Academic Journal of Modern Philology

ISSN 2299–7164
Vol. 23 (2025)
Special Issue
141–159

Katarzyna Mroczyńska
University of Siedlce, Institute of Linguistics and Literary Studies
katarzyna.mroczynska@uws.edu.pl
ORCID: 0000-0003-0367-1056

Collocations of *Sex* and *Gender* in Legal and General Corpora

Abstract

The present study attempts to contribute to the research of collocations in both general and specialized language. Thus, it fits into Trajectory 1 of research in corpus-based studies in legal language carried out by Biel (2010) that explores how legal language differs from general language or other languages for special purposes. Specifically, it aims at looking at variation in candidate collocations of the two selected terms, *i.e. sex* and *gender*, as extracted from a general corpus and a legal one. We investigated the senses the items carry in the specialized and general corpora, the overlap of candidate collocates, and their ranks in frequency lists generated for each corpus. The findings suggest that the two terms in question display considerable differences in their collocational profiles and their combinatory potential. We offer an explanation for the observed significant variations in the way the collocations are represented and conclude that the type and nature of corpora affect the collocational profile of a lexical item. We also discuss the implications this may have for conducting research into terminology or lexicography and for devising resources for language learning and teaching.

Keywords: corpus studies, equal opportunities, legal English, general English, gender, sex, EU regulations

1. Introduction

Research into collocations, and the term *collocation* itself, goes back to Firth and his often-quoted statement "You should know a word by the company it keeps" (1968: 179), which inspired numerous scholars to investigate the area. This interest has resulted in a vast number of publications applying various methodologies and aiming to reach individual objectives. These varying methodologies and objectives have resulted in a situation where researchers find it difficult to adopt a uniform, widely accepted definition

doi: 10.34616/ajmp.2025.23.11

of *collocation* which will cover all the linguistic features of the phenomenon[1]. However, we may attempt to compile a list of criteria of collocability based on characteristics appearing across various studies on collocations. The features which will allow us to classify a given phraseological unit as collocation are as follows:

a) frequency of co-occurrence;
b) combinatory restrictions;
c) degree of compositionality;
d) degree of transparency;
e) span of words between node and collocate, or collocational window. (Patiño 2014: 122–124)

Initially, research into collocations focused on general language word combinations and took advantage of large corpora, *e.g.* The British National Corpus – BNC, The Corpus of Contemporary American English – COCA. The advent of widely-available information and communication technology (ICT) and computer-based tools opened up new opportunities allowing researchers to harvest huge amounts of data, create customized databases, and conduct more in-depth analyses of corpora of a size and on a scale never seen before. Thus, researchers became able to actively collect word combinations and organize them in lexical and terminological resources such as general language and specialized dictionaries, word lists, glossaries or other reference works, printed and more and more often online.

As L'Homme and Azoulay (2020: 150–151) argue, culling word combinations from corpora may yield varying results due to a number of factors such as the adopted definition of collocation and consequently types of word combinations accepted, the nature of a reference work (*e.g.* general or specialized one), methods and tools used for retrieving word combinations and, last but not least, the nature of the corpus used. It may be worth noting that the results may also differ because of the perspective a researcher takes; a lexicographer, a terminologist or a language teacher devising teaching resources will analyze and consider the same language material from different angles.

In this study, we make an attempt to examine the variation in word combinations caused by the nature of the corpus used and its effect on the collection of collocations. For the purpose of this study, two types of corpora are used, *i.e.* a general language corpus (the readily available BNC) and a specialized corpus containing legal texts dealing with equal opportunities and published by the EU.

The EU English corpus constitutes an interesting area for research as EU legal English is unique. On the one hand, it demonstrates the qualities of a legal English genre such as lack of transparency and obscurity, frequent use of formal words, complex syntactic structures, deliberate use of expressions with flexible meanings or the opposite, *i.e.* attempts at extreme precision (Northcott 2012: 215)[2], and on the other hand, the EU law and the EU legal English, one of the 24 official languages, habitually used for

1 An in-depth discussion of research frameworks regarding collocations is beyond the scope of this study as the phenomenon has been widely investigated in linguistic literature on numerous occasions. The readers who are less familiar with the topic are referred to Sinclair (2004), Kjellmer (1994) or Lehecka (2015) for details of a frequency-based approach, to Cowie (1994), Mel'čuk (1998), Hausmann (1997) or González-Ray (2002) for a semantic-oriented view, and to Siepmann (2005, 2006) for a relatively new, pragmatically-driven approach. A concise overview of various approaches is offered in Michta, Mroczyńska (2022).

2 For further discussion of characteristics of the legal English genre see among others Melinkoff (1963), Danet (1980), and Maley (1987).

drafting proposals, can be viewed as "a melting pot for national legal systems, languages and cultures" (Biel 2015: 142). Taking all that into consideration, it may be argued that the EU law is a multilingual result of hybrid translator-mediated communication and it affects the language of the EU law when it comes to its conceptual structure, lexis, grammar, stylistics, creating a hybrid construct of some sort (Biel *et al.* 2018: 251–252). The English language used in EU institutions is not standard English as used in the UK or Ireland but a "Europeanized" variety of the national legal and administrative English language which is transformed to meet the linguistic needs of the European Union as a supranational organization (Biel 2020: 478).

At this point, it may also be worth mentioning that the concept of collocation does not only refer to textual statistics, but it reflects a mental representation of the lexicon, as collocations are formed through the cognitive process of priming. We may distinguish three elementary types of priming: collocation, colligation, and semantic preference/association, with the priming of collocations in this psycholinguistic sense being the foundation of language structure in general (Hoey 2005: 8–9). Considering these findings, it may be assumed that the knowledge of how words collocate forms an integral part of knowing a language or a genre with a speaker's ability to adhere to collocational conventions demonstrating his/her mastering of the language within a given specific genre. Moreover, collocations, or word patterning, illustrate the non-random nature of language (Kilgarriff 2005) and as such offer a way to gain insight into how language works and reflect a language conceptual structure. Corpus linguistic research has shown that language is highly patterned and more importantly this pattern is cognitively motivated (Stubbs 2004: 111).

