Legal Jargon in an Environmental TKB: Pollution Phraseology (Short Paper)

Arianne Reimerink[1], Pamela Faber[1], Melania Cabezas-García[1], and Pilar León-Araúz[1]

[1] *University of Granada, Departamento de Traducción e Interpretación, C/ Buensuceso, 11, 18071 Granada, Spain*

Abstract

Despite its importance, Environmental Law has largely been ignored in environmental knowledge bases. EcoLexicon (ecolexicon.ugr.es) has recently begun to include information on the domain. This paper takes the methodological perspective of Frame-based Terminology [1, 2, 3] to analyze typical verb collocations in Environmental Law that will be added to the phraseology module of EcoLexicon. Corpus analysis was used to compare the behaviour of verbs collocating with *pollution* in Environmental Science and Environmental Law. Verbs were classified based on the lexical domains and semantic classes in Faber and Mairal [4]. The differences were mostly based on the specificity of the other arguments and the emphasis on the polluter in Environmental Law. This resulted in a proposal for the inclusion and configuration of legal information in EcoLexicon.

Keywords

Environmental Law, TKB, phraseology

1. Introduction

Environmental Law is an important transversal domain that combines Law with Environmental Science. It is impossible to understand the environment without an in-depth knowledge of how nations regulate it. For this reason, EcoLexicon (ecolexicon.ugr.es) has begun to include concepts and terms in different languages that pertain to Environmental Law [5, 6]. This multilingual and multimodal terminological knowledge base (TKB) [7] represents the conceptual structure of the specialized domain of the Environment in the form of a visual thesaurus. It combines conceptual, linguistic, and graphical information to help translators, technical writers, and environmental experts acquire an in-depth understanding of specialized environmental concepts and to help them write or translate specialized or semi-specialized texts. It is the practical application of Frame-Based Terminology (FBT) [1, 2, 3], a cognitive approach to domain-specific language, which directly links specialized knowledge representation to cognitive linguistics and cognitive semantics. In FBT, knowledge acquisition begins at the term-level, progresses to the phrase level, and finally results in the codification of an entire knowledge frame. The data are collected by means of corpus analysis.

In a previous study [6], to expand and improve the information related to Environmental Law in EcoLexicon, comparative corpus analysis was used to identify missing concepts, and explore how the multidimensional nature [8] of Environmental Science might affect the behaviour of other concepts in the subdomain of Environmental Law. Our study focused on the POLLUTION frame and the results showed that a new participant (i.e. the POLLUTER) had to be added when contextualised for the subdomain of Environmental Law. Whereas in Environmental Science the main focus is generally on the polluting substance, in Environmental Law, it is the person/institution/industry responsible (see example 1 and 2, emphasis by the authors). We also discovered that some facets of the concept

2nd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2023, June 29–30, 2023, Lisbon, Portugal

EMAIL: arianne@ugr.es (A. Reimerink); pfaber@ugr.es (P. Faber); melaniacabezas@ugr.es (M. Cabezas-García); pleon@ugr.es (P. León-Araúz)

ORCID: 0000-0002-7264-4580 (A. Reimerink); 0000-0003-0581-0005 (P. Faber); 0000-0002-8622-1036 (M. Cabezas-García); 0000-0002-8520-2749 (P. León-Araúz)

POLLUTION are more prominent in this subdomain as compared to the environmental domain as a whole: time and origin (see examples 3 and 4).

1. The **pollutants** disperse in a downward direction causing substantial air pollution at ground level but cannot escape upwards because of the inversion.
2. …the **polluter** pays principle, the **person** responsible for the pollution cannot be identified or cannot be held liable under Community or national legislation…
3. Indeed, the phenomenon of **historical** pollution represents the result of the convergence and interaction of a number of different factors…
4. Historically the regulation of **vessel-source** pollution has engendered conflict between coastal States…

These results entailed changes in the conceptual networks and the definitions of EcoLexicon. Since these differences at the conceptual level also affect the linguistic level, namely, the choice of verbs, terms and phraseological structures, this study analyzed verb collocations in Environmental Law to add to the phraseology module of EcoLexicon, which is currently under construction. In this pilot study, we focus on phraseology in English. Future research will also address the topic in Spanish, one of the other major languages of EcoLexicon.

