Int J Semiot Law
https://doi.org/10.1007/s11196-024-10132-4

Strategic Indeterminacy and Online Privacy Policies: (Un)informed Consent and the General Data Protection Regulation

Daniel Green[1]

Accepted: 22 February 2024

Abstract
Article 12 of the General Data Protection Regulation (GDPR) requires data controllers to provide data subjects with any information relating to data processing operations "in a concise, transparent, intelligible and easily accessible form, using clear and plain language." Linguistic inclusivity of privacy policies is no longer a matter of style, but has been a binding legal requirement under the new data protection framework. Article 5 GDPR sets forth the requirements of lawfulness, fairness and transparency and prohibits any data processing operations which do not meet the standards of specification, explicitness and legitimacy of processing purposes [29]. In this study, a quantitative and qualitative analysis of linguistic indeterminacy in a corpus of 350 online privacy policies is presented and it is argued that a considerable number of data controllers continue to make use of strategic vagueness in the context of purpose limitation, therefore potentially prejudicing compliance with the GDPR. A legal-linguistic perspective on the current challenges of informed consent in European data protection law is provided. Finally, it is concluded that while the GDPR has contributed significantly to the linguistic empowerment of the data subject, the framework fails to satisfy the expectation of creating a participatory culture (see [36]) with a high degree of informational self-determination.

Keywords Legal linguistics · Privacy policy · General data protection regulation (GDPR) · Informed consent · Information requirement · Indeterminacy · Transparency

✉ Daniel Green
daniel.green@wu.ac.at

[1] Vienna University of Economics and Business, Vienna, Austria

Published online: 26 April 2024

1 Introduction

Article 12 of the General Data Protection Regulation (GDPR) imposes on data controllers the duty to provide any information relating to data processing operations "in a concise, transparent, intelligible and easily accessible form, using clear and plain language" [29]. The genre of privacy policies is at the centre of this information requirement, since its underlying communicative purpose is to ensure that any consent obtained from a data subject is informed and consequently valid. The notion of informed consent has been subject to much controversy in the context of the "datafication of everything" [42:146] as privacy policies have been described as being embedded in "non-informed consent cultures" (see [6: 21–38]). In this paper, a usage-based investigation of 350 online privacy policies will be presented with a view to relating linguistic findings to fundamental questions of data protection law generally, and policy drafting specifically. This explorative paper seeks to address the following research questions:

(1) How is indeterminacy utilised in policy drafting, which functions do adjectives fulfil in this context, and how can this be accounted for?
(2) Which legal challenges arise as a consequence of indeterminacy, what impact does this have on informed consent, and is the information requirement of the GDPR sufficient?

Section 1 discusses the notions of "transparency by design," (see [35:2–8]) legal literacy and informed consent. In Sect. 2, privacy policies are approached in terms of genre understood as conventionalisation, their underlying communicative purpose and the construct of their intended audience. Section 3 raises the issue of indeterminacy in privacy policies, and how it relates to the ambiguity of (un)informed consent, arguing that theoretical distinctions may be made between legal and linguistic indeterminacy. Section 4 introduces corpus linguistics as a method of quantitative legal analysis, which, despite its limitations, may offer solutions to "real-world problems in which language is a central issue" [10:27]. In Sect. 5, the method, data, analysis and results of the study are presented. Informed by these empirical findings, Sect. 6 is concerned with the questions as to why a void for vagueness doctrine would lead to an empowerment of the data subject, and how privacy education in schools can foster a participatory privacy culture.

2 Transparency by Design, the GDPR and the Notion of Legal Literacy

2.1 Fairness and Transparency

Article 5 of the GDPR provides that any processing of personal data must be fair, lawful and transparent, [29] an obligation that has been specified by the information

requirement in the current European data protection framework. Fairness and transparency are also important parameters in contract formation (see [69.68–88]) and the context of consumer protection (see [38]). The notions of fairness and transparency have gained significant importance, since data controllers can no longer disregard data subjects rights when collecting, managing or transferring personal data. As found in Article 25 of the GDPR, data protection by design and by default should ensure that the data protection principles are implemented efficiently and complied with consistently at all stages of data handling [29]. In this context, "transparency-by-design" may be understood as "a situation in which the requirements on transparency are satisfied by the very nature of the design and that the outcomes of the design process meet these requirements" [35:4]. While the term *transparency* may give rise to numerous diverging interpretations, language and communication play a fundamental role in creating a balanced relationship between data controllers and data subjects. This assumption is rooted in the conviction that in most cases it is through language that individuals are informed about the conditions of data handling practices. Notably, the Information Commissioners Office (ICO) has continuously placed great emphasis on the relationship between the language use of the data controller, and the (un)informed consent of the data subject. This is evident in the following questions raised by the ICO relating to the assessment of (un)fairness:

1. Was the person supplying the data under the impression that it would be kept confidential?
2. Was any unfair pressure used to obtain the information?
3. Was the person improperly led to believe that they must supply the information, or that failure to provide it might disadvantage them? [34]

Paradoxically, while privacy policies are lasting manifestations of unilateral communication (see [3]), they tend to be ignored by data subjects due to their overwhelming length, and the convoluted syntax in which some of them are written. This raises the question as to whether the presence or absence of fair processing may, inter alia, depend on both binding linguistic requirements and the overall legal literacy levels amongst data subjects.

2.2 Legal Literacy, Legal Awareness and Data Protection Law

Even though the term *legal literacy* is commonplace in the literature, there is no consensus regarding the meaning or use of the term. Narrow approaches to legal literacy conceive it as linguistic competence, that is, the ability "to read and write legal arguments, judgements and legislations that are part of the body of law" [55: 1655]. However, this understanding of legal literacy has received much criticism in scholarship, and has resulted in a broader denotation of the term that is not exclusively limited to the written language of legal practice. For instance, legal literacy has been described as a certain "degree of competence in legal discourse required for meaningful and active life in our increasingly legalistic and litigious culture" [55:1655]. This is in line with the definition suggested by the Canadian Bar

Association, which describes legal literacy as "the ability to understand words used in a legal context, to draw conclusions from them, and to use those conclusions to take actions" ([12] cited in [55:1655]). Legal literacy is commonly associated with legal knowledge encoded in written language, but as shown by Walser and Crespo, [62:9–11] there are also other mediums which can be used to convey legal knowledge, such as illustration. However, especially in the context of privacy information, there appears to be no consensus as to how exactly legally-relevant content should be presented by visualisation (Rechtsvisualisierung). For the purposes of this investigation into privacy related information, the latter understanding of legal literacy is extended by the factors of legal awareness and legal participation. It is assumed that the degree of balance between data controllers and data subjects is significantly co-determined by two main factors; first of all, the extent to which data subjects are aware of the divergence between ordinary language use and legal language use and, secondly, the discrepancies in contract interpretation potentially arising in litigation. Adapting Jenkins concept of participation, one may assume a participatory privacy culture to be an information society in which data subjects "believe that their contributions matter," [36:6] and where they assume the role of an agent rather than that of a merely passive recipient. The information requirement of the GDPR and the provisions relating to informed consent constitute a significant contribution to the empowerment of the data subject (see [1]).

2.3 The GDPR and the Illusion of Informed Consent

Article 4 of the GDPR defines consent as "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing" [29]. In addition to this, Article 7 of the GDPR sets out that where processing operations are based on data subjects consent, the data controller should be able to demonstrate such consent was obtained [29]. Recital 32 also stresses the nature of consent as a "clear affirmative act establishing a freely given, specific, informed and unambiguous indication of […] agreement." [29] It could be argued that obtaining consent is a process rather than a short-lived act, since it is surrounded by the boundaries of a pre-consensual and a post-consensual phase between data controllers and data subjects. In contract law, the offer and acceptance rule has a long-standing tradition whereby the offer made by one party must be accepted by the other in order for a binding agreement to be formed (see [73]).

In this context, autonomy-based theories naturally endorse the idea that the two parties are equals, in that they can construe for themselves which aspects of an agreement they consider (dis)advantageous, and whether they want to be legally bound [20:111]. Notwithstanding the need for such unconditional equality to exist in contract formation, autonomy-based accounts must be relativised in a data protection context. This is because they do not sufficiently capture the intrinsic imbalance between data controller and data subjects. This power asymmetry is notably present in privacy-related instructions. Informed consent as the complete disclosure to, subsequent understanding by and permission from the data subject is likely to be an

illusion. While the requirements under the GDPR constitute a framework to streamline how valid consent ought to be obtained, the participatory challenges with regard to written notifications, such as online privacy policies, remain unsolved. In the following section, the role of such informative instructions will be discussed in relation to contemporary genre conventions, their inherent communicative purpose, and the heterogeneity of the audience(s) for which they are intended.

3 Privacy Policies: Genre, Communicative Purpose and Audience

Privacy policies (or privacy notices), are the most frequently used text type to communicate to customers the conditions of data processing operations, i.e. when, how and for which purpose(s) personal data is collected, used and, if applicable, shared with a third party [48:221]. Within consumer protection law, it is commonly understood that personal information is "a key element of the exchange process between consumers and online merchants" and this makes the notion of information ownership a crucial issue [48:221]. In the remainder of this section, the role of privacy policies in the information society will be analysed with regard to three guiding questions:

(1) Are there general genre conventions in privacy policy drafting?
(2) What is the communicative purpose of privacy policies?
(3) Who is/are the (real) audience(s) of privacy policies?

