https://doi.org/10.22364/iscflul.7.2.31

Aleksei Kelli, *Ph.D.*
University of Tartu, Estonia

Arvi Tavast, *Ph.D.*
Institute of the Estonian Language, Estonia

Krister Lindén, *Ph.D.*
Helsinki University, Finland

Ramunas Birstonas, *Ph.D.*
Vilnius University, Lithuania

Penny Labropoulou, *MSc*
ILSP/ARC, Greece

Kadri Vider, *MA*
University of Tartu, Estonia

Irene Kull, *Ph.D.*
University of Tartu, Estonia

Gaabriel Tavits, *Ph.D.*
University of Tartu, Estonia

Age Varv, *Ph.D.*
University of Tartu, Estonia

Vadim Mantrov, *Ph.D.*
University of Latvia

IMPACT OF LEGAL STATUS OF DATA ON DEVELOPMENT OF DATA-INTENSIVE PRODUCTS: EXAMPLE OF LANGUAGE TECHNOLOGIES[1]

Summary

The purpose of this article is to explain the extent to which the legal regime applicable to language data affects the development and use of language technology (LT). The main focus of the paper is on EU law. The article also maps possible text and data mining (TDM) issues. The authors focus on TDM for research purposes outlined in the Digital Copyright Directive 2019/790.

The authors follow a process approach of LT development, which starts from raw data collection and leads to LT products such as a refrigerator with a speech interface. Particular attention is given to language models.

The raw data used in LT often include copyright-protected works, objects of related rights (e.g., performances) and personal data in the form of person's voice or other information stored in non-annotated and annotated databases.

The authors' main argument is that the legal regime of language data does not usually affect the use of language models since copyrighted works are not likely to remain in models. In the process of developing a language technology application, language models are the first intermediate result that can be free from legal restrictions affecting language data.

[1] The article draws on and develops further the authors' previous research. See Kelli A., Tavast A., Lindén K., Vider K., Birštonas R., Labropoulou P., Kull I., Tavits G., Värv A. The Extent of Legal Control over Language Data: The Case of Language Technologies. Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019. Simov K., and Eskevich M. (eds.), CLARIN, pp. 69–74. Available at https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf [last viewed October 29, 2019].

The use of a person's voice as identifiable personal data in a language model can create legal challenges. In some cases, developers of language technology must be careful how to address issues of processing of personal data contained in models.

Keywords: data, data-intensive product, data protection, algorithm, language technologies

Introduction

The development of language technologies (LTs) relies on the exploitation of language data (LD). LD are often covered with several tiers of rights (copyright, related rights, personal data rights). The use of LD can be based on a consent or an exemption model.[2]

The issue we explore in this article concerns the impact of the legal regime of data on LTs. The question is whether legal restrictions applicable to data also apply to the LTs that are developed using them. The article aims to reduce the legal uncertainty regarding how far, in the pipeline of developing LTs, the original copyright and personal data (PD) protection[3] regulations apply. If we take a recorded phone call, for instance, it is evident that copyright and PD protection apply to a copy of that recording. At the other extreme, it is equally apparent that they do not apply to the Voice UI (User Interface) of a new fridge, even though the latter was trained on a data-set containing the former. The line where the original rights cease to apply has to be somewhere between these points, and it is vital for researchers and developers to know where.

To place the legal analysis into the technological context, it is essential to understand the process of development of LTs. The development of LTs can be divided into the following phases:

Collection/Creation of raw data (written texts, speech recordings, photos, videos, etc.). These often contain copyrighted material and personal data. Their development usually does not involve any other activities than the actual recording, initial cleaning and sanity-checking of the data.

---

[2] For further discussion, see Kelli A., Vider K., Lindén K. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. De Smedt K. (ed.), Linköping University Electronic Press, Linköpings universitet, 2015, pp. 13–24. Available at: http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 [last viewed November 2, 2019]; Kelli A., Vider K., Pisuke H., Siil T. Constitutional Values as a Basis for the Limitation of Copyright within the Context of Digitalization of the Estonian Language. In: Chair: Prof. Dr. habil. iur. Kalvis Torgans, University of Latvia, Latvia (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, 2017.

[3] The GDPR defines personal data as "any information relating to an identified or identifiable natural person ('data subject')" (Art. 4 (1)).

Dangers with regard to copyright and PD protection can be very real: re-publication of copyrighted works, surveillance by governments or insurance companies, and so forth.

There is a possibility to identify significant portions of copyrighted works. It is almost impossible to anonymise or pseudonymise completely so that it would become mathematically impossible to identify any persons.

Compiling of data-sets, or collections of data (raw text corpora like Google News, Common Crawl or OpenSubtitles, speech corpora like the Prague DaTabase of Spoken Czech, etc.). Data such as the above, but collected and organised with a specific criterion in mind (e.g. speech recordings on a specific topic by residents of a certain region in order to capture the accent of that region); these data-sets usually come in such quantities that any individual piece of data constitutes a negligible part of the whole, and could, in principle, be removed without affecting the usability of the data-set.

