

Copyright in the context of tooling up Corsican and other less-resourced languages

Laurent Kevers, Stella Retali-Medori

UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli
Avenue Jean Nicoli, 20250 Corte, France
{kevers.l, medori.e}@univ-corse.fr

Abstract

Anyone trying to gather linguistic resources for Natural Language Processing (NLP) will sooner or later be facing the legal aspects, mainly related to copyright, that arise from this activity. These difficulties often occur when collecting corpora, which is generally among the top priorities for processing less-resourced languages. While the current legislative framework is not adequate, it seems that positive developments are emerging. Various actions can also be considered to support this evolution.

Keywords: less-resourced languages, corpora, linguistic resources, copyright, Corsican language

Résumé

Toute personne qui essaye de rassembler des ressources linguistiques pour le Traitement Automatique d'une Langue (TAL) sera tôt ou tard confrontée aux aspects légaux, principalement liés au droit d'auteur, que soulève cette activité. Ces difficultés se matérialisent souvent lors de la collecte de corpus, qui se situe généralement parmi les premières priorités pour le traitement des langues peu dotées. Si le cadre législatif actuel n'est effectivement pas adapté, il semble que des évolutions positives se profilent. Différentes actions peuvent aussi être envisagées pour accompagner ce changement.

1. The Corsican language, a less-resourced language

Corsican is a Latin language and is part of the Italo-Romance domain. It has known various contacts and linguistic influences. From a dialectal point of view, four or even five areas are identifiable (Dalbera-Stefanaggi, 2002; Dalbera-Stefanaggi, 2007), but they constitute a *continuum* and do not prevent interunderstanding between speakers. The spelling of Corsican is, with some adaptations, based on the Italian graphic system¹. However, despite the implementation of a polynomic approach (Marcellesi, 1984) that encompasses all dialectal variants, the writing of the language is not standardized.

Nowadays, Corsican is, with French, part of a diglossic language environment, and its use is declining. The development of tools is necessary for its preservation, enhancement, transmission and promotion². A policy in the service of the Corsican language is active on the island territory, in particular for its development through new technologies. However, if several tools and works exist for the learning and linguistic description of Corsican dialects, their inclusion in the digital humanities domain remains insufficient. In particular, sites and applications dedicated to translation, lexicon and syntax contain little data in comparison with the richness and complexity of the language. On the other hand, this wealth is found on databases such as the *Banque de Données Langue Corse* (BDLC) and *Infcor*³ (*Banca di dati di a lingua corsa*). To our knowledge, there are very few resources and tools designed for Natural Language Processing (NLP) in Corsican. The ELDA 2014 report on

linguistic resources dedicated to the languages of France (Leixa et al., 2014) lists 93 resources for Corsican. More than a third of these are recordings and transcriptions from the BDLC project. Corsican therefore falls into the category of less-resourced languages.

2. Development of NLP resources and tools for Corsican

Given this observation, we have decided to work to improve the situation of the Corsican language with regard to its place in the digital world, and more particularly in the field of Natural Language Processing. To achieve this objective and start tooling up the Corsican language, we rely on the BDLC project. This project⁴ is designed in a scientific context and hosts linguistic data related to Corsican know-how and cultural traditions throughout the island territory. It is naturally linked to the *Nouvel Atlas Linguistique et ethnographique de la Corse* (NALC).

We have defined a roadmap outlining the actions to be undertaken (Kevers et al., 2019). These are generally in line with those put forward by Ceberio Berger et al. (2018). We started by collecting corpora and setting up an online consultation interface in the form of a concordancer, implementing a language detection tool and building an electronic dictionary. In the long term, we plan to work on a part-of-speech tagger.

The question we want to highlight in this article concerns the legal aspects, mainly related to copyright, that any person trying to gather linguistic resources for a language inevitably encounters. These difficulties often materialize

¹See Retali-Medori (2015)

²According to the recommendations of UNESCO Ad Hoc Expert Group on Endangered Languages (2003)

³<http://infcor.adeceec.net>

⁴See <http://bdlc.univ-corse.fr>. A synthesis of the project history is presented by (Dalbera-Stefanaggi and Retali-Medori, 2015).

when collecting corpora, which is generally among the top priorities for processing less-resourced languages.

