CHAPTER 27

Shifting the Norm: The Case of Academic Plagiarism Detection

Mikhail Kopotev, Andrey Rostovtsev, and Mikhail Sokolov

27.1 Introduction

Plagiarism currently tends to be viewed as a problem connected primarily with students, albeit more prominent authors such as William Shakespeare and George Friedrich Handel were accused of it long ago. Plagiarism continues to be widespread in educational institutions, predominantly due to single-click technology, but another contributing factor that helps make it common practice is the tolerance of plagiarism on the part of educators and academia in general. In 2004, for instance, it was estimated that 10 percent of student projects in the United States and Australia involved plagiarism (Oakes 2014, 60). By contrast, in Russia, 36 percent of respondents admitted to having regularly copied the texts of others (Kicherova et al. 2013, 2); as many as 36.7 percent of undergraduate students in 8 Russian universities took personal credit for material they had, in fact, downloaded from the Internet (Maloshonok 2016).

M. Kopotev (✉)
Higher School of Economics (HSE University), Saint Petersburg, Russia
e-mail: mkopotev@hse.ru

A. Rostovtsev
Institute for Information Transmission Problems RAS, Moscow, Russia

M. Sokolov
European University at Saint Petersburg, Saint Petersburg, Russia
e-mail: msokolov@eu.spb.ru

D. Gritsenko et al. (eds.), *The Palgrave Handbook of Digital Russia Studies*, https://doi.org/10.1007/978-3-030-42855-6_27

The problem of plagiarism is certainly not limited to undergraduate students. For example, two cases of plagiarism were documented in PhD dissertations published in Germany in 2011. These cases, which were analyzed in detail by the *GuttenPlag* community, led to the monograph titled *False Feathers: A Perspective on Academic Plagiarism* (Weber-Wulff 2014). However, plagiarism is arguably exceedingly prevalent and more deeply rooted in Russia than in Europe (see Golunov 2014; Denisova-Schmidt 2016). One reason for this may be that the symbolic value of scholarly achievements in Russia has been widely appropriated by politicians, civil servants, businesspersons, and administrators from educational and medical fields. These professionals have been awarded degrees by lenient defense panels for dissertations that have been entirely copy-pasted from other sources. This would be even more prevalent among those in power if strong opposition had not been voiced by the academic community. This led to the establishment of "Dissernet," a network that purports to expose large-scale plagiarism in Russian scientific publications. Our focus in this chapter is on Russian doctoral and post-doctoral dissertations,[1] which constitute merely the tip of an academic iceberg that includes articles, monographs, coursebooks, and other scholarly works. In fact, the post-Soviet publishing market is flooded with texts of questionable originality.

The current availability of material and ease of use raises more general questions. For example, what is the *textual authenticity* and what are the *norms* of textual authenticity for scholars at a time when everything is "a copy of a copy of a copy" (Palahniuk 1996)? Western academic culture presupposes that the origin of the words and ideas in a scholarly text, from the first word to the last, are from the author or authors accredited in connection with the title, with the exception, of course, of properly attributed quotations from other scholarly works, or paraphrases of them. Even within these norms, however, exactly what is meant by "original from the first word to the last" is somewhat ambiguous (Korbut 2013).

One of the principal subjects in sociology since the time of Durkheim is social norms; in other words, the rules of conduct that are considered proper, right, and socially desirable. In recent decades, digitalization has made it possible to analyze compliance with various norms using digital traces of naturally occurring behaviors rather than self-reporting, official statistics, or other less reliable resources. Areas of conduct analyzed in this manner vary—from using dirty words (McEnery 2004) to observing meritocratic principles in the selection of professors (Clauset et al. 2015). Due to the increasing digitalization of Russian society together with emergent methods of analysis, it is now possible to study the level of support for a particular norm, specifically one that requires authenticity in academic writing, and to analyze conditions under which this norm is likely to be transgressed.

The *norm of textual authenticity* requires that any academic text be fully original in compliance with the highest academic standards, which permit quotation and paraphrasing with correct and appropriate attribution to the source. This could be deconstructed into two different norms requiring (1) that the text be written in full by its presumed author and (2) that the text is written for one and only one purpose or publication outlet. The latter, which forbids any

recycling of an academic text, is more restrictive than the former in that it bans all forms of reuse including that of one's own texts. The focus of this chapter is on the first, less restrictive norm of texts that are written entirely by an author. The assumption is that dissertation authors can reproduce sections of their dissertation in articles and that this is universally regarded as a permissible and even a desirable practice.

The norm of textual authenticity requires identification of what constitutes a form of expression, such as widely used terms or stock phrases, and what is the true content of the academic text. Some forms of expressions or presentation style may or may not qualify as unauthorized borrowing. These include the use of certain truisms and clichés such as "to the best of our knowledge," design layouts, and fonts. To apply the norm of textual authenticity thus requires constant discrimination between what is the "mere form" of an academic message and what is "the message itself," with the form being considered part of academic convention and without authorship. The digitalization of scholarly production together with software development facilitate the study of particular variations in the norms of textual authenticity.

