Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 559–563
Marseille, 20-25 June 2022

Ethical Issues in Language Resources and Language Technology – Tentative Taxonomy

Paweł Kamocki, Andreas Witt
Leibniz-Institut für Deutsche Sprache
R5 6-13, 68161 Mannheim, Germany
{kamocki, witt}@ids-mannheim.de

Abstract

Ethical issues in Language Resources and Language Technology are often invoked, but rarely discussed. This is at least partly because little work has been done to systematize ethical issues and principles applicable in the fields of Language Resources and Language Technology. This paper provides an overview of ethical issues that arise at different stages of Language Resources and Language Technology development, from the conception phase through the construction phase to the use phase. Based on this overview, the authors propose a tentative taxonomy of ethical issues in Language Resources and Language Technology, built around five principles: Privacy, Property, Equality, Transparency and Freedom. The authors hope that this tentative taxonomy will facilitate ethical assessment of projects in the field of Language Resources and Language Technology, and structure the discussion on ethical issues in this domain, which may eventually lead to the adoption of a universally accepted Code of Ethics of the Language Resources and Language Technology community.

Keywords: ethics, privacy, language resources, language technology

1. Introduction

Ethical issues are being invoked more and more often in the context of Language Resources (LR) and Language Technology (LT). Frequently, they serve as the ultimate argument used to prevent certain projects from being carried out, or certain developments from taking place. Rather ironically, the very notion of ethics remains quite vague, as there seems to be no objective, commonly agreed-upon definition or framework for ethical considerations that should be taken into account in building LR and developing LT tools – and still the 'ethical' argument is so authoritative that few are those who dare argue with it.

Although the LR and LT community has paid a lot of attention to 'legal and ethical issues' for at least a decade now, the subject seems to have been dominated by legal issues, which is obviously not surprising, as it is much easier to discuss a framework with clearly defined boundaries. While it is usually easy to determine what is and what is not binding law (although the exact interpretation of this law can and should be debated), the same cannot be said about ethical norms, especially those that should apply in the field of LR and LT. With few notable exceptions such as Leidner and Plachouras (2017), meaningful debate on ethical principles has taken (or is taking) place only in some very specific domains of linguistics such as atypical communication (Heuvel et al., 2020) or endangered languages (O'Meara and Good, 2010; Rice, 2011), which gives a false impression that only such fields, and not LR and LT in general, are really concerned with ethical issues.

The purpose of this paper is to attempt to provide a taxonomy of ethical issues in LR and LT, based on the existing literature and the authors' considerable (although still subjective) experience in this sector. It is not its ambition to be exhaustive or authoritative in any way; rather, the authors' aim is to spur a debate on the subject in the language community, and possibly initiate the creation of something that one day would become a commonly agreed-upon Code of Ethics for LR and LT. It should be noted that the authors are not, and do not pretend to be, trained ethicians; instead, their field of expertise are such areas as LR and LT, research data management, research policy and information law.

Among many approaches to ethics, two should be briefly presented in the introduction to this paper, i.e. deontology and consequentialism. In deontological ethics, an action is evaluated on the basis of the action itself and its compliance with pre-defined rules (cf. prescriptivism in linguistics); in consequentialism, the evaluation is based purely on the consequences of the action (cf. descriptivism in linguistics). It is important to specify that this paper adopts the deontological approach, as this is the only one that can be applied *a priori*. It is possible, however, that an action complying with deontological principles may still lead to 'wrong' consequences, and therefore be wrong from the point of view consequentialism.

1.1 Specificity of Ethical Issues in LR and LT

Needless to say, LR and LT do not evolve in isolation. Some relatively well-described ethical frameworks that may seem to include LR and LT in their scope already exist, like scientific ethos, ethics of technology, or Artificial Intelligence Ethics.

Scientific ethos evolves around the normative principles formulated by Robert King Merton (1973). The so-called Mertonian norms are: Communism, Universalism, Disinterestedness and Organised Scepticism. Today, these principles remain the main frame of reference in academia (cf. e.g. Anderson et al. 2010), although some of them (like Communism, embodied in the Open Access/Open Science movement) seem less controversial than others (like Disintrestedness, see Macfarlane, Cheng, 2008). However, both LT and LR exist and thrive also outside of academia. For example, any commercial use of LR could be regarded as contrary to the principle of Communism (common ownership of results) and Disintrestedness (no pursuit of monetary gains). Therefore, Mertonian norms alone are not an appropriate code of ethics for LR and LT.

