International Journal of
Language & Law

Genovese, JLL 14 (2025): 30–62
www.languageandlaw.eu

Building the *Australian Legislative Corpus 2023*

— Combatting Issues and Highlighting Applications of General Legislative Corpora

Emma Genovese[*]

Abstract

This article introduces and details the construction of the Australian Legislative Corpus 2023 ('ALC23'). The ALC23 includes relevant legislation from each Australian jurisdiction that was in force as at 30 June 2023, and is intended to act as a general corpus of a specialised nature. The article begins by providing a brief overview of corpus linguistic applications in the legal sphere, before summarising current legal corpora. Following this, the article details the composition of the ALC23, before moving to discuss how issues in construction were overcome. The article also notes potential applications of the ALC23, including a case study of how the word "gender" is used in the ALC23. The article concludes with some potential limitations that may be of note to both corpus linguists and legal scholars. Crucially, this article is written from the perspective of a legal scholar, which means that the creation, and use, of the ALC23 is intended to be made accessible to scholars who have a limited background in linguistic theory, method, and programming.

Keywords

legal corpora, corpus linguistics, legal linguistics, law and language.

Submitted: 13.08.2024, accepted: 12.12.2024, published online: 21.02.2025

[*] *Emma Genovese*: University of Technology Sydney, Quentin Bryce Law Doctoral Scholar, emma.genovese@student.uts.edu.au

DOI: 10.14762/jll.2025.030

1. Introduction

The examination of the connections between law and language is a broad field that has been approached from a variety of different disciplines (Goźdź-Roszkowski, 2011: 13–14; Leung & Durant, 2018; Wagner & Matulewska, 2023). Prior to the 1970s, research that explored the connections between the law and language often occurred in a disjointed manner, with limited recourse provided to the linguistic field (Williams, 2005: 15). Since that period, scholarship exploring law and language has gained significant insights from linguists, allowing legal researchers to consider various related topics (Galdia, 2023; Williams, 2005: 15). Specifically, analysis has included anything from examining legal language generally (Conley & O'Barr, 2005; Hart, 1994; Mellinkoff, 1963; Salembier, 2018; Tiersma, 1999), legal discourse (Goodrich, 1987), language used in courtrooms (Berk-Seligson, 2012; Danet, 1980; Okawara, 2012; Stygall, 2012; Woodbury, 1984), and legal language's connection with a variety of fields, such as anthropology, sociology, and philosophy (Freeman & Smith, 2013: 4–6). Ultimately, the exploration of law and language has revealed many insights into the importance of language use in the legal sphere.

A prominent sub-field within linguistics that has contributed to the analysis of legal language is corpus linguistics. Briefly, corpus linguistics is a methodology and area of study situated within the broader sphere of linguistics (Tognini-Bonelli, 2001: 1). The approach essentially aims to uncover widespread patterns within language, in order to conduct subsequent analysis and interpretation (Baker, 2004: 346; 2005: 5). In relation to law and language, corpus linguistic "techniques enable certain reading practices over a large data set", (Lukin & Marrugo, 2023: 227) which means that researchers are able to efficiently engage with sizeable texts that were previously too expansive to meaningfully analyse. Corpus linguistics works with corpora, which is a "body of naturally occurring language" (McEnery et al., 2010: 4), that can include thousands to billions of words (Baker, 2005: 5; Baker & McEnery, 2015: 1), and can be made up of various different texts or text types (Baker, 2014: 7–8; McEnery et al., 2010: 4). While the creation of corpora depends primarily upon the intended research question/s and outcome/s (Baker, 2006: 26; McEnery et al., 2010: 18), scholars agree that corpora generally consist of several key features (McEnery et al., 2010: 5). These features include the corpora being: machine-readable, made of an authentic sample of texts, and being representative of a particular variety or "register" of language (Gillings et al., 2023: 8; McEnery et al., 2010: 5). These features ensure that a corpus represents language in use, with the text able to be processed using automated software, rather than requiring manual analysis.

However, the development of accessible legal corpora is somewhat impaired for several issues that will be considered below. Especially relevant to my exploration is that while some legal corpora exist, many do not focus exclusively on legislation. Generally, this is an impediment for researchers who wish to explore how language is used within the particular domain of legislation, rather than across different legal registers like case

DOI: 10.14762/jll.2025.030

law, legal scholarship, and legislation. Legislation is particularly important within the context of legal discourse because it is arguably the foundation in which all legal language relates to or is derived from. For instance, case law can interpret the language of legislation or apply it within context, and legal scholarship can analyse the language of certain provisions or legislation. Additionally, in relation to discourse and the connection between the construction of reality, law provides "reality constructions and world views and their underlying claims to power with legitimacy and, if necessary, makes them compulsory enforceable" (Stückler, 2018: 113). While there is utility in assessing legislation alongside other uses of legal language, I intend to demonstrate that there is much that can be gained, both practically and theoretically, from legal corpora that comprises solely of legislation.

For my research specifically, I set out to explore how sex, gender, and sexuality are constructed and interpreted in Australian legislation. However, the first issue with this exploration was that a corpus of Australian legislation did not exist, meaning that my first step was to build this corpus. However, the second, and more pressing issue, was that I did not have a background in corpus linguistics or computer programming, meaning: the building of this corpus had to occur using the most accessible means possible. This issue tends to be the case for many scholars who are located primarily in law faculties; therefore, if more legal research is to be produced using corpus linguistic methods, the building and processing of corpora must be comprehensible and accessible to legal scholars.

Accordingly, in this article, I seek to contribute to the field of corpus linguistic applications to the law in a number of key ways. First, I aim to provide instructions for legal scholars on how to create legislative corpora without relying on an understanding of coding software, like Python. I provide this perspective through introducing how I built the Australian Legislative Corpus 2023 ('ALC23'). This method is useful not only for legal scholars who also do not have any previous experience in computing, but for scholars who do. That is, the method can provide a means to address some restrictions that may be faced when building legislative corpora more generally. Second, I offer some diagnostic observations: throughout this manual-based process, there were several issues I uncovered which I consider also contribute to the lack of legislative corpora. In overcoming these issues, I outline some associated benefits that have arisen from taking an alternative approach to corpus building. Third, I highlight some potential applications of the ALC23, with a particular focus on statistical processing in corpus linguistic software. This application is supported by an explanation of the use of 'gender' in the ALC23, which is intended to demonstrate the advantages of engaging with legislative corpora, along with identifying some potential limitations of the ALC23 in particular.

2. Corpus Linguistic Methods in Law

A burgeoning area within law and language includes the application of corpus linguistic techniques, methods, or approaches that engage in or assist with the analysis of legal language. Corpus linguistics began to develop around the early 1980s, and while it was initially used solely by linguists (McEnery et al., 2010: 3; Stubbs, 1996), it has more recently been extended into the legal domain. While there is some disagreement as to what may be appropriately classed as corpus linguistic applications to the law (Goźdź-Roszkowski, 2011), I believe it is useful to take a broad interpretation so as to encourage the various contributions that corpus method can add to the analysis of the law. Accordingly, I consider that the sub-field involving applications of corpus linguistics to the law could include anything from corpus-assisted legal linguistics, "Law and Corpus Linguistics" (Goldfarb, 2021), computer-assisted legal linguistics (Vogel et al., 2018), legal corpus linguistics (Egbert & Römer-Barron, 2024; Gries, 2021), or any other application that is derived from corpus linguistics. In other words, while these applications may not always be perceived as falling within the technical domain of corpus linguistics, designation of this term is nonetheless useful for expanding methodological approaches used by legal scholars. Ultimately, the application of corpus linguistics to the legal field is inherently interdisciplinary in nature, but the complexity of linguistics has meant that explorations of the law involving corpus linguistic techniques often do not appeal to both corpus linguistic and legal scholars, or legal scholars in particular (Goźdź-Roszkowski, 2021: 1535). Accordingly, I consider that emphasising the varied applications of corpus method to the law creates scope for interpretations made by predominantly legal scholars, while also holding value for corpus linguists who wish to make their research relevant to legal researchers.

From this broad perspective, there have been reviews conducted relating to the application of corpus linguistic approaches to legal settings more generally (Goźdź-Roszkowski, 2023; Vogel et al., 2018). Such applications tend to occur across a variety of different areas, with anything from forensic linguistic contexts (Gillings, 2022; Woolls & Coulthard, 1998), courtroom engagements (Berūkštienė, 2018) or judgments (Bhatia et al., 2004: 212–213), to the provision of comparative legal history (Laske, 2022), the translation of legal texts (Hu et al., 2021; Pei & Li, 2018; Pontrandolfo, 2019), representational studies (Pérez-Paredes et al., 2017), or citations analysis (See Vogel et al. for an overview of this area: Vogel et al., 2018: 48–49). Practical applications of corpus linguistic techniques have also occurred, that of which is particularly prevalent in legal education. For instance, corpus methods have been used to assist with legal writing (Hafner & Candlin, 2007), teaching legal English more generally (Breeze, 2017: 6), or in teaching specific areas of law, such as contract law (Römer-Barron & Cunningham, 2024). Additionally, a significant application of corpus linguistic approaches has occurred with respect to the judiciary or scholars relying on these methods to assist in interpreting the language of legislation, legal materials, or cases (Egbert & Wood, 2023; Mouritsen, 2017;

Solan & Gales, 2017). For instance, the interpretation of the definition of certain legal terms is a particular focus, such as the definition or interpretation of 'war' (Lukin & Araujo E Castro, 2022; Lukin & García Marrugo, 2024), the use of lexical bundles in case law (Berūkštienė, 2018), or the analysis of the semantic structure of terms in a selection of cases (Vogel, 2017). There may also be benefits to legal practitioners or organisations involved in advocacy and legal reform, given the ability to search legislation across multiple jurisdictions, thereby increasing the potential to quickly identifying problematic provisions. For instance, by searching in a legislative corpus, the South Australian Law Reform Institute Report on discrimination in South Australia could have provided a comprehensive list of 'gendered language' that was recommended to be removed, rather than highlighting certain Acts (South Australian Law Reform Institute, 2015: 9, 23 para 30). Significantly, each of the above examples demonstrate that the use of corpus linguistics in the legal field has significant practical applications, particularly when it comes to identifying common features and structures of legal language (Breeze, 2017: 16), "semiotic patterns" (Lukin & Araujo E Castro, 2022: 2179), or the interpretability and social effect of terms within legal texts (Lukin & García Marrugo, 2024).

Essentially, a broad consideration of corpus linguistic methods in the law highlights that these methods have been used in a variety of different ways to achieve a variety of different outcomes. As such, this exploration assists in emphasising that the application of corpus method to explore legal language in general is a fast-developing field, with legal corpora able to be analysed from an entirely new lens that is only made possible by these methodological techniques. Crucially, the application of corpus linguistic method to the legal field results in insights that would not otherwise be uncovered, primarily because of the large data that is able to be analysed.

3. Legal Corpora

A central aim of the ALC23 is to contribute to the assortment of publicly accessible legal corpora, particularly as it relates to statutes. While scholars continue to build large legal corpora, many of these relate to case law, or include a combination of legal texts (Pontrandolfo, 2012). Crucially, there are minimal corpora available that focus exclusively on legislation, partly due to issues such as narrow corpora design (Vogel et al., 2018: 1351), or copyright restrictions. This means that any analysis that intends to explore legislative language is restricted to using corpora that was constructed with a particular focus in mind, or by conducting a time-consuming and limited manual exploration.

