ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO DI LAUREA MAGISTRALE IN SPECIALIZED TRANSLATION (LM-94)

CORPUS CONSTRUCTION AND ANNOTATION OF LABOUR LAW TEXTS: METHODOLOGY AND COMPARATIVE CASE STUDY ON KENYA AND TANZANIA

Tesi di laurea magistrale in

Corpus Linguistics

Relatrice

Prof.ssa Silvia Bernardini

Correlatrice

Prof.ssa Beatrice Ragazzini

Presentata da

Verdiana Crociani

Sessione 07/2025

Anno Accademico 2024/2025

Abstract

This dissertation develops and tests a methodology for building and annotating a legal corpus of labour law texts. Its primary aim is to integrate a corpus-based discourse analytical approach into the broader Worlds of Labour (WoL) project, a research initiative that investigates legal segmentation in labour law systems, with particular attention to countries of the Global South, adopting the methodological approach of leximetrics. The annotation framework was designed to reflect the structural features of legislative texts and is intended to serve as a resource for future interdisciplinary research in law and language. To test the analytical potential of this methodology, a case study was conducted combining leximetrics and Corpus-Assisted Discourse Studies (CADS) to explore the representation of female labour in legal texts of two East African countries, Kenya and Tanzania, between 1902 and 2011. Drawing on the WoL database and its coding infrastructure, two annotated sub-corpora were compiled for each country: one focused on general labour legislation (WoL-K1, WoL-T1), and one on legal texts specifically concerning women's labour (WoL-K2, WoL-T2). These corpora were analysed using Sketch Engine to uncover discursive dimensions of legal segmentation in recurring expressions, structural patterns, and narrative strategies concerning women and work. The findings show that, within the selected texts, women are consistently depicted as passive subjects, typically framed in terms of vulnerability, protection, or dependency, rather than agency and participation. The study highlights colonial patterns in the marginalisation of female labour and stresses the value of discourse analysis in legal research. It concludes by offering the proposed methodology as a replicable tool for future interdisciplinary work on legal discourse, to complement quantitative leximetric approaches with qualitative insights.

Riassunto

La presente tesi ha l'obiettivo di sviluppare e testare un metodo per la costruzione e l'annotazione di corpora legali composti da testi legislativi in materia di diritto del lavoro. L'obiettivo principale è quello di integrare un approccio analitico di tipo *corpus-based* all'interno del progetto di ricerca di *Worlds of Labour* (WoL), un'iniziativa che studia la segmentazione legale nei sistemi giuridici, con particolare attenzione ai Paesi del Sud Globale, attraverso la metodologia *leximetrics*. Il modello di annotazione proposto è stato progettato per riflettere le caratteristiche strutturali dei testi legislativi, con l'intento di costituire una risorsa utile per futuri studi interdisciplinari nell'ambito

del diritto e della linguistica. A supporto di tale modello, è stato condotto uno studio di caso che combina l'approccio leximetrico con i Corpus-Assisted Discourse Studies (CADS), con lo scopo di analizzare come il lavoro femminile venga rappresentato nei testi giuridici di due Paesi dell'Africa orientale, Kenya e Tanzania, nel periodo compreso tra il 1902 e il 2011. A partire dal database e dal sistema di codifica del progetto WoL, sono stati creati e annotati quattro sub-corpora: due relativi alla legislazione generale del lavoro (WoL-K1 e WoL-T1) e due specificamente dedicati alle normative sul lavoro femminile (WoL-K2 e WoL-T2). L'analisi dei corpora, condotta attraverso la piattaforma Sketch Engine, ha permesso di far emergere le dimensioni discorsive della segmentazione legale tramite l'individuazione di espressioni ricorrenti e strategie narrative associate al lavoro e alla figura femminile. I risultati evidenziano una rappresentazione ricorrente delle donne come soggetti passivi, prevalentemente associate a concetti di vulnerabilità, protezione o dipendenza, piuttosto che a forme di autonomia o partecipazione attiva. Lo studio mette inoltre in luce tratti discorsivi riconducibili al contesto coloniale, evidenziando come la marginalizzazione del lavoro femminile abbia radici storiche profonde. Infine, la presente tesi propone la metodologia elaborata come uno strumento replicabile per future ricerche interdisciplinari sul discorso giuridico, che mirino a integrare approcci quantitativi, come quello leximetrico, con analisi qualitative di tipo discorsivo.

Table of contents

1 Introduction

This dissertation aims to develop a methodological approach for building and annotating a corpus of labour law texts. To illustrate its practical usability and analytical potential, the method is then used in a case study analysing the discursive representation of female labour in legislative texts in two East African countries, Kenya and Tanzania, covering the period 1902-2011.

The work is the result of a collaboration with participants of the *Worlds of Labour* (WoL) project, a research initiative at the *Global Dynamics and Social Policy* research centre of the University of Bremen[1] that investigates the phenomenon of legal segmentation, i.e., how labour law contributes to economic and social inequalities through an unequal treatment of different categories of workers (Dingeldey et al., 2020). WoL aims to historically trace its development, from the late 19th century to the present, and its global diffusion across legal systems, with particular attention placed on the countries of the Global South (Carlino et al., forth.).

More specifically, this thesis stems from work carried out during the *Worlds of Labour Corpus Project* (WoLCP), an initiative involving members of the University of Bologna and the University of Bremen.[2] The WoLCP aimed to integrate corpus-based linguistic methodologies into the existing legal framework of WoL by constructing and annotating a searchable legal corpus of labour law texts, thereby making legal data more accessible and suitable for discursive analysis. To operationalise the concept of legal segmentation, WoL adopts a leximetric approach, a quantitative method that seeks to measure and compare legal systems by transforming the qualitative content of laws into numerical data according to predefined criteria, with the goal of assessing the intensity, scope, or structure of regulation in a systematic way (Konstantin et al., 2021). Within the WoL project, numerical values are assigned to legal provisions based on their potential to produce segmentation.

[1] Globale Entwicklungsdynamiken von Sozialpolitik. SFB 1342, *Worlds of Labour: Coverage and Generosity of Employment Law*. https://www.socialpolicydynamics.de/projects/project-area-a-global-dynamics/project-a03-2022-25- [last accessed: 25 June 2025]. The project team members, listed in alphabetical order: Marina Carlino, Prof. Dr. Irene Dingeldey, Dr. Heiner Fechner, Prof. Dr. Ulrich Mückenberger, Andrea Schäfer.

[2] The *Worlds of Labour Corpus Project* (WoLCP) emerged within the framework of an academic internship organised by the Department of Interpretation and Translation (DIT) at the University of Bologna – Forlì Campus. Professor Silvia Bernardini facilitated contact with the research team based in Bremen, enabling six students enrolled in the master's degree in Specialized Translation to take part in the project. This experience constitutes a fundamental basis for the development of the present master's thesis, which aims to explore a specific aspect of the project through an independent, methodologically oriented analysis.

Building on this broad approach, the present contribution aims to develop a method for constructing and annotating corpora of legislative texts which might complement the leximetric quantitative method with a corpus-based discourse perspective (Partington et al., 2013). It further seeks to apply this method to a case study on the representation of female labour in legal documents from Kenya and Tanzania. In this way, this dissertation attempts to broad the analytical scope of labour law research, allowing legal texts to be examined not only through coded legal content, but also through the language in which segmentation is articulated. In other words, it seeks to extend the WoL framework by exploring whether and how legal segmentation can be identified and analysed discursively. The method proposed here can also be intended as a flexible and transferable tool for future studies with similar aims. Even though the case study focuses on gender and labour law, the approach could be adapted and expanded for the analysis of other types of legal documents. Its potential makes it a useful starting point for further research in the field of legal discourse.

The methodology applied here builds directly on the tools and resources developed within the WoLCP. All texts selected for analysis are drawn from the WoL dataset, already been structured and coded by the WoL research team. Based on this dataset, a set of legislative texts from Kenya and Tanzania is further processed to create four sub-corpora, two for each country: one containing general labour legislation and one focusing specifically on legislation related to women's labour. These sub-corpora are annotated and subsequently analysed using Sketch Engine.[3] The case study opens with the investigation of the general texts, followed by those with a specific focus on female labour. The analysis combines quantitative and qualitative techniques, starting with keyword extraction to identify statistically significant terms, followed by the examination of concordance lines, collocational patterns, and results from the Word Sketch and Thesaurus tools.[4] The corpora are examined to find discursive patterns and repeated ways of referring to women and their roles in the workplace.

[3] Sketch Engine. https://www.sketchengine.eu/ [last accessed: 25 June 2025]. Founders: Adam Kilgarriff and Pavel Rychlý.

[4] For more information on the Word Sketch and Thesaurus tools, see the official Sketch Engine online guide. Word Sketch: https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words/#toggle-id-2 [last accessed: 25 June 2025]. Thesaurus: https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words/ [last accessed: 25 June 2025]. An explanation of how these tools function and are applied in this research is also provided in Chapter 7.

This dissertation is structured as follows. Chapter 2 introduces the theoretical framework of comparative law and comparative labour law, laying the conceptual foundation for the subsequent analysis. Chapter 3 presents the *Worlds of Labour* (WoL) project, offering an overview of its objectives, its leximetric methodology, and illustrating the concepts of market labour segmentation and legal segmentation. Chapter 4 offers a contextual background on the historical, legal, and gendered dimensions of labour in Kenya and Tanzania. It focuses on how colonial legal systems have constructed and marginalised women's work, and it explains the rationale behind the country selection for the case study. Chapter 5 lays out the theoretical and practical foundations for corpus construction and annotation in the legal domain. It discusses key features of legal discourse and reviews annotation issues relevant to legal corpora. Chapter 6 describes the methodology of the *Worlds of Labour Corpus Project* (WoLCP) on text selection, pre-processing, annotation strategies, metadata design, and tagging conventions. It illustrates the importance of the broader WoLCP methodology for the Kenyan and Tanzanian sub-corpora analysed in this dissertation. Chapter 7 presents the case study that investigates how women's labour is discursively represented in labour legislation from Kenya and Tanzania (1902–2011). It introduces the four sub-corpora, detailing their construction and annotation, and illustrates how Sketch Engine is used to identify recurring patterns of gendered representation. The results are then interpreted comparatively. The chapter concludes by discussing the main limitations of the methodological framework itself and by outlining possible directions for future research. Finally, Chapter 8 revisits the overall aims of the dissertation, summarises the key outcomes, and reflects once again on the constraints of the case study and how the proposed methodology could be further developed and refined in future work

2 The Comparative Method and Labour Law in a Global Context

2.1 Introduction

This chapter explores the crucial role of comparative legal analysis in understanding labour law in a globalised context. Because the present dissertation engages with legislative texts from both Kenya and Tanzania, a comparative perspective is essential.

The chapter is divided into two main parts. The first one investigates the theoretical foundations of comparative law, which is defined as the study of different legal systems whose goal is to understand how legal concepts, rules, and institutions operate in various jurisdictions (Samuel, 1998). More than just comparing laws, comparative law involves the analysis of the underlying structures of legal knowledge to gain insights into how law functions across different cultures and traditions (*ibid.*). In other words, it offers interpretive tools to place legal systems within broader global paths while also acknowledging their specific historical and cultural particularities. With particular attention to its role as translator across cultural and normative worlds, Section 2.2 illustrates how comparative law is an interpretive exercise that requires reflexivity and a deep awareness of context (Curran, 2019; Finkin, 2019; Finkin and Mundlak, 2015).

The second part focuses on the field of comparative labour law. As Deakin *et al.* (2007) explain, comparative labour law can be defined as the study of labour law systems across different jurisdictions that aims to understand how laws regulating employment relationships evolve over time, how they differ across legal traditions, and how these differences interact with broader economic, political, and institutional contexts. It shows how labour law differs between legal systems, such as between civil law and common law traditions, and questions explanations that see legal development as fixed or automatic (*ibid.*). Instead, it highlights how history, geographical context, and practical solutions can shape labour law in different ways (*ibid.*). In this sense, comparative labour law helps us better understand how legislation influences work and society in an increasingly globalised world. Section 2.3 and its related subsections will present the contemporary challenges facing labour law systems worldwide, its methodological approaches and its sources. These reflections lay the groundwork for the case study analysis that will follow in the second part of the thesis (Chapter 7).

Finally, it is important to note that the present chapter is closely linked, both conceptually and methodologically, to Chapter 3 and Chapter 4. Chapter 3 presents the *Worlds of Labour* (WoL)

project developed at the University of Bremen, its aims and methodological approach. This overview is essential for situating the present thesis within the broader research context, as WoL provides the conceptual foundation and the data on which this work builds. At the same time, WoL can be seen as a concrete example of comparative labour law in practice, as it seeks to examine how legal systems around the world contribute to the phenomenon of legal segmentation (see Introduction or Chapter 3 for a definition). Chapter 4 complements the previous two by offering a historical a socio-legal overview of gendered work in Kenya and Tanzania during and after the colonial period. Although it does not claim to provide an exhaustive or definitive account of a such complex issue, Chapter 4 provides a necessary bridge for the subsequent discursive analysis of selected sub-corpora. Together, these three chapters seek to offer the basic theoretical framework required to understand how law mediates work, power, and gender inequality across borders.

2.2 Comparative law as translator

Comparative law has matured from a primarily academic and theoretical discipline into an essential interpretive paradigm for understanding the transformations of legal systems in a rapidly globalising world (Curran, 2019). Originally emerging within the academic sphere as an intellectual exercise focused on exploring foreign legal systems, identifying general principles of law, and classifying legal models, it functioned more as a critical, almost philosophical reflection on law than as a tool for normative intervention (*ibid.*). Over time, however, comparative law has gained growing relevance beyond academia, especially as legal issues increasingly transcend national borders and legal systems become ever more intertwined through transnational regulation, international institutions, and global value chains (*ibid.*). National courts are frequently called upon to interpret or apply foreign and international legal sources (*ibid.*). In this evolving context, comparison has become a necessary method for interpreting legal change, identifying shared challenges, and navigating complex global pressures such as international trade, digitalisation, migration, and shifting social norms. The comparative dimension of law is thus not only relevant, but inevitable (*ibid.*). According to Finkin (2019: 8), comparative description should be "undertaken for the illumination it provides", to challenge the reader's habitual ways of thinking and to demonstrate that other legal systems may offer alternative, perhaps even superior, solutions to shared economic and social problems. Hence, the comparative method is a useful tool for questioning assumptions, disrupting normative expectations, and revealing hidden potential in the law.

A central debate in comparative law concerns the very possibility and limits of comparison between legal systems. Much like in linguistics, where scholars have long discussed whether languages share universal structures or are shaped by irreconcilable cultural differences, comparative legal theorists disagree on the extent to which legal systems are mutually intelligible (Curran, 2019). On the one hand, comparative analysis is often seen as a way to identify successful legal solutions and share them across different legal systems through a process known as transplantation (*ibid.*). The goal is to improve legal frameworks by learning from what works well elsewhere. However, this means assuming that legal arrangements are transferable, which can be problematic as it overlooks the specific social and cultural contexts in which legal norms operate, since laws cannot always be passed on directly from one system to another without significant adaptation (Finkin and Mundlak, 2015). Nonetheless, "it is rare to see recommendations to simply copy and paste arrangements of one country on to the legal system of another" (*ibid.*: 4). Instead, rather than offering direct solutions, comparative law proves most valuable as a method of analysis. By examining legal developments across time and space, comparative analysis uncovers alternative institutional arrangements and illuminates different trajectories of evolution. This process can reveal which elements of legal development are more universal and which are more heavily influenced by local circumstances (*ibid.*).

The analogy between law and language goes in fact beyond a superficial resemblance, since both are culturally embedded systems of signs that mediate between meaning and practice, and both evolve according to complex internal and external dynamics. As Curran (2019: 1) suggests, the "study of language is a cognitive model for comparative law". This unavoidably raises the issue of accessibility: many comparatists are not fluent in the languages of the systems they study, and translations frequently have semantic limitations. Language is never neutral, and legal language is tightly bound to what Curran calls the "inner grammar of legal systems, cultures, and mentalities" (2019: 4-5). For this reason, comparative law cannot be reduced to a mere comparison of equivalent terms. Instead, it should be seen as a complex process of translation between worlds, an act of interpretation that seeks to uncover both similarities and divergences between legal systems. The comparative jurist must navigate spaces where meanings are contextually situated and culturally loaded (*ibid.*: 5): "comparative law's aim is to clarify and communicate the foreign as capably as possible, and not to succumb to the 'domination of words over things'" (*ibid.*: 2).

In this light, comparative law must also account for what is unspoken, implicit, or taken for granted within a legal system: "The polyglot legal comparatist knows that legal orders reside as much beneath and aside from words as they do in the words that purport to embody them" (*ibid.*: 6). Their task, then, is to expose hidden alterities and even to recognize that sameness itself sometimes requires translation, because even what appears familiar may conceal otherness. Therefore, comparison is less about finding direct equivalence and more about tracing similes: approximate correspondences that preserve difference while making it legible (Curran, 2019). From this perspective, comparative legal analysis functions as a lantern, casting light on otherness, both in the foreign and in the familiar. As Curran (2019: 17) observes, "[t]ranslating the foreign into the familiar ends by clarifying the familiar that one discovers also to be foreign". Inevitably, the act of translation lets jurists uncover the peculiarity of their own legal culture, challenging the assumption that legal meaning is never fixed or entirely domestic. Like language, legal systems are dynamic, embedded in historical, cultural, and social contexts, and often perceived most clearly when seen through the prism of external perspectives (Curran, 2019). This insight is fundamental to understand the epistemological significance of comparative law, which is in fact not merely about foreign legal systems, but rather about the process of revealing the conditions under which legal meaning is produced, circulated, and transformed (*ibid.*). In such a context, the legal comparatist must abandon the illusion of fixed categories and be instead flexible and reflexive.

This demands a radical methodological act of humility. Comparative law is not amenable to the scientific method in the traditional sense: it "does not lend itself to formulaic approaches" and offers "no final method of dealing with problems" (*ibid.*: 23-24). Since legal meanings are in constant change, new problems emerge, new values gain prominence, and new global actors come into play, comparative law must continuously invent new tools to remain relevant. It must be willing to "destroy its own past rigidities and manners of perceiving […]", as the comparatist cannot cling to outdated methods or fixed categories and they must have the courage to relinquish tools that, while once useful, have become obsolete (*ibid.*: 24):

> Comparative law evolves differently. Its progress cannot be accomplished in such small linear advances, but requires changing the very perspective from which problems are conceived, and debunking entrenched, established orthodoxies […].
>
> (Curran, 2019: 23)

In other words, the value of comparative legal analysis does not lie in providing definitive answers, but in its ongoing ability to expose the complexities of law and its deep connections to social, cultural, and historical realities. This also means that each generation must develop new approaches to interpret legal meaning in ways that respond to their contemporary challenges (Curran, 2019). At the same time, because legal institutions are shaped by the societies in which they operate, comparative analysis helps reveal how these social forces influence legal norms and practices, and while it cannot produce normative conclusions on its own, it plays a vital role in identifying policy options that deserve further consideration (Finkin and Mundlak, 2015).[5]

Ultimately, legal knowledge is not universal, but a dynamic and interpretive activity that must be continually rethought considering changing social realities (*ibid.*). The following sections will extend these reflections to the specific domain of labour law, exploring its evolution, its current challenges, and the ways comparative analysis can illuminate its transformations in a globalised world.

2.3 Comparative labour law

From the earliest stages of human civilisation, work has been a central aspect of life, first as a necessity for survival, later as a means to improve, organise, and enrich societies (Finkin, 2019). How work is structured and regulated has long been a defining feature of civilisational development and a persistent subject of intellectual investigation (*ibid.*). Labour law, as a distinct legal discipline, emerged relatively recently, born out of the social and economic upheavals brought by industrialisation and the rise of wage labour. It was not until the nineteenth and early twentieth centuries that it became a recognised object of academic study (Finkin and Mundlak, 2015). Although it developed in separate national contexts, labour law arose as a "reaction to common methods of production and common means for the deployment and administration of labour on a global scale" (*ibid.*). This transnational character gave rise to the parallel emergence of comparative labour law, which quickly followed the establishment of labour law itself as a field of academic study (*ibid.*). Yet, the identity of labour law has always been somewhat ambiguous: "even from the beginning, it was far from clear what labour law was" (Finkin, 2019: 1). The discipline has never had a singular, universally accepted definition; its scope, purpose, and internal coherence

[5] In this perspective, it becomes essential to situate the discursive analysis presented in Chapter 7 within its historical and social context. For this reason, Chapter 4 provides a background on the colonial legacy, gendered labour dynamics, and the socio-legal history of Tanzania and Kenya, contextualising the legislative documents selected for the four sub-corpora and supporting a more grounded comparative approach.

remain matters of debate, especially in a rapidly changing world of work. The question takes on even greater urgency today, as the traditional foundations of the field, which are rooted in the Fordist model of industrial employment, are disappearing (Finkin, 2019): the classical labour law model was built on assumptions such as stable, full-time employment for a male breadwinner, long-term service with a single employer, and a clear division between work and leisure (*ibid.*). However, the post-industrial era[6] has seen the disintegration of this model. Labour markets are now characterised by flexibilisation and deregulation, precarious employment, demographic shifts, the impact of digitalisation and diminished state capacity under global economic pressure (*ibid.*).

In this context, comparative labour law is both a reflection of its time and a response to it. It emerged from the same historical forces that shaped modern labour markets, and it continues to evolve alongside them. As previously observed in relation to comparative law more broadly, the future of comparative labour law also depends on its ability to remain methodologically innovative and socially grounded. This adaptability is particularly crucial. As the world of work undergoes profound transformation, comparative analysis will remain an indispensable lens for examining how legal systems adapt and how they might do so more justly. In the following section, this broader reflection will progressively narrow its focus to examine the challenges that comparative labour law must face today.

2.3.1 Current challenges

The post-World War II period, marked by unprecedented economic growth, full employment, and industrial stability, gave rise to what is commonly known as the "standard" or "typical" employment relationship, around which labour law developed as a coherent and relatively stable body of rules (Bronstein, 2009: 1). The economic, social, and legal issues generated by this model formed the basis of labour law and comparative labour law for much of the twentieth century, particularly in advanced economies where it reached increasing levels of sophistication (Finkin, 2019). As stated in Section 2.2, however, in what is now often referred to as a post-industrial world,[7] many of the foundational elements of this system have been, or are in the process of being, dismantled,

[6] The term *post-industrial era* refers to the historical shift from manufacturing-based economies to societies primarily oriented toward services, theoretical knowledge, and information technologies. It is marked by the growing influence of professionals, scientists, and knowledge workers in both economic and political spheres (Bell, 1973).

[7] As outlined in the previous footnote 6, the term *post-industrial world* describes a global context where work is increasingly based on service and technology and shaped by information systems. It features blurred boundaries between work and leisure, a focus on work-life balance, and the growing centrality of education and innovation in economic structures (Lewis, 2003; Bell, 1973).

giving rise to a new set of challenges in the world of work (*ibid.*). As a result, the traditional model of labour law has come under profound stress: as the social, institutional, and economic conditions that once underpinned labour law evolve, so too must its comparative study (*ibid.*).

One of the most critical challenges is the ongoing flexibilization and deregulation of labour markets, driven by employers' growing efforts to adjust labour costs and working conditions in response to volatile market demands (*ibid.*). This has led to the proliferation of so-called atypical forms of employment, including part-time work, fixed-term contracts, on-call work, or temporary agency work, destabilising the historical coherence of labour law and rendering many workers precarious and legally invisible (Bronstein, 2009). This phenomenon is further complicated by fissurisation, a process through which work is fragmented across subcontractors and supply chains, and by the rise of crowdsourcing platforms that obscure the identity and accountability of the employer (Finkin, 2019). These developments challenge the capacity of labour law to define who is protected, who the employer is, and under what conditions protections apply:

> As these forms proliferate globally they are being subject increasingly to comparative law studies with the prospect of mutual learning and potential legal borrowing better to protect those doing such work.
>
> (Finkin, 2019: 17)

A second major issue is the decollectivisation of industrial relations (Finkin, 2019). Historically, labour law was closely tied to the concepts of collective bargaining and union representation. However, union density has declined significantly in some of the advanced economies, such as Germany, Japan, and the Nordic countries, which have always had strong union movements. This decline aggravates problems of wage inequality, weakens mechanisms of worker voice, and contributes to what some call "democratic deficit" in the workplace (*ibid.*: 18-19):

> Absent representation, employers are free unilaterally to adopt any employment policy, any working condition or human resource management policy believed to be economically beneficial, no matter how exploitative or invasive—unless a limit is imposed by law.
>
> (Finkin, 2019: 19)

The decline of collective bargaining institutions raises normative concerns about how to secure worker participation, represent unorganised workers, and protect labour rights in increasingly individualised employment environments.

A third and equally significant issue arises from demographic change. Ageing populations, particularly in countries like Japan and across Europe, put pressure on welfare systems and put current labour market structures to the test. Simultaneously, a greater number of women, migrants, and ethnic or religious minorities are taking part in labour markets, increasing their diversity (Finkin, 2019). This raises new tensions between employer expectations and employee needs, especially regarding work-life balance, cultural integration, and protection against discrimination (*ibid.*). Moreover, technological innovation, particularly automation, robotics, and artificial intelligence, provides both opportunities and threats to labour markets. As jobs are destroyed or transformed, the risks of unemployment and skill obsolescence increase. For many workers, the central concern is no longer job security but future employability (*ibid.*). Perhaps, the most profound challenge is the erosion of national sovereignty in labour regulation due to globalisation (*ibid.*). As capital, goods, and services flow freely across borders, national labour laws remain territorially bounded. This imbalance weakens the states' ability to enforce labour standards and protect workers, especially when companies operate across borders or move their activities to countries with weaker regulations (*ibid.*).

To respond to these challenges, several countermeasures have developed over time. One example is the creation of international labour standards by the International Labour Organization (ILO).[8] However, their real impact is often limited because countries adopt them voluntarily and enforcement remains weak (Bronstein, 2009: 2). Another important step has been supranational regulation, particularly within the European Union, where labour directives and decisions by the Court of Justice have helped build a stronger framework for transnational workers' rights (Bronstein, 2009). In addition, many trade agreements now include social clauses, making trade benefits conditional on respect for labour rights. Corporate social responsibility (CSR)[9] initiatives have also become increasingly common. Even though such measures have had varying degree of success, they represent an attempt to rebalance power in a global economy where businesses often

[8] The International Labour Organization is a specialised agency of the United Nations founded in 1919. It brings together governments, employers, and workers to set international labour standards and promote social and economic justice through decent work. The ILO operates through a unique tripartite structure and plays a key role in shaping global labour policies. See: International Labour Organization. https://www.ilo.org/ [last accessed: 25 June 2025].

[9] Even though CRS initiatives and international agreements are instruments of 'soft law', they aim at promoting labour and social standards through voluntary commitments: "CSR initiatives are developed on the basic assumption that corporations are obliged to consider the interests of customers, employees, shareholders and communities, as well as the environment" (Bronstein, 2009: 3). These initiatives usually entail NGOs, consumer associations, and trade unions, which work together to create and enforce standards without involving the state directly (*ibid.*).

outsmart government control (*ibid.*). Comparative labour law is essential to evaluating the effectiveness, limits, and normative underpinnings of such global mechanisms.

Finally, contemporary labour law's task is also to mediate conflicts between fundamental human rights and the interests of employers. Issues such as privacy, freedom of religion and expression, and non-discrimination frequently clash with employers' needs for productivity, control, and flexibility (*ibid.*). The rise of surveillance technologies, for instance, has raised urgent questions about the right to privacy in the workplace. Courts in many jurisdictions have had to face the challenge of balancing competing rights claims, often in the absence of clear legislative guidance (*ibid.*). One area of reform that is still very active is discrimination law. Although overt discrimination is now illegal in the majority of legal systems, topics like affirmative action and protection against harassment are still highly debated, especially where deep-rooted inequalities persist (*ibid.*).

Looking ahead, the challenges currently facing comparative labour law mirror the historical struggles of earlier periods. Just as nineteenth-century capitalism eventually led to the development of legal solutions for issues like child labour, workplace safety, and wage regulation, today's capitalism brings new pressures linked to alternative work forms, and the erosion of traditional employment relationships (Bronstein, 2009). At the same time, emerging issues such as the invasion of personal privacy through information technology, the demand for full workplace inclusion of marginalized groups, and the call for corporate social responsibility on a global scale, require fresh legal thinking (Finkin, 2019). The development of effective legal responses will take time and will be shaped by political struggles:

> Every period of history brings changes and these come with their share of promises and hopes, risks and anxieties, challenges and answers. It is possible to look at the world today and find that labour law has met with some success.
>
> (Bronstein, 2009).

By providing alternative models and promoting critical thinking across legal systems, comparative labour law will continue to play a crucial role in the future evolution of labour law (Finkin, 2019).

2.3.2 Methodological approaches

At first glance, the methodology of comparative labour law might seem to consist in simply contrasting how different countries address the same legal issue: "country *a* does X, and country *b* does Y", an approach that, if applied without nuance, often risks becoming superficial and therefore unproductive (Finkin and Mundlak, 2015: 6). As Finkin and Mundlak (2015) state, one of the

central challenges for comparatists lies in determining what to compare. To be methodologically more manageable, comparisons should be conducted either between relatively similar legal systems, where multiple contextual variables are controlled, or when the focus is narrowly defined on a specific legal issue. On the contrary, comparisons that span across numerous and highly diverse jurisdictions, or that attempt to address a broad range of topics, are considerably more complex. Importantly, there are no universally applicable rules in this regard, because the scope and structure of comparison must be shaped by the purpose of the research itself (*ibid.*). Finkin (2019) suggests four genres in total that represent the different perspectives through which researchers examine and compare labour laws of various countries: descriptive, predictive, purposive, and multidimensional. Each one reflects different aims and levels of depth.

The descriptive strand of comparative labour law is perhaps the most familiar, since it involves charting how various legal systems regulate a particular issue, offering an inventory of legal provisions, policies, and institutional frameworks (*ibid.*). This genre provides a vital starting point for understanding the global landscape of labour regulation, although it has limitations. There is a risk of conducting it superficially, because merely listing formal norms without considering their social, political, and economic contexts runs the risk of misrepresenting their function and significance (*ibid.*). Nevertheless, when executed with care, descriptive studies can "convey a sense of how the law on the books derived from and plays out in the country's political, economic, and cultural context" (*ibid.*: 7). Still, the "short shelf life" of such work means that descriptive comparisons must be continuously updated to remain relevant (*ibid.*: 7).

Beyond description, comparative labour law can have a predictive function. By observing legal developments across jurisdictions, comparatists can identify early signs of emerging issues: descriptive comparison "may function as an early warning system" that anticipates needs before they become urgent (*ibid.*: 8). Instead, the purposive approach is less about observation and more about normative action. Here, comparative labour law is employed "in search of a better solution to a pressing problem than domestic law currently affords" (*ibid.*: 8): it is logical and often necessary for policymakers and scholars to explore how other countries have addressed similar labour issues to identify legal models, rules, or institutions that can be transplanted or adapted to resolve local problems. However, the success of legal transplantation is far from guaranteed: while some areas, such as unemployment benefits or workplace safety, have proven more suitable to cross-

border adaptation, transplantation efforts depend on both political will and institutional compatibility (Finkin, 2019).

Finally, the multidimensional approach is based on the assumption that labour law is not an isolated doctrinal category, but an interdisciplinary field of study. Therefore, meaningful comparative analysis in this genre requires a grasp of not only multiple legal systems but also their history, cultural logics, economic structures, and social practices (*ibid.*). However, achieving such multidimensional depth is demanding. It requires fluency in legal terminology, familiarity with implicit assumptions within systems, and often, competence in multiple disciplines.

The comparative process also involves other kinds of decisions. Comparisons may begin inductively, starting with an analysis of several countries from which typologies are derived, or deductively, by testing theoretical models against real-world examples (Finkin and Mundlak, 2015). However, while ideal types are helpful in clarifying distinctions, they must be understood as heuristic devices rather than precise interpretations (*ibid.*). They serve to organise the discussion and provide a clear starting point, but their application requires contextualisation. To use ideal types effectively, one must look beyond surface labels and examine how the institution in question is actually structured, how it relates to other parts of the legal system, and how it works in practice, not only in law, but in real life as well (*ibid.*).

Another key methodological question concerns the selection of countries. As previously discussed in the present section, there are no fixed rules; the decision to compare similar or different systems depends largely on the goals of the study (*ibid.*). It is important that researchers choose countries they know well. A solid understanding of national legal systems allows for a deeper analysis of legal complexity. Relying only on secondary sources can lead to mistakes, such as misinterpreting laws, misjudging their impact, or overlooking how legal rules work together within a broader system (*ibid.*). Comparative insights can even be generated within a single country. For example, diachronic comparisons can highlight legal transformations in time, while comparisons between regional courts may reveal structural divergences within federal systems (*ibid.*).

Ultimately, labels matter less than having a clear focus and a solid method. For the epistemic community of comparatists, a genuine comparative method requires more than juxtaposing laws: it entails engagement with national, social, and cultural variation, looking beyond one's legal system to grasp alternative ways of thinking about law and society (*ibid.*).

In this dissertation, the role of context is especially important. While the study is designed in a way that could be extended to other countries in future research, the focus here is only on Kenya and Tanzania. The choice of these two countries reflects a specific interest in their colonial legacies and legal developments, which make them interesting to compare. All decisions regarding the selection of countries and legal texts are explained in Sections 4.4, 6.4, and 7.2.1 and its related subsections.

2.3.3 Sources of labour law

Labour law draws upon a wide and multifaceted array of legislative sources regulating the employment relationship across jurisdictions. The fundamental instruments shaping labour regulation are constitutions, statutes, collective and individual contracts, soft law documents, and international or regional legal instruments (Pittard and Butterworth, 2015).

National constitutions in most countries often serve as a primary legal source for labour rights, either in explicit terms in relation to freedom of association, collective bargaining, and protection against forced labour, or in more general human rights provisions applicable to the field of employment (*ibid.*). However, constitutional protection of labour rights is far from universal because not all constitutions explicitly address labour rights. Countries like Australia are silent, leaving the matter to statute intervention (*ibid.*). Thus, national and sub-national legislation is also a primary source of labour rights. In most legal systems, it is statutory law that constitutes the backbone of labour regulation. It tends to set minimum employment standards such as working time, compensation, health and safety, termination rights, and protection against discrimination. These statues may either complement constitutional provisions or serve as the only regulatory framework (*ibid.*). Collective agreements and individual contracts also play a crucial role in defining the terms and conditions of work. They can be negotiated between employers and employees, often with the involvement of trade unions or other representatives:

> Collective bargaining can generally be described as 'the process by which terms and conditions of employment are determined by negotiations between employers and trade unions' Having said that, some systems permit negotiations between employers and groups of employees, without union involvement.
>
> (Pittard and Butterworth, 2015: 31-32)

However, the way in which collective bargaining is regulated can differ across legal systems. In some contexts, bargaining follows a highly regulated process, while in others it is largely self-

managed. For example, in Sweden there is minimal intervention by the government in the bargaining process (Pittard and Butterworth, 2015). Customs and internal workplace policies also contribute to labour regulation. Customary practices, such as the payment of annual bonuses or the procedures for employment retrenchment, may influence employment relations, but they are less likely to be enforceable unless incorporated into individual contracts (*ibid.*). The emergence of 'soft law' instruments, such as corporate social responsibility codes and global compacts, is another illustration of modern diversification of regulatory mechanisms (*ibid.*). International sources provide an additional level of regulation. The ILO has developed a body of conventions and recommendations aimed at setting minimum labour standards at a global level, in areas such as freedom of association, elimination of forced labour, or abolition of child labour (*ibid.*). Similarly, the United Nations has impacted labour rights with instruments such as the Universal Declaration of Human Rights (1948) and the International Covenant on Economic, Social and Cultural Rights (1966) by affirming the rights to work, to equal pay, and to join trade unions (*ibid.*). Nonetheless, it is important to note that "[r]ules and norms emitted by the UN, the ILO, the OECD, and the CoE (save for the ECHR) have an indirect and voluntary impact" (Jacobs, 2021: 19). Their binding effect is subject to national ratification and implementation.

Finally, at the regional level, supranational bodies like the European Union have also developed a complex system of labour law through directives and regulations, such as the Working Time Directive and the Anti-Discrimination Directives, setting a minimum standard that member states must incorporate into their national legislation (Pittard and Butterworth, 2015). The European Court of Justice makes sure that these standards are interpreted and applied correctly across the UE: "Everywhere in Europe case law of the courts is an important source of labour law. In many countries labour law disputes are dealt with by specialized courts" (Jacobs, 2021: 18). In parallel, the Council of Europe, through instruments such as the European Convention on Human Rights (1950) and the European Social Charter (1961), offers additional protections for workers, safeguarding rights like the freedom to organise or the right to strike.

> Many of the standards set by the European Union have an immediate and mandatory effect in all Member States. This concerns notably the Regulations and Directives and some provisions of the Charter of Fundamental Rights of the EU. The EU Member States no longer have the freedom to choose whether to follow or reject mandatory norms emanated from the EU. They have to implement these norms scrupulously. When the norms of the EU are not applied in the way the European Commission (EC) considers it right, the European Commission may bring the case before the Court of Justice of the EU.

(Jacobs, 2021: 19)

After providing an overview of key methodological approaches in comparative labour law, the next chapter turns to the *Worlds of Labour* (WoL) project, which represents a concrete application of these principles through a distinctive quantitative method: leximetrics. As outlined in the introduction, WoL offers the broader framework within which this dissertation is situated: the legal texts examined in the case study in Chapter 7 are all drawn from the WoL database, which has been systematically coded within the WoL framework to investigate the global dynamics of legal segmentation.