The paper is structured as follows: Section 2 describes the aims, scope and methodology of the contrastive analysis undertaken as part of this project; in Section 3, we analyze the collected language material by comparing lists of collocates extracted from the two corpora under review; and in Section 4, concluding remarks are presented along with some suggestions for further research.

We believe that a contrastive analysis of the combinatory potential of lexical items may contribute significantly to the improvement of knowledge of the legal language and of the workings of the law itself.

2. Aims, Scope and Methodology

2.1. Aims and Scope

The present work aims to make a contribution to the study of specialized legal collocations by offering a comparative analysis of collocations of *sex* and *gender* retrieved from specialized legal and general corpora. The purpose of this study is two-fold, and can be described as follows:

(1) to analyze the combinatory potential of *sex* and *gender* as employed in specialized legal corpora of EU documents regarding non-discrimination issues as compared to the combinatory potential of the two lexical items in a general language corpus (the BNC);
(2) to analyze different facets of collocational behaviour such as polysemy and synonymy of the lexical items, characteristics of collocates, their ranks, and overlap.

For the purpose of this study, two language corpora have been selected: one general and one legal. The legal corpus contains documents regarding equal treatment of men and women. It comprises a set of legal documents of various genres ranging from the EU primary and secondary legislation (such as the Treaty on European Union, the Treaty on Functioning of EU, the European Convention on Human Rights, the Charter of Fundamental Rights of the European Union, and EU Directives), ancillary documents (*e.g.* proposals for directives, strategies, recommendations, action plans, a handbook on European non-discrimination law or other guidelines regarding equal opportunities in the EU) to judgements of the Court of Justice dealing with non-discrimination.

A supranational organization, the European Union, has always felt strongly about non-discrimination on any grounds and equal treatment of men and women was one of the main principles the European Community was founded on. The body of legal regulations in this area has grown considerably over time to supplement the primary legislation that did not always cover the issue in an explicit manner[3]. Consequently, subsequent revisions of the treaties that emphasized human dignity, freedom, democracy, equality, the rule of law and respect for human rights were introduced and led to the Union recognizing them as founding values. These values are embedded in the treaties as well as mainstreamed into all EU policies and programmes. Moreover, new bodies have been established within the EU such as the European Union Agency for Fundamental Rights (FRA) or the European Institute for Gender Equality (EIGE) with the aim of safeguarding and promoting fundamental rights and equality (Council of Europe: European Court of Human Rights 2018: 21–23). This heavy institutional emphasis put on non-discrimination and equal treatment resulted in a relatively vast body of documents regulating the issue and a significant number of new phrases and concepts being introduced.

It seems that the most important developments in the human rights area are happening under the EU auspices and the EU is at the forefront of safeguarding equal treatment (see: Buzmaniuk 2023). Moreover, the way EU legislation is adopted in member states may lead to a situation in which some concepts appear, and consequently terms are coined first in Eurolect and later in national languages. As EU institutions can determine their working languages, a common practice is to resort to the most frequently spoken languages. Therefore, in most institutions there are three procedural languages: English, French, and German, with English having a dominant role in most institutions since the 2000s (Biel 2020: 481). The translation of regulations into all official EU languages is the next stage of the process. Thus, we believe that the English corpus of EU documents regarding equal treatment may represent concepts and reflect tendencies in the language used in this area.

Although we acknowledge that the corpus compiled in such a manner is relatively small, we do hope that our approach will ensure that the language material for the study will be reliable and up-to-date[4]. As the topic the corpus refers to, *i.e.* equal opportunities, is a narrow, specialized area, the number of available relevant texts is also somewhat limited[5]. After all, we need to bear in mind that any corpus is a kind of compromise between what is planned and desired by the designer and what is possible, for

3 [At:] https://www.europarl.europa.eu/factsheets/en/sheet/59/equality-between-men-and-women [date of access: March 16, 2025].

4 The decision which texts to include in the corpus was based on the summaries of EU legislation in the area of equal opportunities found [at:] https://eur-lex.europa.eu/EN/legal-content/glossary/equal-opportunities.html [date of access: March 16, 2025].

5 For a more detailed discussion of building and using small specialized corpora see: Koester (2010).

example in terms of available language input or time restrictions (Hunston 2008: 156–157). It should be emphasized that the corpus in question does not lay a claim to be exhaustive. Therefore, it should not be assumed that a collocation is definitely invalid in a legal or paralegal text covering equal opportunities if it does not occur in our corpus. In terms of size, the corpus contains 75 documents, 467,472 words, and 594,449 tokens.

As the general corpus used for comparative purposes, the British National Corpus (BNC) was chosen as a source of linguistic data. It is a large, balanced corpus (Baker, Hardie, and McEnery 2006: 18) available freely online and integrated with the Sketch Engine corpus linguistic tool. At this point, it may be worth noting how the corpus is made up. The BNC is composed of written texts (books, periodicals, and miscellaneous sources; in total 90% of the corpus) and spoken material (10%) and contains in total 112,338,376 lemmas (Sketch Engine BNC Corpus Description). Summing up, the corpus size, its balanced nature, availability, and functionality (integration with the Sketch Engine tool) were arguments supporting the decision to use it as a reference corpus. However, we need to admit that the BNC is no longer updated and the texts it contains cover the period from 1960 to 1993, with approximately 96% of the content from the period 1984–1993 (calculated based on the statistics available via Sketch Engine). That constitutes a drawback, especially, in view of the fact that the legal corpus used in the study comprises mostly more recent documents drafted mainly at the end of the 20th and in the first two decades of the 21st century. Having considered the advantages and disadvantages, a decision was made to use the BNC as a reference corpus in the absence of another available more relevant source of linguistic material. The two other large corpora, *i.e.* the COCA and EnTenTen were not chosen as the former comprises American English language material which makes it a less desirable candidate for our contrastive analysis due to two main reasons, *i.e.* a language variety which is more distant from the EU English variety than UK English and the fact that the US is not tied to the EU as the UK used to be; thus, we may expect some of the concepts to be absent from the American legal system. When it comes to EnTenTen, another Sketch Engine integrated corpus, it meets linguistic requirements for the corpus to be large, clean, and duplicate-free. It also has rich metadata and offers wide coverage. However, it comprises language material extracted only from websites from different English-speaking geographic areas, which again may not yield representative enough results.