3.1. Case Law

Case law is the typical subject of legal corpora (Vogel et al., 2018: 1351), with collections spanning from historical to modern legal texts (Vogel et al., 2018: 1354). For instance, the House of Lords Judgments Corpus includes 188 judgments made in the House of Lords between 2001 and 2003, and was created to explore techniques of automatic summarisation of judgments (Grover et al., 2004). A related corpus is the British Law Report Corpus, which includes 16,612 judgments from 2008 to 2010 from Northern Ireland, England, Wales, and Scotland, which was intended to be used to assist in enhancing teaching materials and to conduct linguistic analysis (Rea Rizzo & Marín Pérez, 2012: 142). Similarly, the Cambridge Law Corpus includes more than 250,000 judgements across the United Kingdom, with cases spanning from the 16th to 21st century, and it was created to enhance legal research, including in relation to artificial intelligence (Östling et al., 2024).

3.2. Combined Texts

Several legal corpora include a combination of legal texts, often involving a multilingual component. For instance, the CAL^2 Corpus of German Law aims to be a reference corpus, rather than developed for a particular purpose, and it contains various legal texts, such as legislation, legal judgments, and articles (Vogel et al., 2018: 1355). The University of Turin's *Jus Jurium* is a corpus of Italian regulations, cases, and parliamentary reports (Onesti, 2011: 8). Related corpora of a multilingual nature include corpora such as the Bononia Legal Corpus, which is an ongoing project surrounding a legal corpus of both English and Italian legal texts, with the intention that this corpus is continually expanded (Rossini-Favretti, 1998: 57). Similarly, the Romanian Legal Corpus contains Romanian legal documents from 1881 to 2018, including a variety of legislation, and government orders and associated documents (Tufiș et al., 2020: 2774).

In the Australian context, the Open Australian Legal Corpus ('OALC') "is the first and only multijurisdictional open corpus of Australian legislative and judicial documents" (Butler, n.d.-b). The corpus includes primary and secondary legislation, bills, and certain case law from specific Australian jurisdictions (Butler, n.d.-a), and it is frequently updated to encapsulate amendments. However, the purpose of the creation of the dataset was to "train a large language model to solve legal problems" (Butler, 2023) – machine intelligence being a prominent area of research in law and language (See e.g., Hildebrandt, 2018: 27). Due to this purpose, the dataset is stored in a JSON lines file, and operates within the Python library, meaning coding knowledge is necessary to engage in any analysis. To a different extent, there also exists more highly specialised legal corpora that aims to reflect a specific legal area. For example, the Macquarie Laws of War Corpus is a corpus of international war law documents that was in part created to overcome the issues of the texts being made available individually online, which substantially limited the search function and any analysis, to non-corpus linguistic method (Lukin & Araujo E Castro, 2022: 2168). Additionally, the Hong Kong Learner Corpus of Legal Academic Writing in English consists of legal academic writing by students at certain Hong Kong universities and has been used to explore questions related to English-medium instruction, like the use of booster words such as "clearly" (Hafner & Wang, 2018).

Ultimately, combined legal corpora can include anything from larger corpora which are typically intended to be representative of a legal domain more generally, or smaller corpora that is made to respond to a particular research purpose.

3.3. Legislation

Egbert and Wood recognise that "there is a notable gap in our access to corpus linguistic resources for the purposes of investigation of statutory language" (Egbert & Wood, 2023: 2). While there are many issues that may contribute to this gap, scholars are beginning to produce legislative corpora that are representative of certain kinds of statutes. For instance the Swiss Legislation Corpus ('SLC') incorporates the entire classified compilation of Swiss legislative texts (Höfler & Piotrowski, 2011; Höfler & Sugisaki, 2014: 175). This resource includes an "up-to-date collection of statutory laws of the Swiss Confederation, comprising of anything from acts to ordinances and treaties, involving German, French and Italian texts". (Höfler & Sugisaki, 2014: 175). The corpus was primarily created to test automated coding software that could assist in preparing a legal corpus for "legislative drafting, legal linguistics […] translation and […] the evaluation of legislation" (Höfler & Sugisaki, 2014: 175).

Another multilingual resource is the Cadlaws corpus of English and French enactments of Canadian legislation from January 2001 to December 2018 (Sole-Mauri et al., 2021: 497). The purpose of construction of Cadlaws was to produce a parallel corpus that could be used for training Neutral Machine Translation ('NMT') systems (Sole-Mauri et al., 2021: 496–497). The corpus is accessible and can be downloaded in accordance with a Creative Commons Attribution International Licence (Sole-Mauri et al., 2021: 498). The utility of the corpus to extend beyond NMT use has not yet been noted, but the possible uses were stated to include translation studies, "linguistic research and natural language processing applications" (Sole-Mauri et al., 2021: 504).

The CorUSSS is the most comprehensive freely accessible corpus of statutes. This legislative corpus "contains the entire population of state-level statutes in the United States" (Egbert & Wood, 2023: 3). The primary purpose of the corpus was to contribute to accessible statutory corpora, with the corpus available through the Brigham Young University Law & Corpus Linguistics suite of corpora (Egbert & Wood, 2023: 3). The corpus was specifically intended to allow legislative words and phrases to be explored, especially providing for exploration "across states and *within* individual states – a tool that has thus far not been available" (Egbert & Wood, 2023: 3). It also allows greater flexibility to address legislative interpretation questions, overcoming limitations to answering these questions using legislative databases (Egbert & Wood, 2023: 3). This particular application was demonstrated through the example of exploring how the meaning of the word “information” could be construed through using the corpora, providing an alternative means to determining the ordinary meaning of terms across legislation (Egbert & Wood, 2023: 3).

Accordingly, there is minimal legislative corpora available that aims to be representative of statutes. Even more particular, there is limited corpora that can be used for more general law and language purposes, or specific legal linguistic purposes.

4. The ALC23

The ALC23 includes the entire population of legislation from each Australian jurisdiction.[1] The entire population is defined to include in force legislation as at 30 June 2023 that was available on legislative websites, and is relevant for interpretation purposes. Specifically, the corpus includes all available pieces of legislation that were machine-readable, and certain type of subordinate legislation that is necessary to understand the interpretation of acts. The ALC23 can be utilised as a whole, or as a variety of sub-corpora – subject to copyright restrictions.

4.1. In Force Legislation

The relevance of considering the concept of in force legislation in depth, and how it relates to the ALC23, is to ensure that researchers are aware of what the corpus encompasses. In the above noted legislative corpora, the introduction of Cadlaws and the Open Australian Legal Corpus is the only resource that provides information as to currency of the included legislation.

In the Australian context, in force legislation essentially means legislation that is current. However, what is deemed in force varies according to the individual legislative website, as different websites have different procedures to updating legislation due to the differences across jurisdictions. This meant that there were differences in the accessible legislation per jurisdiction, which had to be accounted for when downloading the legislation, to ensure it was current as at a certain date. As such, the ALC23 encompasses

[1] The included Australian jurisdictions are the Australian Capital Territory, Commonwealth, New South Wales, Northern Territory, Queensland, South Australia, Tasmania, Victoria, and Western Australia. Norfolk Island legislation, which is an external Australian territory, was not included for several reasons, which are outlined in discussions of corpus design.

legislation that was deemed by the individual legislative websites as being in force, as of 30 June 2023.

Further, the only occasions where legislation was not included as part of the ALC23 was where legislation was available, but not technically in force as of 30 June 2023, including because it had expired, been rescinded, or had not yet commenced. Other legislation not included involved that of which was unavailable on the relevant website, either because the jurisdiction does not make available that type of legislation,[2] or due to errors in links.[3] There were also some jurisdictions that had particular peculiarities, which made the inclusion of this legislation inappropriate. For instance, South Australia also included principal legislation that merely repealed acts,[4] or some of the legislation in the Northern Territory and the Commonwealth were available only as images.[5] However, these issues do not impact the validity of the corpus, given the parameters include what was available on the relevant legislative website, what was machine-readable, and what legislation is necessary to interpret other legislation.

4.2. Relevant Legislation

The selection of the legislation to be included in the ALC23 occurred on the basis of legislation that is current and relevant for the purpose of interpretation. As noted above, all principal acts and certain subordinate legislation were included. In relation to acts, all principal acts were included, rather than both principal and amending acts. This is because principal acts incorporate or consolidate any amendments, which means that only provisions that are technically in force as at 30 June 2023 are included, that of which would not occur if amending acts were also incorporated.

With respect to subordinate legislation, only certain kinds were included. Subordinate legislation refers to legislation that is made under an authority or body with power that is conferred by acts, and this may include statutory instruments, rules, or regulations. Within the ALC23, only rules and regulations were included, as this legislation often provides additional explanations of the provisions in acts, meaning that it is relevant to include for the purposes of interpreting or understanding the operation of acts. Other kinds of subordinate legislation are procedural in nature, and are often enacted

[2] In Western Australia, in force acts on the legislative website do not include treasury acts, reserves acts, road closure acts and some railway acts: (Government of Western Australia et al., n.d.).

[3] There were a small number of links to pieces of legislation that resulted in an error or led to a different piece of legislation.

[4] For example, the *State Procurement Repeal Act 2020* (SA). Inclusion of this type of legislation was inappropriate because the acts that were repealed would merely not appear in the corpus as legislation that is current. Further, no other jurisdiction contained legislation of this kind, and this legislation is merely procedural in nature rather than assisting with interpreting other legislation.

[5] For example, the *Supreme Court (Rules of Procedure) Act 1987* (NT) or the *Mutual Assistance in Criminal Matters (United States of America) Regulations 1999* (Cth). In total, this amounted to approximately 24 pieces of legislation that were not able to be machine-readable.

to comply with technical requirements. The reasons for this decision are outlined below in the discussion of corpus design.

4.3. Sub-Corpora

To increase the viability of future use of the ALC23, the corpus can be divided and analysed according to various sub-corpora, including either per jurisdiction, and/or per acts or subordinate legislation (see Table 1). Similar to the CorUSSS, division of the ALC23 can occur per jurisdiction, and is named according to the relevant jurisdiction. However, the uniqueness of the ALC23 lends attention to the ability to analyse corpora according to acts or subordinate legislation, with each sub-corpora named according to the relevant type of legislation. Additionally, the files containing legislation are split according to jurisdiction and type of legislation. This means that researchers may work with sub-corpora in any number of different ways, according to their specific research purpose. For instance, researchers could examine a particular jurisdiction or combination or jurisdictions, or a particular type of legislation in a particular jurisdiction.

Table 1: Titles and Composition of Sub-Corpora

Jurisdiction	ACL23-A	ACL23-SL	Total
ACTCor23 Australian Capital Territory	328 6,530,073	213 2,084,017	541 8,614,090
CthCor23 Commonwealth	1301 26,398,527	1112 10,594,300	2413 36,992,827
NSWCor23 New South Wales	864 13,027,507	435 4,451,374	1299 17,478,881
NTCor23 Northern Territory	381 4,747,923	261 1,731,558	642 6,479,481
QldCor23 Queensland	566 12,930,370	345 4,249,089	911 17,179,459
SACor23 South Australia	555 7,435,354	469 2,447,688	1024 9,883,042
TasCor23 Tasmania	593 5,814,030	320 1,614,596	913 7,428,626
VicCor23 Victoria	803 14,439,288	465 4,005,337	1268 18,444,625
WACor23 Western Australia	813 10,983,348	655 4,927,553	1468 15,910,901
Total	6203 102,306,420	4274 36,183,442	10478 138,411,932

Note: The jurisdiction column and associated rows represent the sub-corpora arranged by jurisdiction. The ACL23-A and ACL23-SL columns represent the sub-corpora arranged via acts or subordinate legislation. The larger text size represents the number of texts, and the smaller text size represents the word counts.

4.4. Restrictions

There are some copyright restrictions on the ALC23. That is, all legislation was downloaded from individual authoritative legislative websites; however, there are different copyright implications for certain jurisdictions.