We begin the analysis for this chapter by describing the challenge that academic plagiarism poses for digital humanities in an era when sophisticated tools make it possible to detect inappropriate academic activity, and we focus specifically on Russian dissertations. Second, we examine the changing norms of academic integrity in terms of the sociology of science. Thus, in Sect. 27.2, we describe the various types of plagiarism and the computational tools that have been created to detect fraudulent texts. Section 27.3 comprises a review of available digitized resources, including dissertations, articles, and abstracts published by the Russian academic press. In Sect. 27.4, we provide an overall picture of the Dissernet findings when these tools were applied to large-scale (greater than 50%) plagiarism in dissertations that have been defended in Russia. Section 27.5 presents a case study of small-scale plagiarism based on the same academic genre. This study analyzes and traces the shifting authenticity norms in Russia since post-Soviet times. Finally, Sect. 27.6 concludes the chapter.

27.2 Types of Plagiarism and Tools Enabling its Detection[2]

The *Modern Language Association (MLA) Style Manual and Guide to Scholarly Publishing* defines plagiarism as follows:

> Forms of plagiarism include the failure to give appropriate acknowledgment when repeating another's wording or particularly apt phrase, paraphrasing another's argument, and presenting another's line of thinking. (Modern Language Association 2008, 166)

Two types of plagiarism are commonly distinguished in the scholarly literature, which Bela Gipp refers to as *copy&paste* versus *shake&past* (Gipp 2014, 12; see also Potthast et al. 2010). The former refers to copying someone's text unchanged without proper acknowledgment, whereas the latter implies minor modifications, such as varying the word order or using synonyms—again without acknowledging the source. Several services are currently available that can detect plagiarism in Russian-language texts (see Nikitov et al. 2012). Below we describe several of the most advanced technologies applicable to textual plagiarism. We do not address evidence of fraudulent publication such as image and diagram falsification, carbon-copied lists of references, or data manipulation (for example, wild data or loose correlation).[3]

Copy-and-paste, or *cut-and-paste* refers to "involving or relating to the cutting and pasting of printed material, or (Computing) the 'cut' and 'paste' functions on a computer" (OED, c.v. *cut-and-paste*). Technically, the basic commands available on any computer can create the simplest form of plagiarism, and hence the most alluring, is when a source is used but not cited properly. This is easy to identify, even when the text under suspicion has been—intentionally or otherwise—modified or corrupted. Detection is based on identifying similar chains of symbols and their possible modifications. Some of these modifications reflect deliberate distortions by the borrower-creator, such as Cyrillic letters replaced with identical Latin ones, whereas others derive from optical character recognition (OCR) (see Table 27.1).

The plagiarism in each of these cases can be detected by conducting a basic similarity test or by using a more sophisticated technique such as the *Levenshtein distance*, which is the number of required symbol substitutions for one word to be changed into another (Levenshtein 1966). This approach is exemplified by a tool called *Disserorubka* (literally "the Thesis-grinder") and was developed by the Dissernet community. Another service that is available online, albeit a commercial one, is antiplagiat.ru, which is specifically designed to detect plagiarism in Russian texts. The available techniques and services allow copy-and-paste plagiarism to be effectively detected by taking into account specific issues related to the Cyrillic alphabet, such as the Cyrillic "P" replaced with Latin "P," and the confused recognition of "Ф" as "%."

Table 27.1 A source text (left) and the copy-pasted text after OCR (right)[a]

Специфика воинской деятельности в сочетании с высочай**ш**им напряжением всех духовных и физических си**л**, с возможностью и **нео**бходимостью **самопо**жертвовани**я** во имя **Р**одины, определяют значимость духовного **ф**актора для армии.	**С п е ц и ф и к а в о и н с к о и д е я т е л ь н о с т и в с о ч е т а н и и** с высочаишим напряжением всех духовных и физических си**n**, с возможностью и **не о**бходимостью **само по**жертвовани**п я b**о имя **Q**одины, определяют значимость духовного **%**актора для армии.

The distortions are in bold.

[a]The examples are fictional and were constructed by the authors: any correspondence to actual texts is accidental.

In the case of *paraphrasing*, different linguistic techniques are used to rework the source texts, including word removal, word replacement, synonym substitution, word-order modification, grammatical changes, and patchwriting (for example, by combining fragments from several texts) (Oakes 2014, 60). The nature of these changes depends on whether the paraphrase had been generated by means of manual text editing or automatically (Gupta et al. 2011, 1), as shown in Table 27.2.

Dictionary-based methods are used to detect this type of plagiarism, requiring a lexicon that contains all possible changes, substitutions, and transformations. All modifications are weighted, with the slighter ones prioritized, and those that are more substantial being downgraded. For instance, word-order modification and word replacement are both automatically detectable, but the former is weighted more heavily than the latter because it preserves more of the original source. An application of this approach to the Russian, Ukrainian, and English languages, developed by K. Kuznetsov and M. Kopotev, can be found online at http://dissercomp.ru. Thus far, the service is able to detect paraphrased plagiarism in Russian, Ukrainian, and English texts.

Another case of paraphrasing is *interlingual plagiarism*, when a text is "paraphrased" in a sense from one source language to another. This process may involve manual or automatic translation. When automatic translation is involved, the output of the machine translator usually undergoes post-editing, along with obfuscation, which makes a comparison of the sources with the plagiarized text substantially more difficult while at the same time displaying evidence of translation (Table 27.3).

Detecting this type of plagiarism poses a challenge and tests the very limits of the methods available to scholars in digital humanities. Those engaged in this endeavor have turned to distributional neural net modeling, and specifically to distributional semantics.