3 The Worlds of Labour Project

3.1 Introduction

As outlined in the Introduction, this dissertation builds on the broader research framework developed within the *Worlds of Labour* (WoL) project. WoL provides both the conceptual foundation and the empirical resources that have made it possible to construct the annotated corpus at the centre of this study. The legislative texts discursively analysed here are, in fact, part of the project's annotated database. The central aim of this thesis is to contribute to the development of a methodology suitable for the construction and annotation of corpora of legal texts. Specifically, the study focuses on legislative texts concerning labour law in two East African countries, Kenya and Tanzania, enacted during the colonial and post-colonial period,[10] with the aim of creating an annotated corpus that contributes to the study of legal segmentation (see Section 3.2) within a precise historical and geographical context.[11]

WoL is an integral part of the Collaborative Research Centre 1342 (CRC 1342) *Global Dynamics of Social Policy*, a broad interdisciplinary research hub based at the University of Bremen. CRC 1342 is a research initiative involving eight institutes from the University of Bremen, in collaboration with Constructor University Bremen, Bielefeld University, and the University of Duisburg-Essen. It was funded by the *Deutsche Forschungsgemeinschaft* (DFG) [12] in 2018. The main objective of CRC 1342 is to analyse the evolution and global dynamics of public social policies, aiming to understand how these have developed over time, why and how countries across the world decide to support their citizens, and what external influences play a key role in their decision-making. The approach goes beyond the traditional focus on OECD (Organisation for Economic Cooperation and Development) countries and includes those in the Global South as well.

[10] The annotated and analysed texts cover the period between 1932 and 2011 and consist of a total of 22 legislative documents. More detailed information on the full composition of the project's corpus is provided in Chapter 7. For a contextualisation of the colonial and post-colonial period, see the brief historical background on Kenya and Tanzania in Chapter 4. The same chapter also explains the rationale behind the selection of these two countries for testing the methodology developed for the *Worlds of Labour Corpus Project* (WoLCP) through a comparative discourse analysis of the texts.

[11] Interest in the *Worlds of Labour* (WoL) project emerged within the framework of an academic internship organised by the Department of Interpretation and Translation (DIT) at the University of Bologna – Forlì Campus. Professor Silvia Bernardini facilitated contact with the research team based in Bremen, enabling six students enrolled in the master's degree in Specialized Translation to take part in the project. This experience constitutes a fundamental basis for the development of the present master's thesis, which aims to explore a specific aspect of the project through an independent, methodologically oriented analysis.

[12] "German Research Foundation". Unless otherwise specified, translations into English are provided by the author of this thesis.

This perspective is based on the awareness that a country's social policies are not solely shaped by internal factors but are also influenced by international relations and global processes such as migration, war, geopolitical conflict, the circulation of ideas and legal norms, colonialism, and economic interdependence (CRC 1342: Global Dynamics of Social Policy, n.d.).

Among the various ongoing projects within CRC 1342, *Worlds of Labour* belongs to Project Area A and is officially titled *Worlds of Labour: Coverage and Generosity of Employment Law*. The WoL team consists of five researchers: Prof. Dr. Irene Dingeldey, Prof. Dr. Ulrich Mückenberger, Dr. Heiner Fechner, Marina Carlino, Andrea Schäfer.

This chapter presents the *Worlds of Labour* project as developed by the team in Bremen, which provides the conceptual foundation for the broader collaborative *Worlds of Labour Corpus Project* (WoLCP). The methodology adopted for collecting and annotating legal texts within the WoLCP framework has been applied to the development of the sub-corpora focused on Kenyan and Tanzanian labour law. Chapter 5 will further discuss the WoLCP, using it to illustrate the methodology adopted in the present thesis. Hence, this approach, initially implemented within the WoLCP, serves as the basis for the creation of the Tanzanian and Kenyan sub-corpora and will be tested within the framework of the illustrative case study presented in this thesis.

3.2 The phenomena of labour market segmentation and legal segmentation

The concept of legal segmentation serves as the starting point of the WoL project, as it represents the core of the analysis of regulatory dynamics that influence labour market governance on a global scale. The entire WoL study is developed from this perspective, aiming to examine how labour laws have historically contributed to creating, reinforcing, or mitigating divisions between different categories of workers, thus generating structural inequalities in access to rights and protections (Dingeldey et al., 2020).

It is essential to distinguish between labour market segmentation and legal segmentation, as they are closely related but conceptually distinct notions. Labour market segmentation refers to the division of the labour market into separate sub-sectors, each characterised by different working conditions, regulatory frameworks, and behavioural dynamics (Deakin, 2013; Eurofund, 2019). These divisions largely depend on the specific context in which workers operate and may arise from various institutional factors, such as contract types (permanent vs. temporary employment),

the degree of enforcement of labour regulations, and the categories of workers involved (e.g. migrants, domestic workers, dispatch workers;[13] Deakin, 2013). These factors lead to unequal access to rights, protections, and job opportunities, contributing to the persistence of disparities between labour market segments and producing negative social and economic effects. Key inequalities involve wage gaps and differences in employment conditions, such as job security and contract duration (*ibid*.). In industrialised countries, segmentation is often associated with the divide between standard and non-standard employment, involving part-time contracts, fixed-term contracts, temporary agency contracts, and flexible contracts (*ibid*.). In developing countries, segmentation mainly manifests in the distinction between the formal and informal sectors: while the former is regulated and protected by law, the latter operates outside the scope of legal frameworks and lacks safety and social protection (*ibid*.). Segmentation is thus central to debates on labour law and social protection policies, as it directly affects employment quality and the structure of inequality in global labour markets.

Legal segmentation, in contrast, refers to how labour laws themselves contribute to creating, reinforcing, or reducing inequalities inherent in labour market segmentation (Dingeldey et al., 2020). Through provisions that favour certain groups of workers over others, labour law, rather than providing equal protection, often plays an active role in shaping and perpetuating inequality: "Beyond labour market processes, a cause for segmentation can be found in legislation itself, both by neglecting factual disparities and privileging 'insiders'" (*ibid.*: 1). Some labour regulations fail to address structural differences between worker categories, favouring the so-called insiders, typically those in standard forms of employment such as full-time, permanent contracts, and excluding outsiders, like precarious, part-time, or informal workers (Dingeldey and Gerlitz, 2022). Studying legal segmentation is therefore crucial to understanding how law influences the labour market and shapes inequalities among workers.

A key concept in the study of legal segmentation is the Standard Employment Relationship (SER), a legal paradigm for the employment relationship between employer and employee that has

[13] Dispatch workers are employed by private labour-dispatch agencies under fixed-term or open-ended contracts and are assigned to perform work for a third-party company (the user enterprise), which supervises their tasks but does not formally employ them. Also known as *temporary agency work* or *labour hire*, this non-standard employment model raises issues concerning job security and workers' rights. Its three-way arrangement involves two contracts: an employment contract between the dispatch worker and the agency, and a commercial agreement between the agency and the user enterprise. The user enterprise pays a fee to the agency, which in turn pays the worker's wages (Lo, 2023; Liu, 2014).

historically influenced both labour law and social protection policies, especially in industrialised nations (Mückenberger, 1985, 1989; Bosch, 1986). It emerged relatively recently, around the mid-twentieth century, and has taken different forms depending on national contexts, reflecting each country's specific paths of industrialisation and models of industrial relations (Deakin, 2013). Nevertheless, it is possible to identify core characteristics that are common to its various forms. In particular, the SER, regarded as an ideal type of employment relationship, is centred around stable employment, typically involving a full-time, permanent contract, regulated by a single employer and carried out at a single workplace (Dingeldey et al., 2020). However, it did not become dominant merely because it was widely practiced, but primarily because it served as a reference model for laws and regulations, both in labour law and in social security legislation (*ibid.*). Its status as a 'guiding principle' has influenced, for instance, how protections and social benefits are provided, the definition of gender roles in the labour market (that is, how men and women gain access to employment and who is better protected), and the meaning of citizenship and nationality in relation to social rights:

> [T]he fundamental meaning of the SER derived not necessarily from its quantitative empirical relevance in industrialised countries, but it served as the central point of reference and model for legislation [...], not only in labour law, but often also in social security law.
>
> (Dingeldey et al., 2020: 4)

Although the SER has historically provided a high level of legal protection for workers who fall within its framework, it has also contributed to creating a divide between those who benefit from it and those who are excluded. Non-standard workers, such as the self-employed, temporary or part-time employees, as well as migrants and women, are often excluded from the SER and receive fewer protections: "The Standard Employment Relationship (SER) in industrialised countries is associated with strong protection for employees who fulfil its criteria but tends to neglect those who do not" (Dingeldey et al., 2022: 560).

More recent developments in reflexive labour law theory (Rogowski, 2013)[14] highlight the role of internal factors within the legal system in generating segmentation. The key idea is that the

[14] Reflexive labour law is a theoretical approach that sees labour law not as a system that directly controls workplaces, but as one that supports other social systems, such as industrial relations, in regulating themselves. In this view, labour law does not impose strict rules, but creates general legal frameworks that help employers, workers, and unions make their own agreements. The theory was developed in response to the limits of traditional labour law and aims to deal with the challenges of globalisation and corporate self-regulation through more flexible and cooperative methods (Rogowski, 2015).

legal system does not respond to economic transformations in an immediate or linear way but rather possesses its own autonomy and follows internal evolutionary paths, while remaining connected to the broader external context. As a result, law co-evolves with the economy and society, rather than simply adapting to market demands (*ibid.*). Instead of viewing law as a mirror of the economy or the economy as the sole driver of legal change, the reflexive labour law perspective emphasises mutual influence: the SER was shaped by the economic context, but once codified into law, it in turn influenced economic and social processes, sometimes even reinforcing rigidities. Therefore, it is a model that the legal system constructed and legitimised over the course of the twentieth century, and once established, it continued to serve as a reference point for new legal norms, actively shaping their development (*ibid*.).

Today, the SER is considered partially obsolete, as since the 1980s the structure of the labour market has grown increasingly complex: "With globalisation, tertiarisation, and digitalisation of labour also the flexibilisation of employment and the privatisation of protection were emphasised (Dingeldey et al., 2020: 4-5). The growing heterogeneity of society has contributed to the de-standardisation of labour market participation and the transition into professional life. While the SER continued to revolve around a specific, traditionally conceived form of employment, based on the male breadwinner model, full-time and stable, it implicitly excluded women, migrants, and workers in non-standard forms of employment (Dingeldey et al., 2020). Over time, less traditional employment forms, such as part-time, fixed-term, or agency work, began to gain importance (Dingeldey et al., 2022). In this sense, there is a structural link between the phenomenon of segmentation and the legal structure of the SER itself, as the exclusive protection of standard employment tends to reinforce mechanisms of exclusion for non-standard workers (Carlino et al., forth.). The proliferation of distinct and atypical forms of employment can thus be seen as a response to the problems triggered by the emergence of the SER as the dominant employment paradigm: when the protections associated with the SER are particularly strong, employers may be incentivised to adopt atypical contractual forms in order to avoid costs and labour rigidities (Deakin, 2013). Precisely because of its central role in regulating employment and defining labour protections, the SER has become an essential point of reference in the debate on legal segmentation.

Building on this foundation, *Worlds of Labour* develops its research project with the aim of achieving a broader understanding of labour market differentiation dynamics on a global scale.

The development of the WoL dataset is based on a core assumption: "that national labour regulation can prevent, encourage and counteract labour market segmentation" (Dingeldey et al., 2020: 1). In other words, labour law can influence labour market segmentation in three distinct ways. On the one hand, regulations can have a preventive function, aimed at ensuring equal treatment for all workers and at limiting pre-existing forms of segmentation. On the other hand, they may promote segmentation, for example through provisions that favour specific categories of workers, such as granting greater benefits to full-time employees compared to part-time ones. Finally, laws may take a tolerant stance toward existing segmentation by choosing not to intervene to change or reduce it (Dingeldey et al., 2020). In light of this, WoL has developed an approach to quantitively detect these three distinct ways in which labour law can influence labour market segmentation. The following two sections will outline the project's objectives and describe the methodologies employed.

3.3 The WoL research goals

The *Worlds of Labour* (WoL) project is built around a series of fundamental questions aimed at studying and measuring legal segmentation on a global scale. The focus lies on identifying the necessary data for a comparative analysis of workers' legal protection and on defining the most appropriate variables for constructing a research-oriented dataset. From this perspective, WoL aims to move beyond the focus on industrialised countries[15] and conduct a broader analysis that includes the Global South.[16] Comparative research on labour regulation has so far primarily focused on employment protection in countries of the Northern Hemisphere (Dingeldey et al., 2022). There are few analyses that include the Global South, a region where labour market segmentation takes on specific characteristics linked to colonial legacies, the widespread presence of informality, and the influence of the International Labour Organization (ILO). This is further compounded by

[15] The Industrial Development Organization (UNIDO) (2024) classifies countries into groups based on their stage of industrial development, using objective criteria linked to manufacturing performance. First, economies are categorised as "industrial" or "industrialising" according to their average manufacturing value added (MVA) per capita and the historical maximum of MVA as a share of gross domestic product (GDP) (1970–2022). Then, each group is further divided by income level (high, middle, low) following the World Bank's classification. This results in five main groups, updated annually and used strictly for statistical purposes.

[16] Mahler (2017) explains that the term *Global South* has three main meanings: (i) it traditionally refers to economically disadvantaged states and emerged as a post–cold war alternative to "Third World"; (ii) in a post-national sense, it describes populations and spaces marginalised by capitalist globalisation, including those within wealthy countries; (iii) finally, it denotes a transnational political subjectivity based on shared experiences of subjugation and solidarity among the world's "Souths."

the lack of accessible data for quantitative studies on the historical roots of such phenomena, particularly regarding regulations on worker protection worldwide and segmentation (Dingeldey et al., 2020).

WoL's starting point is the recognition that, although the SER represents the dominant model of labour regulation in industrialised countries, it shows significant structural differences across legal systems. For instance, in France, Germany, and the United Kingdom, labour law and the mechanisms for protecting workers vary considerably at both the individual and collective levels (Carlino et al., forth.). Nevertheless, WoL hypothesises that European paradigms, each shaped by its own legal and institutional history, have all had a significant influence on the regulatory models adopted in countries of the Global South. This influence, however, has not always been coherent or consistent with local employment structures and has, in some cases, clashed with or even contradicted the employment structures of those regions (Dingeldey et al., 2020). In other words, the analysis of the collected data could shed light on the role of the Western SER in the legal systems of the Global South, assessing its impact through factors such as global trade dynamics, colonialism, and ILO conventions and recommendations (*ibid.*). These processes have led to the adoption of Western legal models in different economic and social contexts, shaping labour policies in Global South countries and, in some cases, reinforcing pre-existing forms of segmentation.

> It may seem surprising that we take the criteria of the SER as reference of analysis of global employment law, despite the high amount of informal and non-standard work particularly in the Global South. We assume, however, that formal employment patterns all over the world follow certain paradigms of labour regulation, hence a variety of 'SERs'. And we assume that the European paradigms of employment law, though differing in kind, have had a great impact on the paradigms in the Global South […].
>
> (Dingeldey et al., 2020: 1)

WoL hypothesises the existence of three patterns that may have influenced the standardisation of labour and the evolution of the SER in Global South countries (Dingeldey et al., 2020). The first involves the legal transplantation of norms through colonialism or the actions of international organisations. A second suggests that some territories preserved or independently developed their own legal frameworks, which may later have merged with European standards. Finally, a third model consists of a clear dichotomy between the formal and informal economies: in formal sectors,

European norms and their subsequent developments were adapted to local contexts, while in informal areas, local, indigenous, or traditional normative standards, either inherited or newly developed, continued to prevail (*ibid.*).

As previously mentioned, no systematic attempt has yet been made to compare national labour legislation using quantitative data, despite the extensive literature on the SER in Western contexts. Existing research has largely been limited to describing common phenomena in certain Western countries, without providing tools to measure and compare similarities and differences in a structured way (*ibid.*). WoL was therefore developed to address this gap by collecting and analysing data useful for tracing the evolution of labour legislation and its impact on labour market segmentation: "We address these issues and ask how legal segmentation in employment law can be captured and measured at a global level to sketch the framework of a dataset designed to fill that gap" (Dingeldey et al., 2020: 1). The development of the dataset, however, is not only aimed at reconstructing regulatory changes over time, but also at highlighting their key functions, such as the safeguarding of workers, the differentiation between categories with varying levels of protection, and the fight against discrimination (Carlino et al., forth).

For the design of this database, the WoL research team posed a series of fundamental questions: How can a dataset be developed to map legal segmentation? What data should be collected, and which variables should be used? And how should the dataset be structured to support comparative analysis on a global scale? (*ibid.*). The next section will present the methodology adopted by the WoL team, briefly describing the process of data collection and organisation, as well as the criteria followed to ensure comparability across different legal and historical contexts.

3.4 The methodological approach

To analyse on a global scale how labour laws influence segmentation in the labour market, not merely as neutral rules, but as factors that directly affect who receives greater protection and who does not, WoL has collected data on over 150 countries, with temporal coverage that, for some of them, dates back as far as 1880 (see Section 3.5 for details on the time frame and the data collection of the project; Dingeldey et al., 2020). The main challenge of a project of this scale is to make legal frameworks from different jurisdictions and historical periods comparable. In other words, to identify an analytical method that allows legislative changes to be measured in a systematic and objective way.

From this perspective, WoL identified the leximetric approach (*leximetrics*) as the most suitable strategy for quantifying legal language and transforming the qualitative content of laws into numerical data. Leximetrics is a branch of economic analysis that emerged in the early 1990s and focuses on the quantitative measurement of laws across countries and over time (Konstantin et al., 2021). It aims to transform legal provisions into numerical indices through codification of legal texts or surveys with the goal of assessing the intensity and structure of regulation in various sectors and explaining their socioeconomic effects (*ibid.*). In other words, this method is specifically designed for the analysis of legal texts and aims to quantify the normative content of legal rules by assigning numerical values based on pre-defined criteria. By adopting this approach, it becomes possible to translate qualitative content into quantifiable data by assigning values to legal terms, phrases, or concepts according to predefined criteria:

> Leximetric methods offer the means to measure and quantify various features of legal textual materials. This involves assigning numerical values to legal terms, phrases or concepts based on predefined criteria. The most common approach to measurement is coding laws, involving a six-step process for dataset construction.
>
> (Carlino et al., forth.: 66)

Carlino et al. (forth.) explain that this methodology is based on a structured process of legal coding, organised into six key stages. The first step involves selecting a latent concept, namely, an aspect of the law to be measured, such as legal segmentation in the labour market, followed by the dimensionalisation of that concept through the construction of a theoretical framework that defines its various facets. In the case of labour market, the concept may refer to how labour laws provide different levels of protection or benefits depending on factors such as workers' seniority. This requires specifying in concrete terms what constitutes segmentation, (e.g., the existence of legal provisions that differentiate entitlements to severance pay, dismissal notice, or protection against unfair termination based on years of service). Next, it is necessary to identify indicators, which are essential for making the concept measurable, and to define the variables, each accompanied by precise coding instructions, the "coding algorithm" (Carlino et al., froth.: 66). This is followed by the selection of the legal database (that is, the legal texts to be analysed) and finally, the extraction, interpretation, and assignment of data, during which legal provisions are classified and translated into numerical values (see the end of this section for a clearer explanation on the practical use of this model).

Although it represents an innovative method for measuring and analysing legal norms, leximetrics has been the subject of criticism highlighting certain limitations (see Deakin 2018; Teklé 2024). These are primarily focused on two aspects: on the one hand, the reliability and transparency of data coding; on the other, the validity of the indicators – that is, their ability to measure legal norms without disregarding the legal and cultural context. It is important to acknowledge that law is inherently interpretive, and quantifying it risks oversimplifying legal complexity, which is shaped by essential contextual elements (Carlino et al., forth.). Reducing legal norms to numerical variables may lead to a loss of linguistic complexity, overlooking nuance and semantic richness, as well as detaching texts from the broader legal structures in which they are embedded, ultimately compromising accurate interpretation (*ibid*.).

However, as Carlino et al. (forth.) explain, despite these criticisms, the leximetric method offers undeniable advantages. Its ability to ensure reproducibility and systematicity allows for greater consistency in analysis and facilitates comparison across different studies (*ibid*.). Transforming legal texts into numerical data accelerates the analysis of laws and makes it possible to identify recurring patterns and evolutionary trends in regulation (*ibid*.). Moreover, the method enhances comparability across legal systems and over time, enabling researchers to trace legal developments and construct historical datasets (*ibid*.). Integration with data visualisation tools improves the clarity of results, and when the data obtained are used in statistical analyses, it becomes possible to identify recurring patterns and significant relationships between legal norms and the dissemination of legal concepts (*ibid*.). In this way, leximetrics contributes to testing and refining existing theoretical models, providing empirical support for legal research and making normative analysis more rigorous and grounded in concrete data (*ibid*.).

Initially drawing on the CBR-LRI dataset[17] developed by the University of Cambridge as a key reference, the WoL team developed its own model built around three main functions, each highlighting a distinct and fundamental aspect of protection and segmentation: "In functional terms, we call these the standard-setting (S), privileging (P) and equalising (E) function" (Dingeldey et

[17] The CBR Labour Regulation Index (CBR-LRI) is a dataset developed by the Centre for Business Research at the University of Cambridge, which analyses and codes labour laws from 117 countries over the period from 1970 to 2013 using a leximetric approach. The CBR-LRI is structured into five sub-indices, each designed to measure specific aspects of labour regulation (Deakin et al., 2017). The Worlds of Labour project adopts the CBR-LRI as a methodological reference but expands its scope. While recognising its advanced capacity for comparative analysis, WoL identifies certain limitations in the index's ability to adequately represent legal segmentation in the labour market. As a result, WoL incorporates the variables developed by the CBR into its own analytical framework, extending them with the addition of new variables related to differentiated protection and anti-discrimination aspects (Carlino et al., 2025).

al., 2022: 563). The standard-setting function establishes the legal foundations for ensuring a minimum level of protection for all workers, contributing to the creation of a uniform regulatory framework and defining basic rights such as open-ended contracts, minimum wages, working hours, or paid leave. The privileging function, by contrast, refers to rules that favour certain groups of workers at the expense of others, leading to the exclusion of the latter. In this case, the law itself reinforces segmentation within the labour market. Finally, the equalising function introduces anti-discrimination provisions aimed at reducing legal segmentation by addressing inequalities and ensuring protection and equal access to employment for all, for example, through safeguards for migrants or persons with disabilities (Carlino et al., forth).

Based on these three functions, WoL proposes a classification of labour law systems known as the SPE typology, which aims to broaden traditional labour law studies that often focus exclusively on employment protection. This typology considers two fundamental aspects. First, the extent to which certain groups of workers are actively privileged or supported by the law; second, the extent to which other groups are excluded from protection or even disadvantaged by specific legal provisions. By examining both inclusion and exclusion, the SPE typology helps uncover segmentation dynamics that emerge from a legal system that does not guarantee equal protection for all workers:

> [I]t measures not only the level of employment protection, but also informs about the degree of active legal promotion of certain parts of the labour force and – at the same time – the development of 'reactive' legal inhibition of discrimination against others.
>
> (Dingeldey et al., 2020: 2)

In essence, the SPE typology does not merely assess how protective a legal system is, but also examines how it incentivises, excludes, or equalises different categories of workers, offering a completely new perspective for the comparative analysis of labour law (Dingeldey et al., 2022; see Chapter 2 for a discussion on the comparative approach to labour law).

To operationalise its analysis, WoL developed a set of 35 variables grouped according to the three main functions of the SPE typology (Carlino and Fechner, 2025). Each variable is designed to measure a specific aspect of labour regulation (for example, the minimum number of paid leave days guaranteed by law) and is quantified using a scale typically ranging from 0 to 1, where 1 represents the highest level of protection or regulation provided, and 0 indicates the absence of protection. In this way, the variables translate legal norms into numerical data that can be compared

over time and across different legal systems, providing an empirical basis for analysing legal segmentation in the labour market (*ibid.*). For a concrete example of how legal provisions are converted into numerical values, see Appendix A, which includes an excerpt from the coding template for South Africa.

3.5 The database

Data collection is one of the most complex and time-consuming phases of a project aiming to locate the most relevant legislative texts from several countries. The selection focused on normative acts adopted by public authorities that had a direct impact on employment relationships (Carlino and Fechner, 2025). Norms that lacked general applicability were excluded unless they invalidated an existing law (*ibid.*). This phase was led by a thorough review of existing datasets (e.g., World Bank, OECD, ILO, CBR-LRI), whose strengths and limitations helped define the scope of relevant legislation. It was ensured that the selected texts could be compared across countries and historical periods, and that they provided a reliable basis for coding the indicators associated with the SPE typology. In other words, WoL developed a strategic approach to legal text identification: choosing laws that were both meaningful for the analysis of legal segmentation and fit to coding (*ibid.*).

Although modern legislation is often assumed to be readily accessible through libraries, archives, and online platforms (Carlino et al., forth.), the process may be hindered by numerous factors, including the incompleteness of global legal databases, the lack of systematised national historical archives, and the challenges posed by linguistic differences (*ibid.;* see forthcoming footnotes for a brief overview of the main databases consulted by the WoL team). The main sources used by the WoL research team included international databases such as the ILO Legislative Series (1919–1988),[18] ILO Labour Law Documents (1990–1995), and NATLEX, a database of national labour, social security and related human rights legislation.[19] However, none of these sources proved to be fully comprehensive, particularly concerning colonial legislation. To fill these gaps,

[18] Legislative series. International Labour Organisation. Geneva. ILO; 1919-1989 https://labordoc.ilo.org/discovery/fulldisplay?docid=alma991861063402676&context=L&vid=41ILO_INST:41ILO_V2&lang=en&search_scope=ALL_ILO&adaptor=Local%20Search%20Engine&tab=ALL_ILO&query=any,contains,Legislative%20series.&offset=0 [last accessed: 25 June 2025].

[19] ILO NATLEX Database, Database of national labour, social security and related human rights legislation, International Labour Organization. https://natlex.ilo.org/dyn/natlex2/r/natlex/fe/home [last accessed: 25 June 2025].

the team turned to national archives and digitalised historical collections, as well as research projects such as HDTCOL, which focuses on French colonial regulation (Le Crom, 2017). These sources were further supplemented through direct consultation of physical archives and government databases. To overcome language barriers, the team implemented the use of artificial intelligence for automatic translation, using tools such as DeepL and ImTranslator, which made legal texts to be coded accessible in English (Carlino et al., forth.). To ensure the reliability of the information collected, WoL adopted a validation system based on secondary sources, including legal texts and reports by the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), the International Convention on the Elimination of All Forms of Racial Discrimination (CERD), and the ILO. It is important to note that the database focuses exclusively on formal legislation enacted by national governments, excluding local regulations, collective agreements, and case law (*ibid.*). This choice, driven by considerations of feasibility and international comparability, made it possible to create a consistent dataset suitable for longitudinal analysis (*ibid.*). Despite the challenges, the project led to the creation of the largest digitised collection of labour legislation currently available and represents a valuable resource for comparative studies on legal segmentation.

To ensure consistency in the assignment of values during the legal coding process, WoL also developed a template accompanied by a detailed description of the indicators, along with guidelines outlining the geographical scope, time frame, legal sources to be considered, objectives and measurement principles, and the rules for coding the information (see Carlino and Fechner 2025). The database comes in two versions: the first covers the period from 1970 to 2013, while the second, currently being finalised, extends the analysis from 1880 to 2023 (Carlino et al., forth.). For each country, data are collected and organised in Excel sheets, following strict rules for the citation of legal provisions, classification of sources, and assessment of information reliability. A double-check verification system was adopted to ensure accuracy and consistency. At present, the database includes 35 indicators, with the first version covering 115 countries and the second aiming to expand the analysis to 165 countries (Carlino et al., forth).

The *Worlds of Labour* project thus stands as an ambitious initiative aimed at studying and measuring legal segmentation through a comparative and quantitative approach. By systematically collecting and coding labour legislation in over 150 countries, WoL has developed a dataset that enables the analysis of regulatory developments over time and their impact on the protection and

differentiation of worker categories. The adoption of the leximetric methodology has made it possible to transform legal language into numerical data, allowing for empirical analysis and more balanced comparisons across different legal systems. The importance of the project also lies in its ability to move beyond the traditional focus on advanced economies by including the realities of the Global South, contexts often overlooked in comparative studies. The resulting database represents an unprecedented resource for research on labour regulation and social protection policies, providing a solid foundation for the study of legal segmentation.

To better situate the empirical investigation within its socio-historical context, the next chapter provides a contextual overview of how women's labour has been regulated and represented in Kenya and Tanzania, highlighting the colonial roots of gendered labour segmentation and the relevance of these two countries for the comparative case study.

4 Contextual Backgrounds: Legal and Gender Perspectives on Kenya and Tanzania

4.1 Introduction

This chapter provides a contextual framework for better understanding how women's labour has been historically constructed and represented in Kenya and Tanzania. Its inclusion within the present thesis serves two core purposes. First, it enriches the project's gender dimension by shedding light on how colonial employment laws have affected women in the two countries of East Africa.[20] Second, it offers a key insight into the relationship between gendered labour, informal employment, and the concept of legal segmentation.

The chapter is structured into three main sections. Section 4.2 explores the connection between gender, colonial labour, and law in sub-Saharan African countries, introducing the concept of "double segmentation" as described by Fechner (2022: 619) and providing historical information about how European colonisers shaped and influenced the marginalisation of women within African societies. Section 4.3 briefly presents an overview of the social and economic condition of women in the colonial and postcolonial periods in both countries, specifically focusing on structural inequalities. Finally, Section 4.4 outlines the rationale for selecting the Kenyan and Tanzania texts for the case study presented in Chapter 7, taking into consideration linguistic, legal, and historical reasons for their selection that make their comparison relevant.

4.2 Gender, law, and labour in colonised East Africa

In sub-Saharan Africa, the relationship between gender, law, and work has been shaped by the legacy of colonial rule (Fechner, 2022). Instead of being neutral or fair, legal systems were often used to exclude and control, reinforcing gender roles that still affect the world of work today (*ibid.*). This section looks at how employment laws in Kenya and Tanzania have played a role in shaping these patterns and why it is important to look closely at the language used in legal texts to truly understand how women's labour has been defined and regulated.

4.2.1 Reasons for a discourse approach to female labour

The choice to investigate the discursive construction of female labour in legislative texts from Kenya and Tanzania originates directly from the work of Fechner (2022), which served as a key

[20] The term *colonial employment* as used here refers to the legislative documents included in the Kenyan and Tanzanian sub-corpora. However, it is important to clarify from the outset that both colonial and post-colonial texts were selected for corpus construction, annotation, and analysis. This expression is nonetheless retained because, as discussed in this chapter, the legacy of colonial legal structures continues to influence the representation of female labour throughout the timeline considered.

source in understanding how gendered labour in sub-Saharan African countries was historically shaped and expressed in employment law. His research provided critical insight into the intersection of gender, race, and legislation and helped clarify the socio-legal background required to contextualise the linguistic analysis, playing a fundamental role in shaping the direction of this study. Fechner's (2022) work is centred around the notion of legal segmentation,[21] which he describes as

> the normative concept underpinning employment-related legislation that, either explicitly or through its effects, differentiates between persons based on racial origin or gender for other than compelling reasons.
>
> (Fechner, 2022: 617).

In his contribution (2022), Fechner discusses legal segmentation in connection with the concepts of formal and informal labour and gender discrimination. He argues that the legal structures imposed and enforced in Africa by European colonisers not only created a divide between regulated (formal) and unregulated (informal) labour but also built this divide on a hierarchy of gender and race.

To clarify what is meant by formal and informal work, a brief definition should first be provided. In terms of juridical norms, the International Labour Organization (ILO) initially defined informal work as undocumented, unregulated, and primarily self-employed activities (Barchiesi, 2019). However, the concept is now to be understood as a series of employment arrangements that, either legally or in practice, are not foreseen by national labour laws, the tax system, or social protection measures, meaning that workers are typically not entitled to employment benefits such as notice before dismissal, severance payments, or paid leave for holidays or illness (ILO, 2023).[22] As Barchiesi (2019) explains, the concept of informality nowadays considerably overlaps with that of precarity, especially in the context of global market liberalisation of employment. In sociological discourse, the term "precariat" (*ibid.*: 45) commonly refers to socially and economically vulnerable groups who rely on insecure and unprotected forms of work. Precarity has increasingly come to reflect the concept of the informal economy, particularly as it is used in analyses of the

[21] The concept of legal segmentation has already been presented in Section 3.2. Here, however, it is further contextualised through a historical and social lens by considering the legacy of colonialism in Africa, outlining how formal and informal labour structures intersect(ed) with gendered divisions of work, ultimately resulting in a system of "double segmentation" (Fechner, 2022: 619).

[22] International Labour Organization. (2023). *What is informal employment*? ILO Brief. https://www.ilo.org/media/5481/download [last accessed: 25 June 2025].

Global South (*ibid.*). Thus, informal labour encompasses both the absence of legal status and the socio-economic disadvantages of unstable and insecure work conditions.

Fechner (2022) argues that the distinction between formal and informal labour in African colonial contexts was constructed in a racialised and deeply gendered manner from the very beginning. In other words, colonial laws did not regulate the labour market in a neutral way; rather, they actively defined who could participate in the formal economy and who was to be excluded: only the able-bodied, tax-paying African man was considered a suitable subject for legal labour regulation through formal contracts, tax obligations, and pass rules (*ibid.*). Women, by contrast, were systematically excluded from regulated wage labour and relegated to marginal or invisible economic roles, such as unpaid domestic or agricultural work (*ibid.*). This legal exclusion of women was not accidental, but the result of a deliberate design of colonial law, which institutionalised what Fechner (2022: 619) calls "double segmentation": a first segmentation between formal (regulated) and informal (unregulated) labour, and a second, intersecting one, between male and female workers.

It is important to underline that during colonial rule, the idea of "formal" employment was shaped by the needs of the colonial system and was often exploitative by nature: legal measures were mainly aimed at recruiting African men for European employers. Forced labour, forced labour migration, and taxation helped ensure a constant flow of male workers and kept them tied to colonial employers, often under harsh and restrictive conditions (Fall and Roberts, 2019). Tax systems targeting only men, recruitment laws designed around male mobility, and employment contracts tailored for male workers established economic categories and social hierarchies: "Hut and poll taxes hence not only served to finance the colonial administration but also introduced a gendered duty (and privilege) to participate in the money system" (Fechner, 2022: 623). This legal marginalisation of women's labour, far from being incidental, was a deliberate consequence of colonial governance. As a direct outcome of this legal exclusion rural poverty increased: "Since relatively few women participated in these labour markets, the many women who remained in rural areas experienced increasing poverty precisely because so much male labour was absent working in the capitalist sectors" (Fall and Roberts 2019: 105).

Colonial legislation ensured that most recognised wage labourers were African men, employed in service of colonial interests and engaged in a constant struggle to uphold living conditions (Fechner, 2022). Female labour was not regulated because it did not align with the economic

interests of the colonial administration. Most forms of labour associated with women or considered peripheral to colonial economic goals were systematically excluded from formal regulation and instead pushed into informality, where they were either governed by local customary norms or explicitly prohibited by law (*ibid.*). As Tsikata (2011: 325) highlights, these policies relegated women "to the role of housewives and to the margins of the colonial economy", rather than recognising them as independent workers.

What makes Fechner's work (2022) especially significant for discourse analysis is the idea of how law functions as a producer of social meaning and not as a mere reflection of economic policies. He conceptualises the law as a form of power that shapes how people see themselves and defines who is allowed to take part in the economic life of society. Hence, language and law are instrumental in constructing the boundaries of economic participation. By analysing the discursive patterns that shape representations of female labour in the legislative texts of Kenya and Tanzania, this study seeks to uncover how colonial and postcolonial legal systems have encoded and sustained structural gender inequalities through language.

4.3 The colonial and post-colonial condition of women in Kenya and Tanzania

A look at the paths of female labour in Tanzania and Kenya reveals how colonial legacies have shaped women's roles in the workforce. Beginning with Tanzanian women's employment conditions, both during the colonial and contemporary periods, the focus will then shift to Kenya.

The historical and contemporary condition of female labour in Tanzania reveals a long-standing pattern of marginalisation and structural inequality. During the colonial period, women and children formed the cheapest and most exploited segment of the labour force (Shivji, 1986). They represented a minority of the total workforce, making up approximately 10–15% of wage labour between 1947 and 1951, with child labour accounting for the majority of that figure (*ibid.*). Most of this work was concentrated on agriculture and low-skilled tasks such as weeding, coffee and tea picking, or working on sisal plantations and in ginneries (*ibid.*). Wages for women and young people were extremely low, "[n]ot that the males earned particularly high wages either; but the point is that within the colonial system of extremely low wages, female and juvenile labour was the lowest on the wage scale (*ibid.*: 66).