For Tasmania, South Australia, Western Australia, Queensland, New South Wales, and the Commonwealth, all content on the legislative websites is provided under the Creative Commons Attribution 4.0 International license. This means that content is able to be reused freely subject to sufficient attribution and links (Australian Government, n.d.; Government of South Australia, n.d.-c; Government of Western Australia, n.d.; New South Wales Government, 2021; Queensland Government, 2020; Tasmanian Government, 2023).[6] As such, the sub-corpora related to these jurisdictions is accessible in accordance with the relevant license.

For the Northern Territory, copyright permission is granted subject to certain conditions being complied with, primarily that "the publication must not indicate directly or indirectly that it is an official version of the material" (Northern Territory Government, n.d.-b).[7] However, the permission may be revoked, varied, or withdrawn "on reasonable notice", and should this occur it will be removed from Sketch Engine (Northern Territory Government, n.d.-b).

For the ACT and Victoria, there are restrictions on this material. For the ACT, there is no statement on the legislative website as to the copyright of legislation. For Victoria, the legislative website states that "[n]o part may be reproduced by any process except in accordance with the provisions of the Copyright Act 1968 of the Commonwealth" (Victorian Government, 2020), with the authorised electronic versions of legislation available "for personal use only" (Victorian Government, 2020). Accordingly, these jurisdictions were directly contacted to obtain permission to include the relevant texts in the ALC23.

[6] The ALC23 and TasCor23 is sourced from material from the Tasmanian Legislation website at 29 June 2023. For the latest information on Tasmanian Government legislation please go to legislation.tas.gov.au. The ALC23 and SACor23 is sourced from content from the South Australian Legislation website at 2 and 4 July 2023. For the latest information on South Australian Government legislation, please go to legislation.sa.gov.au/. The ALC23 and WACor23 is sourced from content from the Western Australian Legislation website at 30 June 2023. For the latest information on Western Australian legislation, visit legislation.wa.gov.au. The ALC23 and QLDCor23 is sourced from content from the Queensland Legislation website at 30 June 2023. For the latest information on Queensland Government legislation please go to legislation.qld.gov.au. The ALC23 and NSWCor23 is sourced from content from the New South Wales Legislation website at 7 July 2023. For the latest information on New South Wales Government legislation please go to legislation.nsw.gov.au. The ALC23 and CthCor23 is sourced from content from the Federal Register of Legislation at 5–6 July 2023. For the latest information on Australian Government Legislation please go to legislation.gov.au/. The Creative Commons Attribution 4.0 International licence can be accessed via this link: creativecommons.org/licenses/by/4.0/. Changes made to any legislation include those which occur automatically by Sketch Engine's programs to convert the legislation into readable .txt files.

[7] Accordingly, the legislation contained in the NTCor23 and corresponding portion in the ACL23, is in no way an official version of legislation, but exists as a collection of available acts, regulations, and rules, downloaded on 30 June 2023.

4.5. Access

The ALC23 is accessible on Sketch Engine,[8] via the main corpus page, to paid users of Sketch Engine. Access is subject to accepting the relevant terms noted in the dialogue box.

5. Overcoming Issues

There are several issues that may contribute to the lack of useable legislative corpora. I will focus on three central issues that I identified when building the ALC23, including providing an explanation of how these issues could be overcome. The first issue that impacts corpora generally relates to the restrictive nature of corpus design and particularities of research questions. I will emphasise that a view to create general legislative corpora could effectively address this concern. The second issue is that there is typically a lack of information provided about the currency and availability of legislation. I consider that improving the extent of information provided when introducing corpora can increase utility. Finally, the data collection and collation process can cause significant problems, particularly where a researcher is not familiar with computational activities, such as web scraping. I will present how I overcame this issue by engaging with manually downloading legislation, along with using the software Sketch Engine. The combination of these processes also highlights issues in the collection and collation of data, such as the drawbacks of extensive tagging. Ultimately, there are many insights that can be gained regarding building legislative corpora, particularly from my perspective as a legal scholar.

5.1. Corpus Design

In building corpora, the design of a corpus and the scope of a research question often means a corpus is created according to a specific purpose, thereby limiting its reuse (Vogel et al., 2018: 1351). Accordingly, one reason for the scarcity in general legislative corpora is that corpora design is too narrow, often “serv[ing] one research question” (Vogel et al., 2018: 1351). This means that there are limited corpora that can be reused for multiple different research questions, particularly those that are broader in nature. An exception to this limitation is the CorUSSS, whereby the corpus was created with the intention of being representative of the entire population of legislation in the United States, with particular consideration provided to future utility. Similarly, in creating the

[8] The ALC23 can be accessed via the following link: url.au.m.mimecastprotect.com/s/W6juC3QNp4spAJw12UgfOFQzc_C?domain=app.sketchengine.eu.

ALC23, I aimed to overcome the issue of limited reuse by broadening my corpus design to ensure the outcome produced was a general corpus of Australian legislation. While "no one corpus can answer every research question" (Phillips & Egbert, 2017: 1589), I consider that constructing more general legislative corpora can assist in answering a broader range of research questions that are otherwise limited by smaller, or combined, legal corpora.

Essentially, my intention was to produce a general corpora that was representative of the "whole language" of Australian legislation, at a specified point in time (McEnery & Brookes, 2022: 36). As such, a related consideration in building corpora is whether the corpus is truly representative of the intended texts of which it is sampled. However, "there is no well-defined conception of what the sample is intended to represent" (Biber, 2008: 63), which makes it difficult to determine whether a corpus is truly representative of language. Crucially, Biber recognises that "[i]f we adopt the ambitious goal of representing a complete language, the population boundaries can be specified as all of the texts in the language" (Biber, 2008: 65). In the context of the goal of representing the complete language of legislation in Australia, the population boundaries can be specified as available legislation on authoritative legislative websites, each jurisdiction, and the types of legislation available in each jurisdiction. In relation to available legislation, only legislation that was accessible at the time of the download, and in a machine-readable format, was included. While this only amounted to a miniscule number of pieces of legislation, issues with formatting and related errors are necessary to account for when creating legislative corpora.

In relation to the jurisdiction, I opted to include every major jurisdiction within Australia, which incorporates both state and federal legislation. The only jurisdiction not included was Norfolk Island legislation, which is an external Australian territory. The primary reason that this jurisdiction was not included was because it operates differently to other jurisdictions within Australia, including with respect to legislation that is in force, and the use of ordinances.

In relation to the types of legislation, as noted above, all principal acts and certain subordinate legislation were included. Specifically, only rules and regulations were included, which meant that other statutory instruments, such as proclamations or orders, were not included. Only rules and regulations were incorporated, as this type of legislation often provides additional explanation of the provisions in acts, meaning that it is relevant to include for the purposes of interpreting or understanding the operation of acts. Essentially, subordinate legislation beyond rules and regulations holds limited relevance for examining the language of Australian legislation, as these instruments typically comply with procedural requirements, such as in the case of notifiable instruments. Additionally, most legislative websites only had rules and regulations available as part of their database of subordinate legislation.

5.2. Currency

A factor that has not extensively been considered in scholarship outlining the creation of legislative corpora is the matter of currency. Currency refers to information related to how recent the legislation included in corpora is. In the instance of the ALC23, I have chosen to include legislation at a certain point in time and have explicitly detailed this currency information above. I consider that when creating legislative corpora, an extensive consideration of currency is essential to increase utility.

With respect to the legislative corpora outlined above, information provided about currency is limited. For instance, while the Cadlaws corpus includes information about the inclusion period, being January 2001 to December 2018, it is unclear whether this is all that was available on the legislative websites, or if this period was selected for a particular reason. Similarly, while the SLC includes the Classified Compilation of Swiss Federal Legislation, information about when this compilation was last updated, or the scope, has not been explicitly discussed. The CorUSSS notes that "[t]he 2019 versions of the state codes were collected as they were the most recent at the time of corpus construction" (Egbert & Wood, 2023: 2), but no further information as to currency is provided. To a different extent, the Open Australian Legislative Corpus includes explicit information regarding when the data sources were last updated and the types of documents collected, but this is presented in the form of time stamps when web scraping occurred (Butler, n.d.-a). This information is also included within the files themselves. Comparatively, I have extensively outlined above the information related to currency, taking into account when the individual jurisdiction websites update their data. This detailed currency information is useful to make the corpus accessible for broader purposes. Additionally, I outline the kinds of issues that may impact currency, using the Australian context as an example.

Extensive currency information related to legislation that is collected and used in corpora is necessary to extend the utility of a corpus for broader purposes. That is, it is not that a lack of information related to currency is an issue per se, but rather, the lack of information about currency restricts the ability for corpora to be used for general questions related to law and language. For instance, when legal scholars write about contemporary law, the matter of currency is of particular importance to ensure that the most recent version of legislation is considered. If extensive information about currency and the associated matter of accessibility is not included, this could limit instances where corpora are utilised. For example, a broad audit of legislation would require explicit information about currency, to assist in making conclusions about whether particular provisions or legislation have since been amended, or at what point of time the relevant audit has been conducted. Additionally, by ensuring the legislation has been collected at a specified point in time, any future corpora that may be utilised for comparisons, otherwise termed as diachronic analysis, could ensure greater consistency between corpora.

In collecting legislation to create corpora, issues arise with respect to currency when working with official legislative websites. In the Australian context, these issues arose due to the varied sophistication and publicly available information related to legislation. As noted above, the meaning of in force legislation differed on account of the individual legislative website. For instance, some websites noted that the available legislation was only that of which were assented to, made, or in force, after a certain date (Government of South Australia, n.d.-a, n.d.-b; Northern Territory Government, n.d.-a; Tasmanian Government, 2021).[9] Additionally, the incorporation of amendments to principal legislation occurred over a different time period (Australian Capital Territory Government, n.d.-a; Government of Western Australia et al., n.d., n.d.; New South Wales Government, n.d.),[10] which meant that to be truly representative of a point in time, collection had to occur after a certain date and account for additional changes. Again, there was variation in legislative websites with regard to how amendments were tracked, meaning that some jurisdictions provided extensive updates in relation to specific legislation (Australian Capital Territory Government, n.d.-b), and others did not.

5.3. Data Collection and Collation

While the collection of files for use in a corpus has been improved through computer programming, this posed an issue for me because I do not have technical knowledge of software or coding. This means that the collection of texts, and the associated cleaning and tagging process, is somewhat inaccessible to scholars who also do not have this knowledge, or, like me, did not have the time or resources to gain this knowledge. Accordingly, there are several stages throughout the data collection and collation process where issues can arise, and I suggest that these issues can actually be combatted through manual downloading legislation and relying on the software Sketch Engine. In effect, this manual process enabled me to create the corpus, but may also be of use to corpus linguists who do rely on automated software.

5.3.1. Manual Downloading

Typically, when written corpora are created, the process for collecting texts is automated through web-scraping. For instance, the construction of the Romanian Legal Corpus involved web crawling to collect legislation, with HTML-tag and other mark ups cleaned

[9] In the NT, legislation included acts that were assented to from 1 July 1978. In Tasmania, legislation included acts that were consolidated as of 1 February 1997, and statutory rules that were made after 1 May 198. In South Australia, acts and regulations and rules included those in force as of 1 January 2003. Other jurisdictions did not include specific information.

[10] For the ACT, incorporation occurred the day after amendments came into force or shortly after. For NSW, incorporation occurred within three days of commencement. For Qld, amendments were incorporated as soon as possible. For WA, amendments were incorporated within two working days. Other jurisdictions did not include specific information.

to ensure raw text and specific metadata remained (Tufiș et al., 2020: 2774). In the Australian context, when constructing the OALC, Butler explained that the process turned into a "year-long journey", entirely due to the web-scraping process (Butler, 2023). A significant impediment was that permission from the individual legislative websites was often required to use web-scraping tools to scrape data, whereby negotiation was necessary (Butler, 2023). Once permission was received, indexing and downloading of documents had to occur, but there were difficulties with respect to application programming interfaces, which meant that complex coding was required (Butler, 2023). Similarly, the CorUSSS involved collecting legislation through web scraping using Python, with modifications made to the code, due to issues arising from variation across different states (Egbert & Wood, 2023: 2). This process resulted in individual files for each legislation, with Python scripts then used to clean the files to ensure only text that was part of the original legislation was contained (Egbert & Wood, 2023: 3). As such, these examples demonstrate that the purpose of web-scraping is to collect texts, but additional cleaning is required to ensure that only the raw text is included.