The rest of this paper is organized as follows: Section 2 explains the phraseology extraction method and the results; Section 3 provides a proposal for the representation of these results in the phraseology module; and Sections 4 summarizes the conclusions that can be derived from this research.

2. Phraseology extraction

2.1. Extraction method

When completed, the phraseology module of EcoLexicon will be one of the most important for the representation of Environmental Law data because of its legal terminology. The phraseology module is based on a wide interpretation of the concept of collocation and at its core are verb collocations. In FBT, verb collocations are frequent combinations of two or more lexical units composed of a noun + verb or a verb + noun where the meaning of the verb is limited by the meaning of the noun. However, at the same time, the verb restricts the type of noun with which it can combine [9]. For example, in the collocation "the fire burns", the verb only allows for arguments that can be on fire, whereas the argument "fire" needs a verb that refers to the process of combustion [10]. In this module, verbs will be classified based on their meaning in combination with the terms with which they collocate. Verbs will not have their own entry in EcoLexicon but will be included as additional information in the term entries. The inclusion of a phraseme in EcoLexicon is essentially based on frequency of occurrence in the corpus. However, as will be shown, frequency changes when comparing different subdomains. Therefore, different phrasemes and examples will be shown depending on the context the end user is focussing on in EcoLexicon.

To compare the collocational behaviour of *pollution* in Environmental Science and the subdomain of Environmental Law, Sketch Engine (https://www.sketchengine.eu/) [11] was used. As a reference corpus, we used the EcoLexicon Environmental Corpus (EEC, 23 million words) available in the Open Corpora section of Sketch Engine and compared it to a corpus specifically created for this purpose: the Environmental Law corpus, composed of EEC texts, tagged with the domain of Environmental Law, as well as additional texts from the same domain harvested from the Internet (enLaw, 9.7 million words). The EEC and enLaw were both compiled in Sketch Engine with the Penn Treebank tagset and the EcoLexicon Semantic Sketch Grammar (ESSG) [12].

The ESSG is a Corpus Query Language (CQL)-based grammar [13] as is the default grammar used for word sketches in Sketch Engine. Whereas Sketch Engine's default grammar provides grammatical relations, such as verb-object, modifiers, and prepositional phrases, the ESSG was developed for the extraction of semantic word sketches based on some of the most common semantic relations in terminology: generic-specific, part-whole, location, cause, and function. The Sketch Engine functionalities used to compare the two corpora were the following: Word Sketch and Concordance. For the Word Sketch function the default settings provided by Sketch Engine were used.

After extraction, verbs were categorized according to the lexical domains in Faber and Mairal [4]. They analyzed and categorized the semantic and syntactic structure of 12,000 general language verbs through definition factorization, as described in the Lexical Grammar Model (LGM), and validated by corpus analysis. This resulted in the following general lexical domains: EXISTENCE (*be*, *happen*), CHANGE (*become*, *change*), POSSESSION (*have*), SPEECH (*say*, *talk*), EMOTION (*feel*), ACTION (*do*, *make*), MENTAL PERCEPTION (*know*, *think*), MOVEMENT (*move*, *go*, *come*), PHYSICAL PERCEPTION (*see*, *hear*, *taste*, *smell*, *touch*), MANIPULATION (*use*), CONTACT/IMPACT (*hit*, *break*), and POSITION (*put*, *be*). Other smaller classes included LIGHT, SOUND, BODY FUNCTIONS, WEATHER, etc.

2.2. Extraction results and discussion

The data extracted are in Tables 1-4 in the Annex. Table 1 shows that the verbs that collocate with *pollution* in both corpora mostly belong to the domain of CAUSATIVE EXISTENCE, more specifically to cause something to exist (*cause*), to cause something to cease to exist (*eliminate*), and to cause something to not happen (*prevent*, *avoid*). Other important lexical domains are CHANGE, more specifically, to cause something to change by decreasing it (*abate, reduce*, *minimize*, *mitigate*, *decrease*, *limit*) and MANIPULATION (*control*, *monitor*). Finally, the lexical domains of VISUAL PERCEPTION, COGNITION, and SPEECH are present with verbs such as *consider*, *define*, *regard*.