The introduction of cash crops like sisal, cotton, and coffee for export transformed local economies and created a distorted development model where "we grow what we do not eat and eat what we do not grow" (Mbilinyi, 1975: 403). Women were largely excluded from the modern

economic sectors. They were not recruited to work on plantations, in mines, in administrative roles, nor were they recruited as small-scale producers of export crops. Although they had long played a central role in agriculture, colonial authorities focused their attention on men when introducing new crops and modern farming techniques (Mbilinyi, 1975). Colonial authorities made minimal legal provisions for women in the workforce. Apart from a 1938 ordinance banning night work for women in industrial settings,[23] colonial labour law largely ignored female labour. Existing regulations, such as machinery safety rules, were poorly enforced, leaving women vulnerable to exploitation and unsafe conditions (Shivji, 1986). Interestingly, in contexts where men controlled the cultivation of export crops, many women actively avoided unpaid labour by shifting their efforts toward food production. Over time, this led to a gendered pattern in which men focused on growing cash crops, while women became primarily responsible for food crops used within the household or sold in local markets (Mbilinyi, 2023).

Education followed a similar pattern. Colonial administrations invested in boys' education to employ semi-literate workers for low-level administrative jobs, while girls were largely excluded from schooling: "They were only interested in *male* manpower, and set up government school for boys, to cater for this need. During the German period in Tanzania, not a single government school for girls was established" (Mbilinyi, 1975: 403).

Since becoming independent, Tanzania has made progress in dealing with gender issues. The 2004 Employment and Labour Relations Act (ELRA) "protects women in the workplaces […] through prohibition of discrimination, favourable working conditions and protection against unfair termination" (Ackson, 2015: 32). In addition, efforts to improve women's access to education, employment, and fair wages were supported by initiatives like the National Strategy for Gender Equality (NSGE) and the National Strategy for Growth and Reduction of Poverty (NSGRP) (Ackson, 2015). However, significant challenges remain. Women continue to earn substantially less than men, and around 60% of Tanzanian women live in extreme poverty (UN Women, 2023).[24] Despite making up 70% of the agricultural workforce, only 9% of them own land, and 12% have access to banking services (*ibid.*). Moreover, women carry out approximately three times more domestic care work than men (*ibid.*).

[23] Tanganyika Territory. *Employment of Women Ordinance, 1938*, No. 14 of 1938. Assented by Governor Mark Young on 28 October 1938. Retrieved from the *Worlds of Labour* repository.

[24] UN Women. (2023). Gender pay gap in Tanzania. https://africa.unwomen.org/sites/default/files/2024-03/un_women_tanzania_gender_pay_gap_report.pdf [last accessed: 25 June 2025].

In terms of political participation, Tanzania has taken important steps toward enhancing women's representation in public institutions. The 1977 Constitution introduced a quota system that initially reserved 30% of parliamentary seats for women,[25] which was later increased to 40% following the 2015 general elections (UN Women Africa, n.d.).[26] As a result, women currently hold around 37% of parliamentary seats and cabinet positions, with 18% of deputy ministers and 38% of judges also being women (UN Women Africa, n.d.; IPU, n.d.).[27] These figures place Tanzania among the countries with higher female political representation in the region and reflect progress toward the Southern African Development Community's (SADC) target of 50% representation (UN Women Africa, n.d). However, gender parity remains a challenge, particularly in elected rather than appointed positions. In this context, various initiatives, including the project *Strengthening Meaningful Participation, Leadership and Economic Rights of Women and Girls at the Local Level in Tanzania* (UN Women, 2024),[28] are being implemented to support women's leadership at the local level, improve access to gender-responsive laws and services, and promote socio-cultural conditions for fair participation (UN Women Africa, n.d). Nonetheless, these efforts demonstrate that while legal and institutional progress has been made, the gap between legislation and reality remains wide.

Similarly to Tanzania, the contemporary condition of women in Kenya's labour market is deeply influenced by the country's colonial history and the post-independence economic system. Presbey (2022: 36) explains that before colonisation, many Kenyan women held important social roles "that could not be taken over by men". They served as healers, food producers, and custodians of indigenous knowledge and held usufruct rights to land within communal systems. However, colonial power dynamics drastically altered this balance. During colonial times, British policies on the agriculture economy were focused on exports like coffee, tea, and cotton (Stichter, 1977). These were grown mainly by men, who obtained land titles by the colonial government, which

[25] The United Republic of Tanzania. Office of the Solicitor General. eLibrary. Constitution of the United Republic of Tanzania of 1977. https://www.constituteproject.org/constitution/Tanzania_1977.pdf [last accessed: 25 June 2025].
[26] UN Women Africa. Women's Leadership and Political Participation (WLPP). https://africa.unwomen.org/en/where-we-are/eastern-and-southern-africa/tanzania/womens-leadership-and-political-participation [last accessed: 25 June 2025].
[27] Inter-Parliamentary Union (IPU), Parliament: United Republic of Tanzania. https://www.ipu.org/parliament/TZ [last accessed: 25 June 2025].
[28] UN Women Tanzania (2024). Strengthening Meaningful Participation, Leadership and Economic Rights of Women and Girls at the Local Level in Tanzania https://africa.unwomen.org/sites/default/files/2024-04/abridged_version-strengthening_meaningful_participation_leadership_local_level_in_tanzania120124.pdf [last accessed: 25 June 2025].

dispossessed women ignoring the land systems they were used to (Presbey, 2022). This marked a major shift in women's status and control over resources in an economy that prioritised male labour, marginalising women's work and leading to their economic and social relegation (*ibid.*).

As Sticher (1977) illustrates, women's participation in formal wage labour before 1945 was limited. When employed, they mostly worked in European-owned agricultural estates, performing low-paid and unskilled tasks such as coffee and maize harvesting, while men had better-paid roles. Outside agriculture, options were restricted to domestic work or informal, often criminalised, income sources like prostitution and illicit beer brewing. Women's labour was undervalued: "Despite a new administrative concern for the education and training of women in the 1950s, women's earnings were uniformly less than those of men in all branches of industry and in domestic service, while in agriculture they were about half" (*ibid.*: 21).

Moreover, Christian marriage laws introduced by colonisers further impacted women's freedom: "For example, the 1902 law against divorce took away African women's power by making it impossible for them to return (or threaten to return) to their natal family, as was often done when a woman was mistreated by her husband" (Presbey, 2022: 35-36). Pre-colonial societies in East Africa, such as the Kwaya in Tanzania, were not uniformly patriarchal: matrilineal inheritance was common, giving women considerable freedom (*ibid.*). However, Christian missionary activities started to focus on dismantling these traditions and replacing them with male-dominated structures (*ibid.*). Even so, practices such as woman-to-woman marriages, which are still followed among some ethnic groups like the Akamba, Gikuyu, and Nandi, offered alternative models of female self-rule. In these unions, women could become heads of households and landowners (Presbey, 2022).

Following independence in 1963, Kenya made some progress in recognising women's rights, although some patriarchal structures persist. The 2010 Constitution prohibits gender discrimination and guarantees human rights for all.[29] However, it also recognises customary law, which often disadvantages women in matters of marriage, inheritance, and family life: for example, some customary laws prevent daughters from inheriting land from their fathers (*ibid*.).

[29] Republic of Kenya, The Constitution of Kenya, 2010 (Nairobi: National Council for Law Reporting with the Authority of the Attorney-General). http://www.parliament.go.ke/sites/default/files/2023-03/The_Constitution_of_Kenya_2010.pdf [last accessed: 25 June 2025].

Despite these challenges, women have continued to find ways to participate in economic life. The informal sector remains one of the most important areas for female employment. As Atieno (2006) reported, nearly half (48%) of informal sector businesses in 1999 were owned by women, despite women comprising 53% of the total labour force in the modern sector but holding less than 30% of wage employment (*ibid*.). According to Atieno (2006: 157), "traditional roles, occupational segregation by gender, and lack of access to technology and credit are the main factors that restrict women's access to formal employment". All in all, the informal sector is for many women in Kenya an opportunity to earn a living, but it also comes with serious challenges, like unstable work, low pay, and little to no labour protection.

On top of that, the gender pay gap is still a major issue. As UN Women (2023) reports,[30] Kenyan women make about 73 cents for every dollar earned by men. This gap does not just affect their salaries, it also means lower pensions, weaker social safety nets, and greater exposure to poverty and even domestic violence. Gender inequality in Kenya starts during childhood "when many [girls] are unable to concentrate on their education due to their workload" of chores in the household (Presbey, 2022: 36). Nonetheless, women have made notable gains in political participation: by 2014, the number of women in Parliament had risen to 21.5%, up from 9.9% in 2007, following the application of the affirmative action principle introduced in the 2010 Constitution (Presbey, 2022; Karanja, 2016).[31] In the 2017 elections, three women were elected as county governors and three others as senators, marking a symbolic step forward in women's access to high political offices (Akwei, 2017). Even though Kenya holds the 106th position globally regarding the threshold of one-third female representation in its national legislatures, it appears to be on a path of transformation in gender equality (Presbey, 2022).

4.4 Reasons for country selection: Kenya and Tanzania in a comparative perspective

The choice to centre the case study presented in Chapter 7 on Kenya and Tanzania was based on a combination of historical, linguistic, and practical considerations. Both countries share a colonial

[30] UN Women. (2023). Gender pay gap in Kenya. https://africa.unwomen.org/sites/default/files/2024-03/un_women_kenya_gender_pay_gap_report.pdf [last accessed: 25 June 2025].

[31] Affirmative action, as established in Article 27(8) and Article 81(b) of the 2010 Constitution of Kenya, requires that no more than two-thirds of the members of elective or appointive bodies shall be of the same gender. This provision aims to promote gender equality in political representation and public service. See: Republic of Kenya, The Constitution of Kenya, 2010 (Nairobi: National Council for Law Reporting with the Authority of the Attorney-General). http://www.parliament.go.ke/sites/default/files/2023-03/The_Constitution_of_Kenya_2010.pdf [last accessed: 25 June 2025].

history under British rule, and they continue to adopt English as an official language (Middleton, 2008; Appiah and Gates, 2010). They also share a similar legal history, as they both implemented "the whole set of English laws as they were operating during the colonial times" (Bakari, 1991: 545). Considering their geographic proximity within East Africa as well, they offered a meaningful basis for comparison. Nevertheless, it is important to acknowledge that these countries also present differences: their colonial experiences and independence trajectories differ in significant ways. The following section briefly outlines these historical distinctions to provide a clearer contextual framework for the subsequent case study.

Tanzania, formerly known as Tanganyika on the mainland, became a German colony in 1884 following the Berlin Conference[32] (Middleton, 2008). German rule was characterised by forced labour and exploitation, as well as resistance from the local population against German colonisers (*ibid.*). After Germany's defeat in World War I, Tanganyika became a British mandate under the League of Nations, while Zanzibar remained a British protectorate governed through the Omani sultanate (*ibid.*). Tanzania's path to independence was shaped by collaboration between nationalist movements in Tanganyika and Zanzibar. Julius Nyerere led the Tanganyika African National Union (TANU), while Abeid Amani Karume led the Afro-Shiraz Party (ASP) in Zanzibar (*ibid.*). Tanganyika gained independence from British control in 1961, Zanzibar followed in 1963, and the two united in 1964 to form the United Republic of Tanzania after a violent revolution (*ibid.*). Nowadays, Tanzania remains one of the world's least developed countries, afflicted by external debt and high poverty levels (*ibid.*).

Kenya, on the other hand, became a British protectorate in 1895 after the Imperial British East Africa Company was unsuccessful in establishing economic control (Appiah and Gates, 2010). The construction of the railway from Mombasa to Lake Victoria, completed in 1901, enabled British settlement and military control throughout the region (*ibid.*). Colonial rule in Kenya was characterised by land expropriation, forced labour, and ethnic segregation: the British created what are today known as "White Highlands", land to be made available for white settlers, pushing away

[32] The Berlin West Africa Conference (1884–1885), hosted by German Chancellor Otto von Bismarck, brought together representatives of 14 countries to establish rules for colonial expansion in Africa. Its main objectives were to guarantee free trade and navigation along the Congo and Niger Rivers and to set legal principles, such as the one of the so-known *effective occupation*, for territorial claims, thereby avoiding armed conflict among European powers. No African representatives were involved, and the agreements contributed to intensified exploitation and long-lasting colonial oppression across the continent. See: Federal Foreign Office Political Archive. General Act of the Berlin West Africa Conference, 26 February 1885 (n.d.). https://archiv.diplo.de/arc-en/the-political-archive/general-act-2684414 [last accessed: 25 June 2025].

African communities into reserves (*ibid.*). To secure a steady supply of labour, the colonial administration forced African people into work through taxation and identification laws. As Appiah and Gates (2010: 642) explain: "The Native Registration Act of 1915 helped prevent laborers from fleeing by requiring all African adult males to carry identification whenever they left the native reserves". After decades of settler colonialism, forced labour, and land expropriation, nationalist movements led to Kenya's independence in 1963 and the formation of a republic the following year (Appiah and Gates, 2010).

While Kenya and Tanzania differ in various cultural, social, and political aspects, they share key structural similarities: their geographic proximity in East Africa, their colonial history under British rule, the use of English as an official language, and similar legal systems offer a meaningful basis for comparative discursive analysis. Methodologically, Kenya will serve as the starting point for the analysis due to its high number of legislative texts related to women's labour available within the *Worlds of Labour* leximetric database. Opening the investigation with a broader dataset will provide a more solid foundation for identifying discursive patterns associated with legal segmentation, as a larger volume of material might increase the likelihood of observing a wider range of linguistic tendencies. The analysis will then move to the Tanzanian corpus to see whether similar patterns also appear in a different, though comparable, context. This approach allows for a gradual development of the analysis, starting with a broader base and then testing the method on a smaller dataset. Together, they will allow for a critical analysis of how language both represented and promoted employment relations relative to gendered labour division.

5. Features of Legal Discourse and the Annotation of Legal Texts

5.1 Introduction

As outlined in the introduction to this thesis, and in direct connection with the methodological account presented in Chapter 6, the present chapter lays the theoretical groundwork for the corpus design, selection criteria and annotation strategies that underpin both the collaborative corpus and the four national sub-corpora of Kenya and Tanzania that make the object of the present dissertation. The latter were conceived as a vertical extension of the broader WoL project, to test and apply its methodological framework in greater depth and serve as empirical foundation for the case study presented in Chapter 7.

Chapters 5 and 6 jointly provide the methodological core of this thesis. Particular attention is paid to the challenges encountered during both the collaborative phase and the creation of the four sub-corpora, which are situated within broader theoretical discussions concerning legal discourse. Implementation details are discussed in the next chapter, while this one offers a conceptual overview of the issues that shaped the annotation process and informed the analytical choices made throughout the project.

The chapter is divided into two main parts. The first explores key aspects of legal discourse analysis, focusing on the structural, functional, and phraseological features of legal texts, and discussing how these impact methodological decisions in corpus design. The second delves into the specific challenges posed by legal text annotation.

In framing the methodological challenges discussed in the second part of the chapter, particular emphasis is placed on two contributions, due to their direct focus on structural annotation in legal corpora and their close alignment with the objectives and constraints of this research: Darji *et al*. (2023) and Santosuosso and Pinotti (2020). Their work on annotation procedures offers concrete insights that have proven particularly relevant in the context of the WoLCP.

5.2 Discourse analysis of legal texts

Legal discourse encompasses a wide variety of text types, ranging from legislative texts enacted at different jurisdictional levels to judicial decisions, contracts, legal reports, wills, academic writing, and even spoken courtroom interaction (Goźdź-Roszkowski, 2021: 1519). While each type serves different communicative functions and reflects different institutional constraints, all are unified by a distinctive reliance on language as a vehicle of authority, interpretation, and norma-

tivity. Because of this, the legal field is particularly well-suited for linguistic investigation, especially through corpus-based methodologies that allow for the empirical analysis of recurring patterns across large collections of texts:

> It soon became very clear that Corpus Linguistics has had much to offer many other areas apart from linguistics, especially those where language and other disciplines are intimately bound up. This is particularly true of the disciplinary discourse of law, where language is central to its construction and interpretation.
>
> (Goźdź-Roszkowski, 2021: 1515-1516)

In this context, language is not merely a vehicle of communication, but a core element of law itself. The study of legal discourse is thus essential to both linguistic and legal scholarship.

Legal discourse is typically characterised by linguistic complexity, both at the lexical and structural levels. While this complexity can partly be attributed to the historical evolution of legal language, as shaped by layers of Anglo-Saxon, Latin, and French influences in the case of legal English, it also reflects the institutional pressures and strategic purposes behind legal writing (Tiersma, n.d.: 4). This defensive strategy, shaped by the adversarial nature of the legal process, contributes to the formal complexity that distinguishes legal texts from other genres:

> Virtually any legal document is liable, at some point in its existence, to be picked apart by an opponent eager to exploit a loophole or ambiguity in hopes of wiggling out of an agreement or contesting a will. [...] Those who draft such documents must anticipate these attacks. Therefore, they obsessively try to cover every base, plug every loophole, and deal with every remotely possible contingency. The result is ever longer, denser, and more complicated prose.
>
> (Tiersma: n.d.: 8)

Such levels of detail and redundancy pose specific challenges for discourse analysis. Recognising legal texts as inherently strategic and formally codified discursive constructs is essential for shaping both the methodological approach and the interpretative lens in corpus-based legal discourse analysis.

Thus, the integration of corpus linguistics and legal studies often reveals methodological challenges. In recent years, there has been a noticeable increase in research on legal discourse using corpus-based approaches. This work has been carried out not only by linguists, but also by legal scholars and professionals looking for empirical insights into how language operates within legal contexts (Goźdź-Roszkowski, 2021; see for example Peruzzo 2013, 2014; Scarpa, 2017; Wiesmann 2002, 2006, 2004, 2004b). Much of this research has focused on areas such as legal

terminology, translation or phraseology, especially in relation to how the ordinary meaning of terms is interpreted in legal contexts (*ibid.*). However, as Goźdz-Roszkowski (2021) points out, many legal corpora are designed with linguistic priorities in mind, which may not always align with the needs of legal researchers. This is particularly relevant to interdisciplinary projects like the *Worlds of Labour Corpus Project* (WoLCP), where linguistic and legal goals must be balanced within a shared methodological approach. To address its dual focus on both legal and linguistic goals, the WoLCP adopted a methodological framework that highlights the structural features of legal texts to ensure that its legal relevance is preserved. In projects like this, which aim to uncover discursive patterns in a wide range of legal texts, structural annotation helps bridge the gap between legal and linguistic objectives by providing a shared framework that supports interdisciplinary studies.[33]

However, it is important to note that the value of corpus-based insights in legal research is not absolute. As Goźdz-Roszkowski (2021) explains, its effectiveness depends on whether the legal system, or its legal actors, is prepared to treat patterns of word use and co-selection as meaningful evidence within legal reasoning. In other words, corpus data becomes valuable only if the experts are open to interpreting distributional patterns as indicators of legal meaning or practice. Furthermore, the increasing number of specialized legal corpora has widened the range of research possibilities. Even relatively small but well-constructed datasets, those with just a few million words, can offer valuable insights into specific types of legal discourse (*ibid.*).

Although Biel (2009) outlines four main types of corpora specifically in relation to their applications in legal translation, her typology is also helpful for understanding legal corpora more broadly. She distinguishes between monolingual corpora, monolingual comparable corpora, bilingual or multilingual comparable corpora, and parallel corpora. Parallel corpora, which align original texts with their target texts, are particularly relevant for studying translation processes. In contrast, monolingual and monolingual comparable corpora are better suited for exploring patterns within or across legal systems. The corpora employed in this thesis fall under the monolingual comparable corpora category, as they only consist of English-language texts.[34] According to Biel

[33] The annotation strategy adopted by the WoLCP and for the Kenyan and Tanzanian sub-corpora reflects this dual orientation and is discussed in detail in Section 6.5 of the present thesis, while Chapter 7 presents a case study that concretely illustrates how legal and linguistic perspectives can converge in practice.

[34] Although the WoLCP includes a limited number of translations from Arabic, Amharic and Somali into English, these are treated as if they were original texts due to the lack of accessible source texts and knowledge in those three

(2009), monolingual corpora are particularly useful for studying variation within a single language, for example, by comparing different registers or genres. Goźdź-Roszkowski (2021) adds that the type of corpus chosen should reflect the research aims. Importantly, he also points out that discourse analysis is inherently comparative: to describe any legal discourse meaningfully, it must be compared with other types of discourse, whether legal or non-legal. In this light, the WoLCP and its four national sub-corpora, covering materials from diverse jurisdictions and across different time periods, enable both intralingual synchronic and diachronic analysis. While the legal genre of labour law remains constant, the data allow for the examination of cross-jurisdictional and temporal variation within a single language.

Legal discourse shows a high level of variation, "depending on geographical location, degree of formality, speaking versus writing, and related factors. The language and style of lawyers also differs substantially from one genre of writing to another" (Tiersma, 1999: 139). Biel (2009: 4-4) outlines four main levels at which this variation can be examined: external variation (comparing legal discourse with general or other specialised languages), internal variation (comparing different legal text types), diachronic variation (tracing changes over time), and cross-linguistic variation (comparing legal language across languages). It is important to note that variation analysis is most effective when it considers both the statistical presence of linguistic features and their functional significance, as frequency data can reveal recurring linguistic patterns, but the findings must be interpreted in terms of the communicative roles they play within legal contexts (Goźdź-Roszkowski, 2021).

Another essential aspect of corpus-based legal research is the choice and availability of appropriate analytical tools. The interpretive power of a corpus depends not only on its design but also on the software used to explore it: "It is often pointed out that no matter how good a corpus can be in terms of representativeness, balance, size, etc., it is practically worthless without a suit-

languages, determining the de facto monolingual framing of the corpus. Although these texts are treated as English within the corpus for analytical purposes, the metadata structure developed for the WoLCP allows for clear distinction between texts originally written in English and those translated into English. In the metadata schema (see Appendix E for an overview), each document includes specific fields such as *translated_status*, *translation_source_language*, and *translation_method*, which ensure transparency about the origin and status of each text. This system allows users to identify translations as non-original English texts and to distinguish them from genuinely English-language documents.

able software to explore it" (Goźdz-Roszkowski, 2021: 1520). Concordancer and frequency calculators remain core tools in this field, and applications like AntConc[35] or platforms like Sketch Engine are widely adopted due to their multilingual resources and support for both standard and custom corpora (Goźdz-Roszkowski, 2021). In the present project, Sketch Engine was selected as the main environment for corpus exploration, allowing for keyword extraction, collocation analysis, and structural searches across annotated legal texts (see Chapter 7 for further details).

A key concept in the linguistic study of texts in general, and legal texts in particular, is phraseology:

> '[C]orpus linguistics phraseology' […] prioritizes differential frequencies as a way to identify patterns of repetition and patterns of co-selection as significant elements in a corpus of texts under study. This means that studies of phraseology should not only pay attention to lexical and grammatical co-occurrence (collocation and colligation, respectively) but also to textual co-occurrence, when lexical items occur differentially in different parts of a text, such as in paragraph or text initial position.
>
> (Goźdź-Roszkowski, 2021: 15521-1522)

The distinction between corpus-based and corpus-driven approaches, originally formulated by Tognini-Bonelli (2001), represents a foundational methodological choice in corpus linguistics. In a corpus-based approach, the analysis starts from pre-selected linguistic features, such as specific expressions, collocations, or grammatical structures, and uses corpus data to test hypotheses or verify existing claims. In contrast, the corpus-driven approach adopts a more inductive perspective, aiming to let patterns emerge directly from the data without prior assumptions, thus allowing for the discovery of previously unnoticed linguistic characteristics. Goźdź-Roszkowski (2021) mentions both approaches in relation to the study of legal discourse as well.

One of the most widely studied phrase types in legal discourse is lexical bundles: frequent, uninterrupted sequences of words that are often semantically transparent and serve a structural role in organising legal information (*ibid.*). They help create cohesion and provide a framework for presenting legal arguments and definitions (*ibid.*). However, their rigid structure can limit their effectiveness in capturing the full range of variability typical of legal language. To address this issue, researchers have introduced more flexible models like concgrams and semantic sequences,

[35] AntConc. Laurence Anthony's Website. A freeware corpus analysis toolkit for concordancing and text analysis. https://www.laurenceanthony.net/software/antconc/ [last accessed: 25 June 2025].

which can account for variation in word order and allow for a deeper understanding of legal phrasing (*ibid.*).[36] These models are especially useful for identifying formulaic language that varies structurally while maintaining its function. Corpus research plays a crucial role in classifying these expressions and understanding how they contribute to the communicative goals of different legal genres. Indeed, cross-genre studies have shown that legal genres tend to rely on different kinds of lexical bundles, reflecting the specific purposes and audiences of each text type (*ibid.*).

The next section will focus specifically on the challenges associated with annotating legal texts. It will examine both technical and conceptual difficulties, and, as in the current discussion, it will draw on examples from the WoLCP and the corpus developed for this thesis to explore how annotation choices impact the outcomes of legal-linguistic research.

5.3 Challenges of legal text annotation

Although data annotation is a widespread practice across many domains, the legal field poses a distinct set of challenges that require specific approaches. This is largely due to the unique nature of legal texts, which are characterized by complex structures, highly specialized terminology, and a strong dependence on contextual accuracy (Darji et al., 2023: 1).

As Goźdz-Roszkowski (2021: 1519) explains, legal corpora typically involve two main types of annotation: part-of-speech (POS) tagging and structural mark-up. POS tagging "involves assigning to each tokenized word a label that minimally identifies the part of speech of the word but that typically also includes some grammatical category information" (Gries and Berez, 2017: 383). Structural mark-up, on the other hand, involves "indicating structural units, such as introductions or closing sections in a document" (Goźdź-Roszkowski, 2021: 1519). While POS tagging

[36] Concgrams are a type of linguistic pattern that focuses on how two or more words appear together in different ways within texts. Unlike more rigid patterns, concgrams are flexible: they do not require the words to be next to each other or in a fixed order and they include all the possible ways the chosen words can occur together, no matter their position or how the sentence is structured. For example, if we look at the words 'increase' and 'expenditure', a concgram would include: (i) increase in expenditure (straightforward co-occurrence), (ii) increase in the share of expenditure (where more words appear between them), (iii) expenditure would inevitably increase (where the order of the words changes). So, concgrams allow researchers to capture a full range of how word pairs show up in real language use, regardless of how they are arranged in a sentence (Goźdź-Roszkowski, 2021). Meanwhile, semantic sequences emphasise meaning relationships and represent an even more flexible way of analysing patterns, especially in legal genres that tend to use more idiosyncratic language (*ibid.*). For instance, in judicial opinions, Goźdź-Roszkowski (2021) examines the *N that* construction (e.g., the idea that), which signals legal propositions. These sequences are not frequent in their exact wordings, but they perform consistent functions. For example, numerous unique phrases like *the argument that*, *the theory that*, or *the claim that* all serve to introduce evaluative or justificatory statements in legal reasoning. Thus, semantic sequences offer a methodologically flexible tool to identify regularities in legal texts, especially where language is highly variable and genre-specific.

is often automated, structural annotation is a more manual and time-consuming task that requires a solid understanding of legal document formats and variation (Goźdź-Roszkowski, 2021).

Darji et al. (2023) outline a list of issues that corpus linguists or legal experts typically encounter when annotating legal documents. Some of these issues also became evident during the text selection and annotation phases conducted within the framework of the WoL collaborative project and during the development of the present research. Nonetheless, it is important to note that the type of annotation described by Darji et al. (2023) differs noticeably from the approach adopted in the latter contexts. Their work primarily focuses on the extraction and annotation of legal references within texts, like "§ 256 Abs. 1 ZPO" or "§ 14 Abs. 3 Satz 2 SchVG". In their process, each legal reference is first identified, then broken down into its components (e.g., article, paragraph, legal code), and finally linked to its official definition retrieved from authentic online legal sources. Specific tags are then added to indicate the structural elements of each citation. The aim of this approach is to create annotated datasets that can be used to train NLP models to predict the appropriate legal reference from a given textual input (*ibid.*).

In contrast, the annotation strategy employed for the present thesis is not concerned with extracting references to external legal sources. Instead, it focuses on marking the internal structure of the legal texts themselves. This involved identifying the beginning of structural units, such as articles, chapters, or parts, recording section titles, and embedding metadata. The primary goal of this structural annotation was to make the hierarchical organization of each legal document explicit, rather than to extract intertextual legal references. However, even if some of the specific tools and models, like the fine-tuned BERT model, used by Darji et al. (2023) are not directly applicable to the structural-oriented goals of the *Worlds of Labour Corpus Project* (WoLCP), their discussion remains relevant, as their work highlights the broader complexity of legal annotation and underscores the fact that different annotation tasks require tailored methodological and technical solutions depending on the goals of the project and the nature of the legal data being processed.

When it comes to the practical challenges of building and analysing a legal corpus, one of the main difficulties lies in the availability and quality of legal datasets. Darji et al. (2023) highlight how in the German legal domain, for example, existing corpora often focus on entire legal cases, while datasets providing structured representations of legal references or well-segmented legal texts remain rare (*ibid.*).

As discussed in Section 3.4, this represented one of the major issues that also affected the development of the original WoL legislative dataset. In less-resourced legal systems, like some across the African continent, the problem is even more pronounced, making the data collection phase a particularly demanding preliminary step.

Moreover, the nature of the source material itself can often be problematic. Legal documents typically include footnotes, cross-references, and formalized yet inconsistent structures, which complicate both the extraction and cleaning of text for annotation (*ibid.*). The lack of standardisation in formatting, particularly across different jurisdictions and legal traditions, further exacerbates these difficulties and necessitates extensive pre-processing before annotation can begin. This was also one of the most significant challenges encountered during the annotation process, both in the collaborative phase and in the construction of the four sub-corpora. In the context of the WoLCP, the heterogeneity of structural conventions made it especially difficult to determine how to segment and annotate texts consistently. If legal documents had followed similar formatting practices, such as using uniform terminology for hierarchical units or adopting predictable heading structures, the annotation could have focused more directly on the semantic layer, particularly the thematic labelling of headings. Instead, substantial time was required to identify and reconcile differences in how structure was expressed, not only across different countries, but sometimes even within individual texts.[37] The annotation strategy adopted to address these inconsistencies, particularly regarding the segmentation of structural levels, is presented in Section 6.5.3.

The importance of structure homogeneity is indeed widely acknowledged. As highlighted by Santosuosso and Pinotti (2020), legal documents that have a rigid and predictable structure are far more suitable for text-based analytical approaches. A standard internal structure not only enhances text readability and usability but also creates a more robust foundation for future applications in corpus-based linguistic research. The judgments of the European Court of Human Rights (ECHR) serve as an example: their decisions are mandatorily divided into well-defined sections, such as facts, legal questions, and operative parts, which makes their document structures standardised. By contrast, in jurisdictions where drafting practices have traditionally been less formalised, such as Italy until recent reforms, additional annotation work is often needed to impose a

[37] For instance, in the case of the Equatorial Guinea file in the WoLCP corpus (later excluded due to language criteria, see Section 6.5 for further details), headings such as "Part," "Title," "Section," or "Chapter" were used inconsistently, and they appeared to belong to the same hierarchical level. This was only revealed through close inspection of the numbering system, which showed a consistent, sequential structure across otherwise differently labelled units.

coherent structure retrospectively (*ibid.*). However, in Italy progress has been made through institutional reforms: in 2018, a Memorandum of Understanding[38] between the judiciary and the legal profession introduced guidelines for improving the clarity and structure of judicial decisions (*ibid.*). Such measures support the notion that imposing structure during drafting phase reduces the need for costly manual annotation later, allows for smoother integration of metadata and, in the context of the WoLCP, enables easier annotation.

Another widespread problem lies in the incompatibility between available annotation tools and the specific requirements of legal texts. Software like Microsoft Word[39] or general-purpose NLP libraries are not optimised for the format of legal texts (Darji et al., 2023). Annotation tasks may require reading and editing XML files, writing XPath expressions, or developing Python scripts to identify and tag structural units (*ibid.*). As a result, even the selection of an appropriate annotation environment becomes a complicated process (*ibid.*). While this did not represent a major challenge for the WoLCP, since the annotation environment was defined early on as plain text files compatible with Sketch Engine, this choice nevertheless imposed specific constraints on the overall data preparation workflow. In particular, the annotation process was preceded by technical difficulties due to the format and accessibility of the source files, which often required Optical Character Recognition (OCR) processing to produce machine-readable text. These aspects will be further discussed in Chapters 6 and 7.

An additional issue highlighted by Darji et al. (2023) lies in the frequent need of manual annotation of sections or specific elements that cannot be addressed programmatically. This task is a labour-intensive and knowledge-dependent process, often necessary due to the complexity and semantic richness of legal texts. Darji et al. (2023) explicitly address this issue, noting that certain types of metadata, such as the law title or specific paragraph text, cannot be extracted automatically and must instead be retrieved manually from official legal sources. This underscores how annotation not only requires technical precision but also relies heavily on domain-specific expertise. As Santosuosso and Pinotti (2020) point out, this constitutes a key aspect of the legal annotation 'bottleneck': the indispensable role of human experts in interpreting and formalising legal content for computational applications.

[38] Protocollo d'intesa tra il Consiglio Superiore della Magistratura e il Consiglio Nazionale Forense, 2018. https://www.consiglionazionaleforense.it/documents/20182/462917/Protocollo+d%27intesa+CNF+-+CSM+%2819-7-2018%29.pdf/81d49d12-8cb6-4573-a8b1-976cf09501bc [last accessed: 25 June 2025].
[39] Microsoft 365. https://www.microsoft.com/de-ch/microsoft-365/word?market=ch [last accessed: 25 June 2025].

Both Darji et al. (2023) and Santosuosso and Pinotti (2020) operate within the fields of artificial intelligence and law, and their work explicitly aims to build annotated datasets for training NLP models. In contrast, neither the WoLCP nor the present thesis involved the development of NLP tools or predictive systems. As previously explained, the primary focus was on the tagging of structural elements and metadata within legal texts. Nevertheless, the issue of manual annotation emerged as a significant challenge here as well, particularly due to the lack of standard formatting in the available documents. In several cases, manual revision was necessary because the Python scripts developed for the WoLCP annotation process, which were also used in the creation of the four sub-corpora discussed in Chapter 7, were unable to consistently detect structural boundaries or correctly place the annotation tags. The scripts were designed to insert XML tags delimiting structural units (e.g., articles and sections) based on predefined regular expression patterns. However, their logic assumed that every new structure would automatically signal the end of the previous one. This assumption led to errors in tag placement, particularly in documents with irregular formatting or overlapping structural hierarchies. For example, closing tags were often inserted immediately before each new opening tag, regardless of whether the preceding structural unit had actually ended. In some cases, such as at the beginning or end of a document, or when titles of different levels appeared in succession, this resulted in misaligned or misplaced tags that required manual correction (see Section 6.5.3 for further details on problematic placing of level 0 and level 1 structural tags).

Darji et al. (2023) also note that, although regular expressions (RegEx) can be powerful tools for extracting structured information from legal texts, their effectiveness is limited by the inconsistency of legal drafting conventions. Legal references often follow variable syntactic formats depending on jurisdiction, legal code, or historical period. Their study highlights several cases in which regex-based extraction fails to generalise across variants, noting that slight differences in formatting, such as the use of "§", "§§", "Art.", or their unabbreviated forms, can significantly impact pattern recognition and retrieval accuracy. They conclude that while a regex can work efficiently on well-defined patterns, its precision declines sharply when confronted with heterogeneous or non-standardised input (*ibid.*).

Similar challenges emerged in the annotation of the WoLCP dataset, as well as in the four Kenyan and Tanzanian sub-corpora. While RegEx were successfully implemented to semi-automate the tagging of structural elements such as articles and sections, their application required

extensive preparatory analysis. The internal headings of each document, serving as anchors for level 0 and level 1 segmentation, had to be manually scanned to identify recurring formatting conventions. Custom regex patterns were then developed for each type of heading. However, due to frequent inconsistencies, such patterns often captured only a portion of the relevant elements. In documents where multiple styles of headings coexisted or where exceptions disrupted otherwise regular patterns, manual intervention was required to ensure annotation accuracy. This process, consisting of manual review, regex calibration, and manual correction, proved time-consuming but necessary, especially in the absence of standardised formatting across and within texts.

This chapter has provided a comprehensive overview of the theoretical and methodological foundations that underpin the annotation and analysis of legal texts within the *Worlds of Labour Corpus Project* (WoLCP) and its four sub-corpora. Key issues included the segmentation of documents, the use of regular expressions (RegEx) for semi-automated tagging, and the difficulties caused by inconsistent formatting across different legal systems. The chapter also highlighted the essential role of manual intervention in correcting tagging errors. The next chapter moves from the conceptual to the operational level, describing the concrete steps taken during data preparation, annotation, and corpus construction, further elaborating on the challenges outlined here and the solutions developed to address them.

6. The Worlds of Labour Corpus Project: From a Legal Database to an Annotated Corpus

6.1 Introduction

As outlined in the Introduction to this thesis, the methodological approach presented in this chapter was developed within the context of a collaboration between the Department of Interpretation and Translation (DIT) of the University of Bologna and members of the research team of the *Worlds of Labour* (WoL) project at the University of Bremen. The goal was to integrate corpus-based linguistic methodologies with the existing legal database developed within the WoL framework. This collaboration laid the groundwork for the corpus construction and annotation procedures that were later adapted and refined in the development of the sub-corpora of Kenya and Tanzania described in Chapter 7. Rather than offering a detailed account of the collaborative project, the present part exclusively focuses on the shared methodological procedures and strategies that informed the creation of the four sub-corpora.

This chapter is thus structured around the following key procedural components: the selection and digitisation of legal texts, the design and implementation of metadata fields, the development of annotation layers (macro-, micro- and omitted-content tagging) and of Python scripts for automation purposes. Certain contextual elements and general background information on the original research setting are included for clarity and introduced only when essential for comprehension. Where appropriate, they are supplemented by explanatory footnotes to maintain consistency.