The initial idea behind this approach reflects the understanding of meaning through context, as proposed by J. R. Firth: "You shall know a word by the company it keeps" (Firth, J. R. 1957, 11). The main objective in distributional semantics is to analyze the co-occurrence of linguistic entities (usually words)

Table 27.2 A source text (left) and the paraphrased text (right)

Некоторая часть **начальников** и **преподавательского состава**, обладая **неплохими** теоретическими знаниями, сами имеют **слабые** практические навыки, поэтому они не могут **правильно** учить **курсантов**.	Некоторая часть **командиров** и **учителей**, обладая **хорошими** теоретическими знаниями, сами имеют **плохие** практические навыки, поэтому они не могут **хорошо** учить **студентов**.
English translation Some of the **heads** and **faculty**, possessing **goodish** theoretical knowledge, have themselves **weak** practical skills, so they can not **properly** teach the **cadets**.	Some of the **commanders** and **teachers**, possessing **good** theoretical knowledge, have themselves **poor** practical skills, so they cannot teach **students well**.

The paraphrasing is indicated in bold.

Table 27.3 A source text (left) and the translated text (right)

In a crisis, the whole educational system was reformed: the structure of educational institutions changed; the throughput of schools increased.	В условиях кризиса была проведена реформа всей системы образования: изменилась структура учебных заведений, увеличилась пропускная способность училищ.

and to summarize this distribution statistically on multidimensional "semantic spaces." For example, the English noun *plagiarism* regularly collocates with the same words as the nouns *falsification, obscenity, and misbehavior*:

…**accused of** plagiarism/falsification/obscenity/misbehavior **in**…

Among the many applications for this paradigm, one that is based on the word2vec modeling was specifically developed to expose translated plagiarism. The authors call their method "semantic fingerprinting" (see Kutuzov et al. 2016); the service is also available online: www.dissernet.org/dissemsearch.

27.3 Available Electronic Resources

A well-functioning computational tool does recognize plagiarism effectively. If they are to achieve results, experts also need access to the relevant textual data. Numerous (preferably all) academic texts are required in order to compare the plagiarized text with potential sources by applying an algorithm that can make searches. The full range of texts, both online and offline, would be available in a perfect world, but real life poses additional challenges. An accepted presupposition here is that both the copycat who scans for a suitable source to rewrite, and the unmasker who is intent on revealing the copycatting are most likely to be relying on the same resources, in other words (publicly), available digitized texts.

How many scientific text documents in Russian have been digitized and made available to the public? In answer to this question, we consider different categories of academic texts. The first category includes doctoral and post-doctoral dissertations which are referred to as *autoreferats*, a formal abstract of the dissertation. An *autoreferat* is a summary of the main results reported in a work that the author compiles and it usually consists of 20–30 pages abstracted from the full text. These abstracts also contain basic information on the formal public defense such as the date and place of the event, the name of the academic supervisor, the official opponents, and so on. The degree candidate in Russia is required to deposit both the dissertation and the abstract in the main libraries of the Russian Federation. The RSL (*Rossijskaâ gosudarstvennaâ biblioteka*, Russian State Library) in Moscow has been a major repository for these texts from 1944 onwards. In 2003, the RSL management decided to ensure broad public availability and preservation of dissertations electronically. Thus far, this has led to the creation of the most comprehensive electronic collection of abstracts (*autoreferats*) of domestic doctoral and post-doctoral dissertations

in the world. To date, the collection incorporates more than 919,000 full texts. The dissertations defended in 1994 and thereafter were digitized rather systematically, whereas the collection of abstracts (*autoreferats*) covers the time period from 2007 up to the present. Most, but not all, dissertations and abstracts from previous years have also been digitized.

All of the aforementioned documents are available in the Digital Dissertation Library at http://diss.rsl.ru upon registration. Registered visitors receive free and unrestricted, open access to the abstract collection. Access to the copyright-protected part of the Digital Dissertation Library is provided at the RSL in Moscow or in its virtual reading rooms, of which there are more than 600 in Russia and worldwide. Most of the reading rooms located abroad are accessible through local university libraries. Readers who are registered individually are also offered the opportunity to access the full texts remotely. However, they are limited to viewing at most five dissertations per day, and no more than fifteen per month. Beginning in 2014, prior to their public defenses, all post-graduate students have been required to publish their dissertations and their abstracts online and in open-access forums. As a result, the number of available dissertations is increasing annually by approximately 30,000 texts. The RSL with its Digital Dissertation Library nevertheless remains the only central collection of these documents in Russia.

All types of scientific publications apart from dissertations are accessible in many electronic libraries, both in Russia and beyond. Russia's most comprehensive and ambitious repository is the Russian Scientific Electronic Library, available at elibrary.ru, which also offers many other categories of scientific publications. Another category comprises books and book chapters, of which more than 122,000 full texts are available in the Electronic Library, and more than 55,000 of them are open access. Collected papers constitute a further category of digitized documents available at the same website. There are also more than 127,000 volumes and papers available, and approximately 87,000 of them are open access. Conference and similar short papers are assigned a separate category among the digitized documents: there are more than 982,000 of them with 779,000 being open access. The last of these groups consists of academic articles or publications in scholarly periodicals, and this group naturally represents the largest category of digitized scientific documents with approximately 4.5 million papers written in Russian available at elibrary.ru, and of these, about 3.3 million are open access.

