The Aston Forensic Linguistic Databank (FoLD)

Marton Petyko, Lucia Busso, Tim Grant & Sarah Atkins

Aston Institute for Forensic Linguistics, Aston University, UK

https://doi.org/10.21747/21833745/lanlaw/9_1a1

Abstract. *The Aston Forensic Linguistic Databank (FoLD) is a permanent, controlled access online repository for forensic linguistic data. We broadly understand forensic linguistics as any academic research with a potential to improve the delivery of justice through the analysis of language. FoLD thus comprises a wide range of datasets with relevance to forensic linguistics and language and law, including commercial extortion letters, investigative interviews in police and other contexts, legal documents, forum posts from far-right online groups, and comment threads from political blogs. This paper outlines how FoLD works and its potential impact on the general discipline of forensic linguistics.*

Keywords: *Databank, Repository, Resource, Data sharing, Data access.*

Resumo. *O Aston Forensic Linguistic Databank (FoLD) é um repositório online de dados de linguística forense, de acesso restrito. Entendemos linguística forense no seu sentido lato, em que toda a investigação académica possui potencial para melhorar a administração de justiça através da análise da linguagem. O FoLD inclui, assim, uma série de "datasets" relevantes para a linguística forense linguagem e direito, incluindo cartas comerciais de extorção, entrevistas policiais e noutros contextos, documentos legais, publicações de grupos de extrema-direita em fóruns online e comentários em blogs políticos. Este artigo apresenta o funcionamento do FoLD e discute o seu potencial impacto na área da linguística forense em geral.*

Palavras-chave: *Banco de dados, Repositório, Recurso, Partilha de dados, Acesso a dados.*

Introduction – What is FoLD?

This paper introduces the Aston Forensic Linguistic Databank (FoLD), and explores the academic rationale and decision-making behind its provision and design. FoLD is a permanent, controlled access online repository for forensic linguistic data (available at fold.aston.ac.uk). FoLD has been developed in the Aston Institute for Forensic Linguistics (AIFL)[1] at Aston University, Birmingham, UK since 2019[2] and it was officially launched at the 15th Conference of the International Association of Forensic Linguists (IAFL) in

September 2021. FoLD is an innovative resource that makes it easier for researchers around the world to access a variety of forensic linguistic datasets. Furthermore, FoLD also accepts data contributions from researchers external from Aston, with the aim of populating a wide and diversified collection of forensic linguistic datasets for the advancement of the discipline.

In this paper we will outline how FoLD works and its potential impact on the general discipline of forensic linguistics. Specifically, the second looks at why openly available forensic linguistic datasets are scarce, why more forensic linguistic data is needed and how FoLD can contribute to addressing this challenge. The third section focuses on existing data sharing practices in corpus linguistics and computational linguistics, two areas that have continuing methodological impact on forensic linguistics. The fourth section has a particular focus on the policies and procedures that inform data submission and data access. Finally, the last section discusses our future plans for FoLD.

Why is FoLD needed? – Access to data in forensic linguistics

In the broadest sense, forensic linguistics can be defined as the application of linguistic knowledge, theory, and methods to legal and criminal contexts (Perkins 2021: 68) with the aim to improve the delivery of justice through the analysis of language (MacLeod and Grant 2017: 173). Although it is evident that forensic linguistics can only fulfil this aim by conducting evidence-based, reliable, and replicable research, access to relevant forensic linguistic data has been notoriously challenging since the conception of the discipline in the 1960s (Larner 2018).

The need for specific datasets for forensic linguistic analysis was recognised in the first edition of the journal then known as *Forensic Linguistics*, which changed its name to *Speech Language and Law*. Coulthard (1994: 29) gives mention to the setting up of a corpus to be "whimsically entitled the Habeas Corpus, which will include suicide notes, threatening letters, transcriptions of threatening and obscene telephone calls, court transactions, witness statements and police interview records".

Unfortunately, this corpus was never built but the plan is striking in the breadth of genre under consideration for such a corpus. Within the call is perhaps the recognition that different types of forensic linguistic analysis have different corpus requirements. On the one hand, analysis of evidential texts, such as 'suicide notes' and 'threatening letters' is considered to be required perhaps to contribute to investigative forensic linguistics and/or providing language as evidence. On the other hand, provision of 'court transactions, witness statements and police interview records' might be used to address the study of linguistics in legal contexts and its institutional practices. Notwithstanding various efforts including databases of threats, such as the now defunct *CTARC* and *FTARC* databases (see e.g., Gales (2011)), and more recent efforts, such as *TextCrimes* (held at AIFL) and *ForensicLing.com* (an open-source collection of resources maintained by Tammy Gales), this early call for a need for such texts has never been fully met.

The urgency for provision of forensic texts has grown. Within investigative forensic linguistics, there is growing recognition (Grant 2022) for the need of validation corpora to meet the needs of, for example, the English and Welsh Forensic Science Regulator (2020), or the United States NAS: National Research Council of the [United States] National Academy of Sciences' Committee on Identifying the Needs of the Forensic Sciences Community (2009) / PCAST: President's Council of Advisors on Science and

Technology (2016) critiques of forensic science. With growing recognition of the challenge of cross genre authorship analysis (Litvinova *et al.* 2018) it may also be that validation is required across different genres of forensic texts as used in evidence. Furthermore, with increased interest in academic forensic linguistic analysis beyond investigative casework, there is commensurate need for spoken and written texts from within the legal system, to enable description and analysis in areas of language and law and critical linguistic approaches to forensic and legal interactions.

The need for more forensic linguistic data is therefore clear. Yet, for a variety of reasons, data sharing is not a regular practice in forensic linguistics, which is a problem that hinders the entire discipline. The scarcity of freely available forensic linguistic data is due to at least three, partly related issues. First, forensic linguistic data often comes from highly institutionalised external parties, such as courts or law enforcement agencies, who must adhere to strict requirements on data protection. Building a working relationship with these external partners and going through the administrative processes necessary to transfer and share data takes a large amount of time and effort, meaning that data collection in forensic linguistics is perhaps more challenging and time-consuming than in many other fields of linguistics. In addition, these external partners can normally only provide the research data with legally binding restrictions on how it can be used, which often prevents researchers from sharing the data with others.