A more detailed account of the sub-corpora construction process, including concrete examples drawn from the legal texts selected for the case study analysis, will be provided in Chapter 7. To avoid unnecessary repetition and to maintain ease of understanding, cross-references between the two chapters are included where relevant. The decision to separate these sections reflects a deliberate effort to distinguish between the collaborative methodological foundation and its application in the context of the thesis project that builds upon it.

As a final point worth noting, this chapter was developed in parallel with the drafting of the paper currently being prepared by the DIT group as part of the WoLCP. As the project and its methodology were jointly carried out and discussed during the collaboration, many of the conceptual elements presented here naturally reflect the content of the forthcoming paper. However, while

the analytical foundations are shared, the formulation, structure, and wording of this chapter are entirely original and have been independently developed for the purposes of this dissertation.

6.2 WoLCP's aims

The *Worlds of Labour Corpus Project* (WoLCP) was established as a complementary initiative to support and extend the objectives of the broader *Worlds of Labour* research project. Its main purpose was to apply corpus-based discourse analysis as a means of moving beyond the purely doctrinal or codified content of legal texts to allow a linguistic investigation of their discursive features. As such, the collaboration aimed to transform part of WoL legislative database, with a particular focus on countries from the Global South, into a structured and searchable annotated corpus to make legal data more accessible and analysable. This would help standardise the resource to improve accessibility and facilitate its use and dissemination, not only from a comparative labour law perspective, but also within the field of discourse analysis. Furthermore, the compilation and annotation processes represented an opportunity to explore the challenges of working with legal texts of under-resourced contexts. Although the project incorporated additional components, such as training sessions on corpus tools and the drafting of technical documentation for internal and external use, these are not discussed here, as they fall outside the scope of the present thesis.[40]

6.3 Beginning phase: defining the WoLCP dataset

The dataset used for the *Worlds of Labour Corpus Project* (WoLCP) derives from the legislative repository compiled by the WoL research team, which covers 165 countries and 5 territories and spans the period 1880-2023 (Carlino and Fechner, 2025).[41] To create a manageable corpus for linguistic annotation, a selection protocol was established, taking into account the volume of available material and the time and resource constraints of the project.

[40] These activities included a presentation delivered to the Bremen research team on the use of Sketch Engine to equip them with the tools needed to share this knowledge with future collaborators, as well as the drafting of a user manual for the annotated corpus. As already briefly mentioned, the group also initiated the drafting of a research article describing the internship project and its applications.

[41] The database includes a wide range of legal instruments, such as codes, decrees, acts, proclamations, and amendments, all containing provisions that "should reflect political decisions with direct legal impact on workers and employers" (Carlino and Fechner, 2025: 7). Access to the full legislative archive and supporting documentation was facilitated by the WoL research team through the University of Bremen's Nextcloud server.

Consistent with the WoL project's emphasis on legal segmentation in the Global South, the WoLCP corpus concentrated exclusively on African countries.[42] During the initial planning phase, two complementary approaches were taken into consideration: a 'vertical' and a 'horizontal' strategy. On the one hand, the vertical strategy involves the collection of numerous texts from a limited number of countries and allows for a diachronic study of legal developments within those specific jurisdictions. On the other hand, a 'horizontal' approach favours a smaller number of documents from a larger pool of countries, allowing for a broader comparative analysis of contemporary legal frameworks while maintaining the feasibility of the corpus construction process. For the purposes of WoLCP, the vertical approach was ultimately set aside for two main reasons. First, it risked producing a corpus that was poorly representative of the African or Global South context, given the overrepresentation of just a few jurisdictions. Second, it posed significant practical challenges, particularly in managing interdependent legal texts, such as matching amendments with their corresponding parent laws. At the same time, the horizontal strategy does not exclude vertical analyses. On the contrary, the WoLCP corpus was designed to support future extensions, such as the present dissertation that seeks to apply a vertical perspective to the Kenyan and Tanzanian subcorpora, reflecting the need for a more detailed analysis of legal discourse within specific national contexts. In this way, the WoLCP also serves as a valuable reference corpus that can support both broad overviews and country-specific research.

Another methodological choice concerned the level of inclusion: each legal text was retained in its entirety, rather than extracting only the sections directly related to the WoL project's coding variables. This choice was made to ensure the integrity of the texts and to allow the corpus to be repurposed for future research beyond the immediate scope of the project. Moreover, the corpus reflects the linguistic diversity of African legal systems. While a significant portion of the texts were in English, many were originally written in French and Portuguese. The multilingual nature of the dataset raised important considerations for the annotation process, particularly in relation to tagging, which will be discussed in Section 6.4.

[42] Among all African countries, only Guinea-Bissau was excluded from the project's scope due to its population being below the 500,000-threshold set by the original research parameters (Carlino and Fechner, 2025), and was therefore absent from the WoL database

6.4 Text selection and pre-processing phase

The legislative material selected for the annotation was drawn from the country-specific folders available in the WoL project's internal Nextcloud repository. Each folder included legal texts and accompanying Excel templates containing coded provisions, compiled as part of the leximetric analysis conducted by the research team. These templates listed individual norms identified as relevant for the WoL project, extracted from a variety of legal sources and supplemented by additional contextual information.[43] A sample excerpt of one such coding template is included in Appendix A to provide a concrete example of the structure and content of these documents.[44]

In line with the horizontal approach, the identification of legislative texts to be included in the corpus relied on a frequency-based selection strategy. The *legal references* column in the coding templates that could be found in the WoL Nextcloud database served as a central resource for the identification of the most frequently cited and thematically relevant legal documents in each national context. This frequency-based selection strategy was not only adopted for the compilation of the WoLCP corpus but also served as a foundational criterion in the creation of the four sub-corpora analysed in the next chapter. In particular, the texts identified were used to construct a reference corpus against which the Tanzanian and Kenyan corpora were compared during the case study analysis to form the basis for the extraction of keywords. Nevertheless, a two-step process was implemented to extract and rank legislative references based on their frequency of citation. First, regular expressions were used to clean and standardise the list of references, isolating the titles of the legislations. Then, a Python script was developed to calculate the frequency of each reference.[45] The resulting ranked list of legislative texts was recorded in an Excel file, which was

[43] No completed Excel template was available for Angola, Botswana, Comoros, Djibouti, Gambia, Liberia, Libya, Madagascar, Mauritius, Mozambique, Seychelles, and Togo, as the codification process for these countries had not yet been carried out at the time this task was conducted.

[44] A more detailed explanation of the coding criteria and methodology developed by the research team in Bremen can be found in Section 3.3.

[45] The script was developed by fellow students Angela Forzatti and Joanna Giacobbe during their internship as part of the collaborative WoLCP project and was kindly shared with the team. It was implemented during the text selection phase for the WoLCP and the four national sub-corpora of the present thesis in order to automatically count the frequency of legislative references extracted from the *legal reference* column in the Excel templates inside the WoL repository. The purpose was to identify the most frequently cited legislative texts for each country to support the selection process of texts to be included in the corpora.

then compared against a separate list of key texts identified by the WoL research team.[46] Discrepancies and overlaps were reviewed and resolved through consultation with the Bremen group.

Once a final list of texts had been consolidated, each document was evaluated in terms of its technical usability to convert it into a machine-readable format. Many files were available only as image-based PDFs, making automated text extraction unreliable. Optical Character Recognition (OCR) tools were used to recover the machine-readable text. In cases where OCR produced corrupted outputs, manual corrections were required, and when no usable version could be recovered, the document was excluded from the initial corpus.[47] A record of all included texts was compiled in a working document tracking the corpus contents (see Appendix C). The corpus includes texts in multiple languages, primarily English, French, and Portuguese; and to preserve linguistic consistency and facilitate annotation, it is intended to be organised into three separate, non-aligned sub-corpora corresponding to each language. Texts identified as translations were excluded, except for English versions of documents originally issued in Arabic, Amharic, and Somali.[48] This means that the English sub-corpus includes both original and translated versions from the other two languages, with translation sources recorded in the metadata.[49]

6.5 Annotation phase

Following the selection and pre-processing of the legislative texts, the next step involved the development of the annotation framework. As mentioned in Section 6.4, in order to streamline the process, the dataset was further restricted to include only legal documents originally drafted in English, as verified through the official language(s) in use at the time of publication (see Appendix E).

During the annotation design phase, specific issues emerged concerning the treatment of non-linguistic or structurally peripheral elements, such as tables of contents, signatures, seals,

[46] These variations were primarily due to three factors: (i) some legislative references in the Excel templates served as historical foundations for more recent laws (e.g., Benin, Chad, Rwanda); (ii) others corresponded to earlier versions of currently valid legislation (e.g., Democratic Republic of Congo, Ethiopia, Ivory Coast, Tunisia); and (iii) in a few cases, the references pointed to colonial-era legislation that was no longer applicable (e.g., Cameroon).

[47] For example, the Tanzanian Employment Act (No. 11 of 2005) required extensive manual intervention to extract text from image files.

[48] These translations were included only when the original texts were inaccessible or unintelligible to the team. Their translation status and source languages are explicitly recorded in the corpus metadata.

[49] While the initial annotation work focused on English-language documents, this multilingual structure allows for future expansions into the remaining French and Portuguese sub-corpora.

forms, and images. These elements were systematically excluded from the annotated corpus, although for different reasons. Some, like tables of contents and repeated headers, would have skewed frequency-based analyses by artificially inflating the presence of certain keywords or phrases. Others, such as signatures, seals, and form templates, are irrelevant from a discursive perspective and risk introducing noise into the analysis. The decision to omit them reflects the functional aim of the corpus, which is not philological but directed toward the study of legal discourse and the extraction of meaningful linguistic patterns.

Omission was achieved through a preliminary manual annotation phase, referred to as light annotation, which introduced a tagging convention designed to signal omitted content without disrupting the textual structure. This allows users to identify and locate excluded content in the original PDF versions, should they wish to consult the complete document. Moreover, the tags contributed to preserving document composition and internal organisation, while ensuring that only discursively relevant content was retained in the corpus. This strategy has also informed the development of the sub-corpora analysed in Chapter 7.

tags	**explanation**
<omitted type="unreadable"/>	illegible or corrupted text
<omitted type="seal"/>	institutional seals or emblems
<omitted type="image"/>	embedded images or non-textual elements
<omitted type="signature"/>	handwritten or typed signatures
<omitted type="table"/>	tabular data not suitable for linguistic analysis
<omitted type="table of contents"/>	redundant or structurally repetitive content
<omitted type="other language"/>	texts in languages other than English
<omitted type="form"/>	legal forms or templates
<omitted type="explanatory note"/>	supplementary notes or comments
<omitted type="link"/>	hyperlinks or references to external content
<omitted type="unreadable"/>	illegible or corrupted text

Table 1: Omitted tags used in the light annotation process to annotate excluded elements.

Beyond the removal and tagging of peripheral elements and the automatic insertion of metadata (see Section 6.5.1), the annotation strategy also included the identification of internal structural levels within the legal texts. Specifically, the project's annotation system reflects the hierarchical organisation of the texts, capturing elements such as articles, sections, and titles, enabling targeted corpus-based analysis. One major benefit of structural markup is that it allows researchers to limit their queries to specific parts of a legal text. For instance, they can focus only on section headings or isolate opening and closing parts, thus supporting both qualitative and quantitative explorations

of how particular concepts (e.g., female labour) are framed within different legal systems. Structural annotation also facilitates comparisons between functionally equivalent units across texts, and it supports the automatic extraction of defined elements like section titles, which can then be analysed for recurring terminology or thematic concerns. Moreover, structural annotation makes it possible to create coherent sub-corpora based on document structure. For example, a sub-corpus might consist exclusively of preambles, provisions, or sections related to specific legal areas, such as labour or gender. This allows for more controlled comparative analysis, both within and across documents. In projects like the WoLCP, which aim to uncover discursive patterns in a wide range of legal texts, structural annotation helps bridge the gap between legal and linguistic objectives by providing a shared framework that supports interdisciplinary research.

The decision to introduce structural annotation was driven by the observation that, despite considerable variation in formatting across legal systems, all documents shared a common foundational unit: the article. As a result, a tagging scheme was introduced to mark article-level divisions consistently across the corpus. This initial structural layer was later expanded to include higher-level divisions when available, particularly in cases where article titles were absent and thematic organisation could be inferred from section headings. These structural annotations form the basis for the micro-level markup scheme that will be discussed in the following sections and are also applied to the sub-corpora developed for the case study presented in Chapter 7.

The annotation scheme designed for the WoLCP English corpus (WoLCP-EN1) was organised around two principal layers: metadata and structural markup. The structural layer was further subdivided into macro- and micro-structural elements. The following section presents the technical implementation of these layers in the corpus.

6.5.1 Metadata

Metadata are standardised contextual details "about the texts in the corpus: for example, year of publication, author name, publishing house, medium (written, spoken), register (formal, informal) etc." (Sketch Engine, n.d.) that allow users to create targeted subsets and narrow the scope of a search (e.g., filtering for legislative acts enacted after or during a specific year). A distinction was introduced between core and non-core metadata,[50] considering the full set of legislative texts that

[50] Non-core metadata will be incorporated in future versions of the corpus and adjusted as necessary. Examples include attributes such as "institution or enacting body", "amendments or modifying laws", and "colonial history". See Appendix D and Appendix G for an overview of all core metadata considered for both WoLCP and WoL-T, WoL-K.

will constitute the entire corpus in the future, not only the English documents included in Appendix E. This allowed for the integration of additional fields in anticipation of the future inclusion of French and Portuguese sub-corpora. Among these, specific metadata were designed to indicate whether each text was translated and to specify the translation method. This foresight aimed to ensure coherence and traceability throughout the dataset from the early stages of development. The complete set of metadata attributes, values, and their corresponding descriptions is presented in Appendix D.

Before the automated insertion of metadata could be performed, the corpus underwent another manual structural annotation phase. In addition to the light annotation described in Section 6.5, each document was also annotated with macro-structural elements (see Section 6.5.2 for further details). This included the identification and markup of bigger structural divisions such as introductions and conclusions. These annotations served both to preserve document organisation and to account for any structural components lost during the text conversion process. Only after the completion of this manual phase was a Python script employed to automate the integration of metadata.[51] The script extracted metadata from the Excel file displayed in Appendix E, matched each entry to the corresponding text document, and inserted a metadata header in XML format at the beginning of the file, along with a closing tag at the end.

6.5.2 Macro-structure

In the context of the WoLCP, the term *macro-structure* refers to the division of each legal document into three main sections: the front matter (including elements such as the country name, title of the legislation, and preamble), the body (containing the normative provisions), and the back matter (which typically includes signatures, enactment dates, schedules, and related formal elements). This tripartite structure was introduced to facilitate targeted searches within specific sections of the texts and to restrict subsequent semi-automatic micro-structural tagging to the body

[51] The Python script used for the automatic insertion of metadata matches each *.txt* document with its corresponding metadata values extracted from a table and wraps the document in an XML-style <doc> tag. This was used for the preparation of the final version of the corpus. Before applying this script, the documents had already undergone light annotation with macro-structural tags (see Section 6.5.2). The script was developed by fellow students Angela Forzatti and Joanna Giacobbe during their internship as part of the collaborative WoLCP project and was kindly shared with the team.

section, which contains the core legislative content. Each document was manually annotated using XML tags to mark the three macro-structural sections, as shown below:

tags	explanation
<section type="front"> ... </section>	refers to the front matter of the document (titles, preambles, introductory statements, or official headers preceding the main legal content)
<section type="body"> ... </section>	denotes the main body of the text, i.e., the core legal content of the document (articles, sections, chapters, different provisions)
<section type="back"> ... </section>	refers to the back matter of the document (official signatures, closing formulae)

Table 2: Macro-structural section tags.

This structural markup serves to preserve document organisation and support more precise, layered annotation in downstream processing stages. The same annotation strategy was applied to the sub-corpora of Kenya and Tanzania in order to ensure consistency with the methodology adopted in the WoLCP and to enable comparative structural analysis across datasets.

6.5.3 Micro-structure

As briefly explained in Section 6.5, in addition to the light and macro-structural annotations, a structural segmentation system was introduced to improve the usability and searchability of the legislative texts in the corpus. The main goal was to clearly mark and separate the key parts of each legal document, so that users could focus on specific sections, especially the articles, which in this case represent the basic units of legal content. Since the article level was the only structure consistently found across all documents, it was chosen as the primary level to annotate.[52] This allows users to search or filter texts at the article level and supports further annotation work.

The micro-structural layer applied to the macro-structural part of the body of each legislative text captures the hierarchical subdivisions typically found in legal documents, such as parts, chapters, sections, and paragraphs, which differ substantially across jurisdictions due to variations in legal drafting traditions and document formatting.[53] To systematically account for variation in

[52] To better understand how the legislative documents are structured and annotated, see Appendix H. There, a sample annotation of one document from *WoL-K1*, one of the two Kenyan sub-corpora compiled and analysed in Chapter 7, is displayed. It is important to remind that the annotation strategy is the same for WoLCP and the four national sub-corpora of Kenya and Tanzania.

[53] This refers to the structural heterogeneity observed across the legislative texts from all the considered countries (see Appendix F for the full list). Document organisation varies significantly in the number of hierarchical levels and in the terminology used to label them. For instance, the document selected for Eritrea during the beginning phase is

structural organisation across texts and standardise them to the extent possible, a comparative mapping of internal document structures was developed. This mapping was compiled in a dedicated spreadsheet, where each row represented a hierarchical level (denoted as level 0, 1, 2, etc.) and each column corresponded to a specific document. For each structural level, the label used in the source text, its visual formatting, and the associated regular expressions for automated identification were recorded. The article level, designated as level 0, emerged as the only consistently recurring unit across all documents and was therefore selected as the foundational element for structural segmentation. A Python script was implemented to insert opening and closing XML tags around each article based on regular expressions, thereby allowing for level 0 segmentation across the documents.

However, annotating articles alone proved insufficient in several cases, especially when article titles were missing or did not carry meaningful thematic information. To enhance the linguistic utility of the corpus, a second layer of annotation, referred to as level 1, was introduced. This level captures higher-order textual divisions, such as parts or chapters, and includes their corresponding titles where available. While one Python script supports the insertion of both level 0 and level 1 structural tags,[54] a second dedicated script was implemented to further enrich level 1 segments by annotating their thematic headings.[55] This approach facilitates content navigation

organised into chapters, sections, and subsections preceding the individual articles. In contrast, the Egyptian text is structured into books, parts, chapters, and articles.

[54] This Python script was designed to semi-automate the structural annotation of legal texts included in the first release of the *Words of Labour Corpus Project* (WoLCP). Specifically, it inserts opening and closing structural tags to mark hierarchical divisions within the body section of the documents, such as articles (level 0) and larger groupings (level 1). It operates on corpus files that have already undergone light annotation, the insertion of metadata headers and macrostructural tags identifying the front, body, and back sections of each document. Its purpose is to facilitate consistent structural markup by automatically detecting recurring patterns using regular expressions. Once matched, each occurrence is preceded by an opening tag of the appropriate level. The script additionally inserts a corresponding closing tag before each new opening tag, based on the assumption that one structural unit ends where another begins. Despite its usefulness, the script is subject to some known limitations. When it comes to initial and final closing tags, the algorithm introduces an extra structural tag immediately after *<section type="body">* and omits the final closing tags for the last structural elements. Moreover, when level 1 and level 0 headings are adjacent, the inserted closing tag for level 0 may incorrectly encompass the level 1 header. Therefore, manual post-processing is required to correct the automatic annotation. The script was developed by fellow students Angela Forzatti and Joanna Giacobbe during their internship as part of the collaborative WoLCP project and was kindly shared with the team.

[55] The Python script was developed to enhance the structural markup of corpus files by enriching level 1 structure tags with a title attribute. This process applies to corpus documents that have already undergone previous stages of preparation, including macro-structural segmentation, metadata wrapping, and initial structural annotation (e.g., level 0 and level 1 tags). Files must be named according to the format *##_Country_Year_A.txt* and must contain level 1 headings already identified and tagged. The script identifies each level 1 structure tag and captures the heading that follows it. It then extracts the descriptive portion of the heading and inserts it as a value of the title attribute within the level 1 *<structure>* tag. This automated procedure facilitates future structural analysis by making headings searchable. The

without imposing a uniform classification scheme, which would be incompatible with the diversity of legal drafting conventions across jurisdictions. An example of how this tagging system was applied, using a legislative text from the Kenyan sub-corpus, is provided in Appendix H.

6.6 WoLCP-EN1: corpus overview, limitations and future work

The first version of the WoLCP-EN1 corpus compiled within the scope of this project consists of 30 legislative texts from 25 African countries, covering a range of document types including acts, proclamations, laws and codes. All texts are in machine-readable format and annotated according to the structural and metadata schemes outlined in this chapter. The corpus is segmented by article (level 0), and by thematic structural divisions (level 1), with additional macro-structural and omitted content tagging applied. While the present thesis focuses on a case study involving four sub-corpora, the broader WoLCP dataset serves as a reference corpus for comparative analysis and keyword extraction.

WoLCP-EN1 provides a flexible resource for the application of both leximetric and discourse-based approaches. It establishes a replicable methodological framework for corpus construction and annotation, including document selection, metadata assignment, and multi-level structural markup, while supporting comparative legal analysis on labour regulation and targeted exploration of discursive patterns across jurisdictions.

Nonetheless, the current version presents several limitations. First, it is restricted in both geographical and structural scope. Only a portion of the original WoL legislative database has been processed and annotated, and the corpus is currently limited to English-language texts. The annotation focused primarily on basic metadata and document-level segmentation, without further coding of legal provisions or semantic content. In addition, the thematic labels introduced through structural segmentation are based on internal headings rather than on the normative content identified in the WoL coding templates (see an example in Appendix A). As a result, the link between the annotated corpus and the corresponding variables and values in the leximetric dataset remains undeveloped. However, the existing selection protocol and supporting materials provide a foundation for future expansion in both vertical and horizontal directions, including the potential integration of non-African countries. Future work may also include the addition of French and Portuguese

script was developed by fellow students Angela Forzatti and Joanna Giacobbe during their internship as part of the collaborative WoLCP project and was kindly shared with the team.

texts, as well as their English translations, to reflect the multilingual nature of the original legal sources and enhance linguistic coverage.

Even though the current version of the WoLCP corpus does not allow for a comprehensive diachronic analysis of legal developments within individual jurisdictions considered, the Kenyan and Tanzanian sub-corpora presented in this thesis represent a first step in that direction, illustrating how the general framework can be adapted and expanded to support country-specific investigations. The procedures and tools described in this chapter form the methodological foundation upon which the four sub-corpora of Kenya and Tanzania were built. As outlined in the introduction to this thesis, the present work was conceived with the aim of conducting a vertically oriented investigation into the legal discourse of selected countries. It seeks to test and refine the methodologies introduced within the WoLCP framework, applying them in a focused manner to enable a context-sensitive case study analysis. While the WoLCP was designed as a broad, multilingual resource for comparative legal research, the present study examines how legal language shapes and reflects regulatory approaches to labour protection within two distinct national contexts.

The following chapter presents a detailed account of the corpus-building process specific to the Tanzanian and Kenyan datasets, illustrating how the annotation strategies introduced here were applied and adapted in view of a targeted discourse-based analysis of labour legislation.

7 Kenyan and Tanzanian Sub-Corpora: A Case Study on Female Labour Representation

7.1 Introduction

This chapter presents an exploratory case study on the discursive representation of female labour based on a set of four national sub-corpora compiled for the purpose of this analysis. Specifically, this set includes two general corpora, reflecting broader labour legislation in Kenya and Tanzania, and two thematically focused corpora, centred on gender-specific legal texts. Hence, the analytical framework is built on four distinct corpora, two per country. The rationale behind this design, as well as the criteria adopted for corpus construction, are outlined in Section 7.2.

While these collections are not directly derived from the current version of the corpus of the *Worlds of Labour Corpus Project* (WoLCP-EN1), which includes only one text from Kenya and four from Tanzania, they were informed by the methodological framework developed during the *Worlds of Labour* (WoL) collaborative project. They were built using legislative texts retrieved from the WoL repository managed by the University of Bremen. The primary objective of this case study was to examine how the annotation and analysis strategies introduced in Chapter 6 can support a more in-depth discursive investigation of legal language on a specific thematic area: female labour. More specifically, the case study explores how key terms, extracted in a corpus-driven manner, such as *woman*, *female*, *domestic servant*, *gender* and related notions are linguistically expressed, discursively framed, and patterned across Kenyan and Tanzanian legislative texts. The historical, geographical, and theoretical context that motivated the choice of this topic has been outlined in Chapter 4.

For ease of understanding, the sub-corpora will be referred to as WoL-K1, WoL-K2, WoL-T1 and WoL-T2. The text selection process was based on criteria such as frequency of reference within the WoL database, thematic relevance to women's work, and historical representativeness. While the WoLCP-EN1 is designed around a horizontal, comparative approach across more than two countries, WoL-K and WoL-T reflect a vertical approach, aimed at tracing the evolution of legal discourse over time within each national context.[56]

[56] For details on the meaning of vertical and horizontal approach within the context of the present thesis, please see Section 6.3.

As anticipated in the introduction to Chapter 6, the present chapter is closely interconnected with Chapter 5 and Chapter 6, which provide the methodological foundations upon which the present case study builds. As such, this chapter includes frequent references to Chapter 6, particularly in relation to the criteria for text selection and annotation practices. Additionally, this chapter introduces a dedicated part (Section 7.6) focusing on concepts of corpus-assisted discourse studies (CADS). It will address, among other aspects, the process of keyword extraction and the notions of absolute and relative frequency, which were fundamental for the case study's methodological approach.

The chapter is structured into two main parts. The first part (Sections 7.2, 7.5) presents the design and construction of the four national sub-corpora, provides an overview of their formal characteristics along with technical challenges encountered during corpus compilation, and describes the annotation procedures in line with the *WoLCP* standards. The second part first outlines the analytical framework adopted in the case study, then presents the results of the discursive analysis, which is also structured in two parts: first, an examination of the general sub-corpora (WoL-K1 and WoL-T1), and second, a focused investigation of the gender-specific sub-corpora (WoL-K2 and WoL-T2). Section 7.8 offers a comparative interpretation of the findings, and Section 7.9 concludes the chapter by reflecting on the methodological effectiveness of the study, its limitations, and potential directions for future research.

7.2 Corpus construction

This section describes the creation of the four national sub-corpora compiled for the present case study on the legal discourse surrounding female labour. The corpora were developed in two distinct but complementary phases. In the first phase, two general-purpose corpora were constructed for Kenya and Tanzania by selecting the most frequently referenced legislative texts in the *Worlds of Labour* database.[57] These aimed to capture the dominant discursive trends in each country's labour legislation. In the second phase, two thematic corpora per country focusing specifically on female labour were created to further explore gendered legal discourse. These additional corpora were designed to expand upon the findings from the first phase, allowing for a more targeted investigation into how legal texts address women's roles and rights in the workplace. The following section

[57] For a detailed explanation of the reasons behind the selection of Kenya and Tanzania for this case study, as well as the methodological rationale for beginning the analysis with Kenya before turning to Tanzania, please refer to Section 4.4.

outlines the criteria used to select the countries and texts included in these corpora and explains the methodological rationale behind this two-step design.

7.2.1 Text selection criteria

The corpora were constructed in two phases. First, two general corpora, WoL-K1 for Kenya and WoL-T1 for Tanzania, were compiled by identifying the most frequently cited legislative texts in the WoL repository. In the second phase, two additional sub-corpora, WoL-K2 and WoL-T2, were compiled to focus more explicitly on female labour. This decision was mainly prompted by the limited visibility of gender-related terms in the general corpora that could satisfy the purpose of the present case study.

The selection for the first two sub-corpora was based on frequency rankings generated via a Python script, as described in Section 6.4. In brief, legislative references were extracted from country-specific Excel coding templates available in the WoL project's internal database (see Appendix A for an excerpt from the WoL coding template for South Africa). These templates, originally compiled for leximetric analysis, included a *legal references* column that recorded the legal documents most frequently cited in connection with labour-related provisions. A Python script was then implemented to clean and standardise these references and to generate a ranked list of documents based on citation frequency. This ranked list served as the foundation for the initial selection of documents to be included in each national sub-corpus. Rather than aiming to analyse labour legislation in general, the selection process was intended to identify the most institutionally central labour laws in each country to then investigate the extent to which female labour was addressed within these texts.

The complete frequency lists extracted from the *legal references* column of the WoL coding templates of both Kenya and Tanzania are provided in Appendix F. Table F.1 and Table F.2 present the full range of legislative documents cited in the source templates, ranked by the number of occurrences within each national dataset. Each entry includes the full title of the legal act and the number of times it was referenced across the provisions coded by the WoL research team. To complement the comprehensive frequency lists presented in Appendix F, the tables below display only those legislative documents that were selected for inclusion in the general sub-corpora WoL-K1 and WoL-T1. The criteria guiding this selection process, as well as the reasons for the exclusion of certain texts, are outlined in the subsequent paragraphs.

After the following two sections on WoL-K1 and WoL-T1, the text will turn to the selection criteria for the thematic corpora WoL-K2 and WoL-T2 as well.

	WoL-K1	**WoL-T1**
1	The Employment Act, 2007 (No. 11 of 2007)	Employment and Labour Relations Act, 2004
2	Regulation of Wages and Conditions of Employment Act, 1982	Security of Employment Act, 1964
3	The Employment Act. Chapter 226 (No. 2 of 1976) [revised edition]	Employment Ordinance, 1955
4	Employment Ordinance, No. 2 of 1938	Severance Allowance Act, 1962
5	Shop Hours Ordinance, 1925	Regulation of Wages and Terms of Employment Ordinance, 1951
6	Master and Servants Ordinance, 1910	Master and Native Servants Ordinance, 1923
7	Native Porters and Labour Regulations, 1902	-

Table 3: Selected legislative texts for WoL-K1 and WoL-T1 corpora, listed in chronological order.

7.2.1.1 Text selection criteria for WoL-K1

For the general Kenyan sub-corpus, the selection was based primarily on two factors: citation frequency and historical coverage. Only documents that appeared at least four times in the WoL coding templates were considered. Consequently, all texts ranked below the *Regulation of Wages and Conditions of Employment Act, 1982* were excluded.[58] Despite its high frequency (27 citations), the *Master and Servants Ordinance (Ord. 8 of 1906)* was discarded as well, in favour of the *Master and Servants Ordinance, 1910*, which, although less frequently cited (15), was more complete and substantially longer. Given the minimal temporal gap between the two versions, and to avoid redundancy, only the 1910 edition was retained.

7.2.1.2 Text selection criteria for WoL-T1

The selection process for Tanzania was more complex. Several of the most frequently cited texts were originally written in German and therefore not eligible for inclusion in the English monolingual sub-corpus. As a result, the selection could not rely solely on citation frequency. Instead, a more flexible approach that balanced frequency, historical coverage, and thematic relevance was adopted. For instance, the *Regulation of Wages and Terms of Employment Ordinance, 1951* was included despite being cited only once, because it contributed to the corpus's thematic balance. Some documents were excluded after it was discovered that they had been cited in English but

[58] Even though the *Labour Regulations 1898* are cited four times within the WoL repository, they were ultimately discarded, as these regulations no longer exist in archival form. As the WoL coding templates report, they are referenced in the literature (Anderson, 2000) but the original legislative documents are lost.

were originally drafted in German. Examples include the *Ordinance Concerning the Legal Position of Native Workers, 1909* and *1913*, as well as the *Ordinance Concerning the Making of Labour Contracts with Coloured Persons, 1896*. These exclusions were necessary to maintain linguistic consistency across the corpus.

It is important to note that the selection for Tanzania was still grounded in the original WoL codings: only documents explicitly referenced in the templates were considered, even if their citation frequency was low. No additional texts were arbitrarily retrieved from the repository, since expert legal knowledge would have been required to independently assess their relevance. Even low-frequency documents cited within the WoL coding system were assumed to carry legal or thematic significance, as the coding methodology aimed to identify key features of labour protection or lack thereof (Carlino and Fechner, 2025).

7.2.1.3 Text selection criteria for WoL-K2 and WoL-T2

As explained in Section 7.2.1, two additional sub-corpora (WoL-K2 and WoL-T2) were subsequently compiled during the second phase of the case study to focus explicitly on legal texts concerning female labour, such as acts addressing gender equality, or women's employment in general. These thematic corpora allowed for a more focused discursive analysis of gender-specific legal frameworks and were constructed through a manual selection of texts from the WoL database based on their titles, content, and relevance to the topic. Table 4 below provides an overview of the legislative texts included in the thematic sub-corpora, WoL-K2 and WoL-T2.

	WoL-K2	**WoL-T2**
1	National Gender and Equality Commission Act, 2011	Employment of Women (Restriction) Decree, 1953
2	Employment of Women, Young Persons and Children Ordinance, 1961	Employment of Women and Young Persons Ordinance, 1947
3	Employment of Women, Young persons and Children Ordinance, 1948	Employment of Women and Young Persons Ordinance, 1940
4	Employment of Women, Young Persons and Children Ordinance, 1933	Employment of Women Ordinance, 1938
5	-	The Employment of Women, Children and Young Persons (Restriction) Decree, 1932

Table 4: Ultimately selected legislative texts for both WoL-K2 and WoL-T2 corpora listed in chronological order.

As mentioned, compared to the general texts included in WoL-K1 and WoL-T1, the texts selected for WoL-K2 and WoL-T2 differ both in their selection procedures and in their nature and thematic focus. Unlike the general sub-corpora, which were based on references found in the leximetric

coding templates, the thematic corpora were compiled through an independent manual search conducted directly within the general WoL legislative repository. This selection was based on a principle: only texts whose titles explicitly included terms such as *gender* or *women* were retained. Additional searches on the titles were also carried out using related keywords, such as *female* or *maternity*, to identify further relevant material. However, these additional terms did not return any results.

It is essential to emphasise that the selection of these topic-specific texts was necessarily limited to those available within the WoL repository. In other words, the compilation of WoL-K2 and WoL-T2 did not rely on an unrestricted or comprehensive survey of all legal texts regulating female labour in Kenya and Tanzania but was confined to the corpus of documents that have been formally taken into consideration according to the *Worlds of Labour Project*'s methodology. Since the overarching aim of this thesis is to explore the interplay between legal segmentation and discursive structures, the inclusion of external material would have undermined both the analytical coherence and the methodological replicability of the study. For this reason, all selected texts were drawn exclusively from the WoL database, in alignment with the WoLCP standardised corpus construction and annotation framework.

This methodological constraint is reflected in the structure of the two thematic sub-corpora. Immediately, it becomes evident that WoL-T2 displays very limited temporal variation, containing a narrow selection of legal texts dealing explicitly with female labour concentrated between the 1930s and the 1950s (see Table 4). This restricted chronological span does not result from any intentional exclusion during corpus construction. Rather, it may reflect a gap in the coverage of the WoL repository for Tanzanian legislation on this topic.

The selection criteria were consistent across both thematic sub-corpora. In both cases, all relevant legislative documents available in the respective country folders of the WoL repository were included, with only one type of exclusion applied: amendments, particularly in the case of Kenya. Amendments typically refer to pre-existing legislation and contain only partial modifications or additions. Including them would have resulted in content redundancy and would not have contributed substantial new material to the discursive analysis, given that their meaning is often dependent on the full context of the original act. Apart from this, no further exclusions were made; all texts explicitly addressing women's rights and female labour were retained for analysis.

The following section provides a detailed overview of the structure and size of the compiled corpora, presenting the main quantitative and formal characteristics of each sub-corpus.

7.3 Corpus structure and size

This section provides an overview of the structure and quantitative characteristics of the four sub-corpora compiled for this study. The corpora vary in size, time span, and document types, reflecting the thematic and methodological distinctions outlined in the previous sections. All texts are in English and were converted into plain *.txt* format for consistency and compatibility with corpus analysis tools. Table 5 below summarises the key features of each sub-corpus, including the number of documents, total token count, and time coverage. The general sub-corpora (WoL-K1 and WoL-T1) span a broader time range and include core labour laws, while the thematic corpora (WoL-K2 and WoL-T2) are narrower in scope and focus specifically on gender-related legal texts.

Sub-corpus	No. of documents	Time span	No. of tokens	Language	Text types
WoL-K1	7	1902-2007	68,494	English	Regulation, Ordinance, Act
WoL-T1	6	1923-2004	68,465	English	Ordinance, Act
WoL-K2	4	1933-2011	18,891	English	Ordinance, Act
WoL-T2	5	1932-1953	8,297	English	Ordinance, Act, Decree
Combined	22	1902-2011	164,147	English	Regulation, Ordinance, Act, Decree

Table 5: Key characteristics of WoL-K1, WoL-K2, WoL-T1, WoL-T2.

Across both the general and thematic corpora, Kenya and Tanzania are each represented by a total of 11 legislative documents. As shown in Table 5, however, the four sub-corpora compiled for this study differ in size and temporal coverage. While WoL-K1 and WoL-T1 contain a similar number of tokens, ensuring a balanced comparison between the general labour legislation of Kenya and Tanzania, a noticeable discrepancy emerges between WoL-K2 and WoL-T2. This difference is primarily due to the limited availability of gender-focused legal texts for Tanzania in the WoL repository. While every relevant document was included, the Tanzanian dataset is more restricted in time and quantity. However, the decision to include Tanzania was motivated less by the volume of available material and more by a desire for methodological consistency and geographical comparability. Although Tanzania's database proved more limited, particularly with regard to gender-specific legislation, it nonetheless offers a valuable point of comparison for tracing legal discourse in a neighbouring jurisdiction. This imbalance is acknowledged and taken into account in the comparative analysis presented in Sections 7.7.1 and 7.7.2.