3.1 Genre

Though a large number of online texts refer to themselves as privacy policy, the term itself remains opaque in the light of genre theory. It seems that the specificity of legal genres is the cause of many disagreements about the characteristics of such texts and how exactly these are defined. After all, to categorise privacy policies as a specific genre of text means to understand this genre as "an abstract conception rather than something that exists empirically in the world" [13:1]. Such abstraction is an integral part of legal thought and categorisation, and the fact that privacy policies may differ considerably from one another is not a sufficient argument for the non-existence of this particular genre. There may be a problem of universals as to the shared properties of these texts; however it is important to stress that the text type of privacy policies "as a whole, is conditioned by external considerations" [45:28], e.g. it functions as a shield in litigation (see [37]). Therefore, any text may constitute a privacy policy, provided that it contains more or less specific information on data collection, the purposes of data processing, the sharing of data, and that it is likely to employ linguistic strategies in order to prevent privacy litigation (see [48]). While frequently featuring keywords, such as *personal data* (595 times), *information* (169 times) and *communication* (29 times), the GDPR does not contain the terms privacy policy or privacy notice, or a specific definition of binding structural or linguistic criteria that such texts must fulfil. Article 13 of the GDPR,

however, provides a more specific taxonomy of the controllers duties under the information requirement that could potentially instigate international standardisation processes in policy drafting by means of normalisation. Notwithstanding the normative impact of the GDPR as far as policy drafting is concerned, the genre of privacy policies itself is also largely influenced by the numerous opportunities for computational representation and mediation. This may be termed the digital folding of privacy (see [7]). The representation and mediation of legally relevant information is no longer bound solely to the plasticity of linguistic structures, but also to their arrangement and accessibility in the digital space. Multi-layered privacy policies, for instance, exemplify how the digital space transforms conventional modes of communication. Nevertheless, the notion of genre is inextricably linked to the expectations, or socio-cultural pretext, that both policy drafters and policy readers hold towards the text type and its interpretation. Essentially, privacy policies and any other privacy-related text type do not constitute normative sources of law, such as contracts, but constitute accompanying informative instructions that set out the context of data handling. At this stage of the enquiry, the communicative purpose(s) which privacy policies serve as a distinct genre of instructive texts should also be considered.

3.2 Communicative Purpose as Discourse in Privacy Policies

A privacy policy is written legal text that is intended to last [3:125], as it should provide accountable evidence for compliance with the information requirement, particularly with Articles 12 and 13 of the GDPR. At this point, a distinction ought to be made between the notions of text and discourse, as they are used to describe language interpretation in this paper. Text refers to the formal structures of language that triggers "the recall of some familiar state of affairs" in a language users mind [67:6]. In contrast, the term discourse signifies the actual communicative purposes underlying a certain text and its production [67:6]. Widdowsons understanding of discourse assumes that it is not texts which carry meaning, but that individuals adopt an agentive role in which they interpret texts "as a discourse that makes sense to them" [67:6]. Texts themselves do not construe or contain context, but they merely underlie the activation of such contextual meaning in the mind of the reader [67:23]. Thus, the primary or intended discourse of a given privacy policy may diverge considerably from the secondary discourse derived by a data subject at a later point in time. This is because privacy policies are legal documents that primarily record the intended discourse and legal interests of data controllers "who can only account for second-person reaction by proxy" [68:11]. The primary discourse or intended communicative purpose, it seems, is not necessarily the provision of abundant information to the data subject, but rather the construction of a linguistic wall that effectively prevents privacy litigation. Indeterminacy, that is, the inassignability of a truth value to the relationship between a linguistic sign and a referent, is likely to be utilised by data controllers. A probabilistic legal realist argument [17:143] stands to reason according to which data controllers utilise strategic indeterminacy in order to balance out the obligations imposed by the information requirement and to obtain the

highest possible degree of litigation prevention (privacy policies as a legal shield). The use of such strategic indeterminacy does not necessarily require an intention to present untrue or misleading privacy-related information on the part of the data controller. However, it is arguably sufficient that the controller accepts the potential consequences arising due to the use of such indeterminacy.

3.3 The Audience(s) of Privacy Policies: Between Scylla and Charybdis

Article 12 of the GDPR provides that any information and communication relating to the processing of personal data must adhere to the linguistic requirements of concision, transparency, intelligibility and accessibility [29]. The framework also specifies that the language used in such communication must be clear and plain, "in particular for any information addressed specifically to a child." [29] Importantly, the legislative intent underlying the GDPR appears to be the removal, or at least levelling, of power asymmetries between data controllers and the specific audience targeted. The presence of linguistically induced power asymmetries is notably evident in data collection by means of implicit consent, i.e. where individuals are assumed to agree to an online privacy policy when visiting a website before having had the opportunity to read and understand the agreement [48:222]. In contrast, explicit informed consent, as prescribed by the GDPR, seemingly alleviates unethical practices by placing more control in the hands of the data subject. The audiences of privacy policies are not homogenous; rather they differ in age, education, legal literacy and awareness, privacy attitude, value orientation and world knowledge, yet the conventionalised language use encountered in privacy policies seems to suggest otherwise. While overt falsification is diametrical to the ethics of business communication, data controllers may utilise indeterminacy to induce the inference of "unstated meaning, beyond that derivable from the literal content explicitly stated message" [49:149]. To this end, data controllers must necessarily balance direct, lucid communication and face-saving communicative devices that comply with interpersonal norms of politeness in communication [49:169]. The space between privacy-related precisification and deprecisification is thus likely left negotiable for legal manoeuvres, and to open the privacy policy for legal assessment [26:159]. The reality of unknown heterogenous audiences requires, or even forces upon data controllers, the choice "between Scylla and Charybdis" (see [41]), between explicitness and trust-winning politeness regarding privacy information, such as purpose specification and purpose limitation. It is conventionalisation that defines what the text is, what the text means and for whom it is written. Following Barthes, the meaning and interpretation of privacy policies is embedded in a multidimensional space "in which are wedded and contested various kinds of writing […] the text is a tissue of citations" [5:4]. The construction of a privacy policy is accompanied by the immediate death of its code, which is subsequently re-animated by the individual seeking to engage in a meaning-making process (see [68]).

4 The ambiguity of (Un)informed Consent: Legal and Linguistic Indeterminacy

4.1 Linguistic Indeterminacy and Privacy Policies

In the previous section, obtaining individuals informed consent was described as a process rather than a short-lived act. In this section, the ambiguity residing in the concept of informed consent will be argued with reference to the process–product dichotomy [66:224] and the distinction between legal and linguistic indeterminacy [8]. For instance, informed consent relating to data processing purposes may be conceptualised as both the process of language interpretation and the overt result of meaning-making [66:224]. Notably, the ambiguity between informed and uninformed consent may be explained by virtue of the interplay of meanings involved in the interpretation process, namely "meaning in the world", "meaning in the mind" (see [18:1–14]) and meaning in context see ([30:135–145]). First, the code or scaffolding of privacy policies may be meaningful in relation to the mental link between language and the readers knowledge of the world, or data handling processes specifically (meaning in the world). Secondly, the code may allow for the recognition and assignment of mental representations in the readers mind towards abstract and non-abstract referents, such as automated decision-making (meaning in the mind). Thirdly, since context-dependence is an undeniable feature of any written text, some of the meaning of privacy policies may be generated in relation to "the use to which [language] is put" (meaning in context) [18:41]. It is therefore reasonable to conceive the linguistic indeterminacy of privacy policies as a multi-componential phenomenon of truth value assignment, which arises due to the ubiquity of blurry boundaries in reference, cognition and context. While it is reasonable to assume that there is a potential for linguistic indeterminacy to occur on multiple levels of written and spoken legal discourse, it is a logical fallacy to equate legal and linguistic indeterminacy.

4.2 Legal Indeterminacy and Privacy Policies

Linguistic indeterminacy often induces legal indeterminacy in privacy policies, but the occurrence of the former does not per se justify the latter [23:9]. The distinction to be made between the concepts is, inter alia, related to the parameters of code and context. Legal indeterminacy manifests itself either "on the face of the instrument," [50:140] that is as a co-product of vagueness, indefiniteness and ambiguity, or it arises due to an uncertainty in relation to extralinguistic aspects (context). To distinguish between the linguistic code of privacy policies and their context is pivotal. While the semantic meaning or code of privacy policies relates to the shared linguistic knowledge of a speech community, the context of a certain privacy policy "is not what is perceived in a particular situation, but what is conceived as relevant" [67:19]. The following shows two examples of legal indeterminacy that is not caused by the linguistic code itself, but what is regarded as relevant by the data subject.

(1) "we do not make your …email addresses available to third parties (except for subsidiaries, subcontractors or agents acting on our behalf in compliance with this Privacy Policy)" [48:228].
(2) "We reserve the right to disclose information […] as necessary or appropriate, in our view, to operate the Services, process orders or registrations" [74]

(1) and (2) lay out the conditions for disclosure of identifiable information to third parties, and contain different forms or degrees of legal indeterminacy. In (1), the data subject is assured that personal information will not be disclosed to third parties except for unidentified subsidiaries, subcontractors or agents. In (2), information may be disclosed if this is deemed necessary or appropriate by the data controller. It is not defined or exemplified under which circumstances such disclosure will take place. Therefore, the consent obtained from a reasonable data subject is unlikely to be informed in these examples due to the lack of contextual information that causes legal indeterminacy. In (2), it becomes clear that adjectives may function as powerful deprecisification tools in privacy policies, and that they can seriously prejudice informed consent. The collection of qualitative examples as seen above may be a fruitful starting point as such data relates to actual language use in the information society. However, it cannot provide insights on general tendencies in privacy policy drafting. The following section will argue why and how "computer-assisted legal linguistics" (see [61]) may prove a useful approach to gain comparable and verifiable insights into privacy policy drafting that goes beyond a mere introspective inquiry.