The impressive collections of academic texts described above have become available, thanks to public funding. They are key sources of successful scientific work in Russia and/or of data in Russian for projects ranging from conducting basic bibliographic searches to discovering trends in Russian science. These data provide the groundwork for the detection of plagiarism in academic texts. Plagiarism detection rests on two crucial conditions: effective algorithms and the availability of source texts to which a suspicious text is compared in order to find similarities. The available data in Russian meets both conditions that allow the effective detection of plagiarism and deal with this social

phenomenon in depth. In the next two sections, we explore two case studies that utilize available resources. The first case concerns large-scale plagiarism that involves the copying of more than half of the source text, which provokes general observations of fake academic activity in Russia. By contrast, the second case focuses on small-scale plagiarism and discusses cross-cultural variation in interpretations of authenticity norms.

27.4 The Best Practices of Dissernet in the Detection of Large-scale Plagiarism

The volunteer network known as Dissernet was established in 2013 to counter fraud and dishonesty in academia, specifically in fabricated dissertations and in the conferring of false university degrees. According to its manifesto, Dissernet is "a networking community of experts, researchers and reporters seeking to unmask swindlers, forgers and liars," whose members "oppose abusive practices, machinations and falsifications in the fields of scientific research and education, in particular in the process of defending theses and awarding academic degrees in Russia" (English translation from https://en.wikipedia.org/wiki/Dissernet).

It is now possible to detect plagiarism in thousands of dissertations, primarily through the application of in-house tools, introduced in Sect. 27.2, to the data described in Sect. 27.3 of this chapter. The abstract, or *autoreferat*, serves as a prerequisite for identifying suspected cases of plagiarism in that it is available online and is thus indexed by search engines such as Google and Yandex. This works even when the dissertation itself is not indexed, based on the assumption that when a dissertation contains a large amount of plagiarized text, its *autoreferat* will retain fragments of the plagiarized sources. Dissernet software is able to pick up the abstracts one by one by utilizing search-engine indices to search for textual coincidences within the entire, publicly available mass of Russian digitized texts, including articles, monographs, and dissertations as well as their abstracts. This is essentially how the technological part of the process works, and hundreds of thousands of texts are automatically checked in this manner. Dissernet is principally aimed at detecting large-scale plagiarism, which is determined to be the illegal use of equal to or greater than 50 percent of a text. In an extreme but real-life example, a source text was utilized in full, with the automatic replacement of "dark chocolate" with "local beef," and "confectionery" with "meat and dairy." As at beginning of 2020, Dissernet had identified almost 9,000 plagiarized dissertations, both doctoral and post-doctoral, that had been defended in the previous two decades.

At the next level of its investigation, Dissernet exposes established practices that are corrupt, such as when an *omertà*-like community repeatedly produces fraudulent dissertations. Dissernet findings clearly indicate that as soon as rampant plagiarism is detected in one dissertation, plagiarism is likely to be discovered in other dissertations defended before the same defense panel or under

the same supervision. Many of those who produce these dissertations work in a "conveyor-belt" mode by using exceedingly limited sets of scientific texts as sources. The graph below (Fig. 27.1) demonstrates the density of such practice that one dissertation-defense panel established at MGPU (*Moskovskij pedagogičeskij gosudarstvennyj universitet*, Moscow Pedagogical State University). This panel approved more than 90 "doctored" dissertations from 2001 to 2012, with the same actors playing interchangeable roles first as *kandidat nauk* (doctoral degree candidate) or *doctor nauk* (post-doctoral degree candidate) and later as *naučnyj konsul'tant* (supervisors) or official opponents (see Fig. 27.1).

First and foremost, Dissernet activity targets plagiarism among top-ranked Russian politicians and administrators, both in academia and beyond. Thus, the results cannot be interpreted as representing the whole landscape across all disciplines over the entire country. However, the number of dissertations tested (more than 20,000) allows us to draw a number of preliminary conclusions. First, the number of heavily plagiarized dissertations varies significantly depending on the academic field. Most of the identified fake dissertations (44%) were in the field of economics. Other academic fields deeply infected by fraud include pedagogy (16%) and law (12%), followed by the medical sciences, political science, engineering, and the social sciences. However, this type of fraud is less common in the natural sciences. It is important to mention that this

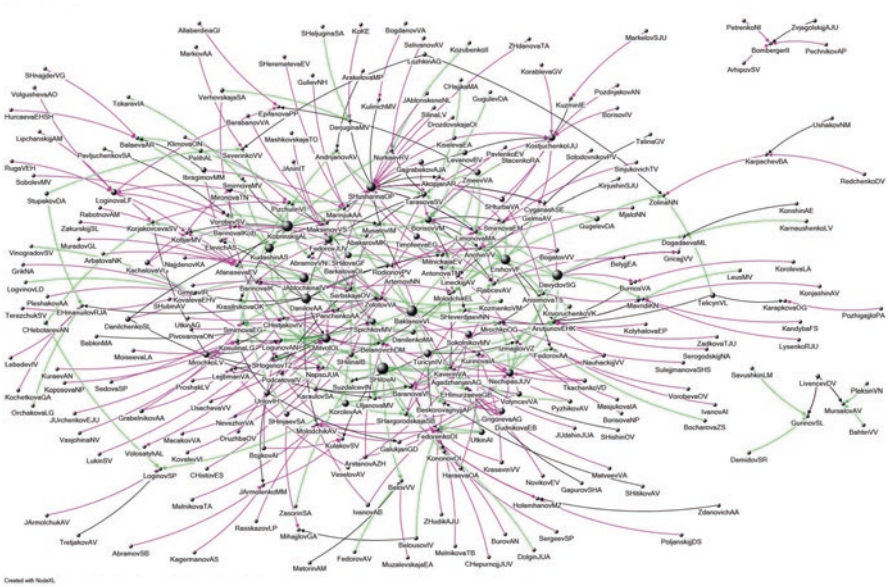

Fig. 27.1 A network in the MGPU producing large-scale plagiarism (A. Abalkina, Dissernet.org). The full interactive graph is available at: https://www.dissernet.org/publications/mpgu_graf.htm

distribution is symptomatic because it represents the main bottlenecks in modern Russia: economics, law, and education.

