Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-71

The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources

Jane KLAVAN[a,1], Arvi TAVAST[b], and Aleksei KELLI[a]
[a] *University of Tartu*
[b] *Qlaara Lab*

Abstract. The present paper focuses on linguistic experiments as a source of language resources (LRs). It addresses some of the legal requirements regulating their collection. The primary focus of the paper is on processing personal data (PD), especially how the General Data Protection Regulations (GDPR) defines personal data, and what it means to anonymize personal data.

Keywords. Language resources, linguistic experiments, crowdsourcing, personal data, General Data Protection Regulation (GDPR)

1. Introduction

Linguistic experiments are a potentially powerful source for creating language resources (LRs). However, researchers need to be aware of the relevant legal requirements regarding their collection and use. In our article, we address some of the legal and practical issues that pertain to the collection of data for linguistic research using different sourcing models, including crowdsourcing. The focus is on processing[2] personal data. The paper draws on previous research [2][3][4] and develops it further to offer insights and recommendations to facilitate language research. The paper contributes to the field of creating and using language resources by raising awareness of the opportunities and limitations of using crowdsourced experimental data for linguistic research. We aim to provide a more interdisciplinary approach to the field of human language technologies.

The first part of the paper discusses experimental data as a source for linguistic resources. It lists some of the possible types of data that can be gathered using linguistic experiments, such as responses to questions, reaction times, eye movements, articulographic data, audio and video recordings, and demographic data. We focus on three possible ways of getting people to participate in a linguistic experiment – using a laboratory setting, a general-purpose crowdsourcing platform available online and finding volunteers online. We then highlight some of the advantages and disadvantages that come with these possible ways of collecting experimental data. Once the linguistic

[1] Corresponding Author: Jane Klavan; E-mail: jane.klavan@gmail.com.

[2] The General Data Protection Regulations defines processing as "any operation or set of operations which is performed on personal data or on sets of personal data" [1]. This means that all activities related to data are regarded as processing.

data have been gathered, they need to be tied with demographic data like gender, age, level of education, and so forth. The language data often constitutes personal data. The second part of the paper, therefore, highlights some of the issues related to processing personal data. We mainly focus on the General Data Protection Regulations (GDPR) and dissect how the GDPR defines personal data, and what it means to anonymize personal data.

2. Using experimental data for linguistic research

The role of experimental data is to build or evaluate statistical models about language. A language model is the probability distribution of certain types of linguistic units, most commonly ngrams of words or characters, in the training material. Unless trained on a single speaker (see below), the objective of the model is to describe language, not any particular speaker. To the contrary, the overrepresentation of a single source in the training material would be a severe flaw and reduce the usefulness of such material. A case in point is the proportion of European legislation in current corpora. It has been included since it is readily available and free of legal restrictions[3], but the result is that models trained on those corpora model European legal language, not language as a whole. Such potential flaws notwithstanding, a language model is still a collection of probabilities, frequencies or weights. It does not contain any representation of the training material, and there exists no method of reverse engineering the model to recreate the training material.

The following is a non-inclusive list of the possible types of data that are processed using linguistic experiments:

- Responses to multiple-choice questions. Examples: Is this a word (yes/no or Likert scale)? How similar are these two expressions? Which of these expressions sounds more natural?
- Test questions to measure understanding. In reading experiments, it is common to ask questions about the content of the text to determine whether participants understood what they read.
- Responses to open questions, like mini-essays.
- Reaction times. Responses to multiple-choice questions can be timed, as can self-paced reading (how long it takes to read a piece of text).
- Eye-tracking data. In reading experiments, the location in the text that the participant is looking at.
- Articulographic data. In research of speech production, movement data from sensors placed on the participant.
- Audio recordings of language production. Example: word naming, where participants pronounce words shown on screen.
- Audio recordings of metadata. Example: think-aloud protocols used in translation process research, where participants comment on their actions while translating.

