CORPUS

Corpus

18 | 2018
Les petits corpus

Discourse on climate and energy justice: a comparative study of Do It Yourself and Bootstrapped corpora

Camille Biros, Caroline Rossi and Inesa Sahakyan

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/corpus/3376
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Camille Biros, Caroline Rossi and Inesa Sahakyan, « Discourse on climate and energy justice: a comparative study of Do It Yourself and Bootstrapped corpora », *Corpus* [Online], 18 | 2018, Online since 24 September 2018, connection on 08 September 2020. URL : http://journals.openedition.org/corpus/3376

Discourse on climate and energy justice: a comparative study of Do It Yourself and Bootstrapped corpora

Camille Biros, Caroline Rossi and Inesa Sahakyan

1 The concept of environmental justice was coined in the United-States to refer to racial minorities being impacted unfairly by environmental degradation (Bullard 1990). Recently, two hyponyms of this term created by analogy have appeared in environmental debates: climate justice and energy justice. Like the term they are derived from, they are used to reflect on inequalities between different populations when facing environmental problems. They signify a will to build an energy transition to a low-carbon future that is inclusive of minorities and fair in its outcome. However, because these terms have been coined relatively recently, because they refer to complex ethical questions, and because they are being taken up across disciplines, a certain amount of controversy exists surrounding their use and definitions. It seems essential to clarify the policy objectives these terms suggest (Heffron & McCauley 2017: 7). Our aim is to shed light on how climate and energy justice are elaborated by different discourse communities. Our approach is variationist as we contrast the treatment of these questions in three sub-corpora representing Non-Governmental Organisations, United-Nations and the Renewable Energy Sector (RES) i.e. companies and representative institutions. We seek to characterise the specificities of the treatment of this issue by each of these discourse communities, according to their different objectives. This can only be done through the constitution of corpora which are specific to each discourse community. We began by constituting a DIY corpus, gathering reports and newsletters identified through a web-search by organisations representative of each of these communities. Further details on the methodology that guided our endeavour are given in what follows. The corpora obtained are relatively small corpora, based on three distinct discourse communities. The limitation in size is linked with the time-consuming nature of our methodology and the small length of the documents published by two of our discourse communities. We decided to extend the corpus by adding the content of automatically-retrieved, relevant web-pages. We used

BootCat to create three sub-corpora from the web, and we hypothesised that by using distinctive keywords as seeds, they would be representative of our three discourse communities. Beyond the characterisation of the specialised lexis of each of these discourses, this work enables us to compare the results between a DIY corpus and a corpus bootstrapped from the Web. The overarching question is whether the characterisation of linguistic specificities in each discourse community under review is possible through a web-based automatically generated corpus or whether small DIY corpora offer more reliable tools for our objectives.

1. The DIY Corpus

1.1. Methodology

2 The first step in the constitution of our corpus was to identify different discourse communities concerned by the concepts of "climate justice" and "energy justice". We base our understanding of a discourse community on Swales' text-based concept, broadly defined as communities with "a broadly agreed set of common public goals [...] mechanisms of intercommunication among its members" and resorting to "one or more genres in the communicative furtherance of its aims" as well as "some specific lexis" (Swales 1990: 24-27). The concepts of "climate justice" and "energy justice" come from the activist community. The first Climate Justice Summit took place in The Hague in 2000. It coincided with the Conference of the Parties 6 (COP6). The Energy Justice Network was created in 1999 to bring together NGOs with a particular focus on how the energy transition could be acquired with more concern for distribution of burdens and benefits (Heffron & McCauley 2017: 2). Because of the activist origin of the terms, the first discourse community we focused upon was that of Non-Governmental Organisations. We consider their discourse through reports in which they deal with this issue.

3 The second discourse community that seemed to us particularly relevant to consider these issues was that of UN institutions, because of their role in negotiating a binding international agreement to fight climate change. The fact that the term "climate justice" appears in the Paris Agreement, tended to confirm the interest of this discourse community for the question. UN organisations also seemed a good option to show contrasting views with NGOs. During the COP 15 conference in particular, the high level of activist mobilisation and the weak outcome of official negotiations led to a clash between the two with lasting effects in terms of opposition in communication strategies and objectives (Dahan & Aykut 2015: Chapter 7). We consider this discourse community through reports published by different United-Nation organisations on the theme of climate change and its impact on human welfare.

4 The third discourse community we identified as having an essential role to play to define and work on these concepts is that of the Renewable Energy Sector (RES). Companies, and institutions representing them, may use climate and energy justice as arguments to promote their products by contrasting them with conventional energy sources that produce a large amount of carbon dioxide. They also may show a willingness to differentiate themselves from big energy companies who have little regard for social impacts by showing an interest in developing fair sources of energy.

5 To constitute each sub-corpus representative of each discourse community, we began by using our prior knowledge of the issue, based on "horizontal reading" (Rühlemann & Aijmer 2014) of sources that we did not include in the corpus but kept as secondary sources, to focus on the discourse of organisations that were most likely to communicate on these issues. A simple Internet query with the keywords and the organisations thus identified enabled us to gather a first series of documents. However, our aim was to gather not only documents that used these terms but also those that dealt with the issue without using the terms. Indeed, as these terms come from the activist community, it seemed likely that the UN and RES would be less willing to use them. To extend the list of keywords for our query we used a list of definitions extracted from the documents gathered in the first stage, but this time we read them "vertically" (*ibid.*) using the concordance *Antconc* (Anthony 2014). From this sub-corpus of definitions, we were able to find new keywords to extend our corpus. We completed this work with defining extracts from academic publications (Rhaman 2016, Heffron *et. al.* 2015, Nicholson & Chong 2011, Heffron & McCauley 2017). In each definition, a semantic analysis enabled us to single out associated terms. Here are two examples of defining extracts and their use.

