey Borodikhin

Abstract: This paper presents a project of digital editions and research corpora of francophone diaries written by two Russian aristocrats, Alexandre Chicherine (1793–1813) and Olga Orlova-Davydova (1814–1876). The original French texts of these diaries have never been published in spite of their considerable value for research in linguistics, literature, social history and history of the arts. Alexandre Chicherin's diary contains a great number of the author's drawings that are closely related to the text of the diary entries. Olga Orlova-Davydova's diary contains interesting details on the everyday life of the Russian aristocracy and on some historical events and well-known figures. Both diaries provide rich data on the interaction of the Russian and French languages and cultures. The workflow of the project consists of a primary transcription with Microsoft Word using styles and some special characters (micro-syntax). This transcription is automatically converted to TEI XML and then imported into the TXM corpus analysis and publication platform (http://textometrie.org). Prototype corpora of both diaries are available on the TXM demo portal (http://portal.textometrie.org/demo/?command=page&path=/JournauxFrancophones).

Keywords: diaristic literature, French, TXM, TEI, cultural interaction

1. Introduction

Personal diaries written in French by Russian aristocrats in the 19th century are an interesting object for research in the fields of linguistics, literature, social history and history of the arts. Whereas studies of diaries created in monolingual contexts (both Russian and French) have been carried out for a few decades,[2] only in recent years have the first works fo-

[1] This research has been funded by the Russian Foundation for Humanities (project No 16-24-08001) and the French *Fondation Maison des sciences de l'homme*.

[2] See Egorov 2018 for bibliographic references.

cused on the bilingual aspect of this literature been published (Gretchnaia & Viollet 2008; Gretchnaia et al. 2012). Many valuable documents are still unpublished, and technologies developed in digital humanities offer a unique opportunity to make these documents available to the academic community in the form of complex digital resources including facsimile, multi-layer transcription, critical apparatus, linguistic annotation and tools for search and visualization. In this article we present the work done on a corpus of two diaries written in the beginning and in the middle of the 19th centuries that were analysed and published using the TXM corpus analysis software.

2. Source documents

The first diary of the project corpus was written by Alexandre Chicherin (1793–1813, referred to as "AC" hereafter), a young Russian officer who took part in the war against Napoleon. He kept his diary from 6 September 1812 to 13 August 1813, when he was mortally wounded in a battle. The manuscript is held in the Russian Public Historical Library. It consists of 270 pages of text and illustrations (83 in total). These illustrations play a very important role in the diary because they interact with the text in many ways, and these relations are particularly interesting to study (Panina 2018). A Russian translation of the diary was published in 1966 (Engel & Perper 1966), but only a few of the illustrations were reproduced. In some cases the translation is not quite faithful to the original, and some passages were deliberately omitted (Mischenko 2001). Therefore, a digital edition of the original French text accompanied by all the illustrations and equipped with tools for analysis of text-to-image relations would be of great use to the research community.

The second diary belongs to Olga Orlova-Davydova (born Bariatinskaya, 1814–1876, referred to as "OD" hereafter), a member of a well-known Russian aristocratic family. Most of the entries correspond to the period from 1830 to 1847, with just one note added in 1849. (OD kept a separate diary, written in Russian, from 1869 to 1870, but this document is beyond the scope of our project.) The original manuscript is stored in the Russian State Library (Orlov-Davydov fund, F. 219, box 92), and there exists a manuscript copy which was probably ordered and revised by her daughter Maria Orlova-Davydova in the late 1890s or early 1900s. This

copy is stored in Novosibirsk State Public and Technical Library of the Siberian Branch of the Russian Academy of Sciences as a part of the Tikhomirov collection.[3] This copy is an interesting document for the study of Franco-Russian bilingualism, as it contains traces of language interference of both the author and the copyist. It also has the advantage of being more legible and easier to access for the project team based in Novosibirsk. For this reason the project started with the transcription of the copy, whereas the transcription of the original was added at a later stage.