2.2. Methodology

The two aims introduced above require the application of a mixed methodology, *i.e.* corpus linguistics quantitative methods for aim 1 and the mixed quantitative/qualitative methodology of corpus linguistics and discourse studies for aim 2[6].

Having chosen the corpora for analysis, we uploaded these documents to Sketch Engine to allow their investigation. Sketch Engine offers a range of sophisticated functionalities that are useful for retrieving collocations based on selected criteria, including the word sketch, *i.e.* a condensed description of a word's grammatical and collocational behaviour and the word sketch difference (Kilgarriff *et al.* 2014: 9). The minimum frequency threshold for retrieving word combinations and potentially identifying them as

6 A study of *sex* and *gender* collocational profiles in the EU equal opportunity corpus was the subject of investigation presented in Mroczyńska (2024). The methodologies applied in Mroczyńska (2024) and in the present study are very similar as research involves the same two lexical items, the difference being the two corpora used to compare LSP and LGP occurrences.

collocations was set at five occurrences, *i.e.* a collocation needs to occur at least five times to be included in the study. This was done to eliminate potentially invalid word combinations[7]. In the subsequent step, the results produced by Sketch Engine were subject to manual verification. In this step, candidate collocates suggested by Sketch Engine were removed from further analysis if they turned out not to act as modifiers in the corpus. The results were then sorted according to the grammatical pattern in which they appear. Moreover, sketch difference functionality, which compares the behaviour of two selected words or lemmas, broken by their collocational patterns, proved extremely helpful in this study as it allowed for the comparison of LSP and LGP uses. The results obtained were the point of departure for the analysis presented in Section 3 below.

The texts included in the corpus deal with a wide area of equal opportunities as presented in the EU legal and paralegal texts. The frequency list generated for nouns shows that *equality* ranks 6th, *woman* 13th, and *man* 37th, whereas *gender* and *sex* were placed in the 31st and 53rd positions, respectively, in terms of absolute frequency, *i.e.* an actual number of hits in the corpus. When it comes to the BNC, the frequency list results for the lexical items were as follows: *man* ranked 5th, *woman* 18th, *sex* was placed in the 596th position, as for *gender* and *equality* they were not included in the top 1000 items, and consequently they were not available in Sketch Engine version intended for academic use. The Sketch Engine wordlist functionality may provide frequency lists, absolute frequencies, and relative frequencies of lexical items. Absolute (raw) frequency is rarely a useful value, especially with periods or/and corpora of uneven sizes. When comparing corpora varying in size, relative frequency seems to be a more reliable and standardized measurement than absolute frequency as it refers to the number of occurrences of an item per million tokens (i.p.m., *i.e.* instances per million). Relevant statistics for the lexical items connected with equal opportunities are presented in the table below. As we can see, rather unsurprisingly, all of the words considered show higher frequencies, both raw and relative ones, in the specialized corpus than in the general one, which is the natural consequence of the area the specialized corpus deals with.

Table 1. Sketch Engine Wordlist analysis results for *equality*, *sex*, *gender*, *man* and *woman*

Lexical item	**LSP corpus**				**LGP (BNC) corpus**		
	rank	absolute frequency	relative frequency	rank	absolute frequency	relative frequency	relative frequency (1994)
equality	6	1,764	2,967.45	n/a	1,535	13.66	47.624
woman	13	1,394	2,345.03	18	58,431	520.13	10.583
gender	31	806	1,355.88	n/a	1,916	17.06	26.458
man	37	655	1,101.86	5	94,645	842.50	111.123
sex	53	500	841.12	596	8,834	78.64	21.166

The aim of the present study was to focus on an analysis of *sex* and *gender* acting as collocation nodes in our corpora using Sketch Engine. The software allows for the extraction and presentation of search results by different collocational patterns, *e.g.* (1) premodifier + noun, (2) noun + noun, (3) verb + noun, (4) noun + verb, (5) preposition + noun, (6) noun + preposition (see: Hausmann 1989).

7 That assumption goes in line with Evert's (2008: 1244) who recommends that a frequency threshold of ≥ 5 be applied so as to "weed out potentially spurious collocations".

The patterns above may be analyzed in pairs due to their structural similarity, for example in pattern (1) above a modifier may be an adjective, a noun or a participle, whereas pattern (2) allows for modifications using a noun only (Michta, Mroczyńska 2022: 40). The results obtained in the Sketch Engine search both in the LSP corpus and in the BNC showed that the largest number of collocates may be found with *sex* and *gender* acting as modifiers for nouns, *i.e.* as in (1) and (2) above. When it comes to the modifier of *sex* and *gender* pattern, the LSP search returned only 2 results while the LGP corpus revealed more possible collocates. The remaining patterns were only scarcely identified, with the software often yielding just a couple of possible word combinations. Therefore, patterns (1) and (2) were the first candidates for more in-depth research.

For the purpose of this comparative analysis, the Sketch Engine word sketch difference function was used as it allows us to juxtapose collocations of two selected lemmas/words. The list of possible collocates returned by the software shows that the collocability of both words rather does not overlap, i.e. *sex* will usually modify a different set of nouns than *gender* will. We also noticed that the lexical items as used in the LSP and the LGP corpora may display various meanings. Consequently, we will be dealing with polysemy and there may be different collocate candidates depending on the meaning the analyzed word carries. The analysis and its findings are presented in Section 3 of the present study.