Technology of ethics (or technoethics) is a relatively new field; although its origins can be traced back to Plato's dialogue *Phaedrus* (critically discussing the 'invention' of writing) or to 19th century American pragmatism, it was only established as an independent branch of ethics in the 1970s (Bunge, 1977). It studies both the ethical implications of developing new technologies, and the effect that technology has on ethical questions and human condition in general. While very interesting for anyone involved in new technologies, technoethics is too abstract to serve as a useful set of guidelines for LR and LT. Moreover, it tends to adopt a consequentialist (or: descriptive) approach, and the LR and LT community is in need of practically implementable deontological (or: prescriptive) rules. Finally, technoethics is too broad in scope and too future-oriented to be more than a distant source of inspiration for ethics in LR and LT.

LR and LT can, however, draw some inspiration from the field of Artificial Intelligence (AI) ethics. The origins of AI ethics can be traced back to Asimov's (1942) Three Laws of Robotics, but the field has developed rapidly in recent years, as landmarked with the adoption of such documents as the Asilomar AI Principles[1] or, very recently, the UNESCO Recommendation on the Ethics of Artificial Intelligence (2022). Further, possibly even more normatively ambitious developments in AI ethics are expected after the adoption of the EU Artificial Intelligence Act proposed in 2021, which emphasizes the role of self-assessment of AI application providers. In its current state, however, AI ethics is not fully applicable to LT and/or Machine Learning for Natural Language Processing and other linguistic applications. Since AI ethics addresses a much broader field, it necessarily overlooks the specificity of LR and LT; on the other hand, some issues in LR, e.g. those concerning fieldwork, are not (or only very remotely) related to AI, and therefore are not within the scope of AI ethics.

LR and LT, therefore, need their own code of ethics, which demonstrates the utility of research in this domain, to which this paper wants to contribute.

1.2 Ethics and Law

The relation between law and ethics is a complex subject to say the least. A simple distinction between legal norms and ethical (or moral) norms may be that the respect of legal norms is guaranteed by the coercive power of the State, whereas moral norms only have a social sanction (e.g. coming to a social dinner empty-handed may eventually lead to ostracism), or even no sanction at all (e.g. various sorts of 'evil thoughts' are perceived as morally wrong, but for obvious reasons cannot be sanctioned, unless by the thinker himself). In this approach, which can be traced back to Kelsen (1991), law and ethics are two disjoint sets of norms, i.e. a norm either is guaranteed by state coercion (legal norm), or it is not (moral norm).

Another view on the relation between law and ethics consists of seeing them as two overlapping sets of norms. In this approach, human acts can be: a) ethically wrong, but not illegal (e.g., excessive use of swear words), or b) both illegal and ethically wrong (e.g., theft) or, c) illegal, but not ethically wrong (the proponents of this approach could quote certain tax offences as an example).

The authors of this paper do not subscribe to either of these points of view. Rather, they believe that legal norms are a sub-section of ethical norms, i.e. every illegal act is necessarily morally wrong. Debatable as this approach might be, this is the premise of this paper.

Another characteristic that distinguishes ethics from law is its universalism. Law generally has a limited territorial scope, whereas ethical principles, if not strictly identical from one territory to another, are more widely shared.

2. Overview of Ethical Issues in Language Resources and Language Technology

This section provides an overview of ethical issues to consider at various stages of building LR and developing LT tools.

2.1 Conception Phase

Some ethical issues must be considered already at the conception phase, i.e., before data is collected for a LR, or before an LT tool is developed.

