When Size Matters. Legal Perspective(s) on N-grams

Paweł Kamocki
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
kamocki@ids-mannheim.de

Abstract

N-grams are of utmost importance for modern linguistics and language technology. The legal status of n-grams, however, raises many practical questions. Traditionally, text snippets are considered copyrightable if they meet the originality criterion, but no clear indicators as to the minimum length of original snippets exist; moreover, the solutions adopted in some EU Member States (the paper cites German and French law as examples) are considerably different. Furthermore, recent developments in EU law (the CJEU's *Pelham* decision and the new right of press publishers) also provide interesting arguments in this debate. The paper presents the existing approaches to the legal protection of n-grams and tries to formulate some clear guidelines as to the length of n-grams that can be freely used and shared.

1 Introduction

N-grams are generally defined as sequences of n items from a sample of text. In the field of linguistics, these items can be letters, phonemes or syllables, but perhaps most importantly: words. In this paper, we will discuss n-grams in a narrowed-down sense, i.e. as sequences of n words.

The use of n-grams in language research can be traced back to Shannon (1948), or even further back to Markov (1913). Today, n-grams are fundamental for computational linguistics and language technology, and lists of n-grams are a valuable resource used especially, but not exclusively, in developing language models for Machine Translation purposes. The importance of n-grams is best illustrated by the popularity of Google N-gram Viewer (https://books.google.com/ngrams), launched in 2009, and by the fact that its use by researchers has become commonplace, despite its questionable quality and lack of metadata (Koplenig, 2017).

Many linguists attempt to compile their own re-usable lists of n-grams. In doing so, they are confronted with the question of legality. Well aware of the fact that copying and sharing of text in principle requires permission from the copyright holder, they are wondering if, and to what extent, this also applies to very short excerpts of text. Indeed, the question about the number of words below which an excerpt becomes copyright-free is among the questions that legal experts are asked the most by language researchers.

The only possible *in abstracto* answer to this question is unfortunately disappointing: while in general very short n-grams can be used and shared without consequences (at least from the copyright perspective), there are no clear rules as to the length of copyright-free n-grams. The decision on where to draw the line should be made for every project on a case-by-case basis, and depend on several parameters such as the language of source texts, their genre, the primary purpose for which the list n-grams will be re-used and - last but not least - the 'risk appetite' of the research team. This paper will hopefully provide some guidance on this issue.

Before our analysis can start, it should be noted that it only concerns situations where the use of source material is not based on a licence[1] (the licence may, e.g., allow sharing of long excerpts, or

[1]For cases where the source material is, like this paper, available under a Creative Commons license, see Eckart de Castilho et al. (2018) for guidance.

Paweł Kamocki 2021. When Size Matters. Legal Perspective(s) on N-grams. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 122–128.

prohibit any sharing whatsoever) — in such cases, the license should generally prevail.[2] Our analysis is also context-independent — in situations covered by statutory exceptions, such as citation or research exception, the rules will to differ. Without entering into the intricacies of statutory exceptions in the many national laws of CLARIN members, it can generally be said that the citation exception allows even for long passages to be quoted, as long as the citation is included in an independent work (e.g. a research paper) and justified by its content. For example, it is justified to quote a set of sentences from a corpus to illustrate or disprove a hypothesis. When it comes to the research exception, it may allow compilation of lists of long n-grams, but these lists can subsequently only be shared within a research team, and they cannot be re-used for commercial purposes.

The guidelines in this paper are for all those who, rather than relying on a context-specific statutory exception, want to compile a list of n-grams that can be shared as an independent resource (not as part of a paper or an annex to a book) and in Open Access conditions, i.e. re-usable by anyone and for any purpose (including for commercial purposes).

2 The Traditional Approach, or 'Originality, You Fool!'

In copyright law, the traditional approach is that parts of works (including literary works) are protected as long as they are themselves original. In this approach, a snippet is regarded as a work in its own right: if it is original, then its reproduction and communication to the public require authorisation of the rightholder, unless they are allowed by a statutory exception.