5 Corpus Linguistics as a Method in Policy Drafting: Chances and Limitations

While the methods and procedures associated with the area of corpus linguistics "are still developing, and remain an unclearly delineated set", [44:1] the quantitative investigation of legal texts, such as contracts, wills, terms of service and privacy policies may provide comparable and verifiable data on authentic legal language use. In contrast to qualitative introspective research, legal corpus linguistics is not concerned with individual linguistic choices, but instead investigates how often a certain linguistic feature occurs and whether any statistically significant tendencies can be identified. Legal corpus linguistics has great potential to function as a useful tool in privacy policy drafting, but it is also pivotal to stress potential limitations of this method, particularly those relating to the commensurability and generalisability of findings [60:292].

5.1 Online Privacy Policies and the Sample Corpus Approach (SCA)

In Sect. 2, the communicative purpose of privacy policies was related to the aim of providing accountable evidence for compliance with the GDPR, and the information requirement specifically. Any quantitative inquiry into the language use of

such legal documents is conducted based on the assumption that a collection of legal texts, such as privacy policies, represents a "particular type of language over a specific span of time" [44:8]. Importantly, the degree of balance and representativeness within a certain sampling frame is essential in order to provide a diagnostic description of the characteristics of a certain population [44:8]. A snapshot corpus, such as the one presented in this paper, provides authentic language data from a specific period of time and allows for the study of diachronic change in legal language use. The compilation and analysis of snapshot corpora may function as a valuable point of reference in academic and non-academic contexts. The idea of utilising such corpora in relation to the identification of common genre conventions and audience-oriented drafting for legal professionals is inviting. However, a snapshot corpus, like any other language collection, does not capture the language of privacy-related information per se, but provides a time-dependent and context-dependent window into such language use and its related characteristics, e.g. the occurrence of linguistic indeterminacy in and around the use of adjectives. A sample corpus approach may be used to identify tendencies or probabilities for certain linguistic units to occur, but it does not generally allow for underlying intentions, or the primary discourse at the creation of a text.

5.2 The Quantifiability of Imprecision

Indeterminacy or imprecision is a ubiquitous characteristic of human language and, paradoxically, it is often a necessary catalyst for efficient and successful legal communication (see [23:27–48]). After all, if every element in such an exchange received full specification, any communicative process would turn out to be rather cumbersome [27:25] or even break down entirely. In the context of privacy-related texts, "overspecification" (see [2:555–574]) of information may in fact be detrimental to effort to comply with the GDPR, since privacy policies already contain an overwhelming amount of information and yet lack the requirements of clarity and concision. It is an established insight in legal linguistics that certain constructions, such as adjectival phrases, have particular precisification qualities that regulate the space between specification or deprecisification within a legal document (see [26]).

While the phenomenon of imprecision itself cannot be quantified by means of corpus analysis, the use of this method provides valuable linguistic data on the frequency and distribution of imprecise or indeterminate expressions in privacy policies. As previously established, the communicative choices made by policy drafters are often the result of balancing explicitness and trust-winning politeness (see [16]). In a second step, it is reasonable to assume that in the drafting of privacy policies a precisification interdiction is operative according to which in certain contexts "the line [of precisification] is not to be drawn" [71:330]. The use of deprecisification tools, such as indeterminate adjectival phrases, is thus likely to constitute a communicative strategy in order to prevent privacy litigation. In other words, a probabilistic argument can be made that some data controllers make use of strategic indeterminacy in privacy policies, and that language corpora can provide statistically verifiable and comparable data on the linguistic realisation of the information requirement.

Finally, such language collections "do not only allow for verification of current analyses" but "will in time, provide answers to as yet unknown research questions" [57:2] in other areas, such as legal theory, legal sociology as well as linguistics. In the following section, the privacy policy corpus compiled for the purposes of this project will be introduced.

6 Empirical Study: A Corpus-based Investigation of Privacy Policies

6.1 Method

In this section, the procedure of selection, annotation and analysis of the Privacy Policy Corpus will be presented. For the purposes of this study, online privacy policies were selected from businesses of varying sizes within and outside of the European Economic Area (EEA). The three selection criteria were "relevance-based website pre-selection", [70:1332] "section-based sub-sampling", [70:1332] and randomisation of the online privacy policies origin. All privacy policies were collected between 1st and 2nd June 2018, and subsequently, the linguistic data obtained were used to compile a specialised language corpus. The dataset may therefore be described as an authentic snapshot of international legal language use in the period of GDPR enforcement in Europe. This may be particularly valuable as it offers insights into the multiple legal and linguistic facets of privacy policy drafting at the time when GDPR enforcement began.

The linguistic resources were annotated by TreeTagger, the multilingual parts-of-speech tagger developed by Schmidt, (see [51]) and provided online by the University of Lancaster [59]. The language corpus was then quantitatively and qualitatively analysed using the analysis toolkits LancsBox (v3.02) and Anthony's Antconc, which allow for descriptive as well as comparative analysis of naturally occurring linguistic data. The corpus-based investigation sought to detect re-occurring patterns of linguistic indeterminacy. Particular focus was placed on the tension between the precisification and deprecisification qualities of indeterminate adjectives in the privacy policies in the context of purpose limitation and purpose specification. A corpus-driven approach was chosen so as to allow for the emergence and description of regularities and irregularities from the data rather than to impose a previously formed legal evaluation on the part of the observer [9:196].

6.2 Data

The corpus compiled for the purposes of this study consists of 350 English online privacy policies from the following 35 countries: the United States, the United Kingdom, Germany, France, Austria, Switzerland, Sweden, the Netherlands, Canada, Finland, Ireland, China, Malta, Russia, Spain, Italy, Israel, Luxembourg, Japan, Denmark, Malaysia, Slovakia, Belgium, Taiwan, Australia, Lithuania, Poland, Estonia, Latvia, South Korea, New Zealand, Lichtenstein, Nigeria, India and the

Czech Republic. Figure 1 shows the distribution of privacy policies selected across countries:

In total, the language data consists of 1,158,180 tokens and 20,128 types. The notions tokens and types are used in the sense of McEnery and Hardie, who define the former as "any instance of a particular wordform" [44:50] and the latter as "a particular, unique wordform" [4442:50]. The policies included in the corpus diverge in length, structure and lexical density. While the shortest policy displays a minimum of 346 tokens, the maximum length measured is 16,524 tokens. The length of an average policy is 3309 tokens, with a median of 2998 tokens. The absolute frequency of tokens across all policies shows a standard deviation of 2059.23. Figure 2 shows the frequency intervals of all attested tokens across the corpus:

The policy with the lowest value shows a minimum of 154 types, the maximum lies at 1933 types. The average number of types is 716.47 with a median of 718 and a mode of 575. The absolute frequency of types displays a standard deviation of 278.95. 72 privacy policies show a frequency interval between 1346 and 2346 tokens, which points towards a normal distribution of the data. Figure 3 shows the frequency intervals of types across the privacy policies:

38% of policies (N = 133) show a frequency interval between 574 and 854 types, which also confirms the normal distribution of types across the language corpus. The type-token ratio as the variation in word choice lies at 0.01738 and is depicted in Fig. 4:

The diversity in length and lexical density becomes evident in the stark contrast between the total number of tokens and types. This confirms the assumption that privacy policies potentially have a fixed conventionalised set of words, i.e. a linguistic comfort zone that is rarely left. Lancsbox classifies 30.66% of tokens as nouns (N),