Second, universities have been predominantly responsible for faking academic production, whereas the research institutions of the Russian Academy of Sciences, the RAS, have produced relatively small numbers of detected plagiarism cases. The two most prominent universities in terms of producing faked material during the last fifteen years are Moscow State Pedagogical University and the Russian Presidential Academy of National Economy and Public Administration. Yet other "leading contenders" include the Russian State University for Humanities and the Russian State Social University, as well as the country's leading seat of learning, Moscow State University. By way of contrast, the RAS, which comprises hundreds of research institutions across the country, was ranked 23rd on the plagiarism list—the frauds being exclusively represented by its Caucasus-based branch.

Finally, the majority (approx. 50%) of those holding questionable academic degrees are working as administrative staff in universities. Not coincidentally, large-scale plagiarism was detected in 66 dissertations (21.22%) defended by rectors (311 of those awarded during the last fifteen years in Russia were checked). Politicians and businessmen fell behind in this regard with only about fifteen percent of their numbers engaging in plagiarism.

Large-scale plagiarism in Russia is, by its very nature, a special case when the numbers are compared to those recently disclosed in Western Europe (for example, see Weber-Wulff 2014). Whereas a Western plagiarist endeavors to present a text that has been copied from others as original research, the high-profile swindler in Russia may well not have even seen the plagiarized text prior to the public defense, having received it ready for publication from ghost-writers. When this occurs, wholesale plagiarism is not disguised; instead, the "dissertation" is composed with a crazy quilt of texts with fully automated replacements.

This pervasive academic corruption inevitably raises various questions. For example, does the widespread occurrence of badly adapted texts indicate a local trend that exclusively features pseudo-academics who attempt to enhance their value among their own kind? Or does it foretell greater changes in acceptable norms that academic communities have faced thus far? We address these questions in our second case study, presented in Sect. 27.5 below.

27.5 Small-scale Plagiarism and Shifting Norms of Textual Authenticity

While the detection of small-scale plagiarism also involves the same tools and collections as those described above, it is more dependent on manual processing in that a small piece of text may be a legitimate quotation or a paraphrase with a valid reference. This challenge calls for deeper conceptual reasoning on the shifting norms of textual authenticity.

As is the case with many other norms, *justifications* for the norm of textual authenticity are subject to deeper disagreement than the norm itself. Those who attempt to provide grounds for accepting this norm tend to present one of two arguments. The first is that either copy-pasting from the texts of other persons is defined as an infringement of these authors' intellectual property and thus as a type of theft, or they regard copy-pasting as a fraudulent way of obtaining intellectual distinction that is not actually deserved, and thus akin to cheating on an exam. The latter interpretation is based on the assumption that an individual with a university-level degree is able, single-handedly, to produce a text that meets certain stringent requirements. Nonetheless, both justifications can be disputed in specific cases. In contrast to more obvious cases of theft, dissertation plagiarism does not necessarily damage the rightful owner of the property, who probably loses little in terms of professional recognition given that dissertations are rarely read. Moreover, as a reason for condemning plagiarism, it becomes irrelevant if an author of the borrowed source raises no objections. The Dissernet studies nevertheless revealed that a person's supervisor and/or opponents are the most likely sources of unauthorized large-scale borrowing (see Sect. 27.4 for details). In all probability, in such cases, the text is borrowed with the author's full consent, thus in the true sense of the word, no theft occurs of intellectual property. As for the second justification, although the copy-pasting of an entire text by another person is obviously incompatible with originality, borrowing some parts of it (such as the literary review or descriptions of procedures) is apparently possible without compromising the originality of the research results. One could therefore argue that the authentic reproduction of the whole text is much less serious than producing substantive original results, particularly in light of the aforementioned disagreements regarding the meaning of originality and authenticity. Despite a certain shakiness concerning the grounds on which it rests, the norm requiring full textual authenticity evolved in Western publishing, and it was officially supported by the VAK (*Vyssšaâ attestacionnaâ komissiâ*, All-Russian Attestation Committee)—a state agency based in Moscow that verifies both doctoral and post-doctoral degrees.[4]

Researchers who use the software and data described above could determine how closely the norm of textual authenticity was adhered to by large numbers of academics and identify the deviants who did not follow it. Two hypotheses could be posited here. The first is the "weakness hypothesis" that deviation from the norm of textual authenticity is associated with academic weakness. In other words, this concerns those authors who are unable to produce texts of an acceptable quality and therefore accept the risk associated with plagiarizing. In a slightly different form, this hypothesis predicts that when academics decide whether or not to plagiarize, they self-sort themselves into two groups. There are those for whom the costs of writing an authentic text are greater than the costs of being revealed as a plagiarizer, multiplied by the estimated probability of such a revelation. The second group consists of those for whom the opposite is true (Spence 1973, 2002). The "convention hypothesis," on the other hand,

holds that some academics disregard the norm because they disagree with its justifications, and may not be fully aware that others support it.