This approach was adopted *inter alia* by the Court of Justice of the European Union in the 2009 *Infopaq* case. In this seminal decision, the Court ruled: *'it should be borne in mind that there is nothing in [EU law] indicating that [parts of works] are to be treated any differently from the work as a whole. It follows that they are protected by copyright since, as such, they share the originality of the whole work. (...) the various parts of a work thus enjoy protection under [copyright], provided that they contain elements which are the expression of the intellectual creation of the author of the work [i.e.: which are original (see below) — PK]'.*[3]

Further analysis of the question requires a brief presentation of the notion of originality. Originality (sometimes also called 'individuality') is a fundamental concept in copyright law — only original works can be protected by copyright. However, since the scope of copyright is extremely broad, encompassing all sorts of human creations from an opera to a piece of software to a cartoon drawing, the notion of originality is necessarily very vague, and its application to a specific case often uncertain. It also used to differ considerably between jurisdictions and time periods: 'original' was interpreted as 'originating from the author', i.e. not copied (in the US), but also as showing a degree of 'labour, skill and judgement' (in England), or even 'an imprint of the personality of the author' (in France).

The above-mentioned judgement of the Court of Justice of the European Union in the *Infopaq* case was also the Court's first attempt to harmonise the notion of originality across EU Member States (Rosati, 2013). The Court ruled that a work is original if it constitutes 'the author's own intellectual creation': this definition existed in some EU Directives,[4] but incidentally it is also the traditional definition of originality (*Individualität*) in German copyright law (*persönliche geistige Schöpfung*).[5] The Court further elaborated on this definition in the *Painer* case,[6] where it ruled that a work is an intellectual creation of the author if it reflects his personality and expresses his free and creative choices.

The possibility to exercise choice in the creative process is therefore a necessary (if not sufficient) condition of originality. One could then hypothesise that the choice of a single word can be sufficient to meet this requirement. The 2008 French Court of Cassation decision[7] declaring a work consisting of a single word (*Paradis*) original could be quoted to support this statement. However, it should not be

---

[2]The question remains whether violation of a license that would prohibit sharing of short excerpts would constitute copyright infringement, or just breach of contract (which are quite different from legal perspective, with different consequences). The answer may vary between jurisdictions, however see: CJEU, judgement of 18 December 2019, case C-666/18 (IT Development vs. Free Mobile) stating that the breach of an IP clause of a software licence constitutes copyright infringement.

[3]CJEU, 16 July 2009, C-5/08 (*Infopaq*)

[4]Art. 6 of the Term Directive (2006/116/EC); Art. 1.3 of the Software Directive (2009/24/EC); Art. 1.3 of the Database Directive (96/9/EC)

[5]Section 2(2) of the German Copyright Act.

[6]CJEU, 1 December 2011, C-145/10 (*Painer*)

[7]Cour de cassation, 1 Civ., 13 novembre 2008, no. 06-19.021

forgotten that the word was not only written in a very specific font, but also - more importantly - placed in a very specific context: above the toilet of a mental hospital. In other words, the protection was not granted to the word *'Paradis'* as a literary work, but rather to the whole setting, which constituted an artistic (and not literary) work.

In 2009, once again in the *Infopaq* ruling, the Court of Justice of the European Union seems to have definitively denied copyright protection of single words, stating that: *'considered in isolation, [words] are not as such an intellectual creation of the author who employs them. It is only through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation'.*[8]

In this part of the ruling the Court referred to another approach to originality, which states that originality manifests itself in 'selection and arrangement' of various elements constituting a work. This definition formally concerns only compilations, but it can be applied more generally (after all, literary works for linguists are but compilations of words). Interestingly, this approach stems from a canonical US copyright case *Sarony* concerning a lithography of Oscar Wilde;[9] today, it appears not only in the Berne Convention (Article 2.5), but also, as an alternative (selection OR arrangement), in the TRIPS Agreement (Article 10.2), the EU Database Directive (Article 3.1) and numerous national laws (e.g. Article L112-3 of the French Intellectual Property Code or Section 4 of the German Copyright Act). In its original form (as a conjunction), known from the Berne Convention or from the *Sarony* case, this criterion would imply that mere choice (selection) is not enough to constitute originality, and that another aspect - arrangement - is also necessary. In other words, the constitutive elements of a work not only have to be chosen by the author, but also placed in a particular order, which in the context of snippets would imply that two words are the absolute minimum for a snippet to be original.