One of such issues for LT applications is the impact on user privacy (cf. Privacy by Design (Kamocki, Witt, 2020), and Art. 25 of the GDPR). It should not be mistaken for protection of personal data as defined in the GDPR. In fact, contrary to a common misconception, personal data protection is not a synonym of privacy. Personal data may have nothing to do with what is generally considered as 'private sphere' of an individual's life (e.g., one's professional phone number is personal data), and privacy infringement may have nothing to do with processing of personal data (e.g., knocking on one's door in the middle of the night, or observing one's driveway). Privacy in the ethical sense can be defined as both *'freedom from unauthorised intrusion'*[2] and as a *'right to keep* [one's] *personal matters and relationships secret'*[3]. Therefore, an LT tool (such as a spell checker, a chatbot or a voice assistant) should be designed in such a way as not to be unnecessarily intrusive: it should not collect and process more information than necessary (e.g., a voice assistant should not record when not activated by the user's command), and should not try to interact with the user without being expressly asked to, e.g. by sending excessive push notifications. Moreover, the user should be enabled to quickly and easily deactivate certain features of the tool (e.g., voice recording) temporarily or permanently. By the same token, if in a data collection process participants are asked to fill out a questionnaire, the questions should be designed in such a way as not to be unnecessarily intrusive (e.g., it would rarely be justified to ask for date of birth or full address).

In LRs, data selection can also be problematic from an ethical standpoint. The principles of representativeness and

[1] https://futureoflife.org/2017/08/11/ai-principles/ (access: 27 April 2022).

[2] Merriam-Webster Dictionary, 'privacy' definition 1.b (https://www.merriam-webster.com/dictionary/privacy).

[3] Cambridge English Dictionary (https://dictionary.cambridge.org/dictionary/english/privacy).

balance in corpus design (e.g., Kupietz, 2015), well known in the language research community, also have an ethical dimension, since an imbalanced or non-representative corpus can have discriminatory effect, e.g. if used for training language models.[4] All sorts of linguistically relevant characteristics that may lead to potential biases should be considered: gender, race, dialect, age, etc. An important dilemma in data selection is whether the corpus should reflect the reality as it is (e.g., in a corpus of corporate executives' speech most data should come from the most represented group, presumably white males), or as it should be in a perfectly balanced reality (e.g., the same amount of male/female and black/white/Asian corporate executives should be represented). The decision is difficult to make; it should depend on what the corpus is to be used for, and it should be properly documented for future uses. For example, a voice assistant should be trained on a corpus in which many characteristics are equally represented (even if this does not reflect the reality), so that no group of users (such as people with a strong regional accent) is discriminated against when using the assistant.

Apart from data selection, the selection of people for tedious tasks such as fieldwork may also lead to biased results. In academic context, most fieldwork is performed by students, usually still in their twenties, which may impact the results (e.g. elderly speakers may feel intimidated, and younger speakers may be more eager than average to participate in a survey conducted by a young researcher).

In both LR and LT, it is also important to properly document the conception phase. In academia, this is part of the requirement of reproducibility of research results, and one of the aspects of Open Science; when personal data are being processed, transparency *vis-à-vis* the data subjects is also required (Art. 5.1(a) of the GDPR). But the importance of transparent documentation exceeds the above-mentioned requirements; it is a foundation of 'explainable' technology requirements which can help prevent the 'AI blackbox' phenomenon (i.e. a situation where the functioning of AI is completely obfuscated to the user, who has no way to falsify the experiment and can only blindly trust its result – see Rudin and Radin, 2019), thereby reinforcing users' trust in the LT.

2.2 Creation Phase

Creation of LR and LT tools usually involves the use of third-party intellectual property (IP). Language data are, as a general rule, protected by intellectual property: individual pieces of data, even as short as 3-grams (Kamocki, 2020) are usually protected by copyright; larger datasets can be protected cumulatively by copyright and the *sui generis* database right. In the authors' approach as explained *supra*, infringement of any of these legal frameworks is to be seen as a wrong (unethical, immoral) act. Moreover, IP is a form of property and enjoys the same protection as a fundamental right, which gives it an even stronger ethical dimension. However, ethical handling of someone else's IP is not reduced to respecting applicable legal norms. For example, in the monist approach to copyright (in countries like Germany and Austria), moral rights (including the right of paternity, i.e. recognition of authorship) are limited in time and expire together with economic rights (70 years after the death of the author); in such context, attributing authorship to such authors like Leibniz or Goethe is a purely ethical requirement, albeit so deeply rooted in common moral sense that it would be nearly insane to even think about overlooking it.

Some aspects of the IP legal framework are also particularly sensible from the point of view of ethics; this is, for example, the case of orphan works, i.e. works protected by copyright whose rightsholders cannot be identified or located despite 'diligent search'. In the EU, under Directive 2012/28/EU, orphan works can be used by certain organizations such as libraries and educational establishments. Although the notion of 'diligent search' is quite specifically defined in the law, it still has a strong ethical component, and people at beneficiary institutions should use their moral compass to evaluate when their efforts to find the rightsholders meet the 'diligent' threshold.