Before turning to the annotation procedures and the analysis, the next section briefly outlines some of the practical challenges encountered during the corpus compilation phase. This is then followed by an account of the annotation strategies applied to the four sub-corpora.

7.4 Practical challenges

The compilation of the four sub-corpora involved some technical challenges, particularly in relation to the formatting of legal documents. In several cases, copy-pasting the content directly from the source document led to misaligned structures: for example, section headings that were visually presented in the margins or at the top of pages were displaced or truncated when extracted as plain text. This resulted in considerable manual intervention to recover and reassign structural elements accurately, particularly to preserve document hierarchy during annotation and to enable a proper analysis in Sketch Engine, which would otherwise return errors. Preliminary keyword list compilations, for instance, revealed truncated forms, such as *empl-* instead of *employment* or *employ*, resulting from hyphenation in the original PDFs. Therefore, fragmented word forms had to be manually reconstructed by reattaching hyphenated segments to their corresponding lexical items.

Another frequent challenge involved the use of Optical Character Recognition (OCR) for documents that could not be easily copied and pasted due to their format, such as scanned images or non-editable PDFs. While OCR enabled the conversion of these files into machine-readable text, it also introduced some spelling and character recognition errors. Although particular attention was given to correcting major issues, especially those affecting legal terminology and general keywords, it is not possible to guarantee complete accuracy across all documents, and occasional misspellings may still be present within the corpora.

7.5 Annotation strategy

All legislative texts included in the four sub-corpora were annotated using the structural and metadata scheme developed for the WoLCP English corpus (WoLCP-EN1) in order to ensure methodological consistency and facilitate future integration into the larger collaborative project. The annotation followed a multi-layered approach, consisting of (i) light annotation, (ii) metadata insertion, and (iii) micro and macro-structural segmentation.[59]

[59] All details regarding the annotation procedures and the rationale behind the choices made are presented in Section 6.5.

Light annotation was first applied to replace structurally peripheral or non-linguistic content, such as signatures, tables of contents, institutional seals, or illegible elements that could distort keyword frequency or interfere with corpus-based analysis (see Table 1, Section 6.5). These excluded components were replaced with XML placeholder tags (e.g., *<omitted type="table"/>*) to preserve the logical structure of the documents.

Next, each document was segmented into three macro-structural sections: front matter, body, and back matter (see Section 6.5.2). This allowed for the isolation of the core legal content, the body section, which then became the focus of further structural annotation. Within the body, a micro-structural tagging layer was introduced to mark hierarchical units, based primarily on the article level (level 0) and higher-order divisions such as chapters, parts, or sections (see Section 6.5.3). To ensure consistency despite naming variation across jurisdictions, these headings were uniformly referred to as level 1 elements. This dual-level segmentation was implemented through Python scripts[60] using regular expressions to automate the placement of XML tags. To enhance semantic traceability, level 1 headings were also annotated with their corresponding titles, allowing for thematic navigation based on the internal logic of the texts themselves, rather than on externally imposed classification schemes.

In parallel, each document was tagged with a standardised metadata header containing key contextual information such as country, year, document title, and language (see Section 6.5.1). These metadata fields were first compiled in a reference spreadsheet and then inserted automatically into each file via a custom script (see Section 6.5.1). All metadata information used for the annotation of WoL-K1, WoL-T1, WoL-K2, and WoL-T2 is provided in full in Appendix G.

This annotation protocol was systematically applied to all the texts to ensure coherence and comparability across both the general WoL and Kenyan and Tanzanian corpora. The segmentation into articles and thematic sections facilitates targeted corpus queries, while the metadata layers support filtered comparisons across countries, time periods, and document types. An example of a fully annotated document segment is provided in Appendix H to illustrate the structure and tagging conventions.

[60] As outlined for all the previous Python scripts, these were developed by fellow students Angela Forzatti and Joanna Giacobbe during their internship as part of the collaborative WoLCP project and were kindly shared with the team.

With four complete annotated corpora, the next step involved defining the analytical strategy through which discursive patterns could be extracted and interpreted. The following section outlines the tools and methods adopted to carry out this investigation.

7.6 Methodological reflections and analytical approach

This section first provides an overview of the methodological principles that form the conceptual and practical foundation for the case study. While the approach draws on the broader framework of Corpus-Assisted Discourse Studies (CADS), the focus here is on the concept of keyness, frequency measures, and the use of a reference corpus, all central to the present analysis. After, the analytical workflow will be described in Section 7.6.2.

7.6.1 Theoretical framework: Discourse analysis and corpus-based keyword studies

This study draws on a corpus-assisted discourse studies (CADS) approach to examine how women are framed in labour law discourse in Kenya and Tanzania. CADS represents a hybrid methodology, defined by Partington *et al*. (2013: 10) as "that set of studies into the form and/or function of language as *communicative discourse* which incorporate the use of computerised corpora in their analyses." Unlike traditional discourse analysis, CADS combines qualitative methods with quantitative tools, enabling the detection of patterns that may not be visible to the naked eye. It provides a methodology for uncovering "non-obvious meaning[s]" (*ibid.*: 11) grounded in recurring lexical and grammatical behaviours across texts.

At its core, discourse analysis involves the study of "language in use" (Brown et al., 1983: 1). Definitions vary, but a common functional view is that discourse is "language that is doing some job in some context" (Halliday, 1985: 10), and that discourse analysis entails analysing how language constructs social activities, identities, and relationships (Gee, 1999). Stubbs (2007) further contends that discourse analysts infer these functions from textual traces, making them, in essence, "text-to-discourse" analysts (Partington et al., 2013: 3). Thus, language is not studied in isolation but as embedded in socio-political contexts, revealing ideological positions and power dynamics.

CADS builds upon corpus linguistics as a methodological paradigm. Corpus linguistics is defined as "a collection of tools and techniques for linguistic analysis" that are evidence-driven, relying on naturally occurring texts rather than introspective examples (Partington et al., 2013: 7).

Its focus on recurrence, frequency, and comparison enables researchers to identify both grammatical norms and thematic structures in discourse types. Moreover, it permits the investigation of discourse change over time, supporting both synchronic and diachronic analyses (Partington et al., 2013).

In CADS, comparison plays a crucial role. One corpus is typically compared with another, a reference corpus, to highlight distinctive features of the discourse under investigation. Baker (2006: 76) explains that reference corpora act as benchmarks of *normal* language use, against which a specialised corpus can be contrasted. These comparisons allow analysts to observe which lexical or semantic items are statistically overrepresented, offering insights into salient themes or ideologies. For example, when analysing a corpus of East African women labour legislation, comparing it to a general legal reference corpus might reveal that terms such as *maternity*, *household*, or *vulnerable* appear significantly more frequently. Such overuse could suggest a discursive pattern primarily focused on domesticity or protection, a theme that might not be as prominent in the reference corpus.

Keyness analysis is a central concept in corpus-assisted discourse analysis. According to Baker (2006), keyness involves comparing word frequency lists from different corpora to identify items that occur significantly more frequently in one corpus relative to another. This statistical measure moves beyond simple frequency to reveal lexical saliency. Baker (2006: 165) writes: "A keyword list therefore gives a measure of saliency, whereas a simple word list only provides frequency". This is particularly helpful to analysts that want to uncover which words, and by extension, which discourses, dominate a given corpus. Tools such as *AntConc* compute log-likelihood scores, "number[s] which indicat[e] the extent that a word is distinctive in one text" (Baker, 2006: 167), or within a corpus, compared to another. These keywords then form the basis for subsequent qualitative analysis through concordance and collocation techniques.

In this context, it is important to distinguish between absolute frequency and relative frequency. Absolute frequency refers to the raw count of how often an item appears in a corpus,

whereas relative frequency standardises this count, for instance per million tokens, allowing comparison across corpora of differing sizes (Sketch Engine, n.d.).[61] This is essential in ensuring comparability, particularly when one corpus is significantly smaller than the reference corpus.

Furthermore, Baker (2006: 183) emphasises that keyness can and should be explored beyond the lexical level. Some discourses manifest not in the repeated use of a single term, but through a cluster of semantically or grammatically related items. For instance, even if the word *strong* is not a keyword on its own, the presence of synonyms like *tough*, *robust*, and *resilient* might point to a discourse of strength. These semantic groupings, or key categories, can be identified manually, or else using automatic semantic annotation tools (e.g. USAS in Wmatrix) and are particularly useful in detecting evaluative prosodies that extend across a text (Baker, 2006: 184). The analysis of modal verbs, personal pronouns, or adjectives, when grouped into functional or thematic categories, can further reveal identity construction, stance, or persuasive strategies.

The methodology applied in this study is inductive and iterative. Following the steps laid out by Baker (2006: 104; 189-190), the analysis began by building two specialised corpora composed of Kenyan and Tanzanian labour law texts. These were compared against a larger, general reference corpus to extract keyword lists. Keyness thresholds were set using statistical tests, such as log-likelihood, and results were refined by inspecting concordance lines and collocates to identify recurring patterns. Where appropriate, key semantic and grammatical categories were also analysed to capture higher-level discourse strategies, particularly those contributing to the representation of gender roles and ideologies.

In conclusion, CADS provides a robust framework for investigating the discursive construction of women in labour law texts. By integrating quantitative techniques like keyness analysis with qualitative interpretation of concordance data, it becomes possible to trace how lexical and grammatical choices reflect broader socio-legal narratives. As Partington et al. (2013: 11) suggest, corpus techniques do not replace traditional discourse analysis but enhance it, particularly by uncovering patterns that are not immediately visible.

[61] Sketch Engine, "Frequency". https://www.sketchengine.eu/glossary/frequency/ [last accessed: 25 June 2025]. Sketch Engine, "Relative frequency, frequency per million". https://www.sketchengine.eu/glossary/freqmill/ [last accessed: 25 June 2025].

The main concepts outlined in this section serve as the foundation for the in-depth qualitative and quantitative analyses presented in Section 7.7, while the following part will present the analytical workflow developed for this case study.

7.6.2 Analytical workflow

This section outlines the methodological strategy developed to analyse how female labour is discursively represented in legal texts from Kenya and Tanzania between 1902 and 2011. It illustrates the corpus tools used for processing the annotated data and explains the rationale behind the followed techniques. The findings will be presented in Section 7.7.

To conduct the investigation, Sketch Engine (reference) was chosen as the corpus analysis platform for several practical reasons. Firstly, its user-friendly interface and accessibility via institutional student accounts made it a suitable choice for facilitating independent exploration by the Bremen team within the collaborative framework of the WoL project. Furthermore, Sketch Engine offers a wide range of analytical tools suitable for both quantitative and qualitative investigation, such as keyword extraction, collocate analysis, frequency lists, and concordance generation. Beyond these standard functionalities, it provides more advanced features that proved particularly useful for this study. Among them, the Word Sketch[62] and Thesaurus[63] functions were especially relevant. The Word Sketch tool made it possible to explore how specific terms function syntactically across the corpora, revealing whether they were typically used as subjects, objects, or in other grammatical roles. These patterns were further examined through direct access to the corresponding concordance lines, allowing for immediate qualitative inspection of the discursive context. Equally valuable was the ability to classify and filter texts based on multiple metadata criteria, such as year of publication or document title.

The first phase of the analysis involved generating keyword lists for both WoL-K1 and WoL-T1 using Sketch Engine's keyword extraction function. The aim was to identify lexical items

[62] The Word Sketch function in Sketch Engine provides an automatically generated summary of a word's typical grammatical and collocational behavior in a corpus. It organises common word combinations into grammatical relations (e.g. subject, object, modifier) and ranks them by their strength using scores like logDice, enabling rapid identification of how a term is used in different contexts. For further details on this, please see Sketch Engine, "Word Sketch — collocations and word combinations" https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words/#toggle-id-2 [last accessed: 25 June 2025].

[63] The Thesaurus function in Sketch Engine automatically identifies words that occur in similar contexts across a corpus, generating a list of semantically related terms based on distributional similarity. Unlike traditional thesauri, it is corpus-specific and reflects actual usage patterns rather than pre-defined lexical relations. For further details on this, please see Sketch Engine, "Thesaurus – synonyms, antonyms and similar Words" https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words/ [last accessed: 25 June 2025].

that occur with statistically significant higher frequency in the two national datasets compared to a broader reference corpus: WoLCP-EN1, the English corpus of the *Worlds of Labour Corpus Project*. Two documents were excluded from the reference corpus to avoid skewing results in case of overlaps: *06*_WoL-T1_*2004* and *07*_WoL-K1_*2007*, named respectively *51_Tanzania_2004* and *26_Kenya_2007* in the WoLCP-EN1 (see Appendix E and Appendix G).[64]

The resulting first two keyword lists featured very few women-specific terms, which may indicate a limited lexical visibility of female labour in the general legislative discourse. However, this could also be due to the relatively balanced frequency of such terms in both the focus and reference corpora, making them statistically non-salient despite being present. Despite this, the analysis still proceeded with a brief investigation of some keywords' concordances and collocates in WoL-K1 and WoL-T1, as this was deemed valuable for understanding broader discursive framings of colonial and postcolonial labour, and for uncovering potentially implicit gendered patterns not immediately evident through keywords alone.

Following the same procedure adopted in the earlier stage of the analysis, the first step in the investigation of the thematic sub-corpora WoL-K2 and WoL-T2 involved the extraction of keywords using WoLCP-EN1 as the reference corpus. Although gendered terms appeared in both lists, the range and variety of such terms was still limited.

Across both corpora, the relatively low salience of explicitly gendered language suggested two possible interpretations. The sub-corpora focusing on women's labour may contain a limited number of such terms, raising questions about the nature and extent of their representation and the methods by which female labour can be meaningfully analysed. Else, the reference corpus WoLCP-EN1 includes a sufficient volume of gendered terms overall, thereby diminishing their prominence within WoL-K2 and WoL-T2.

Therefore, before expanding the keyword analysis to include lower-ranking terms from the keyword list extracted with WoLCP-EN1 as the reference corpus and broaden the investigation, a more methodological cautious step was first opted for. To verify whether the low salience of gendered terms is a by-product of the reference corpus used, an additional keyword extraction using

[64] These two texts were included in both *WoL-T1* and *WoL-K1* because they were among the most frequently cited documents in the WoL leximetric database. For this reason, they are part of *WoL-T1* and *WoL-K1* as well as *WoLCP-EN1*, since the selection strategy for all three corpora was the same (i.e. based on frequency) (see Section 7.2.1). Their removal ensured that the comparison reflected true contrastive differences between the national sub-corpora and the broader dataset.

a different reference corpus was conducted. This allowed me to assess whether a change in comparative baseline would yield different results or bring more explicitly gendered terms into sharper relief. This approach not only aligns with the need for methodological rigour but also helps clarify whether the lexical patterns observed so far are corpus-specific or indicative of a broader discursive trend.

For the purposes of this comparative test, the *British Law Report Corpus* (BLaRC)[65] was selected as the second reference corpus. This choice was informed by several criteria. Firstly, BLaRC is composed of British English texts, which aligns closely with the linguistic context of the Kenyan and Tanzanian documents examined in this study. Given that these legal texts were originally drafted in British English during and after the colonial period, a corpus reflecting the same national variety of English was considered more appropriate than, for instance, a corpus of American English. Secondly, BLaRC is readily accessible through Sketch Engine. This availability, combined with its unrestricted institutional access and lack of copyright limitations, made it a practical and reliable tool for comparison. Thirdly, and most importantly, BLaRC is a domain-specific corpus of legal English. While it does not specialise in labour law, its legal focus ensures that any emerging keyword patterns are more likely to reflect discursive characteristics related to gender or social roles rather than merely highlighting genre-specific features of legal discourse, which would be the case if a general-purpose corpus such as the British National Corpus[66] or enTenTen[67] were used instead.

[65] The British Law Report Corpus (BLaRC) is an 8.5-million-word corpus of judicial decisions issued by British courts and tribunals between 2008 and 2010. Compiled by Dr. María José Marín Pérez and hosted by the University of Murcia, it serves as a resource for the analysis of legal English in a British context. Sketch Engine, "BLaRC: British Law Reference Corpus": https://www.sketchengine.eu/blarc-british-law-reference-corpus/#:~:text=The%20British%20Law%20Report%20Corpus,by%20British%20courts%20and%20tribunals [last accessed: 25 June 2025].

[66] The British National Corpus (BNC) is a 100-million-word collection of written and spoken British English from the late 20th century. Approximately 90% of the corpus consists of written texts (e.g. newspapers, academic publications, essays), while the remaining 10% comprises spoken language (e.g. informal conversations, radio broadcasts), some of which is available in audio format. The corpus is searchable by various criteria, including publication date, region, media type, and genre, using the David Lee classification system. Sketch Engine, "British National Corpus", https://www.sketchengine.eu/british-national-corpus-bnc/ [last accessed: 25 June 2025].

[67] The English Web Corpus (enTenTen) is part of the TenTen family of web-based corpora, built using automated tools that target linguistically valuable online content. The most recent version (enTenTen21) contains approximately 52 billion words, collected between October 2021 and January 2022. Low-quality or spam content was filtered out through semi-manual checks of the most prominent web domains. Sketch Engine, "enTenTen – English corpus from the web" https://www.sketchengine.eu/ententen-english-corpus/ [last accessed: 25 June 2025].

That said, the BLaRC is not without limitations. Ideally, a reference corpus for this analysis would span a broader temporal range, including material from the colonial and immediate post-colonial periods, to better reflect the discursive norms of the time. Moreover, the BLaRC consists of judicial decisions rather than legislative texts, which may differ in linguistic formality, structure, and functional purpose from the documents that make up the WoL corpora.

It is also important to acknowledge that BLaRC was processed using a different tokenisation and part-of-speech tagging system compared to the WoL sub-corpora (Marcus et al., 1993). Unlike the user-uploaded corpora, which are automatically POS-tagged and lemmatised by Sketch Engine upon upload using its default internal models,[68] the BLaRC was tagged using the Penn Treebank tagset (Marcus et al., 1993), a different annotation scheme. As indicated by Sketch Engine itself, the use of corpora processed with different tokenisation or POS-tagging conventions may result in unreliable keyword identification. This is explicitly flagged by the platform through a warning message: "KW: This reference corpus was processed with different tokenization or POS tags. Text tokenized or tagged differently may be incorrectly identified as keywords",[69] alongside the notification that no term extraction (T) will be performed. Hence, such annotation mismatches may compromise the statistical comparability of frequency and salience measures. In this specific case, however, the methodological risk was mitigated by the nature of the research design. The target corpora are relatively small and were manually compiled, allowing for a high degree of familiarity with the content and structure of the source texts. This enables a more informed and critical engagement with the data, particularly when evaluating the accuracy of keyword extraction. Additionally, a strategy that may reduce the potential effects of tagging inconsistencies would be to inspect the absolute frequencies of key terms. This cross-verification allows for the detection of misleading keyword salience during the qualitative stages of the analysis. Despite some constraints, therefore, BLaRC was deemed suitable for assessing whether gendered language is more or less

[68] Upon upload to Sketch Engine, user corpora are automatically lemmatised and annotated for part-of-speech (POS) using internal taggers specific to each language. A POS tag is a label assigned to each word to indicate its part of speech (e.g. noun, verb, adjective), and sometimes additional grammatical features such as number, tense, or case. The complete set of POS tags used in a corpus is known as a *tagset*. Sketch Engine uses different tagsets depending on the language and corpus; in the case of user-uploaded English texts, a default internal English model is applied. Automatic POS annotation enables corpus tools to distinguish grammatical functions and support linguistic pattern detection (e.g. finding plural nouns following modal verbs). The information presented in this note is based on the Sketch Engine POS tagging guide "POS tags" https://www.sketchengine.eu/blog/pos-tags/ [last accessed: 25 June 2025].

[69] Sketch Engine: https://www.sketchengine.eu/ [last accessed: 25 June 2025].

foregrounded when WoL-K2 and WoL-T2 are compared against a broader legal English background.

In this context, the analytical perspective could have been also altered by modifying keyword extraction settings (e.g., lowering the minimum frequency threshold from 5 to 10), a strategy that might have highlighted other relevant lexical items. However, this alternative was not pursued in the present study to maintain comparability across the different sub-corpora and avoid introducing additional variation into an already complex process.

The results from the BLaRC comparison confirmed most keywords across both sub-corpora (see Section 7.7.2.1). However, a few exceptions emerged, highlighting additional terms of potential discursive relevance. Consequently, the analytical strategy was adapted to incorporate not only statistically prominent items, but also terms with high absolute frequency and those displaying thematic or contextual relevance, even if they did not rank highly in the keyword list. While the overall impact on the keyword lists was limited, the use of BLaRC nonetheless proved methodologically valuable, offering a means to confirm findings and reduce the risk of interpretative bias linked to a single reference source.

Once the analytical strategy had been finalised, a set of target terms was selected for detailed investigation through collocation analysis, concordance inspection, Word Sketch, and Thesaurus functions.

7.7 Discursive representation of female labour in WoL-K and WoL-T

As outlined in Section 7.6.2, the analysis is structured in two main parts: the first focuses on WoL-K1 and WoL-T1, while the second is dedicated to WoL-K2 and WoL-T2. This division reflects not only the distinct nature of the documents included in each sub-corpus, but also the order in which the analysis was conducted, allowing for a clearer progression from broader legal discourse to more explicitly gendered legal texts. Since the entire analytical procedure, including the rationale for all methodological choices, has been detailed in Section 7.6.2, the next part will proceed directly to illustrating the initial findings from phase one, focusing on the two general corpora. The discussion will then transition to phase two in Section 7.7.2.

7.7.1 Observations in WoL-K1 and WoL-T1

7.7.1.1 Keyword analysis

The keyword extraction was conducted using the advanced settings interface in Sketch Engine. Minimum frequency was set at 5, and keywords were filtered to exclude country names (*kenya* for WoL-K1 and *tanzania* for WoL-T1), as well as *ordinance*. The attribute selected for keyword matching was *lemma*, to account for inflectional variation across different word forms. The tables below show the first 30 identified keywords for WoL-K1 and WoL-T1.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	porter	79	**11**	sub-section	24	**21**	arrest	17
2	caravan	56	**12**	comoro	24	**22**	malagasy	17
3	native	56	**13**	islander	24	**23**	provincial	16
4	protectorate	48	**14**	arab	23	**24**	pound	15
5	recruiter	43	**15**	occupier	19	**25**	reside	15
6	africa	38	**16**	like	18	**26**	thereunder	15
7	governor	37	**17**	closing	18	**27**	licence	15
8	servant	369	**18**	baluchi	18	**28**	corr	13
9	colony	27	**19**	blanket	18	**29**	intoxicate	13
10	somali	25	**20**	aforesaid	17	**30**	punishable	12

Table 6: Most frequent keywords identified in WoL-K1 with *WoLCP-EN1* as reference corpus.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	governor	48	**11**	protectorate	18	**21**	exact	11
2	lockout	43	**12**	thereunder	17	**22**	apparent	11
3	native	41	**13**	african	16	**23**	mediator	11
4	recruiter	26	**14**	aforesaid	16	**24**	ship	10
5	ticket	25	**15**	ration	16	**25**	voyage	9
6	licence	25	**16**	amalgamate	15	**26**	secondary	9
7	sub-section	23	**17**	united	15	**27**	like	9
8	tanganyika	22	**18**	servant	149	**28**	territory	88
9	protest	21	**19**	organisational	12	**29**	blanket	8
10	services	19	**20**	east	12	**30**	stop	8

Table 7: Most frequent keywords identified in WoL-T1 with *WoLCP-EN1* as reference corpus.

One of the first things likely to be noticed is the interesting high degree of lexical overlap between the two. They share a significant number of common terms: *native*, *recruiter*, *governor*, *africa*/*african*, *protectorate*, *servant*, *blanket*, *aforesaid* and *thereunder*. Even if it is only an initial observation, it suggests a shared foundational vocabulary in the discursive representation of labour. However, closer inspection reveals instances where certain terms exhibit a disproportionate frequency in one corpus over the other. For example, words like *porter* and *caravan* appear with high

prominence in WoL-K1 but are absent from WoL-T1. This discrepancy is due to the inclusion of texts such as the 1902 *Native Porters and Labour* document. In the context of relatively small sub-corpora, the presence of such specialised texts can easily skew keyword frequency results, potentially exaggerating the thematic salience of specific terms unique to those documents. Despite this probable skewing effect, both terms will still be briefly analysed later to explore what they might reveal about the specific context and labour practices they represent.

Moving to a more detailed examination, the keyword list for WoL-K1 reveals a strong lexical concentration around colonial administrative structures and labour categories specific to the Kenyan context. As mentioned, high-frequency keywords include *porter*, *caravan*, *recruiter*, which might suggest a regulatory focus on mobile, physically intensive labour, often tied to colonial infrastructure and logistics. Furthermore, terms such as *native*, *protectorate*, *colony*, *governor*, and a series of ethnic identifiers like *somali*, *arab*, *comoro*, *baluchi*, and *malagasy* reflect colonial circumstances, where racial and ethnic classification may have played a central role in the legal framing of labour. The high frequency of the term *servant* further emphasises a discourse of subordination and hierarchical employment structures. Terms such as *sub-section* and *aforesaid* reflect formal and administrative language, while *licence*, *arrest*, and *punishable* suggest mechanisms of legal enforcement and control. Overall, WoL-K1 appears to foreground a legal regulatory discourse oriented around race, geographic origin, and physical labour. Although direct references to women or gender do not appear among the top-ranked keywords, the lexical profile highlights structures within which gendered labour roles may have been implicitly shaped.

Even a preliminary review of the entire keyword list for WoL-K1 reveals that strictly gendered terms such as *woman*, *female*, *maternity*, and *her* do not feature among the top 100 keywords, nor do they exhibit relatively high absolute frequencies (see Table 8). This suggests that references to women or female-specific issues are not discursively salient in the corpus, even though they are occasionally present. At the same time, masculine pronouns such as *his*, *he*, and *him* are quite numerous. Without further keyness testing against the reference corpus, though, their frequency alone cannot be taken as an indication of discursive prominence. Their presence may reflect not necessarily a thematic emphasis on men, but rather the common use of the generic masculine.

The table below does not show all keywords in relation to gender: not all terms that could be relevant to gendered labour were considered as such. Lexical items like *domestic* or *cook*, although not strictly gendered, could be investigated more closely to ascertain whether they may

reflect historically feminised labour roles (see Chapter 4) and should be investigated closer. The list includes words that refer directly to gendered subjects (e.g., *he*, *woman*, *male*), reproductive roles (e.g., *maternity*, *child*), and job titles marked by gender (e.g., *foreman*, *herdsman*). Terms like *child* were included due to their discursive proximity to maternity and the regulation of family-related employment conditions.

Item	Keyword Rank Position	Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)	Score
his	376	392	254	5723,12891	2488,1958	2,3
he	384	245	166	3576,95557	1626,14368	2,199
him	340	128	74	1868,77686	724,90741	2,576
child	499	103	102	1503,78137	999,19672	1,504
himself	235	35	11	510,99365	107,75651	4,708
woman	121	29	2	423,39474	19,59209	20,61
maternity	877	17	42	248,19693	411,43396	0,604
female	826	16	33	233,59711	323,26953	0,723
sexual	578	13	15	189,79765	146,94069	1,29
her	1016	13	163	189,79765	1596,75549	0,119
male	375	9	6	131,39838	58,77628	2,215
man	144	8	1	116,79855	9,79605	10,911
headman	66	6	0	87,59892	0	88,599
herdsman	75	6	0	87,59892	0	88,599
sex	698	6	9	87,59892	88,16441	0,994
foreman	186	5	1	72,99909	9,79605	6,854

Table 8: Distribution and keyword scores of gender-related terms in WoL-K1 compared to the reference corpus.

These findings suggest that further analysis is necessary, ideally through thematically focused sub-corpora which specifically address gendered labour and female representation. Furthermore, a qualitative examination of some of the most prominent keywords identified in the Kenyan corpus may provide valuable insights into discursive patterns and the implicit ways gender and related issues are constructed within the texts. Such an approach would help move beyond mere frequency counts, offering a deeper understanding of the underlying representations and discursive strategies at play. This phase of the analysis will be undertaken in the subsequent section, following the discussion of the main keywords extracted from WoL-T1.

The keyword list for WoL-T1 reveals a partially overlapping but distinct lexical landscape when compared to WoL-K1. Notable keywords include *lockout*, *protest*, *services*, *ticket*, *licence*, which may reflect a focus on regulations of collective labour relations, potentially indicative of legal responses to organised labour activity, strikes, or disputes within more urbanised employ-

ment sectors. Furthermore, terms like *governor*, *protectorate*, *tanganyika*, *African,* all found almost exclusively in texts from the colonial period,[70] continue the colonial-administrative thread also observed in WoL-K1, while words such as s*ub-section*, *thereunder*, *aforesaid* reflect the bureaucratic and formal nature of the texts.

Of particular interest is the lemma *servant*, which dominates the list with 149 occurrences (see Table 7). This exceptionally high frequency may highlight a legal discourse focused on subordinate labour roles, probably linked to domestic service, institutional hierarchies, or low-skilled employment sectors. While these initial hypotheses suggest that legal categories of servitude played a significant role in shaping how labour relations were described in Tanzanian legislation, they will be further examined in the subsequent qualitative analysis of the term *servant* in Section 7.7.1.2.1.

Similar to the WoL-K1, the WoL-T1 keyword list does not prominently feature female-related terms among its most statistically salient items. As observed for Kenya, gender-specific terms such as *female*, *maternity*, and *her* do not exhibit notably high absolute frequencies (see Table 9) and are generally absent from the top ranks of the keyword list. Nevertheless, it is worth noting that terms such as *woman* and *wife* do appear among the first 100 keywords (positions 88 and 84, respectively). Moreover, some gender-indexical nouns like *mother*, *widow*, *wife* are attested in WoL-T1 but entirely absent from the WoL-K1 keyword list. This contrast may suggest a slightly broader discursive scope in Tanzanian texts, potentially allowing for references to women in relation to familial or marital status, roles traditionally associated with legal dependency or protection. However, it would be reductive to infer discursive prominence solely from the presence of such terms. Interestingly, the reverse is also true: keywords such as *sexual* and *sex*, which are present in WoL-K1, do not appear in the Tanzanian list. While this absence might hint at differing legal focal points or silencing mechanisms in each national context, such isolated lexical observations risk introducing interpretative bias if not measured in broader patterns.

Therefore, while gender-specific terms are intuitively attractive analytical targets, focusing exclusively on them risks skewing the analysis toward thematically expected but potentially non-salient elements. As discussed earlier, many discursively relevant gendered dynamics may instead

[70] The terms *African* and *tanganyika* are also found in a 1962 text (*04_WoL-T1_1962*), though only 4 and 3 times respectively.

be encoded in apparently neutral or functional vocabulary (e.g., *domestic*, *servant*, *cooking*, *parent*).

Item	Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)	Score
his	287	254	4191,92285	2488,1958	1,684
he	160	166	2336,96045	1626,14368	1,437
child	90	102	1314,54028	999,19672	1,315
him	60	74	876,36017	724,90741	1,209
himself	31	11	452,7861	107,75651	4,172
her	20	163	292,12006	1596,75549	0,183
woman	15	2	219,09004	19,59209	10,688
maternity	10	42	146,06003	411,43396	0,357
wife	9	1	131,45403	9,79605	12,269
male	8	6	116,84802	58,77628	1,971
chairman	7	4	102,24202	39,18419	2,569
she	7	79	102,24202	773,88763	0,133
widow	6	1	87,63602	9,79605	8,21
headman	5	0	73,03001	0	74,03
mother	5	5	73,03001	48,98023	1,481
female	5	33	73,03001	323,26953	0,228

Table 9: Distribution and keyword scores of gender-related terms in WoL-T1 compared to the *WoLCP-EN1* reference corpus.

In sum, the keyword analysis of WoL-K1 and WoL-T1 reveals overlapping colonial-administrative words and a shared focus on regulatory language concerning subordinate labour roles. Despite this, notable divergences also emerge, including corpus-specific lexical items and some patterns of female-related term distribution. While the presence of words such as *wife*, *mother*, and *widow* in WoL-T1 but not in WoL-K1 suggests subtle differences in the discursive framing of gender, the overall absence of prominently gendered terms in both lists indicates that female labour is not foregrounded at the level of statistically salient vocabulary.

However, as discussed, relying solely on surface-level frequency and comparative salience risks overlooking deeper discursive structures. For this reason, the next section will move beyond keyword counts to explore a selection of concordances and collocational patterns from both WoL-K1 and WoL-T1. This qualitative phase of the analysis seeks to understand how labour is discursively constructed more broadly, and whether gendered dimensions of work begin to surface implicitly through word associations, syntactic patterns, or semantic framing.

7.7.1.2 Exploring keywords further: Concordances and collocations

To avoid a purely intuition-based selection of keywords, a mixed-method criterion was adopted. Keywords were selected from the top 30 items based on three parameters: (i) their statistical salience in the keyword list, (ii) their thematic relevance to the legal-discursive construction of labour, and (iii) their potential to reveal implicit dimensions of power and gender.

Servant, *porter* and *native* were selected for the analysis. The keyword *servant* was included due to its possible application to both male and female labourers, making it a useful entry point for examining whether and how gender distinctions are embedded within occupational categories. The term *porter* was selected because of its frequent appearance and its association with physically demanding, low-status work, often delegated to indigenous populations. Analysing this term might help illuminate how labour roles were discursively constructed along both racial and gendered lines. The keyword *native* was included for its high relevance in both WoL-K1 and WoL-T2 as well as what it might reveal on labour duties or tasks based on ethnicity. Its analysis offers might offer insights into how legal language encoded broader structures of domination intersecting with class and gender hierarchies.

Moreover, the overall aim of this stage was not only to identify discursive patterns relevant to the regulation of labour but also to assess whether the chosen methodology proved effective in capturing meaningful linguistic trends. This step thus serves both as an exploratory investigation and as a methodological test case for the deeper analysis conducted on the thematic sub-corpora WoL-K2 and WoL-T2.

7.7.1.2.1 Concordances and collocations in WoL-K1

Word Sketch data for the term *porter*, which appears as a top keyword in WoL-K1 and appears in three different documents, reveals that it is most frequently modified by *native*, immediately pointing to a racially marked categorisation in language. Furthermore, *porter* is almost exclusively found as the object of regulatory verbs, such as *engage* (8 times), *provide* (2), *permit* and *register* (1), what could be an indicator of a somewhat passive legal subjecthood. In contrast, very few verbs portray porters as grammatical subjects, and when they do, these are often limited to existential or minimal-action verbs (e.g., *be*, *have*, *fall, die*). Frequent co-occurrence with *servant* (23) suggests a legal assimilation to other subordinated roles, while the dominant prepositional struc-

tures *of porter* (22), *to porter* (5), *for porter* (3) reflect a language of control and logistical provision rather than agency or participation. Altogether, *porter* emerges not only as a physical labourer but also as a racialised and dependent legal entity.

One of the most prominent collocations involving *porter* in the WoL-K1 corpus is the verb *engage*. As shown by the concordance lines in Table 10, *porters* are repeatedly framed as individuals to be "engaged for a journey" or "engaged for service," frequently by an employer or with the involvement of a colonial administrative officer.

Year	Left context	KWIC	Right context
1902 (1)	their breach, notwithstanding any contract or agreement that he may make to the contrary. 10. No person who has **engaged**	*porters*	may transfer them to any other person without the consent of the porters, testified before an authorized registering
1902 (2)	in accordance with any special directions of the Secretary of State. 13. The registering officer shall explain to the	*porters*	**engaged** for service with the caravan :- (a.) The place to which the porters are to go, or if the engagement is by time, the
1902 (3)	made of them on the expiration of the journey both to the Registrar and to the Collector of the District in which the	*porters*	were **engaged**. 32. Any porter who enters into an agreement to accompany or to hold himself in readiness to accompany a
1902 (4)	the following additions. 46. Any employer may apply to a registering officer for a permit authorizing him to **engage**	*porters*	or servants to leave or serve without the Protectorate. 47. In the application shall be stated the place to which it is
1910 (1)	. An employer shall when necessary and if requested by a servant, supply him with a suitable blanket, and in the case of a	*porter*	**engaged** for a journey, also with a jersey and water bottle. In any such case unless expressly agreed to the contrary the
1910 (2)	sufficient Tent accommodation. 28. An Employer shall when necessary provide sufficient tent accommodation for his	*porters*	**engaged** for a journey. Employer to provide Medicine and Medical attendance. 29. Every employer shall provide his
1938 (1)	(2) An employer shall when necessary and if requested by a servant, supply him with a suitable blanket, and in the case of a	*porter*	**engaged** for a journey also with a jersey and water-bottle. In any such case unless expressly agreed to - the contrary the reasonable cost of the article or articles supplied shall be paid by the servant and may be deducted from the remuneration

Table 10: Concordance lines for the lemma *porter* as the object of *engage* in WoL-K1.

This usage positions the porter as an object of contractual or logistical regulation, rather than as an active participant in the labour process. The verb *engage*, in these legal contexts, does not denote voluntary employment negotiations, but rather top-down assignments, often requiring permits or being conditional upon administrative approval. Hence, examples like the ones in Table 10 also imply that porter labour was subject to strict institutional control, including geographic displacement and coordination with the caravan system. The concordance lines further reveal that porter occurs alongside references to tent accommodation, medical provision, and blankets, reinforcing a discourse of regulated dependency. The porter is thus legally constructed not as an autonomous worker, but as a logistically dependent, physically mobile, and administratively controlled labour

unit. Interestingly, concordance line 1902 (1) shows the formulation "without the consent of the porters", which might suggest a degree of autonomy. Still, the framing of the clause still implies that porters are objects of transaction between parties. The use of the verb *transfer* evokes a legal logic more affiliated to property exchange than to labour, placing the porter within a system where consent does not equate to full agency but merely modulates control.