It is indeed our opinion that in 'pure' international copyright (i.e., without taking into account national laws and case law) two words can be enough - still, only in extremely limited cases - to constitute an original work, or an original snippet. The position according to which two-word snippets can (in very rare cases) be original is also supported by Article 2.1 of the Berne Convention, according to which copyright protection is independent from 'the mode or form' in which the work is expressed, which arguably includes also very short forms. This same rule can also be found in Article L112-1 of the French Intellectual Property Code.

An example of a potentially original two-word sequence is *'krwista krówka'* (*'bloody (rare) little cow'* in Polish). There are several reasons to consider this sequence original: first of all, both words bear some phonetic similarities (the /'kr-/ at the beginning, the /v/ in the middle, and the /a/ at the end). Secondly, the use of the adjective *'krwisty'* (*bloody* or *rare*), usually associated with a cooked steak, to describe a little cow can be perceived as shocking or humorous, and is highly unusual. On top of this, *'krówka'* is also the name of a traditional Polish fudge candy, which makes the association with the adjective *'krwisty'* even more unusual. Moreover, this two-word sequence has 0 occurrences in the National Corpus of Polish (`http://www.nkjp.uni.lodz.pl`), which makes it highly probable that it was indeed created by its author, and not merely copied from another source.

Obviously, only a court can authoritatively decide whether *'krwista krówka'* is indeed original or not, and admittedly a lot would depend on the talents of the attorneys representing both parties in the procedure. Nevertheless, it is the author's opinion that it could potentially qualify for protection. This leads to the conclusion that original 2-grams can exist, albeit they are extremely rare.

This, however, does not seem to be the position of French judges who relatively often find two-words combinations to be original. For example: *du rififi*[10], *Charlie Hebdo*[11], *Bourreau d'enfants*[12] or

[8]CJEU, 16 July 2009, C-5/08 (*Infopaq*), no. 45.

[9]U.S. Supreme Court, 17 March 1884, *Burrow-Giles Lithographic Company v. Sarony*, 111 U.S. 53; in short, the Court ruled that although the lithographer did not create Oscar Wild, his attire and the scenography of the famous lithograph, he did select these elements and arranged them in a particular way, thereby creating an original work.

[10]Cour d'appel de Paris, 4e ch., 24 janvier 1970, RTD com. 1971, p. 94, obs. H. Desbois

[11]Cour d'appel de Paris, 4e ch., 25 octobre 1995, JurisData n° 1995-024506

[12]Tribunal de Grande Instance de Seine, 3e ch., 2 février 1960, RTD com. 1960, p. 844, obs. H. Desbois

Paris Canaille[13] were declared protected by copyright as original titles.[14] These decisions, however, are rather old. More recently, French courts denied copyright protection to such slogans as *Le marketing du désir* (desire marketing),[15] or *Le permis libre* (Free/open permit),[16] but admitted copyright protection of *À fond la forme*[17] (roughly translatable as 'Full fitness', a play on words *fond* (content) and *forme* (form, physical shape)) or *Un nom pour un oui* (a name for a 'yes').[18]

German courts seem to be more demanding when it comes to originality of very short literary works. The shortest work declared copyright-protected by a German court that we have been able to identify is a four-word slogan *Ein Himmelbett im Handgepäck*[19] (a canopy bed in hand luggage); however, the decision dates back to 1964. Modern German case law seems to generally refuse copyright protection of slogans and titles, which are deemed too short to constitute original works.[20] This is also the position in the United States, where the Copyright Office states that *'short phrases, such as names, titles, and slogans, are uncopyrightable because they contain an insufficient amount of authorship'*.[21]

So when does a literary work become 'long enough' to be considered for copyright protection according to modern standards? As stated above, there is no definitive answer, but some guidance has been provided by the CJEU in the above-mentioned *Infopaq* case. The Court ruled that snippets of 11 consecutive words can be original (although the evaluation of their actual originality was, of course, left to the national court).[22] The 11-word *limes* resulted simply from the facts of the case (this was the length of snippets used by *Infopaq*, an early news aggregator service), and the decision should not be interpreted as meaning that 10-word or shorter snippets are free from copyright; it does, however, provide an argument in the discussion. The 10-word limit for snippets should, in our opinion, be considered as very liberal. The truth lies therefore somewhere between 2 and 10.

