https://doi.org/10.12697/JI.2019.28.03

Ilya Ilin
Doctoral student
University of Tartu

Aleksei Kelli
Professor of Intellectual Property Law
University of Tartu

The Use of Human Voice and Speech in Language Technologies:

The EU and Russian Intellectual Property Law Perspectives

1. Introduction

Language technologies (LTs) have become an integral part of our everyday lives.[*1] This article focuses on the legal aspects of these technologies. Several legal challenges related to LTs have already been extensively addressed (e.g., issues related to personal data, dissemination models, constitutional bases).[*2] The authors draw on previous research and extend it. In this paper, the authors concentrate on the legal status of voice and speech from an intellectual property[*3] (IP) perspective and on compatibility of the respective EU and Russian legal regimes. Because of this different focus of the article, it does not cover issues related to the protection of voice and speech in terms of personal data rights in the EU and Russia. These issues are analysed in a separate paper.[*4]

---

1 Examples of such technologies are automatic text translation, various services that provide language checks for writing, and applications that vocalise text with an integrated speech-to-speech translation function. In October 2017, Google demonstrated its brand-new headphones (Pixel Buds), which have an integrated speech-to-speech translation function.

2 See, e.g., J. Klavan, A. Tavast, A. Kelli (2018). The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. – *Frontiers in Artificial Intelligence and Applications* (307), pp. 71–78. Abstract available at http://ebooks.iospress.nl/volumearticle/50306 (18.5.2019); S. Calamai, C. Kolletzek, A. Kelli (2018). Towards a Protocol for the Curation and Dissemination of Vulnerable People Archives. In: Inguna Skadin, Maria Eskevich (eds.). *CLARIN Annual Conference 2018 Proceedings* (CLARIN Annual Conference 2018, 8–10 October 2018, in Pisa, Italy). CLARIN (pp. 77–81). Available at https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf (18.5.2019); A. Kelli, K. Lindén, K. Vider, P. Labropoulou, E. Ketzan, P. Kamocki, P. Straňák (2018). Implementation of an Open Science Policy in the Context of Management of CLARIN Language Resources: A Need for Changes? In: *Selected Papers from the CLARIN Annual Conference 2017* (pp. 102–111). Linköping, Sweden: Linköping University Electronic Press. Available at https://www.ep.liu.se/ecp/147/009/ecp17147009.pdf (18.5.2019); A. Kelli, K. Vider, H. Pisuke, T. Siil (2017). Constitutional Values As a Basis for the Limitation of Copyright within the Context of Digitalization of the Estonian Language. In: Kalvis Torgans (ed.). *Constitutional Values in Contemporary Legal Space II, 16–17 November 2016: Collection of Research Papers in Conjunction with the 6th International Scientific Conference of the Faculty of Law of the University of Latvia* (pp. 126–139). Riga, Latvia: University of Latvia Press. – DOI: https://doi.org/10.22364/cvcls.2.2016.

3 Intellectual property (IP) is defined as rights resulting from intellectual activity in industrial, scientific, literary, or artistic fields. Convention Establishing the World Intellectual Property Organization (signed in Stockholm on 14 July 1967 and as amended on 28 September 1979). Available at https://wipolex.wipo.int/en/text/283854 (10.5.2019). IP is traditionally divided into three main categories: 1) copyright, 2) related rights, and 3) industrial property. The article addresses copyright and related rights.

4 See I. Ilin, A. Kelli (2019). The Use of Human Voice and Speech for the Development of Language Technologies: The EU and Russian Data Protection Law Perspectives (forthcoming).

The article discusses the features of the regulatory framework regarding voice and speech in the EU and Russia from a copyright and related rights perspective.[*5] Russia and the EU have been chosen to explore the possibilities of co-operation in the field of LTs. Because the authors have an in-depth understanding of the Estonian copyright system, the Estonian Copyright Act[*6] (CA) is used as an example of implementation of the EU copyright directives.

Both the EU member states and Russia are parties to the majority of international conventions dealing with intellectual property regulation, including the Berne Convention on Literary and Artistic Works[*7] (Berne Convention), and the World Intellectual Property Organization (WIPO) Copyright Treaty[*8], which ensures common ground for intellectual property regulation.

The authors evaluate whether the EU and Russian IP laws are compatible with treating voice and speech as an input to the development of language technologies. This is relevant since co-operation between EU and Russian language-technology developers (firms, research institutions, etc.) is inevitable. The Russian language cannot be ignored in the development of contemporary and competitive LTs. Therefore, it is crucial to identify potential barriers to co-operation and map legal risks. The authors argue that the differences between the EU and Russian IP regulatory frameworks do not constitute major obstacles in joint activities to develop LTs.

The human voice and speech are legally complex phenomena in both jurisdictions. While a plethora of scholars have engaged in discussion of the legal nature of multimedia works[*9] that frequently assumes use of voice and speech, very little attention in academic literature is paid to the issues of voice and speech application particularly in the rapidly developing language technologies.

Voice and speech should be differentiated in terms of their origin. The voice refers to sound creation and speech to phoneme creation.[*10] Hence, a problem of delineating how these objects tie in with the intellectual property concept arises. Should they be considered a single object or two different objects? Speech without an oral component becomes written language (text) that in most cases is subject to copyright protection as a literary work.[*11] Voice without speech becomes a personality characteristic that involves a unique combination of voice patterns (vocal qualities, volume, speed, and so forth).