3. Workflow

The workflow for both diaries was generally the same, although each source required some special processing. The first, and the most important and time-consuming, stage consisted of the transcription of primary sources with Microsoft Word. At this stage customized styles and special characters were applied to prepare automatic conversion to TEI XML and import into TXM. At the second stage the document files were converted to TEI XML using the OxGarage service,[4] and finally a series of XSLT transformations was applied during the TXM import process. The choice to do all the work of editing and pre-annotating the transcriptions with Microsoft Word (and not directly in TEI XML) was due to the fact that it was impossible to organize a training in this technology for the Russian team within the deadlines of the project. The high cost of user-friendly XML editing software like Oxygen XML Editor also influenced the decision to use Microsoft Word. The project was an occasion to test how far it is possible to go in text annotation using Microsoft Word.

4. Document structure

Microsoft Word's "Heading X" styles were used to encode the text structure. The basic level in both diaries is a daily entry, but there are some differences at higher levels of the structure.

AC's diary is contained in a single volume, so encoding the structure was rather straightforward: a single file for the diary and a division for

[3] See Borodikhin & Dolgushin Acc. for a detailed presentation of the source documents.

[4] http://oxgarage.tei-c.org.

each daily entry. The situation is much more complex in the case of OD. Both the original and the copy were written in several notebooks, and the chronological order of the entries does not always correspond to the archival order of the documents. In 1834 OD wrote a long autobiographical chapter starting with her birth and ending in 1833, so there is an overlap with some earlier entries. This chapter marks, in our opinion, the beginning of a "new diary" for several reasons: the story of OD's life starts from the beginning, she tries to organize records in chapters (although she abandons this effort some years later), and she calls the chapter "Chapter 1" and writes it (and the following chapters) in a new notebook. In addition to regular entries, OD used a separate notebook to write down some spiritual thoughts in 1843, 1847 and 1849. Therefore, we believe that this diary consists of three separate "works": (1) the early diary (1830-1834), (2) the new diary (1834-1845) and (3) spiritual thoughts (1843, 1847, 1849).

The primary transcription of the diary was made in the archival order of the source documents. Each notebook was transcribed in a separate document file. The call number is given in "Heading 1" style in the beginning of each document. "Heading 2" is used to indicate to which of the three "works" of OD's the following entries belong.

A normalized date in {YYYY-MM-DD} format (according to the Julian calendar used in Russia until 1918) was added in "Heading 3" style for each daily entry in both diaries. This was necessary to allow automated analysis and comparison of records according to their date.

In OD's diary, some dates were given in the Gregorian calendar (especially during her trips to Western Europe), and some dates are erroneous. A correct normalized date according to the Julian calendar was provided in all cases. Some records describe events of several days; in this case a "+" sign was added to the normalized date (e.g. {1830-11-23+}). The dates as they appear in the source document were transcribed and marked up with "date" character style.

In AC's diary most of the records were accompanied by illustration. For each image, we added a table with some metadata including the title of the drawing, the date of the drawing and the corresponding entry, the drawing genre and style, the people and places depicted, etc.). Whenever possible, the inner structure of the entries was marked up using custom-

ized paragraph styles such as “1_preamble”, “2_action”, “3_retrospection”, “4_comparison” and “5_conclusion”.

5. Segment-level markup

Names of people and places were marked up using character styles named “persName” and “placeName” respectively. An “unclear” style was applied to hardly legible segments. All these styles are automatically converted to proper TEI XML tags by OxGarage. In other cases we had to use project-specific style names and some special characters to ensure more precise TEI XML markup. These are processed and converted to TEI XML by XSLT scripts during import into TXM.

For instance, the styles named “sic-ortho”, “sic-lex” and “sic-gramm” were used to mark spelling, vocabulary and grammatical errors, respectively. If necessary, a particular letter or group of letters was marked as erroneous using curly braces: e.g., *G{é}nève* (the whole word is marked using “sic-ortho” style).