> (1) "to deepen the understanding of climate justice which, at the nexus of climate change, development and human rights, seeks to ensure that climate action is fair and people-centred" (*Mary Robinson Foundation Climate Justice Annual Report 2014*: 3).
> (2) "a range of groups are even now beginning to strategically utilize human right institutions, practices and discourses under the umbrella of "climate justice", in debates about climate change" (Nicholson & Chong 2011: 122).

6 In definition (1), the words "fair and people-centered" associated to "climate action" are presented as close synonyms of "climate justice". In both definitions (1) and (2), "human rights" are associated to "climate justice": in definition (1) there is a centre/periphery relation while in definition (2) human rights are associated to climate justice as a means to an end. This semantic analysis of definitions enabled us to identify the following terms as linked to climate and energy justice: human rights, fairness, equality, equity, accessibility, safety, distribution, people-centered, representation. The corpus we gathered with this method presents the following characteristics.

Table 1. DIY Corpus created through vertical and horizontal methodologies

Source	Number of Reports	Number of Words	List of Organisations (number of documents collected)
UN organisations	11	1,317,576	IPCC (2), REDD+ (2), UNEP (6), WorldBank (1)
NGOs	49	661,559	350.org (2) Action 2015 (1), CAN (2), Care (2), Climate Institute (1), Environmental Defence Fund (5), Environmental Justice Foundation (2), Greenpeace (5), Friends of the Earth (12), Mary Robinson Foundation (3), Oxfam (5), WWF (9),
Renewable Energy Sector	43	243,022	ACORE (3), Clean Choice Energy (2), Ebico (6), Ecotricity (1), Good Energy (4), GreenEnergy UK (9), irena (13), loco (3), REA (3)

7 To constitute each of these three corpora, the documents we gathered were mostly reports. Many of these were published around the Conference of the Parties in Paris in 2015. The discrepancy in sizes is explained by the fact that the UN publishes long reports whereas the documents in the RES and NGO corpora are much shorter. It was more difficult to find company publications on the topic. Although the RES corpus is quite small in terms of number of words, it seems representative as it includes

43 documents published by nine different organisations. We were attentive to maintain a certain generic uniformity and included only official material published by an identified organisation and dated. The problem of the discrepancy in sizes of the three corpora is somewhat toned down by the fact that most of the tools we adopt for our analysis use statistic tests (Log Likelihood Feature on *Antconc*, UCREL Semantic Analysis System on *WMatrix*) rather than raw counts. Besides, we added information on words per million to raw frequencies, so that comparison is possible across corpora of differing sizes.

1.2. Keyword analysis

8 Our object is to better understand how each discourse community conceptualises issues of climate and energy justice. With this in mind, we built the corpus using thematic and institutional criteria to identify meaningful documents. Although a more refined qualitative approach to the texts would be necessary to fulfil our objectives, in a first stage, we wanted to question the pertinence of our corpus. Keyword analysis (Baker 2004) offers great means for this. Using automatic software, keywords indeed make it possible to characterise the corpus broadly and check its thematic homogeneity. To identify keywords in the three sub-corpora we loaded a sample of the Corpus of Contemporary American English (Davies 2010) as a reference corpus in the software *Antconc* and used the Log-Likelihood (LL) feature[1]. We compared the fifty first keywords in each corpora to determine similarities and differences. The twenty most significant keywords for each are presented in Table 2 below.

Table 2. Keywords in each sub-corpus

Keywords in NGO corpus		**Keywords in UN corpus**		**Keywords in RES corpus**	
Key-ness	**Keyword**	**Key-ness**	**Keyword**	**Key-ness**	**Keyword**
29304.907	energy	36500.484	climate	16925.196	energy
10645.142	climate	27921.888	change	8253.902	renewable
10370.203	renewable	14874.951	adaptation	3643.826	electricity
7120.264	electricity	9850.471	impacts	3621.238	solar
6748.711	billion	7642.051	global	2799.404	power
6314.403	evolution	5263.802	changes	2379.888	pv
5752.826	power	4923.773	risk	2347.745	grid
5514.051	scenario	4613.720	development	2267.199	efficiency
5328.067	gas	4427.212	environmental	2147.299	renewables
5083.653	emissions	4352.179	confidence	2063.174	sector
5035.613	fossil	4105.441	water	2059.533	generation
4974.768	solar	4079.518	vulnerability	1928.637	projects
4792.705	heat	3991.846	chapter	1841.208	capacity
4495.632	global	3895.942	journal	1690.145	costs
4202.111	coal	3798.035	systems	1529.136	development
3875.741	biomass	3590.019	coastal	1519.042	irena
3762.596	total	3492.357	ocean	1486.247	cost
3700.559	generation	3468.763	temperature	1405.193	technologies
3655.399	geothermal	3438.598	energy	1402.704	africa
3561.308	demand	3414.469	high	1301.164	biomass

9 The colour code enables a quick comparison between the content of the list of the first fifty keywords for each corpus as they appear differently according to whether you find them in only one corpus, in all three, or in a combination of two. They appear in blue if present in the three corpora, in green if present in the NGO and RES corpora, in red if

present in the NGO and UN corpora, in yellow if present in UN and RES corpora and in black if specific to one corpus. Five keywords are common to the three corpora: CLIMATE, ENERGY, EMISSIONS, GLOBAL and DEVELOPMENT, i.e. which one can easily identify as pertaining to the field of climate change.

10 If you compare the colours in each list, the first striking feature is the proximity between the NGO and RES corpora as compared to the UN corpus. Especially if you take the first twenty keywords, you can see that a majority of keywords from the UN appear in black, while only six appear in black in the NGO corpus and five in the RES corpus. Furthermore, nine appear in green in the NGO corpus and twelve in the RES corpus. Those in green tend to be lexical items linked to the energy sector and the production of electricity.