[3] Legal acts are not copyright protected and they do not usually contain personal data. For instance, the Estonian Copyright Act (*Autoriõiguse seadus*) does not apply to legislation and administrative documents and court decisions (§ 5) [5].

- Measurements of audio recordings. It is often sufficient to use numerical data of the recording (e.g. latency, duration, pitch, intensity), rather than the recording itself.
- Video recordings. In research of speech production and multi-modal communication, video of the participant's face while speaking.
- Demographic data (e.g. age, education, dialect background) for sociolinguistic research.

For researchers conducting linguistic experiments, there are mainly three ways of getting people to participate in an experiment: in a laboratory setting [6][7], using general-purpose crowdsourcing platforms available online (see [8] for a comprehensive overview of using Amazon Mechanical Turk), and finding volunteers online [9]. Each of these ways come with their own set of advantages and disadvantages that are discussed briefly below.

The laboratory setting comes with the apparent benefit of the researchers having control over the conditions in which the experiment is being conducted (something that may be essential depending on the nature of the research question). Another benefit of the laboratory setting is pertinent when the researchers need to use specialized equipment for measuring reaction times or eye movements. In such cases, it may be essential that all the participants be measured with the same device to control for any variation that may be due to the measurement device. As for the participants themselves, university students of psychology and linguistics tend to be overrepresented in linguistic experiments conducted in laboratory settings. It is more difficult (or less convenient) to obtain data on participants who are older, have different educational backgrounds or are not located within commuting distance of a well-equipped research centre. To target samples from other populations besides the university students, it may be necessary for the researchers to go out in the field. The development of technical equipment makes it easier to use portable devices, but even then researchers may need to keep the room of the experiment a constant. The upside of using student participants is that they understand the necessity of conducting experiments and may, therefore, be more willing to participate in such studies and do not object to their data being used for research purposes.

One of the most manifest disadvantages of using general-purpose online crowdsourcing platforms like the Amazon Mechanical Turk [8] is that for the participants there is no demographic data available - data which for linguistic research is highly relevant. For example, [10] point out that while they explicitly asked only Turkers whose first language was English to work on their task, they had no method of enforcing this. Furthermore, such crowdsourcing platforms may be useful for a language like English, but their availability for smaller languages is limited. [11] have pointed out that they had to be very creative to find Korean transcribers and that "cultivating workers for a new language is definitely a 'hands on' process". An additional concern with such online platforms is the motivation of crowdsourced participants. The participants provide data in exchange for a monetary payment, and in some cases, we can no longer talk about "naive" participants in experiments but rather "professional" participants for whom participating has become a way to earn an income. Still, the allure of having access to a vast amount of data is tempting and convenient, and the value of these general-purpose online crowdsourcing platforms should not be underestimated.

The third way of getting people to participate in a linguistic experiment is to get them to participate in a large-scale online experiment on a voluntary basis [9]. In such a

case, the participants are self-selected. One of the disadvantages is that the researchers have no control over the conditions of the experiment. Even though researchers nowadays have means to measure reaction time for an experiment that is done online and outside the laboratory setting in the participants own time, the data may not be reliable since the differences in reaction time may be due to the speed of the web browser or other technical matters. Still, it is far more comfortable for people to participate in an experiment at the time most convenient for them rather than taking the trouble of going down to the laboratory.

Our own experience with crowdsourced linguistic data come from conducting an online experiment about morpho-syntactic alternations in Estonian with the Qlaara platform. We combined the following two approaches to recruit participants for the study - having students participate in the experiment in a controlled setting and finding volunteers online. We were able to recruit 145 participants for the study during roughly 10 months when the experiment was open. One of the problems we encountered was that only 103 of the participants had completed the whole questionnaire with 60 items.

Once the data have been gathered - in whichever way is deemed the best - the data need to be tied with demographic data like gender, age, level of education. Information that is of interest to a linguist constitutes often personal data. We now turn to the issues related to processing personal data collected via linguistic experiments.