If correction marks were present in the source document, the styles “add”, “del” and “subst” were used for additions, deletions and replacements, respectively, and square brackets indicate the added or deleted characters. For instance, *pren[aie/a]nt* means that the word form *prenaient* was changed to *prenant*.

Russian words were marked with “lang-ru” if written in Cyrillic and “translit” if they were transliterated (e.g., *téléga*, Russian for ‘cart’).

The situation is more complicated where several different phenomena co-occur. For example, a personal name could be spelled in Cyrillic and a place name could contain an error. In this case combined style names such as “placeName-sic” were used, but it is clear that the use of styles in Microsoft Word meets its limits here, as unlike TEI XML elements, the character style of word-processing software cannot nest.

6. Pagination

In order to connect the transcription to the facsimiles of the source documents and to facilitate browsing through the document, special markup was used for page breaks. Some pages are numbered in the source documents, some are not, and in some cases the numbers appearing in the source document are erroneous. Therefore, in the digital edition each page had two

numbers: an "original" number, included whenever present in the source document, and an "archival" number systematically added. If both numbers are identical, simple notation in angle brackets was used: e.g., <10>. If the original number is different from the archival one, the latter was given after a vertical line: e.g., <14|15>. If the original page is unnumbered, the following code was used : e.g., <|12>.

If a page break occurred inside a word in the source document, the mark is placed exactly where it was found in the source document: e.g., *der-*<11>*nière*.

7. Import into TXM

TXM is a free open-source software platform designed for corpus compilation, analysis and publication (Heiden 2010). Thanks to its "XML XTZ + CSV" import module, it is possible to create corpora from any XML-encoded source documents through a series of XSLT transformations and under certain conditions generate synoptic pagination displaying side-by-side the text and a facsimile of the source document.

XSLT processing is available at different import stages:

1. Splitting or merging XML source files in order to optimize the corpus structure;

2. Pre-processing XML files to prepare for tokenization;

3. Post-processing XML files after tokenization to fix possible errors and create token-level annotations;

4. Generating one or several custom page layouts for reading the texts of the corpus.

In our project, only OD's diary required split-merge stage processing, due to the complexity of its text structure described earlier. After a conversion of transcription Word documents to TEI XML with OxGarage, the data was reorganized to create six files (three for the original and three for the copy) corresponding to three separate "works" we identified earlier.

At the next stage, a series of transformations was applied to obtain a TEI-conformant XML file ready for tokenization. As a matter of fact, text-processing software like Microsoft Word or LibreOffice Writer may create artificial divisions between text segments with identical markup. It may also introduce segments with additional formatting information (such

as font size) as a result of copy and paste or other editing actions. These segments and divisions are invisible to users but they may cause errors in tokenization and element indexing. Therefore, we had to remove superfluous segments and merge artificially divided adjacent elements before any further processing. Once the merger of adjacent identical elements is complete, it is possible to convert project-specific styles and special characters into proper TEI tags. Any styles not recognized as TEI tags are rendered as <hi rend=”styleName”> by OxGarage. Their conversion to proper TEI is pretty straightforward with XSLT. Transforming special characters into XML tags requires a sequence of templates that parse text nodes with regular expressions.

Tokenization is operated by a Groovy script internal to the TXM platform. With this script, the user may adjust the list of word-separating characters and pre-tokenize some segments that require complex processing. The user can also provide a list of elements that should not be tokenized (for instance, editorial notes).

The tokenized text is once again submitted to a sequence of XSLT transformations. At this stage it is possible to merge or split some tokens and to add token-level annotations to facilitate corpus queries. In our case, an attribute “error” was introduced for various types of errors marked up in the Word document. Another attribute, “crochets”, re-introduces some special characters to the word form (such as brackets for additions and deletions) in order to allow queries on the word-internal markup.