Second, many forensic linguistic datasets, such as dark web fora dedicated to the illegal exchange of child sexual exploitation and abuse material, threatening communications, extortion letters, hate speech corpora, police interviews, and courtroom discourse, contain highly sensitive or disturbing data, which poses an important ethical dilemma for researchers. Whilst data-sharing is seen as good academic practice, publishing the datasets, such as the abovementioned ones, without any restrictions on access runs the risk of making data subjects from these highly sensitive contexts potentially identifiable or may cause serious psychological or other harms to researchers. For these reasons, forensic linguists understandably often abandon the idea of publishing these datasets altogether.

Finally, open access data sharing platforms are simply not tailored to the discipline-specific needs of forensic linguists. Due to the challenging nature of many forensic linguistic datasets, a blanket open-access approach to data sharing in forensic linguistics is unrealistic. Researchers are understandably reluctant to publish their datasets if they are unable to control who gains access to their data and how it is used, but existing standard platforms, such as university repositories or the UK Data Service, do not always have the capacity and discipline-specific expertise to develop the policies and procedures needed for providing controlled access to highly sensitive datasets.

The main purpose of FoLD is to address the challenge posed by the scarcity of available and reusable forensic linguistic datasets by providing a permanent, controlled access platform for sharing all kinds of forensic linguistic data. The uniqueness of FoLD originates from the fact that it is developed and maintained by the Aston Institute for Forensic Linguistics, which has the capacity and discipline-specific expertise to tailor the repository to the needs of the wider forensic linguistic community.

Why not use an existing platform? – Data sharing in corpus linguistics and computational linguistics

Research data, as primary source material that supports academic enquiry and technical analysis, are the foundation of most linguistic projects. The digital age has seen growth not only in the amount of data available to researchers, but also in the infrastructure for sharing data between researchers, with the drive for 'open science' meaning that many studies now make the data that underpins their work available to others.

Since both corpus linguistics and computational linguistics are concerned with the computer-assisted analysis of large and structured collections of electronically stored texts (McEnery and Hardie 2011: 2), the importance of data sharing has long been emphasised in these fields (Ädel 2020: 16). Practices of data sharing developed in corpus linguistics and computational linguistics are potentially relevant to forensic linguistics since these areas have had a continuing methodological impact on forensic linguistics (Coulthard 1994; Cotterill 2010; Wright 2020).

The practice of publishing corpora is considered good academic practice in these fields for at least three reasons. Firstly, the accuracy and reliability of any empirical study can only be scrutinised by other researchers if they can access the data that the findings are based upon (Resnik and Shamoo 2017). This drive for 'open science' and good practice in data sharing is a key reason that many universities, funders and academic journals now expect or at least encourage researchers to make their data freely available (Corti *et al.* 2019). Secondly, building a corpus (or indeed any dataset) takes time and effort, but only published corpora can be reused by others for their own research projects. Finally, some corpus techniques, such as keyword analysis, require large reference corpora, which are often unfeasible to compile for the purposes of a single research project (Pojanapunya and Watson Todd 2018).

Several online platforms have been developed over the years to publish corpora and other datasets (Mieskes 2017). For example, both the *British National Corpus* and the *National Corpus of Contemporary Welsh* have their own web-based interface called *BNCweb* and *CorCenCC online*, respectively, while the web-based corpus analysis system *CQPweb* provides access to more than a hundred corpora. Some other well-known and widely used online platforms for accessing corpora from multiple languages include *Sketch Engine* (Kilgarriff *et al.* 2004), the *Oxford Text Archive*, and *CLARIN*, just to name a few. These platforms often include not only language resources but also software tools for the preparation, collection, and management of corpora.

In the field of computational linguistics, instead, data sharing is often practiced by individual scholars – rather than at the institutional level. Computational linguists who want to make their code or datasets available mostly gravitate toward the use of online public platforms, such as the *Open Science Framework (OSF)* (Foster and Deardorff 2017) or *GitHub* (https://github.com/). However, it is also increasingly common for individual researchers to build their own interfaces and repositories to hold and share specific types of data, such as corpora, datasets, or pipelines. In turn, the presence of this incredible amount of data available online generates unprecedented large-scale linguistic research, and the creation of new platforms and databases (for a discussion see inter alia Bender and Good (2010); Forkel *et al.* (2018)).

Creating new, publicly available resources is an academic endeavour and as such researchers publish peer-reviewed papers to present the new resources to the academic world. For example, the journal *Language Resources and Evaluation* is "the first publication devoted to the acquisition, creation, annotation, and use of language resources, together with methods for evaluation of resources, technologies, and applications." (journal website). In parallel, the biennial conference *LREC* (Language Resources and Evaluation Conference) reunites scholars working in language technologies to share and present advancements in the field.

Academic publications on computational linguistic resources tend to fall into one of two broad categories. Most commonly, resources are built with the specific aim to answer a theoretically motivated research question, or at the very least are grounded in a specific theory/framework. Hence, resulting publications mainly deal with the theoretical implications, analysis tools provided, and case studies (see among many others for example Pustejovsky *et al.* (2017); Petruck (2018); Brunato *et al.* (2020). However, there is also another type of publication for available datasets and resources. This second line of literature describes more in detail the architecture of the resource itself, providing flowcharts and more technical details on the process of data sharing per se. In other words, these publications describe the pipeline, dataset or website and its internal workings and mechanisms (see among others Reichel *et al.* (2016); Shu *et al.* (2020)).

The sheer existence of a wide range of existing data sharing platforms raises the question why forensic linguists would need their own platform in the first place? Whilst developing a new platform arguably requires much more resource than using an existing one, we argue that FoLD is needed because several challenges affecting data sharing in forensic linguistics have not been resolved in corpus linguistics or computational linguistics.