In both word sketches, *air* is high up on the list. However this is a tagging mistake as in these cases *air* is a noun in an adjectival position and not a verb. The tagging mistake occurs in constructions where the tagger is incapable of interpreting *to* as a preposition.

In the word sketch of verbs with *pollution* as subject (see Table 2), there are fewer results for the EEC because the numbers of collocations with *pollution* did not exceed a certain threshold. This makes sense because the EEC is a corpus on the environment. Pollution is thus only one of the aspects to be considered. In contrast, in the enLaw corpus, pollution is a central concept, and that is why collocations with pollution are statistically more relevant. The lexical domain of the verbs that predominate in both corpora is EXISTENCE: *originate*, *occur*, *arise*, *be*, *emanate*, *become*, *include*. Another lexical domain present in both corpora is CHANGE (*reduce*, *increase*), to cause something to change by making it worse (*destroy*, *damage*, *harm*, *threaten*) and more general causative verbs such as *cause*, *affect*, *derive*, *result*.

The verb *flush* in the EEC word sketch of pollution is the result of the term *pollution flushing*, which is a process through which pollution is removed from a water body through natural or artificial currents or tides. It can be classified as to cause something to cease to exist (EXISTENCE) or as MOVEMENT [5].

After analysing *pollution*, we also analysed the verb *pollute* and the noun *polluter* in Word Sketch. When looking at the results for the word sketch *object_of*, there were no obvious differences between the verb's behaviour in enLaw and EEC, apart from the difference in the number of results. Table 3 shows *polluter* as the object of verbs. Once again, the enLaw corpus provides more results, some of which are directly related to the legal domain: *prosecute*, *sue*. Another important lexical domain is MANIPULATION: *implement*, *regulate*, *oblige*, *force*, *compel*, *deter*, *require*, etc. Finally, the word sketch *polluter subject_of* showed the verb *pay* as the very first result for both corpora. This is of course because one of the most important principles of Environmental Law is the polluter-pays principle.

Apart from the fact that there are more results for pollution in enLaw, the lexical domains of the verbs collocating with pollution were very similar in both corpora. The differences pertained to the arguments of the verbs.

Figure 1 (see Annex) shows an extract of the concordances of the CQL *abate* + *pollution* in enLaw. The second argument that collocates with this combination is an institutional body (*state*, *UK*), a company (*industries*, *firms*), measure (*measures*) or cost (*expenditures*, *costs*). The second argument for the CQL *minimise* + *pollution* (Figure 2) is mostly measure (*requirements*, *directive*, *measures*). The second argument for the CQL *control* + *pollution* (Figure 3) includes institutional body (*state*, *administration*, *agencies*) and measure (*strategies*, *measures*, *regulations*, *laws*).

In Environmental Law, the verbs *abate*, *minimise* and *control* would be included in the phraseology module under the term *pollution* in the following phrasemes: INSTITUTIONAL BODY/COMPANY/MEASURE/COST + CHANGE [decrease] + POLLUTION; INSTITUTIONAL BODY/MEASURE + MANIPULATION + POLLUTION.

One of the participants that is specific to the POLLUTION frame in Environmental Law is evidently the POLLUTER. Figure 4 shows an extract of the concordances of the CQL *pollution caused_by* in enLaw. The cause is evidently the polluting industry (*ship*, *operational discharges*, *activities*) or the person or entity responsible (*polluters*, *manufacturers*, *persons*, *parties*, *corporation*).

3. Phraseology representation

The results showed that the lexical domains of the verbs that collocate with *pollution* were quite similar in the EEC and enLaw corpora. The differences are mostly based on the specificity of the other arguments and the emphasis on the POLLUTER in the Environmental Law subdomain. To represent this in the phraseology module, under the term *pollution*, the choice of example sentences provided for the subdomain would be the following (Figure 5).