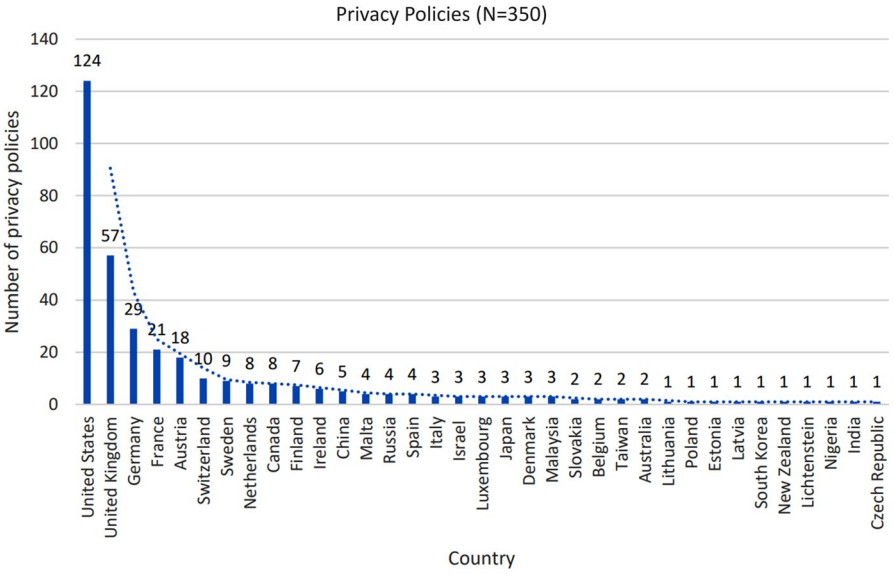

Fig. 1 Distribution of privacy policies across the privacy policy corpus

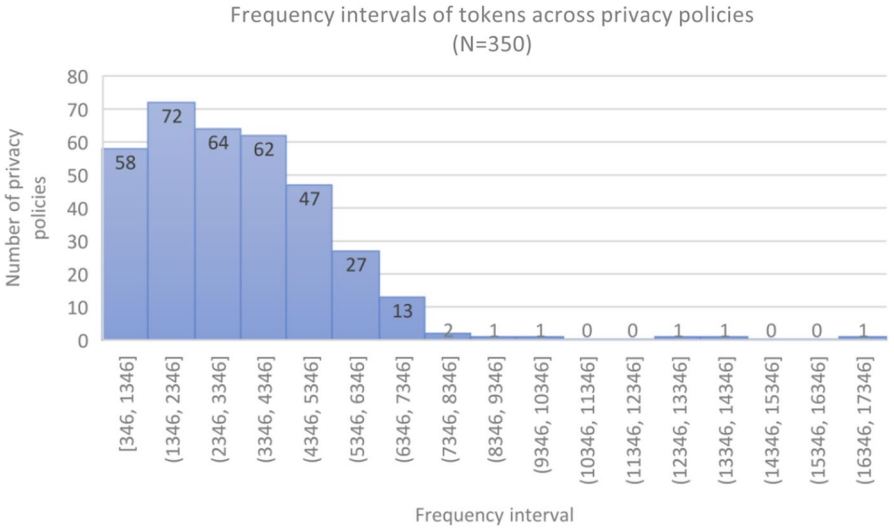

Fig. 2 Distribution of privacy policy length across the corpus

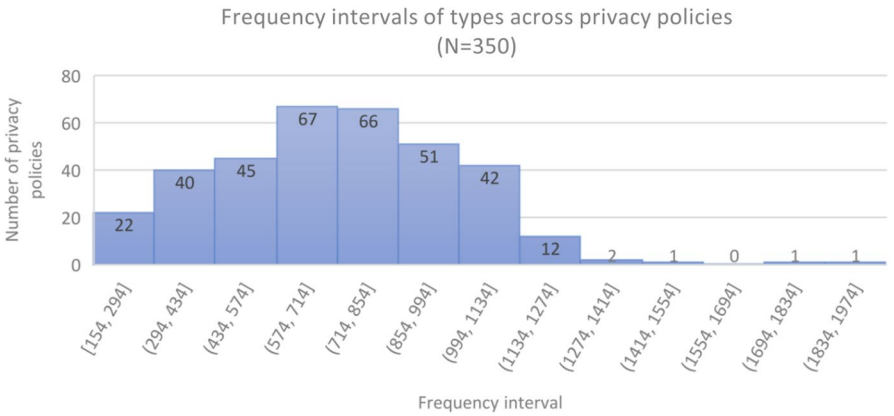

Fig. 3 Distribution of types across the corpus

16.09% of tokens as verbs (V), 25.06% of tokens as prepositions and subordinating conjunctions (PRP + SUB), 7.05% of tokens as adjectives (AJ), 3.70% of tokens as adverbs (AV), 10.75% of tokens as coordinating conjunctions (CC) and 6.69% of tokens as determiners (DT). The distribution (absolute frequency) of these parts of speech is shown in Fig. 5:

In this section, a general overview of the main characteristics of the data collected was provided. The data suggests a strong resemblance between the language of privacy policies and that of normative texts, e.g. statutes or contracts, due to the frequent nominalisation and complex syntax by means of subordination. In the following section, a quantitative and qualitative analysis of

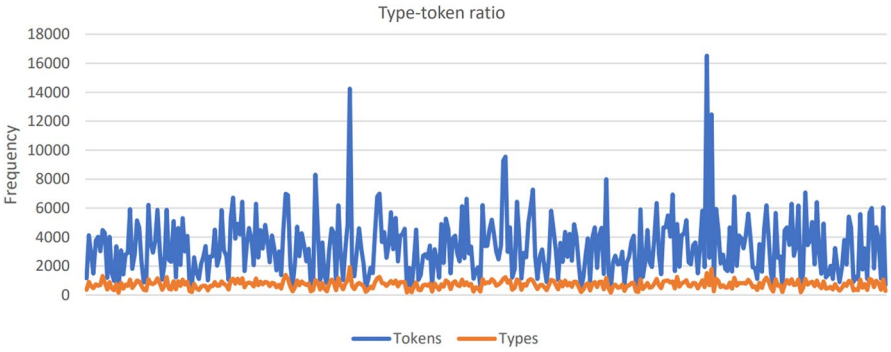

Fig. 4 Distribution of tokens and types across the corpus in comparison

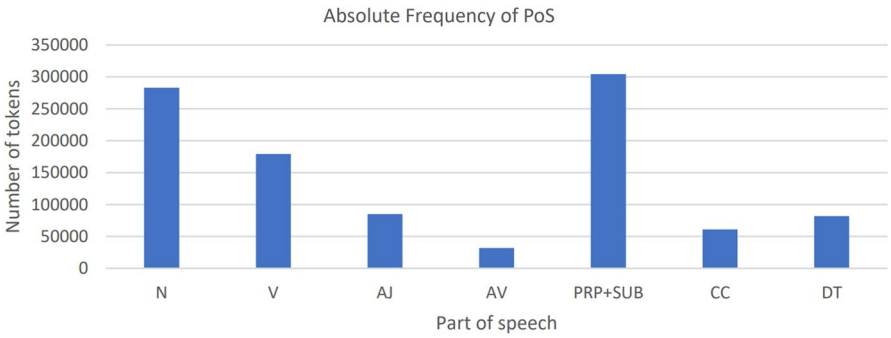

Fig. 5 Distribution of parts of speech (PoS) across the corpus

indeterminate adjectival phrases from the policies will be presented. The findings of the analysis will then be related to relevant legal questions in the context of transparency, informational self- determination and the issue of (un)informed consent.

6.3 Analysis

This section sets out to ascertain which forms of linguistic indeterminacy arise in the policy corpus, and how this indeterminacy may be relevant to the information requirement under the GDPR. First, a quantitative analysis of the data will be presented with a view to showing both frequent and infrequent adjectival phrases employed by policy drafters. Secondly, the most salient instances of indeterminacy arising from such phrases will be discussed in detail with regard to their legal relevance in the light of the language norms imposed by the GDPR.

6.3.1 Quantitative analysis

The corpus shows various manifestations of linguistic indeterminacy, originating from both the linguistic expressions used and an unclear referential relationship towards the external world. While adjectives only constitute approximately 7.3% of all linguistic forms, previous research shows that they are powerful tools of both precisification and deprecisification (see [47]), with some authors explicitly warning of their use in legal contexts (see [22]). Adjectives may thus also play a crucial role with regard to providing privacy-related information, e.g. processing purposes. In the following subsections, the general tendencies found in the data will be described. Then, the absolute and relative frequency of the most frequently attested adjectives will be presented and the quantitative distribution of frequently occurring word partnerships containing these adjectives will be shown. These findings will subsequently be contextualised with the legal challenges of informed consent.

6.3.1.1 General tendencies As previously stated, the data shows a strong tendency towards nominalisation and complex sentence structure by means of frequent subordination. Notably, *consent* occurs 2293 times in 309/350 policies, *agreement* occurs 460 times in 177/350 policies, *approval* is attested 27 times in 20/350 policies. In contrast, *complaint* occurs 376 times in 199/350 policies and *withdrawal* is only featured 107 times in 66/350 policies. In the context of consent, frequently occurring verb forms are *allow* (694 times in 240/350 policies), *agree* (355 times in 160/350 policies), *withdraw* (152 times in 105/350 policies) and *object* (546 times in 204/350 policies). The verb form *complain* is only attested 74 times in 54/350 policies. Constructions such as *allow you* (123/350 policies) and *allow us* (106/350 policies) both occur 179 times in the corpus. *You agree* to occurs 105 times in the overall corpus, but only in 73/350 policies, while the construction *you consent* to is attested 137 times in 96/350 policies. *Object to processing* (58 times in 41/350 policies) and *object to the processing* (73 times in 63/350 policies) are only rarely used constructions. The expression *withdraw your consent* (129 times in 128/350 policies) is marginally used and underrepresented in the data. A general insight from these preliminary findings is that the process of obtaining consent, informed or uninformed, heavily relies on such "performative formulas" [4:70] that are, despite their varying forms, constructions used consistently by data controllers to establish the conditions for agreement. The corpus suggests a tendency for data controllers to be able to obtain consent much more easily than for data subjects to withdraw it, which is contrary to the requirements set out in Article 7 of the GDPR [29]. This raises the question as to why policy drafters nevertheless favour expressions relating to the acquisition of consent, even though the communicative requirements of privacy policies are made explicit by the GDPR. The adjectives in the corpus are frequently used in attributive position and frequently involve types such as *personal* (12,454 times in 346/350 policies), *certain* (1707 times in 291/350 policies), *applicable* (1391 times in 284/350 policies), *necessary* (1413 times in 302/350 policies), and *legitimate* (989 times in 236/350 policies). Table 1 shows the absolute and relative frequency of fourteen of the most frequently attested adjectives in the corpus. These form the basis for the collocational analysis presented below.