Without relation to speech, voice is not a result of creation by the human mind, and, therefore, it usually cannot be regarded as an object of intellectual property. There are some cases wherein voice is protected as intellectual property (e.g., the voice of a fictional character is protected by copyright[*12] or by trademark[*13] law). Most frequently, voice is treated as a personality characteristic, which is not covered as intellectual property.[*14] In this paper, voice is considered as a vocal element of the speech and is examined alongside it.

---

5 Within the context of the article, the reference to copyright should be interpreted as encompassing copyright and related rights both.

6 The Estonian Copyright Act (Autoriõiguse seadus). Entry into force: 12.12.1992. English translation available at https://www.riigiteataja.ee/en/eli/504042019001/consolide (18.5.2019).

7 Berne Convention for the Protection of Literary and Artistic Works (as amended on 28 September 1979). Available at https://wipolex.wipo.int/en/text/283698 (18.5.2019).

8 WIPO Copyright Treaty (adopted in Geneva on 20 December 1996). Available at https://wipolex.wipo.int/en/text/295166 (18.5.2019).

9 See, e.g., A.M. Eskicioglu (2003). Protecting Intellectual Property in Digital Multimedia Networks. – *Computer* 36 (7), pp. 39-45. – DOI: https://doi.org/10.1109/mc.2003.1212689; T. Aplin (2005). *Copyright Law in the Digital Society: The Challenges of Multimedia*. Bloomsbury Publishing; A.L. Moorthy, C.R. Karisiddappa (2005). The Relevance of Intellectual Property Rights in the Digital Millennium (International Conference on Information Management in Knowledge Society). Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.614.8596&rep=rep1&type=pdf (6.5.2019); S. Hideyasu, Y. Kiyoki (2008). Frameworks for Intellectual Property Protection on Multimedia Database Systems. – *Frontiers in Artificial Intelligence and Applications* 166, p. 181; E. Dementieva, Е. Дементьева (2016). Проблемы правового регулирования цифровой формы произведения. Интеллектуальная собственность ["Problems of Legal Regulation of the Digital Form of the Product: Intellectual Property"]. – *Авторское право и смежные права* [*Copyright and Related Rights*] 8, pp. 35–44; A. Nazarenko, А. Назаренко (2016). Проблемы правовой квалификации мультимедийных продуктов ["Problems of Legal Qualification of Multimedia Products"]. – *Интеллектуальная собственность. Авторское право и смежные права* [*Intellectual Property: Copyright and Related Rights*] 9, pp. 27–34.

10 A. Behrman (2017). Speech and Voice Science (3rd edn., Plural Publishing), p. 4.

11 Article 2(1) of the Berne Convention.

12 Kamina Pascal (2016). Film Copyright in the European Union. – *Cambridge Intellectual Property and Information Law* 33, pp. 111–113.

13 K. Foley (2009). Protecting Fictional Characters: Defining the Elusive Trademark–Copyright Divide. – *Connecticut Law Review* 41(3), pp. 923–960.

14 The voice as personality characteristic in some countries may refer to the image rights (e.g., Germany, Spain, France) and to the data protection legislation.

As objects of intellectual property, voice and speech have a dual meaning. On one hand, voice and speech might be used to create works or make works available to the public (e.g., interpretations, translations, performances). Then they mainly bring in the copyright and related rights concepts. When one speaks about the use of voice and speech in LTs, in this scenario they are processed by LT applications. In other words, voice and speech are considered to be input to the LT applications and, hence, become subject to copyright or related rights protection.

On the other hand, the samples of the human voice and speech are used for the creation of language resources[*15] (language datasets), and they constitute an element of the database. Language resources (LRs) are used to create LTs.

For reasons of space and different focus, the article does not address the entire process of the development of language technologies (from raw data to LT products) and the impact of the legal regime associated with the material used to create LT products. These issues are covered in other publications.[*16]

2. Protectability of voice and speech via copyright and related rights

To foster co-operation in the field of language technology between Russia and the EU, the treatment of voice and of speech from the perspective of copyright and related rights has to be similar between the two jurisdictions. This section comparatively analyses these regulations to identify potential incompatibilities.

The authors' aim is not to provide a comparison of all legal norms regulating copyright and related rights protection in the EU and Russia. The similarity of the legal grounds for such protection allows presuming that the regulations are similar to each other. For instance, a brief comparison of the legislation of Estonia (an example of an EU member state) and Russian regulation exemplifies that in both countries copyright does not require any official registration[*17], software and databases are protectable[*18], and the duration of copyright is the author's life plus seventy years after his or her death (70 years *post mortem auctoris*).[*19]

In this section, the authors concentrate on the material used to develop language technologies. The legal basis for the use of the material is analysed and compared in the next section.

2.1. Copyright protection

The key concept behind copyright protection is the originality of the work. Under the Berne Convention, the work may be protected by copyright if it fulfils the requirement of originality.[*20] The level of originality required is sometimes debated.[*21] The EU and Russian copyright legislation do not define originality.

15 For further discussion of the nature of language resources, see A. Kelli, K. Vider, K. Lindén (2015). The Regulatory and Contractual Framework As an Integral Part of the CLARIN Infrastructure. In: Koenraad de Smedt (ed.). *Selected Papers from the CLARIN Annual Conference 2015* (14–16 October 2015, Wroclaw, Poland). Linköping, Sweden: Linköping University Electronic Press (pp. 13–24). Available at https://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (6.5.2019).