3 A Few Words About Trademarks, or 'Some Rights Reserved'

A careful reader might be wondering — if an 1-gram turns out to be a registered trademark (like *Mercedes* or *CLARINS*), can it still be lawfully included in a list of n-grams and shared with the community?

This time the answer is a clear yes. It is so because the exclusive rights conferred by trademark law are in fact much more limited than those conferred by copyright. The copyright holder can in principle prevent others from accomplishing any acts of reproduction (copying) and communication to the public (sharing) of his work (unless they are expressly authorised by a statutory exception); the trademark holder, however, can only prevent others from using his trademark in the course of trade for the purposes of distinguishing goods or services.[23] The inclusion of an n-gram in a list, even if the list is then used for commercial purposes (e.g., developing a language model for a commercial Machine Translation service) is irrelevant from the point of view of trademark law.

In conclusion, there is no need to remove trademarks, or what appears to be trademarks, from lists of n-grams.

---

[13]Cour d'appel de Paris, 1er ch., 30 mai 1956, *Léo Ferré c/ Sté Océan Films et a.*, JCP G 1956, II, 9354

[14]Article L112-4 of the French Intellectual Property Code states that *'Le titre d'une oeuvre de l'esprit, dès lors qu'il présente un caractère original, est protégé comme l'oeuvre elle-même'*; however, it can be argued that titles are protected by *sui generis* copyright, which only restricts the use of an original title as a title of another work.

[15]Cour d'Appel de Paris, 7 novembre 2017

[16]TGI de Paris, 7 juillet 2016

[17]TGI Paris, 8 janvier 2002.

[18]CA Paris, Ch. 2, 17 juin 2011, RG n°10/12092.

[19]Oberlandesgericht Düsseldorf, 28 Februar 1964 – 2 U 76/63.

[20]E.g. both *DEA – hier tanken Sie auf* (OLG Hamburg, Urteil vom 09.11.2000, Az. 3 U 79/99) and *Für das aufregendste Ereignis des Jahres* (OLG Frankfurt, Beschluss vom 04.08.1986, Az. 6 W 134/8) were denied copyright protection.

[21]US Copyright Office, Circular 33: Works Not Protected by Copyright.

[22]CJEU, 16 July 2009, C-5/08 (*Infopaq*), no. 48.

[23]Recital 18 of the Trademark Directive of 16 December 2015 (2015/2436); for more details, see also Article 10 of the same Directive.

4 The New Approach, or 'Name that Tune'

A new approach to 'reproduction in part' was adopted by the CJEU in a recent case *Pelham*.[24] The facts involved not a literary work, but a sound recording (phonogram) by Kraftwerk, a short part of which ('approximately two seconds rhythm sequence') was used as a sample in another recording. The relevant aspect of the case was decided not on the grounds of copyright, but on the grounds of a related right of phonogram producers, a legal framework that is independent from originality.

The CJEU ruled that the use of a short excerpt of a sound recording constitutes 'reproduction in part' (and therefore an act that in principle requires authorisation of the rightholder) if the excerpt is 'recognisable to the ear'.

Can this approach be applied to text snippets? Arguably, it would mean that n-grams can be freely reused as long as they are not *hapax legomena* (i.e., as long as they occurred independently in more than one text in the language, to the extent that this can be established), and therefore their exact source cannot be identified with certainty. This approach can be viewed as stricter than the one based on originality — a purely descriptive, banal paragraph (e.g., a relation from a football match) will at some point become a *hapax legomenon* if it is long enough, but it will still lack originality. However, the 'recognisability' approach has the advantage of being more objective, and therefore perhaps easier to apply *in abstracto*.

It also presents a practical advantage, as it is quite commonsensical: if the excerpt is not recognisable to the rightholder, then he will not sue for copyright infringement, and if it is not recognisable to the judge, its use will not be qualified as copyright infringement.

This approach may also be particularly appealing in compiling lists of n-grams to be used for training language models (e.g., for Machine Translation purposes). *Hapax legomena* should not be used for such purposes, as the resulting language model would lack the necessary context (Kamocki et al., 2016). Therefore, eliminating *hapax legomena* from such compilations of n-grams is not perceived as a constraint.