The Thesaurus function in Sketch Engine adds depth to the analysis of *porter* by highlighting lexical items that frequently occur in similar contexts, such as *task*, *servant*, *child*, *conviction*, *penalty*, *regulations*, *food*, *death*, *assistant*. These associations further suggest that porters are embedded within a semantic field characterised by subordination. The presence of legal-criminal vocabulary such as *penalty*, *conviction*, and *death* is interesting. These terms do not necessarily indicate that porters were themselves the primary subjects of criminalisation, their recurrence could likely reflect a discursive context in which legal responsibility, liability, and the regulation of porter-related activities were highly codified. The focus may therefore extend to those who employed, managed, or engaged porters, suggesting a broader legal framework concerned with preventing abuses or illegal practices involving this category of labour.

As previously mentioned, further insights into the discursive positioning of *porter* emerge from the recurrent use of the prepositional phrase *of porters*. As shown in Table 11, these concordance lines reinforce a regulatory framing: porters are typically discussed in terms of their recruitment, registration, identity, or engagement, which not only erases individual agency but also embeds porters within a system of legal surveillance and colonial labour management.

Year	Left context	KWIC	Right context
1902 (1)	for the recruitment of porters within the Province of Ukamba, dated the 11th June, 1896. Order as to **enlistment**	*of porters*	, dated the 22nd July, 1896. Order authorizing punishment of deserting porters, dated the 13th October, 1896. Orders
1902 (2)	He shall produce the signed list of porters or servants, and the officer, if he is satisfied as to the **identity**	*of the porters*	or servants, shall countersign the list and return it to the employer; until the list is so countersigned, the employer
1902 (3)	employer shall furnish to the registering officer a list in duplicate showing the **name**, **village**, **and district**	*of every porter*	or servant engaged, and the place to which he is to proceed, and the place of exit by which the porters or servants are to
1902 (4)	to Travel or Serve within the Protectorate. 42. Parts I and II of these Regulations shall also apply to the **engagement**	*of porters*	and servants respectively in places without the Protectorate to travel or serve within the Protectorate with the
1902 (5)	event of any dispute between the porters and the caravan leaders. 6. Such fees shall be charged for the **registration**	*of porters*	as the Registrar, with the approval of the Commissioner, may from time to time notify ; and until further notification

Table 11: Concordance lines for the prepositional phrase *of porters* in WoL-K1.

Proceeding to the analysis of the lemma *servant*, which appears in four documents in total, the Word Sketch reveals that among the most frequent modifiers are *domestic* (13) and *railway* (2), which point to two key domains of employment: private households and state infrastructure. However, the significantly higher frequency of *domestic* indicates that domestic service represented a far more prominent employment domain within the corpus. While railway suggests the presence of state-related employment, its limited occurrence compared to *domestic* may reflect the greater legal-discursive visibility, or perhaps institutional concern, around domestic labour during the analysed period. No other modifiers referring to distinct infrastructure domains appear with comparable frequency. The co-occurrence with *deceased* (2) suggests bureaucratic procedures involving death, succession, or legal responsibility, reinforcing the notion of the servant as a legalised and regulated category. Moreover, *servant* is very often used as an object with verbs like *recruit* (20), *engage* (14) and *employ* (15), indicating that servants seem to be primarily construed as legal subjects to be managed or sourced. Interestingly, verbs like *cause* (8) *induce* (6), *convict* (5), *harbour* (4), *prevent* (3), *follow* (3), *return* and *order* (2) may reveal an underlying worry around the movement, intent, and compliance of servants. These actions seem to portray them not just as passive recipients of employment but as subjects of surveillance and behavioural regulation susceptible to legal intervention or correction.

In the subject position, *servant* frequently appears with verbs such as *be* (37), *have* (11 times), *pay* (8), *enter* (4), *die* (3), *request*, *leave*, *work* (2). This seems to reflect a dual discursive structure: on the one hand, servants are described through their working status, contractual obligations, or eligibility for wages ("shall be paid", "has entered into a contract", "servant is working"); on the other hand, their presence in clauses involving death ("If a servant dies during") or inability ("a servant who is unable to read and", "servant was ill", "on which case the servant is absent from his place") reflects their vulnerability, highlighting the impact of their absence on the functioning of the household or workplace. In this sense, illness and death are labour-related contingencies that affect management responsibilities and potentially disrupt the organisation of work. This is reinforced by the three adjectival predicates listed by Word Sketch: *absent*, *ill*, and *unable*. In sum, the lemma frequently appears in clauses involving physical or cognitive weakness or in connection with mechanisms of control (*prevent*, *return*, *convict*). In these cases, the servant seems to be constructed as a passive or problematic figure in need of monitoring, regulation, or discipline.

Exploring further the range of terms that most frequently collocate with *servant* in WoL-K1 also proves to be highly valuable, especially when paying particular attention to those that may not initially appear relevant but could nonetheless offer interesting insights. Using the collocations tool in Sketch Engine, a search was conducted within a span of ±3 words around *servant*, with the attribute set to *lemma*, and with thresholds set to a minimum frequency of 5 in the corpus and 3 in the given range. The results were ranked using T-score, MI score, and logDice as statistical association measures.

As previously observed with the Word Sketch tool, preliminary results indicate that among the most salient verbal collocates are *recruit* (24), *engage* (22), and *employ* (18), each of which underscores the framing of *servant* within recruitment and contractual discourses. Additionally, the collocates *provide* and *supply* (12) are also noteworthy (see Table 12). Both collocates seem to be indicative of a framework in which the employer is cast in the role of caretaker rather than contractor, and the servant is conceptualised not as a rights-bearing subject but as a dependent recipient of basic necessities. The examples below, selected from colonial-era texts, underline the concern of ensuring minimal physical well-being for servants, especially in cases of illness or termination.

Year	Left context	KWIC	Right context
1910 (1)	engaged for a journey. Employer to provide Medicine and Medical attendance. 29. Every employer shall **provide** his servants	*servants*	with proper medicines during illness and also (if procurable) medical attendance during serious illness, and any
1910 (2)	desired by the employed, and (e) In the case of a foreign contract of service a stipulation by the employer to **provide** the	*servant*	with sufficient means of returning if he shall desire to do so, at the termination of the contract, to the place at which
1910 (3)	etc., when necessary to be supplied if requested by a Servant. 27. An employer shall when necessary and if requested by a	*servant*	, **supply** him with a suitable blanket, and in the case of a porter engaged for a journey, also with a jersey and water bottle
1910 (4)	, in like manner return the servant to the place of engagement should the servant wish to return. To **supply** food for	*servant*	‘s consumption when returning to place of engagement. 32. Every employer shall on the termination of the contract of
1938 (1)	such cases– (a) the employer shall, except when it is impossible for him to do so by reason of any default on the part of the	*servant*	, **provide** thirty tasks for such servant or shall **provide** thirty days' work for such servant; (b) the employer shall **provide** food for the servant or payment in
1938 (2)	during journey to be supplied. Where under the provisions of this Ordinance any person is required to **provide** a	*servant*	with transport to the place of employment of such servant, to return a servant to his home or to the place where he was
1938 (3)	means and is otherwise unable to obtain food for himself pending the determination of his complaint, he may cause such	*servant*	to be **supplied** with necessary food at the expense of the Government, but in such case the cost thereof shall be a debt due to
1938 (4)	of food for the servants consumption and the way back to the place of recruitment or engagement or shall **supply** to such	*servants*	such amount of money as will purchase a sufficient supply of food. RECRUITING. 38. Professional or private recruiter

Table 12: Concordance lines for the lemma *servant* collocated with the verbs *provide* and *supply* in WoL-K1.

At first glance, the apparent concern with ensuring that servants receive essential goods might seem to reflect a protective or benevolent legal stance. However, the consistent pairing of *servant* with concrete nouns such as *food*, *blanket*, *medicine*, *accommodation*, and *transport* should not be mistaken for a recognition of labour rights or legal equality. These lexical choices suggest a consciousness that prioritised the logistical management of labour rather than inclusion. Moreover, the discourse does not frame servants as independent subjects with entitlements. This is evident not only in the use of verbs such as *supply* and *provide*, which structurally position the servant as the object of the action, but also in other frequent constructions involving the verb *to be* (a brief analysis of this pattern is presented further below).

Notably, the term *servant* is used here instead of a more neutral or modern alternative such as *employee*. In the WoL-K1 sub-corpus, while *employee* appears only 4 times in the 1938 legislation, it rises to 141 occurrences in the 1976 act, 62 in the 1982 document, and 431 in the 2007 legislation. It is entirely absent from earlier texts (1902, 1910, and 1925).[71]

The analysis now proceeds with a further investigation of *servant* in WoL-K1, before concluding the examination of this lemma with a focus on its collocate *domestic* discussed in a subsequent section. Notably, high-frequency auxiliary verbs such as *be* (8), *shall* (28) *may* (21) also occur consistently in proximity to *servant*. The concordance lines for the verb *to be* seem to reinforce the framing of the servant as a passive, regulated subject rather than an agentive worker (see Table 13). While *be* is, by nature, a stative verb and often structurally passive, its contextual usage here consistently places the servant at the receiving end of legal determinations and administrative conditions.

[71] While this lexical shift is historically and discursively significant, likely reflecting broader changes in the socio-legal framing of work and legal subjecthood in postcolonial settings, its detailed exploration falls outside the scope of the present study. This thesis is specifically concerned with the representation of female labour, and a systematic analysis of general labour terminology in colonial and postcolonial Kenya and Tanzania would require a dedicated investigation. Nonetheless, these patterns are worth flagging as they further contextualise the linguistic and ideological frameworks in which women's labour was encoded.

Year	Left context	KWIC	Right context
1902	of these Regulations, shall have power to cancel or modify the contract, award damages to either party, and to order the	*servant*	to **be** conveyed home at the employer's expense. PART IV. Engagement within to Travel or Serve without the Protectorate.
1910 (1)	purpose of obtaining a contract or service he shall give a false name or address. Offences by Servants Class II. 48. Any	*servant*	may **be** fined any sum not exceeding the amount of two months' wages and in default of payment may **be** imprisoned with or
1910 (2)	contract the tenor and execution of which are not in conformity with this Ordinance shall be enforced as against a	*servant*	who **is** unable to read and understand writing. Any such contract shall be deemed executed in conformity with this
1910 (3)	himself or departing from service period of absence may be added to term of service. 51. When the offence of which any	*servant*	shall **be** convicted under this Ordinance, shall be the offence of absenting himself from or of departing from the
1910 (4)	for him to return to his home at the conclusion of his daily work, the employer shall at his own expense cause such	*servant*	to **be** properly fed and to be supplied with sufficient and proper cooking utensils and means of cooking. Provided,
1938 (1)	the provisions of section 42 of this Ordinance, not been complied with, at the expense of the employer. 41. Recruited	*servants*	to **be** brought before magistrate or justice of the peace. Recruited servants shall, as soon as possible after being
1938 (3)	magistrate or a justice of the peace, as near as may be convenient to the place of recruiting, who before permitting such	*servants*	to **be** taken or transported to the place of employment shall satisfy him self that the requirements of this Ordinance and

Table 13: Concordance lines for the lemma *servant* collocated with the verb *to be* in WoL-K1.

The servant is "to be conveyed home", "to be brought before magistrate", "to be taken or transported", "to be properly fed". In this narrative, the subject is acted upon and rarely exercises agency or initiates action. These constructions suggest that the colonial legal language encodes servitude in grammatical structures that foreground control. Servants are institutionally framed as logistical concerns to be managed by others. It seems that the verb *to be* not only to describe states or outcomes but also emphasises a power asymmetry, where action is imposed upon servants rather than begun by them.

A closer examination of the collocational pattern reveals that *servant* frequently appears in constructions such as *domestic servant*. Concordance lines for *domestic servant* (see Table 14), reveals a discourse centred around mechanisms of exclusion and exceptionality. In some instances, the domestic servant is not addressed as an active subject within the labour system, but rather as a category exempted from legal obligations, protections, or penalties that would otherwise apply to other workers. Phrases such as "does not include a domestic servant", "shall not render himself liable", and "alone shall not be deemed to be a foreign contract of service" suggest that the legislative texts may define this role negatively, by specifying when and how it is not included in general provisions. In other words, *domestic servant* is primarily a reference to an entity that the law does not cover. When paired with other types of roles (e.g. "caretaker", "cleaner") or conditions of employment (e.g. non-manual labour, indoor work), the domestic servant is mentioned only to

be exempted from the scope of certain legal regulations. In the line 1938 (2), the term is juxtaposed with skilled trades, individuals may be apprenticed either to a trade "in which art or skill is required" or, alternatively, as domestic servants. This contrast seems to imply a perception of domestic service as a fallback category of work.

It is important to note that the choice to examine *domestic servant* was motivated by a desire to understand whether this figure could in some way be associated with gendered roles, given that women in colonial periods were often employed as domestic servants (see Chapter 4.2.1). However, this has not been the case, and the results do not appear to reveal any specific linguistic patterns directly related to women's labour. Nonetheless, what emerges instead may itself be indicative of its peripheral status. The frequent pairing of the terms with other types of roles or the exclusion from certain legal provisions suggests a lack of formal importance or legal recognition of the job, which could be a noteworthy aspect to consider in further analysis. This pattern may reflect broader societal and legal perceptions of domestic work as a less regulated or marginal category, an insight that remains valuable for understanding the socio-legal framing of this employment sector.

Year	Left context	KWIC	Right context
1902	who is a native of Africa and who is engaged as an artificer, workman, or manual labourer, but does not include a **domestic**	*servant*	engaged for indoor work or any porters as hereinbefore defined. (i.) "Collector" means a Collector of a district and any person acting as such.
1910 (1)	service shall not render himself liable to the aforesaid penalties by inducing or attempting to induce such **domestic**	*servant*	or sailor to proceed to any place within the Uganda Protectorate or within the dominions of the Sultan of Zanzibar
1910 (2)	for a term of one year or to a fine of 1,000 Rupees or to both. Proviso. Provided, however, that an employer of a **domestic**	*servant*	or sailor engaged under a contract of service shall not render himself liable to the aforesaid penalties by inducing or
1925	a shop; but does not include an occupier or any person employed solely as a caretaker or as cleaner or other **domestic**	*servant*	: Provided that in any shop where not more than three persons are employed, the occupier of that shop not being the owner
1938 (1)	and any contract for service with a foreign state: Provided, however, that a contract for employment of a **domestic**	*servant*	for service in the Uganda Protectorate or in the Tanganyika Territory or within the dominions of the Sultan of Zanzibar […] alone shall not be deemed to be a foreign contract of service.
1938 (2)	deed of the apprenticeship, apprentice him to a trade or employment in which art or skill is required, or as a **domestic**	*servant*	, for any term not exceeding five years. 20. Apprenticeship of children without known relatives or a guardian Whenever
1938 (3)	a prescribed distance from any place of employment; and (c) operations for the engagement of personal and **domestic**	*servants*	and servants for the performance of non-manual labour. BREACH OF CONTRACT AND DISPUTES BETWEEN AND OFFENCES BY
1938 (4)	building or other structure whatsoever when he has reasonable cause to believe that any servant, other than a domestic	*servant*	, is living, residing or is employed thereon or therein and may make such inquiry and examination as may be necessary to

Table 14: Concordance lines in WoL-K1 of the lemma *servant* with *domestic* as a modifier.

After a brief analysis of the selected keywords in the WoL-K1 sub-corpus, attention now turns to the Tanzanian data.

7.7.1.2.2 Concordances and collocations in WoL-T1

The keyword chosen for closer inspection within WoL-T1 is *native*. When examined through the Word Sketch tool, the term *native* emerges both as a noun and as an adjective, revealing subtle but important differences in function and discursive framing. As a noun, native appears predominantly as the object of transitive verbs such as *recruit* (3), *induce*, *employ* (2), or *supply* (1). Such collocational patterns also seem to suggest that natives may be consistently cast as the passive recipients of external action or regulation, rather than as autonomous agents. Interestingly, *native* appears only once in a coordination structure (*servants* or *natives*), which further highlights its marginal presence in roles of equivalence or agency. As an adjective, *native* is used to modify a narrow set of institutional nouns: *authority* (8), *ordinance* (5), *vessel* (2), and *servant* (3). This pattern may reveal how the term normally functions not to describe individual identity or social status, but to signal a colonial administrative categorisation tied to control and differentiation from European subjects. Taken together, these patterns suggest that *native* in WoL-T1 is less about ethnic or cultural identity itself, and more about how colonised individuals are positioned within a framework of legal and labour-related control.

A closer inspection of the concordance lines for *native* (see Table 15) reveals that it is rarely used to denote individuals in their own right; rather, it is embedded within administrative structures aimed at defining employment and classification of African workers. In all the 1923 entries, the lemma *native* is repeatedly framed as an object to control ("engagement of natives", "native to be exhibited in a circus", "for each native employed", "natives to be employed as servants"). It is particularly noteworthy that the passive voice is used so frequently in these contexts, further emphasising the lack of agency attributed to the referent.

Interestingly, a shift begins to emerge in the later entries. In 1955 and 1962, the term *native* persists but it is no longer used predominantly in the passive voice; instead, it functions as the agent of the sentence. Even more striking, however, is the emergence of negated constructions such as "no native authority" and "nor any local or native authority," which signal a legal discourse that still challenges or restricts the native authority. The 1962 example marks a further evolution, with *native* appearing alongside the more neutral term *employee*, reflecting a slowly shifting legal

discourse toward more formalised labour rights, though without eliminating racialised terminology altogether.

Year	Left context	KWIC	Right context
1923 (1)	to regulate the relations between Employers and Native Servants and to control the recruiting and engagement of	*Natives*	for service within or without the Territory. Be it enacted by the Governor and Commander-in-Chief of the Tanganyika
1923 (2)	, sailor, boatman, porter, messenger, or in any employment of a like nature to any of the foregoing, and includes a	*native*	to be exhibited in any capacity in a circus, show, or exhibition; the expression "employer" means any person who enters
1923 (3)	employer resides in the Territory, the amount of a bond, if demanded, shall not exceed one hundred shillings for each	*native*	employed. (3) A bond entered into for the purposes of this section shall be enforceable by the administrative officer
1923 (4)	, procure or attempt to procure, seek for engagement, conduct, take charge of, supply, or undertake to supply	*natives*	to be employed as servants. Provided that the expression "labour agent" shall not apply to any person who procures or
1923 (5)	. Provided that the expression "labour agent" shall not apply to any person who procures or engages or conducts	*natives*	for his own bona fide domestic or personal service or business exclusively, or to any messenger or servant who procures
1955 (1)	recruitment, be deemed to be an authorisation to remain with him for the full duration of his term of service. 115. (1) No	*Native*	authority or headman shall– (a) engage in recruiting; (b) exercise pressure on persons to engage or not to engage for
1962 (1)	, whether or not such employee was employed in a substantive appointment. (3) Neither the Government nor any local or	*native*	authority shall be liable to pay any severance allowance to, or in respect of, any person who was employed in any of the

Table 15: Concordance lines of *native* in WoL-T1.

What emerges is a discursive pattern that aligns with colonial governance: natives are spoken of primarily in terms of their availability as labour and their relationship to state-controlled systems of employment. There is no suggestion of individual legal subjectivity or agency. Instead, native functions as a legal-administrative category, invoked to delimit obligations, permissions, and constraints in labour legislation.

In sum, the analysis of the keywords *porter*, *servant*, and *native* in WoL-K1 and WoL-T1 reveals a consistent discursive pattern of subordination and control. These terms are rarely framed as agents within legal discourse. Instead, they are predominantly constructed as passive recipients of regulation. The grammatical structures in which these appear underscore a framing in which labour is heavily mediated by colonial authority. Notably, even when the terms appear in subject positions, the associated predicates tend to reflect vulnerability, dependence, or limitation. Taken together, these findings highlight the centrality of administrative control and hierarchical differentiation in the legal representation of labour during the colonial period, offering a foundation for understanding how gendered and racialised dimensions of work may have been similarly constructed in the subsequent women-focused sub-corpora.

The focus will now shift to the core analytical section of this study: the investigation of the two thematic sub-corpora, WoL-K2 and WoL-T2. As in the previous sections, the analysis will begin with an exploration of the keywords in each sub-corpus.

7.7.2 Observations in WoL-K2 and WoL-T2

7.7.2.1 Keyword analysis

As was done for the investigation of WoL-K1 and WoL-T1, keyword extraction was carried out using the advanced settings interface in Sketch Engine. A minimum frequency threshold of 5 was applied, and results were filtered to exclude three terms: *kenya*, *tanzania* and *ordinance*. The analysis was conducted at the lemma level to capture different inflectional forms of the same lexical item. Tables 16 and 17 present the top 30 keywords identified for the WoL-K2 and WoL-T2 sub-corpora.

	Lemma	**Frequency**		**Lemma**	**Frequency**		**Lemma**	**Frequency**
1	ship	48	**11**	organ	7	**21**	woman	56
2	authorized	32	**12**	nomination	6	**22**	commission	231
3	rights	13	**13**	crew	6	**23**	juvenile	33
4	freedom	13	**14**	aforesaid	6	**24**	assembly	14
5	native	12	**15**	cabinet	6	**25**	convention	13
6	pound	9	**16**	governor	6	**26**	equality	32
7	selection	8	**17**	nominee	5	**27**	s	8
8	reside	8	**18**	trimmer	5	**28**	panel	8
9	away	8	**19**	stoker	5	**29**	shilling	20
10	women	7	**20**	convene	5	**30**	underground	7

Table 16: Most frequent keywords identified in WoL-K2 with *WoLCP-EN1* as reference corpus.

	Lemma	**Frequency**		**Lemma**	**Frequency**		**Lemma**	**Frequency**
1	decree	46	**11**	zanzibar	6	**21**	servant	5
2	ship	18	**12**	native	6	**22**	demolish	5
3	governor	16	**13**	british	5	**23**	transform	5
4	stoker	8	**14**	woman	48	**24**	transformation	5
5	women	8	**15**	juvenile	28	**25**	transmission	5
6	trimmer	8	**16**	master	16	**26**	ornament	5
7	port	7	**17**	also	6	**27**	vessel	13
8	rupee	7	**18**	generation	5	**28**	mine	12
9	legislative	7	**19**	harbour	5	**29**	inland	8
10	clock	6	**20**	electricity	5	**30**	dock	8

Table 17: Most frequent keywords identified in WoL-T2 with *WoLCP-EN1* as reference corpus.

As expected, gender-related lemmas such as *woman* and *women*[72] are present in both lists, occurring 63 times in WoL-K2 and 56 times in WoL-T2. It is a notably balanced distribution, especially considering that the Tanzanian texts contain significantly fewer tokens overall (see Table 5). However, despite their prominence, these terms appear alongside a wide range of lexemes that suggest other semantic domains are equally, if not more, salient. In WoL-K2, the presence of terms like *rights*, *freedom*, *equality*, *commission*, *assembly, organ,* and *nomination* is particularly noteworthy and likely reflects shifts in legal discourse over time: they may be associated with more recent documents. Nevertheless, the co-occurrence of these lemmas with terms such as *authorized*, *juvenile*, and *underground* also suggests a continued emphasis on regulatory and normative frameworks in earlier texts. Furthermore, terms like *native* and *juvenile* may point to an intersectional approach to legal classification, where gender intersects with age and ethnicity in the legal framing of labour.

In WoL-T2, by contrast, the keyword profile appears less oriented toward rights-based terminology and more toward administrative and colonial governance. Terms such as *decree*, *juvenile*, *master*, *servant*, *governor*, *legislative*, *rupee*, *zanzibar*, and *british* seem to reflect the colonial context in which these laws were formulated. This pattern may in part be influenced by the composition of the Tanzanian sub-corpus, which includes a high number of documents from the colonial period. As such, some discursive features may reflect corpus design choices as much as historical legal realities. The recurrence of *stoker*, *trimmer*, and *vessel* alongside *port*, *harbour*, and *dock* suggests a regulatory focus on maritime labour, raising questions about the implications of such a framework for women; specifically, whether these laws imposed restrictions or had other effects. Furthermore, the presence of terms like *mine* may relate to industries historically considered unsuitable or dangerous for women, potentially reflecting patterns of labour segregation or exclusion (see Table 25).

To further interrogate the discursive patterns identified and determine whether the limited lexical visibility of gendered terms in WoL-K2 and WoL-T2 is a result of corpus-specific features or the influence of the reference corpus, a supplementary keyword extraction was conducted using the *British Law Report Corpus* (BLaRC) as the comparative baseline. The keyword extraction

[72] Although *woman* and *women* are the same lemma, Sketch Engine does not display them as such in the keyword output, despite *lemma* being selected under keyword settings. To preserve the integrity of the results, no manual merging or alteration of keyword data was performed.

parameters mirrored those applied in previous analyses: a minimum frequency threshold of 5, lemma-based matching, and the exclusion of country names and generic legal markers. Tables 18 and 19 below present the top 30 keywords identified for the two thematic sub-corpora using BLaRC as the reference corpus.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	chairperson	30	**11**	vacancy	10	**21**	hundred	14
2	authorized	32	**12**	authorize	13	**22**	aggrieve	8
3	shilling	20	**13**	ship	48	**23**	eighteen	7
4	gazette	11	**14**	gender	24	**24**	persons	8
5	juvenile	33	**15**	therefor	6	**25**	contravention	21
6	native	12	**16**	nomination	6	**26**	fourteen	7
7	trimmer	5	**17**	equality	32	**27**	commission	231
8	stoker	5	**18**	sixteen	9	**28**	pound	9
9	women	7	**19**	organ	7	**29**	industrial	51
10	organization	8	**20**	mine	14	**30**	labour	30

Table 18: Most frequent keywords identified in WoL-K2 with BLaRC as reference corpus.

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	stoker	8	**11**	transformation	5	**21**	harbour	5
2	trimmer	8	**12**	rupee	7	**22**	sixteen	5
3	juvenile	28	**13**	eighteen	11	**23**	industrial	50
4	shilling	8	**14**	decree	46	**24**	undertaking	60
5	zanzibar	6	**15**	mine	12	**25**	transmission	5
6	gazette	5	**16**	earth	6	**26**	ship	18
7	ornament	5	**17**	demolition	5	**27**	extraction	6
8	women	8	**18**	dock	8	**28**	generation	5
9	waterway	6	**19**	demolish	5	**29**	clock	6
10	native	6	**20**	transform	5	**30**	woman	48

Table 19: Most frequent keywords identified in WoL-T2 with BLaRC as reference corpus.

A preliminary comparison of the keyword extractions using the two different reference corpora reveals some differences in both the nature and prominence of the terms identified. The use of BLaRC appears to highlight certain keywords related to the discourse around female labour, especially for WoL-K2, that were not detected when WoLCP-EN1 was used.

Table 18 presents the top 30 keywords in WoL-K2 using BLaRC and includes terms such as *chairperson*, *gender*, *persons*, *labour*, *contravention*, *industrial*, as well as age-related terms like *fourteen*, *sixteen*, and *eighteen*. These do not appear in Table 16, which is based on the WoLCP-EN1 comparison. Conversely, *freedom* and *rights*, which are present in Table 16, are absent from Table 18. This may suggest that these terms occur more frequently in the BLaRC

corpus, therefore reducing their relative salience in WoL-K2 when compared to a broader legal corpus. Other notable absences from Table 18 include *ship*, *selection*, *reside*, *away*, and importantly, the singular *woman*, which appears in Table 16, likely for the same reason of diminished salience due to its higher baseline frequency in BLaRC.

In the case of WoL-T2, Table 19 contains keywords not found in Table 17, such as *shilling*, *gazette*, *waterway*, *earth*, *eighteen*, *sixteen*, *industrial*, *undertaking*, and *extraction*. In contrast, keywords present in Table 17 but absent from Table 19 include *governor*, *legislative*, *british*, *master*, and *servant*, which seem more broadly associated with colonial labour rather than specifically with female labour.

These results suggest that, in the case of WoL-T2, the choice of reference corpus, whether WoLCP-EN1 or BLaRC, does not significantly alter the profile of gender-related keywords, as both tables reveal only limited lexical markers explicitly connected to female labour, aside from *woman*/*women* and *juvenile*. For WoL-K2, however, the shift in reference corpus does produce slightly more variation, with BLaRC highlighting additional terms of explicit thematic relevance to gender discourse, such as *chairperson*, *persons* and *gender.* It is also important to note that WoL-T2 is considerably smaller in size compared to WoL-K2 (8,298 vs. 18,907 tokens), which may partly explain the less pronounced differentiation in keyword output across reference corpora. The exercise confirmed the salience of several keywords that may have been overlooked or deprioritised when relying exclusively only WoLCP-EN1 as the baseline. The outcome reinforces the view that understanding which terms are genuinely characteristic of the target corpus through multiple reference comparisons is a valuable step in building a coherent analytical framework.

At this stage, however, the analysis revealed an unexpected lexical pattern. Even though all texts included in WoL-K2 and WoL-T2 explicitly reference women or gender in their titles (see Appendix G for a full list of the headings), the actual range and frequency of gender-indexical terms within the body of these texts was strikingly limited. For instance, a search across the entire WoL-K2 corpus using the concordance function in Sketch Engine, based on the set of gendered terms previously identified in WoL-K1 and WoL-T1 (see Table 8 and Table 9), shows that *woman*/*women* appear 56 times, *female* 6, *her* 14, *gender* 24, *herself* and *she* only once. Other expected terms such as *widow*, *wife*, *pregnancy*, *mother*, and *maternity* are entirely absent.

This lexical sparsity of explicit gender references is itself a meaningful discursive finding. It may suggest that, even in legislation overtly concerned with women or gender, gender is not

consistently articulated through direct lexical markers. Rather, it could be often implicitly embedded within legal formulations or hidden behind specific terms (*ship*, *authorized*, *rights*, *freedom*, *reside*, *away*; see Table 16) that emphasise regulation, restriction, or protection. In this sense, the absence of gendered language becomes a salient feature, prompting further investigation into how female labour is framed through alternative discursive strategies.

Moreover, a closer examination of the keyword list generated for WoL-K2 reveals that gender-specific terms like *female* and *her* appear in relatively low positions. Whereas other words, though not directly gendered, could potentially be relevant to understanding the discursive framing of female labour (e.g. *parent*, *child*, *discrimination*, *employ*; see Table 20). These patterns suggest that by strictly focusing on the highest-ranked gendered keywords one would risk overlooking important lexical items with thematic significance.

As a result, the analytical strategy was adjusted to follow a combined approach. Instead of relying solely on keyword salience derived from comparison with the WoLCP-EN1 reference corpus, the selection was expanded to also include terms with high absolute frequency and strong discursive or thematic relevance. This mixed criterion enables the inclusion of conceptually significant items that may not stand out statistically but nonetheless contribute to the legal-discursive construction of female labour.

Term	Keyword Rank Position	Absolute Frequency in Sub-Corpus
parent	32nd	32
chairperson	38th	30
gender	45th	24
child	59th	141
young	61st	63
age	64th	54
male	65th	8
discrimination	67th	18
persons	76th	8
employ	85th	119
female	305th	6
work	374th	38

Table 20: Examples of lower-ranked keywords of WoL-K2 with potential relevance to the discursive framing of female labour.

Given these observations, the subsequent stage involves selecting a core set of terms for in-depth analysis of their collocational and contextual patterns through concordance lines. While a wide range of terms emerge as strong candidates for further investigation due to their semantic and

statistical relevance, the scope of the current study requires a selective approach to maintain analytical clarity and ensure feasibility. For this reason, only 11 keywords in total were selected for qualitative analysis: *woman*, *female*, *rights*, *freedom*, *equality*, *chairperson*, *ship*, *undertaking* for WoL-K2; *woman*, *mine*, *juvenile*, *her* for WoL-T2. This final list includes items that (i) directly reference female identity or gender, (ii) evoke broader legal or institutional principles linked to gender equality and rights, and (iii) point to employment contexts or categories that intersect with issues of gender and labour segmentation.

7.7.2.2 Exploring keywords further: Concordances and collocations

7.7.2.2.1 Concordances and collocations in WoL-K2

The analysis focuses first on the term *woman*. To initiate the investigation, the Word Sketch tool was used to identify the most frequent lexical and grammatical collocates of *woman*, which appears 56 times in the corpus. The results highlight a marked asymmetry in the syntactic positioning of *woman*, which is predominantly found in object position (13 times) and more rarely functions as a subject (4 times). Among the verb-object constructions are *employ* (5), *engage*, *affect* (2), and *kill* (1). These collocates frame *woman* as the object of legal regulation ("any person who employs a woman", "a woman engaged in health"), socio-political consequence ("as it affects women and male young persons"), or even physical vulnerability ("woman is killed"). Such patterns already suggest a discursive construction in which women are not active participants in labour processes but are primarily positioned as recipients of normative, protective, or punitive action. Conversely, only two verbs were found with woman as subject: *hold* (3) and *be* (1). The verb *hold* appears in expressions such as "a woman holding a position", which may point to exceptional contexts of empowerment, but its overall rarity reinforces the subordinate positioning of women within the legal discourse. The imbalance between agentive and non-agentive uses may indicate that the concept of female labour is framed through grammatical passivity, aligning with a legal logic in which women are regulated rather than empowered. This framing is further reinforced by the analysis of modifiers and coordination patterns. The term *woman* is most often modified by *disability* and *person* and co-occurs in coordinated structures with terms such as *juvenile*, *youth*, and *person(s)*, suggesting a classification that assimilates women to other socially "vulnerable" groups.

The results from the Thesaurus function for *woman* provide a useful reinforcement of earlier findings, particularly the frequent co-occurrence with terms such as *person* (213) and *child*

(141), already identified with the Word Sketch tool. It also revealed further relevant lexical associations, including *undertaking* (60), *ship* (46), *work* (36), and *employer* (33). Some of these may point to legal restrictions on female employment in specific sectors, especially maritime or industrial contexts, potentially reflecting gendered patterns of occupational exclusion. The presence of *chairperson* (30) is also noteworthy, as it suggests discursive spaces where women are represented in institutional or leadership roles: this term appears exclusively in the 2011 document, reflecting a shift toward more gender-inclusive language in the legal discourse. Overall, however, the discursive landscape remains predominantly shaped by logics of classification, control, and protection, rather than active participation.

Hence, the fact that women's labour is frequently regulated alongside that of children is in itself indicative of a discursive tendency to infantilise women. Notably, despite being explicitly referenced in the titles of the sub-corpus texts, women are mentioned considerably less frequently in the legal texts themselves than children (149 times in WoL-K2) or youth-related categories, such as young persons or juveniles. An examination of the concordance lines for *woman* further supports this observation.

The lemma often co-occurs with references to legally or socially defined vulnerable groups, suggesting a classificatory logic that positions women within a framework of oversight and special regulation, rather than one that emphasises empowerment, agency, or economic participation.

Year	Left context	KWIC	Right context
1933 (1)	of Women, Young Persons and Children An Ordinance to carry out certain Conventions relating to the employment of	*women*	**, young persons and children**. No. XIV of 1933. Assented to 5th May, 1933. 1. This Ordinance may be cited as "the
1933 (2)	of fourteen years; "young person" means a person who has ceased to be a child and who is under the age of eighteen years; " woman " means	*woman*	a of the age of eighteen years or upwards; "employment" means employment in any labour exercised for the
1933 (3)	be in addition to and not in derogation of any of the provisions of any other Ordinance restricting the employment of	*women*	**, young persons, or children**. (2) Nothing in this Ordinance contained shall apply to an industrial undertaking or a
1933 (4)	, in the opinion of a duly authorized officer, a fit and proper person to have charge of such child. (5) No young person or	*woman*	shall be employed at night in any industrial undertaking, except to the extent to which and in the circumstances in
1933 (5)	of this Ordinance, that parent shall be liable to a fine not exceeding two pounds. (3) If any person employs a	*woman*	in contravention of this Ordinance he shall be liable to a fine of twenty pounds. (4) If any person who is by this
1933 (6)	is, in the opinion of a duly authorized officer, a fit and proper person to have charge of such child. (5) No young person or	*woman*	shall be employed at night in any industrial undertaking, except to the extent to which and in the
1948 (1)	. 7. Restriction on employment of women and young persons. Subject to the provisions of the next succeeding section, no	*woman*	**or young person** shall be employed between the hours of 7 p.m. and 6 a.m., in any industrial undertaking except– (a) young
1948 (2)	employer, and shall at all times be open to inspection by any duly authorized officer. 10. Restriction of employment of	*women*	in mines. No female shall be employed on underground work in any mine except in the following circumstances:– (a) a
1961 (1)	guilty of an offence and liable to a fine not exceeding five hundred shillings. 15. Penalty for unlawful employment of	*women*	. 12 of 1956, s. 11(b). Any person who employs a woman in contravention of the provisions of this Ordinance shall be
1961 (2)	specified in the charge or information is not the actual employer, shall lie upon the person alleging such fact. (3) If a	*woman*	**or juvenile** is found in any industrial undertaking, or workings as aforesaid, or in any mine as aforesaid at any time at
2011 (1)	under the National Commission on Gender and Development Act, 2003; "gender" means the social definition of	*women*	and men among different communities and cultures, classes, ages and during different periods in history; "gender
2011 (1)	policies, laws and administrative procedures in all political, economic and societal spheres; so as to ensure that	*women*	and men benefit equally, and that inequality is not perpetuated; "marginalised group" means a group of people who,
2011 (3)	from discrimination and relating to special interest groups including minorities and marginalized persons,	*women*	, **persons with disabilities, and children**; (d) co-ordinate and facilitate mainstreaming of issues of gender,

Table 21: Concordance lines of *woman/women* in WoL-K2.