16 A. Kelli, A. Tavast, K. Lindén, K. Vider, R. Birštonas, P. Labropoulou, I. Kull, G. Tavits, A. Värv (2019). The Extent of Legal Control over Language Data: The Case of Language Technologies. In: CLARIN conference proceedings (forthcoming); A. Kelli, A. Tavast, K. Lindén, K. Vider, I. Kull, G. Tavits, A. Värv, V. Mantrov, R. Birštonas (2019). Impact of Legal Status of Data on the Development of Data-Intensive Products: The Example of Language Technologies. In: *Latvian University Conference "Legal Science: Functions, Significance and Future in Legal Systems"* (forthcoming).

17 The Estonian Copyright Act, §7 (3); Article 1259 (4) of the Civil Code of the Russian Federation. – *The Civil Code of the Russian Federation* (Part I of IV) (Гражданский кодекс Российской Федерации (часть первая)) N 51-FZ, dated 30.11.1994. Adopted by the State Duma on 21 October 1994, signed by the President of the Russian Federation on 30 November 1994. Entry into force: 1.1.1995. Unofficial English translation available at http://www.wipo.int/edocs/lexdocs/laws/en/ru/ru083en.pdf&gt (6.5.2019).

18 The Estonian Copyright Act, §4 (3); Article 1259 of the Civil Code of the Russian Federation.

19 The Estonian Copyright Act, §38 (1); Article 1281 of the Civil Code of the Russian Federation.

20 WIPO Intellectual Property Handbook: Policy, Law and Use 2 (2004), paragraph 5.171.

21 See E. Rosati (2013). Originality in EU Copyright: Full Harmonization through Case Law. Edward Elgar Publishing; J. Street, K. Negus, A. Behr (2018). Copy Rights: The Politics of Copying and Creativity. – *Political Studies* 66 (1), pp. 63–80. – DOI: https://doi.org/10.1177/0032321717706012; A. Lukoseviciene (2017). On Author, Copyright and Originality: Does the Unified EU Originality Standard Correspond to the Digital Reality in Wikipedia. – *Masaryk UJL & Tech*. (11), pp. 215–242. –

The EU copyright rules are harmonised by the CJEU case law, where originality is mainly conceptualised with reference to the author's creativity. The concept of the author's creativity[*22] is also used in the Russian copyright rules, which presume the work should be a manifestation of the author's intellect and personality. Under this concept, the work should be new, original, unique, and creative. Therefore, voice and speech can be copyright protected if they are a part of the creative work. The quality of the work and its cultural and artistic merits do not create valid grounds for the exclusion of copyright protection in either of the jurisdictions. However, these works should express an idea or be a derivative work[*23] (e.g., translation, a work's adaptation).

The author's creativity – and, therefore, the originality requirement – is connected to the author's personality. In this regard, the question of the authorship of the work created by or with the help of language technology needs to be resolved. There are three scenarios for how the work can be created: language technology applications are used as a tool to create a work (e.g., usage of speech-to-text applications to dictate a novel), a language technology application creates a work with a human contribution (e.g., the human analyses the outcomes or selects the valuable results), and the language application creates a work by itself (e.g., sound and music creation[*24], automatic paper generators[*25], painting generation[*26], machine translation without human interaction).[*27]

In the first scenario, the author of the work is a person who used a language technology application to create a work. In this scenario, the creativity of the author and, therefore, the originality of the work can be easily identified; that makes the created work in most cases copyright protected. In the case of a work created by a technology application with a human contribution or on its own, the question of authorship becomes more complicated. The following example can be provided to illustrate the problem. The combination of speech synthesis, speech analysis, and speech recognition may create a situation wherein the voice from a video (e.g., a lecture, a movie, a performance) is captured, transformed into subtitles, translated, and then vocalised by a synthesised voice without any human interaction in this process. The process described involves three stages, with different results at the end. The first stage is transforming voice into text; the result is the initial text in a written form. The second stage is text translation; the result is translated text. The last stage is transforming the text into a voice; the result is vocalised text.

To be copyright protected, the result of every stage needs to be related to the author's creativity and be original. Human interaction needs to be evaluated for identification of the author. The majority of national jurisdictions in the EU rely on the concept under which the work might be protected by copyright only if it was created with a connection to the author's mind and personality (see the Estonian Copyright Act's §4 (2)), and some jurisdictions state that only humans can be the authors of a copyright-protected work (e.g., France, Germany, Spain, Estonia). The Russian copyright legislation too clearly identifies the author of the copyrighted work as a human.[*28] In this regard, the current EU and Russian copyright regulation do not deem computer-generated works copyright protected.[*29] If the minimum effort is put in, the person who

DOI: https://doi.org/10.5817/mujlt2017-2-2; K. Peifer (2014). "Individualität" [individuality] or Originality? Core Concepts in German Copyright Law. – *GRUR Int*. 63 (12).

22 For further discussion, see J. Street, K. Negus, A. Behr (2018). Copy Rights: The Politics of Copying and Creativity. – *Political Studies* 66 (1), pp. 63–80. – DOI: https://doi.org/10.1177/0032321717706012; G. Adomaitytė, V. Žilinskaitė, Ž Sederevičiūtė-Pačiauskienė, I. Valantinaitė, V. Navickienė (2018). Shift of Creativity Concepts: From Mysticism to Modern Approach. – *Filosofija. Sociologija* 29 (3). – DOI: https://doi.org/10.6001/fil-soc.v29i3.3777; N. Kawashima (2010). The Rise of "User Creativity" – Web 2.0 and a New Challenge for Copyright Law and Cultural Policy. – *International Journal of Cultural Policy* 16(3). – DOI: https://doi.org/10.1080/10286630903111613.

23 Article 2 of the Berne Convention.

24 E.g., Google research project Magneta, details available at https://opensource.google.com/projects/magenta (6.5.2019); Sony CSL music research project, details available at https://www.sonycsl.co.jp/tag/music/ (6.5.2019).