One of the most frequently mentioned challenges is how to deal with sensitive data in linguistic datasets and corpora (Rock 2001; Anthony 2013; Leedham *et al.* 2021). In the social sciences generally, and particularly perhaps in linguistics, commonly used data, such as written texts, recordings of interactions and interviews, nearly always originate from individuals, which becomes particularly problematic when using data from sensitive contexts. A handful of qualitative studies have met these challenges for data-sharing through explicitly consenting data subjects for the use of their data and outlining the level of anonymity that will be achieved, but such methods are not always appropriate or feasible in forensic and legal contexts, particularly when working with secondary data from external organisations, and there are no direct research 'participants' in the traditional sense.

In the context of corpus linguistics, sensitive data almost exclusively refers to personal information of named individuals, such as names, addresses, phone numbers and other contact details (Leedham *et al.* 2021). The standard method for mitigating this problem is anonymisation, i.e., replacing personal information with standard placeholders in a corpus (Rock 2001). Anonymisation is widely used in corpus linguistics, especially in published language resources but it also presents its own challenges. Due to the sheer size of many present-day corpora, manual anonymisation or the manual inspection of the output of automated anonymisation tools has become unfeasible, which means that it is practically impossible to ensure that all personal information has been removed from published corpora (Baker 2018).

An equally important but less often discussed challenge is how to publish resources that contain distressing or otherwise sensitive language with the potential to cause harm to the researcher or others (de Maiti and Fiser 2021). When facing this challenge, researchers generally follow one of three strategies. The first strategy involves the complete avoidance of publishing corpora containing distressing language. This practice of course minimises the risk of causing harm, but it also prevents other researchers from reusing a potentially valuable resource for their own projects. The second strategy is to remove all potentially distressing language from the corpus and share a heavily redacted version of the dataset with others. This approach again minimises the risk of causing harm at the expense of the integrity and academic usefulness of the corpus. Finally, rather than publishing the corpus as an open-access resource, researchers can provide others with controlled, often heavily restricted access to the dataset they have compiled. This option, however, requires substantial policy and procedure development as the data owner needs to define who can gain access to the corpus and under what circumstances. Policy development is a challenging endeavour in its own right, which is probably one of the reasons why the above-mentioned online corpus platforms do not provide an effective environment for sharing sensitive or potentially distressing language resources with others. This in turn means that researchers often stay on the safe side by not publishing their corpora in any form if they feel there is potential risk involved.

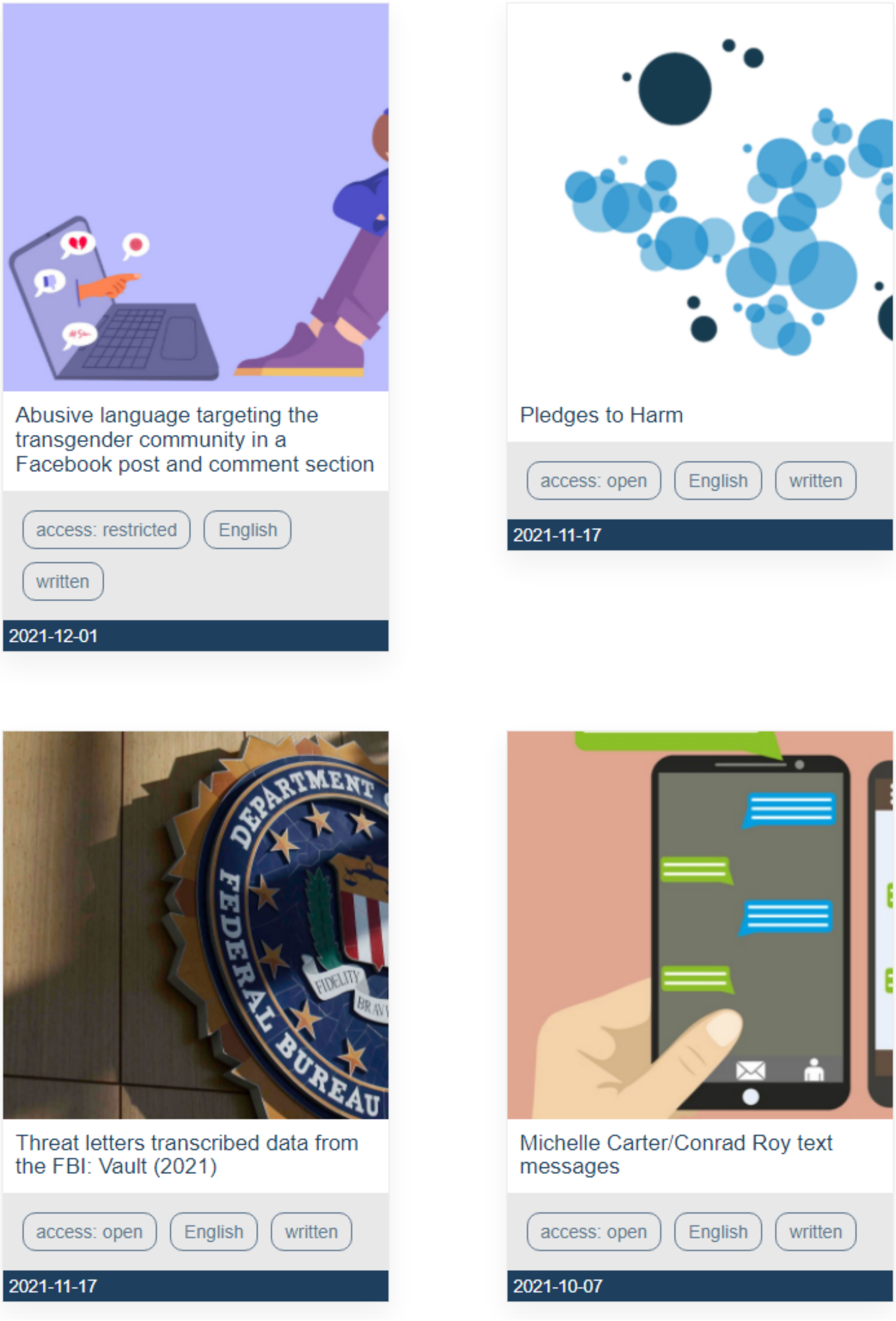
How does FoLD work?