Cleaning can involve anything from background information, such as headlines, to duplications of text, being removed (McEnery & Brookes, 2022: 43). For instance, in the CorUSSS, text such as "website contact numbers, ads, and disclaimers" were removed, in a process that was described as "particularly time consuming, as each state varied in the information included about the statute" (Egbert & Wood, 2023: 3). Cleaning is an arduous process, but necessary to ensuring only the raw text is included in the data. However, cleaning requires a level of understanding about code, as this process occurs within plaintext files (TXT). Ultimately, the purpose of cleaning is to ensure only information relevant to the original texts are included. In using web-scraping to collect individual legislation, a lot of unnecessary information can be incorporated. To avoid unnecessary information, and tiresome cleaning, a useful workaround is manually downloading legislation, should time allow.

I opted to manually download all relevant acts and statutory instruments from each Australian jurisdiction legislative websites. This process overcame a lot of issues that occur when dealing with legislation. In particular, complexities associated with web-scraping and Python were eliminated. That is, no permission was required from legislative websites because I was not using web-scraping software. Additionally, no clean-up of the files was required because the individual pieces of legislation were presented in their singular form, without any additional metadata or information from websites.

There were also several unforeseen benefits that arose when using this manual process. First, legislative websites are the most authoritative and freely accessible source for accessing legislation in Australia. This means that the manual process is able to be replicated by any person, and they can be assured that amendments have been effectively incorporated. Second, the manual process meant that checks and decisions were able to occur in real time. For instance, there were many documents that were incorrectly linked or duplicated, or just not available. In an automated process, duplication of legislation could occur but under a different name, meaning it may evade detection by software. Additionally, there were variations in jurisdictions, which meant that certain types of legislation could be excluded for consistency purposes. On a related point, reviews meant that duplications of Commonwealth legislation within other jurisdictions could be addressed on an as needed basis. For instance, national laws that were implemented across jurisdictions, such as the Heavy Vehicle National Law, were duplications, but these were included to ensure that sub-corpora could be split into individual jurisdictions. Third, in individually downloading and renaming the files to present the name of the legislation, an associated result was a collection of legislation at a certain point in time. This meant that these files could be referred back to in order to more easily identify the relevant provision. This would be particularly useful whereby more information is required rather than just the name of legislation or the general position within text.

The only drawback I identified in this manual process was the time constraints. The process occurred over several days, where many hours were spent downloading, reviewing, and taking notes to ensure all available texts were included in the corpus. However, the issues experienced when constructing the OALC, by a person who has significant experience in computing, arguably justified my time spent manually downloading and reviewing. Additionally, my manual process limited the time spent on cleaning a corpus, as only raw files were included, rather than any additional information that could have been picked up while web-scraping.

An alternative process to engaging with individual downloads would be to work from an already completed compilation of legislation, like that which occurred when building the SLC (Höfler & Piotrowski, 2011). However, in Australia, compilations are currently inaccessible for corpus construction. This is because compilations are not freely available on legislative websites. The only database that includes compilations is AustLII, where not only is the currency information somewhat unclear, but these compilations are subject to copyright, due to additional annotations. Further, utilising compilations would require knowledge of coding or linguistic software. For instance, in the SLC, Höfler and Sugisaki "developed a tool that automatically detects the boundaries of [...] structural units and marks them in the XML representation" (Höfler & Sugisaki, 2014: 176). As such, working from a compilation would require checks to ensure the individual pieces of legislation were demarcated and labelled correctly – something which can instead be assured through collating individual pieces of legislation.

A related matter when working with legislative websites is associated with converting all files to a standardised format. That is, there was variation in legislative websites with respect to the available file types. For instance, the range of available documents includes anything from PDF, RTF, HTML, or DOCX files. Further, some PDFs available were only images of older legislation, which were excluded because they could not be automatically transcribed.[11] With respect to variation in file types, Butler noted issues with converting to a standardised plain text format, specifically highlighting difficulties in "presev[ing] […] spacing, indentation, line breaks and tables" (Butler, 2023). For my purposes, preserving these features were unnecessary, as the focus of the corpus is to assist in exploring the language of legislation, rather than training language models. As such, I opted for consistency by downloading all files in PDFs. However, there still remained the issue of how to convert these files to plaintext files (TXT), which is required for corpus linguistic software. While file conversion tools exist, I opted to use the built-in feature within the Sketch Engine software, which could also assist with further inclusions to the files.

5.3.2. Sketch Engine and Tagging, Annotation, and Mark-Up

There are several different software available to conduct linguistic analysis, each with their own features and benefits (Baker, 2023: 47–49). In selecting software, a key factor for my consideration was whether it could assist with building my corpus. An additional consideration was whether it was user-friendly, including explaining how linguistic analysis could occur, alongside providing support. For this reason, I selected Sketch Engine (Kilgarriff et al., 2014).

Sketch Engine has broader scope than other software to build a corpus, which could occur either through inbuilt web-scraping capabilities, or through uploading multiple PDFs – thus bypassing the need to engage with additional text convertor tools that would convert files to text format (Lexical Computing, n.d.-c). A further benefit of Sketch Engine is that they offer a Boot Camp, which is a course that is designed to assist in navigating the Sketch Engine system and using corpus tools (Lexical Computing, n.d.-b).

A significant benefit of Sketch Engine for my purposes was that the system provides automatic part-of-speech tagging and lemmatisation (Lexical Computing, n.d.-a). Tagging refers to information that is encoded into corpora (McEnery & Brookes, 2022: 43), but this could be perceived as an issue that impacts the useability of legal corpora. That is, new corpora will typically be tagged, annotated, or marked-up with additional contextual or linguistic information, particularly according to what the relevant research question requires (Cock et al., 2024; McEnery et al., 2010: 29). This process is intended to assist in the analysis of the corpus, such as by including annotations that provide further contextual information, such as when the text was created, the relevant target audience, demographic information, or date of publication (Baker, 2010: 15; 2014: 8). Additional marking-up or tagging of a corpus could involve part-of-speech tagging (McEnery et al., 2010: 29), syntactic parsing, semantic annotation, error tagging, lemmatization, or coreference annotation (McEnery et al., 2010: 29; 35–42), or the incorporation of

[11] These only occurred in Western Australia, but these pieces of legislation were significantly outdated.

context using schemes such as the Text Encoding Initiative (Barnard et al., 1996), or the Dublin Core Metadata Initiative (Dekkers & Weibel, 2003).

In relation to legal corpora, the variation in tagging depends upon the particular corpus. In the SLC, along with part-of-speech and lemmas, annotation of text segments were required, along with morphosyntactic and partial syntactic information (Höfler & Sugisaki, 2014: 175–176). The difference in information was determined according to what the corpora was intended to be used for – for instance, greater annotation was needed for the German-language texts, as it was to be "used as a testing environment for the development of an automatic style checker for legislative drafting" (Höfler & Sugisaki, 2014: 175). Similarly, the Romanian Legal Corpus also included significant annotation, with anything from part-of-speech tagging to dependency parsing (Tufiș et al., 2020: 2774). In contrast, the CorUSSS was part-of-speech tagged to allow for sortable concordance lines (Egbert & Wood, 2023: 3), and Cadlaws was tokenised and aligned using various software (Sole-Mauri et al., 2021: 497).

Crucially, McEnery and Brookes have clarified that while annotation can assist in searching for linguistic purposes, it "is not essential for corpus analysis […] [including because] it can be [a] time consuming and resource-draining process" (McEnery & Brookes, 2022: 45). Further, for a corpus to have broader use, the text should ideally be unprocessed or free of mark-up, thus permitting it to be annotated according to the relevant research question (McEnery et al., 2010: 29; Sinclair, 1991: 21). For these reasons, I did not engage in any further annotation, tagging, or mark-up with respect to the ALC23, aside from the automatic part-of-speech tagging and lemmatisation. These automatic functions allow the ACL23 to act similarly to the CorUSSS, where they can be assessed using sortable concordance lines. This means that the ALC23, in its current form on Sketch Engine, is currently fit for purpose to address broader explorations on the language of the law, particularly as it relates to the use of concordancing. Further, this minimal approach will allow other scholars to engage in more particular annotation, tagging, or mark-up, depending on the research question.

6. Applications

A specific benefit of the ALC23 is that it provides an alternative method to conducting research on the language of legislation, particularly when compared to legal databases, such as Westlaw or AustLII, or legislative websites. Significantly, there are some limitations with respect to search functionality of websites or databases. There are issues related to legislative websites, whereby some websites in Australia have limited search functionalities, limited terms indexed, or there are issues in indexations which restricts the ability to search for certain terms within legislation. For instance, Genovese's research into gendered language in legislation revealed that the Commonwealth legislation website at the time did not index certain pronouns, or revealed inaccurate results (Genovese, 2023: 680 fn 226). With respect to legal databases, while there may be increased functionality, any subsequent analysis must occur on a manual basis (See, e.g., Grey & Severin, 2021). Additionally, Vogel et al. have specifically noted the benefits of broad legal corpora in overcoming issues with legal databases (Vogel et al., 2018: 1351). These benefits include that the material is "accessible for statistical processing", whether "annotation layers can be added that enable new search capabilities, such as retrieving multiword expressions or syntactic structures" (Vogel et al., 2018: 1351). The potential for the ALC23 to be subject to additional annotation has already been discussed above, but even without further annotation, there are significant benefits that arise from statistical processing. I will demonstrate these benefits by using Sketch Engine to search for the word "gender" in the ALC23.

6.1. Statistical Processing

A substantial advantage of innovations in corpus linguistic analysis is the development of different software programs that can assist in rapidly organising and analysing large corpora (Baker & McEnery, 2015: 1). The types of analysis that could be conducted depends upon the specific software that is used, as different programs offer different tools. For instance, AntConc4 is a free program that provides anything from keyword extraction, concordancing, and wordlist generation (Anthony, n.d.). WordSmith Tools also allows concordancing with advanced search capabilities, including permitting multiple words to be searched at one time (Lexical Analysis Software & Oxford University Press, n.d.). The program also provides for complementary approaches in the presentation of data, such as through visual dispersion plots (Baker, 2010: 28). Sketch Engine is especially useful, in that along with typical tools, the 'Word Sketch' and 'Word Sketch Difference' functions are accessible, those of which provide users with organised information about specific grammatical patterns surrounding one or multiple words (Sketch Engine, n.d.).

Different software and corpus linguistic techniques can be used alongside a particular corpus approach. These approaches are best described as existing on a continuum from corpus-based to corpus-driven perspectives (Baker, 2010: 8). A corpus-based approach involves a researcher analysing a corpus with preconceived theories in mind, thus influencing the type of corpus selected, and the tools and strategies used (McEnery et al., 2010: 10; Tognini-Bonelli, 2001: 65). For example, some researchers will have preselected terms or hypothesis that will be investigated within the corpus (Baker, 2023: 16). As further noted by Breeze in relation to the law,

> [c]orpus-based methods [in particular] offer a useful way to approach specialised genres, since their strength lies in their ability to detect what is characteristic about texts of a conventionalised nature […]

> [making it] possible to understand more about how language is used – and therefore how meaning is made – in those texts (Breeze, 2019: 79).