3. Collocational Profiles of *Sex* and *Gender* in the Specialized and General Corpora

3.1. Different Meanings of *Sex* and *Gender* in the LSP Corpus Versus the LGP Corpus

While analyzing the combinatory potential of *sex* and *gender* as used in the EU documents regarding equal treatment of men and women in the earlier study, we focused only on their meanings referring to being male, female or neutral as these were the only meanings these terms adopted in the EU equal opportunities corpus (see: Mroczyńska 2024). The terms deal with the issue of being male or female from different angles – the biological or psychological and socio-cultural ones, respectively. Consequently, they represent different concepts and we do not expect them to act as synonyms and be used interchangeably. In the present comparative study of the terms in question as used in the specialized corpus and in the general one we will also attempt to investigate other senses of the lexemes. A good starting point may be a brief overview of definitions of *sex* and *gender* extracted from selected dictionaries and presented in Table 2.

Table 2. Selected dictionary definitions of *sex* and *gender*

Source	*Sex*	*Gender*
Oxford English Dictionary (online)[8]	1a. a1382 Either of the two main categories (male and female) into which humans and many other living things are divided on the basis of their reproductive functions; (hence) the members of these categories viewed as a group; the males or females of a particular species, esp. the human race, considered collectively. 2. Quality in respect of being male or female, or an instance of this; the state or fact of belonging to a particular sex; possession or membership of a sex. 2.a. c 1475 With regard to persons or animals. […] 4.a. a1631 The distinction between male and female, esp. in humans; this distinction as a social or cultural phenomenon, and its manifestations or consequences; (in later use esp.) relations and interactions between the sexes; sexual motives, instincts, desires, *etc.* 4.b. 1899 Physical contact between individuals involving sexual stimulation; sexual activity or behaviour, *spec.* sexual intercourse, copulation. to have sex (with): to engage in sexual intercourse (with). 5. 1664 *slang* or *euphemistic*. A person's genitals.	1.*Grammar*. 1.a.c1390 In some (esp. Indo-European) languages, as Latin, French, German, English, *etc.*: each of the classes (typically masculine, feminine, neuter, common) of nouns and pronouns distinguished by the different inflections which they have and which they require in words syntactically associated with them […] 3.a. (1474) *gen.* Males or females viewed as a group; = sex n.1 1. Also: the property or fact of belonging to one of these groups. 3.b. (1945) *Psychology* and *Sociology* (originally *U.S.*). The state of being male or female as expressed by social or cultural distinctions and differences, rather than biological ones; the collective attributes or traits associated with a particular sex, or determined as a result of one's sex. Also: a (male or female) group characterized in this way.

8 Due to the limitations of space, obsolete uses, phrases including head words or additional comments were left out.

Source	*Sex*	*Gender*
Cambridge Dictionary (online)[9]	1a the physical state of being either male, female, or intersex 1b all males considered as a group, or all females considered as a group 2 physical activity between people involving sexual organs	1a a group of people in society who share particular qualities or ways of behaving which that society associates with being male, female, or another identity 1b the condition of being a member of a group of people in a society who share particular qualities or ways of behaving which that society associates with being male, female, or another identity 1c used to refer to the condition of being physically male, female, or intersex (= having a body that has both male and female characteristics) 2 the grammatical arrangement of nouns, pronouns and adjectives into masculine, feminine, and neuter types in some languages
Collins English Dictionary online[10]	1 The two sexes are the two groups, male and female, into which people and animals are divided according to the function they have in producing young 2 The sex of a person or animal is their characteristic of being either male or female. 3. the physical activity between people that involves the sexual organs	1 Gender is the state of being male or female in relation to the social and cultural roles that are considered appropriate for men and women 2 You can use gender to refer to one of a range of identities that includes female, male, a combination of both, and neither 3 Some people refer to the fact that a person is male or female as his or her gender 4. Some people refer to all male people or all female people as a particular gender. 5. In grammar, the gender of a noun, pronoun, or adjective is whether it is masculine, feminine, or neuter. A word's gender can affect its form and behaviour. In English, only personal pronouns such as 'she', reflexive pronouns such as ‚itself', and possessive determiners such as ‚his' have gender.

9 Cambridge Dictionary online [at:] https://dictionary.cambridge.org/dictionary/english [date of access March 16, 2025].

10 Collins English Dictionary online [at:] https://www.collinsdictionary.com/dictionary/english [date of access March 16, 2025].

From the definitions provided above we may infer that the senses *sex* revolves around may be broken into two broad categories, *i.e.* the first one generally referring to the fact of being male or female (in a physical sense) and the other one regarding physical activity, whereas *gender* may convey the meaning referring to the state of being male or female (in relation to social and cultural roles) or the grammatical concept. Usually, the order in which dictionaries list the meanings is supposed to reflect the frequency of use of a given meaning, with the first one presented being the most frequent or the logic, *i.e.* the first one carrying a literal meaning, and other ones having a more metaphorical sense a word may convey. Unlike the Cambridge Dictionary and Collins Dictionary, The Oxford English Dictionary Online (The OED) presents senses a lexeme may have in chronological order as they entered the language. Thus, it offers valuable insight in terms of a diachronic perspective. The lexeme *sex* originally entered the language in the 14th century and meant "either of the two main categories (male or female) into which humans [...] are divided based on their reproductive functions" (OED online). About a hundred years later, it acquired the meaning of "quality in respect of being male or female [...]; the state or fact of belonging to a particular sex with regard to persons and animals". However, around the 1960s, this sense of *sex* started to be replaced by *gender* in situations when speaking about humans. Referencing *sex* to "physical contact between individuals involving sexual behaviour" (meaning 4b) dates back to 1899 and as the authors of OED argue, meaning 4b is now the most common general sense of the lexeme. As far as *gender* is concerned, it started as a word referring to a grammatical category (around 1390). Later, this sense was extended in general language to "males or females viewed as a group" (1474) and the specialist psychological and sociological sense we attribute to *gender* most commonly nowadays, *i.e.* "the state of being male or female as expressed by social and cultural distinctions [...]" was originally used in the US in 1945. Having considered sense 2 for *sex* and sense 3a for *gender*, respectively, it seems that for a period of time there was an overlap of senses that *sex* and *gender* carried.