FoLD as an online platform

FoLD (fold.aston.ac.uk) is a simple and easy-to-use online repository that provides access to forensic linguistic datasets (Figure 1). At AIFL, we broadly understand forensic linguistics as any academic research with the potential to improve the delivery of justice through the analysis of language (see Gibbons and Turell (2008); Coulthard *et al.* (2011); Rock (2011)). FoLD thus comprises a wide range of datasets with relevance to the "the application of linguistic knowledge and theory to forensic, legal, or criminal contexts" (Perkins 2021: 68).

Each dataset is represented as a separate tile and has its own dataset page on the website (Figure 2). Dataset pages provide detailed information about every dataset FoLD holds, including a title, a 50-word summary, a 50–500-word detailed description focusing on the content, structure and collection methods of the dataset, up to 5 subject keywords, data collection start and end dates, the language(s) of the dataset, data type(s) (written, spoken-audio, spoken-video, spoken-transcript, and other), access category (open, restricted, controlled, and external; see subsection ***Access categories*** for details), publication license, name and affiliation of the data donor(s), funding information, the AIFL research centre (e.g., Centre for Forensic Text Analysis) that the dataset has most relevance to, and the date of upload. We are also planning to provide every dataset with a Digital Object Identifier (DOI) and a standard citation so that they can be properly referenced in academic publications.

As explained in more detail in subsection ***Access categories***, the access category of a dataset determines whether or not the data files themselves are directly available from the dataset page. Data files for open datasets can be downloaded directly from the



Abusive language targeting the transgender community in a Facebook post and comment section

access: restricted English written

2021-12-01

Pledges to Harm

access: open English written

2021-11-17

Threat letters transcribed data from the FBI: Vault (2021)

access: open English written

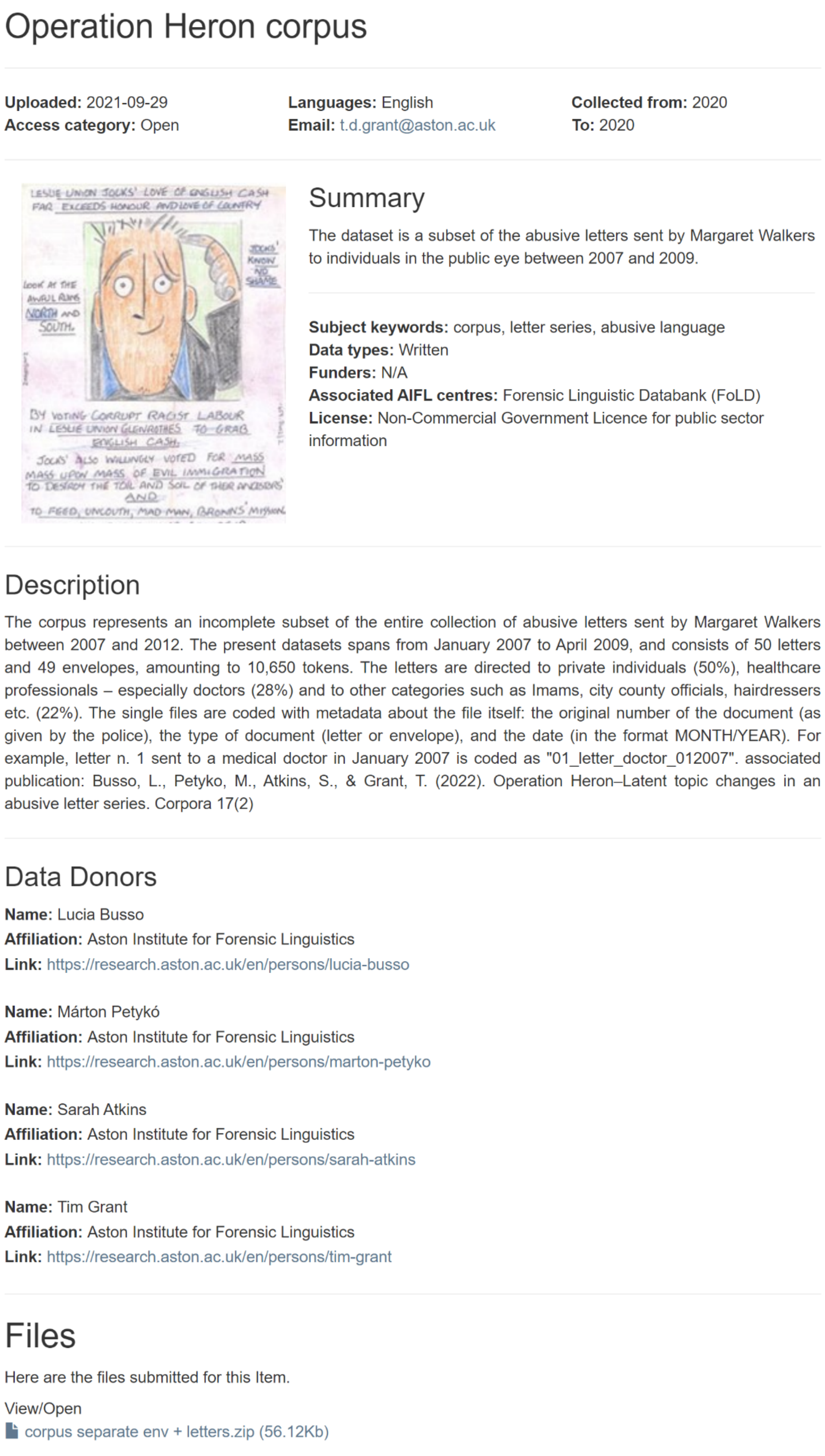
2021-11-17

Michelle Carter/Conrad Roy text messages

access: open English written

2021-10-07

Figure 1. Dataset tiles on FoLD


LESLIE UNION JOCKS' LOVE OF ENGLISH CASH
FAR EXCEEDS HONOUR AND LOVE OF COUNTRY
LOOK AT THE
AWAILING
NORTH AND
SOUTH.
JOCKS'
KNOW
NO
SHAME
BY VOTING CORRUPT RACIST LABOUR
IN LESLIE UNION GLENROTHES TO GRAB
ENGLISH CASH.
JOCKS' ALSO WILLINGLY VOTED FOR MASS
MASS UPON MASS OF EVIL IMMIGRATION
TO DESTROY THE TOIL AND SOIL OF THEIR ANCESTORS'
AND
TO FEED, UNCOUTH, MAD MAN, BROWN'S MISSION.