A recurrent formulation across the documents is: “women, young persons or children”, and a significant portion of the concordances centres on explicit restrictions placed on women’s employment, particularly regarding time, location, and sector (prohibitions on night work, restrictions on mining work). Constructions such as “no young person or woman”, “shall not be employed”, “may be employed”, and “shall be liable to a fine” clearly illustrate the linguistic encoding of passivity attributed to women within the legal discourse. The general tone of the text is not of empowerment or inclusion. Instead, it focuses on the limits, conditional access, and legal responsibility.

The 2011 *National Gender and Equality Commission Act* represents a clear discursive shift, moving away from narrow and protective ideas toward a more inclusive, rights-based perspective.

The concordance lines reveal that *woman* primarily appears in contexts that emphasise equality, participation, and anti-discrimination. Importantly, the text does not treat women as a fixed biological group but as a social category shaped by different contexts. This shift in meaning also shows up in how the term is used: *woman* is no longer found in passive or restricted phrases, but rather in active ones like "women and men benefit equally".

Furthermore, it is worth noting that the words *rights*, *freedom*, *equality*, and *chairperson*, all key terms in the top 30-word list for the WoL-K2 sub-corpus (see Table 16 and Table 18), only appear in this 2011 legislative document.[73] Namely, the frequent use of these four terms points to a shift in discourse: from control and limitation to inclusion and recognition. These words are not isolated: they often occur in combinations like "gender equality", "freedom from discrimination", and "human rights", establishing a semantic field of a progressive form of governance rooted in constitutional values. The term *equality* appears both in the name of an institution (e.g., "national gender and equality commission") and within normative clauses that describe the Commission's goals: "promote gender equality and freedom from discrimination", "monitor the integration of the principles of equality in all policies". Here, *equality* functions as a guiding principle rather than a descriptive condition, it is something to be achieved and put in practice. This denotes a performative function: the legal language does not just reflect reality; it helps to shape a future where equality becomes real.

As mentioned, the word *freedom* appears almost exclusively in the phrase "freedom from discrimination", which is repeated consistently throughout the document (e.g., "create a culture of respect for the principles of equality and freedom"). Unlike a more general or abstract idea of freedom, as in freedom *to* do something, freedom *from* implies removing barriers and protecting people from injustice.

Although *chairperson* does not occur in close association with references to women in the concordance lines, this absence should not be interpreted as a lack of relevance. Rather, the adoption of *chairperson* in place of the more traditionally *chairman* signals an institutional intention to promote gender equality through inclusive language practices. Its presence alone is noteworthy as

[73] Even though the 2011 *National Gender and Equality Commission Act* is not about women per se and its title does not expressively mention women, the document was intentionally added to the corpus because of its institutional role and its focus on gender issues.

evidence of a broader discursive move toward gender-neutral representation in official legal discourse. However, the concordance lines show that *woman*/*women* seem to be mentioned only alongside other marginalised groups ("freedom from discrimination and relating to special interest groups including minorities and marginalized persons, women…"). This shows a commitment to inclusion but also raises a question whether the placement of women in this broad list risk reducing their specific experiences to general issues of vulnerability.

As shown in Table 22, *female* occurs only 6 times in the whole WoL-K2 sub-corpus.

Year	Left context	KWIC	Right context
1948	at all times be open to inspection by any duly authorized officer. 10. Restriction of eompyment of women in mines. No	*female*	shall be employed on underground work in any mine except in the following circumstances:– (a) a woman holding a
1961 (1)	by the Minister to be an authorized officer for the purposes of this Ordinance; "child" means a person, male or	*female*	, who has not attained the age of sixteen years; "employment" means employment in any labour exercised for the purpose
1961 (2)	meaning assigned to it in the Native Vessels Ordinance; "parent" includes a guardian: Provided that, in relation to a	*female*	child who is married and is living with her husband, the expression shall be construed as meaning her husband; "public
1961 (3)	a ship of war; "woman" means a woman of the age of eighteen years or upwards; "young person" means a person, male or	*female*	, who has attained the age of sixteen years but has not attained the age of eighteen years. 3. Application. 12 of 1956, s.3
1961 (4)	required to be registered under subsection (1) of this section. 10. Restriction on employment of women in mines. No	*female*	shall be employed on underground work in any mine except in the following circumstances– (a) a woman holding a position
1961 (5)	with the conditions (if any) endorsed upon the consent. (4) No person shall employ, or cause to be employed, any	*female*	child of or above the age of thirteen years, in circumstances which are calculated or are likely to cause, or do cause,

Table 22: Concordance lines of *female* in WoL-K2.

Although the number of concordance lines is limited, they offer clear insight into the broader discursive patterns surrounding women's labour. Most of these lines appear in legislative contexts where *female* is used in reference to employment restrictions, legal definitions, and age limits. One prominent theme is the regulation of female labour in mines, particularly concerning underground work. The exceptions allowed (e.g. holding a position of management) highlight a logic of exclusion, women's access to certain types of work is not assumed but treated as an exception to the rule. Moreover, the document also uses *female* in connection to marital status, implying a legal subordination of the married minor to her husband, reflecting a strong patriarchal context.

Another notable linguistic feature in the concordance lines for *female* is the frequent use of negation. Phrases like "no female shall be employed" or "no person shall employ, or cause to be employed, any female child" are clearly exclusionary, marking certain categories of people as

subjects to be restricted or barred from full participation in specific sectors of the labour market, reinforcing a discourse of exclusion.

Table 23 shows the concordance lines of the terms *ship* and *undertaking*. When looking at the most frequent collocates of *ship* through the collocation tool in Sketch Engine, *master* (8), *contravention* (6), and *except* (5) are the first three emerging terms. This may suggest that also the term *ship* appears predominantly within contexts of restrictive regulation and legal exceptions. The same seems to be valid for the term *undertaking,* appearing in sentences that restrict access to women and young children in industrial contexts.

Year	Left context	KWIC	Right context
1933	.the employment of any child in any specified trade or industrial undertaking. (4) No child shall be employed in any	*ship*	**except** to the extent to which and in the circumstances in which such employment is permitted under the Convention set
1961	of children in ships. 12 of 1956, s. 8, 15 of 1961, Sch. No child under the age of fifteen years shall be employed in any	*ship*	**except** a ship approved by the Minister as a school or training ship: Provided that– (i) the Minister may, subject to such
1948	and where a child or young person is taken into employment in any industrial or other undertaking or in any trade or in any	*ship*	in **contravention** of the provisions of this Ordinance on the production by the parent or with the privity of such parent,
1933	to this Ordinance, no young person shall be employed on work as a trimmer or stoker in any ship. (2) The **master** of every	*ship*	shall, if young persons are so employed therein, keep a register of those persons with particulars of their ages or
1961	this subsection shall not apply to the master of a native vessel. (2) An authorized officer may require the **master** of a	*ship*	who employs one or more juveniles as aforesaid to produce for inspection any register maintained by him under
1933	and proper person to have charge of such child. (5) No young person or woman shall be employed at night in any industrial	*undertaking*	, except to the extent to which and in the circumstances in which such employment is permitted under the conventions set
1948	of gain whether the gain be to a child, young person or woman, as the case may be, or to any other person; "industrial	*undertaking*	" has, with respect to employment, the following meaning– (a) mines, quarries, and other works for the extraction of
1961	8 of this Ordinance, no woman or juvenile shall be employed between the hour of 6.30 p.m. and 6.30 a.m. in any industrial	*undertaking*	: Provided that– (i) women or male young persons may be so employed in cases of emergencies which could not have been

Table 23: Concordance lines of the terms *ship* and *undertaking* in WoL-K2.

What is particularly noteworthy, however, is the near-total absence of explicit reference to women in these contexts as well. The legal provisions overwhelmingly address children and young persons, while women are almost always entirely omitted. Nonetheless, as mentioned for the analysis of previous terms, this discursive silence is meaningful. The juxtaposition of women with children and juvenile can be read as a form of legal and discursive infantilisation, as women are regulated not in their own right, but often in parallel with or through association with minors. In this way, labour law does not construct women as autonomous subjects, but rather as figures requiring the same forms of oversight and restriction traditionally reserved for children and youth. The absence

of a clear, distinct legal discourse around women in these domains reinforces the broader historical pattern of their marginalization in the workforce.

7.7.2.2.2 Concordances and collocations in WoL-T2

The Word Sketch for *woman* in the WoL-T2 sub-corpus reveals linguistic patterns that mirror those already observed in WoL-K2. At first glance, it does not appear to be substantial differences between the two corpora in terms of the roles and associations assigned to woman. Here, *child* is the only noun modified by *woman*, appearing in combinations like "women, children". As an object, *woman* is mostly linked to the verb *employ*, often in regulatory or even punitive contexts (see Table 24). Even when *woman* appears as the subject of a verb, it is in restricted forms (e.g. "woman is employed in contravention). Again, the linguistic pattern points to a figure that is more often acted upon than acting. The adjective predicates are particularly telling: those ones attached to *woman* are only *absent* and *contrary*.

The concordance lines for *woman*/*women* clearly reflect a discourse of restriction and legal exclusion. Most occurrences appear within negative constructions, such as "no woman shall be employed", or are framed around restricting language ("employs a woman in contravention", "is prohibited under the provisions"), which recur frequently across the data. This pattern reveals a legal narrative where the presence of women in the workforce is systematically controlled. Women are either explicitly blocked from certain forms of labour, such as night work or employment in mines, or allowed only under exceptional conditions, reinforcing their marginal status in industrial contexts. Furthermore, *women* rarely appears as active grammatical subject. When it does, it is in passive constructions (e.g. "any woman is employed in contravention"), underscoring a lack of agency. More commonly, woman is the grammatical object, as in "if a person employs a woman" or "employment of women", highlighting how legal language positions them as subjects to be governed and restricted.

Year	Left context	KWIC	Right context
1933 (1)	actual offender in the first instance without first proceeding against the employer or master. 6. If a person **employs** a	*woman*	in **contravention** of this Decree, he shall be liable on conviction to a fine not exceeding two hundred rupees. 7. If it
1938 (1)	the commencement of this Ordinance it shall not be lawful except as expressly provided in this Ordinance to **employ**	*women*	in night work in any industrial undertaking in the Territory. 4. Exemption in certain circumstances. The provisions
1938 (2)	three months, or to both. 6. Penalty. The proprietor, owner or manager of any industrial undertaking in which any	*woman*	is **employed** in night work **contrary** to the provisions of this Ordinance shall be guilty of an offence and shall on
1947 (1)	provisions of this Ordinance. 16. Penalties for contravening provisions of the Ordinance. Any person who **employs** any	*woman*	, or any person under the age of eighteen years, in a manner which is contrary to any of the provisions of this Ordinance or
1953 (1)	that the notice would expire during such absence. 7. Penalty for unlawful employment of women. (1) If a person employs a	*woman*	in **contravention** of any of the provisions of this Decree he shall be liable on conviction therefor to a fine not
1953 (2)	it appears to any magistrate on the complaint of an authorised officer that there is reasonable cause to believe that a	*woman*	is **employed** in contravention of this Decree in any place, whether a building or not, such magistrate may by order under
1953 (3)	age. 3. Restriction on employment of women in industrial undertakings and mines. (1) Save as hereinafter provided, **no**	*woman*	**shall be employed** at night in any industrial undertaking unless the industrial undertaking is one in which only
1953 (4)	is one in which only members of the same family are employed. (2) Except in such cases as may be prescribed, **no**	*woman*	**shall be employed** in any mine. 4. Exemption from provision of section 3 in certain circumstances. Notwithstanding
1953 (5)	six months, or to both such fine and imprisonment. (2) Where the contravention of this Decree consists in taking a	*woman*	into employment in any industrial or commercial undertaking, the person who so takes her into employment shall,
1953 (6)	is prohibited under the provisions of any other written law for the time being in force relating to the employment of	*women*	

Table 24: Concordance lines for *woman*/*women* in WoL-T2.

The concordance lines for the terms *juvenile* and *mine* also offer relevant insights (see Table 25). Even here the focus on women appears to be somewhat overshadowed by the emphasis on child labour. First, this is particularly evident in the frequent use of words like *juvenile* (28) and *young* (54) in the whole WoL-T2 sub-corpus, which occur more often than *woman* (48) or *female* (2), even though the texts are explicitly concerned with women as well. Second, in the concordance data, *juvenile* is linked to a female subject only twice. Moreover, as in the Kenyan corpus, women are mostly mentioned in restrictive or prohibitive legal contexts and are not presented as active participants. The figure of the woman remains largely passive, framed by rules and limitations rather than inclusion or agency.

Year	Left context	KWIC	Right context
1953 (1)	of the female sex without distinction of age. 3. Restriction on employment of women in industrial undertakings and	*mines*	. (1) Save as hereinafter provided, no woman shall be employed at night in any industrial undertaking unless the
1953 (2)	members of the same family are employed. (2) Except in such cases as may be prescribed, no **woman** shall be employed in any	*mine*	. 4. Exemption from provision of section 3 in certain circumstances. Notwithstanding anything contained in section
1953 (3)	Notwithstanding anything contained in section 3 a woman may be employed in an industrial undertaking (not being a	*mine*	) at night. (a) in unavoidable cases when an interruption of work occurs which could not be foreseen and which is not of a
1940 (1)	not to be employed in industrial undertaking No child shall be employed in any industrial undertaking. 10. No **woman** or	*juvenile*	to be employed on night work No woman and no juvenile shall be employed in any industrial undertaking between the hours
1940 (2)	shall be employed in any industrial undertaking. 10. No **woman** or	*juvenile*	to be employed on night work No woman and no juvenile shall be employed in any industrial undertaking between the hours of 5 p.m. and 7 a.m. 11. Employers to keep registers of
1940 (3)	by rules made by the Governor in Council under the provisions of section 18, nothing in this section shall apply to a	*juvenile*	employed in domestic service. 6. Juveniles not to be employed in certain capacities (1) No juvenile shall be employed–
1940 (4)	case. 7. No juvenile to be employed against the wishes of the parent or guardian No employer shall continue to employ any	*juvenile*	after receiving notice, either verbally or in writing, from the parent or guardian that the juvenile is employed

Table 25: Concordance lines of the terms *mine* and *juvenile* in WoL-T2.

As a final lemma, the pronoun *her* was selected for closer analysis. Several patterns emerge from the concordance lines for *her* in the WoL-T2 corpus (14). *Her* appears almost exclusively in legal constructions related to pregnancy, confinement, and maternity leave (e.g., "following her confinement," "absent from her work," "notice of dismissal during such absence"). This strongly anchors the female representation in biological and reproductive roles. The recurrence of passives and modal verbs ("shall not be permitted", "shall be allowed", "it shall not be lawful") reflects a legal discourse in which the woman is primarily a regulated figure rather than an acting subject. Her actions (e.g., absence, leaving work) are framed within permissions, restrictions, and compliance conditions set by others. Although these provisions may appear protective (e.g., safeguarding the right not to be dismissed during maternity leave), the concordances suggest a paternalistic framing. The law does not necessarily "empower" women but rather delineates narrowly defined situations where their rights are acknowledged. Moreover, the referent *her* is not used broadly to describe a generic worker but is tied specifically to reproductive functions. This reveals a semantic narrowing, where female workers are constructed primarily in terms of motherhood and domesticity, not economic participation per se.

Year	Left context	KWIC	Right context
1953 (1)	any industrial or commercial undertaking- (a) shall not be permitted to work during the period of six weeks following	*her*	confinement; (b) shall have the right to leave her work if she produces a medical certificate issued by a duly
1953 (2)	not be permitted to work during the period of six weeks following her confinement; (b) shall have the right to leave	*her*	work if she produces a medical certificate issued by a duly registered medical practitioner that her confinement will
1953 (3)	the right to leave her work if she produces a medical certificate issued by a duly registered medical practitioner that	*her*	confinement will probably take place within six weeks; (c) shall in any case, if she is nursing her child, be allowed
1953 (4)	practitioner that her confinement will probably take place within six weeks; (c) shall in any case, if she is nursing	*her*	child, be allowed half an hour twice a day during her working hours for this purpose. (2) Where a woman is absent from her
1953 (5)	2) Where a woman is absent from her work in accordance with paragraph (a) or (b) of subsection (1), or remains absent from	*her*	work for a longer period as a result of illness certified by a duly registered medical practitioner to arise out of
1953 (6)	practitioner to arise out of pregnancy or confinement and to render her unfit for work, it shall not be lawful, until	*her*	absence shall have exceeded a maximum period of six months for her employer to give her notice of dismissal during such
1953 (7)	her unfit for work, it shall not be lawful, until her absence shall have exceeded a maximum period of six months for her employer to give	*her*	notice of dismissal during such absence, nor to give her nouce of dismissal at such a time that the
1953 (8)	Decree consists in taking a woman into employment in any industrial or commercial undertaking, the person who so takes	*her*	into employment shall, whether or not he is the employer, be deemed to be the employer for the purposes of this section

Table 26: Concordance lines of *her* in WoL-T2.

All occurrences of *her* are concentrated in a single document from 1953. This absence from other texts prevents any meaningful diachronic analysis of its usage. At the same time, this restricted distribution serves as a significant indicator of the limited discursive visibility of female subjects and, in general, the broader absence of active female agency.

Following the analysis of the keywords, the next section offers a discussion of the findings on how the discursive representation of women's labour in the Kenyan and Tanzanian sub-corpora is framed.

7.8 Comparative interpretation

The comparative analysis of the Kenyan and Tanzanian legal sub-corpora reveals a set of converging discursive patterns that offer important insights into how female labour was historically constructed and regulated. Across both jurisdictions and corpus phases (general and gender-specific), women are consistently positioned at the margins of legal discourse, rarely as autonomous legal subjects and frequently as passive recipients of regulation. This marginality is not only evident in the relatively low frequency of explicitly gendered terms, but also in the analysed terms' structural patterns of collocation and syntactic role.

Women are discursively excluded from specific forms of work (e.g. night shifts or mining), through restrictive clauses, negative formulations, and exemption-based language. Employment in these sectors is not only regulated but also selectively prohibited, and women's access to them is constructed as exceptional. Such exclusions are reinforced through grammatical structures that foreground institutional control and frame the female worker in relation to reproductive functions or dependency status. Furthermore, women are often juxtaposed to juveniles and young persons, promoting a framing where women are not simply underrepresented but are actively assimilated to vulnerable, dependent categories, suggesting a discourse of infantilisation. However, WoL-K2 reveals more direct lexical markers of empowerment and has terms like *rights*, *freedom*, *equality*, and *chairperson*, but these stem from only one contemporary legislative act, raising questions about whether such inclusive language constitutes a systemic discursive shift or remains isolated.

Another important dimension is the limited diachronic visibility of gender-indexical terms. Many of the most revealing collocates (e.g., *her*, *female*) are highly localised in one or two documents, which restricts the ability to observe changes over time. Nonetheless, this patchy presence may only reflect archival gaps. In contrast, male pronouns and words like *servant*, *master*, *governor*, and *recruiter* occur robustly across the corpus, reinforcing the idea of male labour as the normative reference point in colonial and post-colonial legal constructions of work. Taken together, the findings underscore a recurring dynamic in which women's labour is regulated more through exception and protection than through agency and inclusion.

These results reveal how law not only reflects but helps produce segmentation through linguistic choices and lexical silences. In this sense, this legal archive is the evidence of how discursive form can illuminate the underlying power dynamics that shape women's historical and contemporary place in the world of work.

7.9 Concluding remarks

This study has presented a corpus-assisted discourse analytical investigation into the representation of female labour in colonial and postcolonial legal texts from Kenya and Tanzania, with a particular focus on the development, annotation, and analysis of dedicated sub-corpora. While the case study offered insights into how women are discursively positioned within labour law, it also functioned as a test case for assessing the methodological viability of constructing, annotating, and analysing a legal corpus within the framework of the *Worlds of Labour* (WoL) project.

Several areas for future development and expansion of the project emerged over the course of the research. Firstly, a significant opportunity lies in expanding the geographical scope of the corpus: including additional countries, particularly from other areas formerly under British colonial rule, would allow for a richer comparative analysis and a better understanding of how gendered labour discourses travelled across legal systems. Secondly, the current analysis could be enriched by comparing the usage of selected keywords from the two female labour sub-corpora (WoL-K2 and WoL-T2) with their occurrences in the general sub-corpora (WoL-K1 and WoL-T1), wherever applicable. By doing so, it would be possible to ascertain whether the discursive patterns found are specific to those texts or if they also occur in a larger legal discourse, albeit with possibly different connotations. This would offer a deeper understanding of how gendered meanings are used (or absent) in general labour law narratives. Thirdly, for reasons of time and scope, only a limited number of terms were examined. Hence, a future iteration of this research could involve a more extensive selection of keywords.

Moreover, this study engages with diachronic dynamics only to a limited extent. Although the sub-corpora include texts from both the colonial and postcolonial periods, the analysis mainly focused on thematic and lexical patterns rather than historical change. This was largely due to the limited time range covered by the Tanzanian sub-corpora. A diachronic analysis of Kenya alone would not have allowed for a meaningful comparison with Tanzania, given the lack of Tanzanian texts from the same range of years. Hence, another critical area for expansion involves enriching the sub-corpora themselves. However, while it would be valuable to include additional texts on female labour, any such additions would need to be carefully assessed against the selection criteria used in the original WoL corpus to ensure methodological consistency.

Additionally, the present study chose not to compare the Tanzanian and Kenyan sub-corpora to each other directly, opting instead for an internal, keyword-based exploration. This methodological choice, though more labour-intensive, proved valuable in uncovering subtle semantic and ideological dimensions that might have been flattened in a contrastive approach. However, it remains a viable and complementary methodological pathway for future research, as it could help balance depth and contrast.

Another main key limitation of this study was the superficial level of annotation. While basic metadata (e.g., document title, year, country) was consistently recorded, structural annotation remained limited. Only level 0 (articles) and level 1 (sections and section headings above article

level) structures were marked, without further hierarchical or thematic tagging. As a result, the analytical potential of structural filtering, especially through Sketch Engine's functionalities, was only partially exploited: it was primarily used to filter documents by country or year, rather than to isolate specific provisions within texts. Future work should explore deeper levels of annotation which would enable more granular analyses. This would enhance its usability by other members of the WoL research team involved in the leximetric studies as well.

In sum, the project showed that using a corpus-based approach anchored in careful methodological decisions about corpus design and annotation can offer meaningful contributions to the study of gender and labour in legal discourse. The tools used proved effective in managing, querying, and interpreting the data. At the same time, there is a need for improvement, expansion, and critical reflection, especially if this approach is going to be used in future studies comparing legal language across countries or over time.

8 Conclusion

The primary methodological aim of this dissertation was to develop and test an approach for the construction and annotation of labour law texts, with the goal of enabling the analysis of the phenomenon of legal segmentation from a discourse-oriented perspective that complements a leximetric approach. Specifically, the present contribution is the result of a collaboration with members of the *Worlds of Labour* (WoL) project, a research initiative at the *Global Dynamics and Social Policy* research centre of the University of Bremen, which investigates the phenomenon of legal segmentation, i.e., how labour law contributes to economic and social inequalities by granting uneven levels of protection to different categories of workers. Adopting a leximetric approach, a quantitative method that assigns numerical values to laws in order to assess their scope and intensity in a systematic way, WoL seeks to quantitatively measure the extent to which legislation contributes to legal segmentation. In this context, the present work has contributed to the expansion of the WoL project's objectives by introducing a complementary discourse-analytical perspective, aimed at identifying patterns of legal segmentation not only through numerical indicators, but also through the language and discursive structures embedded in the legal texts themselves.

Building on the broader initiative of the *Worlds of Labour Corpus Project* (WoLCP), involving members of the University of Bologna and the University of Bremen, this work adopted and extended the WoLCP framework to create four annotated sub-corpora based on labour legislation from two East African countries, Kenya and Tanzania, from 1902-2011. These were compiled using legislative texts retrieved from the WoL database and selected through a frequency-based approach: the most cited documents in the WoL coding templates were identified, cleaned, and ranked through regular expressions and Python scripts. The final selection was organised into two sub-corpora per country: one with general labour legislation and one focused specifically on female labour. Each text was made machine-searchable and semi-automatically annotated using a layered tagging system that included macro-structural segmentation (dividing each document into front matter, body, and back matter), micro-structural segmentation (identifying articles, sections, parts, chapters, and headings), metadata insertion, and the exclusion of non-linguistic or analytically irrelevant content through omission tags, such as tables, dates and signatures or tables of content. The case study conducted with Sketch Engine on female labour representation in the four

sub-corpora served to test the methodological framework and demonstrated how discourse analysis can reveal linguistic manifestations of legal segmentation.

Keyword extraction was used to identify statistically prominent terms, which were then analysed through concordance lines, collocation patterns, thesaurus results, and word sketches. Throughout this process, methodological choices were documented and reflected upon in detail, particularly in relation to corpus size, reference corpus selection, and the interpretation of frequency patterns. The results showed that legal segmentation is not only measurable through codified legal content but is also linguistically constructed and reinforced through the discursive framing of the legal subjects. The analysis revealed how in the selected documents women labour was mostly associated with that of minors, suggesting an infantilisation of adult female workers. Moreover, the texts showed a tendency to depict women in passive roles, lacking individual agency, and often referred to them in the context of protection rather than participation. The collocational behaviour of terms like *female*, *woman*, *her* indicated that female labour was discursively framed as a problem to be managed or regulated, rather than as an active economic contribution.

In this sense, the dissertation contributes new knowledge to the intersection of corpus linguistics, legal discourse analysis, and comparative labour law. It demonstrates that corpus-based methods can support the investigation of legal segmentation, offering an additional analytical layer to a leximetric approach. Furthermore, it provides a replicable methodological model for constructing annotated legal corpora. Future work could build on this foundation by expanding the corpus both horizontally and vertically, adding more countries and texts across time. This would make it possible to compare discursive patterns in a broader context. As for the findings, one current limitation of the present study is that interpretations are based on a limited number of texts and would ideally need to be supported by additional comparable material from other national or regional contexts, especially from African countries. To better support such findings, future studies should include legislative documents that allow for a discursive investigation of the representation of women's employment. The annotation model could also be enriched by adding semantic tagging linked to WoL's coding variables, thus strengthening the bridge between discourse analysis and legal data codification. Finally, the methodological framework developed here could be applied to new case studies that aim to explore other forms of segmentation, for example, based on migration status, ethnicity, or age.

In conclusion, this dissertation has shown that a corpus-based approach can meaningfully contribute to the goals of the *Worlds of Labour* project by uncovering the discursive dimensions of legal segmentation. The methodology developed and tested here not only provides a framework for the construction and annotation of labour law corpora but also offers a useful starting point for further interdisciplinary research at the intersection of language, law, and social policy.

Bibliography

Ackson, T. (2015). "Gender equality and labour law: Protecting working mothers, girls and female persons with disabilities in Tanzania". Law in Africa. 17–39. https://www.nomos-elibrary.de/10.5771/2363-6270-2015-1-17.pdf?download_full_pdf=1 [last accessed: 25 June 2025].

Adams, Z., Bastani, P., Bishop, L., & Deakin, S. (2017). The CBR-LRI dataset: Methods, properties and potential of leximetric coding of labour laws.

Akwei, I. (2017). "Kenya's history-making women elected governors, senators," 11/08, africanews https://www.africanews.com/2017/08/11/kenya-s-history-making-women elected-governors-senators// [last accessed: 25 June 2025].

Anderson, D. (2000). "Master and Servant in Colonial Kenya", 1895–1939. *The Journal of Afri-can History*. 41: 459-485.

Appiah, K. A., & Gates, H. L., Jr. (Eds.). (2010). Encyclopedia of Africa (Vol. 1). Oxford University Press.

Atieno, R. (2006). Female Participation in the Labour Market: The Case of the Informal Sector in Kenya.

Bakari, A. H. (1991). "Africa's paradoxes of legal pluralism in personal laws: comparative case study of Tanzania and Kenya". *African Journal of International and Comparative Law.* 3(3): 545-557.

Barchiesi, F. (2019). "Precarious and informal labour". In S. Bellucci & A. Eckert (Eds.), *General labour history of Africa: Workers, employers and governments, 20th–21st Centuries.* 45-75. Boydell & Brewer.

Bell, D. (1973). The coming of post industrial society. New York: Basic Books.

Biel, Ł. (2009). Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In C. Heine & J. Engberg (Eds.), R*econceptualizing LSP: Online proceedings of the XVII European LSP Symposium 2009*. 1-15. Aarhus School of Business.

Bosch, G. (1986). "Hat das Normalarbeitsverhältnis eine Zukunft?", in *WSI-Mitteilungen*, 3(90). 163-176.

Bronstein, A. (Ed.). (2009). International and Comparative Labour Law: Current Challenges. International Labour Organization.

Carlino, M., Fechner, H., & Schäfer, A. (Forthcoming). "Using Leximetrics for Coding Legal Segmentation in Employment Law: The Development and Potential of the Worlds of Labour Dataset." In *Constructing Worlds of Labour: Coverage and Generosity of Labour Law as Outcomes of Regulatory Social Policy*. Mückenberger, U., Fechner, U., Dingeldey, I. Palgrave Macmillan.

Carlino, M., Fechner, H. (2025). Coding Legal Segmentation in Employment Law. The Worlds of Labour (WoL) Dataset. Technical Paper Series, 22 Bremen: SFB 1342.

CRC 1342: Global Dynamics of Social Policy. (n.d.). https://www.socialpolicydynamics.de/about-the-crc-1342 [last accessed: 25 June 2025].

Curran, V. G. (2019). "Comparative Law and Language". In M. Reimann & R. Zimmermann (Eds.), The Oxford Handbook of Comparative Law. 680–709 (2019). Oxford University Press.

Darji, H., Mitrovic, J., & Granitzer, M. (2023). Challenges and Considerations in Annotating Legal Data: A Comprehensive Overview.

Deakin, S. (2013). Addressing labour market segmentation: The role of labour law. Governance and Tripartism Department. International Labour Organization.

Deakin, S. (2018). The Use of Quantitative Methods in Labour Law Research: An Assessment and Reformulation. Social & Legal Studies, 27(4): 456-474.

Deakin, S., Adams, Z., Bastani, P., & Bishop, L. (2017). The CBR-LRI Dataset: Methods, Properties and Potential of Leximetric Coding of Labour Laws. Centre for Business Research, University of Cambridge Working Paper No. 489. https://www.jbs.cam.ac.uk/wp-content/uploads/2023/05/cbrwp489.pdf [last accessed: 25 June 2025].

Deakin, S., Lele, P., & Siems, M. (2007). The evolution of labour law: Calibrating and comparing regulatory regimes. International Labour Review, 146(3–4). 133–162.

Dingeldey, I., & Gerlitz, J. (2022). Not just black and white, but different shades of grey: Legal segmentation and its effect on labour market segmentation in Europe. International Labour Review, 161(4). 593–613.

Dingeldey, I., Fechner, H., Gerlitz, J.-Y., Hahs, J., & Mückenberger, U. (2020). Measuring legal segmentation in labour law. SOCIUM; SFB 1342, 5.

Dingeldey, I., Fechner, H., Gerlitz, J.-Y., Hahs, J., & Mückenberger, U. (2022). "Worlds of Labour: Introducing the Standard-Setting, Privileging and Equalising Typology as a Measure of Legal Segmentation in Labour Law". *Industrial Law Journal*, 51(3): 560–597.

Eurofound. (2019). Labour market segmentation: Piloting new empirical and policy analyses. Publications Office of the European Union, Luxembourg.

Fall, B. & Roberts, R. (2019). "Forced labour". In S. Bellucci & A. Eckert (Eds.). *General labour history of Africa: Workers, employers and governments, 20th–21st Centuries*: 77–115. Boydell & Brewer.

Fechner, H. (2022). Legal segmentation and early colonialism in Sub-Saharan Africa: informality and the colonial exploitative legal employment standard. International Labour Review, 161(4). 615–634.

Finkin, M. W. (2019). "Comparative Labour Law". In M. Reimann & R. Zimmermann (Eds.), The Oxford Handbook of Comparative Law (2019). 1109–1136. Oxford University Press.

Finkin, M. W., & Mundlak, G. (Eds.). (2015). Comparative Labor Law. Edward Elgar Publishing.

Goźdź-Roszkowski, S. (2021). "Corpus Linguistics in Legal Discourse". *International Journal for the Semiotics of Law-Revue Internationale de Sémiotique Juridique*, 34(5): 1515–1540.

Gries, S. Th., & Berez, A. L. (2017). "Linguistic Annotation in/for Corpus Linguistics". In N. Ide & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (2017). 379–409. Springer Netherlands.

International Labour Organization. (2023). What is informal employment? ILO Brief. https://www.ilo.org/media/5481/download [last accessed: 25 June 2025].

Inter-Parliamentary Union (IPU), Parliament: United Republic of Tanzania. https://www.ipu.org/parliament/TZ [last accessed: 25 June 2025].

Jacobs, A. T.J.M. (2021). A Key to Comparative Labour Law in Europe. Open Press TiU. https://openpresstiu.pubpub.org/comparative-labour-law-europe [last accessed: 25 June 2025].

Karanja, Samuel. 2016 "More Needs To Be Done To Protect Women's Rights-Leaders." *Daily Nation*, March 8.

Konstantin A., Pfeiffer K. and L. (2021). Measuring Unmeasurable: How to Map Laws to Numbers Using Leximetrics. DIW Berlin Discussion Paper No. 1933. https://ssrn.com/abstract=3810489 [last accessed: 25 June 2025].

Le Crom, J.-P. (Ed.). (2017). Histoire du Droit du Travail dans les Colonies.

Lewis, Suzan. (2003). The Integration of Paid Work and the Rest of Life. Is Post-Industrial Work the New Leisure? Leisure Studies - LEIS STUD. 22. 343-345.

Liu, G. (2014). Private Employment Agencies and Labour Dispatch in China. Sectoral Activities. Working Paper No. 293. Geneva: International Labour Office.

Lo, V. I. (2023). "Labour dispatch in China: flexibility and security in employment". *Peking University Law Journal*, 11(2): 161–183.

Mahler, A. G. (2017). "What/Where is the Global South?" *Oxford Bibliographies in Literary and Critical Theory*. Eugene O'Brien. https://www.globalsouthstudies.org/what-is-the-global-south/ [last accessed: 25 June 2025].

Mbilinyi, M. (1975). Tanzanian women confront the past and the future. Volume 7, Issue 5: 400-413.

Mückenberger, U. (1985). "Die Krise des Normalarbeitsverhältnisses". *Zeitschrift für Sozial-reform*, 7: 415-33 and 8: 457-74.

Mückenberger, U. (1989). "Non-standard forms of work and the role of changes in labour and social security law". *International Journal of the Sociology of Law*, 17: 381-402.

Peruzzo, K. (2014). “Term extraction and management based on event templates: An empirical study on an EU corpus”. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 20(2): 151–170.

Peruzzo, K. (2013). Terminological Equivalence and Variation in the EU Multi-Level Jurisdiction: A Case Study on Victims of Crime. Università degli studi di Trieste.

Pittard, M. J., & Butterworth, S. (2015). “The rich panoply of sources of labor law: National, regional and international”. In M. W. Finkin & G. Mundlak (Eds.), *Comparative Labor Law* (2015). Edward Elgar Publishing.

Presbey, G. (2022) Women’s Rights in Kenya since Independence: The Complexities of Kenya’s Legal System and the Opportunities of Civic Engagement. *The Journal of Social Encounters*: Vol. 6. Iss. 1: 32-48.

Republic of Kenya, The Constitution of Kenya, 2010 (Nairobi: National Council for Law Re-porting with the Authority of the Attorney-General), Article 7(2). http://www.parliament.go.ke/sites/default/files/2023-03/The_Constitution_of_Kenya_2010.pdf [last accessed: 25 June 2025]

Rogowski, R. (2013). Reflexive Labour Law in the World Society. Edward Elgar Publishing Limited.

Rogowski, R. (2015). “The emergence of reflexive global labour law”. *Industrielle Beziehungen. The German Journal of Industrial Relations*, 22(1): 72–90.

Samuel, G. (1998). “Comparative Law and Jurisprudence”. The International and Comparative Law Quarterly, 47(4): 817–836.

Santosuosso, A., & Pinotti, G. (2020). Bottleneck or Crossroad? Problems of Legal Sources Annotation and Some Theoretical Thoughts. Stats, 3(3). 376–395.

Scarpa, F., Peruzzo, K., & Pontrandolfo, G. (2017). Methodological, terminological and phraseological challenges in the translation into English of the Italian Code of Criminal Procedure: What’s new in the second edition. In M. Gialuz, L. Lupària, & F. Scarpa (Eds.), *The Italian Code of Criminal Procedure. Critical Essays and English Translation*. Milano: Wolters Kluwer Italia. 53–80.

Shivji, I. G. (1986). Law, state and the working class in Tanzania: c. 1920–1964. James Currey; Heinemann. Tanzania Publishing House.

Sketch Engine. (n.d.). Metadata. https://www.sketchengine.eu/glossary/metadata/ [last accessed: 25 June 2025].

Stichter, S. (1977). Women and the labor force in Kenya, 1895–1964 (Discussion Paper No. 258). Institute for Development Studies, University of Nairobi.