On the other hand, the approach also carries some risk of overgeneralisation, as it may lead to deleting n-grams that are certainly not copyright-protected. To illustrate, this claim, let us come back to our example from the previous section. The 2-gram discussed above as potentially original - '*krwista krówka'* - would also be a *hapax legomenon* in the National Corpus of Polish. As such, it should be deemed as non-reusable both in the 'originality' approach, and in the 'recognisability' approach. However, *'szczęśliwy Zenobiusz'* (*happy Zenobiusz* — Zenobiusz being a highly uncommon first name in Polish) is definitely not original, as there is nothing creative about the use of a basic adjective and a known (albeit uncommon) first name; and yet, this 2-gram also has no occurrences in the National Corpus of Polish. Therefore, one can hypothesise that it would be reusable under the 'originality' approach (as it is non-original), but not under the 'recognisability' theory (as its source would be easy to establish).

Another difficulty lies in the fact that it is indeed difficult to establish with enough certainty whether an n-gram is a *hapax legomenon* at the level of the entire language. It is possible (albeit rather unlikely) that the expression *'krwista krówka'* has never been uttered in Polish before (due to its creative, playful and humorous nature). It is, however, virtually impossible that the words *'szczęśliwy Zenobiusz'* have never been uttered, as there have been people named Zenobiusz, and at least some of them must have been described as happy at some point in their lives. Still, both expressions are absent from the National Corpus of Polish, which is the largest corpus of the language. Threfore, the fact that an expression does not appear in a corpus, even very large, can hardly be regarded as proof of its truly unique nature.

5 New Related Right of Press Publishers: A Trench War Has Begun

Article 15 of the new Directive 2019/790 on Copyright in the Digital Single Market (DSM Directive) introduced and harmonised a new related right of press publishers. The right protects said publishers against parasitic use of press articles by commercial news aggregators (Papadopoulou, Moustaka, 2020), and as such is not directly relevant for language research. However, this new right is only triggered when an online service uses more than 'individual words or very short extracts of a press publication'. It will therefore be necessary to define precisely, via case law or via a collective agreement, the maximum length of a 'very short extract'. In Germany, where this right was first introduced in

[24]CJEU, 29 July 2019, C-476/17 (*Pelham*)

2013 (before the DSM Directive), the Patent and Trade Mark Office initially (2015) recommended 7 consecutive words as a freely reusable 'very short extract' of a press publication. However, this recommendation no longer appears on the Office's website.[25]

In Germany, the debate on this question has recently been revived by the publication of a series of government bills on the implementation of the DSM Directive.[26] The first bill did not define the maximum length of a 'very short extract', but instead stated that headlines (*Überschriften*), which can sometimes be quite long (even longer than 7 words), should be considered as such. In the consultation process, three German associations of press publishers (BDZV, VDZ and VDL) issued a joint statement starkly disapproving the inclusion of headlines in the definition of 'very short extracts', and arguing that the freely-reusable extracts should be limited to three consecutive words (BDZV et al., 2020).

Although made in a specific context, very different from language research, the statement may be interpreted as indicating that German press publishers will generally not oppose the use of 3-grams extracted from their articles, as they consider them of little value, at least from the commercial standpoint. If press publishers are ready to let news aggregators use three-word snippets, *a fortiori* they should let language researchers do at least as much.

It should be noted here that in German a 3-gram (not to mention a 7-gram) can carry quite a lot of information, as this language uses compound nouns and is known for its unique ability to convey complex meanings in single words (e.g. *Schadenfreude*[27] or *Futterneid*,[28] to quote the most common examples). It is therefore imaginable that e.g. English- or French-language publishers would be ready to accept free use of slightly longer extracts.

Unfortunately, the most recent developments seem to indicate that there will be no clear answer to the question of the exact length of 'very short extracts' under the right of press publishers. The German government has recently published another bill on the same issue, which regrettably does not contain any definition of 'very short extracts'. In France, Google and an association of press publishers signed an agreement on the application of this new right;[29] while it is likely that it defines the minimum length of a freely re-usable snippet, its content has not yet been made public (as of 7 February 2021).

6 Conclusion

As promised in the introduction, we will try to formulate some guidelines concerning the use of short snippets of text without permission from the rightholders:

- 3-grams should be regarded as generally free from copyright.[30] Only very exceptionally and only in some jurisdictions expressions of 3 words or shorter can be found original — even then, such very short original expressions seem to occur almost exclusively in very specific contexts, as titles, slogans or in poetry. Even in very large corpora, original 3-grams are likely to be *hapax legomena;* therefore, eliminating *hapax legomena* from the list of n-grams (which

[25]But cf. the 2015 news report at: https://www.internet-law.de/2015/09/leistungsschutzrecht-dpma-schlaegt-einigung-vor.html (retrieved 4 September 2020).