In that article, Breeze used part-of-speech analysis of academic writing, case law, legal documents, and legislation to uncover legal language and the related grammatical patterns that distinguish each type of genre (Breeze, 2019: 80–81). In exploring the frequency of terms, Breeze was able to uncover the key features of each legal genre, thus highlighting the similarities and differences between different uses of legal language (Breeze, 2019: 99–100). An additional example of a corpus-based approach arguably occurs when legal corpora are used to interpret legislation. For example, Egbert and Wood used the CorUSSS to uncover the meaning of the term 'information' in legislation (Egbert & Wood, 2023: 3–4). They specify that the CorUSSS could have been used to consider the ordinary meaning of 'information' in a case where the Corpus of Contemporary American English was instead used (Egbert & Wood, 2023: 3). Given the ALC23 is constructed similarly to the CorUSSS, there is potential for the ALC23 in its current form to be used for similar corpus-based inquiries that relate to legislative interpretation. In fact, as noted in the introduction, the ALC23 was created for my purpose of conducting a corpus-based exploration of terms related to sex, gender, and sexuality in Australian legislation.

In comparison to a corpus-based approach, a corpus-driven approach instead takes the corpus as the starting point, with the data collected from the corpus guiding the direction of the theories developed (Baker, 2010: 7). In this process, the statistical tools and techniques that are utilised directly result in linguistic categories, allowing somewhat of an objective outcome that is subjectively interpretated (McEnery et al., 2010: 10). On this continuum, there also exists corpus-assisted approaches, those of which are typically used alongside other methodology to improve results and address limitations. For instance, an emerging field includes corpus-assisted discourse studies, whereby analysis of discourse occurs with the assistance of corpus method (Mautner, 2022: 250). Goźdź-Roszkowski has specifically detailed how each of these approaches on the continuum have "been used to investigate legal discourse more broadly" (Goźdź-Roszkowski, 2021: 1517–1518). Accordingly, while it is clear that there are many specific corpus approaches that can be used to analyse the ALC23, I will demonstrate below some benefits of using the corpus for legal research, through the example of exploring the word "gender".

6.2. Gender

While the advantages of statistical processing are extensive, one application for my research purpose is that the ALC23 is able to effectively provide insight into how certain terms related to sex, gender, and sexuality are used across legislation. This insight can extend to the individual uses of terms, the identification of case studies, or comparisons between and across jurisdictions. An associated advantage is that the corpus is general in nature, which means the subject matter is not restricted to legislation that contains certain keywords. In particular, an issue related to the exploration of sex, gender, and sexuality under the law is the matter of silence. That is, historically, certain identities have not been explicitly referred to under the law, such as using the terms 'sodomy' or 'buggery' to describe homosexual offences rather than merely using the term 'homosexual' (Moran, 1996: 8; 9; 33–38). While legislation today is clearer in identifying some queer identities, the need to draw from a general legislative corpus is thus useful when examining queer issues, or other issues where the scope of legalese may be less clear. Had a corpus merely been built according to certain keywords, the results would be limited only to that terminology, even if further terms that represented these concepts were uncovered throughout analysis.

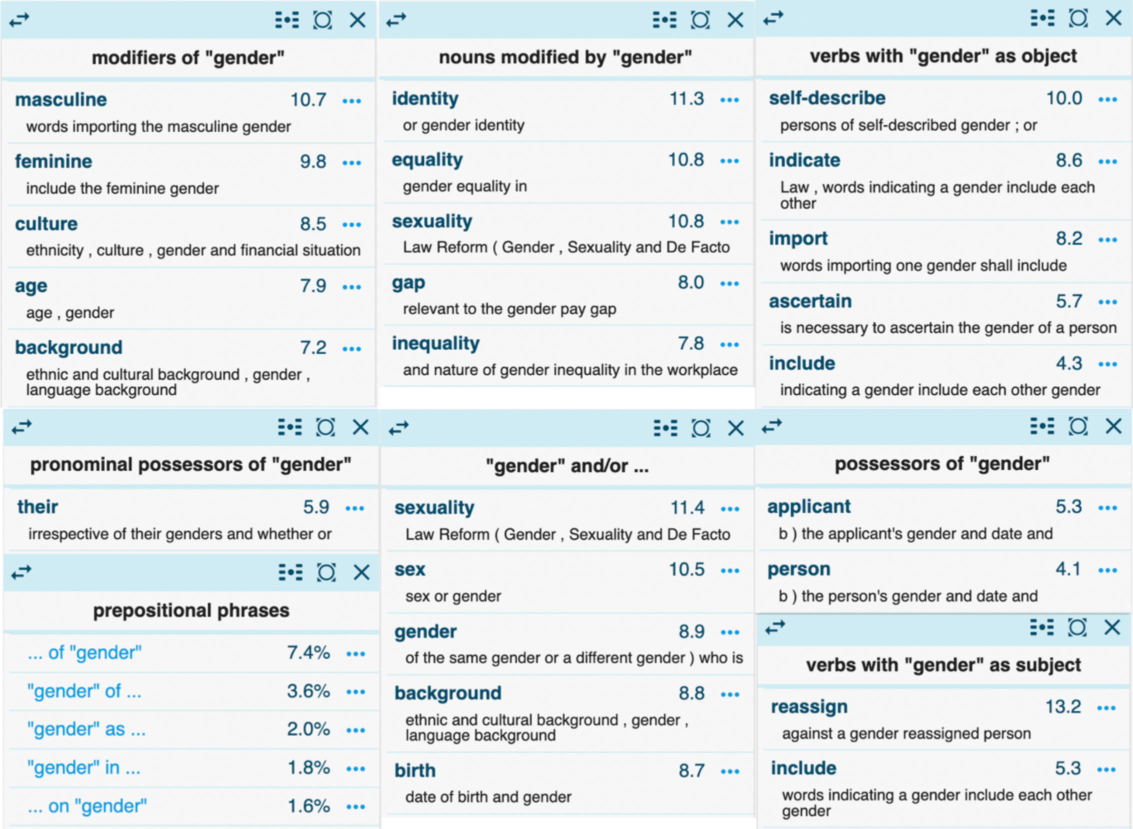
These advantages can be demonstrated through engaging with corpus linguistic tools in Sketch Engine to analyse the word "gender", with insight to be provided on the use of the word itself, alongside assisting with identifying potential avenues for further research. In current literature, uncovering the contemporary meaning of sexed/gendered concepts within legislation tends to occur using manual analysis (Genovese, 2023; Haigh, 2018), which may in part be attributed to the lack of legislative corpora. Accordingly, these kinds of analysis are confined to certain jurisdictions (Genovese, 2023), languages, or terms (Haigh, 2018). The result is that the conclusions drawn about sexed/gendered concepts provide an incomplete picture as to the distinctions across jurisdictions, or in relation to certain concepts. As a result, this brief overview into the meaning of the word "gender" provides an introductory snippet as to some perceptions that may be gained from corpus linguistic tools.

A simple search of concordances in revealed 2,760 hits, across 674 different pieces of legislation, with the most frequent appearance of the term occurring with 238 hits in the *Gender Equality Act* 2020 (Vic), and a total of 504 hits in Victorian acts. Merely reviewing the frequency of these terms according to either the individual piece of legislation where hits appear, or a particular jurisdiction and legislation type, could suggest avenues for in depth investigation. However, the co-text of "gender" is also significant, which can be uncovered by reviewing the concordance lines or examining collocations of the word "gender". This contextual examination would be particularly useful in relation to gender, given the frequent use aligns closely with equality between women and men, rather than relating to a particular identity. To examine context efficiently, a useful feature is the Word Sketch function, which assists in analysing collocations by grouping the collocates on the basis of their grammatical relations (Sketch Engine collocates). This function revealed that gender is used as a noun 2,688 times, and as a verb 72 times. While "gender" has been incorrectly tagged as a verb by Sketch Engine, the results are nonetheless useful for identifying alternative relations to gender. The top five results from each category sorted according to logDice score have been replicated in Image 1 and Image 2 below. The logDice score indicates how strong the collocation is, relying on relative frequencies,

DOI: 10.14762/jll.2025.030

with a "very high score of the collocate mean[ing] that there is little competition from other collocates" (Lexical Computing, n.d.-d; Lexical Computing, 2015; Rychlý, 2008).

Image 1: Gender as a Noun



modifiers of "gender"		
masculine	10.7	words importing the masculine gender
feminine	9.8	include the feminine gender
culture	8.5	ethnicity , culture , gender and financial situation
age	7.9	age , gender
background	7.2	ethnic and cultural background , gender , language background

nouns modified by "gender"		
identity	11.3	or gender identity
equality	10.8	gender equality in
sexuality	10.8	Law Reform (Gender , Sexuality and De Facto
gap	8.0	relevant to the gender pay gap
inequality	7.8	and nature of gender inequality in the workplace

verbs with "gender" as object		
self-describe	10.0	persons of self-described gender ; or
indicate	8.6	Law , words indicating a gender include each other
import	8.2	words importing one gender shall include
ascertain	5.7	is necessary to ascertain the gender of a person
include	4.3	indicating a gender include each other gender

pronominal possessors of "gender"		
their	5.9	irrespective of their genders and whether or

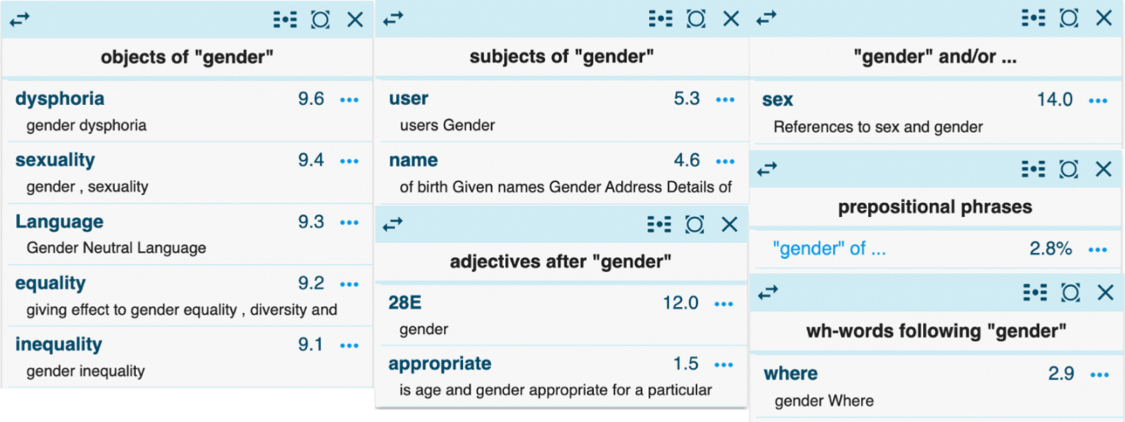
prepositional phrases	
... of "gender"	7.4%
"gender" of ...	3.6%
"gender" as ...	2.0%
"gender" in ...	1.8%
... on "gender"	1.6%

"gender" and/or ...		
sexuality	11.4	Law Reform (Gender , Sexuality and De Facto
sex	10.5	sex or gender
gender	8.9	of the same gender or a different gender) who is
background	8.8	ethnic and cultural background , gender , language background
birth	8.7	date of birth and gender

possessors of "gender"		
applicant	5.3	b) the applicant's gender and date and
person	4.1	b) the person's gender and date and

verbs with "gender" as subject		
reassign	13.2	against a gender reassigned person
include	5.3	words indicating a gender include each other gender

Image 2: Gender as a Verb



objects of "gender"		
dysphoria	9.6	gender dysphoria
sexuality	9.4	gender , sexuality
Language	9.3	Gender Neutral Language
equality	9.2	giving effect to gender equality , diversity and
inequality	9.1	gender inequality

subjects of "gender"		
user	5.3	users Gender
name	4.6	of birth Given names Gender Address Details of

adjectives after "gender"		
28E	12.0	gender
appropriate	1.5	is age and gender appropriate for a particular