3. Legal aspects : corpora and copyright

3.1. Introduction

In addition to documenting the language, there are many uses for corpora, starting with comparing the intuition and linguistic knowledge of language specialists with large “real” datasets. Corpora can also be useful for building lexical resources, for creating automatic processing tools, especially through machine learning, or even in the educational field.

This task faces two main obstacles: the availability of documents, preferably in a digital form, and their legal terms of use.

Apart from the question of the existence of the documents, the first difficulty is essentially technical. The first step is to identify existing resources and process them according to their nature. In the case of printed documents, it will be required to digitize them. If they are already in a digital format, conversion operations⁵ or even “harvesting”⁶ may be necessary.

The second difficulty lies in respecting the rights that apply to this content. Indeed, the copyright laws do not generally allow their free and complete use, even for research purposes. This obstacle constitutes a real limitation for research in general, and for the digital development of less-resourced languages in particular, and has therefore been highlighted on many occasions, including by Zayed et al. (2016) : *One of the big obstacles for the current research is the lack of large-scale freely-licensed heterogeneous corpora in multiple languages, which can be redistributed in the form of entire documents. [...] due to the restrictive license of the content, many corpora cannot be re-distributed because of the risk of copyright infringement.*

The task of automatic corpora building from the web⁷ is particularly affected by this problem. Tools proposed for this purpose, such as BootCaT⁸ (Baroni and Bernardini, 2004) or Sketch Engine⁹ (Kilgarriff et al., 2014), will be difficult to use if it is planned to redistribute the resources and tools created from these corpora.

3.2. Current situation

The legal analyses that are reproduced later in this article come mainly from Geiger et al. (2019). Our objective here is to summarize them by highlighting the main points, so that actors in the world of Text and Data Mining (TDM) who are not aware of the issues, can take note of it.

TDM research is faced with a legal situation that does not allow it to proceed smoothly, given its potential involvement in copyright issues : [...] *during the chain of activities*

enabling TDM research, technically some IPR relevant actions are necessary so that in the absence of a specific permission within the legal framework, TDM can lead to an infringement, (Geiger et al., 2019, p.7). In particular, copying and modification of copyrighted works may be problematic : TDM usually involves some copying, which even in case of limited excerpt might infringe the right of reproduction. [...] any reproductions resulting in the creation of a copy of a protected work along the chain of TDM activities might trigger copyright infringement. In this respect, pre-processing to standardize materials into machine-readable formats might trigger infringement of the right of reproduction, (Geiger et al., 2019, p.7-8). In addition to textual documents, these limitations also apply to the database protected by the sui generis right.

This situation also poses a problem in terms of scientific approach : [...] *contemporary research practices, striving for verifiability of TDM research results, require the ability of researchers to store source materials and to communicate them at least to their peers. From a legal perspective, this conduct could most likely trigger the infringement of the right of communication to the public, (Geiger et al., 2019, p.9). Similarly, the diffusion of models learned and derived from non-free sources, which constitute a transformed state of the original work, places researchers in a legitimate position of uncertainty about the legal implications of their work. The only elements that would be risk-free to communicate would be the final results produced by the TDM procedure : it is to be noted that the TDM output should not infringe any exclusive rights as it merely reports on the results of the TDM quantitative analysis, typically not including parts or extracts of the mined materials, (Geiger et al., 2019, p.9).*

Even if some exceptions exist and can be used, the current European legal framework does not allow the development of TDM projects in a serene manner : *All in all, the possibility of relying on existing provisions — including temporary acts of reproduction, scientific research, private use, normal use of a database, and extraction of “insubstantial parts” from a database protected by the sui generis right — without adoption of additional interpretative norms or judgements of high instances was doubtful, (Geiger et al., 2019, p.17).*

Finally, it should be noted that various European countries have taken initiatives to develop copyright exceptions for TDM. For example, France allows the : *reproduction from “lawful sources” (materials lawfully made available with the consent of the rightholders) for TDM as well as storage and communication of files created in the course of TDM research activities, (Geiger et al., 2019, p.25). However, this exception to copyright is only granted if it occurs as a part of a scientific writing. In general, without going into the details about the exceptions provided by the various countries, these legal developments ultimately retain a certain degree of uncertainty as to their real ability to meet the legal needs faced by the TDM community. Moreover, the lack of a uniform approach should be underlined.*

⁵Such as switching from PDF to text format.

⁶For content published in the form of websites.