Several predictions follow from the "weakness hypothesis" as to where plagiarism is to be found. In the case of Russia, one would expect plagiarism to occur primarily in disciplines that were the least developed during the Soviet period, but which expanded after the collapse of the Soviet Union, that is, the social sciences. Second, one might expect less borrowing in institutions in which the prime research forces are concentrated, namely, the Academy of Sciences and the top universities. Third, individuals who conduct highly esteemed research are presumably less likely to borrow than those whose results are less prominent.

The "convention hypothesis" does not generate predictions, but it does explain why expectations based on the "weakness hypothesis" may be falsified. If no correlation occurs between borrowing and intellectual weakness, then the social sciences may not differ from the natural sciences, and the best institutions and scholars may not differ from their weaker counterparts. In this case, the principal variable deciding who plagiarizes and who does not is the degree of contact with Western academia and its standardized norms. Indeed, institutions which conduct the highest quality research are also likely to be more globalized. However, this correlation is probably weak, given that there are a few intervening variables.

To determine which hypothesis has more support, we analyzed 2,468 post-doctoral dissertations (*Doktor nauk*, see note 1 above), which were randomly selected from the pool of all dissertations defended in Russia in the years 2006–2015.[5] We utilized the antiplagiat.ru online service, which allowed us to assess the selected texts against many sources, including the Digital Dissertation Library of the RSL.

Figure 27.2 presents the overall distribution of plagiarism that occurred across disciplines. The figure in the graph is a boxplot. It divides the amount of borrowed materials found in each discipline into four quartiles, from the highest to the lowest, and indicates where the boundaries of each of them are situated. The band inside the box corresponds to the median, crosses (X) stand for averages, and points outside of the upper "whisker" are outliers with an extraordinarily high amount of borrowing for a given discipline. Three aspects of plagiarism immediately become apparent. First, inappropriate borrowing is almost universally present.[6] Second, the disciplines differ dramatically in what an "extraordinary" amount of borrowing means to them, such that the exceptional case of borrowing around 30 percent of a text in philology would be close to the average in agriculture. Third, cases of large-scale plagiarism similar to those discovered by Dissernet are rather rare. Thus, from the sample of 2,468 post-doctoral dissertations, we determined that 44 contained borrowing that exceeded 60 percent (1.7%). We checked these 44 manually, and three cases were false positives. Overall, large-scale plagiarism exceeding 50 percent was found in 149 of the 2,468 dissertations (6%). Thus, we further focus on relatively small-scale plagiarism.

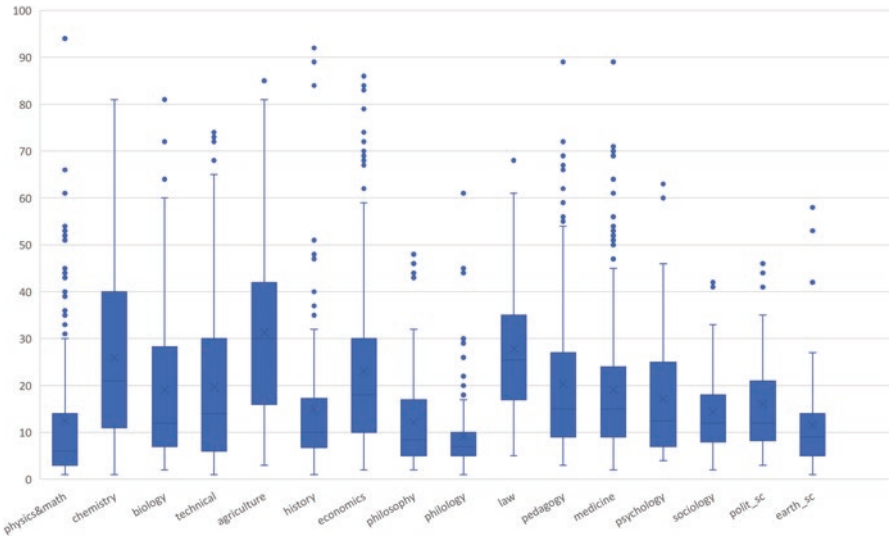

Fig. 27.2 The overall distribution of small-scale plagiarism across disciplines

In contrast to what is posited in the weakness hypothesis, no straightforward connections were discovered between the character of a discipline (humanities, social sciences, or natural sciences; predominantly theoretical or predominantly applied), the degree of its expansion in post-Soviet times, and the degree of plagiarism. Thus, of the three disciplines with the highest levels of unauthorized borrowing, agriculture (natural sciences, predominantly applied) has witnessed moderate expansion, chemistry (natural sciences, including both theoretical and applied subfields) is shrinking, and law (social sciences, both theoretical and applied) is expanding enormously. In the case of specific disciplines, apparently traditions play a key role, which sometimes differ in otherwise closely related subjects such as chemistry and biology, or economics and sociology. In general, it seems that neither the lower-level development of scholarship in a given field in Russia nor its recent expansion played a prominent role in tolerating unauthorized borrowings. The weakness argument does not appear to be valid for the moderate infringement of the norm for textual authenticity. It is interesting, however, that the logic does seem to be applicable in another sense: among the relatively sizable disciplines, philology (in Russia, this includes both literary studies and linguistics) displayed the least amount of borrowing.