Operation Heron corpus

Uploaded: 2021-09-29
Access category: Open

Languages: English
Email: t.d.grant@aston.ac.uk

Collected from: 2020
To: 2020

Summary

The dataset is a subset of the abusive letters sent by Margaret Walkers to individuals in the public eye between 2007 and 2009.

Subject keywords: corpus, letter series, abusive language
Data types: Written
Funders: N/A
Associated AIFL centres: Forensic Linguistic Databank (FoLD)
License: Non-Commercial Government Licence for public sector information

Description

The corpus represents an incomplete subset of the entire collection of abusive letters sent by Margaret Walkers between 2007 and 2012. The present datasets spans from January 2007 to April 2009, and consists of 50 letters and 49 envelopes, amounting to 10,650 tokens. The letters are directed to private individuals (50%), healthcare professionals – especially doctors (28%) and to other categories such as Imams, city county officials, hairdressers etc. (22%). The single files are coded with metadata about the file itself: the original number of the document (as given by the police), the type of document (letter or envelope), and the date (in the format MONTH/YEAR). For example, letter n. 1 sent to a medical doctor in January 2007 is coded as "01_letter_doctor_012007". associated publication: Busso, L., Petyko, M., Atkins, S., & Grant, T. (2022). Operation Heron–Latent topic changes in an abusive letter series. Corpora 17(2)

Data Donors

Name: Lucia Busso
Affiliation: Aston Institute for Forensic Linguistics
Link: https://research.aston.ac.uk/en/persons/lucia-busso

Name: Márton Petykó
Affiliation: Aston Institute for Forensic Linguistics
Link: https://research.aston.ac.uk/en/persons/marton-petyko

Name: Sarah Atkins
Affiliation: Aston Institute for Forensic Linguistics
Link: https://research.aston.ac.uk/en/persons/sarah-atkins

Name: Tim Grant
Affiliation: Aston Institute for Forensic Linguistics
Link: https://research.aston.ac.uk/en/persons/tim-grant

Files

Here are the files submitted for this Item.

View/Open
corpus separate env + letters.zip (56.12Kb)

Figure 2. FoLD dataset page

dataset page. In contrast, dataset pages for restricted or controlled datasets only provide information on how users can gain access to the data files by submitting a data access request (see subsection ***Data requests*** for details). In addition, dataset pages for external datasets provide a link to the external website where users can access the data.

Finally, there are ethical considerations for a repository, such as FoLD, that go beyond the potential harm to data subjects – but also the potential for material to cause distress to researchers and users of the site. A great deal of data used in forensic linguistics is likely to contain content that can be disturbing to those that view it. Responsibilities to provide content warnings for such material has been well argued in teaching contexts (e.g., Stringer 2016) and there is a case for providing similar warnings in a repository, so that users have a choice before downloading material.

Access categories

Every dataset in FoLD falls into one of four access categories: open, restricted, controlled, or external. When reviewing newly submitted datasets, the FoLD editorial team makes a decision as to whether the dataset is in scope for FoLD, whether copyright may prevent publications and the most suitable access category for each dataset. Where necessary, for example, if the dataset contains sensitive material, the editorial team can also seek advice from the Aston Institute for Forensic Linguistics Research Ethics Committee. Data donors have the option to select a provisional access category for their own dataset and explain why they think the selected category is the most appropriate, but this category needs to be ultimately approved through the editorial process before the dataset gets published.

The access category of a dataset has significant implications on how the dataset is stored and how researchers can use it.

Open datasets are stored on the FoLD web server, are freely available for download from the FoLD website, and users do not need to be registered on the website or submit a data access request to gain access to these datasets. As a result, researchers can use any software tools of their choice to analyse open datasets. FoLD holds various open datasets reflecting a variety of research interests within and outside the Aston Institute of Forensic Linguistics (AIFL). For example, corpora of legal language (Busso forthcoming; Makouar 2021), datasets of trolling in comment threads (Petykó 2019), police interrogations (Szczekulska and Haworth 2021), or data from old casework such as the *Operation Heron Corpus* (Busso *et al.* forthcoming). FoLD also holds the *Drill Rap Slang Glossary* and the *Youth Slang Glossary* compiled by Professor Tony Thorne (Thorne 2014a,b).

Given the profound importance of data sharing in academia in general and in forensic linguistics in particular, we aim to make as many datasets openly available as possible. However, providing full and unrestricted access to some datasets is not possible because of ethical concerns, copyright or license issues, or constraints established by a Data Sharing Agreement. This is why we have introduced the restricted and controlled access categories.

Similarly to datasets in the open category, restricted datasets are also stored on the FoLD web server. However, they are not immediately available for download, but require some checks to be made on the user's research purpose before being made available for use. Users must register with a FoLD account, submit a data access request describing their research purpose, and agree and sign up to the terms of use (which may vary according to different conditions under donors Data Sharing Agreements, or under European and UK law Data Sharing Impact Assessments). Once approved, users will be able to download and use the dataset from the FoLD website. In line with these arrangements, users who gain access to restricted datasets are required to adhere to the terms of use when conducting their research project and publishing the results. Restricted datasets currently held on FoLD (as of March 2022) include data from old casework, such as the Amanda Birks' case text messages (Grant 2012), or the 100 Idiolects Dataset, a multichannel corpus of 100+ English speakers used for research on cross-genre idiolect (Heini *et al.* 2021).