11 To delve deeper into the lexical fields that are present in each corpus, it may be useful to turn to semantic differences. We used the semantic tagging feature available in the *WMatrix* (Rayson 2008). Based on the UCREL Semantic Analysis System[2] (USAS), this software enables researchers to identify key semantic domains in a corpus by comparing it to a corpus of reference[3].

1.3. Semantic Categories and Lexical Units

1.3.1. Overview

12 The results obtained concerning the first twenty-five semantic categories in each corpus are presented in Table 3 below. As in the previous part, the colour code enables a comparison of the content of the three corpora.

Table 3. Semantic categories in each sub-corpus

Semantic Categories in NGO corpus		Semantic Categories in UN corpus		Semantic Categories in RES corpus	
LL	**Semantic Tags**	**LL**	**Semantic Tags**	**LL**	**Semantic Tags**
18829.26	Numbers	34375.26	Unmatched	7647.21	Interested/excited/ energetic
16613.12	Unmatched	33423.61	Numbers	3556.77	Money and pay
8332.99	Interested/excited/ energetic	30492.22	Personal names	3077.01	Time: New and Young
6949.80	Weather	21991.49	Weather	2609.16	Electricity and electrical equipment
5441.51	Electricity and electrical equipment	20673.61	Change	2348.50	Money: Cost and price
4394.23	Substances and materials: Gas	12827.22	Geographical terms	1371.25	Business: Selling
4281.02	Quantities	8499.59	Green issues	1330.71	Places
4206.56	Change	6923.14	Cause&Effect/ Connection	1158.47	General actions/making
3913.44	Substances and materials generally	5624.04	Geographical names	1027.75	Wanted
3565.09	Geographical terms	4453.04	Science and technology in general	981.03	Helping
2496.69	Places	4285.03	Danger	955.62	Giving
2367.84	Time: New and young	3789.77	Places	873.44	Change
2258.12	Money generally	3144.32	Education in general	849.10	Usefulness
2255.32	Temperature: Hot/on fire	3005.12	Farming & Horticulture	827.80	Substances and materials generally
2228.62	Money and pay	2604.76	Temperature	791.84	Weather
2159.84	The universe	2160.48	Substances and materials: Liquid	752.36	Able/intelligent
1884.56	Substances and materials : Liquid	2043.54	Other proper names	717.36	The universe
1742.32	Science and technology in general	1964.86	Weak	677.48	Money generally
1718.91	Green issues	1854.01	Confident	660.54	Mental object: Means, method
1620.53	Giving	1821.90	Measurement: Size	617.15	Green Issues
1393.52	Substances and materials: Solid	1800.48	Temperature: Hot/on fire	610.02	Using
1277.74	Using	1740.20	Investigate, examine, test, search	568.56	Science and technology in general
1272.20	Geographical names	1732.65	Quantities: Little	566.08	Government
1186.73	Industry	1731.75	Quantities: Many, much	551.94	Cheap
1142.99	Belonging to a group	1445.09	Entire; Maximum	546.97	Measurement: Size

13 One can start by stating that one fifth of these lists is common to all three corpora, i.e. the following five categories: “Weather”, “Change”, “Place”, “Green issues” and “Science and Technology”. This tends to confirm the focus of our corpora on climate change, which is obviously related to the weather, describes a type of change, considers effects of change in different places, and can be labelled as a green issue. Science and technology are central to discourse on climate change both to better its understanding and to find solutions to fight it. A phenomenon which was already apparent in the comparison of keywords appears with much more clarity in the comparison of semantic categories: the proximity of the NGO corpus to the other two and the strong variation between the UN and RES corpora. The NGO corpus shares nearly all of its semantic categories with at least one other corpus, with only four semantic categories that are specific to this corpus. It could be seen as a combination of categories essential to the UN corpus and of categories essential to the RES corpus.

14 The high number of lexical items linked to the energy sector in the NGO and RES corpus is confirmed when looking at the categories in green. Among the first are “Interested / excited / energetic” and “electricity and electrical equipment”. Of course the polysemy of the lexical unit ENERGY clearly appears in the first category where it is associated with lexical units like INTEREST, ACTIVE, ENTHUSIASM and KEEN. However, one can readily check by looking at the list of lexical units in this category that its importance is mainly due to the presence of an important lexical field of the energy sector. In our corpora, the lexical unit ENERGY is mainly used as a reference to production of electricity rather than as a state of mind. A focus on economic matters also seems prevalent in these two corpora with the categories “Money and Pay” and “Money Generally”. The categories “Giving” and “Using” suggest a pragmatic approach. Overall, the impression that these two corpora offer a concrete focus on the energy transition dominates.

15 Conversely, the categories shared by the NGO and UN corpora suggest a descriptive approach of the effects of climate change with “Geographical Terms” and “Geographical Names”, “Temperature”, “Quantities” and “Numbers”. The unmatched category appearing right at the beginning reveals a high number of technical terms, neologisms and references to institutions and organisations. With respect to categories specific to the UN corpus, the focus on description of problems is confirmed with the categories “Danger” and “Weak”. The abstract character of this corpus appears with the categories “Cause&Effect / Connection” and “Investigate, Examine, Test, Search”. The category “Confident” is also significant in the UN corpus. Indeed, our prior knowledge of the corpus enables us to assert that the expression “degrees of confidence” is used to qualify the certainty and uncertainty of statements on the possible effects of climate change, namely in the International Panel on Climate Change publications. A quick search in *Antconc* retrieves 54 concordances and confirms that the term is present throughout all four IPCC5 reports.

16 To compare the content of the three corpora using the semantic tags, one can also go into more detail and look at the lexical units that compose the semantic categories in each corpus to see if they are related. In the next sections, we will compare the lexical units found in the five categories shared by each sub-corpus.