For copyright and PD purposes, data-sets are not different from raw data[4]. The main practical difference is that the sheer volume of data may make it technically difficult for individuals to become aware that their data have been included in the data-set.

Creation of a data-set often involves a nontrivial contribution in gathering, organising, indexing, presenting, hosting, etc. of the data.

Creation of annotated data-sets (POS-tagged corpus of written texts like the ENC17, syntactically parsed corpora like the Universal Dependencies treebanks, etc.). The above category augmented with some analysis.

Again, annotated data is not different from raw data in terms of copyright and PD, although the copyright holders of the raw data and the annotations may be different. The annotation layers may be stored separately and may even have some use on their own, but the usual practice is to produce copies of the original data together with the annotation layers so that the resulting dataset contains all of the original data.

Creation of an annotated data-set includes analysis of the data, either manual, semi-automatic or automatic.

Models. Data products developed from some processing of the above, but not necessarily containing the above, which try to *model*, i.e. represent or describe, language usage[5]. Examples: dictionaries, wordlists, frequency distributions, n-gram

[4] In fact, it can be argued that datasets qualify for database protection (for further discussion, cf. Eckart de Castilho et al. 2018).

[5] It should be noted that for the legal purposes of this article we use a broad definition of models, while in the literature of Natural Language Processing, the term "model" is usually used for Machine Learning models mainly.

lists like Google ngrams, pre-trained word embeddings like in Grave et al.[6], pre-trained language models like in Devlin et al.[7]

The creation of a model involves significant amounts of work, expertise and (computational) resources. Steps include, at least, creation and/or selection of the algorithm, implementation of the algorithm in software, hardware setup (which may even include custom hardware development), hyperparameter optimisation, and model validation.

In rare cases, some model types may be consumer products of their own (e.g., dictionaries). Mainly, however, models are used in downstream tasks to create other products.

Semi-finished products (text-to-speech engine or a visual object detector) and **finished products** (talking fridge). These are out of scope for the current analysis, because their status as original works should be beyond doubt.

The authors' main argument is that the legal regime of LD does not usually affect the use of language models. LD may be covered with different rights (copyright, related rights, PD protection). However, after language models are developed using the referred data (e.g., relying on research exception), they can be used without copyright and PD law restrictions, unless models contain identifiable material protected by copyright and PD.

To comprehensively address this crucial issue, the international team of researchers consists of experts with different backgrounds covering law and technology. Therefore, it is possible to discuss the latest technological developments and relevant regulatory framework. The main focus is EU law. Particular attention is given to the Directive on Copyright in the Digital Single Market[8] (Digital Copyright Directive, DCD) since it introduces a new regulation on text and data mining (TDM).

1. Copyright protection of language data and definition of models

The authors explore the impact of copyright law on LD. The first essential principle, which is well-established in international and national copyright law, is that the mere data are not copyrightable. Here it is appropriate to note, that the concept of "data" can be interpreted in a broad and a narrow way. Interpreted broadly, the concept of data encompasses copyrightable works, data in a narrow

[6] Grave E., Bojanowski P., Gupta P., Joulin A., & Mikolov T. Learning word vectors for 157 languages, 2018. ArXiv Preprint ArXiv:1802.06893.

[7] Devlin J., Chang M.-W., Lee K., & Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. ArXiv:1810.04805 [Cs].

[8] 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. OJ L 130, 17.5.2019, pp. 92–125. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790 [last viewed October 29, 2019].

sense and all kinds of other materials.[9] Data, understood in a narrow sense, means only non-copyrightable pieces of information, such as numbers, names, addresses, single words or sounds, and so forth. Since data in a narrow sense are not protected by copyright, they could be freely used as LD. However, there is a problem. Single pieces of data are not usually found in isolation, but they are in combination with or within the copyrightable materials and their separation in practice can be very complicated or even impossible.

The second long-established requirement is that of **originality**. Work is protected if, and only if, it is original. Therefore, the originality requirement defines the copyright status of the input data. Oddly enough, this general requirement was never defined in international treaties or European *acquis*.[10] The task to define the legal meaning of originality for copyright purposes was mainly taken by the Court of Justice of the European Union (CJEU). As was explained in the seminal decision of the *Infopaq* case[11], originality means the author's own intellectual creation. In turn, the "author's own intellectual creation" presupposes the expression of the author's creative abilities in the production of the work by making free and creative choices.[12] In one of the last decisions, CJEU has explained, that in order to determine the originality of the textual material, the national court should ascertain whether, in drawing up such materials, the author was able to make free and creative choices capable of conveying to the reader the originality of the subject matter at issue, the originality of which arises from the choice, sequence and combination of the words by which the author expressed his or her creativity in an original manner.

On the contrary, if the materials under consideration constitute purely informative documents, the content of which is primarily determined by the information which they contain, so that such information and the expression of those materials become indissociable and that those materials are thus entirely characterised by their technical function, originality is missing.[13]

Another important statement in the *Infopaq* case was that an extract consisting of eleven words could constitute an original work. The Court has also explained that a single word cannot be regarded as original and protectable work.