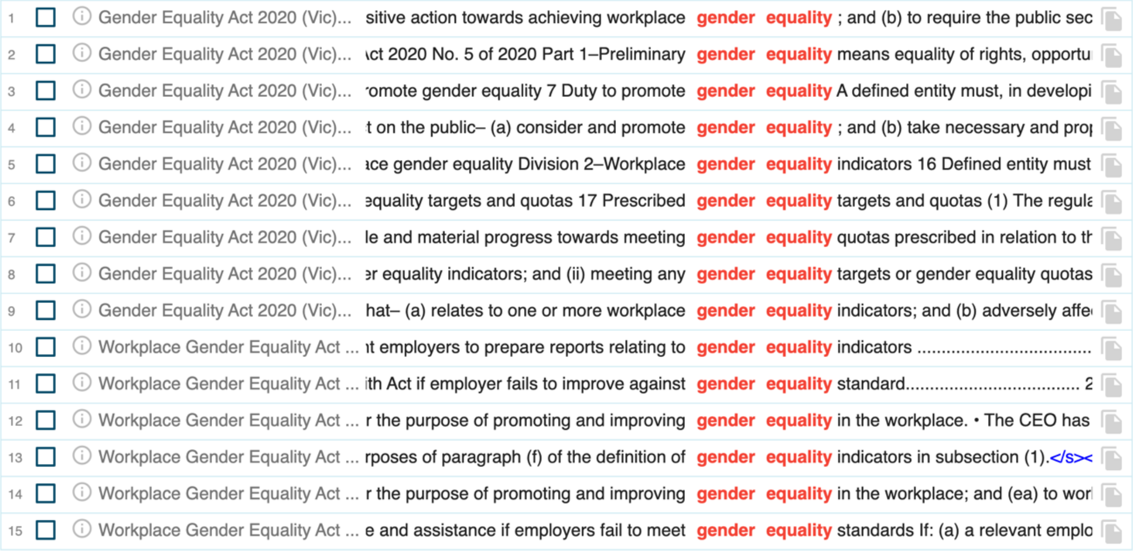
"gender" and/or ...		
sex	14.0	References to sex and gender

prepositional phrases	
"gender" of ...	2.8%

wh-words following "gender"		
where	2.9	gender Where

A brief review of these images allows several conclusions to be drawn. As indicated by the previous examination of frequency, terms related to equality between women and men, such as "equality", "inequality", and "gap", typically appears alongside "gender" (see Image 1 – "nouns modified by 'gender'"; see Image 2 – "objects of 'gender'"). A review of the co-occurrences of "equality" and "gender" specifically in a range of three words either side of the node confirms this association of equality between women and men. Specifically, this co-occurrence is found across 18 pieces of legislation,[12] most of which relate to employment, occurring most frequently in the *Gender Equality Act 2020* (Vic), and the *Workplace Gender Equality Act 2012* (Cth) (see Image 3).

Image 3: A Random Sample of 15 Concordance Lines of the Co-Occurrence of 'gender' and 'equality'



1	□	ⓘ Gender Equality Act 2020 (Vic)...	sitive action towards achieving workplace	**gender**	**equality**	; and (b) to require the public sec
2	□	ⓘ Gender Equality Act 2020 (Vic)...	ıct 2020 No. 5 of 2020 Part 1–Preliminary	**gender**	**equality**	means equality of rights, opportu
3	□	ⓘ Gender Equality Act 2020 (Vic)...	·omote gender equality 7 Duty to promote	**gender**	**equality**	A defined entity must, in developi
4	□	ⓘ Gender Equality Act 2020 (Vic)...	t on the public– (a) consider and promote	**gender**	**equality**	; and (b) take necessary and pro
5	□	ⓘ Gender Equality Act 2020 (Vic)...	ıce gender equality Division 2–Workplace	**gender**	**equality**	indicators 16 Defined entity must
6	□	ⓘ Gender Equality Act 2020 (Vic)...	ɘquality targets and quotas 17 Prescribed	**gender**	**equality**	targets and quotas (1) The regul
7	□	ⓘ Gender Equality Act 2020 (Vic)...	le and material progress towards meeting	**gender**	**equality**	quotas prescribed in relation to tl
8	□	ⓘ Gender Equality Act 2020 (Vic)...	ɘr equality indicators; and (ii) meeting any	**gender**	**equality**	targets or gender equality quotas
9	□	ⓘ Gender Equality Act 2020 (Vic)...	hat– (a) relates to one or more workplace	**gender**	**equality**	indicators; and (b) adversely affe
10	□	ⓘ Workplace Gender Equality Act ...	ıt employers to prepare reports relating to	**gender**	**equality**	indicators ....
11	□	ⓘ Workplace Gender Equality Act ...	ith Act if employer fails to improve against	**gender**	**equality**	standard.................................... 2
12	□	ⓘ Workplace Gender Equality Act ...	r the purpose of promoting and improving	**gender**	**equality**	in the workplace. • The CEO has
13	□	ⓘ Workplace Gender Equality Act ...	rposes of paragraph (f) of the definition of	**gender**	**equality**	indicators in subsection (1).</s><
14	□	ⓘ Workplace Gender Equality Act ...	r the purpose of promoting and improving	**gender**	**equality**	in the workplace; and (ea) to worl
15	□	ⓘ Workplace Gender Equality Act ...	e and assistance if employers fail to meet	**gender**	**equality**	standards If: (a) a relevant emplo

However, this association between "equality" and "gender" only appears in legislation within Victoria, Commonwealth, Western Australia, and South Australia. This would suggest that in these jurisdictions, there is a continued reliance on relating gender to matters of equality between women and men. Comparatively, a search of the ACTCor23 for collocations of gender in a range of three words either side of the node reveals only 121 hits, with gender appearing predominantly in relation to matters of sexuality and conversion therapy practices (see Table 2). This example suggests that in a jurisdiction where gender is not substantially related to equality, gender is instead referred to with respect to an identity that exists beyond a binary conceptualisation. The association between "gender" as an identity also appears across the whole corpus, with gender also typically used alongside terms like "dysphoria", "self-describe", and "reassign" (see Image 1 – "verbs with 'gender' as object", "verbs with 'gender' as subject"; Image 2 – "objects of 'gender'").

[12] In order of frequency, this includes: Gender Equality Act 2020 (Vic); Workplace Gender Equality Act 2012 (Cth); Local Government Act 2020 (Vic); Fair Work Act 2009 (vol 1) (Cth); Financial Framework (Supplementary Powers) Regulations (1997) (Cth); Jobs and Skills Australia Act 2022 (Cth); Local Government (Governance and Integrity) Regulations 2020 (Vic); Appropriation Act (No 1) 2021-2022 (Cth); Appropriation Act (No 1) 2020-2021 (Cth); Appropriation Act (No 3) 2021-2022 (Cth); South Australian Motor Sport Act 1984 (SA); Procurement (Debarment of Suppliers) Regulations 2021 (WA); Supply Act (No 1) 2020-2021 (Cth); Prevention of Family Violence Act 2018 (Vic); Appropriation (Coronavirus Response) Act (No 1) 2021-2022 (Cth); Supply Act (No 1) 2022-2023 (Cth); Supply Act (No 3) 2022-2023 (Cth).

DOI: 10.14762/jll.2025.030

Table 2: Collocations of 'gender' in the ACTCor23

Word*	Cooccurrences	Candidates**	LogDice
Sexuality	26	42	12.35
sexuality	22	39	12.14
Practices	21	38	12.08
conversion	24	106	11.76
Conversion	21	96	11.63

* Note: The capitalised words appear within the title of legislation, which is frequently repeated throughout the corpus.

** Note: Candidates refers to the total number of occurrences of the collocate in the whole corpus.

Additionally, there are several words that relate to the interpretation of gender in legislation, such as "masculine", "feminine", "indicate", "import", "include", and "Language" (see Image 1 – "modifiers of 'gender'", "verbs with 'gender' as object", "verbs with 'gender' as subject"; see Image 2 – "objects of 'gender'"). These collocates reveal that references to gender within legislation also relate to gender-neutral drafting responses, which either attempt to require using inclusive language when drafting legislation, or address "'the masculine rule": "[which is] a sexist drafting technique whereby masculine language and pronouns are exclusively utilised in legislation" (Genovese, 2023, 674). In fact, the earliest use of the term "gender" appears in section 2 of the *Infants Property Act 1830* (Imp) (WA), where a reference to "the masculine gender only [...] shall be understood to include and shall be applied to several persons as well as one person, and females as well as males". This was able to be uncovered by reviewing and organising the text types of each result. One conclusion that may be drawn from these considerations is that there is a long history within Australian jurisdictions to address the masculine rule within legislation through both interpretative provisions and encouragement to draft with gender neutrality. Given the initial presence within an imperial act, this is likely a legacy carried over from the United Kingdom legislative drafting principles.

On a different note, "gender" is also listed as something that is required as part of personal information (see Image 1 – "possessors of 'gender'"; Image 2 – "subjects of 'gender'"). In these examples, gender is requested alongside information like 'date' of birth (see Image 1 – "'gender' and/or"; see Image 2 – "subjects of 'gender'"), 'name', and address (see Image 2 – "subjects of 'gender'"). Reviewing the concordances of these instances highlights the particular legislation where these provisions appear; however, to understand the context surrounding for what purposes this information is requested, the legislation itself must be reviewed, including to identify the particular provision in which it occurs. Accordingly, a general conclusion drawn from the Word Sketch function is that gender typically occurs with respect to other kinds of personal information, such as name, address, or date of birth.

Similarly, "gender" also appears to operate as a characteristic, existing alongside words like "culture", "age", "background", and "sexuality" (see Image 1 – "modifiers of 'gender'"; "'gender' and/or"; Image 2 – "objects of 'gender'"). While it is reasonable to infer from the images that gender is also used within legislation as a characteristic, an examination of concordance lines can qualify this conclusion. That is, the lines associated with "gender" and "sexuality" actually reveal that contextually, the typicality between these two terms is largely impacted by the repetition across multiple pieces of legislation in the name of amending legislation in the Northern Territory: *Law Reform (Gender, Sexuality and De Facto Relationships) Act 2003* (NT).[13] While there are instances where 'sexuality' is used as a characteristic alongside "gender",[14] which continues to support the use of gender as a characteristic, examining the concordances demonstrates that a lesser emphasis should be placed on the typicality of the association with 'sexuality'. This may indicate that drawing conclusions from legislative corpora should also be balanced alongside the features of legislative corpora, whereby repetition of words or phrases could occur more frequently due to naming conventions, and amendment information.

It is imperative to note that there are a myriad of other conclusions that could be drawn from this data alone, let alone further examinations in Sketch Engine or other linguistic software. For instance, changing the collocation or statistical examination could impact the conclusions drawn (Baker, 2006: 102; Brezina et al., 2015), or different researchers could produce different interpretations of the data presented here – due to their own views and biases. (Baker, 2006: 18; 2008: 23). Rather than acting as a detracting factor, the varied interpretations and avenues for additional research should be viewed as beneficial for legal scholars, as it ultimately provides different avenues for further research. Essentially, the primary purpose of this section is to demonstrate the benefits of using general legislative corpora, that of which could include examining how a term is used across many or individual jurisdictions.

7. Limitations

There are several minor limitations to the ALC23 in its current form that have been highlighted by the above example. First, as noted in relation to Image 2, the automatic tagging that Sketch Engine conducts is not entirely accurate with respect to distinguishing between nouns and verbs. However, it is well accepted that the use of automated software for tagging cannot be entirely accurate, with Sketch Engine noting 95% accuracy (Lexical Computing, 2021). While this may pose an issue for certain corpus linguistic inquiries, the corpus remains fit for purpose for general enquiries made by legal scholars.