Teklè, T. (Forthcoming). "The Leximetric Methodology Applied to the Analysis of Legal Segmentation in Labour Law: A Critical Appraisal". In *Constructing Worlds of Labour: Coverage and Generosity of Labour Law as Outcomes of Regulatory Social Policy*. Mückenberger, U., Fechner, H., & Dingeldey, I. Palgrave Macmillan.

The United Republic of Tanzania. Office of the Solicitor General. eLibrary. Constitution of the United Republic of Tanzania of 1977. https://www.constituteproject.org/constitution/Tanzania_1977.pdf [last accessed: 25 June 2025].

Tiersma, P. (n.d.). The nature of legal language. http://grammar.ucsd.edu/courses/lign105/student-court-cases/Tiersma.pdf [last accessed: 25 June 2025].

Tiersma, P. (1999). Legal language. Chicago University Press.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Amsterdam/Philadelphia: John Benjamins.

Tsikata, D. (2011). "Toward a Decent Work Regime for Informal Employment in Ghana: Some Preliminary Considerations". *Comparative Labor Law & Policy Journal* 2 (2): 311–342.

UN Women Africa. (n.d.) Women's Leadership and Political Participation (WLPP). https://africa.unwomen.org/en/where-we-are/eastern-and-southern-africa/tanzania/womens-leadership-and-political-participation [last accessed: 25 June 2025].

UN Women Tanzania (2024). Strengthening Meaningful Participation, Leadership and Economic Rights of Women and Girls at the Local Level in Tanzania https://africa.unwomen.org/sites/default/files/2024-04/abridged_version-strengthening_meaningful_participation_leadership_local_level_in_tanzania120124.pdf [last accessed: 25 June 2025].

UN Women. (2023). Gender pay gap in Kenya. https://africa.unwomen.org/sites/default/files/2024-03/un_women_kenya_gender_pay_gap_report.pdf [last accessed: 25 June 2025].

UN Women. (2023). Gender pay gap in Tanzania. https://africa.unwomen.org/sites/default/files/2024-03/un_women_tanzania_gender_pay_gap_report.pdf [last accessed: 25 June 2025].

UNICEF. (2017). Tanzania. The impact of language policy and practice on children's learning: Evidence from Eastern and Southern Africa 2017. Country Review. https://www.unicef.org/esa/sites/unicef.org.esa/files/2018-09/UNICEF-2017-Language-and-Learning-Tanzania.pdf [last accessed: 25 June 2025].

United Nations Industrial Development Organisation. (2024). Methodological document. UNIDO Country Classification. Edition 2024. https://stat.unido.org/portal/storage/file/publications/country-classif-report-2024.pdf [last accessed: 25 June 2025].

Wiesmann, E. (2022). Der Ausdruck der Konditionalität. Korpusgestützte Überlegungen zu Rechts- und Gemeinsprache. Syntax in Fachkommunikation, Berlino, Frank & Timme. 175 – 211.

Wiesmann, E. (2006). Terminologia e fraseologia del diritto. «MEDIAZIONI». Numero monografico (atti di: La formazione in terminologia, Portico di Romagna, 29).

Wiesmann, E., (2004). JUSLEX oder Die elektronische Verwaltung rechtssprachlicher Terminologie und Phraseologie. «JUR-PC». 1-45.

Wiesmann, E. (2004b). Neue Wege der Beschreibung von Rechtstermini: der semasiologisch-onomasiologische Ansatz der Begriffsbestimmung und seine Bedeutung für den Rechtsübersetzer. «LINGUISTICA ANTVERPIENSIA». LA - NS 3/2004: 37-51.

Appendix A – Excerpt from the Worlds of Labour Coding Template for South Africa[74]

This excerpt illustrates the structure and format used by the *Worlds of Labour* (WoL) research team to record, quantify, and interpret national labour legislation. The table reproduces a portion of the Excel template created as part of the leximetric coding process that was used internally by the project team. Each column corresponds to a specific variable or data point relevant for coding legal provisions. Among others, these include the variable name, the assigned value, the legal reference, a brief justification for the value, and a textual extract from the original legal source. The table shown here is not complete: for reasons of space and readability only the first row has been included and 3 out of 13 columns in total have been omitted.

Among the excluded columns are "Coding instructions (for CBR-LRI: reconstructed)", "Secondary source/literature", and "Commentary to Secondary source/literature or de facto application". The excerpt shown includes a selection of legal provisions from the Section 20(2) of the Basic Conditions of Employment Act 75 of 1997, used as part of the codification process for South Africa. For reasons of space and clarity, several sections of the text have been replaced with square brackets to indicate omitted portions of the original legal extract. The full text remains available in the WoL repository. This format preserves the structure and coding logic of the original document while improving readability in the printed appendix.

This excerpt comes from the document that marked the starting point of the WoLCP. As such, it represents a direct link between the leximetric coding carried out by the Bremen team and the subsequent stages, namely text selection, conversion, and annotation, undertaken during the WoLCP. For a full overview of the WoL methodology applied in the codification process, please refer to Chapter 3. Further discussion of how the indicators were used for the present thesis can be found in Chapter 7.

[74] The table continues on the following landscape page for legibility.

Version: 18.05.2022	Variable	Value assessment	CBR-LRI values	CBR-LRI explanation	CBR-LRI comment (jurisp./coll.agr./err.)	WoL Value	Year / period	Extract of norm(s)	Legal reference (law / official source)
S.1 = CBR-LRI 9	Annual leave entitlements	Measures the normal length of annual paid leave guaranteed by law or collective agreement. The same score is given for laws and for collective agreements which are de facto binding on most of the workforce (as in the case of systems which have extension legislation for collective agreements). The score is normalised on a 0-1 scale, with a leave entitlement of 30 days equivalent to a score of 1.	1997: 0,5	BCEA 1997: 3 weeks.	Mistake: BCEA 75 of 1997 came into force in 1998 and not 1997.	0,5	1998	CHAPTER THREE Leave 19. Application of this Chapter (1) This Chapter does not apply to an employee who works less than 24 hours a month for an employer. (2) […]. 20. Annual leave (1) In this Chapter, "annual leave cycle" means the period of 12 months' employment with the same employer immediately following- (a) an employee's commencement of employment; or (b) […]. (2) An employer must grant an employee at least- (a) […]; or (b) […]; (c) […]. (3) An employee is entitled to take leave accumulated in an annual leave cycle in terms of subsection (2) on consecutive days.	Section 20(2) of the Basic Conditions of Employment Act 75 of 1997

Appendix B – Final List of Legislative Documents for Corpus Inclusion[75]

The table presented here includes the final list of legislative documents forming the basis of the WoLCP corpus. These texts were identified through a comparative process involving two sources: (i) the frequency-based selection using a Python script (see Section 6.4) to analyse the *legal references* column in the WoL Excel templates, and (ii) the list of key national laws provided directly by the Bremen research team. The resulting harmonised list guarantees that, for each country, at least one core legislative document is represented. The only exception is Guinea-Bissau, which was excluded from the analysis due to its population being under 500,000 inhabitants. In the subsequent stages of the project, these documents were prepared for conversion to *.txt* format and annotation.

This table served as the main working document for compiling and managing the corpus materials. Originally, the spreadsheet included internal-use columns for task assignment, document status tracking, and workflow management. These supplementary columns were:

a) "Problematic Case": flagged when the most frequently cited legislation (identified via script) did not align with initial Bremen input, indicating the need for review;
b) "Analysis + Annotation": assigned each country to a specific user, enabling consistent tracking across the processing stages;
c) "Transformable document": indicated whether the legislation was accessible in a format suitable for *.txt* conversion (via copy-paste or OCR), including notes in case of technical issues.

[75] The table continues on the two following landscape pages for legibility.

Country	Main legislation currently in force	Document's language
Algeria	Code du travail	FR
Angola	General Labor Law of Angola	EN
Benin	Code du Travail 1998	FR
Botswana	Employment Act	EN
Burkina Faso	Code du Travail 2008	FR
Burundi	Loi n° 1/11 du 24 Novembre 2020 portant revision du Decret-Loi n° 1/307 du 7 juillet 1993 portant revision du Code du Travail du Burundi	FR
Cabo Verde	O Código laboral cabo-verdiano, Decreto-Lei n.º 5/2007	PT
Cameroon	Code du travail 1992	FR
Central African Republic	Loi n°08/004 portant de la Republique Centrafricaine & Act No. 60187, to prescribe legal public holidays in the Central African Republic. Dated 23 January 1961.	FR
Chad	Tchad code 1996 du travail	FR
Comoros	LOI N° 84-108/PR portant Code du Travail	FR
Congo Republic	Code du travail (Loi n°45-75 du 15 mars 1975)	FR
Djibouti	Code du travail Loi n°133/AN/05/5ème du 26 janvier 2006	FR
DR Congo	Democratic Republic of the Congo Labour code 1	FR
Egypt	Egyptian Labour Law 2003	EN
Equatorial Guinea	Law No. 10/2012 dated 24 December, on the Reform of the General Labour Law	EN
Eritrea	The Labour Proclamation of Eritrea. Proclamation No. 118/2001.	EN
Eswatini	Eswatini employment act of 1980	EN
Ethiopia	Labour proclamation no.1156 2019	Amharic & EN
Gabon	Gabon code travail 2021	FR
Gambia	Labour Act 2007	EN
Ghana	Labour Act 2003 Act No 651	EN
Guinea	Code du travail 2014	FR
Ivory Coast	Code du travail ivorien	FR
Kenya	The Employment Act Cap226 No 11 of 2007	EN
Lesotho	Labour Act 2024	EN
Liberia	Liberia Labour Law	EN
Libya	Law No. (12) of 1378 FDP (2010 AD)	EN
Madagascar	LOI N° 2003 -044 Portant Code du Travail	FR
Malawi	Malawi Employment Act 2000	EN
Mali	Mali Code du travail 1992	FR
Mauritania	Loi N° 2004-017 portant code du travail	FR
Mauritius	The Workers' Rights Act 2019 – Act No. 20 of 2019	EN
Morocco	Loi n° 65-99 relative au Code du travail	FR
Mozambique	Labour Law Mozambique 2023	PT
Namibia	Labour Act 11 of 2007	EN

Niger	Décret n° 2017-682/PRN/MET/PS du 10 août 2017 portant partie réglementaire du Code du Travail	FR
Nigeria	Nigeria labour act	EN
Rwanda	Establishing the labour code law no.51 of 2001	EN
São Tomé and Príncipe	Law no.62019 of November 16 2018	PT
Senegal	Code du travail	FR
Seychelles	Employment Act Cap. 69	EN
Sierra Leone	Employment act 2023	EN
Somalia	Code labour draft version 3 & Law No. 65 of 18 October 1972 to promulgate the Labour Code	EN
South Africa	Basic Conditions Employment Act Basic condonations of Employment Act amendments & Labor Relations Act 1995	EN
South Sudan	Labour act 2017	EN
Sudan	Labour code 1997	EN
Tanzania	Employment Act No-6-2004 & Employment Act (No. 11 of 2005) & Employment and Labour Laws (Miscellaneous Amendments) Act, 2015 & Employment and Labour Relations (General Regulations), 2017	EN
Togo	Loi n° 2006-010 du 13 décembre 2006 portant Code du travail	FR
Tunisia	Code du Travail 2018	FR
Uganda	Employment Act 2006 & Employment Regulations, 2011	EN
Zambia	Employment Code Act, 2019	EN
Zimbabwe	Labour Act as amended 2016	EN

Appendix C – WoLCP corpus documents (first version)[76]

This table presents the definitive list of legal documents selected for inclusion in the first version of the WoLCP corpus. Unlike previous tables organised by country, this one is structured by document. Each entry is assigned a unique identifier and a standardised name that will remain consistent throughout all phases of the project, including annotation and analysis.

Originally, this table also included two additional columns that were used internally to track technical issues, such as non-machine-readable PDFs or OCR errors, encountered during the text extraction phase. This version contains only the core fields needed for documentation purposes: the country of origin, the full name of the legislative text, the standard file name adopted for project use, and the document's language. These documents constitute the full set of texts that were successfully converted and officially included in the initial version of the WoLCP corpus.

[76] The table continues on the following landscape pages for legibility.

File nr.	Country	Legislation	Standard document name	Language
01	Algeria	Code du travail	01_Algeria_1996	FR
02	Angola	Lei General do Trabalho	02_Angola_1981	PT
03	Benin	Code du Travail de la République du Bénin	03_Benin_1998	FR
04	Botswana	Employment Act	04_Botswana_1984	EN
05	Burkina Faso	Loi n° 028 -2008/an portant code du travail au Burkina Faso	05_Burkina_Faso_2008	FR
06	Burundi	Loi n° 1/11 du 24 Novembre 2020 portant revision du Decret-Loi n° 1/307 du 7 juillet 1993 portant revision du Code du Travail du Burundi	06_Burundi_2020	FR
07	Cabo Verde	O Código Laboral Cabo-Verdiano Revisto (Decreto-Legislativo nº 5/2007, de 16 de Outubro, alterado pelo Decreto Legislativo nº 5\16, de 16 de Junho e Decreto-Legislativo nº 01\16 de 03 de Fevereiro)	07_Cabo_Verde_2007	PT
08	Cameroon	Cameroun Code du Travail Loi n°92-007 du 14 août 1992	08_Cameroon_1992	FR
09	Central African Republic	Loi n°09-004 du 29 janvier 2009 portant Code du travail de la République centrafricaine	09_Central_African_Republic_2009	FR
10	Central African Republic	Act No.60187, to prescribe legal public holidays in the Central African Republic. Dated 23 January 1961.	10_Central_African_Republic_1961	EN
11	Chad	Tchad Code du travail Loi n°038/PR/96 du 11 décembre 1996	11_Chad_1996	FR
12	Comoros	LOI N° 84-108/PR portant Code du Travail	12_Comoros_1984	FR
13	Congo Republic	Code du travail (Loi n°45-75 du 15 mars 1975)	13_Congo_Republic_1975	FR
14	Djibouti	Code du travail (Loi n°133/AN/05/5ème du 26 janvier 2006)	14_Djibouti_2006	FR
15	DR Congo	Loi n° 015/2002 du 16 octobre 2002 portant Code du Travail & Loi n°016/2002 portant création, organisation et fonctionnement des Tribunaux du Travail	15_DR_Congo_2002	FR
16	Egypt	Law No. 12 of the Year 2003 Promulgating Labour Law	16_Egypt_2003	EN
17	Equatorial Guinea	Law No. 10/2012 dated 24 December, on the Reform of the General Labour Law	17_EQ_Guinea_2013	EN
18	Eritrea	Proclamation No. 118/2001 - The Labour Proclamation of Eritrea	18_Eritrea_2001	EN
19	Eswatini	The Employment Act	19_Eswatini_1980	EN
20	Ethiopia	Proclamation No. 1156/2019	20_Ethiopia_2019	EN
21	Gabon	Code du travail 2021 (Loi n°022/2021 du 19 novembre 2021)	21_Gabon_2021	FR
22	Gambia	Labour Act	22_Gambia_2007	EN
23	Ghana	Labour Act (Act 651)	23_Ghana_2003	EN
24	Guinea	Code du Travail de la Republique de Guinee LOI N°L/2014/072/CNT Du 10 Janvier 2014	24_Guinea_2014	FR
25	Ivory Coast	Code du Travail Ivoirien	25_Ivory_Coast_2021	FR
26	Kenya	Employment Act Chapter 226	26_Kenya_2007	EN
27	Lesotho	Labour Act 2024	27_Lesotho_2024	EN
28	Liberia	Labour Law	28_Liberia_1972	EN

29	Lybia	Law No. (12) of 1378 FDP (2010 AD) on issuing the Labour Relations Law	29_Libya_2010	EN
30	Madagascar	LOI N° 2003 -044 Portant Code du Travail	30_Madagascar_2003	FR
31	Malawi	Employment Act No. 6 of 2000	31_Malawi_2000	EN
32	Mali	Loi n°1992-20 du 18 août 1992 portant Code du travail	32_Mali_1992	FR
33	Mauritania	Loi N° 2004-017 portant code du travail	33_Mauritania_2004	FR
34	Mauritius	The Workers' Rights Act 2019 – Act No. 20 of 2019	34_Mauritius_2019	EN
35	Morocco	Loi n° 65-99 relative au Code du travail	35_Morocco_2004	FR
36	Mozambique	Lei n.º 13/2023: Lei do Trabalho e revoga a Lei n.º 23/2007, de 1 de Agosto.	36_Mozambique_2007	PT
37	Namibia	Labour Act 11 of 2007	37_Namibia_2007	EN
38	Niger	Décret n° 2017-682/PRN/MET/PS du 10 août 2017 portant partie réglementaire du Code du Travail	38_Niger_2017	FR
39	Nigeria	Nigeria Labour Act Chapter 198 Laws of the Federation of Nigeria 1990	39_Nigeria_1971	EN
40	Rwanda	Law N° 51/2001 OF 30/12/2001 Establishing The Labour Code.	40_Rwanda_2001	EN
41	São Tomé and Príncipe	Lei n.º 6/2019 Aprova o Código do Trabalho	41_Sao_Tomé_2018	PT
42	Senegal	Code du travail Loi N° 97-17	42_Senegal_1997	FR
43	Seychelles	Employment Act Chapter 69	43_Seychelles_2020	EN
44	Sierra Leone	Employment Act 2023	44_Sierra_Leone_2023	EN
45	Somalia	Code labour draft version 3	45_Somalia_2018	EN
46	Somalia	Law No. 65 of 18 October 1972 to promulgate the Labour Code	46_Somalia_1972	EN
47	South Africa	Basic Conditions Of Employment Act, 1997	47_South_Africa_1997	EN
48	South Africa	Labor Relations Act 1995	48_South_Africa_1995	EN
49	South Sudan	Labour Act 2017 Act No. 64	49_South_Sudan_2017	EN
50	Sudan	Labour Code 1997	50_Sudan_1997	EN
51	Tanzania	Employment Act No-6-2004	51_Tanzania_2004	EN
52	Tanzania	Employment and Labour Laws (Miscellaneous Amendments) Act, 2015	52_Tanzania_2015	EN
53	Tanzania	Employment Act (No. 11 of 2005)	53_Tanzania_2005	EN
54	Tanzania	Employment and Labour Relations (General) Regulations, 2017	54_Tanzania_2017	EN
55	Togo	Loi n° 2006-010 du 13 décembre 2006 portant Code du travail	55_Togo_2006	FR
56	Tunisia	Code du Travail 2018	56_Tunisia_2018	FR
57	Uganda	Employment Act 2006	57_Uganda_2006	EN
58	Uganda	Employment Regulations, 2011	58_Uganda_2011	EN
59	Zambia	Employment Code Act, 2019	59_Zambia_2019	EN
60	Zimbabwe	Labour Act Chapter 28:01	60_Zimbabwe_2016	EN

Appendix D – Metadata Overview for the WoLCP and WoL-T, WoL-K[77]

The metadata attributes listed in the table below describe key characteristics of each legislative document included in the WoLCP corpus and the four sub-corpora. The first set of fields, up to and including *translation_source_language*, are relatively straightforward and can be easily interpreted with the help of the accompanying explanations. However, from *translation_method* onwards, including *document_origin* and *document_source*, the choices involved are more complex and reflect the need to differentiate between texts originally produced in English and those that were translated, either manually or automatically, within the scope of the project.

For *translation_method*, four values were established: manual, automatic, MTPE (Machine-Translation Post-Editing), and N/A. The N/A value is used exclusively for texts considered original in English and therefore not subject to translation. For translated texts, the method used is specified accordingly.

The attribute *document_origin* indicates whether a document was obtained externally (for example, sourced from an institutional partner such as the University of Bremen or official websites) or produced internally within the WoLCP framework (typically translations). This classification is crucial for future developments of the corpus, especially as new translated texts are integrated.

Finally, *document_source* captures the specific URL or platform from which the document was retrieved in the case of external sources. If the text was produced internally, and thus has no external provenance, this field is marked as N/A.

[77] The table continues on the following landscape page for legibility.

attributes	values		explanation
ID	-		unique ID for each text made up of a serial number, the country the text belongs to and the year of publication
country	-		the country the text belongs to
year_of_publication	-		the year when the text was published
heading	-		the official title of the legislation
legislation_type	Act, Code, Law, Proclamation		the type of legislation the text belongs to
legislation_language_status	official, non-official		whether the language in which the text is written can be considered official or not
translated_status	yes, no		whether the text is a translation or not
translation_source_language	FR, PT, AR, SO, AMH, N/A		the original language the text has been translated from
translation_method	automatic, manual, MTPE	N/A	the method with which the translation was carried out
document_origin	internal, external		"internal" indicates documents produced within the WoLCP project (e.g., translations); "external" refers to documents sourced from outside
document_source	-	N/A	URL of the official source if externally obtained; marked as "N/A" for internally produced documents

Appendix E – WoLCP English Corpus (WoLCP-EN1)[78]

The table presented here provides an overview of all the documents officially included in the first version of the WoLCP in English (WoLCP-EN1). This initial release contains only legislative texts that were considered originally produced in English, meaning no translated documents are included at this stage. The table includes only the first four metadata fields as an example: ID, country, year of publication, and heading. Additional metadata fields, such as translation status, translation method, and document origin, have been omitted here. A complete metadata record for the Tanzanian and the Kenyan sub-corpora can be found in Appendix G.

For a complete description of the metadata architecture for each text and access to the full dataset, readers are encouraged to consult the forthcoming technical documentation and research article, which are currently (2025) being finalised by the WoLCP team. The full version of the corpus, along with its documentation, will be made publicly available in the near future.

[78] The table continues on the following landscape page for legibility.

ID	country	year_of_publication	heading
04_Botswana_1984	Botswana	1984	Employment Act
16_Egypt_2003	Egypt	2003	Law No. 12 of the Year 2003 Promulgating Labour Law
18_Eritrea_2001	Eritrea	2001	Proclamation No. 118/2001 - The Labour Proclamation of Eritrea
19_Eswatini_1980	Eswatini	1980	The Employment Act
20_Ethiopia_2019	Ethiopia	2019	Proclamation No. 1156/2019
22_Gambia_2007	Gambia	2007	Labour Act
23_Ghana_2003	Ghana	2003	Labour Act (Act 651)
26_Kenya_2007	Kenya	2007	Employment Act Chapter 226
27_Lesotho_2024	Lesotho	2024	Labour Act
28_Liberia_1972	Liberia	1972	Labour Law
29_Libya_2010	Libya	2010	Law No. (12) of 1378 FDP (2010 AD) on issuing the Labour Relations Law
31_Malawi_2000	Malawi	2000	Employment Act No. 6 of 2000
34_Mauritius_2019	Mauritius	2019	The Workers' Rights Act 2019 – Act No. 20 of 2019
37_Namibia_2007	Namibia	2007	Labour Act 11
39_Nigeria_1971	Nigeria	1971	Nigeria Labour Act Chapter 198 Laws of the Federation of Nigeria 1990
40_Rwanda_2001	Rwanda	2001	Law N° 51/2001 OF 30/12/2001 Establishing The Labour Code
43_Seychelles_2020	Seychelles	2020	Employment Act Chapter 69
44_Sierra_Leone_2023	Sierra Leone	2023	Employment Act
45_Somalia_2018	Somalia	2018	Federal Republic of Somalia Draft Labour Code
46_Somalia_1972	Somalia	1972	Law No. 65 of 18 October 1972 to promulgate the Labour Code
48_South_Africa_1995	South Africa	1995	Labor Relations Act 1995 (Act No. 66 of 1995, as amended up to Prevention and Combating of Corrupt Activities Act 2004)
49_South_Sudan_2017	South Sudan	2017	Labour Act 2017 Act No. 64
50_Sudan_1997	Sudan	1997	Labour Code 1997
51_Tanzania_2004	Tanzania	2004	Employment and Labour Relations Act, 2004
52_Tanzania_2015	Tanzania	2015	Employment and Labour Laws (Miscellaneous Amendments) Act, 2015
54_Tanzania_2017	Tanzania	2017	Employment and Labour Relations (General) Regulations, 2017
57_Uganda_2006	Uganda	2006	Employment Act 2006
59_Zambia_2019	Zambia	2019	Employment Code Act, 2019
60_Zimbabwe_2016	Zimbabwe	2016	Labour Act Chapter 28:01

Appendix F - Frequency Lists of Legislative References for Kenya and Tanzania[79]

This appendix presents the frequency-ranked lists of legislative documents cited in the WoL coding templates for Kenya and Tanzania. The data were extracted from the *legal references* column in the Excel files compiled by the WoL research team as part of the project's leximetric analysis. Each entry corresponds to a specific legislative text and is accompanied by the number of times it was cited across the provisions recorded in the templates. The frequency values provide an empirical indicator of the legal documents most frequently referenced in connection with labour-related norms in each national context. The tables are presented separately for Kenya and Tanzania and are sorted by frequency, from highest to lowest.

These lists were used as the basis for selecting the core documents included in the general sub-corpora WoL-K1 (Kenya) and WoL-T1 (Tanzania), as described in Section 7.2.1. Not all the documents were included in the final sub-corpora. In some cases, this was due to lack of access or their relatively low frequency, while in others it was because the documents represented redundant or closely overlapping versions of the same law, for instance, different editions of ordinances with minor changes (e.g., the Kenyan Master and Servants Ordinance from 1906 and 1910). This frequency-driven overview helped to ground the corpus selection in empirical data and ensured that the general sub-corpora reflected the core legal sources most frequently referred to in labour-related context.

[79] The tables continue on the following landscape pages for legibility.

Legislative text	Frequency
The Employment Act, 2007 (No. 11 of 2007)	31
The Employment Act. Chapter 226 (No. 2 of 1976) [revised edition]	29
Master and Servants Ordinance (Ord. 8 of 1906)	27
Employment Ordinance, No. 2 of 1938	24
Master and Servants Ordinance, 1910	15
Native Porters and Labour Regulations, 1902	9
Shop Hours Ordinance, 1925	5
Regulation of Wages and Conditions of Employment Act, 1982	4
Labour Regulations 1898	4
Master and Servants (Amendment) Ordinance, 1924	3
Labour Institutions Act, No. 12 of 2007	3
Regulation of Wages and Conditions of Employment Ordinance, 1951	2
Minimum Wage Ordinance, 1946	2
Minimum Wage Ordinance, 1932	2
Shop Hours Amendment Ordinance, 1937	2
Employment and Labour Relations Court Act, 2011	2
Trade Disputes Act (Cap. 234)	2
National Cohesion and Integration Act, No. 12 of 2008	2
Regulation of Wages and Conditions of Employment Act, 1972	2
Mombasa Shop Hours, 1952	1
Public Holidays Act, 2018	1
Public Holidays (Amendment) Act, 1990	1
Public Holidays Act, 1984	1
Public Holidays (Amendment) Act, 1964	1
Shop Hours Act (Cap. 231)	1
The Public Holidays Ordinance 1912	1
Trade Disputes Act (Cap. 234) (Rev. 1972) = Law 22 of 1971, s. 6	1
Employment of Servants Ordinance, 1937	1
Employment of Natives Ordinance, 1910	1
The Finance Act 1994	1
Employment Agents Licensing Act, No. 2 of 1979	1

Table F.1 – Kenya: Frequency of cited legislative texts in the WoL leximetric database.

Legislative text	Frequency
Employment and Labour Relations Act, 2004	30
Employment Ordinance, 1955	8
Verordnung des Kaiserlichen Gouverneurs von Deutsch-Ostafrika, betreffend die Abschliessung von Arbeitsverträgen mit Farbigen, Deutsches Kolonialblatt, Jg 8 (1897), p.160-162.	5
Security of Employment Act, 1964	6
Ordinance Concerning the Making of Labour Contracts with Coloured Persons, 1896	6
Ordinance Concerning the Legal Position of Native Workers, 1913	5
Ordinance Concerning the Legal Position of Native Workers, 1909	5
(Verordnung des Gouverneurs von Deutsch-Ostafrika, betr. die Rechtsverhältnisse der eingeborenen Arbeiter. (Arbeiterverordnung). Vom 5. Februar 1913, Deutsches Kolonialblatt 1913, S. 396)	1
(Verordnung des Gouverneurs von Deutsch-Ostafrika, betreffend die Rechtsverhältnisse der eingeborenen Arbeiter. (Arbeiterverordnung). Vom 27. Februar 1909, Deutsches Kolonialblatt 1909, S. 367)	1
Master and Native Servants Ordinance, 1923	5
Regulation of Wages and Terms of Employment, 2010	2
Severance Allowance Act, 1962	2
Decree Repealing the 1896/97 Labour Contract Ordinance (1899)	2
Eduard von Liebert, Runderlass, 16 August 1899 (copy). German Federal Archives, R 1001/118, folio 107-108	2
Empioyment and Labour Relations (Code of Good Practice) Rules, 2007	2
Employment (Amendment) Ordinance 1960	1
Commencement unclear.	1
Public Holidays Ordinance (Amendment) Act, 1966	1
Employment Ordinance (Amendment) Ordinance 1960	1
Employment Ordinance (Amendment) Act 1962	1
CEACR, Direct Request on Equal Remuneration Convention, 1951 (No. 100), adopted 2006, published 96th ILC session (2007)	1
Regulation of Wages and Terms of Employment Ordinance, 1951	1

Table F.2 – Tanzania: Frequency of cited legislative texts in the WoL leximetric database.

Appendix G – Complete Metadata Records for WoL-K and WoL-T[80]

This appendix presents the complete metadata records used in the annotation of the four sub-corpora developed for the case study: WoL-K1, WoL-T1, WoL-K2, and WoL-T2. Each entry corresponds to a single legislative document and includes all the attribute values assigned during the metadata tagging process. These values were used to populate the metadata headers inserted at the beginning of each corpus file (see Section 6.5.1 for details on metadata design and implementation).

The metadata fields contain core information such as document title, country, year, language, as well as attributes related to translation status and translation method. Although the current sub-corpora only have texts originally drafted in English, the latter fields were included to ensure compatibility with future stages of the collaborative WoL project methodology. These fields are essential for tracking translation provenance, especially in view of the planned inclusion of translated legislative texts originally written in French and Portuguese. While the development and integration of these multilingual components fall outside the scope of the present study, the metadata structure was designed to accommodate them and ensure long-term consistency across the broader WoL corpus (see Section 6.5.1).

The metadata tables below are organised by country and sorted chronologically by sub-corpus.

[80] The tables continue on the following landscape pages for legibility.

ID	country	year_of_publication	heading	legislation_type	legislation_language_status	translated_status	translation_source_language	translation_method	document_origin	document_source
01_WoL-K1_1902	Kenya	1902	Native Porters and Labour	Regulation	official	no	N/A	N/A	internal	N/A
02_WoL-K1_1910	Kenya	1910	Master and Servants Ordinance	Ordinance	official	no	N/A	N/A	internal	N/A
03_WoL-K1_1925	Kenya	1925	Shop Hours Ordinance	Ordinance	official	no	N/A	N/A	external	https://gazettes.africa/akn/ke/officialGazette/government-gazette/1925-11-20/1/eng@1925-11-20/source
04_WoL-K1_1938	Kenya	1938	Ordinance No. 2 of 1938	Ordinance	official	no	N/A	N/A	internal	N/A
05_WoL-K1_1976	Kenya	1976	The Employment Act. Chapter 226 (No. 2 of 1976)	Act	official	no	N/A	N/A	internal	N/A
06_WoL-K1_1982	Kenya	1982	Regulations of Wages and Conditions of Employment Act	Act	official	no	N/A	N/A	external	https://new.kenya-law.org/akn/ke/act/ln/1982/120/eng@2022-12-31
07_WoL-K1_2007	Kenya	2007	The Employment Act	Act	official	no	N/A	N/A	external	https://kenya-law.org/kl/fileadmin/pdfdownloads/Acts/EmploymentAct_Cap226-No11of2007_01.pdf
01_WoL-K2_1933	Kenya	1933	Employment of Women, Young Persons and Children	Ordinance	official	no	N/A	N/A	external	https://digital.library.lse.ac.uk/Documents/Detail/kenya-1-ordinance-employment-of-women-young-persons-and-children/221116
02_WoL-K2_1948	Kenya	1948	Employment of Women, Young Persons and Children	Ordinance	official	no	N/A	N/A	internal	N/A
03_WoL-K2_1961	Kenya	1961	The Employment of	Ordinance	official	no	N/A	N/A	internal	N/A

			Women, Young Persons and Children Ordinance							
04_WoL-K2_2011	Kenya	2011	The National Gender and Equality Commission Act	Act	official	no	N/A	N/A	external	https://www.ngeckenya.org/Downloads/The_National_Gender_and_Equality_Act_2011.pdf

Table K.1: Metadata fields for legislative texts in WoL-K1 and WoL-K2 (Kenya).

ID	country	year_of_publication	heading	legislation_type	legislation_language_status	translated_status	translation_source_language	translation_method	document_origin	document_source
01_WoL-T1_1923	Tanzania	1923	Master and Native Servants Ordinance, 1923	Ordinance	official	no	N/A	N/A	internal	N/A
02_WoL-T1_1951	Tanzania	1951	Regulation of Wages and Terms of Employment Ordinance, 1951	Ordinance	official	no	N/A	N/A	internal	N/A
03_WoL-T1_1955	Tanzania	1955	Employment Ordinance, 1955	Ordinance	official	no	N/A	N/A	internal	N/A
04_WoL-T1_1962	Tanzania	1962	Severance Allowance Act, 1962	Act	official	no	N/A	N/A	external	https://www.chragg.go.tz/uploads/documents/sw-1669280827-57-1962%20The%20Severance%20Allowance%20Act.pdf
05_WoL-T1_1964	Tanzania	1964	Security of Employment Act, 1964	Act	official	no	N/A	N/A	internal	N/A

06_WoL-T1_2004	Tanzania	2004	Employment and Labour Relations Act, 2004	Act	official	no	N/A	N/A	external	https://webapps.ilo.org/static/english/inwork/cb-policy-guide/tanzaniaemploymentandlabourrelationsact2004sec626to7.pdf
01_WoL-T2_1932	Tanzania	1932	The Employment of Women, Children and Young Persons (Restriction) Decree, 1932	Decree	official	no	N/A	N/A	internal	N/A
02_WoL-T2_1938	Tanzania	1938	Employment of Women Ordinance, 1938	Ordinance	official	no	N/A	N/A	internal	N/A
03_WoL-T2_1940	Tanzania	1940	Employment of Women and Young Persons Ordinance, 1940	Ordinance	official	no	N/A	N/A	internal	N/A
04_WoL-T2_1947	Tanzania	1947	Employment of Women and Young Persons Ordinance	Ordinance	official	no	N/A	N/A	internal	N/A
05_WoL-T2_1953	Tanzania	1953	Employment of Women (Restriction) Decree	Decree	official	no	N/A	N/A	internal	N/A

Table K.2: Metadata fields for legislative texts in WoL-T1 and WoL-T2 (Tanzania).

Appendix H – Sample Annotation from *07*_WoL-K1_*2007*

This appendix presents a representative excerpt from one of the annotated legislative texts included in the WoL-K1 sub-corpora. The sample is intended to illustrate the full range of annotation layers applied to the corpus and reflects the standardised protocol described in Chapter 6 and Chapter 7.

The excerpt begins with a *<doc* tag, which marks the opening of each document file in the corpus. Immediately following this, the metadata section appears, containing all relevant attributes such as country of origin, year of publication, and translation status. The sample is structurally divided through macrostructural tags, which identify the three main sections of each legislative document: front, body, and back. Omitting tags like *<omitted type="table of contents"/>* is used to mark sections that were removed due to redundancy, lack of analytical relevance, or unreadability. In the body of the text, microstructural annotation is applied to reflect the internal hierarchical organisation of legal content. Each legal article is marked using level 0 tags, which identify the core provisions and their associated content. Level 1 tags are used to group related articles under broader divisions and include corresponding thematic headings to facilitate semantic navigation of the document.

To indicate that the full content is not reproduced in this appendix, a placeholder *[...]* is inserted. This example provides a comprehensive visual reference of how the documents both in the WoLCP-EN1 and in the Kenyan and Tanzanian sub-corpora were encoded and prepared for discursive analysis.

```
<doc ID="07_WoL-K1_2007" country="Kenya" year_of_publication="2007"
heading="The Employment Act Chapter 226" legislation_type="Act" legis-
lation_language_status="official" translated_status="no" transla-
tion_source_language="N/A" document_origin="external" docu-
ment_source="https://kenyalaw.org/kl/fileadmin/pdfdownloads/Acts/Em-
ploymentAct_Cap226-No11of2007_01.pdf" translation_method="N/A">
<section type="front">
LAWS OF KENYA
EMPLOYMENT ACT
<omitted type="table of contents"/>
CHAPTER 226
Revised Edition 2012 [2007]
Published by the National Council for Law Reporting with the Authority
of the Attorney-General www.kenyalaw.org
CHAPTER 226
EMPLOYMENT ACT
CHAPTER 226
EMPLOYMENT ACT
Date of assent: 22th October, 2007.]
```

```
Date of commencement: 2nd June, 2008.]
An Act of Parliament to repeal the Employment Act, declare and define the fundamental rights of employees, to provide basic conditions of employment of employees, to regulate employment of children, and to provide for matters connected with the foregoing
[L.N. 8/2008, L.N. 61/2008, Corr. No. 1/2008.]
</section>
<section type="body">
<structure type="level_1" title="PRELIMINARY">
PART I – PRELIMINARY
<structure type="level_0">
1. Short title
This Act may be cited as the Employment Act, 2007.
</structure>
<structure type="level_0">
2. Interpretation
In this Act, unless the context otherwise requires—
[…]
</structure>
</section>
<section type="back">
<omitted type="form"/>
</section>
</doc>
```