1.3.2. Lexical Units in categories shared by the three sub-corpora

17 The first semantic category shared by the three sub-corpora is that of "Weather". The list of the thirty most used lexical units in this category appears in Table 4, which gives raw counts of word frequency (w.f.) with words per million (wpm) in brackets.

Table 4. The most used lexical items in the category "Weather"

NGO Corpus		UN Corpus		RES Corpus	
Word	W.f. (wpm)	Word	W.f. (wpm)	Word	W.f. (wpm)
climate	4,451(6,728)	climate	1,8467 (14,016)	climate	337 (1,386)
wind	1,136 (1,717)	climatic	1,162 (881)	wind	273 (1,123)
weather	78 (117)	drought	736 (558)	weather	12 (49)
storm	65 (98)	weather	667 (506)	droughts	7 (28)
drought	45 (68)	flood	632 (479)	drought	7 (28)
rainfall	28 (42)	rainfall	517 (392)	storm	5 (20)
flooding	25 (37)	flooding	430 (326)	rainfall	4 (16)
floods	24 (36)	floods	415 (314)	flood	4 (16)
droughts	23 (34)	storm	286 (217)	sunny	3 (12)
cloud	23 (34)	wind	278 (210)	storms	3 (12)
storms	20 (30)	meteorological	269 (204)	floods	2 (8)
meteorological	14 (21)	snow	241 (182)	flooding	2 (8)
clouds	14 (21)	droughts	225 (170)	climatic	2 (8)
climatic	13 (19)	storms	156 (118)	snow	1(4)
rain	13 (19)	climates	134 (101)	climates	1(4)
winds	13 (19)	monsoon	110 (83)	inclement	1(4)
flood	12 (18)	inundation	96 (72)	draught	1(4)
sunny	10 (15)	winds	68 (51)	sunnier	1(4)
snow	9 (13)	rain	50 (37)	wind_companies	1(4)
climates	8 (12)	hurricane	38 (28)	wind_market	1(4)
wind_market	7 (10)	hurricanes	32 (24)	wind-farm	1(4)
climate-	5 (7)	snowfall	24 (18)	climate-	1(4)
rainwater	5 (7)	cloud	22 (16)	windstorms	1(4)
cloudy	5 (7)	floodplains	20 (15)	rain	1(4)
climate.	4 (6)	humid	20 (15)	rainfalls	1(4)
flooded	4 (6)	rainy	18 (13)	wind_mills	1(4)
windy	4 (6)	clouds	18 (13)	wind_sector	1(4)
cloud_companies	4 (6)	climate_factors	16 (12)	humid	1(4)
rains	3 (4)	avalanche	16 (12)	cloud	1(4)
floodplains	3 (4)	weather_conditions	14 (10)	mist	1(4)

18 What is striking with this list is the uniformity between the different corpora when treating this theme. The fifteen first lexical units of each appear in the three corpora, with only two exceptions. From sixteen onwards, there is more diversity. The table also confirms the links already observed between the NGO corpus and the two others and the greater discrepancy between the UN and RES corpora. The bigger focus on energy markets in the NGO and RES corpora is confirmed when you see that the two lexical units specific to these two are SUNNY, probably as a reference to the use of solar energy, and WIND MARKET. There are even more lexical units specific to the energy market in the RES corpus: SUNNIER, WIND COMPANIES, WIND FARM, WIND MILLS and WIND SECTOR. The lexical units that are specific to the UN corpus suggest more focus on the negative outcomes of climate change as many are evocative of extreme events: MONSOON, HURRICANE(S), SNOWFALL and AVALANCHES. In the UN corpus we significantly find technical terms like CLIMATE FACTOR and WEATHER CONDITIONS.

19 Another category that shows a strong degree of uniformity between the three corpora is that of "Green Issues", presented in table 5.

Table 5. The most used lexical items in the "Green Issues" category

NGO Corpus		UN Corpus		RES Corpus	
Word	W.f. (wpm)	Word	W.f. (wpm)	Word	W.f. (wpm)
environmental	651 (984)	environmental	2,386 (1810)	environmental	125 (514)
environment	328 (495)	ecosystems	1,868 (1417)	environment	66 (271)
pollution	179 (270)	environment	1,227 (931)	conservation	42 (172)
nature	164 (247)	ecosystem	1,1 54(875)	energy_ resources	41 (168)
ecosyst-ems	136 (205)	nature	1,015 (770)	EPA	32 (131)
conserva-tion	108 (163)	ecology	932 (707)	energy_policy	26 (106)
defores-tation	97 (146)	ecological	674 (511)	air_pollution	26 (106)
energy_policy	62 (93)	conservation	600 (455)	pollution	19 (78)
energy_resources	42 (63)	pollution	246 (186)	energy_con-servation	18 (74)
air_pollution	39 (58)	deforestation	220 (166)	nature	18 (74)
EPA	36 (54)	environments	143 (108)	environmen-tally	16 (65)
ecological	35 (52)	air_pollution	114 (86)	energy_saving	12 (49)
polluting	29 (43)	soil_erosion	60 (45)	polluting	10 (41)
ecosys-tem	28 (42)	energy_policy	48 (36)	greening	9 (37)
environ-mentally	23 (34)	greening	40 (30)	deforestation	8 (32)
EPAs	15 (22)	desertification	36 (27)	ecosystems	6 (24)
environments	14 (21)	EPA	24 (18)	EPAs	5 (20)
agro-ecological	10 (15)	naturalist	12 (9)	environments	5 (20)
ecology	8 (12)	environmentally	10 (7)	ECO	4 (16)
environmentally_friendly	8 (12)	ecologically	8 (6)	desertification	4 (16)
nature.	8 (12)	ecosystem_types	8 (6)	environmentally_friendly	4 (16)
ecologically	7 (10)	greenness	8 (6)	polluted	2 (8)
energy_saving	7 (10)	natures	6 (4)	ecosystem	2 (8)
energy_conservation	6 (9)	conservation_areas	6 (4)	nature.	1 (4)
environmentalists	5 (7)	ecosystem_based	4 (3)	soil_erosion	1(4)
pollute	5 (7)	environmentally_friendly	4 (3)		
conservancy	4 (6)	agro-ecological	4 (3)		
environmentalist	4 (6)	Palaeoecology	4 (3)		
greening	3 (4)	ecosystems(c)46food	2 (1)		
conservationists	3 (4)	energy_conservation	2 (1)		