3. Personal data in linguistic experiments

The processing of language resources must be compatible with legal requirements. It poses several challenges. From the legal perspective, LRs are a complex phenomenon which can be explained with the following graph:

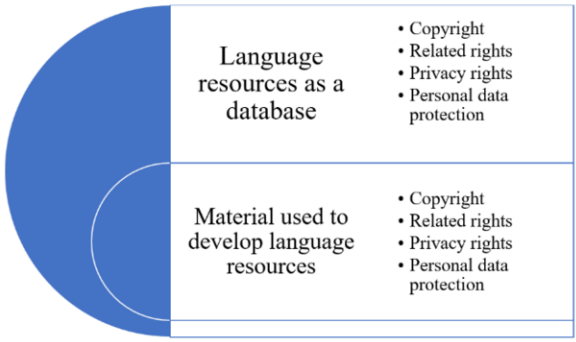

Figure 1. Two tiers of rights covering language resources [2].

The material used as input to LRs often include copyright protected works and personal data (PD). LRs themselves are protected as a database. Recordings of oral speech are often protected as performances. Since the paper concentrates on personal data, intellectual property issues are not addressed (for further discussion see [12] [2]).

Language resources often include personal data (person's voice, psychological characteristic, demographic data and so forth). The General Data Protection Regulation (GDPR) defines personal data as "any information relating to an identified or

identifiable natural person" [1]. According to the Article 29 Working Party (WP29)[4] "Identification is normally achieved through particular pieces of information which we may call "identifiers" and which hold a particularly privileged and close relationship with the particular individual. Examples are outward signs of the appearance of this person, like height, hair colour, clothing, etc… or a quality of the person which cannot be immediately perceived, like a profession, a function, a name etc" [14]. Due to the extensive definition of personal data, GDPR concerns a considerable amount of LRs.

Non-personal data is not subject to personal data protection.[5] According to the literature "For the controller[6], there is a strong incentive to anonymise data. Through anonymisation the data are placed outside the scope of data protection; by making data non-identifiable, the controller is relieved of the burden of compliance with data protection's rules and limitations" [15]. It is even suggested that "the categorization of data as identifiable or non-identifiable is a self-assessment of the controller" [15]. The authors agree that the controller has to decide whether to treat data as anonymous or not. However, the assessment cannot be entirely subjective, and it can be challenged in case of dispute.

Data derived from personal data is also outside of personal data protection. The Court of Justice of the European Union (CJEU) has ruled that the data in the legal analysis contained in that document, are personal data, whereas, by contrast, that analysis cannot in itself be so classified [16]. With reference to the case, it is suggested that "A model derived from big data does not as such constitute personal data" [15]. The situation is more complicated with the information that is not non-personal from day one. This means that anonymization of personal data is considered the processing of personal data and subject to legal requirements. WP29 is also of the opinion that "Anonymization constitutes a further processing of personal data; as such, it must satisfy the requirement of compatibility by having regard to the legal grounds and circumstances of the further processing" [17].

Protection of personal data is not without limitations. Recital 4 of the GDPR emphasis the need to protect linguistic diversity [1]. This sets a conceptual framework for processing personal data in language research. The processing of language resources containing personal data must have a legal ground (consent or research done for legitimate or public interest). The GDPR provides several conditions (freely given, specific, informed, explicit) for consent. The data subject can withdraw the consent at any time [1]. The language research can also be conducted without the consent of the data subject if certain conditions are met (the data subject's rights are respected, safeguards are in place and so forth).

[4] According to the Data Protection Directive [13] WP29 has an advisory status on data protection. The Data protection Directive is repealed with effect from 25 May 2018 (GDPR Art. 94). However, since the basic concept of data protection remain the same, opinions of WP29 remain relevant until they are replaced with new ones.

[5] It is also emphasised in GDPR that "the principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable" [1].