In the context of the current research, the originality requirement is important from two different perspectives. First, if originality is missing, the pre-existing text contained in a data-set is not protected and can be used without authorisation. Therefore, even if parts of this text are reproduced in the model, they are not

---

[9] Art. 1(2) of the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases defines the database as the collection of independent works, data or other materials.

[10] Although it was defined in several EU directives with regard to specific categories of works, such as computer programs or photographic works.

[11] CJEU judgement of 16 July 2009 in case No. C-5/08. *Infopaq International A/S vs. Danske Dagblades Forening*.

[12] CJEU judgement of 1 December 2011 in case No. C-145/10 *Eva-Maria Painer vs. Standard VerlagsGmbH et al.*

[13] CJEU judgement of 29 July 2019 in case No. C- 469/17 *Funke Medien NRW GmbH vs. Bundesrepublik Deutschland*.

protected as well. Second, even if a text as a whole is original and, therefore, protected, the question remains, whether the fragments used in the model are original on their own. If they are not, then again, they can be used without authorisation. Thus, originality must be established not only concerning the original work but also as regards the parts used.

In addition, in its latest case law CJEU has underlined, that, besides originality, a work also must meet the third requirement in order to be copyright-protected, i.e. it

> *must be expressed in a manner which makes it identifiable with sufficient precision and objectivity, even though that expression is not necessarily in the permanent form.*[14]

Arguably, this requirement in practice will be present in the majority of cases, because the texts (or other materials) used for models typically are expressed in a fixed form.

It should also be borne in mind that the protectability of works is usually presumed. For instance, according to the Copyright Act of Estonia[15]

> *The protection of a work by copyright is presumed except if, based on this Act or other copyright legislation, there are apparent circumstances which preclude this. The burden of proof lies on the person who contests the protection of a work by copyright.*[16]

Similarly, the Latvian Copyright Act[17] provides that copyright shall apply to works of literature, science, art and other works referred to in Article 4 of this Act, also unfinished works, regardless of the purpose of the work and the value, form or type of expression.[18] To put it differently, it is up to a person using LT to prove that it is not copyright protected. In practice, this point is very complicated.

LD are used to develop models. Models are the main focus of our study. Language models are a major intermediate result in developing LTs. They aim to describe language, like the models of physics aim at describing physical reality. Like modelling in other research fields, the creation of language models is not possible without extensive data processing, which may often be the last step in creating the model.

Due to their heterogeneous typology and frequent development of new types, models are not easy to define. Broadly, a model is a data product aimed at describing something – like natural language in the case of language models. Traditionally,

---

[14] CJEU judgement of 13 November 2018 in case No. C-310/17. *Levola Hengelo BV vs. Smilde Foods BV*.

[15] The Copyright Act of Estonia. English translation. Available at: https://www.riigiteataja.ee/en/eli/504042019001/consolide [last viewed November 3, 2019].

[16] Copyright Act of Estonia § 4 (6).

[17] Latvian Copyright Act. English translation available at: https://vvc.gov.lv/image/catalog/dokumenti/Copyright%20Law.doc [last viewed November 11, 2019].

[18] Art. 2(2) Latvian Copyright Act.

such descriptions as dictionaries and grammars were created using pen and paper, previously with basic tools, like typewriters and text processors, now increasingly using more complex tools like machine learning software. Examples of models used in language technology include the following:

a) Dictionaries and grammars (both traditional and now increasingly machine-readable) provide information about words in a natural language and how they are used.

b) Frequency lists and co-occurrence lists contain words or short sequences of words with information about how often they occur in texts, including how often they occur next to each other;

c) Word embeddings are currently a popular type of model, listing words like above, but providing each with a set of numbers that try to capture the meaning of the word, based on how it has been used in texts. Words with more similar meanings have more similar numbers, and various interesting operations on the embeddings turn out to be possible, like "king" – "man" + "woman" = "queen";

d) Speech recognition uses several models, one of which is the acoustic model, providing statistical information that relates pieces of audio signals to phonemes or other linguistic units that make up speech.

Developing a model includes substantial intellectual effort on the part of the developer, including one or more of the following depending on the type of model: choice/creation of the dataset, choice/creation of the algorithm, choice/creation of its software implementation to be used for the training of the model and various cycles of testing and validation by tuning the parameters of the software.

Just like it is possible for a text to be too short or trivial or limited in creative choices to qualify as an original work, some models (like a simple frequency list) may also be too simple or too limited in options.[19] In nontrivial cases, the *de facto* situation is that models are made available together with the research papers describing them and the software tools used in their creation. Standard licenses applied to models by their creators include Creative Commons – Attribution-ShareAlike 4.0

[19] Cf. De Castilho R. E., Dore G., Margoni T., Labropoulou P. & Gurevych I. A legal perspective on training models for Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, ELRA. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf [last viewed October 29, 2019.]

International[20], Apache License 2.0[21] and Public Domain Dedication and License v1.0[22].

A question might be raised whether models constitute derivative works. There is no clear definition of derivative work in international or European legal acts, and different jurisdictions have a quite different understanding of this concept.[23] It is not clear how much of the original work should remain to categorise a model as a derivative work. Models cannot be considered derivative works, if representations of linguistic units in the model are kept so short (e.g. individual words) that they cannot be considered original parts of the underlying texts.

To give a definite answer, we should have a closer look into all the model types and the processes and resource types and modalities they have been built upon, which is not possible within the limits of this article. It can be argued, though, that models by definition try to capture *generalities* of language use and *abstract* from the original texts as far as possible, producing mainly patterns with statistical measures.

2. Legal bases to use copyright-protected language data to develop language models

There are several legal grounds to use copyright-protected LD for the development of LTs. These grounds can be visualised in the following figure:

[20] E.g., Grave E., Bojanowski P., Gupta, P., Joulin A., & Mikolov T. Learning word vectors for 157 languages, 2018. ArXiv Preprint ArXiv:1802.06893; Kondratyuk, D and Straka, M. UDify Pretrained Model, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2019. Available at: http://hdl.handle.net/11234/1-3042 [last viewed November 3, 2019].

[21] E.g., Devlin J., Chang M.-W., Lee K., & Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. ArXiv:1810.04805 [Cs]; Ulčar, M. ELMo embeddings model, Slovenian, Slovenian language resource repository CLARIN.SI, 2019. Available at: http://hdl.handle.net/11356/1257 [last viewed November 3, 2019]; Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., & Le Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding, 2019. ArXiv:1906.08237 [Cs]. Available at: http://arxiv.org/abs/1906.08237 [last viewed November 3, 2019].

[22] E.g., Pennington J., Socher R, and Manning C. D. GloVe: Global Vectors for Word Representation, 2014. Available at: https://nlp.stanford.edu/projects/glove/ [last viewed November 3, 2019].

[23] For further discussion, see Birštonas R., Usonienė J. Derivative Works: Some Comparative Remarks from the European Copyright Law. *UWM Law Review,* Vol. 5, 2013; De Castilho R. E., Dore G., Margoni T., Labropoulou P. & Gurevych I. A legal perspective on training models for Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, ELRA, 2018. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf [last viewed October 29, 2019].

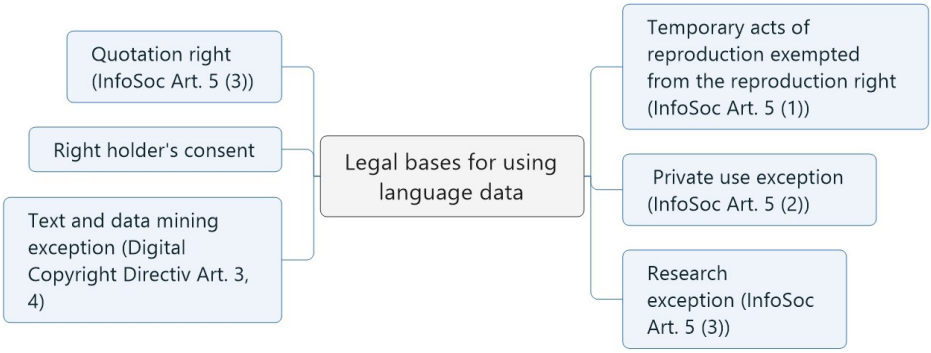

Figure 1. Legal bases for using language data

Generally speaking, these grounds can be divided into two main categories: 1) use of copyright-protected LD is based on consent; 2) use of copyright-protected LD is based on a copyright exception.

The acquisition of consent is the most respectful of the interests of the copyright holder. However, it is not always possible (e.g., anonymous blog posts and comments and so forth) or administratively (large number of works) possible to acquire consent. Therefore, the development of language technology is often based on copyright exceptions.[24] The main focus of the article is on the exceptions used to develop language models. Particular attention is given to the new text and data mining (TDM) regulation in the new Digital Copyright Directive.

From a copyright perspective, the development of LTs involves a TDM process. The Digital Copyright Directive defines TDM 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.*[25]

---

[24] For further discussion see, See Ilin I., Kelli A. The Use of Human Voice and Speech in Language Technologies: The EU and Russian Intellectual Property Law Perspectives. *Juridica International,* No. 28, 2019, pp. 17–27. Available at: https://www.juridicainternational.eu/public/pdf/ji_2019_28_17.pdf [last viewed October 29, 2019]; Kelli A., Vider K., Pisuke H., Siil T. Constitutional Values as a Basis for the Limitation of Copyright within the Context of Digitalization of the Estonian Language. In: Chair: Prof. Dr. habil. iur. Kalvis Torgans, University of Latvia, Latvia (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 Riga, Latvia: University of Latvia Press, 2017, pp. 126–139; Kelli A., Tavast A., Pisuke H. Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. *Juridica International,* No. 19, 2012, pp. 40–48. Available at: https://juridicainternational.eu/public/pdf/ji_2012_1_40.pdf [last viewed November 2, 2019]; Kelli A., Vider K., Lindén K. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. Koenraad De Smedt (ed.), Linköping University Electronic Press, Linköpings universitet, 2015, pp. 13–24. Available at: http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 [last viewed November 2, 2019].

[25] The Digital Copyright Directive Art. 2 (2).

It should be kept in mind that TDM as such is not a relevant activity from the copyright perspective.[26] Since the performance of TDM requires copying of the content (often copyrighted material), the reproduction right needs to be limited. The following legal grounds are explored for the purpose of limiting reproduction rights for TDM:

2.1. Temporary reproduction right

The InfoSoc Directive[27] obliges the EU Member States to limit the reproduction right so that it does not cover temporary technical copies which have no independent technological significance.[28] The use of this legal ground for TDM is also emphasised in the Digital Copyright Directive.[29] The usability of this ground for LT development is also acknowledged by technology experts.[30]

2.2. Private use exception

LT development can be based on the private use exception as well. This is relevant in countries with a very limited research exception. The InfoSoc Directive sets forth private use as an optional exception to the reproduction right that the EU Member States can adopt. The exception can be relied on by a natural person for private use without commercial purpose. The rightholders are entitled to fair compensation (Art. 5 (2) (b)), so the exception is rarely applied. If a country has

[26] It is explained in the Digital Copyright Directive that "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" (Recital 9).

[27] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisa-tion of certain aspects of copyright and related rights in the information society. Official Journal L 167, 22/06/2001 pp. 0010–0019. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 [last viewed October 29, 2019].

[28] The exact provision in the InfoSoc Directive reads: "Temporary acts of reproduction referred to in Article 2 [reproduction right – the authors' addition], which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable:

(a) a transmission in a network between third parties by an intermediary, or

(b) a lawful use of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2" (Art. 5 (1)).

[29] According to Digital Copyright Directive "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).

[30] See De Castilho R. E., Dore G., Margoni T., Labropoulou P. & Gurevych I. A legal perspective on training models for Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, ELRA, 2018, pp. 1272–1273. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf [last viewed October 29, 2019].

enacted a research exception, that is a better ground, since research institutions cannot rely on the private use exception.

2.3. Quotation right

The Berne Convention[31] regulates the quotation right at the international level. Article 10 (1) of the Convention allows quotations but requires, among other things, that quotations must be compatible with fair practice, and that they do not exceed the extent justified by the purpose. The InfoSoc Directive allows the EU Member States to introduce the quotation right, but quotations must be limited to criticism or review.[32] The quotation right is differently implemented in national laws. For instance, the Estonian Copyright Act does not require any purpose for quotation (e.g., criticism) but the author of the quoted work needs to be attributed, the quoted work has to be lawfully published, and the quotation should not exceed the justified extent.[33] The Latvian Copyright Act provides similar regulation by adding that right of quotation shall be permitted in works "created and used in the face-to-face teaching and research process in educational and research institutions for non-commercial purposes".[34] There are jurisdictions where the quotation right has more limitations (e.g., Lithuania). In case the quotation right does not have too restrictive requirements, it can be used to compile data-sets containing LD and use it for the development of LTs.

2.4. Research exception

This exception is often used when a legal ground is needed for TDM. The research exception is provided in the InfoSoc Directive as non-mandatory for the EU Member States. According to the wording of the InfoSoc Directive, EU Member States may allow the use of works

> *for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved.*[35]

The research exception is not a panacea for TDM. The current framework has several limitations, such as the exclusion of a commercial purposes, which has an adverse impact on industry-academia cooperation.

---

[31] Berne Convention for the Protection of Literary and Artistic Works. Available at: https://wipolex.wipo.int/en/text/283698 [last viewed October 29, 2019].

[32] The InfoSoc Directive Art. 5 (3) clause d.

[33] The Copyright Act of Estonia § 19 clause 1.

[34] Art. 21(1) Latvian Copyright Act.

[35] The InfoSoc Directive Art. 5 (3) clause a.

2.5. TDM exception

The Digital Copyright Directive has two mandatory TDM exceptions. One is meant for research and cultural heritage institutions (Art. 3) and the other for everyone (Art. 4). Since the focus of the current article is on the research context and due to limited space, the authors concentrate on TDM for research purposes. The TDM regulatory framework is visualised in the following figure:

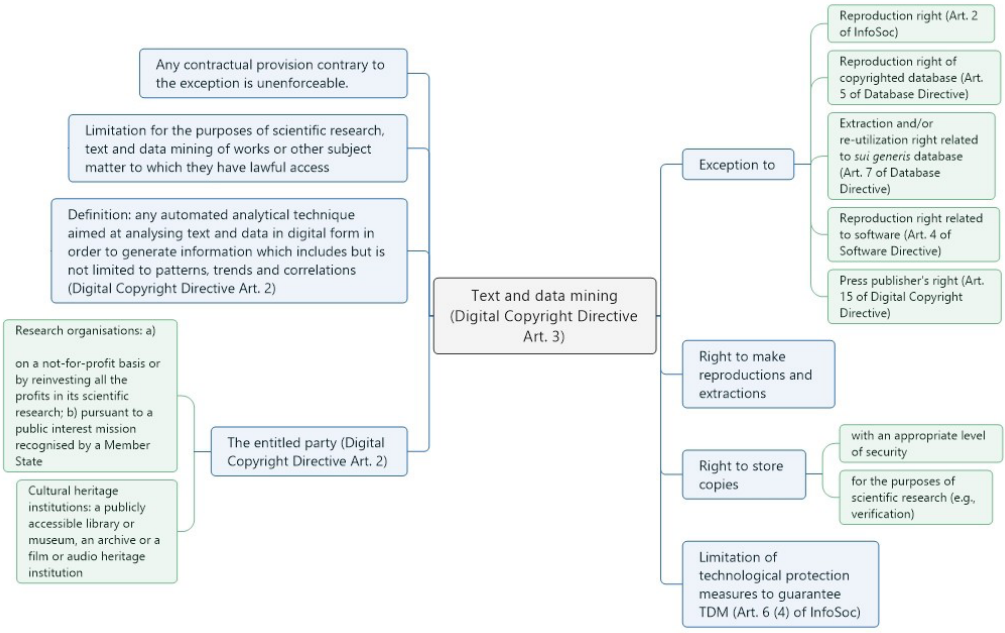

Figure 2. Text and data mining for scientific research

According to the Digital Copyright Directive research, organisations and cultural heritage institutions[36] are entitled to rely on this exception. The Directive defines research organisations extensively. The requirement is that research is conducted

> *on a not-for-profit basis or by reinvesting all the profits in its scientific research; or pursuant to a public interest mission recognised by a Member State in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis.*[37]

[36] The Digital Copyright Directive defines cultural heritage organisations as "a publicly accessible library or museum, an archive or a film or audio heritage institution" (Art. 2 (3)).

[37] The Digital Copyright Directive Art. 2 (1).

The Digital Copyright Directive Art. 3 (1) allows making copies of works, objects of related rights (e.g., performances), press publications[38] and extractions from *sui generis* databases for TDM for scientific research. The key issue here is that access to the material has to be lawful.

There are remedies in case rightholders adopt measures limiting the TDM exception. According to 7 (1) of the Digital Copyright Directive, any contractual provision contrary to the exception is unenforceable. The situation is more nuanced with technological measures.[39] The Digital Copyright Directive Art. 3 (3) allows rightholders

> *to apply measures to ensure the security and integrity of the networks and databases where the works or other subject-matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.*

The question is what happens if rightholders go beyond what is allowed by the Directive. According to the InfoSoc Directive Art. 6 (4) Member States

> *take appropriate measures to ensure that rightholders make available to the beneficiary of an exception or limitation.*

It should be mentioned that the practical application of this requirement is not so efficient. There are few efficient mechanisms to compel rightholders to adopt technological measures to allow the free use prescribed by law.

A key issue for language research relates to the use of compiled data-sets exploited for TDM. The question is, what can be done with data-sets. The Digital Copyright Directive Art. 3 (2) provides that

> *Copies of works or other subject-matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.*

The Directive does not say clearly whether data-sets can be shared among researchers. This is a genuinely crucial issue since research and research infrastructures[40] are based on the ideology of sharing research data. It remains to see how the national legislators implement the provision. The research community should use all possible measures to introduce a regulation which allows at least limited sharing.

---

[38] The right to press publications is introduced with the Digital Copyright Directive Art. 15.

[39] The InfoSoc Directive Art. 6 (3) defines technological protection measures as "any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject-matter, which are not authorised by the rightholder".

[40] E.g., CLARIN (Common Language Resources and Technology Infrastructure). European Research Infrastructure for Language Resources and Technology. Additional information available at: https://www.clarin.eu/ [last viewed November 3, 2019].

The TDM exception is not limited to non-commercial activities. The Directive allows for public-private partnerships. This means that research organisations can collaborate with private partners to carry out the TDM.[41]

3. Protection of personal data remaining in language models

Personal data issues relating to language technology with a particular emphasis on voice have been previously studied.[42] Therefore, PD protection is covered to the extent needed for this article. The following figure summarises the main aspects of PD processing for research purposes:

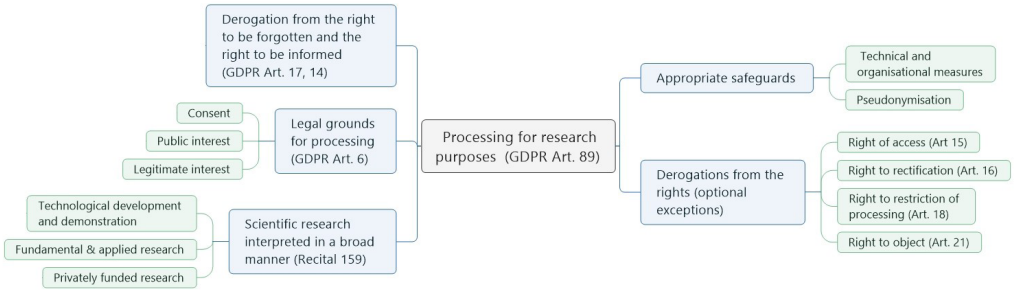

Figure 3. Processing personal data for research purposes

Article 4 (14) of GDPR defines "biometric data" which means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data. The human voice can be considered biometric data as it contains information regarding a person's physiological characteristics which make that person distinct. Technical means make it possible to distort the voice recording in a way that it is not any more possible to identify the speaker. In such a case, the recording may be considered anonymous information (GDPR, recital 26) to which the GDPR does not apply. Article 4 (1) defining personal data refers not only to "identified" but also to "identifiable" natural person. In deciding over "identifiability", account should be

[41] Recital 11 of the Digital Copyright Directive.

[42] Kelli A., Lindén K., Vider K., Kamocki P., Birštonas R., Calamai S., Labropoulou P., Gavrilidou M., Straňák P. Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (eds.), Selected papers from the CLARIN Annual Conference 2018 CLARIN Annual Conference 2018, Pisa, 8-10 October 2018, pp. 72–82. Linköping University Electronic Press, Linköpings universitet, 2019. Available at http://www.ep.liu.se/ecp/article.asp?issue=159&article=008&volume= [last viewed October 29, 2019]; Klavan J., Tavast A., Kelli A. The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. *Frontiers in Artificial Intelligence and Applications,* No. 307, 2018, pp. 71–78. Available at: http://ebooks.iospress.nl/volumearticle/50306 [last viewed October 29, 2019].

taken of all the means reasonably likely to be used, including all objective factors, such as the costs of and the amount of time required for identification, the available technology at the time of the processing and technological developments (recital 26), Thus, it may be concluded that the GDPR does not apply to the recording of the human voice if the recording has been technically processed in a way which makes the speaker's voice unidentifiable and it is not technically possible to reverse the initial voice in the recording.

Regarding PD, it is theoretically possible that small but identifiable bits of information make it to the model. A wordlist might contain a name or e-mail address, for instance. This is easy to avoid using anonymisation or pseudonymisation.

However, it should be kept in mind that for PD, there is no minimum segment in the audio synthesis. Even if the voice is synthesized using neural networks without any remnants of the person's original voice recordings, which, for instance, could be a publicly available radio transmission that has been used for the training of the neural network for research purposes, one is still using the PD of that person if that person can be identified from the synthesized output, despite the fact that there is no single bit in the network which could be attributed to the person's voice.

The main issue here is how to substantiate the processing[43] of PD contained in a model. Generally speaking, the compilation of data-sets containing PD used to create models can be based on the consent, public interest research and legitimate interest (see, GDPR Art. 6 (1) a), e), f)). In case there is consent to process data for research purposes, or processing relies on public interest and the resulting model is used for research purposes as well (i.e., it is not made available to the public or used for commercial purposes), then there is no problem. There is also no problem if consent covers commercial use and public dissemination.

However, the situation becomes complicated when a data-set containing PD is processed based on consent asked for research or on the public interest research exception, but the resulting model (where the PD may remain) is planned to be used for commercial purposes or be made publicly available.[44]

In the described case, there are the following scenarios:

1) Use some technical measure to modulate the speech signal so that it no longer resembles the original;
2) Ask for consent for commercial use.

[43] The GDPR defines processing as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (Art. 4 (2)).

[44] It is not the voice data which could affect the person negatively, but a speech synthesizer could be used to say all kinds of things which would reflect badly on the person whose voice it is, e.g. if people get the impression that the person said something reprehensible in public.

Conclusions

The authors' principal findings can be visualised with the following graph:

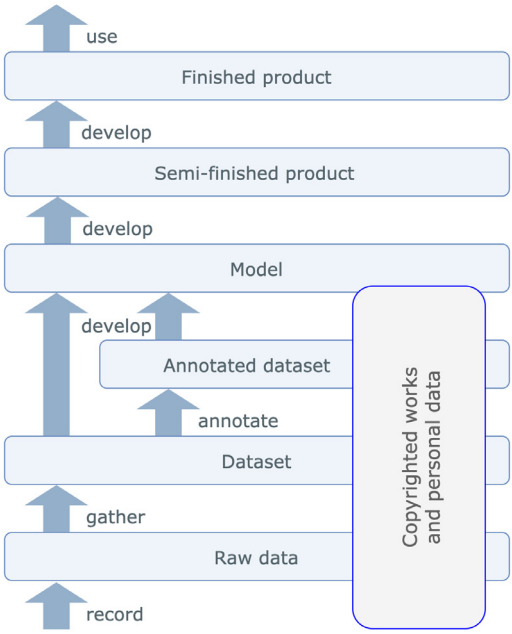

Figure 4. Process of developing language technology

Language data used in language technology may be subject to restrictions arising from copyright and PD protection. In the process of developing a language technology application, language models are the first intermediate result that can be free of such restricted data. This means both that models do not contain any original parts of the data, and that it is not possible to re-create original parts of the data from the model. A potential exception is speech data and the ability of specific models to recreate the voice of a person, in which case PD protection issues need to be addressed.

Our analysis shows that language models cannot be considered derivative works based on the underlying language data. While processing and annotating raw language data is possible only in consent-based or exception-based cases, the use of models as independent scientific results does not presuppose the existence of permission or exception. License terms of the model are at the discretion of the developer of the model, including commercial use and making available to the public.