After the third series of XSLT transformations, the properly tokenized text is submitted to TreeTagger (Schmid 1994) for automatic part-of-speech tagging and lemmatization. The POS tags and lemmas provided by TreeTagger, as well as all other word-level annotations, are finally recorded in TEI-TXM XML format. The TXM extension for TEI provides the possibility to create an unlimited number of annotations at the word level thanks to <txm:ana> element.

At the next stage, importing a corpus, TXM produces indexes for the CQP search engine and generates edition pages. These can be created either by a default Groovy script or by a custom XSLT stylesheet. In our case, XSLT customization was necessary in order to visualize various levels of markup.

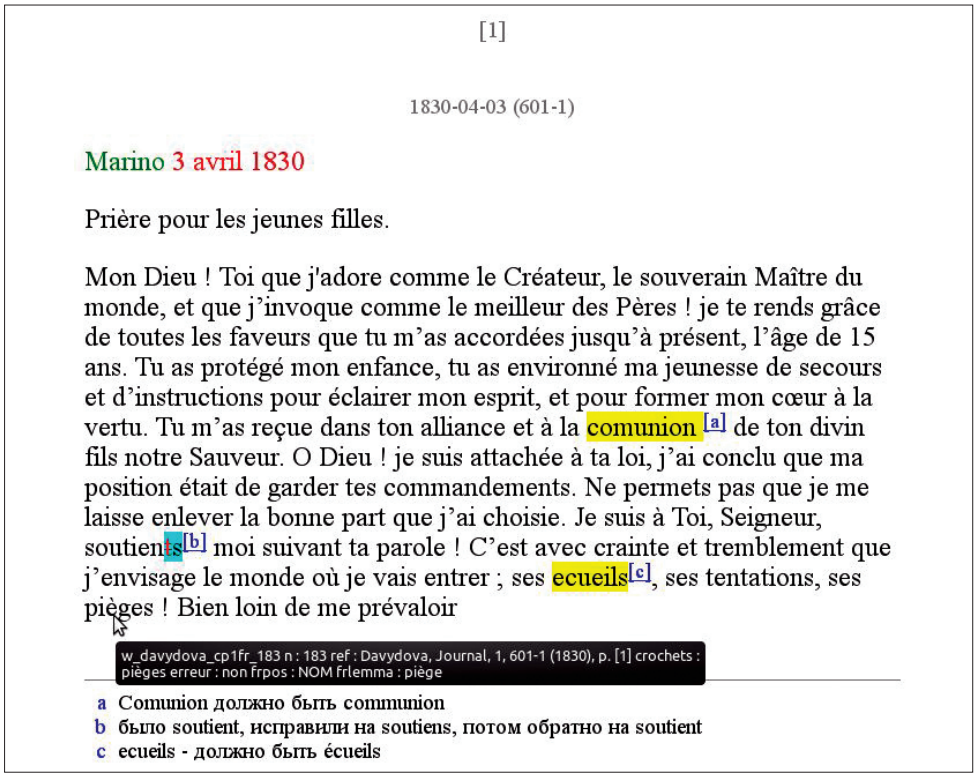


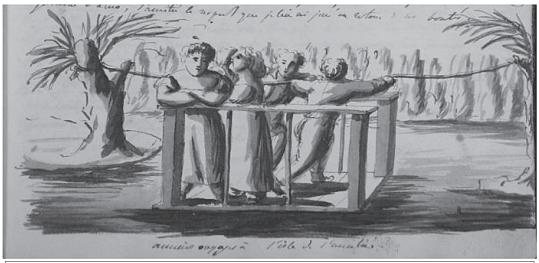
Figure 1. Screenshot of a page transcription from OD's diary

8. Reading and analysing the corpus with TXM

In Figure 1 we provide a screenshot of a page from OD's diary. The page number is given in brackets, as the page is not actually numbered in the source document. The normalized date and the call number of the source document are displayed in gray in the beginning of each entry. In the transcription that follows one can see a place name in green and the date in red, the errors highlighted in yellow, and the correction in the source document highlighted in blue. Additional comments are provided in footnotes. When the mouse cursor is placed over a word, a tooltip with corresponding annotations is displayed.