20 In this list, seventeen lexical units are present in the three corpora. The link between the NGO corpus and the two others is confirmed. Surprisingly, the semantically-linked lexical units DESERTIFICATION and SOIL EROSION do not appear in the NGO corpus as they do in the two others.

21 There is also a strong degree of uniformity in the lexical units included in the category “Places”, presented in Table 6.

Table 6. The most used lexical items in the "Places" category

NGO corpus		UN corpus		RES corpus	
Word	W.f. (wpm)	Word	W.f. (wpm)	Word	W.f. (wpm)
countries	1507 (2277)	areas	1,622 (1,231)	local	361 (1,485)
international	1,194 (1,804)	regional	1,517 (1,151)	countries	287 (1,180)
local	587 (887)	regions	1514 (1,149)	areas	201 (827)
national	587 (887)	international	1486 (1,127)	national	194 (798)
district	418 (631)	local	1,061(805)	city	178 (732)
regions	337 (509)	tropical	956 (725)	regional	140 (576)
regional	328 (495)	urban	900 (683)	international	93 (382)
areas	299 (451)	countries	765 (580)	area	69 (283)
developing_countries	296 (447)	region	741(562)	cities	62 (255)
region	263 (397)	indigenous	434 (329)	municipal	60 (246)
area	126 (190)	area	428 (324)	urban	53 (218)
globally	114 (172)	national	400 (303)	regions	50 (205)
indigenous	112 (169)	cities	349 (264)	region	40 (164)
cities	85 (128)	globally	259 (196)	developing_countries	37 (152)
urban	79 (119)	developing_countries	253 (192)	county	33 (135)
base	76 (114)	zones	214 (162)	villages	29 (119)
foreign	70 (105)	settlements	199 (151)	sites	26 (106)
nationally	61(92)	zone	180 (136)	base	24 (98)
city	54 (81)	locations	176 (133)	zones	23 (94)
sites	49(74)	antarctic	146 (110)	site	19 (78)
places	45(68)	boundary	125 (94)	foreign	19 (78)
location	41 (61)	location	118 (89)	municipalities	18 (74)
tropical	38 (57)	sub-regions	118 (89)	landfill	17 (69)
locations	36 (54)	places	109 (82)	globally	16 (65)
village	35 (52)	sites	100 (75)	outreach	15 (61)
borders	35 (52)	park	86 (65)	location	15 (61)
internationally	34 (51)	campus	82 (62)	locations	15 (61)
site	33 (49)	boundaries	72 (54)	cross-border	13 (53)
municipal	29 (43)	place	68 (51)	district	11 (45)
developing_country	22 (33)	native	66 (50)	town	11 (45)

22 Here, sixteen lexical units out of thirty are common to the three corpora. It confirms the greater proximity of the NGO corpus to the two others. MUNICIPAL and CITY, common to the NGO and RES corpora, could be linked to an interest for the local scale where decisions about energy are or could be taken. BASE and SITE could refer to sites of energy production. The lexical unit INDIGENOUS, common to the NGO and UN corpora, confirms an interest for negative impacts of climate change as indigenous people are often quoted as victims, in particular in the context of deforestation.

23 Although the lexical field of “Change” is central in the three corpora, there is more diversity in ways of expressing it as one can observe in the list of lexical units presented in Table 7.

Table 7. The most used lexical items in the "Change" category

NGO corpus		UN corpus		RES corpus	
Word	W.f. (wpm)	Word	W.f. (wpm)	Word	W.f. (wpm)
change	2,283 (3,450)	change	16,599 (12,598)	development	481 (1,979)
development	1,206(1,822)	adaptation	5,628 (4,271)	change	235 (966)
evolution	844 (1,275)	changes	3,653(2,772)	develop	102 (419)
adaptation	812 (1,227)	development	2,424 (1,839)	transition	73 (300)
transition	318 (480)	adaptive	725 (550)	developing	69 (283)
fluctuating	252 (380)	changing	710 (538)	developed	60 (246)
develop	202 (305)	shifts	443 (336)	become	56 (230)
changes	192 (290)	affected	443 (336)	changes	53 (218)
become	179 (270)	affect	436 (330)	hybrid	41 (168)
developing	168 (253)	adapt	317(240)	transformation	38 (156)
developments	153 (231)	become	296 (224)	changing	36 (148)
developed	147 (222)	evolution	284 (215)	developments	28 (115)
transformation	136 (205)	occur	271 (205)	becoming	26 (106)
affected	130 (196)	developing	226 (171)	became	24 (98)
shift	104 (157)	adapting	210 (159)	reform	23 (94)
adapt	85 (128)	developed	197 (149)	transform	18 (74)
changing	82 (123)	shift	148 (112)	experienced	17 (69)
conversion	66 (99)	transformation	146 (110)	affect	17 (69)
reform	66 (99)	occurrence	136 (103)	replacement	16 (65)
happen	58 (87)	adaptations	132 (100)	switch	14 (57)
hybrid	55 (83)	restoration	132 (100)	momentum	13 (53)
occur	50 (75)	occurred	130 (98)	replace	13(53)
becoming	49 (74)	evolutionary	128 (97)	adjustments	13(53)
replace	47 (71)	altered	125 (94)	switching	13(53)
substitution	43 (64)	alter	120 (91)	adapt	12(49)
became	42 (63)	transition	120 (91)	shift	11(45)
transforming	36 (54)	changed	118 (89)	adapted	11(45)
affect	35 (52)	affecting	112 (85)	adjusted	11(45)
converter	35(52)	oscillation	108 (81)	replaced	11(45)
becomes	32 (48)	occurring	107 (81)	replacing	11(45)

24 The link between the NGO corpus and the two others is confirmed with six lexical units common to the NGO and RES corpora and four common to the NGO and UN corpora.