We discovered some limited support for the hypothesis positing that aversion to plagiarism will be strongest among institutions of the Russian Academy of Sciences and universities that participate in Project 5-100, as their objective is to have at least five Russian universities among the top one hundred in the world university rankings. However, Russia's leading institutions are the most highly internationalized. For example, the top Russian universities are

evaluated according to the number of foreign students and faculty they employ. The leading institutions also serve as the gateways through which international norms find their way into Russia. In this sense, the "convention hypothesis" that resulted from the adoption of international norms may explain the aversion of these institutions to plagiarism (Table 27.4).

Finally, we examined individual publication profiles in the Russian Index for Scientific Citing. We selected 10 percent of the representatives of each discipline with the highest and the lowest number of borrowing and compared their publication profiles. The formulation of the sample thus eliminated the influence of differences in profile. Table 27.5 presents the results. Although some statistically significant differences emerged in the amount of plagiarism among researchers who publish widely in international publications compared to those who publish exclusively in second-rate domestic editions, such differences are relatively minor in absolute terms. Again, one could infer that according to the "convention hypothesis," scholars with the most impressive international publication records are also those with the highest exposure to the norms of international publication.

Overall, our findings cast considerable doubt on the validity of the "weakness hypothesis." It appears that the norm of textual authenticity is not widely accepted in Russia. Although borrowing larger amounts of a text (as in exceeding 50%) is rather rare, recycling the minor parts of other people's texts is almost a universal practice (probably 75% of dissertations include at least a few slightly re-written paragraphs from the works of others, without attribution).

Some Russian scholars justified borrowing by describing a dissertation and its public defense as a "mere formality" and decrying "senseless conventions." Others questioned the possibility of dividing collaborative work into personalized scientific contributions. There are no reasons to believe that the tendency to provide such explanations in any way correlates with the authors' intellectual competency. Regretfully, widespread tolerance toward borrowing in Russia greatly impedes the addressing of more notorious types of plagiarism because it renders the difference between borrowing some technical paragraphs and borrowing the whole text a matter of degree rather than a matter of principle.

Table 27.4 Percentage of borrowing in dissertations defended at various Russian institutions

Organization	*Median percent plagiarized (%)*
Top universities[a]	10.1
Russian Academy of Sciences	10.1
Other	15.9
ALL	*14.4*

[a]This includes 21 participants in Project 5-100, as well as Moscow and Saint-Petersburg state universities, in effect, 23 institutions in all.

Table 27.5 Differences in publication and citation performance among authors demonstrating the highest and the lowest amount of borrowing

Variable	*Averages*		*Average treatment effect (Δ)*
	Plagiarism top 10%	*Plagiarism bottom 10%*	
Publications RISC[a] core, %	19.96	24.99	−5.03*
Citing from RISC core, %	17.79	24.29	−6.50**
Impact factor, published	0.37	0.45	−0.07*
Impact factor, cited	0.42	0.52	−0.10*
Articles in foreign publications, %	3.74	6.84	−3.10***
Citing from foreign publications, %	6.94	10.4	−3.45**

*Statistically significant values; their numbers reflect the degree of confidence from less (*) to most (***) significant

[a]The Russian Index for Scientific Citing, RISC, includes a "core" of editions receiving the highest evaluations in a survey of Russian academics. It is also partially integrated with the Scopus and Web of Science databases, which enable the tracing of publications and citations from non-Russian-language editions.

27.6 Conclusion

The aim of this chapter was to describe the tools and resources that are available to detect plagiarism, as well as to establish how academic plagiarism in Russia, detected by automatic means, can be interpreted from different perspectives. The most visible manifestation of this, and the one that is most hotly debated in the media, is the spread of large-scale plagiarism in dissertations by those in power, who believe that possessing an academic degree will advance their careers. Less commonly discussed, but no less interesting, is the range of interpretations of the authenticity norm that underlies the notion of plagiarism. Tolerance toward utilizing someone else's text, which is evident in Russia, may be the sign of an impending global shift in academia, because it perfectly matches the Zeitgeist of digital post-modernity, or as Roland Barthes once observed: *La mort de l'auteur*, "the death of the author" (Barthes 1968).

Notes

1. Russia has two higher academic degrees: ***kandidat nauk*** (Candidate of Science, roughly equal to a Ph.D.) and ***doktor nauk*** (Doctor of Science, roughly equal to *doctor habilitatus* in some European countries). Henceforth in this chapter, we distinguish doctoral (=PhD) and post-doctoral (=***habilitation***) dissertations respectively.
2. This section is adapted from an article by M. Kopotev et al.; see (Smirnov et al. 2017).
3. There is currently no software that detects mathematical formulas. The possibilities for detecting "borrowed" graphs and figures are also rather limited, although the situation is rapidly changing (see, for example, Acuna et al. 2018; see also the survey by Eisa et al. 2015).