Datasets with controlled access contain highly sensitive material that may come from a third party and have even heavier constraints on access and use. Controlled datasets are therefore stored not on the FoLD web server but on an air-gapped, offline computer in our secure data lab at the Aston Institute for Forensic Linguistics. Users who wish to access these datasets must make a detailed application to FoLD and the data owner, as well as potentially gain additional agreement from an external organisation before they can be approved for access. Although information on controlled datasets is detailed in the FoLD repository for users to search, the data itself is not available for download and users may need to visit Aston or agree a secure means of access. Datasets in this category include, for example, scraped data from white supremacist and dark-web child abuse discussion fora (Kredens and Pezik 2021b,a).

Finally, we also use FoLD to signpost the availability of external datasets that are managed by third parties and are not directly available via the FoLD website. These datasets might be open-access or have various access restrictions at the data owner's discretion. External datasets include for example other data repositories, such as the *ForensicLing.com* website managed by Tammy Gales of Hofstra University.

Data submission

As of December 2021, FoLD holds around 20 datasets which have been donated to us by AIFL academic and research staff, PhD and Masters students, and external researchers. FoLD is completely unrestrictive in the sense that anyone who is willing to share their forensic linguistic datasets via FoLD can become a data donor. Given that our aim is to publish as many datasets as possible, we are constantly looking for new relevant datasets, and we actively encourage potential data donors from within and outside AIFL to contact us at aifl_fold@aston.ac.uk to discuss the suitability of their datasets for inclusion into FoLD.

Publishing a dataset via FoLD has several benefits for data donors. Firstly, data donors retain full ownership of their datasets and can withdraw them from the databank at any time if they wish to do so. We never change the content, structure, or access category of any dataset we hold on FoLD without the data donor's explicit permission. At the same time, we provide a permanent, controlled access platform for the datasets and data donors can work with us to set the access category, publication license and any limitations on use. We also help data donors with the everyday practicalities of storing their datasets online. For example, data donors can regularly update their datasets while we ensure that the integrity of the datasets remains intact.

Secondly, FoLD provides online visibility for the datasets donated to us. We work actively to promote FoLD across the global forensic linguistics community to ensure that our data donors' work reaches those interested, thus maximising the impact potential of their datasets on research and teaching.

Finally, many funders require datasets collected for research projects to be deposited in a repository and made available for other researchers. As explained in section ***Why is FoLD needed? – Access to data in forensic linguistics***, historically this has been difficult in forensic linguistics because of the sensitive nature of much of the data it uses. FoLD aims to make data-sharing possible through managing different access categories, meaning that forensic linguists can meet the requirements of their funders' research data policy by using a platform that is tailored to their discipline-specific needs.

As mentioned above, we invite all potential data donors to have an informal discussion with the FoLD team about the suitability of their dataset for inclusion before they submit the dataset to the FoLD website. During this discussion, we establish whether the dataset holds relevance to forensic linguistics and whether there are any obvious ethical or licensing issues that would prevent us from publishing the dataset via FoLD.

Once we have established that the dataset is in principle publishable on FoLD, we ask the data donor to submit their dataset by completing an online form available on the FoLD website (http://fold.aston.ac.uk/static/documents/item_submission_form.pdf). Data donors are required to provide the following information about their datasets: a title, a 50-word summary, a 50–500-word detailed description outlining the content, structure and collection methods of the dataset, up to 5 subject keywords, data collection start and end dates, the language(s) of the dataset, data type(s) (written, spoken-audio, spoken-video, spoken-transcript, and other), access category (open, restricted, controlled, or external), publication license, name, affiliation and contact details of the data donor, funding information, and – if relevant – the associated AIFL research centre (e.g., Centre for Forensic Text Analysis) that the dataset belongs to.

Given that the access category is one of the key considerations that we make when publishing datasets, we also ask data donors to provide a rationale for their requested access category. If the data donor wishes to publish an open dataset with us, we ask them to upload the data files as well. For restricted and controlled datasets, on the other hand, we ask data donors to share their dataset with us for review via other means. Finally, we also ask all data donors to upload an image for their dataset (distributed with Creative Commons licence), which is used on the data tiles and dataset pages to make the datasets visually more recognisable.

Once the dataset has been submitted, we carry out our editorial review before making the data available on FoLD. During this process, we check the metadata and the data files for any inaccuracies and decide on the most suitable access category for the dataset under review. If we have any ethical concerns about the dataset that we are unable to resolve, we seek advice from the AIFL Research Ethics Committee, which works independently from the FoLD team. Whilst we normally follow the advice received from the Ethics Committee, all publication decisions lie with and are the responsibility of the FoLD team.

The ethical aspect of the editorial process is of great importance to a specialist repository such as FoLD. As stated above, sharing and publishing research data through a repository poses ethical and some legal challenges, with the potential for harm to individuals or communities should they become identifiable or if data were to be misused. These challenges prompted many early decisions by the FoLD team about the structure and management of the repository.

The gold standard for sharing data beyond a project is usually to consent research participants for the storage and reuse of their anonymised data (ESRC 2015). Where researchers at Aston University were conducting projects with participants, a description of this ongoing use was indeed provided in information sheets and consent forms, for example:

> *"During the project your data will be anonymised and will become part of a larger, anonymised dataset. At the end of the project this dataset will be made available in an open access repository (Forensic Linguistic Databank (FoLD)) in the Aston Institute for Forensic Linguistics. Once in the repository, your data will not be able to be re-identified and will be made available for reuse."*

However, it is acknowledged that such consent or anonymisation may not exist for pre-existing datasets or even be feasible for many types of research data and in these instances careful consideration and potentially a review by a research ethics committee is needed before a dataset is archived and reused (Summers *et al.* 2019). This issue has perhaps been particularly acute in forensic linguistic contexts, where data may not originally have been intended for research, might be sourced from criminal or sensitive contexts, or be provided by another organisation that requires strict limits on use – long held difficulties for sharing data in the field.