25 The category “Science and Technology” also presents contrasting lists of lexical units with only ten common to the three corpora as can be seen in Table 8.

Table 8. The most used lexical items in the "Science and Technology" category

NGO corpus		UN corpus		RES corpus	
Word	W.f. (wpm)	Word	W.f. (wpm)	Word	W.f. (wpm)
technologies	757(1,144)	science	1,517(1,151)	technologies	278(1,143)
technology	527(796)	sciences	818(620)	technology	167(687)
nuclear	502(758)	biology	778(590)	technical	90(370)
technical	233(352)	scientific	578(438)	technological	21(86)
science	109(164)	geophysical	451(342)	technically	14(57)
scientific	68(102)	technical	359(272)	laboratory	13(53)
technological	62(93)	hydrology	332(251)	engineers	11(45)
scientists	47(71)	technology	330(250)	nuclear	10(41)
radiation	45(68)	oceanography	216(163)	engineering	10(41)
neodymium	41(61)	technological	186(141)	scientific	9(37)
tellurium	40(60)	environmental_ science	170(129)	radiation	8(32)
engineering	39(58)	technologies	166(125)	evs	8(32)
nd	29(43)	ph	152(115)	scientists	7(28)
radioactive	26(39)	climatology	151(114)	lab	6(24)
technically	23(34)	s.e.	140(106)	science	6(24)
reactors	22(33)	engineering	120(91)	formula	5(20)
scientist	20(30)	experiments	114(86)	LR	5(20)
observatory	16(24)	radiation	108(81)	n.d	5(20)
NPS	16(24)	Biogeography	108(81)	laboratories	4(16)
se	14(21)	Epidemiology	104(78)	electronic_ engineers	3(12)
technology_ transfer	14(21)	chemistry	94(71)	tech	3(12)
state_of_the_art	13(19)	meteorology	89(67)	technicians	3(12)
tes	13(19)	laboratory	87(66)	technology_ transfer	3(12)
reactor	13(19)	T.E	86(65)	nd	3(12)
sciences	12(18)	psychology	84(63)	engineer	2(8)
tech	11(16)	scientists	75(56)	po	2(8)
state-of-the-art	10(15)	physiology	72(54)	ev	2(8)
ev	8(12)	marine_biology	64(48)	n.d.	2(8)
geophysical	8(12)	SES	50(37)	selenium	2(8)
selenium	8(12)	biota	42(31)	geophysical	2(8)

26 The proximity between the NGO and RES corpora is confirmed. The UN appears more specific here with the strong degree of abstraction confirmed by the presence of twelve lexical units referring to disciplinary fields. It suggests a stronger focus on science whereas the two others are dominated by references to technology (TECHNOLOGY TRANSFER, TECH, TECHNICALLY), in particular technology linked to the production of energy (SELENIUM[4], NUCLEAR).

1.4. Lexical Units on Justice

27 Although our starting point for the constitution of the corpus were the terms “climate justice” and “energy justice”, the semantic tagging feature in *WMatrix* does not identify this issue as key in our corpora. The thematic prevalence of climate change is clear but our interest for equity issues in relation to climate change, which led to our selection of sources, is not reflected in the main semantic categories. In the UN corpus, we find no significant semantic category linked to ethics or law. In the NGO corpus we find two. The category Lawful (G2.1), appears 69th in terms of Log Likelihood, and the category “Ethical” appears 119th. In the RES corpus, the category “Ethical” appears 110th. The *WMatrix* software does not really help us to compare the importance of ethical issues in our three corpora. This is the reason why we decided to use the *Antconc* word count feature to find out the number of occurrences of our lexical units of interest in the corpus. Starting from the list of keywords identified through the semantic analysis of definitions in our corpus (Section 1.1), we established frequency counts with *Antconc*: Table 9 gives raw counts, with words per million in brackets (wpm).

Table 9. Frequency counts with *Antconc*

Term of interest	UN	NGO	RES
	W.f. (wpm)	W.f. (wpm)	W.f. (wpm)
Environmental Justice	14(10)	32(48)	1(4)
Climate Justice	16(12)	236(356)	0(0)
Energy Justice	0	2(3)	0(0)
Human Rights	125(94)	649(981)	5(20)
Equity	538(408)	152(229)	19(78)
Inequity	23(17)	3(4)	0(0)
Equality	335(254)	98(148)	17(69)
Inequality	297(225)	48(72)	8(32)
Fairness	27(20)	19(28)	0(0)
Accessibility	53(43)	17(25)	1(4)
Safety	505(383)	77(116)	25(102)
Distribution	1,769(1342)	208(314)	105(432)
People-centered	2(1)	2(3)	0(0)
Representation	107(81)	23(34)	3(12)

28 The first striking feature is the very small word frequencies of our three main terms of interest in the UN and RES sub-corpora. The use of CLIMATE JUSTICE is high in the NGO corpus, suggesting a term that is specific to this activist discourse community. The second important feature is the high use of the alternative terms EQUITY and HUMAN RIGHTS in the UN corpus, terms which are also present in the NGO corpus. In the RES corpus, these terms only remain marginal but one may notice that EQUITY and EQUALITY are the most used. Of course, this simple word count does not enable us to conclude that in all cases where EQUITY is being used, issues linked to justice are being discussed. It is possible that the term is also being used in its economic meaning in our corpora. However, we were able to check that a significant number of occurrences referred to its ethical meaning. There is a high number of occurrences of DISTRIBUTION in the three corpora. Here too, the polysemy of the term tones down the interest of this simple word count. However, issues of distribution of burdens and benefits are central to climate and energy justice and the word count suggests the importance of this term in our corpora, whose collocations could be interesting to study in future research.