Moving on to the present study of occurrences of *sex* and *gender* in the two corpora, lists of modifiers of *sex*/*gender* and nouns modified by *sex*/*gender* were generated using Sketch Engine to conduct the analysis. The first observation that can be made about the lists produced from each corpus is that the collocations from the specialist corpus include collocates where *sex* and *gender* refer only to the fact of being male and female, whereas in the BNC there are also collocates encompassing other senses of the two words in question. Actually, in both of the analyzed categories in the BNC, the majority of collocates of *sex* refer to physical activity, and not the fact of being male or female, which may suggest that this is the most frequent sense in which the lexeme is used. When it comes to *gender*, the LSP corpus provides more instances of collocates of *gender* than the BNC, with all of them falling only into one category, namely nouns modified by *gender*. The BNC also offers a small number of collocates of *gender* in its grammatical sense, whereas in the LSP corpus there are none. The statistics are presented in the table below.

Table 3. Collocates of *sex* and *gender* in the LSP and general corpora

Lexeme	Meaning	LSP corpus	BNC	LSP corpus	BNC
		modifiers of *sex/ gender*		nouns modified by *sex/gender*	
sex	1 (male/female)	2	13	2	19
	2 (physical activity)	-	28	-	49
gender	1 (male/female)	-	2	20	14
	2 (grammar)	-	3	-	2

The data presented in Table 3 above may suggest that the LSP corpus regarding equal treatment of men and women focuses on social and cultural differences rather than biological ones. Hence, we attest greater combinatory potential of *gender* than of *sex* in the specialized texts as compared to the general language, where the situation is the opposite. What is more, the specialized corpus comprises documents which put emphasis on human dignity, freedom, democracy, equality, the rule of law, and the respect for human rights. Consequently, we may assume that the legislator may not have intended to infer and regulate the intimate sphere of sexual life. As a result, *sex* in the sense of physical contact is not attested in this corpus, and the lexeme meaning a biological distinction is also hardly ever used.

3.2. Common Collocates for the Two Lexemes

For our analysis, we singled out only the occurrences of *sex* and *gender* in the sense referring to being male or female, as these are the senses that are attested both in the LSP and the LGP corpora. An observation that can be made here is that collocates of the two lexemes vary significantly as we can see in Table 4 below[11].

11 The section of the study concerning collocate candidate overlap and their ranks was inspired by research conducted by L'Homme and Azoulay (2020) regarding collocate candidates of selected items extracted from the specialized corpora regarding environment protection as compared to the general corpora. Also, an interesting contrastive study of collocational behaviour of the term *evidence* in legal and general corpora was conducted by Michta (2022).

Table 4. Collocates of *sex* and *gender* in the specialized and general corpora

Lexeme		LSP corpus	The BNC
sex	Modifiers of the lexeme	other	opposite
		same	same
			own
			female
			other
			male
			fair
			different
			separate
			single
			weak
			biological
			mixed
	Nouns modified by the lexeme	discrimination	difference
		characteristic	ratio
			discrimination
			hormone
			role
			cell
			determination
			organ
			distribution
			group
			chromosome
			equality
			segregation
			structure
			pheromone
			inequality
			composition
			rate
			bias

Lexeme		LSP corpus	The BNC
gender	Modifiers of the lexeme		different
			natural
	Nouns modified by the lexeme	equality	difference
		gap	relation
		identity	role
		directive	identity
		balance	bias
		reassignment	inequality
		stereotype	division
		inequality	issue
		strategy	politics
		surgery	stereotype
		mainstreaming	imbalance
		perspective	study
		bias	neutrality
		role	discrimination
		expression	segregation
		dimension	consciousness
		impact	line
			schema
			equality
			dimension
			order
			group

Collocates in the table above are presented by their (raw) frequency (starting from the most frequent ones), and the overlapping candidates are marked with greyish cell shading. As we can see, some candidates were suggested for both corpora; for instance *same* and *other* appear as modifiers of the lexeme *sex* in both lists. Actually, these are the only collocates found in the specialized corpora. The BNC, on the other hand, suggests a wider range of possible modifiers such as *opposite*, *female*, *male*, *fair*, or *different*. When it comes to nouns modified by *sex*, the specialized corpus offers just two candidates, i.e. *characteristics* and *discrimination*, whereas the general corpus lists more options, *e.g. determination*, *chromosome*, *distribution*, or *war*.

As far as *gender* and its collocates are concerned, we find no modifiers of *gender* in the specialized corpus while the BNC offers just two, namely *different* and *natural*. We attest many more nouns modified by *gender* in both corpora. Interestingly, the relatively small specialized corpus yields a large number of collocate candidates as compared to results extracted from the much larger general corpus, 17 in the specialized corpus and 22 in the BNC. On the one hand, that may be linked to the fact that EU institutions place gender and equality very high in their agenda, and hence the term *gender* is used extensively in official publications. On the other hand, the BNC contains less recent language material than the specialized corpus used in the present study. The collocations occurring in the small specialized corpus maybe the

reflection of the changing values and beliefs in societies, which in turn affect the language used. As Kjær (2007: 508) argues, the legal language and consequently its typical word combinations are inextricably intertwined with a particular legal system, in this case with the European Union legal regulation system common for all member states.

Table 5. Common collocates of *sex* and *gender* extracted from the BNC

Sex		**Frequency**	
		Gender	
Modifiers of the lexeme	different	14	8
Nouns modified with the lexeme	difference	164	104
	discrimination	61	8
	role	37	52
	group	19	5
	equality	13	5
	segregation	11	7
	inequality	6	33
	bias	5	38

Also, it may be worth noting that in the specialized corpus, there is a clear-cut division between collocates used with *sex* and *gender*, no candidates appear with both, which suggests that the lexemes are not used interchangeably and each of them carries its own different meaning. This division may reflect the legislator's intention to make a distinction between individual's biological role and the socio-cultural one. What is more, the results of the study indicate that in the LSP corpus *gender* has a much greater combinatory potential appearing in a wide range of collocations whereas *sex* appears only in two collocations, namely *sex characteristics* and *sex discrimination*, the latter actually being a well-established term featuring in most dictionaries[12]. Table 5 shows that in the general corpus, we attested a number of collocates which appear with both *sex* and *gender*. The fact that the two lexical items have common collocates may suggest that, in these instances, they act as synonyms.