Figure 2 represents a screenshot from AC's diary containing an illustration followed by a table with its description. In this corpus image, descriptions and the transcription are both searchable, but it is of course possible to work on a subcorpus of pure transcriptions.

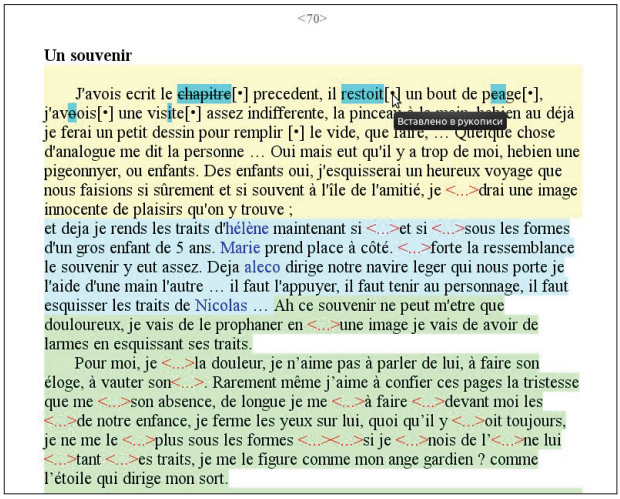
[16 décembre 1812]



Titre	ami <... > voyage à l'ile de l'amitié
dateDessin	1812-12-16
dateNote	1812-12-16
nDessin	NN
nPage	
lieuNote	Vilnius
lieuDessine	demeure familiale
orientation	portrait
typeDessin	fin de chapitre
genreDessin	allégorique

Figure 2. Screenshot of an illustration and its description from AC's diary

A screenshot of the transcription of AC's diary is given in Figure 3. Background colour is used to indicate the semantic structure of the record that was marked up with paragraph styles in the initial transcription. The yellow background corresponds to a preamble, blue to an action and green to retrospection. As in OD's diary, personal names are displayed in blue font colour and corrections are highlighted in blue. Illegible segments are represented by "<…>".



<70>

Un souvenir

J'avois ecrit le chapitre[•] precedent, il restoit[•] un bout de peage[•], j'aveois[•] une visite[•] assez indifferente, la pinceau à [illegible] au déjà je ferai un petit dessin pour remplir [•] le vide, que faire, … Quelque chose d'analogue me dit la personne … Oui mais eut qu'il y a trop de moi, hebien une pigeonnyer, ou enfants. Des enfants oui, j'esquisserai un heureux voyage que nous faisions si sûrement et si souvent à l'ile de l'amitié, je <...>drai une image innocente de plaisirs qu'on y trouve ;
et deja je rends les traits d'hélène maintenant si <...>et si <...>sous les formes d'un gros enfant de 5 ans. Marie prend place à côté. <...>forte la ressemblance le souvenir y eut assez. Deja aleco dirige notre navire leger qui nous porte je l'aide d'une main l'autre … il faut l'appuyer, il faut tenir au personnage, il faut esquisser les traits de Nicolas … Ah ce souvenir ne peut m'etre que douloureux, je vais de le prophaner en <...>une image je vais de avoir de larmes en esquissant ses traits.

Pour moi, je <...>la douleur, je n'aime pas à parler de lui, à faire son éloge, à vauter son<...>. Rarement même j'aime à confier ces pages la tristesse que me <...>son absence, de longue je me <...>à faire <...>devant moi les <...>de notre enfance, je ferme les yeux sur lui, quoi qu'il y <...>oit toujours, je ne me le <...>plus sous les formes <...><...>si je <...>nois de l'<...>ne lui <...>tant <...>es traits, je me le figure comme mon ange gardien ? comme l'étoile qui dirige mon sort.