⁷An ACL *Special Interest Group* (SIG) is dedicated to this domain under the name of *Web AS Corpus* (SIGWAC - <https://www.sigwac.org.uk/>).

⁸<https://bootcat.dipintra.it/>

⁹<https://www.sketchengine.eu/>

3.3. Future developments

However, Directive 2019/790/EU of the European Parliament and of the Council on copyright and related rights in the Digital Single Market¹⁰, adopted on 17 April 2019, should improve the situation.

Indeed, this text introduces new exceptions to copyright, in particular *for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access* (article 3, paragraph 1).

In addition to this exception specific to the field of scientific research, a more general exception is also provided (article 4). However, there is a restriction on the latter which limits its application to cases where the works concerned *has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online* (article 4, paragraph 3).

It should be noted that this directive must be transposed into the national laws of the Member States in order to be implemented, which should be the case by 2021.

4. Discussion

In light of the above, we wanted to identify some points for which actions can be undertaken.

- In the meantime, until a perfectly adapted legal framework is in place, we think it is necessary to initiate or continue work to provide adequate resources for TDM research. In this respect, the use of licenses that allow certain exceptions to copyright — for example the Creative Commons¹¹ family — should be encouraged and intensified. Obviously, the dialogue with the rights holders is a complex, sometimes time-consuming task that goes beyond the strict framework of research activities, but worth leading.
- It could be helpful to take initiatives to make legal information on copyright issues more visible and more easily accessible to a research audience which, while generally of good will, does not necessarily give all the necessary attention to these questions. In a sense, this article is a modest contribution to this goal. Our wish would be that this question could be taken into account in a more in-depth manner, by teams of both lawyers and researchers, and that it gives rise to a wider dissemination.
- Finally, it also seems useful to raise awareness among the legislative bodies in order to change the legal framework as quickly as possible, in particular with regard to the transposition of the Directive.

5. Bibliographical References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*.

Ceberio Berger, K., Gurrutxaga Hernaiz, A., Baroni, P., Hicks, D., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*. The Digital Language Diversity Project. Available at http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf.

Dalbera-Stefanaggi, M.-J. and Retali-Medori, S. (2015). Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse. In Stella Retali-Medori, editor, *Actes du colloque Tribune des chercheurs, études en linguistique*, volume 6 of *Corse d'hier et de demain - Nouvelle série*, pages 17–25, Bastia, France, June. Société des Sciences Historiques et Naturelles de la Corse.

Dalbera-Stefanaggi, M.-J. (2002). *La langue corse*. Number 3641 in *Que sais-je?* PUF, Paris, June.

Dalbera-Stefanaggi, M.-J. (2007). *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*. Comité des travaux historiques et scientifiques - CTHS, Ajaccio : Paris, Alain Piazzola edition, December.

Geiger, C., Frosio, G., and Bulayenko, O. (2019). Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU. Research Paper No. 2019-08, Centre for International Intellectual Property Studies (CEIPI).

Kevers, L., Guéniot, F., Tognotti, A. G., and Retali-Medori, S. (2019). Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC. In *Actes de la 26e conférence sur le Traitement automatique des langues naturelles (TALN)*, Toulouse, France, July.

Kilgarrieff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36, July.

Leixa, J., Mapelli, V., and Choukri, K. (2014). *Inventaire des ressources linguistiques des langues de France*. ELDA, September. Available at http://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf.

Marcellesi, J.-B. (1984). La définition des langues en domaine roman : les enseignements à tirer de la situation corse. In *Actes du Congrès de Linguistique et de Philologie Romanes 5*, pages 307–314, Aix-en-Provence.

Retali-Medori, S. (2015). La documentation corse. In Eugen Roegiest Maria Iliescu, editor, *Anthologies, textes, corpus et sources des langues romanes*, number 7 in *Manuals of Romance Linguistics*, pages 558–564. De Gruyter, Tübingen.

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language Vitality and Endangerment. Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*. United Nations Educational, Scientific and Cultural Organization, Paris. Available at <https://ich.unesco.org/doc/src/00120-EN.pdf>.

Zayed, O., Habernal, I., and Gurevych, I. (2016).

¹⁰<http://data.europa.eu/eli/dir/2019/790/oj>

¹¹<https://creativecommons.org/licenses/>

C4corpus: Multilingual Web-size Corpus with Free License. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.