When we look more closely at Table 4, we notice that the degree of overlap between collocates for *sex* and *gender* is rather low. As we already mentioned, there are two modifiers of *sex* found in both corpora, namely *same* and *other*, and only one noun modified by the lexical item (i.e. *discrimination*) is found in the two lists. As far as *gender* is concerned, we may investigate only nouns modified by *gender* as there are no modifiers of the term found in the specialized corpus. The candidates suggested for both corpora include *gender dimension*, *role*, *bias*, *inequality*, *stereotype*, *identity*, and *equality*. However, the BNC shows that some of the candidates appear interchangeably with *sex* and *gender*, with varying frequencies. *Sex* is attested more often than *gender* in such word combinations as *sex group*, *sex discrimination*, *sex*

12 That is in line with what some researchers point out, namely the fact that modifier + noun combinations may cover not only collocations but also terms. See among others Bergenholtz and Tarp (1994), Michta *et al.* (2009), L'Homme and Azoulay (2020). Distinguishing between collocations and terms may constitute an interesting line of research though it is not the main focus of this study.

difference, *sex equality*, and *sex segregation*, whereas *gender* features more often in combinations such as *gender role*, *gender inequality*, or *gender bias*.

Summing up, we attested 2 modifiers of *sex* in the specialized corpus and 13 in the general one, the overlap being 15% (2 items). We also found 2 nouns modified by *sex* in the LSP corpus and 19 in the BNC; here the overlap is just 5% (only 1 item). When it comes to *gender*, there are no modifiers of the lexeme in the specialized corpus and in the general one we attested two instances. The highest degree of overlap of 32% (7 items) may be found in the group of nouns modified by *gender*, with 17 candidates in the specialized corpus and 22 in the general one. That may lead us to the conclusion that a sizeable number of collocates are specific to each corpus and consequently, the collocational behaviour of the analyzed lexemes displays large differences.

3.3. Ranks of Collocates

As it was shown in Section 3.2, the lists of collocates extracted from the specialized corpus and from the general one do not tend to show a large degree of overlap. Let us take a closer look at how common collocates of *sex* and *gender*, respectively, rank in the two corpora. Candidate collocates presented in the table below were retrieved from the corpora for a window size 1. The order collocates are presented in Table 6 is based on the rank of collocates of the *gender* attested in the corpora.

Table 6. Ranks of common collocates of *sex* and *gender*

Lexeme	Collocate type	Collocate	Rank in the specialized corpus	Rank in the general corpus
sex	modifier	other	1	5
		same	2	2
	modified noun	discrimination	1	3
gender	modifier	equality	1	19
		identity	3	4
		stereotype	7	10
		inequality	8	6
		bias	13	5
		role	14	3
		dimension	16	20

As can be seen in Table 6, common collocates rarely appear at the same rank. When we look at collocates that were assigned high ranks (from 1 to 3), only *same* (*sex*) has an identical rank for the specialized and general corpora and *identity* ranks 3 in the specialized corpus and 4 in the general one. The top collocate in the specialized corpus, i.e. *equality*, appears at much lower ranks in the BNC (rank 19), s*tereotype* ranks 7 and 10, and *dimension* comes 16 and 20, respectively. There are also instances where high-ranking collocates in the BNC appear at lower ranks in the LSP corpus, *e.g. role* ranks 3 in the BNC and 14 in the specialized corpus, *bias* comes 5 and 13, respectively. This observation seems to indicate that there are large differences between corpora when it comes to collocational profiles of the analyzed

items. That may be connected with the fact that even when discussing the same issues different terms tend to be used in the specialized and general texts, which in turn is reflected in the corpora. Accordingly, the character of the corpus has an effect not only on the terms selected but also on the words they go with.

4. Conclusion

Data presented in Section 3 tend to confirm that we are more likely to witness polysemy of lexical items in the general corpus than in the specialist one, or looking from a different perspective, we may expect higher cohesiveness and less polysemy in the specialized corpus as one of the aims of using specialized terminology is to convey the intended meaning clearly. Our specialized corpus comprises legal and paralegal texts and as Jopek-Bosiacka claims in such texts authors should follow the "principles of semantic accuracy and language consistency," which are key to avoiding ambiguity and misunderstandings (Jopek-Bosiacka 2011: 16)[13]. That may account for the lack of polysemy of the analyzed terms in the specialized corpus, they are used in one sense only, and there is no overlap of collocate candidates of the two lexemes in question. Unlike the specialized corpus, the general one contains instances of the uses of *sex* and *gender* where the items carry meanings other than the one assigned in the LSP corpus. Actually, the meaning *sex* is assigned in the specialized corpus is secondary in the general one.

What is more, the study appears to confirm what L'Homme and Azoulay (2020: 162) advocate, *i.e.* the fact that "the nature of the corpus and the topics it addresses has consequences on the terms used and on their collocates". General corpora contain texts which deal with a wide range of topics whereas a specialized corpus focuses on one area only. Consequently, in the specialized corpus lexical items are likely to convey a single meaning, authors avoid polysemy, and lexical variety is rather limited. This in turn may lead to a low degree of overlap between collocate candidates extracted from the specialized and general corpora, as is the case of *sex* and *gender* analyzed in this study. In this specific case, the degree of overlap may also reveal some differences between standard English in the general corpus and the EU English in the specialized corpus. We would expect that collocations retrieved from the latter reflect the EU policies and regulations in the area and the terminology used may also be affected by the so-called Eurolect specific to communication in this institution.

Further differences in collocational behaviour of the terms in question are revealed by the ranks assigned to collocates in each corpus. Hardly ever were collocates assigned the same ranks, only a few ranked close and most of them appeared in different parts of lists extracted from each corpus.