Figure 3. Screenshot of a page transcription from AC's diary

In addition to visualising the edition, TXM provides a wide range of tools for qualitative and quantitative research on the corpus. For instance, the *index* command allows you to obtain a list of personal names in the corpus with the following query: *<persname> []+ </persname>*. There are 7041 occurrences of 1595 different forms of personal names in the copy of OD's diary. Another query used with the *progression* command produces a chart with a curve increasing with every occurrence of a Russian word (*<foreign_lang="ru"> []+ </foreign>*). The steeper the curve, the denser is the use of the searched term over time. As one can clearly see in Figure 4, the density of Russian words increases considerably in 1845 and the following years. The distance between the vertical lines representing years corresponds to the volume of writing. We can see that this volume increases dramatically in 1844 and decreases progressively later.

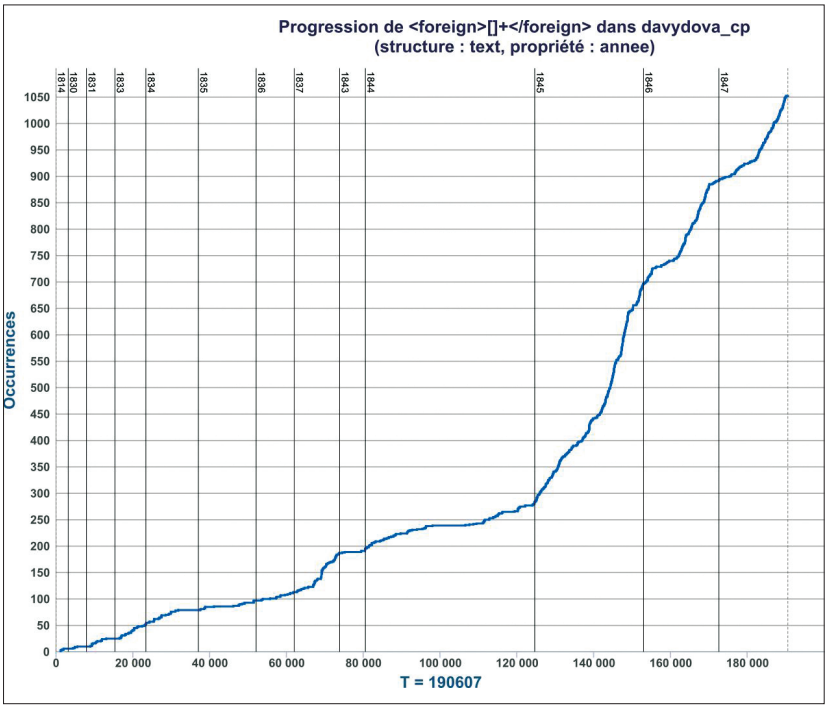


Figure 4. Progression chart of Russian words in OD diary

The full range of research tools provided by TXM is provided in the TXM User Manual and various tutorials available at the TXM project website (http://textometrie.org).

9. Conclusion

This project on Francophone diaries has provided an occasion to set up a workflow for creating a complex digital edition by an international team and to test the opportunities that TXM provides for working with such kind of corpora. The workflow based on transcription and pre-annotation with Microsoft Word and a series of XSLT transformations incorporated in the TXM import process has been efficient, although it is less suitable for complex cases of embedded annotations (e.g., manuscript errors in proper names).

The digital editions AC's and OD's diaries are available online on the TXM demo portal[5]; the access codes to read and search theses resources are provided upon request. Work on their proof-reading and refining is still going on. Once the text of the editions is stabilized, open public access to them will be provided.

References

Borodikhin & Dolgushin 2019: Borodikhin, A. Yu., Dolgushin, D. V. "Dnevniki O. I. Orlovoj-Davydovoj kak chast' Tihomirovskij kollekcii GPNTB SO RAN." *Quaestio Rossica*, Forthcoming (article accepted).

Debrenne 2016: Debrenne, M. "Sopostavitel'nyj deviatologicheskij analiz perepisannyh dnevnikov O. I. Davydovoj i pervichnyh tekstov." *Vestnik NSU. Series: Linguistics and Intercultural Communication* 14-3 (2016), 59–75.