Table 1 Fourteen most frequently attested adjectives

No	Adjective	Abs.Freq	Rel.Freq per 10 k tokens
1	Personal	12,454	107.53
2	Other	6425	55.47
3	Such	5380	46.45
4	Third	3741	32.30
5	Legal	2169	18.73
6	Certain	1707	14.74
7	Necessary	1413	12.20
8	Applicable	1391	12.01
9	Online	1308	11.29
10	Mobile	1192	10.29
11	Social	1181	10.20
12	Legitimate	989	8.54
13	Appropriate 676	5.84	
14	Reasonable	396	3.42
Total	N = 14	39,350	339.76

6.3.1.2 Collocational Analysis Significant word partnerships between adjectives and nouns are *personal data* (6520 times in 279/350 policies), *personal information* (5426 times in 251/350 policies), *third parties* (1973 times in 320/350 policies), *social media* (638 times in 161/350 policies), *applicable law* (417 times in 158/350 policies) and *legitimate interests* (401 times in 144/350 policies). It may be hypothesised that the interpretation of the adjectives in such word partnerships can have a significant bearing on the interpretation of the textual meaning of crucial sections in privacy policies, for instance purpose specification. Table 2 shows the twenty-one most frequently used word partnerships:

Figure 6 illustrates the eighteen collocations with the highest relative frequency across the corpus after third parties, personal data, and personal information, which were excluded from this chart due to their excessive frequency:

It is evident that most data controllers make frequent reference to third parties, with the singular form *third party* being used considerably less (1641 times in 262/350 policies). *Personally identifiable information* is referred to using two main constructions. The collocation *personal data* tends to be more frequently used than *personal information*, whereby the question is raised as to whether the conceptual difference between data and (identifiable) information is made sufficiently clear to the data subject. The corpus also shows that data controllers frequently make use of intertextual references to the construct of *lawful processing*, which is constructed by reference to *applicable law*, *legitimate interests*, *legal obligations* and *legal basis*. This suggests that policy drafters, potentially in an attempt to ensure GDPR-compliance, lift legal constructs from privacy legislation without providing any further specification. A considerable number of adjectival phrases may be described as rather precise, or specified by intertextual reference to statutory definitions, e.g. identifiable information may suggest this assumption. While it seems that overt

Table 2 Twenty-one most frequently occurring word partnerships

No	Collocation	Abs.Freq	RelFreq per 10 k tokens	/350	% of policies
1	third parties	1973	17.04	320	91.43
2	personal data	6520	56.30	279	79.71
3	personal information	5426	46.85	251	71.71
4	social media	638	5.51	161	46.00
5	applicable law	417	3.60	158	45.14
6	legitimate interests	401	3.46	144	41.14
7	legal obligations	225	1.94	132	37.71
8	mobile device	355	3.07	120	34.29
9	legal basis	333	2.88	119	34.00
10	other purposes	115	0.99	86	24.57
11	certain circumstances	136	1.17	80	22.86
12	mobile applications	149	1.29	73	20.86
13	reasonable steps	85	0.73	72	20.57
14	identifiable information	224	1.93	71	20.29
15	online services	92	0.79	63	18.00
21	reasonably necessary	62	0.54	54	15.43
16	certain services	72	0.62	52	14.86
17	social network	143	1.23	50	14.29
18	online advertising	61	0.53	34	9.71
19	such services	32	0.28	29	8.29
20	appropriate measures	28	0.24	24	6.86
21	such purposes	30	0.26	20	5.71
Total:	N = 21	17,517	151.25		

falsification is diametrical to the ethics of business communication, data controllers may utilise indeterminacy to connive the inference of "unstated meaning, beyond that derivable from the literal content explicitly stated message." [49:179].

Non-restrictive adjectives such as *other*, *such* or *certain*, which provide little or no specification as to quality of the concept, only narrow down the entity's epistemological status [26:165]. Ethic adjectives, e.g. reasonable, in the expression reasonable steps, remain both linguistically and legally indeterminate, since they are "related to an ethical standard or moral code", and thus "require a normative or deontic ordering source" to allow for a binding interpretation [26:165]. Modal adjectives, e.g. necessary in the expression necessary data, are sometimes used to form complex phrases such as reasonably necessary, thus linking social expectations of an ethical standard to contextual parameters of necessity. Finally, the meaning of relational adjectives, e.g. appropriate in the expression appropriate measures, is reflected in a relative requirement, e.g. data security, and an "objectively fixed or indisputable […] standard" [26:165]. While the linguistic indeterminacy present in these expressions needs no further commentary, the data seems to corroborate the

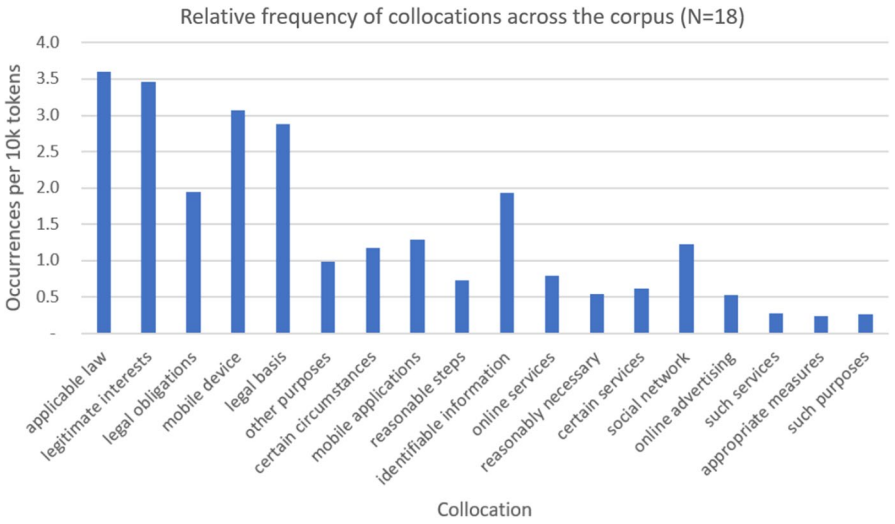

Fig. 6 Eighteen collocations with the highest relative frequency

assumption that such linguistic indeterminacy is indeed utilised for strategic reasons in policy drafting to maintain the interpretability and legal flexibility of the information presented. What is more, there is a tendency for data controllers to reproduce the wording of the GDPR, and with it, the challenges associated with statutory interpretation. In the following section, the qualitative analysis will be presented with a view to showing manually selected examples from the corpus, and their potential to prejudice the requirements of transparency, intelligibility, and most importantly, that of informed consent.

6.3.2 Qualitative Analysis

For reasons of feasibility and research focus, the quantitative findings of the previous analysis will be narrowed down and restricted to a usage-based investigation of modal and ethic adjectives. The aim of this section is to exemplify how the indeterminacy of such adjectival constructions is utilised by data controllers to tell "more than the truth" [49:179]. To allow for a systematic discussion, Fjelds typology of interpretation situations [26:170] was chosen to identify and explain how the interpretation of privacy-related information differs from that of ordinary language use. The following four situations of meaning-making may be used to describe the divergence in interpretation between legally trained policy drafters and the average data subject without legal training:

1. "The layman and lawyer [i.e. policy drafter] make the same interpretation, which means that general language interpretation strategies are adequate.
2. The layman and lawyer make nearly the same interpretation, but the layman feels insecure about his interpretation because of unusual linguistic signals.

3. The layman and lawyer make different interpretations because the text gives too few clues for the necessary recoverability of meaning.
4. The layman can make no sense of the […] text because of unfamiliar linguistic signals." [26:170]

It will be argued that strategic indeterminacy is likely to be found in the continuum between precisification and deprecisification of adjectival phrases (see [47]) and that, in some contexts, policy drafters accept the potential for such indeterminacy to weaken the autonomy of the data subject. In the following, strategic indeterminacy is associated with the occurrence of imprecision in the sense of situation (3), where policy drafters could have provided more precisification devices in a privacy policy, but evidently decided to refrain from doing so.

6.3.2.1 Modal and Ethic Adjectives Policy drafters often utilise modal adjectives to allow for controlled specification of "context demands of purpose and grade" [26:164]. Article 5 (1) of the GDPR states that "[p]ersonal data shall be […] adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" [29]. Importantly, the framework remains indeterminate regarding the meaning of adjectival phrases such as adequate and relevant, which arguably leads policy drafters to use modal adjectives in order to express and stress the necessity of data collection and processing. Example (3) below shows this communicative move potentially involving strategic indeterminacy:

(3) "We limit our uses of data for anti-fraud purposes to those which are strictly necessary and within our assessed legitimate interests to protect our customers and our services" (Sect. 16, Privacy Policy Corpus).

The legal context of the data controllers data processing is established by reference to *anti-fraud purposes*, and then restricted to only such processing that is both strictly necessary and within the boundaries of the data controllers legitimate interests. Whether or not consent is informed depends heavily on the interpretation of strict necessity, as data subjects must have at least basic knowledge of the contemporary technological standards in fraud prevention, and be able to relate it to the undefined legitimate interests of the data controller. Modal adjectives are sometimes also used as binomial expressions, which is an established feature of legal language use in English [32:123]. Example (4) shows how the indeterminacy of binomial expressions in policies may lead to complications in language processing:

(4) "We may share Personal Information with our headquarters and affiliates, and business partners to whom it is reasonably necessary or desirable for us to disclose your data for the purposes described in this Privacy Policy" (Sect. 82, Privacy Policy Corpus).