[6] The GDPR defines controller as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" [1].

4. Management of personal data in linguistic research

The consent is a legal ground for processing personal data [1]. One of the points that the consent form needs to address is that the person retains their right to withdraw their consent at any time [1]. Responsibility for processing of personal data is shared between the researcher and the crowdsourcing platform[7]. Various technical measures (e.g. anonymization, adding noise) can be taken by the crowdsourcing platform to mitigate PD related risks. The objective of such measures is to remove personal data from the linguistic material and only share a "clean" version that is no longer subject to the GDPR. There is no guarantee, however, that the measures will prove to be sufficient. While personal data usually needs to be processed for creating the language model, the model itself can be anonymized so that it no longer contains personal data. The model is then disconnected from the data it is based on, and use of such models in linguistic research and language technology does not count as the processing of personal data. The following are situations when disconnecting the model from the data requires additional procedures of anonymization.

- Personal data explicitly mentioned in the linguistic material. In a word-based model, if an element of personal information is no longer than a word (names, e-mail addresses, phone numbers, etc.), then these words will by default be included in the language model. To avoid this, the data must be anonymised for the model, removing personal information or replacing it with generic placeholders.
- Identifiable idiolect. In rare cases, some ways of expressing may be so characteristic to a person that they can be recognised. Removing or replacing such expressions would be sufficient, but unlike names and phone numbers, they are almost impossible to identify automatically.
- Identifiable combinations of demographic data. Depending on the size of the language community and the granularity of demographic data gathered, it may easily be possible to identify persons based on combinations of their data. A solution may be to gather less demographic data or leave it out of the model.
- Audio and video recordings. In research and development of speech technology, summary statistics are often not sufficient, and it is unavoidable to work with (portions of) the original audio and video recordings. Such work is inevitably processing of personal data. However, the results of this work are not. A text-to-speech system is normally disconnected from the audio recordings it was trained on, except as described in the following paragraph.
- Small corpus sizes. The model describes whatever is provided as training data, so if very little data is given, the model will also be characteristic of that data. A case in point is a text-to-speech system trained on a single speaker, synthesising speech with the voice of that speaker. The solution of using larger datasets may not be feasible due to unavailability of data, or even not desirable

[7] In the same vein, the CJEU found that the concept of 'controller' encompasses the administrator of a fan page hosted on a social network (C□210/16) [18]. In other words, the European Court said that Facebook (FB) and the administrator of a fan page are joint controllers. The concept of joint controllers are defined in the GDPR Article 26.

5. Conclusions

The paper focuses on linguistic experiments as a source of language resources (LRs). It addressed legal requirements regulating the collection and use of LRs. The focus of the paper is on personal data. In the first part of the paper, we discuss experimental data as a source for linguistic resources. Our own first-hand experience of using software intended for crowdsourcing experimental data showed that it is relatively complicated to get people to participate for an extended stretch of time and to complete the entire experiment.

Linguistic data is often personal data. The processing of PD is subject to the General Data Protection Regulation. When conducting linguistic research, it is recommended to remove personal data when possible (anonymize the data). Linguistic data which does not contain personal data has fewer restrictions on use and dissemination. It is essential to bear in mind that anonymization of personal data is considered the processing of personal data and subject to legal requirements [17]. Furthermore, the processing of language resources containing personal data must have a legal ground (consent or research done for legitimate or public interest). The GDPR provides several conditions (freely given, specific, informed, explicit) for consent. Subjects can withdraw the consent at any time [1]. Language research can also be conducted without the consent of the data subject if certain conditions are met (the data subject's rights are respected, safeguards are in place and so forth). Finally, we stress that the responsibility for the processing of personal data is shared between the researcher and the crowdsourcing platform.

References

[1] General Data Protection Regulation (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. 2016. OJ L 119. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1526898974837&uri=CELEX:32016R0679.