Pollution [Environmental Law]			
EXISTENCE [cause to exist]	cause	INDUSTRY	
		INSTITUTIONAL BODY	On the other hand, if state B causes pollution in state A, state A is entitled to invoke its territorial sovereignty
		PERSON/ COMPANY	the cost is borne by the company who causes the pollution or transferred to consumers driving demand for the relevant product
EXISTENCE [cause to cease to exist]	eliminate tackle		
EXISTENCE [cause to not exist]	prevent avoid		
CHANGE [decrease]	abate decrease limit minimize mitigate reduce	INSTITUTIONAL BODY	Courts have allowed a common law suit brought by one state to abate pollution emanating from another state
		COMPANY	Some firms could abate pollution at relatively low costs
		MEASURE	The 1976 BARCON requires parties to take all appropriate measures to prevent and abate pollution caused by dumping from ships and aircraft
		COST	
MANIPULATION	combat control monitor		

Figure 5: Proposal for phraseology module related to the term *pollution* in EcoLexicon

Figure 5 shows the information that will be included in EcoLexicon's phraseology module. Under the term *pollution* within the subdomain of Environmental Law, the different lexical domains will be presented with the verbs identified by corpus analysis. When clicking on each verb, the second argument categories will be shown and, when clicked on, example sentences that illustrate these verbs and arguments will also appear.

4. Conclusions

The results described in this paper show that Frame-based Terminology provides the methodological underpinnings to extract the subtle differences between Environmental Science and its subdomains at the linguistic level. Specifically, verbal collocations in the Environmental Law domain differ from those in the Environmental Science domain in regard to the specificity of the arguments. These differences must be included in terminological knowledge bases in order to provide an accurate representation of environmental knowledge. Differences at the conceptual level pervade the linguistic level because of the choice of verbs and their arguments. Representing this

phraseological knowledge for all the terms in EcoLexicon in English and in Spanish will be one of the challenges for the future development of EcoLexicon.

5. Acknowledgements

This research was carried out as part of the projects PID2020-118369GBI00 and A-HUM-600-UGR20, funded by the Spanish Ministry of Science and Innovation and the Regional Government of Andalusia.

6. References

[1] P. Faber, A Cognitive Linguistics View of Terminology and Specialized Language, De Gruyter Mouton, Berlin/Boston, 2012.
[2] P. Faber, Frames as a Framework for Terminology, in: H. J. Kockaert, F. Steurs (Eds.), Handbook of Terminology, 1, John Benjamins Publishing Company, 2015, pp. 14–33.
[3] P. Faber, Frame-based Terminology, in P. Faber, M.C. L´Homme (Eds.), Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge, volume 23 of Terminology and Lexicography Research and Practice, John Benjamins, Amsterdam, 2022, pp. 353–376. doi:https://doi.org/10.1075/tlrp.23.16fab.
[4] P. Faber, R. Mairal Usón, Constructing a lexicon of English verbs. Berlin: Mouton de Gruyter, 1999.
[5] P. Faber, A. Reimerink, Framing Terminology in Legal Translation, International Journal of Legal Discourse, 4.1 (2019) 15–46. doi:https://doi.org/10.1515/ijld-2019-2015.
[6] A. Reimerink, Pollution in Environmental Law: Comparative Corpus Analysis, International Journal of Lexicography, ecab027, 2021. doi:https://doi.org/10.1093/ijl/ecab027.
[7] P, Faber, P. León-Araúz, A. Reimerink, EcoLexicon: New Features and Challenges, in: I. Kernerman, I. Kosem Trojina, S. Krek, L. Trap-Jensen (Eds.), GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference, Portoroz, 2016, pp. 73–80.
[8] P. León Araúz, Representación Multidimensional del Conocimiento Especializado: El Uso de Marcos desde la Macroestructura hasta la Microestructura, Ph.D. thesis, Universidad de Granada, Granada, Spain, 2009.
[9] M. Buendía Castro, Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources, Ph.D. thesis, Universidad de Granada, Granada, Spain, 2013.
[10] S. Montero Martínez, M. Buendía Castro, Clasificación semántica de colocaciones verbales para la adquisición y codificación de conocimiento experto: El caso de los riesgos naturales, Revista Española de Lingüística Aplicada, 30.1 (2017) 240–272.
[11] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel, The Sketch Engine: Ten Years on, Lexicography 1.1 (2014) 7–36.
[12] P. León-Araúz, A. San Martín, P. Faber, Pattern-based Word Sketches for the Extraction of Semantic Relations, in: Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016), COLING 2016, Osaka, Japan, 2016, pp. 73–82.
[13] M. Jakubíček, A. Kilgarriff, V. Kovářr, P. Rychlý, V. Suchomel, The TenTen Corpus Family, in: 7th International Corpus Linguistics Conference CL 2013, Lancaster, 2013, pp. 125–127.