In the above example, the modal force of terms such as reasonable necessity and desirability remains indeterminate, and is thus unlikely to induce informed consent in data subjects. Notwithstanding their original intentions, the policy drafter calls upon the precisification interdiction to keep the policy open for legal assessment in potentially arising litigation. Article 5 (1) (b) states that personal

data can only be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" [29] It stands to reason whether informed consent as a product is at all arguable in such extreme cases of indeterminacy, due to the lack of clarity in terms of reference and context. It is reasonable to conclude that informed consent is highly unlikely to take place where the data subject is confronted with "pure vagueness", that is to say, an activation of the precisification interdiction where expressions "remain vague even if all the necessary contextual information is available." [26:159] The notion of practicality in the following example causes similar problems in the context of data collection:

(5) We generally collect personal information directly from you where this is reasonable and practical, but may also acquire information from other trusted sources to update or supplement the personal information you provided or which we processed automatically (Sect. 184, Privacy Policy Corpus).

In the above example, the modal force of practicality is linked to an undefined moral standard of reasonability, which does not provide transparent information on the collection process. The standards of reasonability and practicality, as referred to in the example, are thus likely to be a communicative strategy intentionally adopted to increase data subjects dependence on how the process of data collection is framed. (see [56]). Similarly, example (6) shows how the level of data security offered is linguistically realised by reference to an unrestricted territory (and elsewhere), and an unspecified standard of reasonability regarding security measures taken by the controller:

(6) By using the Services, you consent to such collection, storage and processing in the United States and elsewhere, though the United States and other jurisdictions may not afford the same level of data protection as considered adequate in your own country. We will take reasonable steps to protect your personal information (Sect. 47, Privacy Policy Corpus).

This raises serious issues with regard to the requirement of informed consent since data subjects remain uninformed as to the full territorial scope of the data handling and the actual level of security mechanisms provided. At this point, it is important to acknowledge that if strategic indeterminacy really is an empirically recoverable variable, it is likely to be a local rather than a global characteristic of privacy policy drafting [8:13]. This is founded on the assumption that policy drafters tendentiously make use of indeterminate adjectives in sections of particular legal relevance, where litigation is sought to be prevented by linguistic means. In the final example (7), justifiability is framed as depending on the interests of the data controller rather than those of the data subject. This is in line with the tendency found in more recent data that controllers are more likely to provide information regarding their desired entitlements than those of the data subject, e.g. the right to lodge a complaint, as found by Knotzer and Green ([39:186]).

(7) When we have no justifiable business need to process your personal information, we will either delete or anonymize it, or, if this is not possible […] then we will securely store your personal information and isolate it from any further processing until deletion is possible (Sect. 111, Privacy Policy Corpus).

In the example above, the policy drafter uses a negative framing of justifiable business needs that does not allow for a specification as to the circumstances under which such justifiability ceases to exist. "[T]ransparency-by-design", [35:4] as previously introduced, is not present in this example, since the policy is unlikely to satisfy the information requirement of clear language use. Furthermore, the "outcomes of the design process", that is, obtaining fully informed consent for a specific and explicit aspect, remain opaque due to the utilisation of both negation (see [72:498–518]) and indeterminacy. Due to space restrictions, a detailed discussion of the other adjectives is not possible. The use of other adjectives includes examples such as *adequate level of protection*, *appropriate measures*, *considerable importance*, *significant changes* and *sufficient advance notice*. The qualitative analysis has corroborated the assumption that linguistic indeterminacy tends to be a communicative strategy employed by data controllers to maintain interpretability and flexibility of the text. Since privacy policies are presumably drafted with a potentially negative data subject reaction in mind (see [53]) data controllers may naturally seek to maintain a certain prerogative of interpretation over the informative document. This raises the question as to which approach the courts should take in the reinforcement of linguistic norms, particularly where language is used to communicate with and inform vulnerable audiences such as children. In this context, it becomes evident that the current data protection framework lacks a frame of reference for courts to make an empirically founded decision as to whether a certain linguistic feature is or is not suitable for children. Section 6 presents the potential value of multilingual language documentation and annotation, and explores how this can provide a reasonable frame of reference for the courts in assessing audience-sensitive privacy information.

6.4 Results

The quantitative analysis shows that 56.6% of policies in the corpus comprise a range of 1346 to 4316 tokens. 38% of policies employ a range of 574 to 864 types. This shows that while policies may differ in length, a considerable number of texts use a fixed set of expressions. This is particularly evident in the collocations which occur throughout the corpus. The six most frequently occurring adjective-noun word partnerships are *third parties* (91.4%), *personal data* (79.1%), *personal information* (71.7%), *social media* (46.0%), *applicable law* (45.1%) and *legitimate interests* (41.1%). The corpus also displays a strong tendency towards nominalisation and subordination. Particularly the latter indicates the use of complex sentence structures, which may impede the intelligibility of the privacy policies. In the context of granting and withdrawing consent, 68.6% of policies comprise the performative verb *allow*, 45.7% of policies the form *agree* and only 30% the item *withdraw*. The form *object* is attested in 58.3% of policies, and *complain* in just 15.7%. This indicates that data controllers tend to place more emphasis on the collection of data than on the right of data subjects to withdraw consent. This may create the impression that it is easier for consent to be obtained than to be withdrawn, which stands in opposition to Article 7 of the GDPR [29]. Finally, there is a clear tendency of data

controllers to produce unspecified verbatim citations of adjectival phrases found in the GDPR, e.g. *legitimate interests*, *reasonable steps* and *legal basis*. This is significant since in doing so, data controllers seem to appropriate indeterminate phrases commonly found in normative texts in order to present the policies as normative texts themselves. In the qualitative analysis of modal and ethic adjectival expressions, it was found that data controllers use indeterminate adjectival phrases such as *strictly necessary*, *reasonably necessary* or *desirable*, *reasonable* and *practical*, *justifiable business need* and unspecified territorial information (e.g. *and elsewhere*). It may be argued that the occurrence of such indeterminate expressions without any further exemplification could constitute a manifestation of strategic indeterminacy as a means of maintaining the interpretability and legal flexibility of the policies, i.e. they ensure the texts are kept open for legal assessment (see [26]). Policies which contain these or similar expressions thus seemingly approve the inference of secondary discourses that are "beyond […] the literal content explicitly stated" [49:179] By utilising such linguistic constructions, data controllers impose on data subjects a sensitivity and understanding of specific legal constructs which require professional methods of legal exegesis that are beyond the legal literacy levels expectable from a reasonable person (see [28]).

7 Towards Enforceable Language Norms and an Informed Consent Culture

7.1 Void for Vagueness: Towards a GDPR with Teeth

The corpus analysis presented shows that after the beginning of GDPR enforcement, a considerable number of privacy policies contain unfair terms. Importantly, these texts do not constitute contracts ipso facto; instead, they are unilateral informative instructions that set out the conditions of controllers data handling practices which, although not always, are often found in the context of consumer contracts. It was previously argued that the information requirement imposed by the GDPR seemingly aims to remove, or least to level potentially existing power asymmetries between data controllers and data subjects. This endeavour is also visible in Article 3 of the Directive 93/13/EEC, stating that a term "which has not been individually negotiated shall be regarded as unfair if, contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations arising under the contract, to the detriment of the consumer" [21]. The interplay of EU consumer protection law and privacy legislation may be a powerful synthesis of norms which might have the potential to create a more balanced relationship between merchants and customers, between data controllers and data subjects. Privacy policies and the data handling practices described therein may play a significant role in the decision whether or not a consumer contract will be formed. In order to create the basis for a balanced relationship between data controllers and data subjects, it is reasonable to vindicate a case-specific and audience-oriented void for vagueness doctrine that more clearly regulates the interpretation of the information requirement. This suggestion is made on the grounds that if data controllers are given the freedom

of relatively autonomous policy drafting, the supervisory authorities and the courts should be equipped with an appropriate and effective toolkit to counter any unilaterally imposed imbalance and dependence. Legal corpus linguistics may be a valuable addition to this toolkit by means of a multilingual European reference corpus that presents cohesive annotation and documentation of authentic language use in context (see [57]).

7.2 The Applicability of a Common European Reference Corpus (CERC)

The linguistic inclusivity of privacy policies depends largely on the language code used in their construction. In order for an informed consent culture to live, it seems, formalism must die (see [65]), and pave the way for European privacy enforcement mechanisms that emulate validated methods of computer-assisted corpus linguistics. For instance, the compilation of a Common European Reference Corpus (CERC) for policy drafting may provide a comparable and verifiable empirical basis to guide the judiciary, policy drafters, data subjects and the scientific community towards the characteristics of linguistically inclusive, audience-oriented and GDPR-compliant privacy policies. Such a corpus will provide much needed data that allows for a usage- based and norm-referenced legal assessment of concision, transparency, intelligibility and accessibility of privacy-related information. At the same time, it will strengthen legal certainty and reduce arbitrariness in adjudication. The decision as to whether a certain privacy policy adheres to the language norms of the GDPR would largely rely on the judiciary's application of a systematic, yet context-sensitive three-step-test of compositional inclusivity. In this empirical model, the overall transparency, clarity and intelligibility of privacy-related information would be conceptualised as consisting of the sum of its parts. A test of compositional inclusivity would take into consideration a norm-referenced evaluation of the following three questions:

(1) Is the use of the disputed term represented in the audience-specific section of the CERC? If so, how often and in which context?
(2) Does the use of the term deviate from its expectable common meaning?
(3) Is the use of the term likely to induce in the audience improper interpretations as to the nature and the purpose of the data processing?