25 SCIgen – an Automatic CS Paper Generator. Available at https://pdos.csail.mit.edu/archive/scigen/ (6.5.2019).

26 M. Brown (2016). "New Rembrandt" to Be Unveiled in Amsterdam. Available at https://www.theguardian.com/artanddesign/2016/apr/05/new-rembrandt-to-be-unveiled-in-amsterdam (13.5.2019).

27 H.M. Bøhler (2017). EU Copyright Protection of Works Created by Artificial Intelligence Systems. University of Bergen, pp. 1–37. Available at http://bora.uib.no/bitstream/handle/1956/16479/JUS399_V17_183.pdf?sequence=1&isAllowed=y (6.5.2019).

28 Article 1257 of the Civil Code of the Russian Federation.

29 See H.M. Bøhler (2017). EU Copyright Protection of Works Created by Artificial Intelligence Systems. University of Bergen, pp. 1–37. Available at http://bora.uib.no/bitstream/handle/1956/16479/JUS399_V17_183.pdf?sequence=1&isAllowed=y (6.5.2019).

made this effort is deemed the author and the work may be copyright protected.[*30] Although there are also other legal issues related to works created by computers (liability for infringement, identification of the person liable, and so forth), they are not addressed here, on account of the focus of the article being elsewhere.

Moral rights constitute a legal challenge in the field of language technology in both jurisdictions. In Russia and the majority of the EU countries, copyright rights are divided into two separate groups: moral rights and economic rights.[*31] According to the Berne Convention,

> independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honour or reputation.[*32]

In other words, moral rights have a strong connection with the personality of an author.[*33] The scope of moral rights protected by copyright depends on the approach of the national legislation. It is possible to distinguish between Anglo-American (common law) copyright and the Continental-European *droit d'auteur* approach. The Anglo-American copyright tradition applies very limited protection of moral rights in comparison to the Continental-European approach.

The majority of EU countries and Russia belong to the Continental copyright tradition. This means the author's moral rights are integral to the author's person and non-transferable in both jurisdictions.[*34]

The strong connection to the author's person and the absence of a legal mechanism for the transfer of the moral rights creates the problem of how the moral rights might be exercised by the author in the case of the development of language technologies. Under the Russian national legislation, the law protects the following moral rights of an author: the right of attribution, the right to one's own name, a right to the integrity of the work, and a right to publish the work.[*35] That means that there is always the risk that authors (e.g., employees of the company, individuals who contributed to the product development) can claim the infringement of moral rights. This is related mainly to the integrity right, because there is a need to modify copyright-protected works (e.g., add annotations, metadata, and so forth) while developing language resources used to create language technologies. The exercise of moral rights by third parties can be identified as one of the main challenges connected with the development of language technologies.[*36]

The strategy for dealing with potential legal risks depends on the specific situation. If a company is developing the technology itself, then one way forward to address challenges in both jurisdictions is to obtain the author's prior agreement not to exercise his or her moral rights. This is not a clear-cut solution, but it could still mitigate some risks. European copyright scholars have even suggested a model addressing the consent connected with moral rights in the European Copyright Code (ECC).[*37] The European Copyright Code suggests regulating the exercise of moral rights as follows: 'The author can consent not to exercise his moral rights. Such consent must be limited in scope, unequivocal and informed' (Art. 3.5). The model provisions can be relied on as guidelines for drafting.

In the case of language resources acquired to develop LTs, due diligence is required, to clarify the legal situation with regard to moral rights (amendments/adaptations to the original works or agreements on the exercise of moral rights by third parties).

---

30 H.M. Bøhler (2017). EU Copyright Protection of Works Created by Artificial Intelligence Systems. University of Bergen, pp. 1–37. Available at http://bora.uib.no/bitstream/handle/1956/16479/JUS399_V17_183.pdf?sequence=1&isAllowed=y (6.5.2019). For instance, according to the UK copyright law, '[i]n the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken' (Art. 9 (3)). Copyright, Designs and Patents Act 1988. Available at https://www.legislation.gov.uk/ukpga/1988/48/section/9 (10.6.2019).

31 Some of the EU countries in their copyright regulations apply another approach, which presumes that the moral and economics rights are integral (e.g., Germany).

32 Article 6*bis* of the Berne Convention.

33 For further discussion on moral rights, see E. Adeney (2006). The Moral Rights of Authors and Performers. Oxford University Press; M.M. Walter, S. von Lewinski (2010). European Copyright Law: A Commentary. Oxford University Press; S. von Lewinski (2008). International Copyright Law and Policy. Oxford University Press.

34 The Estonian Copyright Act, §11 (2); Article 1265 of the Civil Code of the Russian Federation.

35 Article 1255 of the Civil Code of the Russian Federation.

36 For further discussion, see A. Kelli, T. Hoffmann, H. Pisuke, I. Kull, L. Jents, C. Ginter (2014). The Exercise of Moral Rights by Non-Authors. – *Journal of the University of Latvia* 6, pp. 108–125.

37 European Copyright Code. Available at https://www.ivir.nl/copyrightcode/european-copyright-code/ (14.5.2019).

2.2. Related rights protection

Voice and speech can be protected as objects of related (neighbouring) rights. While copyright protection has a strong connection to the author's personality, which results in the acknowledgement of moral rights, related rights are often connected to the beneficiary of these rights. Related rights mainly constitute economic rights. However, it should be mentioned that performers have moral rights, as well.