All in all, the study provides further evidence for the claim "You shall know a term by the company it keeps" (Firth 1968: 179). While the two words in question may be thought of as sharing a common core meaning, our analysis has shown that these may be successfully teased apart by looking at their collocates. Investigating whether the findings reported in this study can also be extended to other lexical items might be an interesting line of further research.

We are aware that this study was carried out only for two selected lexical items, namely *sex* and *gender*. Still, we believe that it may provide some valuable insights into the workings of specialized and general corpora and reveal some differences that may be of importance for researchers dealing with the

13 It may be worth noting that using synonymy in legal contexts is rather unwelcome though not absent from legal texts. See Góźdź-Roszkowski (2013), Matulewska (2016), and Rzepkowska (2023).

area of terminology and lexicography. The findings may also prove useful in native and foreign language teaching and learning both in general and specialist language contexts as there is a body of literature showing that for both L1 and L2 speakers collocations pose a challenge[14]. Bearing in mind our research findings, it seems advisable that designers of teaching materials have access to appropriate authentic language material and corpora, know how to make the best use of them, and be aware of possible differences in collocational profiles of words as used in their general or specialized meanings so as to guarantee the highest standard of the content they produce.

References

Badziński, Arkadiusz (2019) "Problems in Medical Translation among Professional and Non-Professional Translators: Collocations as a Key Issue." [In:] *Beyond Philology* 16(4); 157–177.

Baker, Paul, Andrew Hardie, Tony McEnery (2006) *A Glossary of Corpus Linguistics* Edinburgh: Edinburgh University Press.

Benson, Morton (1989) "The Structure of the Collocational Dictionary." [In:] *International Journal of Lexicography* 2/1; 1–14.

Bergenholtz, Henning, Sven Tarp (1994) "Mehrworttermini und Kollokationen in Fachwörterbüchern." [In:] Burkhard Schaeder, Henning Bergenholtz (eds.) *Fachlexikographie: Fachwissen und seine Repräsentation in Wörterbüchern*. Tübingen: Gunter Narr Verlag; 385–419.

Biel, Łucja (2010) "Corpus-Based Studies of Legal Language for Translation Purposes: Methodological and Practical Potential." [In:] Carmen Heine, Jan Engberg (eds.) *Reconceptualizing LSP: Online Proceedings of the XVII European LSP Symposium 2009*. Aarhus: Arhus University Aarhus; 1–15.

Biel, Łucja (2015) "Phraseological Profiles of Legislative Genres: Complex Prepositions as a Special Case of Legal Phrasemes in EU Law and National Law." [In:] *Fachsprache* 3–4/2015; 139–160.

Biel, Łucja (2020) "Eurolects and EU Legal Translation." [In:] Ji Menj, Sara Laviosa (eds.) *The Oxford Handbook of Translation and Social Practices*. Oxford: Oxford University Press; 478–500.

Biel, Łucja, Agnieszka Biernacka, Anna Jopek-Bosiacka (2018) "Collocations of Terms in EU Competition Law: A Corpus Analysis of EU English Collocations." [In:] Silvia Marino, Łucja Biel, Martina Bajčić, Vilelmini Sosoni (eds.) *Language and Law*. Switzerland: Springer [https://doi.org/10.1007/978-3-319-90905-9_14 (date of access: March 16, 2025)].

Buzmaniuk, Stefanie (2023) "Gender Equality in Europe: A Still Imperfect Model in the World." [In:] *European Issues* No. 659. Foundation Robert Schuman Policy Paper; 1–7.

Council of Europe: European Court of Human Rights, Handbook on European non-discrimination law. 2018. ISBN 978-92-871-9851-8 [At:] https://rm.coe.int/fra-2018-handbook-non-discrimination-law-2018-en/1680a2b52b [date of access: August 1, 2024].

Cowie, Anthony P. (1994) "Phraseology." [In:] Ronald E. Asher, J.M.Y. Simpson (eds.) *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon Press; 3168–3169.

Danet, Brenda (1980) "Language in the Legal Process."[In:] *Law and Society Review* 14; 445–564.

Evert, Stefan (2008) "Corpora and Collocations." [In:] Anke Lüdeling, Merja Kytö (eds.) *Corpus Linguistics. An International Handbook*. Berlin: de Gruyter; 1212–1248.

14 See among others Benson (1989) and Frankenberg-Garcia (2018) for discussion of collocation competence in native and non-native speakers, Saber *et al.* (2020) (academic genre), Giczela-Pastwa (2021) (legal genre), Badziński (2019) (medical language).

Firth, John R. (1968) "A Synopsis of Linguistic Theory, 1930–1955." [In:] Frank R. Palmer (ed.) *Selected Papers of J. R. Firth 1952–1959*. Bloomington: Indiana University Press; 168–205.

Frankenberg-Garcia, Ana (2018) "Investigating the Collocations Avail-Able to EAP Writers." [In:] *Journal of English for Academic Purposes* 35; 93–104.

Giczela-Pastwa, Justyna (2021) "Developing Phraseological Competence in L2 Legal Translator Trainees: A Proposal of a Data mining Technique Applied in Translation from an LLD into ELF." [In:] *The Interpreter and Translator Trainer* 15/2; 187–204.

González-Ray, Isabel (2002) *La phraséologie du français*. Toulouse: Presses Universitaires du Mirail.

Goźdź-Roszkowski, Stanisław (2013) "Exploring Near-Synonymous Terms in Legal Language. A Corpus-Based, Phraseological Perspective." [In:] *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 12; 94–109.

Hausmann, Franz (1989) "Le dictionnaire de collocations." [In:] Franz Joseph Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (eds.) *Wörterbücher: ein internationales Handbuch zur Lexicographie*. Berlin: De Gruyter; 1010–1019.

Hausmann, Franz (1997) "Tout est idiomatique dans les langues."[In:] Michel Martins-Baltar (ed.) *La locution entre langue et usage*s. Fontanay/Saint-Claud: ENS Éditions; 27–290.