The utilisation of reference corpora and related analysis toolkits in interpretation was successfully applied in a number of US cases, where courts decided to extend their "statutory interpretation tool box" in order to determine the common or ordinary meaning of individual linguistic expressions (see [46]). Similarly, although the contemporary limitations of natural language processing (NLP) must be recognised, (see [14]) big data corpus analysis may enable artificial intelligence with powerful algorithms that could allow for organised linguistic interaction with data subjects. For instance, the "pure vagueness" [26:159] associated with some adjectival expressions could receive specification by means of plain language exemplification. This is not to claim that if the language code or semantic understanding of a data

subjects query is accessible by means of NLP, [41:66] the productivity, complexity and pragmatics of linguistic indeterminacy can be entirely offset. However, as argued by Maxwell and Schafer, "the semantic intent of the query would be helpful" in information retrieval. [43:66] This interactive element to privacy-related information could potentially lead to a more participatory consent culture, in which data subjects assume an agentive rather than a passive role, and where they know and are aware "that their contributions matter" [36:6]. However, whether or not future data subjects perceive themselves as autonomous agents in the information society also depends on the role ascribed to data protection issues in education.

7.3 Tackling the Participation Gap in the Information Society

The previous section introduced an argument for a participatory privacy culture in which data subjects are the architects of their informational self-determination. Yet for individuals to make informed decisions on privacy settings, data protection must be appropriately represented in education. Building on this view, it seems reasonable that legal education in schools could be a powerful tool to tackle the lack of participation in the information society effectively. Paradoxically, while the beginning of GDPR enforcement has generated considerable media interest, evidence for privacy education in national curricula remains scarce. In Austria, for instance, data protection law is a much- neglected child in school subjects such as Economics and Law and Law and Justice [11]. While it is indeed debatable when and how legal literacy in general and "online privacy literacy" [64:655–671] in particular should be acquired, the Austrian School Organisation Act (Schulorganisationsgesetz 1962) clearly provides that education should equip children with the knowledge and competence necessary for their future life and profession. [52] In addition, Article 8 of the European Convention of Human Rights (ECHR) explicitly awards protection to childrens "private and family life" [25] and the Convention on the Rights of the Child (CRC) stresses that the "best interests of the child shall be a primary consideration." [15] The implementation of privacy literacy programmes in schools could lead to higher legal autonomy and awareness amongst students. Further endeavours in the regulation and specification of privacy policy drafting, as evident in the Age Appropriate Design Code in the UK [34], only constitute a first step and should be complemented by educational measures that take into account the fundamental right to privacy, which indisputably is in any childs best interest. The enforcement of privacy legislation and its concurring general preventive effects are not sufficient means to create a participatory privacy culture without the various linguistic barriers for data subjects. The information requirement of the GDPR does allow for a legal empowerment of data subjects, but in order to exercise their rights and entitlements, individuals must be made aware of *actionable* protection of their private and family life. The findings of the present study, for instance, seem to suggest that data controllers tendentially provide more information about their own entitlements than the rights of the data subject. Legal education concerning data protection and privacy issues could foster students' ability to recognise the significance of privacy-related information, to draw informed conclusions on the conditions and extent of

processing operations and to use the conclusions drawn to make informed decisions. In this view, the aim of privacy education is not the transfer of professional legal competence, but to provide students with a toolkit that sufficiently represents the space between privacy legislation as it is and privacy information as it ought to be (see [33:593–629]).

8 Conclusion

This study investigated how indeterminacy is utilised in policy drafting, which functions adjectives fulfil in this context and how this can be accounted for. In this context, the legal challenges for informed consent were discussed and related to the question as to whether the information requirement of the GDPR is sufficient. To this end, a usage-based investigation of 350 online privacy policies was conducted, including a quantitative and qualitative analysis of the texts. It was found that a considerable number of data controllers continue to make use of indeterminacy in the context of purpose limitation, which prejudices compliance with the requirement of informed consent under the GDPR. The corpus shows a strong tendency towards frequent nominalisation and subordination, leading to complex and cognitively challenging syntax in privacy policies. The quantitative analysis revealed that while privacy policies may differ considerably in length, drafters seem to use a relatively fixed set of expressions in written instruction from which they rarely deviate. This linguistic comfort zone is likely established by convention and does not necessarily cater for specific audiences within the information society. Policy drafters frequently make use of indeterminate adjectival phrases from the GDPR, which leads to the conclusion that the language of normative texts is at times unduly appropriated in order to create a unilaterally induced imbalance. In the qualitative analysis, the hypothesised existence of strategic indeterminacy was corroborated by the occurrence of indeterminate phrases and binomials, such as *strictly necessary*, *reasonably necessaryor desirable*, and *reasonable and practical*. It is concluded that by utilising these and similar expressions, data controllers approve the inference of secondary discourses that are beyond what is explicitly stated in the policy ("telling more than the truth" [49]). The appropriation of normative language use requires professional methods of legal exegesis that are beyond the average legal literacy levels of the reasonable person (see [28]). This also shows severe shortcomings in the consideration of the diversity of the audiences targeted. With reference to Article 3 of the Directive 93/13/EEC, a case-specific and audience-sensitive void for vagueness doctrine is proposed in relation to the interpretation of the information requirement under the GDPR. This suggestion relies on the assumption that an interplay of EU consumer protection law and privacy legislation may provide a powerful synthesis which could lead to an empowerment of data subjects. To provide the judiciary, policy drafters, data subjects and researchers with authentic language use in privacy-related information, the compilation of a multilingual, annotated and context-sensitive Common European Reference Corpus (CERC) is suggested. Such data would provide an empirically founded basis for usage-based and norm-referenced legal assessment of language norms, and could be applied in the context of natural

language processing, e.g. by providing specification of indeterminate expressions by means of plain language exemplification. Finally, it is concluded that more restrictive regulation of privacy policy drafting is only a first step. The implementation of empirically informed curricula in secondary education has the potential to provide students with a privacy toolkit that allows for the critical deconstruction of privacy-related information. Legal education in schools could help promote a privacy culture in which the data subjects of tomorrow become architects of their own informational self-determination. In this way, students can develop the ability and confidence to draw informed conclusions from privacy information in order to be able to take action according to their best interests.

Acknowledgements This article constitutes a reworked version of my Master thesis, which I composed under the much appreciated supervision and support by my professors and lecturers at the Edinburgh Law School, University of Edinburgh, South Bridge, Edinburgh. I am deeply grateful to my professors and lecturers in Scotland for their enduring impact on my academic journey.

Funding Open access funding provided by Vienna University of Economics and Business (WU).

References

1. Ales, Edoardo, Ylenia Curzi, Tommaso Fabbri, Olga Rymkevich, Iacopo Senatori, and Giovanni Solinas. 2018. *Working in Digital and Smart Organizations: Legal, Economic and Organizational Perspectives on the Digitalization of Labour Relations*. Cham: Springer.
2. Arts, Anja, Alfons Maes, Leo G. M. Noordman, and Carel Jansen. 2011. Overspecification in Written Instruction. *Linguistics* 49 (3): 555–574.
3. Asprey, Michele M. 2003. *Plain Language for Lawyers*, 3rd ed. Annandale, New South Wales: Federation Press.
4. Austin, John Langshaw. 1975. *How to Do Things with Words*. Oxford: Clarendon Press.
5. Barthes, Roland. 1967. The Death of the Author. In Art and Interpretation: An Anthology of Readings in Aesthetics and the Philosophy of Art, 383–86. Peterborough: Broadview.
6. Bechmann, Anja. 2014. Non-Informed Consent Cultures: Privacy Policies and App Contracts on Facebook. *Journal of Media Business Studies* 11 (1): 21–38.
7. Berry, David. 2011. The Computational Turn: Thinking about the Digital Humanities.
8. Bhatia, Vijay Kumar, and Jan Engberg. 2005. *Vagueness in Normative Texts*. Bern: Peter Lang.
9. Biber, Douglas. 2012. Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. The Oxford Handbook of Linguistic Analysis, 2nd ed., 193–224.
10. Brumfit, Christopher. 1995. Teacher Professionalism and Research. In Principles and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson, 27–41. Oxford: Oxford University Press.
11. Bundesgymnasium, Bundesrealgymnasium und Wirtschaftskundliches Realgymnasium für Berufstätige. 2018. Wahlpflichtfach „Recht und Gerechtigkeit“). Accessed 22 Jan 2024. https://doi.org/10.13140/RG.2.2.26767.69283.