[13] There are 52 pieces of legislation in the Northern Territory where this amending legislation appears.

[14] *Mental Health Act 2014* (WA); *Mental Health Act 2014* (Vic); *Children's Guardian Regulation 2022* (NSW); *Intervention Orders (Prevention of Abuse) Act 2009* (SA).

The second limitation is that Sketch Engine's file conversion does not allow the alignment of provisions to be maintained, which means that the particular provisions where a word or phrase appears will not be readily identified. In essence, if a legal scholar wished to identify the specific provision where the language occurred, this would have to be uncovered by reviewing the legislation in its original form. The matter of alignment has not been an issued addressed or combatted in current legislative corpora, potentially because identification of the specific provision is largely unnecessary, especially if research is conducted from a purely linguistic lens. Nonetheless, this limitation is not a major detriment of the corpus, as the nature of legislative documents often requires recourse to other provisions or pieces of legislation, and legal databases and websites also do not provide specific provisions when searching for key terms. Irrespective of this, a legislative corpus that does automatically provide the particular provision would be an incredibly useful tool for legal scholars.

The final limitation is that the corpus only represents the law at a specified point in time, rather than acting as a monitor corpus that is frequently updated. This may pose an issue for currency, as the law is updated regularly. However, there is still benefit in understanding what legal language is used at a specified point in time, with the potential for future uses involving diachronic analysis. For instance, even in relation to the example, it would be interesting to determine how "gender" is used across legislation five or ten years from now, compared to an assessment conducted with the ALC23.

8. Conclusion

In this article, I have introduced the legislative corpus of the ALC23, and related subcorpora. In detailing the process of building this corpus, I have presented a different approach to building general legislative corpora, particularly as it relates to overcoming issues. I also emphasise the need for this corpus through noting potential applications associated with statistical processing. The particular example I provide relates to a brief examination of the word "gender". I also provide some potential limitations in using the ALC23 that may be of use to both legal scholars and corpus linguistics.

It is my hope that this article assists with encouraging other legal scholars to engage more closely with corpus linguistic techniques, particularly if these scholars do not have any background in this methodology. Further, I also expect the article will provide some insight for corpus linguists as to how their legislative corpora may be more of use to legal scholars. After all, like any interdisciplinary field, corpus linguistic applications to the law can benefit greatly from alternative perspectives, those of which can ideally benefit both fields individually too.

Acknowledgements

I would like to sincerely thank Assoc Prof Annabelle Lukin, Dr Honni van Rijswijk, Prof Penny Crofts and Dr Alexandra Grey for the assistance received while writing this paper. I would also like to acknowledge the advice provided by Dr Isabella Alexander, and Dr Monika Bednarek and the Sydney Corpus Lab group.

References

Anthony, Lawrence (n.d.). *AntConc4*. Retrieved September 20, 2022. Available at laurenceanthony.net/software/antconc/ (accessed 7 Jan 2025).

Australian Capital Territory Government (n.d.-a). About the register — An overview. *ACT Legislation Register*. Retrieved June 30, 2023. Available at legislation.act.gov.au/Static/Help/About/about_the_register.html#3 (accessed 7 Jan 2025).

Australian Capital Territory Government (n.d.-b). *ACT Legislation Register*. Available at legislation.act.gov.au/ (accessed 7 Jan 2025).

Australian Government (n.d.). Terms governing the use of this website. *Federal Register of Legislation*. Available at legislation.gov.au/terms-of-use (accessed 7 Jan 2025).

Baker, Paul (2004). Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, *32*(4), 346–359. DOI: 10.1177/0075424204269894.

Baker, Paul (2005). *Public Discourses of Gay Men*. London: Routledge.

Baker, Paul (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, Paul (2008). *Sexed Texts: Language, Gender and Sexuality*. London: Equinox Pub.

Baker, Paul (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh; Edinburgh University Press.

Baker, Paul (2014). *Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, Paul (2023). *Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, Paul & McEnery, Tony (2015). Introduction. In Baker & McEnery (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora* (pp. 1–19). New York: Palgrave Macmillan.

Barnard, David T.; Burnard, Lou & Sperberg-McQueen, C. Michael (1996). Lessons learned from using SGML in the Text Encoding Initiative. *Computer Standards & Interfaces*, 18(1), 3–10. DOI: 10.1016/0920-5489(95)00035-6.

Berk-Seligson, Susan (2012). Linguistic issues in courtroom interpretation. In Tiersma & Solan (Eds.), *The Oxford Handbook of Language and Law* (pp. 421–434). Oxford: University Press.

Berūkštienė, Donata (2018). A corpus-driven analysis of structural types of lexical bundles in court judgments in English and their translation into Lithuanian. *Kalbotyra*, 70(70), 7–31. DOI: 10.15388/Klbt.2017.11181.

Bhatia, Vijay; Langton, Nicola M. & Lung, Jane (2004). Legal discourse: Opportunities and threats for corpus linguistics. In Connor & Upton (Eds.), *Studies in Corpus Linguistics* (pp. 203–231). Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/scl.16.09bha.

Biber, Douglas (2008). Representativeness in corpus design. In Fontenelle (Ed.), *Practical Lexicography: A Reader* (pp. 63–87).

Breeze, Ruth (2017). Corpora and computation in teaching law and language. *International Journal of Language & Law*, 6, 1–17. DOI: 10.14762/JLL.2017.001.

Breeze, Ruth (2019). Part-of-speech patterns in legal genres: Text-internal dynamics from a corpus-based perspective. In Fanego & Rodríguez-Puente (Eds.), *Studies in Corpus Linguistics* (pp. 79–103). Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/scl.91.04bre.

Brezina, Vaclav; McEnery, Tony & Wattam, Stephen (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 139–173. DOI: 10.1075/ijcl.20.2.01bre.

Butler, Umar (n.d.-a). Open Australian Legal Corpus. *Datasets*. Available at huggingface.co/datasets/umar-butler/open-australian-legal-corpus (accessed 7 Jan 2025).

Butler, Umar (n.d.-b). Open Australian Legal Corpus Creator. *Open Australian Legal Corpus Creator*. Available at umarbutler/open-australian-legal-corpus-creator (accessed 7 Jan 2025).

Butler, Umar (2023, October 28). How I built the largest open database of Australian law. *Umar Butler*. Available at umarbutler.com/how-i-built-the-largest-open-database-of-australian-law/ (accessed 7 Jan 2025).

Cock, Barbara De; Hambye, Philippe & Pedraza, Andrea Pizarro (2024). Annotation and mark up for representation analysis. In Heritage & Taylor, *Analysing Representation* (pp. 84–99). London: Routledge. DOI: 10.4324/9781003350972-6.

Conley, John M. & O'Barr, William M. (2005). *Just Words: Law, Language, and Power* (2nd ed). Chicago: University of Chicago Press.

Danet, Brenda (1980). Language in the legal process. *Law and Society Review*, *14*(3), 445–564.

Dekkers, Makx & Weibel, Stuart (2003). State of the Dublin core metadata initiative. *D-Lib Magazine*, *9*(4). DOI: 10.1045/april2003-weibel.

Egbert, Jesse & Römer-Barron, Ute (2024). Applying corpus linguistics to the law. *Applied Corpus Linguistics*, *4*(2), 100093. DOI: 10.1016/j.acorp.2024.100093.

Egbert, Jesse & Wood, Margaret (2023). The corpus of United States state statutes — Design, construction and use. *Applied Corpus Linguistics*, *3*(2), 100047. DOI: 10.1016/j.acorp.2023.100047.

Freeman, Michael D. A. & Smith, Fiona (2013). Law and language: An introduction. In Freeman & Smith (Eds.), *Law and Language* (pp. 1–7). Oxford: University Press.

Galdia, Marcus (2023). Researching the language of law. In Wagner & Matulewska (Eds.), *Research Handbook on Jurilinguistics* (pp. 17–34). Edward Elgar Publishing. DOI: 10.4337/9781802207248.00009.

Genovese, Emma. (2023). The spectacle of respectable equality: Queer discrimination in Australian law post marriage equality. *University of New South Wales Law Journal*, *46*(2). DOI: 10.53637/NAFG1780.

Gillings, Mathew (2022). How to use corpus linguistics in forensic linguistics? In O'Keeffe & McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 589–601). London: Routledge. DOI: 10.4324/9780367076399.

Gillings, Mathew; Mautner, Gerlinde & Baker, Paul (2023). *Corpus-Assisted Discourse Studies* (1st ed.). Cambridge University Press. DOI: 10.1017/9781009168144.

Goldfarb, Neal (2021). The use of corpus linguistics in legal interpretation. *Annual Review of Linguistics*, *7*(1), 473–491. DOI: 10.1146/annurev-linguistics-050520-093942.

Goodrich, Peter (1987). *Legal Discourse: Studies in Linguistics, Rhetoric and Legal Analysis*. Basingstoke: Macmillan.

Government of South Australia (n.d.-a). A-Z Acts. *South Australian Legislation*. Available at legislation.sa.gov.au/about-this-site/legislation-available-on-this-website/about-this-website (accessed 7 Jan 2025).

Government of South Australia (n.d.-b). A-Z Regulations and Rules. *South Australian Legislation*. Available at legislation.sa.gov.au/about-this-site/legislation-available-on-this-website/regulations-and-rules (accessed 7 Jan 2025).

Government of South Australia (n.d.-c). Copyright. *South Australian Legislation*. Available at legislation.sa.gov.au/copyright (accessed 7 Jan 2025).

Government of Western Australia (n.d.). Copyright and licence. *Western Australian Legislation*. Available at legislation.wa.gov.au/legislation/statutes.nsf/copyright.html (accessed 7 Jan 2025).

Government of Western Australia, Department of Justice, & Parliamentary Counsel's Office (n.d.). *Acts in Force*. Available at legislation.wa.gov.au/legislation/statutes.nsf/actsif_info.html (accessed 7 Jan 2025).

Goźdź-Roszkowski, Stanislaw (2011). *Patterns of Linguistic Variation in American Legal English: A Corpus-Based Study* (1st, New ed ed.). Frankfurt: Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Goźdź-Roszkowski, Stanislaw (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *34*(5), 1515–1540. DOI: 10.1007/s11196-021-09860-8.

Goźdź-Roszkowski, Stanislaw (2023). Corpus linguistics, methodology of jurilinguistics. In Wagner & Matulewska (Eds.), *Research Handbook on Jurilinguistics* (pp. 103–115). Edward Elgar Publishing. DOI: 10.4337/9781802207248.00014.

Grey, Alexandra, & Severin, Alyssa A. (2021). An audit of NSW legislation and policy on the government's public communications in languages other than English. *Griffith Law Review*, *30*(1), 122–147. DOI: 10.1080/10383441.2021.1970873.

Gries, Stefan T. (2021). Corpus linguistics and the law: Extending the field from a statistical perspective. *Brooklyn Law Review*, *86*, 321–356.

Grover, Claire; Hachey, Ben & Hughson, Ian (2004). *The HOLJ Corpus: Supporting Summarisation of Legal Texts*.

Hafner, Christoph A. & Candlin, Christopher N. (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, *6*(4), 303–318. DOI: 10.1016/j.jeap.2007.09.005.

Hafner, Christoph A. & Wang, Simon Ho (2018). Hong Kong learner corpus of legal academic writing in English: A study of boosters as a marked language form in an English-Medium instruction context. *TESOL Quarterly*, *52*(3), 680–691. DOI: 10.1002/tesq.451.

Haigh, Richard (2018). Thirty years with section 15 of the charter: A report on legislative terminology in Canada. *National Journal of Constitutional Law*, *38*(1), 7–34.