There are three groups of beneficiaries of related rights: performers, producers, and broadcasting organisations. The application of the protection of the related rights to the voice and speech in the field of language technology depends on how the voice and speech are treated within the technology. Two distinct scenarios can be outlined. The first scenario involves the voice and speech being used in the process of making works available to the public (e.g., in a situation in which the performance was recorded, the rights of the phonogram producers are protected by the related rights). In the second scenario, voice and speech are considered to be input to a digital language resource (an element of the database), and, therefore, *sui generis* databases[*38] fall within the field of related rights.[*39]

Language technologies (inclusive of those relying on voice and speech) are often used in the process of making works available to the public. Therefore, voice and speech are a part of this process. For instance, in cases of voice and speech that are part of a performance, recording, or broadcasting of an audiovisual work, the identification of legal risks depends on the particular scope of the related rights. This is connected with unlawful usage of the work itself (a publicly available work) and unlawful usage of the recording or broadcasting of the audiovisual work.

In the case of performer's rights, the situation is different, since it is crucial to consider the performer's moral rights as well. The question here is similar to problems of the author's identification as described above for copyright protection.

Voice and speech can also be viewed from the perspective of digital language resources (databases containing language data). The European copyright framework protects databases as copyright-protected works and by *sui generis* database rights. The *sui generis* protection relies on related rights.[*40] The latter options require that 'qualitatively and quantitatively a substantial investment' has been made in regard of the databases created.[*41] The Russian database regulation also presumes two options for database protection, by means of the copyright[*42] and by related rights protection, which refers to the concept of 'substantial investments' here also.[*43]

Samples of voice and speech may constitute content of databases. The EU database directive and the Russian regulation[*44] on databases define a database as a 'collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means'.[*45] The definition presents three main characteristics of a database: 1) independence, 2) systematic order, and 3) individual-level accessibility. The independence of the elements means that every element can be removed from the base without damage to other elements of the database[*46] (this distinguishes databases, for instance, from novels and films, as they too are composed of separate elements, such as chapters and soundtracks).[*47] The systematic order involves the elements being put into and classified in a specific order that allows searching the separate elements. At the same time, the European database directive, CJEU court practice[*48], and Russian database regulation do not define the character of the accessibility by electronic means.

38 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

39 As a matter of fact, language resources can cumulatively be protected as copyrighted and *sui generis* database.

40 Databases can be protected also by trade secret, competition, and contract law.

41 Article 7 of Directive 96/9/EC.

42 Article 1260 of the Civil Code of the Russian Federation.

43 Article 1334 of the Civil Code of the Russian Federation.

44 Part 4 of the Civil Code of the Russian Federation.

45 Article 2(1) of Directive 96/9/EC refers to Article 1260 of the Civil Code of the Russian Federation.

46 Case C-444/02, *Fixtures Marketing Ltd v. Organismos Prognostikon Agonon Podosfairou* [2004] ECLI:EU:C:2004:697; Case C-46/02, *Fixtures Marketing Ltd v. Oy Veikkaus AB* [2004] ECLI:EU:C:2004:694; Case C-338/02, *Fixtures Marketing Ltd v. Svenska Spel AB* [2004] ECLI:EU:C:2004:696); Case C-203/02, *The British Horseracing Board Ltd v. William Hill Organisation Ltd* [2004] ECLI:EU:C:2004:695.

47 E. Derclaye (2007). Intellectual Property Rights on Information and Market Power: Comparing the European and American Protection of Databases. – *International Review of Intellectual Property and Competition Law* 38 (3), pp. 275–298.

48 E. Derclaye (2002). What Is a Database? A Critical Analysis of the Definition of a Database in the European Database Directive and Suggestions for an International Definition. – *The Journal of World Intellectual Property* 5 (6), pp. 981–1011. – DOI: https://doi.org/10.1111/j.1747-1796.2002.tb00189.x.

There are two points that need to be examined from the perspective of language technology: the first one is how these databases are formed and the second is how the technologies that are built on these databases are further distributed to other language technology companies and users. These problems are addressed in part in the following section. However, the entrepreneurial aspects of such distribution fall outside the scope of the article and are investigated in further research.

3. The creation of digital language resources

The development of language technologies relies on the use of language resources. Language resources constitute a database consisting of text in written and oral form further used for the machine learning process. From the IP perspective, LRs may contain copyright-protected works, performances protected as objects of related rights, and personal data. Language resources are covered with two tiers of rights. The first tier of rights covers material containing language data (text, videos, voice samples, and so forth). The second tier of rights is related to the database itself. It is visualised with the following figure:[*49]

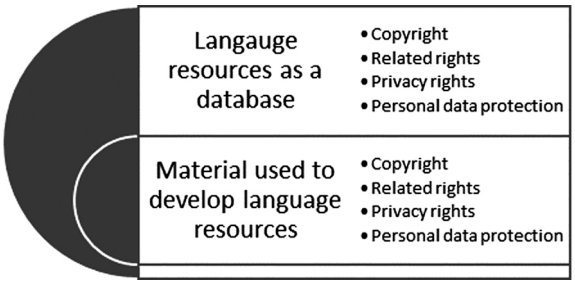

Figure 1: Two tiers of rights covering language resources

It should be emphasised that, to avoid legal risks, it is crucial to address legal issues pertaining to LRs themselves (the database) and the material used to develop LRs. When it comes to LRs as a database, the IP issues could easily be contractually regulated with individuals involved in their development. IP rights can be transferred or extensive licences acquired.