As a means of mitigating these difficulties, an early decision was made to provide datasets through layered access categories (detailed in subsection ***Access-categories***), whereby more complex datasets would not be published openly but could be restricted to authorised researchers, who apply with a particular purpose and can demonstrate ethical approval at their own research institution. This is a process mirrored by the UK Data Service for historic or sensitive datasets such as Dodds *et al.* (2017) (see case study 7.5 in Summers *et al.* (2019)).

Once our editorial review is complete, we inform the data donor about the outcome and we actively liaise with them to reach an agreement on the final version of the metadata, such as the access category and the wording of the title, summary, and description. As a principle, we only publish a dataset if and when the data donor and the FoLD team have managed to agree on the metadata. We also take precautions to ensure that we do not force our opinion about the nature of dataset on the data donor whilst we only accept datasets that we are comfortable to publish. Data donors retain full ownership of their datasets. Finally, data donors can submit new versions of their already accepted datasets. These new versions need to go through the same albeit occasionally simplified editorial process.

Data requests

As explained in subsections ***Data requests*** and ***Access categories***, users who wish to gain access to controlled or restricted datasets need to register on the website and submit an online data access request (http://fold.aston.ac.uk/static/documents/item_request\

_form.pdf). Although the data access request form that users need to complete is the same for both access categories, controlled and restricted datasets differ in that researchers are only allowed to work on controlled datasets on an air-gapped computer in our secure data lab at Aston University while restricted datasets are eventually released to the named requester upon approval.

When completing the data access request form, users need to provide various information about themselves and their research project. This includes the researcher's name, affiliation, academic position or student status, the details of a signatory from their host institution who has the authority to agree to and sign an access agreement on behalf of the requester's host institution, details about the requester's proposed research project with particular focus on the proposed use of the data, plans for publishing and disseminating findings of the project, and ethical approval from the requester's host institution.

Similarly to data submissions, all data access requests are reviewed by the FoLD team. As a team, it was agreed this review should purely assess a legitimate research purpose and not appraise the perceived quality or stance of the research methods. In case of ethical concerns about an application, the editorial team may consult the AIFL Research Ethics Committee and make a decision in line with the conditions or restrictions set by the data donor. When the dataset comes from an external organisation, said organisation might also need to approve access. After all enquiries have been satisfied, we inform the requester about the outcome. If access is granted by all parties, the researcher is then asked to sign an agreement, along with a signatory from either a supervisor (in the case of a student) or the research office at their institution, consenting to the terms of use for the data and any particular limitations in how it can be stored, used and published on. If we are unable to approve the initial request, we provide feedback to the researcher to help them make the necessary adjustments to their proposal and invite them to submit a new request. This practice is to ensure that users, especially students, are not penalised for previously submitted unsuccessful access requests.

Conclusions – What will happen to FoLD now?

Our future plans for FoLD are twofold. Firstly, the website is still under development. Currently, users can manually browse the available datasets, but we are working towards developing a more sophisticated search facility to ensure that users can easily find the datasets they are looking for. We are also in the process to provide all datasets with a Digital Object Identifier (DOI) and a standard citation so that they can be properly referenced in academic publications. In addition, we aim to streamline access to controlled datasets to enable researchers access these highly sensitive but extremely valuable datasets without any unnecessary inconvenience. As mentioned earlier, our utmost priority is to tailor FoLD to the needs of its potential users, including academics and students. In line with this aim, we encourage members of the forensic linguistic community to contact us with their ideas on how FoLD can be improved.

Secondly, whilst FoLD is an innovative data sharing platform that builds on the collective expertise of the AIFL members, it can only become a truly useful resource for forensic linguists and strengthen the discipline if it holds a wide range of relevant forensic linguistic datasets, enabling researchers to build on each other's work. We therefore encourage everyone who wish to share their forensic linguistic datasets with others to consider publishing them with us. After all, FoLD can only be as good as the datasets it holds.

Notes

[1] The Aston Institute for Forensic Linguistics (AIFL) was founded in 2019. It is a substantial expansion of the former Aston Centre for Forensic Linguistics that was founded in 2008. This expansion was funded via a £6M investment including a £5.4M award from Research England's Expanding Excellence in England (E3) fund.

[2] The FoLD online interface and server is managed by Aston University's student-led software enterprise *Beautiful Canoe* (https://beautifulcanoe.com/)

References

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141–161.

Baker, P. (2018). Language, sexuality and corpus linguistics. concerns and future directions. *Journal of Language and Sexuality*, 7(2), 263–279.

Bender, E. and Good, J. (2010). A grand challenge for linguistics: Scaling up and integrating models. In *White paper contributed to NSF's SBE 2020 initiative*. 1–5.

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G. and Montemagni, S. (2020). Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 7145–7151: European Language Resources Association.

Busso, L. (forthcoming). Coriells: A specialised bilingual corpus of lay legal communication. In C. Meluzzi and S. Cenceschi, Eds., *Proceedings of the workshop "Linguistica forense: dalla ricerca empirica alla pratica legale"*.

Busso, L., Petykó, M., Atkins, S. and Grant, T. (forthcoming). Operation heron – latent topic changes in an abusive letter series. *Corpora*, 17(2). http://publications.aston.ac.uk/id/eprint/42497/.

Corti, L., Eynden, V., Bishop, L. and Woollard, M. (2019). *Managing and sharing research data: a guide to good practice*. Sage.

Cotterill, J. (2010). How to use corpus linguistics in forensic linguistics. In A. O'Keeffe and M. McCarthy, Eds., *The Routledge Handbook of Corpus Linguistics*. Routledge, 578–590.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics*, 1(1), 27–43.