12. Association, Canadian Bar, ed. 1992. *Report of the Canadian Bar Association Task Force on Legal Literacy*. Ottawa: Canadian Bar Association.
13. Chandler, Daniel. 1997. An Introduction to Genre Theory. 1997: 1–15. Accessed 22 Jan 2024. www.visualmemory.co.uk/daniel/Documents/intgenre/ chandler_genre_theory.pdf.
14. Clark, Alexander, Chris Fox, and Shalom Lappin. 2013. *The Handbook of Computational Linguistics and Natural Language Processing*. Chichester: John Wiley & Sons.
15. Convention on the Rights of the Child. 1989. Council Directive 93/13/EEC (1993).
16. Culpeper, Jonathan, Michael Haugh, and Dániel. Z. Kádár. 2017. *The Palgrave Handbook of Linguistic (Im)Politeness*. Basingstoke: Palgrave Macmillan.
17. Culver, Keith. 1999. *Readings in the Philosophy of Law*. Ontario: Broadview Press.
18. Cummings, Louise. 2005. *Pragmatics: A Multidisciplinary Perspective*. Edinburgh: Routledge.
19. Cummings, Louise. 2015. Theory of Mind in Utterance Interpretation: The Case from Clinical Pragmatics. *Frontiers in Psychology* 6 (2015): 1–14.
20. Dagan, Hanoch, and Michael Heller. 2017. *The Choice Theory of Contracts*. Cambridge: Cambridge University Press.
21. Council Directive 93/13/EEC of 5 April 1993 on Unfair Terms in Consumer Contracts, Pub. L. No. 31993L0013, OJ L 095 (1993), http://data.europa.eu/eli/dir/1993/13/oj/eng.
22. Eidenmueller, Horst, Florian Faust, Hans Christoph Grigoleit, Nils Jansen, Gerhard Wagner, and Reinhard Zimmermann. 2008. The Common Frame of Reference for European Private Law—Policy Choices and Codification Problems. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.
23. Endicott, Timothy. 2008. The Value of Vagueness. In Vagueness in Normative Texts, 27–48.
24. Endicott, Timothy, Arbitrariness. 2014. Canadian Journal of Law and Jurisprudence 2014, Oxford Legal Studies Research Paper No. 2/2014, available at SSRN: https://ssrn.com/abstract=2378858
25. European Convention of Human Rights (1953).
26. Fjeld, Ruth Vatvedt. 2005. The Lexical Semantics of Vague Adjectives in Normative Texts.
27. Fraser, Bruce. 2010. Pragmatic Competence: The Case of Hedging, in New Approaches to Hedging, ed. Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider (Emerald, 2010), 25.
28. Gardner, John, 2019. The Many Faces of the Reasonable Person, ed. Gardner, John, Torts and Other Wrongs. Oxford: Oxford University Press.
29. General Data Protection Regulation and Recitals (2016).
30. Gerrig, Richard J. 2015. Meaning in Context. *The American Journal of Psychology* 128 (2): 135–145.
31. Günther, Franziska. 2016. *Constructions in Cognitive Contexts: Why Individuals Matter in Linguistic Relativity Research*. Berlin: De Gruyter.
32. Gustafsson, Marita. 2009. The Syntactic Features of Binomial Expressions in Legal English. *Text - Interdisciplinary Journal for the Study of Discourse* 4 (1–3): 123–142.
33. Hart, H.L.A. 1958. Positivism and the Separation of Law and Morals. *Harvard Law Review* 71 (4): 593–629.
34. Information Commissioners Office. 2018. Call for Evidence - Age Appropriate Design Code. Accessed 27 June 2018. https://ico.org.uk/about-the-ico/ico-and- stakeholder-consultations/call-for-evidence-age-appropriate-design-code/.
35. Janssen, Marijn, Ricardo Matheus, Justin Longo, and Vishanth Weerakkody. 2017. Transparency-by-Design as a Foundation for Open Government. Transforming Government: people, process and policy (online), 11(1), 2–8.
36. Jenkins, Henry, Ravi Purushotma, Margaret Weigel, Katie Clinton, and Alice J. Robison. 2009. *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. London: MIT Press.
37. Kahn-Freund, Otto. 1968. Delictual Liability and Conflict of Laws. In Collected Courses of the Hague Academy of International Law, Vol. 124, 1–166. Leiden: Martinus Nijhoff.
38. Klijnsma, Josse. 2015. Contract Law as Fairness. *Ratio Juris* 28 (1): 68–88.
39. Knotzer, Stefan, and Daniel Green [Leisser]. 2020. *Die Datenschutzerklärung Compliance in klarer und einfacher Sprache*. Vienna: LexisNexis.
40. Knotzer, Stefan. 2018. Wissenschaftliche Forschung und Datenschutz: Eine kritische Analyse ausgewählter Aspekte der österreichischen Rechtslage. *ZTR* 4: 202–218.
41. Lakoff, Robin Tolmach. 2017. Context Counts: Papers on Language, Gender, and Power.
42. Macenaite, Milda, and Eleni Kosta. 2017. Consent for Processing Children's Personal Data in the EU: Following in US Footsteps? *Information & Communications Technology Law* 26 (2): 146–197.

43. Maxwell, Tamsin, and Burkhard Schafer. 2010. Natural Language Processing and Query Expansion in Legal Information Retrieval: Challenges and a Response. *International Review of Law, Computers and Technology* 24 (1): 63–72.
44. McEnery, Tony, and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
45. Paltridge, Brian. 1994. Genre Analysis and the Identification of Textual Boundaries. *Applied Linguistics* 15 (3): 288–299.
46. People vs. Harris, No. 149872, 149873, 150042 (Supreme Court of Michigan 22 June 2016).
47. Pinkal, Manfred. 1995. *Logic and Lexicon The Semantics of the Indefinite. Studies in Linguistics and Philosophy*. Dordrecht: Kluwer Academic Publishers.
48. Pollach, Irene. 2005. A Typology of Communicative Strategies in Online Privacy Policies: Ethics, Power and Informed Consent. *Journal of Business Ethics* 62 (3): 221–235.
49. Riley, Kathryn. 1993. Telling More than the Truth: Implicature, Speech Acts, and Ethics in Professional Communication. *Journal of Business Ethics* 12 (3): 179–196.
50. S. H. O. 1866. Patent and Latent Ambiguities in Written Instruments. *University of Pennsylvania Law Review* 14 (3): 140–142.
51. Schmidt, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees.
52. RIS - Schulorganisationsgesetz - Bundesrecht Konsolidiert, Fassung Vom 29.07.2018, section 2, accessed 22 Jan 2024, https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=100092
53. Stark, Tina L. 2003. *Negotiating and Drafting Contract Boilerplate*. New York: ALM Publishing.
54. State of Utah vs. Andy Rasabout, No. 20130430 (Supreme Court of Utah 14 August 2015).
55. Steininger, Katharina, and David Rückel. 2013. Legal Literacy and Users Awareness of Privacy, Data Protection and Copyright Legislation in the Web 2.0 Era. *Wirtschaftsinformatik Proceedings* 2013: 1651–1665.
56. Stocke, Volker. 2002. *Framing und Rationalität: Die Bedeutung der Informationsdarstellung für das Entscheidungsverhalten*. Munich: De Gruyter.
57. Thieberger, Nick, Anna Margetts, Stephen Morey, and Simon Musgrave. 2016. Assessing Annotated Corpora as Research Output. *Australian Journal of Linguistics* 36 (1): 1–21.
58. Traxler, Matthew, and Morton Ann Gernsbacher, eds. 2006. *Handbook of Psycholinguistics*. London: Academic Press.
59. TreeTagger Interface. Accessed 22 Jan 2024. https://corpora.lancs.ac.uk/tree-tagger/.
60. Vessey, Rachelle. 2017. Corpus Approaches to Language Ideology. *Applied Linguistics* 38 (3): 277–296.
61. Vogel, Friedemann, Hanjo Hamann, and Isabelle Gauer. 2017. Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies. *Law & Social Inquiry* 2017: 1–24.
62. Kessel, Walser, and Caroline, and Maria Crespo. 2009. Visualisierung von Rechtsnormen durch Kinder – Darstellung ihres Fairness- und Gerechtigkeitssinns. *Jusletter* 2009: 1–27.
63. Webster, Mandy. 2017. *Data Protection in the Financial Services Industry*. London: Routledge.
64. Weinberger, Maor, Maayan Zhitomirsky-Geffet, and Dan Bouhnik. 2017. Factors Affecting Users Online Privacy Literacy among Students in Israel. *Online Information Review* 41 (5): 655–671.
65. Whittington, Keith E., R. Daniel Kelemen, and Gregory A. Caldeira. 2010. *The Oxford Handbook of Law and Politics*, 2010. Oxford: Oxford University Press.
66. Widdowson, Henry George. 2014. The Role of Translation in Language Learning and Teaching. In Translation: A Multidisciplinary Approach. (2014): 222–40. Basingstoke: Palgrave Macmillan.
67. Widdowson, Henry George. 2011. *Discourse Analysis Oxford Introduction to Language Study Series*. Oxford, New York: Oxford University Press.
68. Widdowson, Henry George. 2004. *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford: Blackwell.
69. Willett, Chris. 2016. *Fairness in Consumer Contracts: The Case of Unfair Terms*. New York: Routledge.
70. Wilson, Shomir, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, et al. 2016. The Creation and Analysis of a Website Privacy Policy Corpus, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. (2016): 1330–40.
71. Wright, Crispin. 1975. On the Coherence of Vague Predicates, Synthese: An International Journal for Epistemology. *Methodology and Philosophy of Science* 30: 325–365.

72. Zeijlstra, Hedde. 2007. Negation in Natural Language: On the Form and Meaning of Negative Elements. *Language and Linguistics Compass* 1 (5): 4983.
73. Zweigert, Konrad, Hein Kötz, and Tony Weir. 2015. *An Introduction to Comparative Law.* 131: 563–584.
74. 1800-Flowers Dallas. Privacy Policy. Accessed 4 July 2018. http://www.1800flowersdallas.com/privacy