4. It is important to note that the verification does not extend to research papers.
5. The study reported in this section was conducted between March and November of 2018, at the Centre for Institutional Analysis of Science and Education in collaboration with the Centre for the Sociology of Education of the Russian Presidential academy. The authors would like to thank Katerina Guba, Alexandra Makeeva, Nadezhda Sokolova, and Anzhelika Tsivinskaya for their help in this project.
6. Antiplagiat produces some false positives. For example, it sometimes counts lists of referenced literature as borrowing, or it may not recognize alternative spellings of an author's name. We checked more than 800 dissertations manually and for most disciplines found medians that were approximately five percent lower than in the case of automatic search. However, the manual check was rather conservative and probably underestimated the scale of the borrowing. The actual statistics are therefore somewhere in between these estimates. There were no significant differences between the relative propensity of disciplines to borrow as estimated by automatic and manual procedures, with one notable exception: automatic checks probably overestimate the borrowing in dissertations on law, most likely due to the highly formulaic forms of speech in this genre.

References

Acuna, D. E., P. S. Brookes, and K. P. Kording. 2018. Bioscience-scale Automated Detection of Figure Element Reuse. Preprint at bioRxiv. Accessed February 1, 2019. https://doi.org/10.1101/269415.

Barthes, R. 1968. La mort de l'auteur. *Manteia* 5: 12–17.

Clauset, Aaron, Samuel Arbesman, and Daniel B. Larremore. 2015. Systematic Inequality and Hierarchy in Faculty Hiring Networks. *Science Advances* 1 (1): e1400005.

Denisova-Schmidt, E. 2016. Corruption in Russian Higher Education. *Russian Analytical Digest* 191: 5–9.

Eisa, Taiseer, Naomie Salim, and Salha Alzahrani. 2015. Existing Plagiarism Detection Techniques: A Systematic Mapping of the Scholarly Literature. *Online Information Review* 39: 383–400. https://doi.org/10.1108/OIR-12-2014-0315.

Firth, J.R. 1957. A Synopsis of Linguistic Theory, 1930–1955. In *Studies in linguistic analysis. A Special Volume of the Philological Society*, 1–32. Oxford: Oxford University Press.

Gipp, B. 2014. *Citation-based Plagiarism Detection. Detecting Disguised and Cross-language Plagiarism Using Citation Pattern Analysis.* Springer Fachmedien Wiesbaden. Accessed February 1, 2019. http://link.springer.com/book/10.1007%2F978-3-658-06394-8.

Golunov, S. 2014. *The Elephant in the Room: Corruption and Cheating in Russian Universities.* Columbia University Press.

Gupta, P., K. Singhal, P. Majumder, and P. Rosso. 2011. Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism. In *Proceedings of ICON-2011*, Macmillan Publishers, India. Accessed February 1, 2019. http://users.dsic.upv.es/~prosso/resources/GuptaEtAl_ICON11.pdf.

Kicherova, M., D. Kyrov, P. Smykova, et al. 2013. Plagiat v studenčeskih rabotah: analiz suŝnosti problemy [Plagiarism in Students' Papers: Toward the Roots of the

Problem]. *Online Journal Naukovedenie (IGUPIT)*, 4. Accessed February 1, 2019. http://naukovedenie.ru/PDF/83pvn413.pdf.

Korbut, A. 2013. Plagiat i konstitutivnyj porâdok dissertacionnogo teksta [Plagiarism and the Constitutive Order of a Dissertation Text]. *Sociologičeskoe obozrenie* 12 (2): 145–171.

Kutuzov, Andrey, Mikhail Kopotev, Tatyana Sviridenko, and Lyubov Ivanova. 2016. Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. Accessed February 1, 2019. https://arxiv.org/pdf/1604.05372.pdf.

Levenshtein, Vladimir I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics* 10 (8): 707–710.

Maloshonok, N. 2016. Kak vospriâtie akademičeskoj čestnosti sredy universiteta vzaimosvâzano so studenčeskoj vovlečennost'û: vozmožnosti konceptualizacii i èmpiričeskogo izučeniâ [How Perception of Academic Honesty at the University Is Linked with Student Engagement: Conceptualization and Empirical Research Opportunities]. *Voprosy obrazovaniâ* 1: 35–60.

McEnery, Tony. 2004. *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Routledge.

Modern Language Association. 2008. *MLA Style Manual and Guide to Scholarly Publishing*. New York: Modern Language Association of America.

Nikitov, A., et al. 2012. Plagiat v rabotah studentov i aspirantov: problema i metody protivodejstviâ [Plagiarism in Under- and Postgraduate Students' Papers: Problem and Actions Against]. *Universitetskoe upravlenie: praktika i analiz* 5: 61–68.

Oakes, M. 2014. *Literary Detective Work on the Computer*. Amsterdam: John Benjamins Publishing Company.

OED Online. June 2020. Oxford University Press. Accessed July 01, 2020. https://www.oed.com/view/Entry/272993?rskey=Ma8aD8&result=2&isAdvanced=false.

Palahniuk, Chuck. 1996. *Fight Club*. New York: W. W. Norton & Company.

Potthast, M., B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters: 997–1005*. Accessed February 1, 2019. http://www.uni-weimar.de/medien/webis/publications/papers/stein_2010p.pdf.

Smirnov, I., R. Kuznetsova, M. Kopotev, A. Khazov, O. Lyashevskaya, L. Ivanova, and A. Kutuzov. 2017. Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue-2017*, 285–297. Moscow: RSUH.

Spence, M. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87 (3): 355–374.

———. 2002. Signaling in Retrospect and the Informational Structure of Markets. *American Economic Review 92* (3): 434–459.

Weber-Wulff, D. 2014. *False Feathers. A Perspective on Academic Plagiarism*. Berlin: Springer.