Hoey, Michael (2005) *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, Susan (2008) "Collocation Strategies and Design Decisions." [In:] Anke Lüdeling, Merja Kytö (eds.) *Corpus Linguistics. An International Handbook*. Berlin: de Gruyter; 154–168.

Jopek-Bosiacka, Anna (2011) "Defining Law Terms: A Cross Cultural Perspective." [In:] *Research in Language. Special Issue on Legal Terminology: Approaches and Applications*; 9–29.

Kilgarriff, Adam (2005) "Language is Never, Ever, Ever, Random." [In:] *Corpus Linguistics and Linguistic Theory*, Vol. 1 No.2; 263–276.

Kilgarriff, Adam, Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel, Suchomel Vít (2014) "The Sketch Engine: Ten Years On." [In:] *Lexicography* 1; 7–36.

Kjær, Anne Lise (2007) "Phrasemes in Legal Texts." [In:] Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn, Neal R. Noerrick (eds.) *Phraseologie/Phraseology: Ein internationales Handbuch zeitgenössischer Forschung / An international handbook of contemporary research*. Walter de Gruyter; 506–516.

Kjellmer, Göran (1994) *A Dictionary of English Collocations*. Oxford: Clarendon Press.

Koester, Almut (2010) "Building small Specialised Corpora." [In:] Anne O'Keeffe, Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics.* London: Routledge; 66–79.

Lehecka, Tomas (2015) "Collocation and Colligation." [In:] Jan-Ola Östman, Jef Verschueren (eds.) *Handbook of Pragmatics*. Amsterdam: John Benjamins Publishing; 1–20.

L'Homme, Marie-Claude, Daphnée Azoulay (2020) "Collecting Collocations from General and Specialised Corpora: A Comparative Analysis." [In:] Gloria Corpas Pastor, Jean-Paul Colson (eds.) *Computational Phraseology*. Amsterdam–Philadelphia: John Benjamins Publishing Company; 151–175.

Maley, Yon (1987) "The Language of Legislation." [In:] *Language and Society* 16; 25–48.

Matulewska, Aleksandra (2016) "Semantic Relations between Legal Terms. A Case Study of the Intralingual Relations of Synonymy." [In:] *Studies in Logic, Grammar and Rhetoric* 45(1); 16–174.

Mel'čuk, Igor (1998) "Collocations and Lexical Functions." [In:] Anthony P. Cowie (ed.) *Phraseology, Theory, Analysis and Application*. Oxford: Clarendon Press; 23–53.

Melinkoff, David (1963) *The Language of the Law*. Boston, Massachusetts: Little Brown.

Michta, Tomasz, Maria Kloza, Jan Łompieś, Mariusz Mela, Wioletta Mela, Magdalena Miąc, Joanna Newska, Liliana Religa (2009) "Studencki Słownik Kolokacji Angielskiego Języka Medycyny." [In:] Marek Łukasik (ed.) *Debiuty Naukowe III*. Warszawa: Katedra Języków Specjalistycznych UW; 89–225.

Michta, Tomasz (2022) "You Shall Know a Term by the Company it Keeps: Collocations of the Term Evidence in General and Legal Corpora." [In:] *Beyond Philology* No. 19/1; 65–96.

Michta, Tomasz, Katarzyna Mroczyńska (2022) *Towards a Dictionary of Legal English Collocations*. Siedlce: Wydawnictwo Naukowe Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach.

Mroczyńska, Katarzyna (2024) "Do Sex and Gender Go Hand in Hand? A Study of their Collocational Profiles in EU Documents Regarding Equal Treatment of Men and Women." [In:] *Crossroads. A Journal of English Studies*, (45), 82–103. doi: 10.15290/CR.2024.45.2.05.

Northcott, Jill (2012) "Legal English." [In:] Brian Paltridge, Sue Starfield (eds.) *The Handbook of English for Specific Purposes*. Malden, Massachusetts: Wiley-Blackwell; 213–226.

Patiño, Pedro (2014) "Towards a Definition of Specialized Collocation." [In:] Gabriel Quiroz, Pedro Patiño (eds) *LSP in Colombia: advances and challenges*. Bern: Peter Lang; 119–133.

Rzepkowska, Agnieszka (2023) "The Collocational Profile of Employment and Work in UK Employment Law." [In:] *Conversatoria Linguistica* XV; 67–87.

Saber, Anthony, Audrey Cartron, Claire Kloppmann-Lambert, Céline Louis (2020) "Towards a Typology of Linguistic and Stylistic Errors in Scientific Abstracts Written by Low-Proficiency Geoscience and Mechanical Engineering Doctoral Students in France." [In:] *Fachsprache* 42/3-4; 90–114.

Sergejeff, Katja, Mariella Di Ciommo (2023) "Gender Equality in EU External Action: Mainstreaming Women's Economic Empowerment." [In:] *Briefing Note* No. 163. The Centre for Africa-Europe Relations. [At:] https://ecdpm.org/application/files/2516/8312/5788/Gender-equality-in-EU-external-action-mainstreaming-womens-economic-empowerment-ECDPM-Briefing-Note-163-2023.pdf [date of access August 7, 2024].

Sinclair, John (2004) *Trust the Text. Language, Corpus and Discourse*. New York/London: Routledge.

Siepmann, Dirk (2005) "Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects." [In:] *International Journal of Lexicography* Vol. 18(4); 409–443.

Siepmann, Dirk (2006) "Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographic Aspects." [In:] *International Journal of Lexicography* Vol. 19(1); 1–39.

Stubbs, Michael (2004) "Language Corpora." [In:] Alan Davies, Catherine Elder (eds.) *Handbook of Applied Linguistics*. Oxford Blackwell; 106–132.

Internet sources

Oxford English Dictionary (2022) online [at:] https://oed.com [date of access August 26, 2024].

Cambridge Dictionary online [at:] https://dictionary.cambridge.org/dictionary/english [date od access January 6, 2024].

Collins English Dictionary online [at:] https://www.collinsdictionary.com/dictionary/english [date of access January 6, 2024].

Received: 25.09.2024
Reviewed: 02.02.2025
Accepted: 27.02.2025