Coulthard, M., Grant, T. and Kredens, K. (2011). Forensic linguistics. In R. Wodak, B. Johnstone and P. Kerswill, Eds., *The SAGE handbook of sociolinguistics*. Sage, 529–544.

de Maiti, K. and Fiser, D. (2021). Working with socially unacceptable discourse online: Researchers' perspective on distressing data. In I. Heindrickx, L. Verheijen and L. Wijngaert, Eds., *Proceedings of the 8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)*, 78–82: Radboud University.

Dodds, C., Keogh, P. and Weatherburn, P. (2017). *Accessing HIV post-exposure prophylaxis: gay and bisexual men in the UK describe their experiences*. Data Collection. Colchester, Essex, United Kingdom: UK Data Service.

ESRC, (2015). Esrc framework for research ethics, economic and social research council. https://esrc.ukri.org/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/.

Forkel, R., List, J. and Greenhill, S. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205), 1–10.

Foster, E. and Deardorff, A. (2017). Open science framework (osf. *Journal of the Medical Library Association*, 105(2), 203–206.

Gales, T. (2011). Identifying interpersonal stance in threatening discourse: An appraisal analysis. *Discourse Studies*, 13(1), 27–46.

J. Gibbons and M. Turell, Eds. (2008). *Dimensions of Forensic Linguistics*. John Benjamins.

Grant, T. (2012). Txt 4n6: method, consistency, and distinctiveness in the analysis of sms text messages. *Journal of Law and Policy*, 21(2), 467–494.

Grant, T. (2022). *On the idea of progress in forensic authorship analysis*. Cambridge University Press.

Heini, A., Pezik, P. and Kredens, K. (2021). The 100 idiolects project. the aston forensic linguistic databank. http://fold.aston.ac.uk/handle/123456789/17.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004). The sketch engine. In G. Williams and S. Vessier, Eds., *Proceedings of the 11th EURALEX International Congress*, 105–116: Université de Bretagne-Sud.

Kredens, K. and Pezik, P. (2021a). Child sexual abuse dark-web discussion forum corpus. the aston forensic linguistic databank. http://fold.aston.ac.uk/handle/123456789/20.

Kredens, K. and Pezik, P. (2021b). White supremacist discussion forum corpus. the aston forensic linguistic databank. http://fold.aston.ac.uk/handle/123456789/19.

Larner, S. (2018). Forensic linguistics. In A. Phakiti, P. Costa, L. Plonsky and S. Starfield, Eds., *The Palgrave Handbook of Applied Linguistics Research Methodology*. Palgrave Macmillan, 703–718.

Leedham, M., Lillis, T. and Twiner, A. (2021). Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the xxx corpus. *Applied Corpus Linguistics*, 1(3), 1–10.

Litvinova, T., Seredin, P., Litvinova, O., Dankova, T. and Zagorovskaya, O. (2018). On the stability of some idiolectal features. In *International Conference on Speech and Computer*, 331–336: Springer.

MacLeod, N. and Grant, T. (2017). go on cam but dnt be dirty": linguistic levels of identity assumption in undercover online operations against child sex abusers. *Language and Law/Linguagem e Direito*, 4(2), 157–175.

Makouar, N. (2021). Cyberhate regulation in france: understanding the semantic negotiation and dynamics in parliamentary debates. In *Presentation at 15th Biennial Conference of the International Association of Forensic Linguists*, Birmingham, UK: Aston University.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics. Method, Theory and Practice*. Cambridge University Press.

Mieskes, M. (2017). A quantitative study of data in the nlp community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 23–29.

NAS: National Research Council of the [United States] National Academy of Sciences' Committee on Identifying the Needs of the Forensic Sciences Community, (2009). Strengthening forensic science in the united states: a path forward. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf.

PCAST: President's Council of Advisors on Science and Technology, (2016). Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Perkins, C. (2021). The application of forensic linguistics in cybercrime investigations. *Policing: A Journal of Policy and Practice*, 15(1), 68–78.

M. Petruck, Ed. (2018). *MetaNet*, volume 100. John Benjamins.

Petykó, M. (2019). *The discursive construction of trolling on British and Hungarian political blogs.* Lancaster University.

Pojanapunya, P. and Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 133–167.

Pustejovsky, J., Ide, N., Verhagen, M. and Suderman, K. (2017). Enhancing access to media collections and archives using computational linguistic tools. In *CDH@ TLT*. 19–28.

Reichel, U., Schiel, F., Kisler, T., Draxler, C. and Pörner, N. (2016). The BAS speech data repository. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 786–791.

Resnik, D. and Shamoo, A. (2017). Reproducibility and research integrity. *Accountability in Research*, 24(2), 116–123.

Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics*, 6(1), 1–26.

Rock, F. (2011). Forensic linguistics. In J. Simpson, Ed., *The Routledge Handbook of Applied Linguistics.* Routledge, 158–172.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3), 171–188.

Stringer, R. (2016). Reflection from the field. trigger warnings in university teaching. *Women's Studies Journal*, 30(2), 62–66.

Summers, S., Bishop, L., Eynden, V. and Corti, L. (2019). Legal and ethical considerations in sharing data. In L. Corti, V. Eynden, L. Bishop and M. Woollard, Eds., *Managing and Sharing Research Data.* Sage, 2nd ed.

Szczekulska, A. and Haworth, K. (2021). *US police-suspect interrogations.* http://fold.aston.ac.uk/handle/123456789/7: The Aston Forensic Linguistic Databank.

Thorne, T. (2014a). Dictionary of contemporary slang.

Thorne, T. (2014b). The new canting crew. In J. Coleman, Ed., *Global English Slang.* Routledge, 72–81.

Wright, D. (2020). Corpus approaches to forensic linguistics. applying corpus data and techniques in forensic contexts. In M. Coulthard, A. May and R. Sousa-Silva, Eds., *The Routledge Handbook of Forensic Linguistics.* Routledge, 2nd ed., 611–627.

Ädel, A. (2020). Corpus compilation. In M. Paquot and S. Gries, Eds., *A Practical Handbook of Corpus Linguistics.* Springer, 3–24.