1.5. Significant Compounds and Clusters

29 When analysing the differences between semantic categories in the three different corpora, we were particularly interested by the fact that in the UN corpus, the category "Weak" and "Danger" featured among the first twenty-five. This tends to suggest a focus on the description of the negative consequences of climate change. The category "Weak" seems particularly significant in dealing with the issue of justice. Most of the lexical units included refer to victims of climate change. The lexical units that are most important in the "Weak" category in the UN corpus according to *WMatrix* are the

following: VULNERABILITY (2142), VULNERABLE (480), VULNERABILITIES (391), WEAK (52), SUSCEPTIBLE (40) and FRAGILE (10). Compared to the two other corpora, where these lexical units do not appear as central, it seems that the UN tends to focus on describing victims of climate change and explaining why they are victims. VULNERABLE and its morphological derivatives are most used for this purpose. To further understand the differentiated use of these lexical items in our three corpora, we used the N-Gram feature in *Antconc* to identify significant clusters and compare them in each corpus. Our hypothesis was that the two other corpora, with their pragmatic approach, would present more of an interest for identifying people responsible for climate change rather than identifying its victims. To consider this issue, we identified the adjective RESPONSIBLE and its derivatives in each sub-corpora following the same method. In the following tables, we consider the ten most significant N-grams in each corpus, with a minimum frequency of three.

Table 10. N-Grams starting with Vulnerab*

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	vulnerable countries	45(68)	vulnerability to climate change	301(228)	vulnerable clients	3(12)
2	vulnerable communities	43(64)	vulnerability and adaptation	163(123)		
3	vulnerable people	28(42)	vulnerability assessment	144(109)		
4	vulnerable to climate change	23(34)	vulnerable to climate change	117(88)		
5	vulnerability to climate change	17(25)	vulnerability and exposure	109(82)		
6	vulnerable groups	14(21)	vulnerability and adaptive capacity	102(77)		
7	vulnerable countries and communities	8(12)	vulnerability assessments	96(72)		
8	vulnerable populations	8(12)	vulnerable populations	81(61)		
9	vulnerable sections	8(12)	vulnerability mapping	58(44)		
10	vulnerable situations	6(9)	vulnerability reduction	57(43)		

30 This table confirms that the lexeme VULNERABLE is much more represented in the UN corpus, with many clusters starting with the adjective or noun. In the NGO corpus, we have some occurrences also but in the RES corpus hardly any. A specificity of the items in the UN corpus is that several refer to tools to measure vulnerability: VULNERABILITY ASSESSMENT(S) and VULNERABILITY MAPPING.

Table 11. N-Grams starting with Responsib*

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	responsibility and capability	16(24)	responsible environmental management	8(6)	responsibility to meet electricity demand	3(12)
2	responsible governance	6(9)	responsible fisheries	8(6)		
3	responsible investment	6(9)	responsible for changing landslide activity	6(4)		
4	responsibilities and respective capabilities	5(7)	responsibilities for climate adaptation	4(3)		
5	responsibility and capacity	5(7)	responsibility, capability and vulnerability	4(3)		
6	responsibility for emissions	4(6)				
7	responsible companies	4(6)				
8	responsible governance of tenure of land	4(6)				
9	responsibility and capability index	3(4)				
10	responsible energy	3(4)				

31 This table tends to show that the issue of responsibility is central in the NGO corpus compared to the two others. The idea is confirmed if we adopt a differentiated approach to the use of the lexical unit RESPONSIBLE, distinguishing its meaning as "taking care of" as opposed to "causing". Most of the occurrences of RESPONSIBLE in the UN table signify that someone is responsible as in "taking care of", whereas most of the occurrences in the NGO corpus are used to refer to causes of climate change. In the NGO corpus, we find a term to refer to a tool to measure responsibility: RESPONSIBILITY AND CAPABILITY INDEX. This analysis tends to confirm that there is more of a focus on measuring vulnerability in the UN corpus and more of a focus on measuring responsibility in the NGO corpus. These issues do not appear as central in the RES corpus.

32 Another interesting distinction we noted between our corpora thanks to the *WMatrix* categories is the fact that the NGO and RES corpora are much more focused on energy and its production. The second type of cluster that we decided to focus on aims to further our understanding of these differences. We consider the first ten occurrences of two-word units having the structure [energy + NOUN] and the first ten occurrences of two-word units having the structure [ADJECTIVE + energy].

Table 12. [ENERGY+NOUN] in the three sub-corpora

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	energy revolution	2,372(3,585)	energy efficiency	484(367)	energy efficiency	407(1674)
2	energy demand	834(1,260)	energy demand	242(183)	energy sector	102(419)
3	energy use	473(714)	energy systems	181(137)	energy use	101(415)
4	energy consumption	461(696)	energy sources	170(129)	energy projects	98(403)
5	**energy outlook**	444(671)	**energy policy**	159(120)	energy technologies	90(370)
6	energy efficiency	418(631)	energy consumption	132(100)	energy mix	78(320)
7	energy sector	284(429)	energy supply	113(85)	energy services	74(304)
8	energy sources	246(371)	energy production	111(84)	energy savings	66(271)
9	energy system	223(337)	energy agency	106(80)	energy system	65(267)
10	energy supply	215(324)	energy use	106(80)	energy demand	63(259)

33 Although the figures for this type of cluster are higher in the NGO and RES corpora, we find similar units in each. They show a willingness to calculate what is being produced and consumed, and to go towards an integrated and efficient system.