The situation is more challenging with the material used to create language resources. It involves copying of copyright-protected works. Since the focus of this article is on voice and speech, the performer's rights and phonogram producer's rights are relevant as well.

The copyright does not protect all works as such, and to be copyright protected, the work needs to constitute original results in the literary, artistic, or scientific domain (e.g., the Estonian Copyright Act, §4 (2)). In this regard, it is possible to specify three categories of works that can be used for the creation of the digital resources: the non-protected works (e.g., legal acts, official documents), 'safe' texts (manuals, technical documents, medical reports, etc.), and copyright-protected works.[*50]

From the technical perspective, on account of the vast volume of works in a language resource, it is a challenging task to identify the particular category of the works that are used for data mining. Even if it were possible, creation of a language resource based only on the non-protected and safe texts would be not sufficient to create a sizeable database of good quality. The development of language technologies requires language samples from everyday language usage, which most likely are copyright protected.

Language technology products (e.g., user interfaces, translation tools) do not necessarily contain copyright-protected content.[*51] However, if one is to create LT products, there is a need to use language resources having IP-protected material for text and data mining.

---

49 A. Kelli, K. Vider, K. Lindén (2015). The Regulatory and Contractual Framework As an Integral Part of the CLARIN Infrastructure. In: Koenraad de Smedt (ed.). *Selected Papers from the CLARIN Annual Conference 2015* (14–16 October 2015, Wroclaw, Poland). Linköping, Sweden: Linköping University Electronic Press, pp. 13–24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (6.5.2019).

50 M. Truyens, P. Van Eecke (2014). Legal Aspects of Text Mining. – *Computer Law & Security Review* 30 (2), pp. 153–170. – DOI: https://doi.org/10.1016/j.clsr.2014.01.009.

51 A. Kelli, A. Tavast, K. Lindén, K. Vider, R. Birštonas, P. Labropoulou, I. Kull, G. Tavits, A. Värv (2019). The Extent of Legal Control over Language Data: The Case of Language Technologies. In: CLARIN conference proceedings (forthcoming).

The Directive on Copyright in the Digital Single Market[*52] (Digital Copyright Directive) introduces the concept of text and data mining (TDM) at the EU level.[*53] The directive defines text and data mining as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations' (Art. 2 (2)).[*54]

Before looking at specific copyright exceptions used for TDM, it is necessary to evaluate the nature of text and data mining from a copyright perspective. Max Planck Institute experts have correctly pointed out that TDM is not copyright-relevant activity. They suggest that 'the automated analysis of these contents must be permitted, just as reading by the human being does not require any separate consent by the rightholder'.[*55] The main issue here is the right to make copies of copyright-protected works and objects of related rights (performances and phonograms).[*56] It should be mentioned that text and data mining could also interfere with the adaptation right (an economic right) and the integrity right (a moral right). The reason is that the material used for TDM should sometimes be annotated.

In a very general way, it can be said that the development of language resources can be based on the exception or consent model.

The Digital Copyright Directive set the following framework for text and data mining for research purposes (Art. 3.7):

1) the exception allows reproductions and extractions to be made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research[*57], text and data mining of works or other subject matter to which they have lawful access;
2) copies of works or other subject matter are stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results;
3) rightholders are allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted;
4) any contractual provision contrary to the exceptions is unenforceable.

The Russian national legislation has general exceptions allowing free use of copyright-protected works if this usage is not commercial (e.g., use for private, scientific, or cultural purposes).[*58] There is no specific text and data mining exception in Russian copyright law. This does not mean, however, that TDM is not allowed under Russian law, since text and data mining as such is not copyright-relevant activity. Basically, TDM means that certain patterns and information are derived from language data (often protected by copy-

52 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (text with EEA relevance). OJ L 130, 17.5.2019, pp. 92–125. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1558383919270&uri=CELEX:32019L0790 (20.5.2019).

53 Some EU member states (e.g., Estonia) have already introduced a provision on text and data mining (Estonian Copyright Act, §19, clause 3^1). These provisions need to be revisited in light of the Digital Copyright Directive.

54 The experts have pointed out that the concept 'text and data mining' is too narrow. Instead, 'data analysis' should be preferred. J.-P. Triaille, J. de Meeûs d'Argenteuil, A. de Francquen (2014). Study on the Legal Framework of Text and Data Mining, p. 8. Available at https://docplayer.net/16673528-Study-on-the-legal-framework-of-text-and-data-mining-tdm.html (6.5.2019). Although their argument is valid, the situation is that official documents use the term 'text and data mining'. Therefore, the latter term should be used.

55 R.M. Hilty, H. Richter (2017). Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules. PART B. Exceptions and Limitations. Max Planck Institute for Innovation and Competition, p. 3. Available at http://www.ip.mpg.de/fileadmin/ipmpg/content/stellungnahmen/MPI-Position-Paper_TDM_2017-01-14-corr_def.pdf (6.5.2019).

56 The Digital Copyright Directive follows the same line of argument: "Text and data mining can also be carried out in relation to mere facts or data that are not protected by copyright, and in such instances no authorisation is required under copyright law. There can also be instances of text and data mining that do not involve acts of reproduction or where the reproductions made fall under the mandatory exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC, which should continue to apply to text and data mining techniques that do not involve the making of copies beyond the scope of that exception" (Recital 9).

57 The Digital Copyright Directive specifies scientific research as follows: "The term 'scientific research' within the meaning of this Directive should be understood to cover both the natural sciences and the human sciences. Due to the diversity of such entities, it is important to have a common understanding of research organisations. They should for example cover, in addition to universities or other higher education institutions and their libraries, also entities such as research institutes and hospitals that carry out research. Despite different legal forms and structures, research organisations in the Member States generally have in common that they act either on a not-for-profit basis or in the context of a public-interest mission recognised by the State. Such a public-interest mission could, for example, be reflected through public funding or through provisions in national laws or public contracts" (Recital 12).