Hart, Herbert Lionel Adolphus (1994). *The Concept of Law* (2nd ed). Oxford, United Kingdom: Clarendon Press/Oxford University Press.

Hildebrandt, Mireille (2018). Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. *University of Toronto Law Journal*, *68*(1), 12–35.DOI: 10.3138/utlj.2017-0044.

Höfler, Stefan & Piotrowski, Michael (2011). Building corpora for the philological study of swiss legal texts. *Journal for Language Technology and Computational Linguistics*, *26*(2), 77–89. DOI: 10.21248/jlcl.26.2011.148.

Höfler, Stefan & Sugisaki, Kyoko (2014). *Constructing and Exploiting an Automatically Annotated Resource of Legislative Texts*. DOI: 10.5167/UZH-96172.

Hu, Ming; Hu, Xitao & Cheng, Le (2021). Exploring digital economy: A sociosemiotic perspective. *International Journal of Legal Discourse*, *6*(2), 181–202. DOI: 10.1515/ijld-2021-2053.

Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel & Tugwell, David (2014). The sketch engine. *Lexicography*, *1*, 7–36.

Laske, Caroline (2022). Corpus linguistics: The digital tool kit for analysing language and the law. *Comparative Legal History*, *10*(1), 3–32. DOI: 10.1080/2049677X.2022.2063510.

Leung, Janny H. & Durant, Alan (2018). Editors' Introduction. In Leung & Durant (Eds.), *Meaning and Power in the Language of Law* (pp. 1–16). Cambridge University Press. DOI: 10.1017/9781316285756.001.

Lexical Analysis Software & Oxford University Press (n.d.). *WordSmith Tools*. Available at lexically.net/wordsmith/> (accessed 7 Jan 2025).

Lexical Computing (n.d.-a). *POS tags*. Available at sketchengine.eu/blog/pos-tags/ (accessed 7 Jan 2025).

Lexical Computing (n.d.-b). Sketch Engine Boot Camp. *Sketch Engine*. Available at sketchengine.eu/bootcamp/ (accessed 7 Jan 2025).

Lexical Computing (n.d.-c). *Sketch Engine*. Available at sketchengine.eu (accessed 7 Jan 2025).

Lexical Computing (n.d.-d). *Word Sketch—Collocations and Word Combinations*. Retrieved August 6, 2024. Available at sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/#toggle-id-7 (accessed 7 Jan 2025).

Lexical Computing (2015). *Statistics Used in the Sketch Engine*. Available at sketchengine.eu/wp-content/uploads/ske-statistics.pdf (accessed 7 Jan 2025).

Lexical Computing (2021). *Why are Some Words in the Corpus Tagged Incorrectly?* Available at support.sketchengine.eu/help/en-us/5-tags-lemmas-taggers-vertical-file/60-why-are-some-words-in-the-corpus-tagged-incorrectly (accessed 7 Jan 2025).

Lukin, Annabelle & Araujo E Castro, Rodrigo (2022). The macquarie laws of war corpus (MQLWC): Design, construction and use. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *35*(5), 2167–2186. DOI: 10.1007/s11196-022-09889-3.

Lukin, Annabelle & García Marrugo, Alexandra García (2024). The 'existential fabric' of war: Explaining the phrase of war in the laws of war. *Applied Linguistics*, amae027. DOI: 10.1093/applin/amae027.

Lukin, Annabelle & Marrugo, Alexandra García (2023). The international laws of war: Linguistic analysis from the perspectives of register, corpus and grammatical patterning. *Journal of International Humanitarian Legal Studies*, *14*(2), 223–249. DOI: 10.1163/18781527-bja10065.

Mautner, Gerlinde (2022). What can a corpus tell us about discourse? In O'Keeffe & McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 250–262). London: Routledge. DOI: 10.4324/9780367076399.

McEnery, Tony & Brookes, Gavin (2022). Building a written corpus: What are the basics? In O'Keeffe & McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 35–47). London: Routledge. DOI: 10.4324/9780367076399.

McEnery, Tony; Xiao, Richard & Tono, Yukio (2010). *Corpus-Based Language Studies: An Advanced Resource Book* (Reprinted). London: Routledge.

Mellinkoff, David (1963). *The Language of the Law*. Eugene, Oregon: Resource Publications.

Moran, Leslie J. (1996). *The Homosexual(ity) of Law*. London: Routledge.

Mouritsen, Stephen (2017). Corpus linguistics in legal interpretation. An evolving interpretative framework. *International Journal of Language & Law (JLL)*, *6*, 67–89 . DOI: 10.14762/JLL.2017.067.

New South Wales Government (n.d.). *In Force Legislation*. Retrieved June 30, 2023. Available at legislation.nsw.gov.au/browse/inforce (accessed 7 Jan 2025).

New South Wales Government. (2021, September 10). Copyright. *NSW Legislation*. Available at legislation.nsw.gov.au/copyright (accessed 7 Jan 2025).

Northern Territory Government (n.d.-a). Help. *Northern Territory Legislation*. Retrieved June 30, 2023. Available at legislation.nt.gov.au/Footer/Help (accessed 7 Jan 2025).

Northern Territory Government (n.d.-b). Terms of Use. *Northern Territory Legislation*. Available at legislation.nt.gov.au/Footer/Terms-of-Use (accessed 7 Jan 2025).

Okawara, Mami Hiraike (2012). Courtroom Discourse in Japan's New Judicial Order. In Tiersma & Solan (Eds.), *The Oxford Handbook of Language and Law* (pp. 381–394). Oxford: University Press.

Onesti, Cristina (2011). Methodology for building a text-structure oriented legal corpus. *Comparative Legilinguistics*, *8*.

Östling, Andreas; Sargeant, Holli; Xie, Huiyuan; Bull, Ludwig; Terenin, Alexander; Jonsson, Leif; Magnusson, Måns & Steffek, Felix (2024). The Cambridge law corpus: A dataset for legal AI research. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4763429.

Pei, Jiamin & Li, Jian (2018). A corpus-based investigation of modal verbs in Chinese civil-commercial legislation and its English versions. *International Journal of Legal Discourse*, *3*(1), 77–102. DOI: 10.1515/ijld-2018-2003.

Pérez-Paredes, Pascual; Jiménez, Pilar Aguado & Hernández, Purificación Sánchez (2017). Constructing immigrants in UK legislation and administration informative texts: A corpus-driven study (2007–2011). *Discourse & Society*, *28*(1), 81–103. DOI: 10.1177/0957926516676700.

Phillips, James C. & Egbert, Jesse (2017). Advancing law and corpus linguistics: Importing principles and practices from survey and content analysis methodologies to improve corpus design and analysis. *Brigham Young University Law Review*, 1589–1619.

Pontrandolfo, Gianluca (2012). Legal corpora: An overview. *Rivista Internazionale Di Technica Della Traduzione*, *14*, 121–136.

Pontrandolfo, Gianluca (2019). Corpus methods in legal translation studies. In Biel, Engberg, Martín Ruano & Sosoni (Eds.), *Research Methods in Legal Translation and Interpreting: Crossing Methodological Boundaries* (pp. 13–28). Routledge. DOI: 10.4324/9781351031226.

Queensland Government (2020). Copyright. *Queensland Legislation*. Available at legislation.qld.gov.au/copyright (accessed 7 Jan 2025).

Rea Rizzo, Camino & Marín Pérez, M. José (2012). Structure and design of the British Law Report Corpus (BLRC): A legal corpus of judicial decisions from the UK. *Journal of English Studies*, *10*, 131. DOI: 10.18172/jes.184.

Römer-Barron, Ute & Cunningham, Clark D. (2024). Applied corpus linguistics and legal interpretation: A rapidly developing field of interdisciplinary scholarship. *Applied Corpus Linguistics*, 4(1), 100080. DOI: 10.1016/j.acorp.2023.100080.

Rossini-Favretti, Rema (1998). *Using Multilingual Parallel Corpora for the Analysis of Legal Language: The Bononia Legal Corpus* (W. Teubert, E. Tognini Bonelli, & N. Volz, Eds.; pp. 57–68). TELRI Association.

Rychlý, Pavel (2008). A lexicographer-friendly association score. In Sojka & Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing* (pp. 6–9). Czechia.

Salembier, J. Paul (2018). *Legal and Legislative Drafting* (Second edition). LexisNexis.

Sinclair, John McHardy (1991). *Corpus, Concordance, Collocation*. Oxford: University Press.

Sketch Engine (n.d.). *Word Sketch*. Retrieved August 19, 2023. Available at sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/ (accessed 7 Jan 2025).

Solan, Lawrence M. & Gales, Tammy (2017). Corpus linguistics as a tool in legal interpretation. *Brigham Young University Law Review*, 1311–1357.

Sole-Mauri, Francina; Sánchez-Gijón, Pilar & Oliver, Antoni (2021). Cadlaws – An English–French parallel corpus of legally equivalent documents. *Mutatis Mutandis. Revista Latinoamericana de Traducción*, *14*(2), 494–508. DOI: 10.17533/udea.mut.v14n2a10.

South Australian Law Reform Institute (2015). *Discrimination on the Grounds of Sexual Orientation, Gender, Gender Identity and Intersex Status in South Australian Legislation* [Audit Report]. Available at law.adelaide.edu.au/system/files/media/documents/2019-01/audit_report_lgbtiq_sept_2015.pdf (accessed 7 Jan 2025).

Stubbs, Michael (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Blackwell Publishers.

Stückler, Andreas (2018). Legislation and discourse: Research on the making of law by means of discourse analysis. In Keller, Hornidge & Schünemann, *The Sociology of Knowledge Approach to Discourse* (pp. 112–132). London: Routledge. DOI: 10.4324/9781315170008-6.

Stygall, Gail (2012). Discourse in the US Courtroom. In Tiersma & Solan (Eds.), *The Oxford Handbook of Language and Law* (pp. 369–380). Oxford: University Press.

Tasmanian Government (2021, March 9). Legislation. *Tasmanian Legislation*. Available at legislation.tas.gov.au/about/legislation (accessed 7 Jan 2025).

Tasmanian Government (2023, February 12). Copyright notice. *Tasmanian Legislation*. Available at legislation.tas.gov.au/copyrightanddisclaimer (accessed 7 Jan 2025).

Tiersma, Peter Meijes (1999). *Legal Language*. Chicago: University of Chicago Press.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company.

Tufiș, Dan; Mitrofan, Maria; Păiș, Vasile; Ion, Radu & Coman, Andrei (2020). *Collection and Annotation of the Romanian Legal Corpus*.

Victorian Government. (2020, February 24). Copyright. *Victorian Legislation*. Available at legislation.vic.gov.au/copyright (accessed 7 Jan 2025).

Vogel, Friedemann (2017). Calculating legal meanings? Drawbacks and opportunities of corpus-assisted legal linguistics to make the law (more) explicit. In Giltrow & Stein (Eds.), *The Pragmatic Turn in Law* (pp. 287–306). De Gruyter. DOI: 10.1515/9781501504723-012.

Vogel, Friedemann; Hamann, Hanjo & Gauer, Isabelle (2018). Computer-assisted legal linguistics: Corpus analysis as a new tool for legal studies. *Law & Social Inquiry*, *43*(4), 1340–133. DOI: 10.1111/lsi.12305.

Wagner, Anne & Matulewska, Aleksandra (Eds.). (2023). *Research Handbook on Jurilinguistics*. Edward Elgar Publishing.

Williams, Christopher (2005). *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*. Bern: P. Lang.

Woodbury, Hanni (1984). The strategic use of questions in court. *Semiotica*, *48*(3–4). DOI: 10.1515/semi.1984.48.3-4.197.

Woolls, David & Coulthard, Malcolm (1998). Tools for the trade. *Forensic Linguistics*, 5(1), 33–57. DOI: 10.1558/sll.1998.5.1.33.