Annex

Table 1
Word Sketch: first 25 verbs with *pollution* as object in enLaw and EEC

	enLaw			EEC	
Collocate	Freq	Score	Collocate	Freq	Score
2997	21.930		918	16.130	
control	299	10.940	air	27	9.670
cause	510	10.860	prevent	51	8.930
prevent	265	10.530	reduce	142	8.440
combat	138	10.300	control	44	8.380
reduce	243	10.120	cause	90	7.660
eliminate	95	9.740	minimize	13	7.580
air	41	8.760	combat	6	7.530
address	91	8.710	abate	5	7.450
regulate	58	8.530	eliminate	10	7.420
avoid	46	8.370	emit	12	7.300
minimise	30	8.170	avoid	10	6.960
abate	24	7.970	address	9	6.800
concern	60	7.880	regard	9	6.770
emit	25	7.860	create	18	6.650
limit	27	7.430	limit	11	6.410
regard	28	7.340	see	52	6.150
produce	26	7.310	increase	36	6.140
generate	20	7.250	indicate	11	5.960
tackle	15	7.200	generate	14	5.920
mitigate	14	7.040	monitor	5	5.920
minimize	14	7.020	associate	16	5.820
include	64	7.000	decrease	6	5.790
define	18	6.790	include	20	4.820
increase	22	6.720	produce	12	4.740
cover	15	6.520	consider	6	4.230

Table 2
Word Sketch: first 25 verbs with *pollution* as subject in enLaw and EEC

	enLaw			EEC	
Collocate	Freq	Score	Collocate	Freq	Score
1724	12.610		566	9.950	
cause	126	9.710	flush	18	9.830
affect	86	9.560	destroy	5	7.460
originate	34	9.150	affect	20	6.960
occur	46	8.680	reduce	8	6.660
arise	42	8.540	result	8	6.450
result	33	8.480	include	17	5.680
include	51	8.040	increase	8	5.670
pose	15	7.680	cause	13	5.650

damage	9	7.320	become	11	5.550
be	671	7.250	take	6	5.510
come	13	7.190	lead	5	5.390
emanate	8	7.150	occur	10	4.900
control	9	7.130	do	8	4.740
contribute	10	7.090	have	51	4.600
remain	14	7.060	be	226	4.030
derive	8	7.020			
permit	9	7.000			
impact	7	6.930			
continue	10	6.930			
threaten	7	6.830			
take	17	6.800			
follow	11	6.790			
harm	6	6.750			
have	142	6.650			
enter	7	6.570			

Table 3
Word Sketch: Word Sketch: first 25 *polluter object_of* in enLaw and EEC

	enLaw			EEC	
Collocate	Freq	Score	Collocate	Freq	Score
346	21.750		13	13.270	
prosecute	20	10.270	divorce	1	11.090
sue	11	9.390	enshrine	1	10.410
deter	6	8.710	motivate	2	8.560
excuse	4	8.460	ascertain	1	8.540
oblige	10	8.430	hold	2	4.990
force	9	8.380	become	1	3.070
order	5	7.970	apply	1	2.550
compel	4	7.860	allow	1	2.440
let	3	7.760			
police	3	7.700			
pay	10	7.560			
allow	17	7.250			
get	3	7.040			
identify	12	7.030			
undermine	3	6.810			
locate	3	6.600			
find	6	6.350			
apply	6	6.230			
incorporate	3	6.220			
encourage	4	6.210			
bring	6	6.140			
require	21	6.030			
regulate	3	5.310			