[2] A. Kelli, K. Vider, I. Kull, T. Siil, K. Lindén, A. Tavast, A. Värv, C. Ginter, and E. Meister, Keeleressursside Loomise ja Kasutamisega Seonduvaid Isikuandmete Kaitse Küsimusi (Data Protection Issues Relating to the Development and Utilisation of Language Resources), *Eesti Rakenduslingvistika Ühingu Aastaraamat* **14** (2018), 77–94.

[3] A. Kelli, K. Vider, K. Lindén, The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure, In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland* (2015), 13-24. http://www.ep.liu.se/ecp/article.asp?issue=123&article=002.

[4] A. Tavast, H. Pisuke, and A. Kelli. Õiguslikud Väljakutsed Ja Võimalikud Lahendused Keeleressursside Arendamisel (Legal Challenges and Possible Solutions in Developing Language Resources), *Eesti Rakenduslingvistika Ühingu Aastaraamat* **9** (2013), 317–32.

[5] Autoriõiguse seadus (Estonian Copyright Act). Entry into force 12.12.1992. English translation, https://www.riigiteataja.ee/en/eli/519062017005/consolide.

[6] D. A. Balota, M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, The English Lexicon Project, *Behavior Research Methods* **39(3)** (2007): 445–459.

[7] E. Keuleers, P. Lacey, K. Rastle, and M. Brysbaert, The British Lexicon Project: Lexical Decision Data for 28,730 Monosyllabic and Disyllabic English Words, *Behavior Research Methods* **44(1)** (2011), 287–304. https://doi.org/10.3758/s13428-011-0118-4.

[8] R. Munro, S. Bethard, V. Kuperman, V. Tzuyin Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily, Crowdsourcing and Language Studies: The New Generation of Linguistic Data, In *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (2010), 122.

[9] E. Keuleers, M. Stevens, P. Mandera, and M. Brysbaert, Word Knowledge in the Crowd: Measuring Vocabulary Size and Word Prevalence in a Massive Online Experiment, *The Quarterly Journal of Experimental Psychology* **68(8)** (2015), 1665–92. https://doi.org/10.1080/17470218.2015.1022560.

[10] S. Tratz, S., and E. Hovy, A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (2010), 678-687.

[11] S. Novotney, and C. Callison-Burch, C. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010), 207-215.

[12] A. Kelli, K. Vider, H. Pisuke, and T. Siil, Constitutional Values as a Basis for the Limitation of Copyright within the Context of Digitalization of the Estonian Language, In *Collection of Research Papers in Conjunction with the 6th International Scientific Conference of the Faculty of Law of the University of Latvia, Constitutional Values in Contemporary Legal Space II 16–17 November, 2016* (2017), 126–139. https://www.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Book_Juristu_6_konf_II-dalja.pdf#page=126 (27.7.2018)

[13] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive). OJ L 281, 23.11.1995, pp. 31-50, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&qid=1532332454516&from=EN

[14] Article 29 Working Party (WP29). Opinion 4/2007 on the concept of personal data. Adopted on 20th June, https://www.clinicalstudydatarequest.com/Documents/Privacy-European-guidance.pdf

[15] M. Oostveen, Identifiability and the applicability of data protection to big data, *International Data Privacy Law* **6(4)** (2016), 299-309.

[16] Court of Justice of the European Union (CJEU). YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel vs M and S (Cases C 141/12 and C 372/12). 17 July 2014, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1532759952949&uri=CELEX:62012CJ0141

[17] Article 29 Working Party (WP29). Opinion 05/2014 on Anonymisation Techniques. Adopted on 10 April 2014, https://www.scribd.com/document/328357978/WP29-Anonymisation

[18] Court of Justice of the European Union (CJEU). Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein vs Wirtschaftsakademie Schleswig-Holstein GmbH (C□210/16). 5 June 2018, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1532850897519&uri=CELEX:62016CJ0210.