Debrenne 2017a: Debrenne, M. "Sozdanie sistemy razmetki francuzskogo teksta dnevnikov O. I. Davydovoj dlja avtomaticheskoj obrabotki teksta." *Vestnik NSU. Series: Linguistics and Intercultural Communication* 15-1 (2017), 34–40.

Debrenne 2017b: Debrenne, M. "The French Language in the Diaries of Olga Davydova." In Strien-Chardonneau, M. van & Kok Escalle, M.-Ch. (eds.), *French as Language of Intimacy in the Modern Age. Le français, langue de l'intime à l'époque moderne et contemporaine.* Amsterdam: Amsterdam University Press, 2017.

Debrenne 2018: Debrenne, M. "Sopostavitel'naja tipologija oshibok i ispravlenij v originalah i kopijah dnevnikov O. I. Orlovoj-Davydovoj". *Vestnik NSU. Series: Linguistics and Intercultural Communication* 16-2, 127–143 (DOI: 10.25205/1818-7935-2018-16-2-127-143).

[5] http://portal.textometrie.org/demo/?command=page&path=/JournauxFrancophones.

Egorov 2018: Egorov, O. G. *Russkij literaturnyj dnevnik XIX veka. Istorija i teorija zhanra. Issledovanie*. 2nd edition. Moscow: Flinta, 2018.

Engel' & Perper 1966: Engel', S. T., Perper M. I. (eds.) *Dnevnik Aleksandra Chicherina, 1812–1813*, Moscow: Nauka, 1966.

Gretchnaia & Viollet 2008: Gretchanaia, E., Viollet, C. *"Si tu lis jamais ce journal…": Diaristes russes francophones, 1780–1854*, Paris: CNRS-Editions, 2008.

Gretchnaia et al. 2012: Gretchanaia, E., Stroev, A., Viollet, C. (dir.) *La francophonie européenne aux XVIIIe–XIXe siècles. Perspectives littéraires, historiques et culturelles*. Bruxelles, Bern, Berlin, Frankfurt am Main, New York, Oxford, Wien: Peter Lang, 2012.

Heiden 2010: Heiden, S. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." in Otoguro, R., Ishikawa, K., Umemoto, H., Yoshimotoand, K. and Harada, Y. (eds.) *24th Pacific Asia Conference on Language, Information and Computation*. Tokyo: Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010, 389–398 (https://halshs.archives-ouvertes.fr/halshs-00549764).

Mischenko 2001: Mischenko, T. K. "A. V. Chicherin – uchastnik voennyh dejstvij pod Malojaroslavcem: iz opyta raboty s podlinnym dnevnikom A.V. Chicherina." in *Otechestvennaja vojna 1812 goda v Kaluzhskoj gubernii i rossijskoj provincii. Materialy nauchnoj konferencii*. Malojaroslavec, 2001, 10–20.

Panina 2016: Panina, N. "Dnevnik Aleksandra Chicherina 1812–1813 gg.: Avtobiograficheskij zamysel v kontekste frankofonii avtora". In Dmitrieva, E.E., Lebedeva, O.B., Stroev, A.F. (eds.) *Sibirsko-francuzskij dialog XVII-XX vekov i literaturnoe osvoenie Sibiri: materialy Mezhdunarodnogo nauchnogo seminara, Tomsk, 11-15 iyunya 2015 / In-t mirovoj literatury im. A. M. Gor'kogo Rossijskoj akad. nauk,* Moscow: IMLI RAN, 2016, 202–216.

Panina 2019: Panina, N. "Francuzskij original dnevnika Aleksandra Chicherina 1812–1813 gg." *Tekst. Kniga. Knigoizdanie*. Forthcoming (article accepted).

Schmid 1994: Schmid, H. "Probabilistic Part-of-Speech Tagging Using Decision Trees." in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger