The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies

Aleksei Kelli
University of Tartu,
Estonia
aleksei.kelli@ut.ee

Arvi Tavast
Institute of the
Estonian Language,
Estonia
arvi@tavast.ee

Krister Lindén
University of Helsinki,
Finland
krister.linden@
helsinki.fi

Kadri Vider
University of Tartu,
Estonia
kadri.vider@ut.ee

Ramūnas Birštonas
Vilnius University,
Lithuania
ramunas.birstonas@
tf.vu.lt

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Irene Kull
University of Tartu
Estonia
irene.kull@ut.ee

Gaabriel Tavits
University of Tartu
Estonia
gaabriel.tavits@ut.ee

Age Värv
University of Tartu
Estonia
age.varv@ut.ee

Pavel Straňák
Charles University,
Czechia
stranak
@ufal.mff.cuni.cz

Jan Hajic
Charles University,
Czechia
hajic@ufal.mff.cuni.cz

Abstract

The authors address the legal issues relating to the creation and use of language models. The article begins with an explanation of the development of language technologies. The authors analyse the technological process within the framework copyright, related rights and personal data protection law. The authors also cover commercial use of language models. The authors' main argument is that legal restrictions applicable to language data containing copyrighted material and personal data usually do not apply to language models. Language models are generally not considered derivative works. Due to a wide range of language models, this position is not absolute.

Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Straňák and Jan Hajic 2020. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 53–65.

1 Introduction

The development of language technologies (LTs) relies on the use of language data (LD). Language data is often covered with several tiers of rights (copyright, related rights, personal data rights). Their use can be based on a contractual (e.g. licence, contract, terms of use/service, etc.) or exception model as regards IPR and a consent or exemption model concerning personal data.

The current paper discusses the impact of language data's legal regime on LTs. The question is whether legal restrictions applicable to language data apply to the language technologies that are developed using them as well. The authors analyse how far, in the pipeline of developing language technologies, the original copyright and personal data protection regulations apply. If we take a recorded phone call, for instance, it is evident that copyright and data 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 corpus containing the former. The line where the original rights cease to apply has to be somewhere between these points, and researchers and developers need to know where.

The authors present arguments that copyright and personal data restrictions covering language data usually do not affect language models.[1]

The article develops further the previous legal research conducted in the field of language technologies (see Eckart de Castilho et al. 2018; Kelli et al. 2016; Kamocki et al. 2019; Kelli et al. 2018a; Ilin and Kelli 2019; Klavan et al. 2018).

2 From language data to language technologies

The development of data-driven/data-based language technologies contains:

1. Collection of raw data (written texts, speech recordings, photos, videos, etc.). These often include 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.

Dangers for both copyright and personal data implications can be real: re-publication of copyrighted works, infringement of privacy by governments or insurance companies, etc.

It is almost impossible to anonymise data entirely so that it would become impossible to identify any persons.

2. Compilation of datasets, or collections of data (raw text corpora like Google News, Common Crawl[2] or Open Subtitles[3], speech corpora like the Prague Database of Spoken Czech, etc.). The above, but collected and organised with a specific criterion in mind (e.g. speech recordings of a particular topic by residents of a specific region to capture the accent of the region). These datasets 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 dataset.

For personal data purposes, data collections are not different from raw data. The main practical difference is that the sheer volume of data may make it technically difficult for an individual to become aware that their data has been included in the dataset.

At the IPR side, the original rights of the individual pieces of data remain as is when included in the dataset. For instance, the copyright of a photo or speech recording is carried over so that the copyright of the dataset consists of the copyright of the individual items. IPR of the individual items must first be cleared to attach a single licence to a dataset. Also, the creation of a dataset often involves a nontrivial contribution in gathering, organising, indexing, presenting, hosting etc. of the data. This reflects on the *sui generis* database (SGDB) right, which governs the *structure* rather than the *contents* of the dataset[4].

[1] The analysis is limited to models containing speech and text.
[2] See also http://commoncrawl.org/
[3] see also https://www.opensubtitles.org/
[4] In fact, it can be argued that data-sets qualify for database protection (for further discussion, cf. Eckart de Castilho et al. 2018; Kelli et al. 2012).

3. Creation of annotated datasets (POS-tagged corpora of written texts like the web13 (etTenTen)[5], syntactically parsed corpora like the Universal Dependencies[6] treebanks, etc.). The above, augmented with some analysis.

Again, annotated datasets are not different from raw data in terms of copyright and personal data, 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. Still, standard practice is to include 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 dataset includes analysis of the data, either manual, semi-automatic or automatic. When this analysis is performed manually, it can be argued that the copyright of the annotations belongs to the annotators (or the organisation that has commissioned the task). On the other hand, we can argue that a strictly automated annotation of a dataset does not create new rights either on the part of the person(s) that have run the annotation tool nor on the part of the person(s) that have developed the annotation tool.

4. Models. Data products developed from some processing on the above, but not necessarily containing the above, which try to model, i.e. represent or describe, language usage. Examples, in this broad sense, include dictionaries, wordlists, frequency distributions, n-gram lists like Google n-grams, pre-trained word embeddings (cf. Grave et al. 2018), pre-trained language models (cf. Devlin et al. 2018).

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 (may even include custom hardware development), hyperparameter optimisation, model validation.

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.

5. Semi-finished products (text-to-speech engine or a visual object detector) and finished products (talking fridge). Out of scope for the current analysis, because their independent status should be beyond doubt.

3 Copyright perspective on the creation of language models

From the copyright perspective, there are three relevant issues. Firstly, whether copyrighted material is used. Secondly, if it is used, whether there is a legitimate ground for this use. Thirdly, how to define models themselves within the copyright framework.

The requirements for copyright subject matter should be briefly outlined before explaining the copyright law impact on models. The primary and long-established requirement is that of originality. A 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*[7]. 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 in the *Infopaq* case (C-5/08), originality means the author's intellectual creation. Another relevant explanation in the *Infopaq* case was that an extract consisting of eleven words could constitute an original work (C-5/08 para 48). The Court has also explained that a single word cannot be regarded as an original and protectable work.

In the context of the current paper, the originality requirement is important from two different perspectives. First, if originality is missing from the dataset used for the creation of the model, the pre-text contained in a dataset is not protected and can be used without authorisation. Therefore, even if parts of this text are reproduced in the model, they are not protected either. Second, even if a text as a whole is original and, therefore, protected, the question remains, whether the fragments used in the model are

[5] http://doi.org/10.15155/1-00-0000-0000-0000-0012EL
[6] See also https://universaldependencies.org/
[7] Although it was defined in several EU directives with regard to specific categories of works, such as computer programs or photographic works.

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.

To answer the question of whether models are copyright protected, we must establish whether they meet the requirement of originality also on their own (irrespective of the input dataset).

One of the criteria that can be used for assessing originality has to do with the degree of human intellectual effort invested in the process: how far is the model a unique product, the result of the intellectual creation of the author? Building a model (as presented in Section 2) includes several choices and actions on the part of the developer: choice/creation of the dataset, choice/creation of the programme to be used for the training and development of the model and various cycles of testing and validation by tuning the parameters of the training programme.

Text can 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 (cf. Eckart de Castilho et al. 2018). 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 (CC) Attribution ShareAlike 4.0 International (e.g., Grave et al. 2018), Apache License 2.0 (Devlin et al. 2018; Yang et al. 2019) and Public Domain Dedication and License v1.0 (e.g., Pennington et al. 2014).

It is also crucial to answer the question about the copyright status of models. The problem is whether they can be considered "derivatives" or "adaptations" of the original (primary or underlying) work. There is no uniform definition of derivative work at EU and international levels. Different jurisdictions have their approaches (for further discussion, see Birštonas and Usonienė, 2013; Eckart de Castilho et al. 2018).

The Berne Convention does not name derivative works but refers to them. According to the convention "[t]ranslations, adaptations, arrangements of music and other alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work" (Art. 2 (3)). The EU case law concerning derivative works makes a reference to the concept of substantial similarities (T-19/07 para 259).

Some examples are provided below to describe national approaches. For instance, although the Estonian Copyright Act provides that derivative works are protected by copyright (§ 35 (1)), it does not define the derivative work. Instead, it says that "*translations, adaptations of original works, modifications (arrangements) and other alterations of works*" are considered derivative works (§ 4 (3) clause 21). The Estonian Copyright Act provides that the derivative work has to be "*derived from the work of another author*" (§ 35 (1)). The Act sets forth non-exhaustive examples of what constitutes the creation of derivative works: "*the transformation of a narrative work into a dramatic work or a script, the transformation of a dramatic work or a script into a narrative work, the transformation of a dramatic work into a script, and the transformation of a script into a dramatic work*" (§ 35 (2)).

The Finnish Copyright Act contains the following regulation "A person who translates or adapts a work or converts it into some other literary or artistic form shall have copyright in the work in the new form, but shall not have the right to control it in a manner which infringes the copyright in the original work" (Art. 4 (1)).

According to the Lithuanian Copyright Act, the subject matter of copyright also includes "derivative works created on the basis of other literary, scientific or artistic works (translations, dramatisations, adaptations, annotations, reviews, essays, musical arrangements, static and interactive Internet homepages, and other derivative works)" (Art. 4 (3) clause 1).

The Czech Copyright Act provides that "A work which is the outcome of the creative adaptation of another work, including its translation into another language, shall also be subject to copyright" (Art. 2 (4)).

Finally, the Greek Copyright Act under 'economic rights' (Art. 3) gives the rightholder of a work the right to authorize or prohibit what we usually term 'derivative works'. In fact, although the term 'derivative' is often used in the Greek related legal literature, it is not used in the law, which prefers to refer to "*the arrangement, adaptation or other alteration of their works*", using the same exact wording as in Article 12 of the Berne Convention. We should also mention here the clause "*the translation of their works*" which is usually deemed a derivative work. The Greek law does not further specify any criteria for assessing what an "*adaptation or alteration*" is, yet similarity to the original work and originality of the new work are typically used for this purpose (Marinos 2018).

Based on the referred copyright laws, it can be concluded that to qualify as a derivative work, it has to include substantial copyright-protected parts of the used primary work.

In addition to national laws, it is useful to look at standard licenses such as Creative Commons (CC) which are used as tools to make copyrighted content (language data) available. For instance, CC Attribution-NoDerivs (CC BY-ND) does not allow to share adapted materials (derivative works). According to CC BY-ND "*Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor*". A key issue here is whether permission is required to create a model.

The authors consider that the creation of the model is done through a data mining activity. The Digital Copyright Directive (DCD) defines text and data mining (TDM) as "*any automated analytical technique aimed at analysing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations*" (Art. 2 (2)).

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.

According to the Digital Copyright Directive research organisations and cultural heritage institutions[8] 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*" (Art. 2 (1)).

The Digital Copyright Directive Art. 3 (1) allows making copies of works[9], objects of related rights (e.g., performances), press publications[10] 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.[11] 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 shall "*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 smooth. 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 datasets exploited for TDM. The question is, what can be done with datasets. 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 datasets can be shared among researchers. This is a genuinely crucial issue since research and research infrastructures such as CLARIN are based on the ideology of sharing research data. It remains to be seen how the national

[8] 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)).

[9] The quotation right also allows to copy parts of a work. However, according to the EU case law "the user of a protected work wishing to rely on the quotation exception must therefore have the intention of entering into 'dialogue' with that work" (C-476/17 para. 71). Since the development of language technology relies on works as language data, then there is no 'dialogue' and the quotation right is not applicable.

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

[11] 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".

legislators implement the provision. The research community should use all possible measures to introduce a regulation which allows at least limited sharing.

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 (Recital 11 of DCD).

To say whether models constitute derivative works, we should classify and analyse all possible model types, the processes and resource types and modalities they have been built upon. It is not feasible 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 lists of words or phrases and patterns with statistical measures. Therefore, they cannot be usually considered derivative works. This conclusion is supported by other researchers as well (see Eckart de Castilho et al. 2018).

4 Database right perspective

Another set of rights which could encumber the initial material and (possibly) the models, is the *sui generis* database makers right (database right). This type of protection was introduced by the Database Directive of 1996 (DD) and mostly remains a peculiarity of European countries. The database right is different from copyright in several important respects. First of all, the subject matter of protection is only a database. A database is defined 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*" (DD Art. 1 (2)). Although this could seem like a technical definition with quite a narrow field of application, the said definition is very wide and encompasses such different subject matter as directories, dictionaries, newspapers, journals, the content of a website and so forth. It is obvious that raw data for the models often, if not most, come from databases.

If the subject matter is qualified as a database, then, according to the Database Directive, it can be protected by copyright or by the database right. Therefore, there are four possibilities: a database can be protected by copyright, by database right, by both of them or by none of them. Since copyright protection was discussed above, we do not address copyright-protected databases further. In all that was said before, copyright equally applies to a copyright-protected database.

The database right should be differentiated from copyright. The requirement for the database protection is not based on originality, but instead, the following three cumulative conditions should be met: 1) there should be an investment, 2) an investment should be "*qualitatively and/or quantitatively substantial"*, 3) an investment should be in the obtaining, verification or presentation of the contents (DD Art. 7 (1)).

The existing court practice (both from ECJ and national courts) shows that these formal requirements can be met quite easily. The concept of investment is interpreted broadly (it includes financial resources, time, energy, labour, time, and so forth) and the threshold of "*qualitatively and/or quantitatively substantial*" is not very high. As a result, a considerable number of datasets, used for language technology purposes, are covered by the database right. In such a case, the maker of a database has the right to prevent extraction and/or re-utilization of the whole or a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database (DD Art. 7 (1)). An act of extraction is defined as "*the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form*", whereas 're-utilization' means "*any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission*" (DD Art. 7 (2)). It should be stressed, that contrary to copyright protection, the database right protects not the original selection or arrangement, but the content of a database itself (or, more precisely, the whole or a substantial part, evaluated qualitatively and/or quantitatively). According to the case law "*[t]he purpose of the protection by the sui generis right provided for by the directive is to promote the establishment of storage and processing systems for existing information and not the creation of materials capable of being collected subsequently in a database*" (C-203/02 para. 31).

Translating these legal rules to the field of language technologies, it could be noted that the database right could have a direct impact on language technology and its results. Many sources, from which data

is taken for the compilation of datasets, are protected by the database right. Further, the same two questions, which are relevant in the context of copyright, arise: first, if there is a legitimate ground for the use of a database, and, second, how a model can be affected by the database right.

Answering the first question, there can be no doubt that normally a collection of raw data from a database (see chapter 2) constitutes an act of extraction. While the Database Directive restricts the scope of the said right to the instances, when the whole or substantial part of the content of a database is copied (Art. 7 (1)), this is exactly the case in a typical raw data collection process. The problem is solved by the data mining exception introduced by the Copyright Directive (Art. 3 and 4), which expressly provides exceptions not only to copyright but also to the database right. Therefore, while normally, the collection of data falls into the sphere of the extraction right, it can still be legitimate if the requirements of the exception of data mining are met. As was said, the collection of data and development of the model is considered data mining. Therefore, additional permission from the database maker is not necessary.

The remaining question concerns models and their status from the database right perspective. As it was seen, models rely on datasets. Arguably, in some cases the interference with the re-utilization right is possible. As was explained, re-utilization means any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission. It should be borne in mind, that ECJ confirmed several times that re-utilization should be interpreted is a wide manner (see C-203/02 para. 51; C-173/11 para 20). In one of its latest decisions, ECJ has reiterated that 're-utilization' refers to *any* unauthorised act of distribution to the public of the contents of a protected database or a substantial part of such contents, while the nature and form of the process used are of no relevance in this respect (C-202/12 para 37). It can theoretically be argued that in a model, a certain amount of data (e.g., discrete words) is transmitted, that is re-utilized. Still, the right of re-utilization is infringed only if the whole or a substantial part, evaluated qualitatively and/or quantitatively, of a database is used. Lawful users of the database have a specific right to re-utilize insubstantial parts of its contents for any purposes whatsoever.

In a model, the whole or a substantial part, evaluated quantitatively, of the content of a database is rarely re-created, but the same cannot be said for a substantial part, evaluated qualitatively. The substantial part evaluated qualitatively is a nebulous concept, which, according to the ECJ, refers to the scale of the investment in the obtaining, verification or presentation of the contents of the subject of the act of re-utilization, regardless of whether that subject represents a quantitatively substantial part of the general content of the protected database (C-203/02 para. 71). Also, a quantitatively negligible part of the content of a database may represent, in terms of obtaining, verification or presentation, a significant human, technical or financial investment (C-203/02 para. 71). So, in principle, even small excerpts of the original data can represent a qualitatively substantial part of the content of the protected database. Furthermore, the repeated and systematic re-utilization of insubstantial parts of the content of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database are not permitted (DD Art. 7(5)). Again, this can potentially pose a problem for language technologies, because it can be argued that in models some insubstantial parts of the database are re-utilized repeatedly and systematically.

In conclusion, the database right causes certain uncertainties and risks when it comes to the creation of models.

5 Personal data protection perspective on the creation of language models used for commercial purposes

Since several issues relating to personal data in language research have been previously addressed (see Kelli et al. 2019; Lindén et al. 2019; Kelli et al. 2018b) then only personal data aspects relevant for models are addressed in this article.

The General Data Protection Regulation (GDPR) defines personal data as "*any information relating to an identified or identifiable natural person*" (Art. 4 (1)).

It is possible that models contain personal data such as a name or an e-mail address. As a result, GDPR becomes applicable. To avoid legal restrictions stemming from the General Data Protection Regulation, it is advisable to anonymise data (for further discussion, see WP29 2014). The GDPR does not apply to anonymous information (Recital 26 of the GDPR).

However, it should be kept in mind that for personal data, there is no minimum segment in the audio synthesis. Even if the voice is synthesised using neural networks without any remnants of the person's original voice recording, having trained the network for research purposes using a publicly available radio transmission as training data, one is still using the personal data of that person when the person can be identified based on the synthesised output although 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[12] of personal data contained in a model. Generally speaking, the compilation of datasets containing personal data 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, 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 dataset containing personal data is processed based on consent asked for research or on the public interest research exception, but the resulting model (where the personal data remains) is planned to be used for commercial purposes or be made publicly available. If the personal data is in the form of speech, then anonymization is rather difficult. In the described case, there are the following scenarios:

1. Argue that voice without any identifying information is not personal data (it is anonymous data). The key here is how to interpret the concept of an identifiable natural person (see Art. 4 (1); WP29 2007);
2. Ask for consent for commercial use (see WP29 2018);
3. Argue that the use of voice in the model is based on the legitimate interest (for further discussion, see WP29 2014a). Especially bearing in mind that the identification is impossible or almost impossible and the voice does not contain any data which would affect the data subject negatively;
4. Technically modulate the voice data so that it no longer resembles the original speaker without destroying the properties of the speech signal essential for the intended application.

The first and third options are somewhat uncertain and pose legal risks.

6 The commercialisation of language models

6.1 The general framework for commercialisation

The goal of commercialisation is to derive profit from adding value to some raw material. The raw material can be any input that is further processed before it is commercialised. A minimal act of further processing is copying. In this work, we explore the conditions for commercialising a language model. In the following, we will compare some practices in different countries for further processing language data to produce language models for commercialisation.

6.2 National approaches to commercial exploitation of language models

Czechia. At LINDAT/CLARIAH-CZ[13], hosted at Charles University, Prague, the choice made was rather cautious. If a source of data contains a non-commercial clause (such as –NC in CC licences), it is considered binding for any derivative work and also for models which are not deemed copyrightable. This is equally applied to data and models Charles University produces itself. That allows us to keep releasing our annotated corpora under CC licenses and yet try to protect their commercial use. That is

[12] 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)).

[13] https://lindat.cz

why we release all UD models (parsing and morphology from >100 treebanks) under -NC, regardless of the original CC clauses attached to the dataset CC licence, and at the same time we cooperate with a commercial NLP company on licensing for the commercial sector under the following procedure:

(i) We say that for the models, we charge only for the convenience of the user, so that s/he does not have to go through the training process, usually just some small money (any commercial user can train the model themselves if they possess the expertise). However, this makes commercial use legal only for models from corpora that are free of further restrictions (e.g. CC-BY and similar).

(ii) For legal use of the models trained from datasets with restrictions (CC-BY-NC(-SA), but also one GPL, etc., see https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.5#) the commercial NLP company contacts the dataset creators one by one and tries to make a deal for the commercial use of the dataset. Based on these deals they offer to their commercial client to clear the models for these datasets by getting them commercial licenses for the underlying datasets. That gets the clients into the situation when they have legal access to the datasets for commercial use, and they have paid us the fee for the convenience of training the models for them.

LINDAT/CLARIAH-CZ itself is one of those dataset creators, so if the commercial use case involves Czech, we get something for training the models and something for the commercial license of the annotated dataset(s).

Estonia. Ideally, all the groundwork from raw data collection to annotated datasets as described in section 2 (steps 1-3) would be done by researchers who have the benefits of using data for research purposes. The result of the first steps would remain a derivative or original work (including *sui generis* databases), strictly managed by the research organisation, subject to copyright and subject to conditions of use determined by the research organisation in accordance with the input data rights.

The model building phase (step 4) could also involve collaboration between researchers and companies, where companies outsource the expertise of researchers in data processing and researchers process or determine, the appropriate parameters in the (automatic) model creation process that meet the business model development goals.

Cases where researchers work (in part, also in their start-up) in a business enterprise or a start-up that has grown out of a research institution are more complicated.

The commercial use of language resources, including the creation of models, should, in any case, remain a tailor-made activity and the value-added for business purposes should be left to the researchers. In doing so, researchers can ensure that the model does not include raw data that conflicts with the conditions of use of the first steps (1-3).

For example, there are three different morphosyntactic analysis models available for the development of language technology in Estonia that can be used (and have been used) under the CC-BY-SA license. Researchers of the University of Tartu create these models by training on a variety of text materials with different legal regimes, but in the resulting model, none of the text used as reference data is in a recognisable form. The same is true with the statistical machine translation model.

Finland. As soon as an organisation has lawful access to a dataset, it can create a model based on the dataset. Provided that the model does not contain substantial reproducible parts of the original copyright-protected data, the model is an independent work which can be given whatever license its creator chooses. However, if a dataset has additional restrictions, e.g. non-commercial use, this means that the organisation cannot engage in such activities for producing models with the dataset unless the restriction is lifted. This is similar to the situation described for Czechia.

Greece. To the best of our knowledge, there are very few models for modern Greek publicly available[14], and these are distributed under a CC-BY-NC-SA licence, most probably because they have been trained on the relevant UD treebank licensed with the same terms. It also seems that there are not yet any discussions on the commercial use of models. As regards language resources of modern Greek, the interest lies mainly on annotated datasets and relevant tools and services.

[14] https://ufal.mff.cuni.cz/udpipe/models and https://spacy.io/models/el

Lithuania. There is no reliable data concerning the commercial exploitation of language models in Lithuania. Models are mainly prepared by educational and scientific institutions, and they tend to make models either publicly open[15], or make them available on request. It is important to note that models are shared only for non-commercial purposes. Requests for commercial use are declined. If models are shared, the specific public licence is applied. It has to be noted, that language resources used in the framework of the CLARIN-LT, are not licensed by one of the Creative Commons licences, but the specific public licence was prepared by Lithuanian team of lawyers.

7 Conclusion

The creation of language models relies on the use of language data. Language data could contain copyright-protected works, objects of related rights (performances, recordings, databases) and personal data. Therefore, its use is restricted by copyright and personal data protection laws. The process of creating a model can involve text and data mining (TDM). From the copyright perspective, TDM in itself is not an activity requiring a legal basis (consent or exception). However, to conduct TDM, there is a need to copy language data (copyright-protected works, objects of related rights) which must have a legal basis. The Digital Copyright Directive introduced a mandatory exception for TDM, which allows making reproductions for TDM.

The creation of a language model involves several complex human intellectual activities, such as choosing and annotating datasets as well as choosing the software and tweaking its parameters. The outcome of the preparatory software activities is applied to a prepared dataset to compile a language model.

If the created language model contains copyrighted language data used to develop the model, then the model is subject to the same copyright restrictions as the data. However, this depends on the model type. In most cases, models do not contain copyright-protected content.

From the perspective of database right, the use of models in certain cases theoretically could infringe re-utilization right, although the legal practice has not settled this issue yet.

There is also the question of whether they contain enough material to breach personal data regulations. For instance, models containing speech need to address personal data issues.

The authors' main conclusion is that language models usually do not have the same legal restrictions as language data used to create them.

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

[15] Some examples can be found here: http://mwe.lt/, https://www.semantika.lt/, etc.

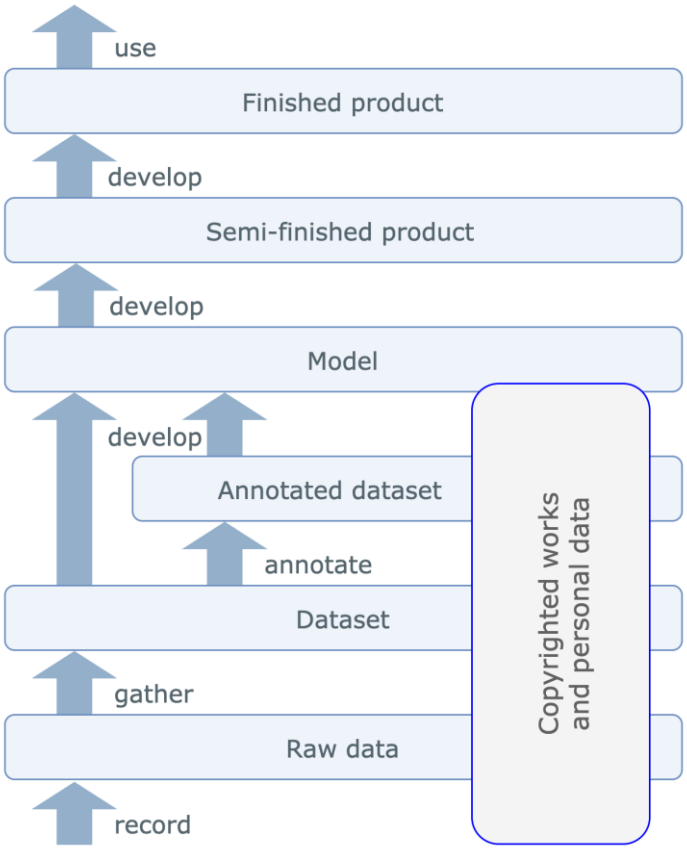

Figure 1: Process of developing language technology

Acknowledgements

Research is supported by the ERDF project "Federated Content Search for the Center of Estonian Language Resources (2014-2020.4.01.16-0134) under the activity "Support for Research Infrastructures of National Importance, Roadmap" and by the Research and Development Program "Estonian Language Technology 2018-2027" of the Ministry of Education and Research.

References

[Birštonas and Usonienė 2013] Ramunas Birstonas, Jurate Usoniene. 2013. Derivative Works: Some Comparative Remarks from the European Copyright Law. *UWM Law Review*, Volume 5.

[Berne Convention] Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979). Available at https://wipolex.wipo.int/en/text/283698 (2.2.2020).

[CC BY-ND] Creative Commons Attribution-NoDerivatives 4.0 International Public License. Available at https://creativecommons.org/licenses/by-nd/4.0/legalcode (2.2.2020).

[Czech Copyright Act] Copyright Act (121/2000). Available at https://www.wipo.int/edocs/lexdocs/laws/en/cz/cz043en.pdf (2.2.2020).

[C-476/17] C-476/17. Pelham GmbH, Moses Pelham, Martin Haas *vs* Ralf Hütter, Florian Schneider-Esleben (29 July 2019). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1576587562212&uri=CELEX%3A62017CJ0476 (2.2.2020).

[C-5/08] Case C-5/08. Infopaq International A/S *vs* Danske Dagblades Forening (16 July 2009). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005 (14.4.2019).

[C-203/02] Case C-203/02. The British Horseracing Board *vs* William Hill Organization (9 November 2004). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580123225598&uri=CELEX:62002CJ0203 (27.1.2020).

[C-173/11] Case C-173/11. Football Dataco *vs* Sportradar (18 October 2012). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580123660404&uri=CELEX:62011CJ0173 (27.1.2020).

[C-202/12] Case C-202/12. Innoweb *vs* Wegener ICT Media (19 December 2013). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580124636913&uri=CELEX:62012CJ0202 (27.1.2020).

[Database Directive = DD] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. OJ L 77/20, 27.3.1996, pp. 20–28. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580123910536&uri=CELEX:31996L0009 (27.1.2020).

[Digital Copyright Directive = DCD] 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 (26.1.2020).

[Devlin et al. 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs].

[Eckart de Castilho et al. 2018] Eckart de Castilho, R., Dore, G., Margoni, T., Labropoulou, P. & Gurevych, I. 2018. 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 (17.4.2019).

[Estonian Copyright Act] Copyright Act (12.12.1992). Available at https://www.riigiteataja.ee/en/eli/504042019001/consolide (2.2.2020).

[Finnish Copyright Act] Copyright Act (404/1961). Available at https://www.finlex.fi/en/laki/kaannokset/1961/en19610404 (2.2.2020).

[Grave et al. 2018] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, Tomas Mikolov. 2018. Learning word vectors for 157 languages. ArXiv Preprint ArXiv:1802.06893.

[GDPR] 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, p. 1-88. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (2.2.2020).

[Greek Copyright Law] Greek Copyright, Related Rights and Cultural Matters (Law 2121/1993 amended by Law 4531/2018). Available at https://www.opi.gr/en/library/law-2121-1993 (5.2.2020).

[Ilin and Kelli 2019] The Use of Human Voice and Speech in Language Technologies: The EU and Russian Intellectual Property Law Perspectives. Juridica International 28, 17-27. Available at https://www.juridicainternational.eu/public/pdf/ji_2019_1_17.pdf (2.2.2020).

[InfoSoc Directive] 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 P. 0010 – 0019. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 (14.4.2019).

[Kamocki et al. 2019] Pawel Kamocki, Erik Ketzan, Julia Wildgans, Andreas Witt. 2019. New exceptions for Text and Data Mining and their possible impact on the CLARIN. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 66-71. Available at http://www.ep.liu.se/ecp/article.asp?issue=159&article=007&volume= (2.2.2020).

[Kelli et al. 2019] Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labrpolou, Maria Gavrilidou, Pavel Straňák. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 72-82. Available at http://www.ep.liu.se/ecp/159/008/ecp18159008.pdf (29.1.2020).

[Kelli et al. 2018a] Aleksei Kelli, Tõnis Mets, Lars Jonsson, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Age Värv. 2018. Challenges of Transformation of Research Data into Open Data: the Perspective of Social Sciences

and Humanities. International Journal of Technology Management & Sustainable Development, 17 (3), 227-251.

[Kelli et al. 2018b] Aleksei Kelli, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki, Pavel Straňák. 2018. Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes? In: Maciej Piasecki (Ed.). Selected papers from the CLARIN Annual Conference 2017. Linköping University Electronic Press, 102-111. Available at http://www.ep.liu.se/ecp/147/009/ecp17147009.pdf (29.1.2020).

[Kelli et al. 2016] Aleksei Kelli, Kadri Vider, Krister Lindén. 2016. 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. Linköping University Electronic Press, 13-24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (29.1.2020).

[Kelli et al. 2012] Aleksei Kelli, Arvi Tavast, Heiki Pisuke. 2012. Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. Juridica International, XIX, 40-48. Available at https://www.juridicainternational.eu/public/pdf/ji_2012_1_40.pdf (3.2.2020).

[Klavan et al. 2018] Jane Klavan, Arvi Tavast, Aleksei Kelli. 2018. The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. Frontiers in Artificial Intelligence and Applications, 307, 71-78. Available at http://ebooks.iospress.nl/volumearticle/50306 (29.1.2020).

[Lindén et al. 2019] Krister Lindén, Aleksei Kelli, Alexandros Nousias. 2019. To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest. Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019. Ed. K. Simov and M. Eskevich. CLARIN, 56-60. Available at https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf (3.2.2020).

[Lithuanian Copyright Act] Law on copyright and related rights (18.5.1999). Available at https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/5f13b560b2b511e59010bea026bdb259 (2.2.2020).

[Marinos 2018] Michael-Theodore Marinos. 2018. The infringement of a copyrighted work with its "al-teration" – the delimitation between free and prohibited usage. Greek Law 1/2018 (59), p. 1-10. (in Greek). Available at https://mklpartners.gr/wp-content/uploads/meletes/metatropi_ergou.pdf (5.2.2020).

[Pennington et al. 2014] Jeffrey Pennington, Richard Socher, Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. Available: https://nlp.stanford.edu/projects/glove/ (3.11.2019).

[T-19/07] Case T-19/07. Systran SA, Systran Luxembourg SA *vs* European Commission (16 December 2010). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580669711496&uri=CELEX:62007TJ0019 (2.2.2020).

[Yang et al. 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019. XLNet: Gen-eralized Autoregressive Pretraining for Language Understanding. ArXiv:1906.08237 [Cs]. Avaialble at: http://arxiv.org/abs/1906.08237 (3.11.2019).

[WP29 2018] Article 29 Working Party (WP29). Guidelines on consent under Regulation 2016/679. Adopted on 28 November 2017. As last Revised and Adopted on 10 April 2018. Available at https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (3.2.2020).

[WP29 2014] Article 29 Working Party (WP29). Opinion 05/2014 on Anonymisation Techniques. Available at https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf (3.2.2020).

[WP29 2014a] Article 29 Working Party (WP29). Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (3.2.2020).

[WP29 2007] Article 29 Working Party (WP29). Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (2.2.2020).