Another issue, this time of more ethical than legal nature, concerns mostly the creation of LRs for endangered languages. The rights of indigenous peoples in their cultural expressions, including language, have attracted considerable attention for more than two decades now. In 2007, the United Nations adopted a Declaration on the Rights of Indigenous Peoples (UNDRIP), Article 31 of which reads as follows: *"Indigenous peoples have the right to maintain, control, protect and develop their cultural heritage (...) and traditional cultural expressions, as well as the manifestations of their (...) cultures, including (...) oral traditions* [and] *literatures (...). They also have the right to maintain, control, protect and develop their intellectual property over such cultural heritage (...) and traditional cultural expressions"*. The Declaration remains, for the most part and in most countries, a soft law instrument, the enforcement of which is based on morality rather than on state coercion. The research community is gradually adopting the CARE principles of Indigenous Data Governance, formulated in 2018, largely to complete and supplant the access- and reuse-oriented FAIR principles (Carroll et al., 2020). The CARE principles are: Collective benefit, Authority to control, Responsibility and Ethics. In particular, the Ethics principle invites not to portray Indigenous Peoples in terms of deficit, to assess any potential benefits and harms from the Indigenous Peoples' perspective, to address imbalance of power in processing Indigenous Data, to involve representatives of the concerned communities in the decisions about the processing of their data, and to take into account potential future use and future harm resulting from the data.

Privacy of individuals who contribute data to a LR also needs to be addressed at the creation stage – some data may need to be pseudonymised or anonymised, or even deleted altogether, if they are judged as presenting too much of a risk from a privacy standpoint. Again, this requirement exceeds pure compliance with the GDPR – for example, information related to deceased persons is not considered

[4] See e.g., Bender, E., A Typology of Ethical Risks in Language Technology, 2020 online lecture available at https://www.youtube.com/watch?v=DLGwIfjoh3w (access: 14 January 2022).

personal data by the GDPR, and yet good practice requires preserving their 'privacy' in LR and LT.

Finally, contributors should be given a chance to get thoroughly informed about the purposes for which the data are collected and how they will be used. Arguably, most people are not interested in, and are not equipped to understand the technicalities, but they nevertheless have the right to know what they are contributing to. Needless to say, all contributions should be voluntary, and all contributors should be given the possibility to change their mind and withdraw their data from the dataset if this is reasonably practicable. This right may even go beyond the GDPR's rights to be forgotten and to withdraw consent, and be applied also to anonymized data and other non-personal data.

2.3 Use Phase

The use of LT tools, even those developed according to the highest ethical standard, should also abide by ethical principles.

Firstly, it should be made clear to the user that he or she is in presence of an LT output when there is no human intervention involved. The user should be aware that the text he or she is reading was machine-translated or otherwise automatically generated (without a quality check by a human), or that he or she is interacting with a chatbot or a voice assistant rather than a human. In fact, LT often produces outputs that are indistinguishable in quality from human production, without providing the same level of trustworthiness. For example, an MT system may generate a grammatically perfect output with semantic mistakes that would never be made by a human (e.g., using 'he' instead of 'she', or translating French *'serviette'* as 'towel' instead of 'briefcase'). Transparency about the use of LT could also help prevent the spread of computer-generated 'fake news', some of which may start as seemingly innocuous jokes.

Secondly, LT should not be used to make decisions or choices that may seriously impact the user – such decisions should always be verified and confirmed by a human (cf. Art. 22 of the GDPR). Whether a decision is 'serious enough' to require human intervention is a matter of circumstances: if a customer is misunderstood by a chatbot and receives her T-shirt in the wrong size with a possibility to return it, little or no harm is done; if, however, the user calls an ambulance which arrives at a wrong address, the consequences might be dire.

Thirdly, all privacy-related concerns that should be envisaged at the conception phase should still actively be addressed during the use phase.