This result contributes to clarifying the legal aspects of creating language technology applications by specifying a point in the development process where the copyright and PD restrictions of raw language data become no longer applicable.

BIBLIOGRAPHY

Literature

1. Birštonas R., Usonienė J. Derivative Works: Some Comparative Remarks from the European Copyright Law. *UWM Law Review*, Vol. 5, 2013.
2. Devlin J., Chang M.-W., Lee K., & Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. ArXiv:1810.04805 [Cs].
3. De Castilho R. E., Dore G., Margoni T., Labropoulou, P. & Gurevych I. A legal perspective on training models for Natural Language Processing. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, ELRA. Available at: http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf [last viewed October 29, 2019].
4. Grave E., Bojanowski P., Gupta P., Joulin A., & Mikolov T. Learning word vectors for 157 languages, 2018. ArXiv Preprint ArXiv:1802.06893.
5. Ilin I., Kelli A. The Use of Human Voice and Speech in Language Technologies: The EU and Russian Intellectual Property Law Perspectives. *Juridica International*, No. 28, 2019, pp. 17–27. Available at: https://www.juridicainternational.eu/public/pdf/ji_2019_28_17.pdf [last viewed October 29, 2019].
6. Kelli A., Lindén K., Vider K., Kamocki P., Birštonas R., Calamai S., Labropoulou P., Gavrilidou M., Straňák P. Processing personal data without the consent of the data subject for the development and use of language resources. In: Skadina I., Eskevich M. (eds.), Selected papers from the CLARIN Annual Conference 2018 CLARIN Annual Conference 2018, Pisa, 8-10 October 2018. Linköping University Electronic Press, Linköpings universitet, 2019, pp. 72–82. Available at: http://www.ep.liu.se/ecp/article.asp?issue=159&article=008&volume= [last viewed October 29, 2019].
7. Kelli A., Vider K., Pisuke H., Siil T. Constitutional Values as a Basis for the Limitation of Copyright within the Context of Digitalization of the Estonian Language. In: Chair: Prof. *Dr. habil. iur.* Kalvis Torgans, University of Latvia, Latvia (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 Riga, Latvia: University of Latvia Press, 2017, pp. 126–139.
8. Kelli A., Vider K., Lindén K. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. De Smedt K. (ed.), Linköping University Electronic Press, Linköpings universitet, 2015, pp. 13–24. Available at: http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 [last viewed November 2, 2019].
9. Kelli A., Tavast A., Pisuke H. Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. *Juridica International*, No. 19, pp. 40–48, 2012. Available at: https://juridicainternational.eu/public/pdf/ji_2012_1_40.pdf [last viewed November 2, 2019].
10. Klavan J., Tavast A., Kelli A. The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. *Frontiers in Artificial Intelligence and Applications*, No. 307, 2018, pp. 71–78. Available at: http://ebooks.iospress.nl/volumearticle/50306 [last viewed October 29, 2019].
11. Kondratyuk D. and Straka M. UDify Pretrained Model, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics,

Charles University, 2019. Available at: http://hdl.handle.net/11234/1-3042 [last viewed November 3, 2019].

12. Pennington J., Socher R, and Manning C. D. GloVe: Global Vectors for Word Representation, 2014. Available at: https://nlp.stanford.edu/projects/glove/ [last viewed November 3, 2019].
13. Ulčar M. ELMo embeddings model, Slovenian, Slovenian language resource repository CLARIN.SI, 2019. Available at: http://hdl.handle.net/11356/1257 [last viewed November 3, 2019].
14. Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., & Le Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding, 2019. *ArXiv:1906.08237 [Cs]*. Available at: http://arxiv.org/abs/1906.08237 [last viewed November 3, 2019].

Legislative acts

1. Copyright Act of Estonia. English translation available at: https://www.riigiteataja.ee/en/eli/504042019001/consolide [last viewed November 3, 2019].
2. Latvian Copyright Act. English translation available at: https://vvc.gov.lv/image/catalog/dokumenti/Copyright%20Law.doc [last viewed November 11, 2019].
3. 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. OJ L 130, 17.5.2019, pp. 92–125. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790 [last viewed October 29, 2019].
4. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. Official Journal L 167, 22/06/2001, pp. 0010–0019. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 [last viewed October 29, 2019].
5. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. OJ L 77, 27.3.1996, pp. 20–28. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572850210125&uri=CELEX:31996L0009 [last viewed November 3, 2019].
6. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, pp. 1–88. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 [last viewed October 29, 2019].

Legal practice

1. CJEU judgement of 29 July 2019 in case No. C- 469/17 *Funke Medien NRW GmbH vs. Bundesrepublik Deutschland*.
2. CJEU judgement of 13 November 2018 in case No. C-310/17 *Levola Hengelo BV vs. Smilde Foods BV.*
3. CJEU judgement of 1 December 2011 in case No. C-145/10 *Eva-Maria Painer vs. Standard VerlagsGmbH et al.*
4. CJEU judgement of 16 July 2009 in case No. C-5/08 *Infopaq International A/S vs. Danske Dagblades Forening*.