58 Article 1229, Article 1273, Article 1274, and Article 1306 of the Civil Code of the Russian Federation.

right or related rights). The key issue here is the legality of making copies of copyright-protected works and objects of related rights (e.g., performances). The Russian law allows scientific use of works, which also covers making copies for research purposes. For instance, the situation was similar in Estonia up until 2016 when a specific TDM exception was enacted.*[59] Prior to the introduction of the exception, copies for TDM were made under the research exception*[60] in the Estonian Copyright Act.*[61]

Although the EU and Russian approaches to research use (incl. text and data mining) of content protected by copyright and related rights are not identical, they are more or less compatible.

The creation of digital language resources can also be based on the consent of the holder of rights to the works and objects of related rights (licence). When language resources (a database) are created for commercial purposes, the contract model should be applied, since the creation cannot be based on the research exception. For example, in the case of Alisa (Yandex voice assistance), the assistant uses samples of voice taken not only from the app but also from other Yandex services, such as the Yandex navigation system and Yandex taxi service.

Individual consent does not always have to be negotiated. A report on a study of TDM points out that permissive Creative Commons (CC) licences facilitate the use of copyright-protected material without a need to rely on statutory exceptions.*[62] Since the focus of this article is on the comparison of the relevant EU and Russian law, contractual models to support TDM are not further explored.

In conclusion, it could be emphasised that the way in which language resources (databases) were created plays an essential role in the further distribution of language technologies. For instance, the speech recognition system developed by Yandex (SpeechKit) is distributed in three ways: as an API, a cloud service, and a program built on the client servers. If language resources are created unlawfully (protected material is used without proper legal basis), the further usage or resale of the products built on the language resources may constitute copyright infringement. Therefore, it is crucial to consider all intellectual property issues when developing language technologies.

4. Managing legal risks related to the creation and use of language technologies

There is no uniform model for how to manage legal risks arising from the use of material protected by copyright and related rights in the development of language technologies. Each case requires an individual assessment and analysis of the protectability of the material used, the legal grounds for use, and so forth. Despite the limitations cited, the authors still offer a model (in Figure 2) applicable in Russia and at the EU level to assess legal risks connected with the use of voice and speech for the development of language technologies:

59 The amendment was enacted with the passing of the Legal Deposit Copy Act (Säilituseksemplari seadus) on 15.6.2016. Available at https://www.riigiteataja.ee/en/eli/514092016001/consolide (10.5.2019).

60 Estonian Copyright Act, §19, clauses 2 and 3.

61 For further discussion, see A. Kelli, A. Tavast, H. Pisuke (2012). Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. – *Juridica International* XIX, pp. 40–48. – DOI: http://dx.doi.org/10.12697/issn1406-1082.

62 J.-P. Triaille, J. de Meeûs d'Argenteuil, A. de Francquen (2014). Study on the Legal Framework of Text and Data Mining, p. 27. Available at https://docplayer.net/16673528-Study-on-the-legal-framework-of-text-and-data-mining-tdm.html (6.5.2019).