2.4 Ethical Issues in LR and LT Evaluation

Evaluation of LR and LT also presents some ethical challenges. In order to produce sound results, it should be carried out according to transparent, objective and up-to-date criteria. The choice of metrics to measure the impact of data (including language data) is one the biggest questions that the academic community is currently facing (Lampert et al., 2017). The fact that practice in the field of data citation is still evolving does not make this task any easier.

3. Tentative Taxonomy of Ethical Issues in Language Resources and Language Technology

Based on the overview provided in the previous section, the authors would like to propose a taxonomy of ethical issues in LR and LT built around the principles of Privacy, Property, Equality, Transparency and Freedom, which can be summarized as follows:

- **Privacy:** stakeholders (data providers, users) should be protected against disproportionate intrusion and allowed to keep certain information secret;
- **Property:** intellectual and cultural property should be handled with respect, in compliance with applicable law, ensuring that any potential harm (evaluated from the owner's perspective) is outweighed by collective benefit;
- **Equality:** no group of stakeholders or contributors should be directly or indirectly discriminated against;
- **Transparency:** LT outputs should be clearly marked as such; stakeholders should be informed about the main principles of, and given a possibility to learn the details about the functioning of LT;
- **Freedom:** data providers should be free to contribute their data to LR&LT, and, to a reasonably practicable extent, to change their mind at any later stage; human intervention should be necessary and decisive in any process involving the use of LT the outcome of which may seriously impact the user.

4. Conclusion

Although ethical issues in LR and LT may appear to be a vague topic, an overview of these issues in different phases of LR and LT shows that they concentrate around five recurring themes: Privacy, Property, Equality, Transparency and Freedom. A tentative taxonomy of these issues proposed by the authors will hopefully facilitate ethical assessment of LR and LT projects, and help structure the discussion on ethical issues in LR and LT, which may eventually lead to the adoption of a universally accepted Code of Ethics of the LR and LT community.

5. Bibliographical References

Anderson, M.. S., Ronning, E. A., DeVries, R., Martinson, B. C. (2010). Extending the Mertonian Norms: Scientists' Subscription to Norms of Research. *Journal of Higher Education* 81(3): 366–393.

Asimov, I. (1942). Runaround. *Astounding* 29(1): 94-103.

Bunge, M. (1977). Towards a Technoethics. *Monist* 60(1): 96–107.

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J. and Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal,* 19(1):43.

Heuvel, H. van den, Oostdijk, N.H.J., Rowland, C.F. & Trilsbeek, P. (2020). The CLARIN Knowledge Centre for Atypical Communication Expertise. In *LREC 2020 Conference Proceedings*, pp. 3312-3316. ELRA.

Kelsen, H. (1991). *General Theory of Norms*. Oxford University Press.

Kamocki, P. (2021). When Size Matters. Legal Perspective(s) on N-grams. In Navaretta, C. and Eskevich, M. (Eds.) *Selected Papers from the CLARIN Annual Conference 2020,* pp. 122-128, virtual event, October. Linköping University Electronic Press, 2021.

Kamocki, P., Witt, A. (2020). Privacy by Design and Language Resources. In *LREC 2020 Conference Proceedings*, pp. 3423-3427. ELRA.

Kupietz, M. (2015). Constructing a corpus. In Durkin, P. (Ed), *The Oxford Handbook of Lexicography*. Oxford University Press.

Lampert, D., Lindorfer, M., Prem, E., Irran, J. and Sanz, F. S. (2017). New indicators for open science - Possible ways of measuring the uptake and impact of open science. *fteval Journal for Research and Technology Policy Evaluation*, 44: 50-56.

Leidner, J. L. and Plachouras, V. (2017). Ethical by Design: Ethics Best Practices for Natural Language Processing, In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pp. 30–40, Valencia, Spain, April. Association for Computational Linguistics

Macfarlane, B. and Cheng, M. (2008). Communism, Universalism and Disinterestedness: Re-examining Contemporary Support among Academics for Merton's Scientific Norms. *Journal of Academic Ethics*, 6: 67–78.

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

O'Meara, C. and Good, J. (2010). Ethical issues in legacy language resources. *Language & Communication* 30(3):162-170.

Rice, K. (2012). Ethical Issues in Linguistic Fieldwork. In Thieberger, N. (Ed.), *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press.

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, *1*(2).

UNESCO (2022). *Recommendation on the Ethics of Artificial Intelligence*. Adopted on 23 November 2021. UNESCO.