implement	4	5.120
see	3	4.710

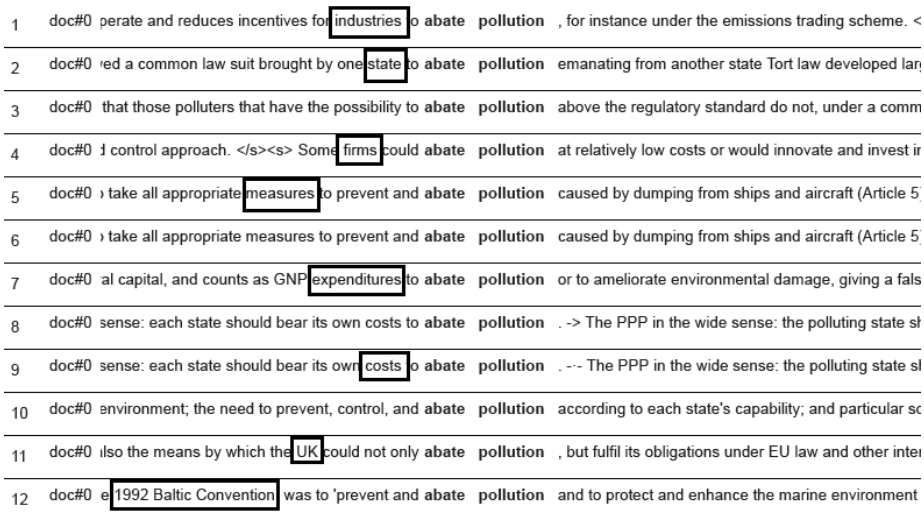

Figure 1: Extract concordances *abate* + *pollution* in enLaw

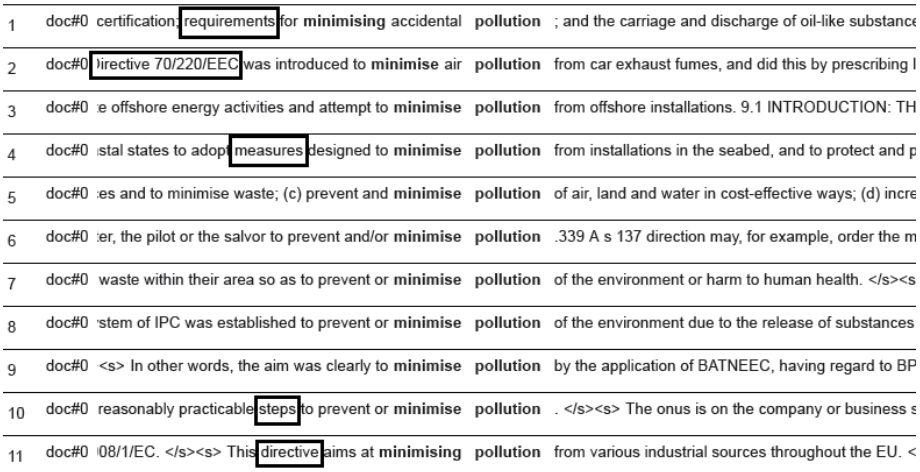

Figure 2: Extract concordances *minimise* + *pollution* in enLaw

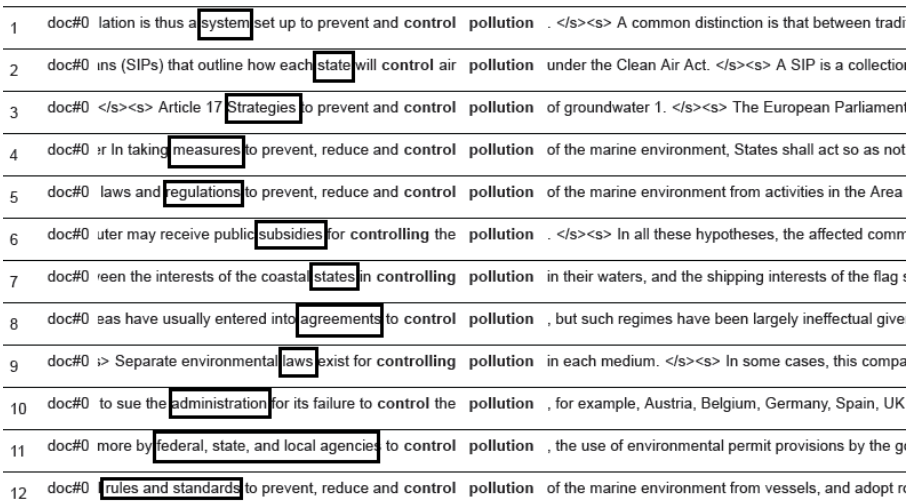

Figure 3: Extract concordances *control* + *pollution* in enLaw

Figure 4: Extract concordances *pollution caused_by* in enLaw