[26]*Diskussionsentwurf des Bundesministeriums der Justiz und für Verbraucherschutz — Entwurf eines Ersten Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts, 15 Januar 2020*, available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/DiskE_Anpassung%20Urheberrecht_digitaler_Binnenmarkt.pdf?__blob=publicationFile&v=1 (retrieved 7 February 2021), followed by Gesetzentwurf der Bundesregierung — Entwurf eines Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, 3 Februar 2021, available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RegE_Gesetz_Anpassung_Urheberrecht_digitaler_Binnenmarkt.pdf?__blob=publicationFile&v=5 (retrieved 7 February 2021).

[27]'Enjoyment obtained from the troubles of others' (definition by Merriam-Webster Online Dictionary)

[28]Literally 'food envy', used to describe a feeling of jealousy towards someone who, while eating at the same restaurant, ordered something that looks more appetising than our meal.

[29]L'Alliance de la Presse d'Information Générale et Google France signent un accord relatif à l'utilisation des publications de presse en ligne, *Le blog officiel de Google France*, 21 janvier 2021, https://france.googleblog.com/2021/01/APIG-Google.html (retrieved 7 February 2021).

[30]Interestingly, Linden (2014) seems to have reached the exact same conclusion.

is sometimes desirable for linguistic purposes) substantially increases legal certainty about the possibility of its re-use;

- the use of 7-grams, while not risk-free, may be seen by many as a reasonable compromise. In order to mitigate the associated risk, one might attempt to reduce the 'recognisability' of text snippets by removing the elements that are 'highly identifying', such as *hapax legomena*, proper names and other named entities, or very unusual syntactic structures (e.g., 'Yoda-Speak');
- the use of 10-grams may be justified under a very liberal interpretation of the *Infopaq* case; in our opinion, however, it would carry significant risk of copyright infringement.

The guidelines presented above only apply if the use of the data is not based on a license (in which case the license should take precedence), or is not covered by a statutory exception (e.g. citation or research exception).

References

BDZV (Bundesverband Deutscher Zeitungsverleger Verband), Deutscher Zeitschriftenverleger Verband, Deutscher Lokalzeitungen. 2020. *Stellungnahme zum Diskussionsentwurf des Bundesministeriums der Justiz und für Verbraucherschutz für ein Erstes Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts vom 16. Januar 2020*.

Eckart de Castilho, Richard, Giulia Dore, Thomas Margoni, Penny Labropoulou and Iryna Gurevych. 2018. A Legal Perspective on Training Models for Natural Language Processing. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kamocki, Paweł, Jim O'Regan and Marc Stauch. 2016. All Your Data Are Belong to us. European Perspectives on Privacy Issues in 'Free' Online Machine Translation Services. In: David Aspinall, Jan Camenisch, Marit Hansen, Simone Fischer-Hübner, Charles Raab [Eds.]. *Privacy and Identity Management. Time for a Revolution?* Springer International Publishing.

Koplenig, Alexander. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169-188.

Linden, Krister. 2014. Update from language resources. Oral presentation at the Legal Issues in Language Resources and Infrastructures Workshop. LREC 2014.

Markov, A.A. 1913. Essai d'une recherche statistique sur le texte du roman "Eugène Oněgin", illustrant la liaison des épreuves en chaîne. *Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg. VI série*, 7 (3): 153–162.

Papadopoulou, Maria-Daphne and Evanthia-Maria Moustaka. 2020. Copyright and the Press Publishers Right on the Internet: Evolutions and Perspectives. In: Tatiana-Eleni Synodinou, Philippe Jougleux, Christiana Markou, Thalia Prastitou [Eds.]. *EU Internet Law in the Digital Era*. Springer: Cham.

Rosati, Eleonora. 2013. *Originality in EU Copyright: Full Harmonization Through Case Law*. Edward Elgar Publishing: Cheltenham, Northampton.

Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication". *Bell System Technical Journal*, 27 (3): 379–423.