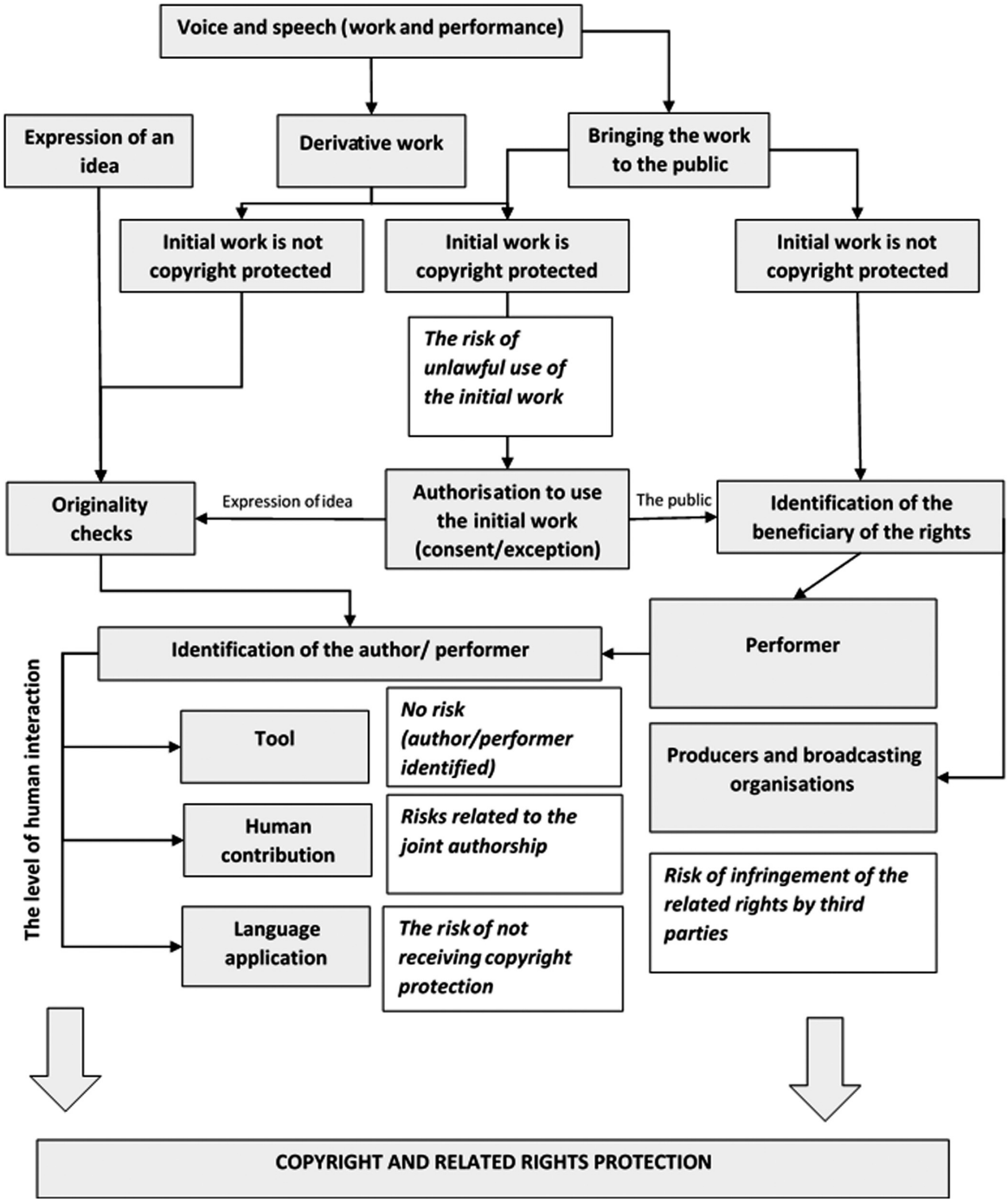

Figure 2: The process of assessing the protectability of the material used

The assessment starts with the identification of works containing voice and speech. Then the main characteristics of such a work are mapped: does it express an idea, or should the work be deemed a derivative work for which voice and speech are used in the process of making the work available to the public? After this, the following steps should be taken:

1) if the work expresses ideas, the originality test should be performed;
2) if the work is a derivative work, the authorisation for the use of the initial work should be checked;
3) if the voice and speech are used in the process of making the work available to the public, (a) the beneficiary of the related rights should be identified and (b) the authorisation for usage of the initial work should be checked.

These steps lead to the possibility of outlining and systemising the taxonomy in Table 1, presenting the groups of legal risks.

Table 1: Mapping of legal risks

Step:	Risk:	Solution:
Originality test	Risk of not acknowledging the copyright protection	(a) Identification of the author (b) Assessment of the level of the human original contribution
Derivative work	Risk of unlawful use of the work	Checking of authorisation
Bringing the work to the public	(a) Risk of not acknowledging the related rights (b) Risk of unlawful use of the work	(a) Identification of the author (b) Assessment of the level of the human original contribution (c) Checking of the authorisation

5. Conclusions

The authors have focused on the legal nature of voice and speech from the perspective of IP law in the EU and Russia. The article was aimed at determining whether there are incompatibilities of legal frameworks that have an adverse impact on the potential for co-operation between language technology developers from the two jurisdictions. Since voice and speech are intertwined with each other, they should be treated as one object. To be IP protected, the voice and speech need to be a part of a work or a database.

The comparison of the EU and Russian regulatory framework for copyright and related rights protection showed that these jurisdictions define the protectable subject matter similarly, and they both vest moral rights in the author. Therefore, the challenge in the development of language technologies covering European and Russian languages does not lie in the differences between the systems analysed but lies, rather, in the international foundation of the copyright system itself.

The authors have also compared the legal grounds applied for the creation of language technologies. The article has not conceptualised the entire process of development of LTs, which starts from collection of raw data (often containing text, speech samples, and so forth) and leads to specific products based on LTs. However, for identifying potential legal challenges, the concept of language resources (LRs) had to be introduced. LRs are database cumulatively protected by copyright (copyright-protected database) and by related rights (*sui generis* database). In a simplified way, it can be said that language resources are used to create language technologies.

LRs contain systematically arranged material protected by copyright and related rights. This means that LRs are covered with two tiers of rights: 1) database rights covering LRs themselves and 2) rights covering material used as input to LRs. Rights covering LRs themselves have to be managed contractually in both jurisdictions analysed (e.g., by means of transfer or licensing of IP rights). The issues related to materials (e.g., speech samples, videos, and so forth) are more complex and problematic.

Two possible legal grounds for the use of protected material in LRs that may be applicable can be distinguished in both jurisdictions: 1) exception and 2) consent. The EU and Russia both allow the use of objects protected by copyright and related rights for research purposes (exception model for the creation of LRs). The Digital Copyright Directive even introduced a specific exception for text and data mining. TDM is a core process for the creation of language technologies. Even though Russia does not have a specific TDM exception, the activities related to copying protected content for TDM are covered with a general research exception. The framework for consent to use material protected by IP is also similar between the jurisdictions.

The acknowledgement of protectability of voice and speech leads to identification of the common legal risks in the field of language technology relevant to both jurisdictions. The main risk is the use of copyright-protected works and objects of related rights without proper legal grounds. The regulatory framework of Russia and of the EU have an exception allowing the use of IP protected material for research purposes. The problem is that language technologies themselves are often used for commercial purposes. This, however, is not an issue of incompatibility of two different legal regimes.

The comparison of the intellectual property frameworks of the EU and Russia exemplifies that the basic understanding of copyright and the related rights concept is the same between these jurisdictions. Therefore, it can be concluded that regulatory incompatibilities in the field of copyright and related rights are not hampering the joint initiatives to develop LTs involving European languages and Russian.