Table 13. [ADJECTIVE+ENERGY] in the three sub-corpora

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	renewable energy	1,872(2829)	district energy	778(590)	renewable energy	1,512(6,221)
2	world energy	781(1,180)	renewable energy	481(365)	clean energy	108(444)
3	final energy	609(920)	international energy	80(60)	sustainable energy	61(251)
4	advanced energy	472(713)	sustainable energy	75(56)	community energy	57(234)
5	ocean energy	455(687)	wind energy	62(47)	modern energy	50(205)
6	primary energy	288(435)	primary energy	50(37)	global energy	49(201)
7	clean energy	267(403)	global energy	45(34)	good energy	40(164)
8	basic energy	256(386)	solar energy	44(33)	geothermal energy	32(131)
9	global energy	192(290)	world energy	35(26)	national energy	29(119)
10	international energy	171(258)	national energy	30(22)	solar energy	27(111)

34 Table 13 confirms the higher number of occurrences of two-word units related to energy in the NGO and RES corpora. It shows RENEWABLE ENERGY is the term that is most used to refer to non-conventional sources of energy. The term CLEAN ENERGY is used by NGO and RES corpora. The term GOOD ENERGY is used only in the RES corpus, which may be linked to the fact that one of the companies we selected documents from for this corpus is named Good Energy. Specific types of energies are mentioned in the list for the three corpora but there is no distinct pattern to explain why some energy sources are being mentioned rather than others in each corpus.

35 On the whole, the first results concerning distinctive characteristics of our three sub-corpora tended to confirm their usefulness to offer a contrasting view of climate and energy justice, according to our three discourse communities. However, the time-consuming character of the corpus constitution and the discrepancy in sizes of our three sub-corpora led us to question whether a bootstrapping tool could be useful to extend our corpus. Of course, the generic specificity would not be entirely maintained with this method. However, some questions concerning terms, their use and their definitions could be considered more efficiently in an extended corpus. Our interest for using this tool also stemmed from a willingness to test its efficiency and determine its possible limits.

2. The BootCat Corpus

36 BootCat stands for "Bootstrapping Corpora And Terms" from the web (Baroni & Bernardini 2004). It is a freely available toolkit which was developed as a simple web-mining device to help translation students, translators and terminologists build specialised corpora. BootCat was thus originally aimed at "users who need relatively large and varied corpora (typically of about 1-2 million words), and who are likely to search the corpus repeatedly for both form- and content-oriented information within a single extended task." (Bernardini 2006). The resulting corpora have been called "DIY corpora" (Zanettin 2002) or "disposable corpora" (Varantola 2003). In this paper, data collection was not conducted primarily with a view to helping translators, and we sought to gather representative DIY corpora that could then be used for various purposes. Our BootCat corpora are thus somewhat different from our DIY corpora. They were designed as possible complements to these relatively small corpora. Bearing in mind the general advantages and limitations of using webpages as a corpus (i.e. "The web is a dirty corpus, but expected usage is much more frequent than what might considered as noise" Kilgarriff and Grefenstette 2003: 342), our overarching aim was to assess their usefulness for the analysis of specialised discourse. In this section we

describe how we went about gathering the data before assessing the relevance of BootCat corpora for a broader set of aims, which could be defined as corpus-based discourse analysis.

2.1. Methods to Build Three BootCat Corpora

37 BootCat starts from seed terms[5] that are automatically combined into tuples[6] to produce a series of web queries. Although the tuples cannot be controlled, there are at least two ways of discarding irrelevant results. The first one consists in carefully choosing distinctive keywords to be used as seeds. The second one has to do with URLs: BootCat allows users to establish restrictions as well as to manually sort automatically retrieved URLs before downloading the corpus.

2.1.1. Keywords

38 We chose to use BootCat within the SketchEngine (Kilgariff et al. 2014) because of recent limitations imposed on the number of free queries with the Bing search engine, and because it allowed us to see more clearly how many words had been retrieved from each URL[7]. This brought about new limitations: the SketchEngine has fixed parameters and can only take 20 seeds to create a corpus that will also be limited in size (1 million words). The parameters, however, were ideal since we wanted the size of our bootstrapped corpora to remain comparable to that of our three DIY corpora.

39 We defined two initial seeds that set the main themes (climate change and energy) and selected the remaining 18 from our comparative table of keywords. First we kept only those keywords that were distinctive of each of our DIY sub-corpus. Then we sorted them according to frequency and kept only the first 18 for each sub-corpus. We used frequency rather than keyness in order to make sure our selected keywords were present in a substantial number of documents (based on Baker 2004: 350-354). The resulting list is presented in Table 14 below.

2.1.2. URLs

40 In selecting URLs, we did not seek to achieve discourse homogeneity of the kind that was present in our DIY corpora. Manually checking that each and every retrieved URL corresponded to speakers belonging to the relevant discourse community would have taken us far beyond regular BootCat procedures, which were precisely what we wanted to test. In order to get rid of data that would not be relevant to any of our subcorpora, we excluded the whole of Wikipedia.org, and took out manually the least relevant sources such as the press and blogs. Table 14 shows the final number of URLs used to build our three subcorpora, and observed differences are worth noticing: longer documents were found for the first two subcorpora, and a quick look at URLs suggests that those documents are indeed longer reports of the kind that were gathered in our DIY corpora.

Table 14. The BootCat Corpora

Targeted discourse community	Keywords used as seeds (n = 20)	Number of URLs retrieved	Number of Words
UN organisations	climate change, energy, assessment, confidence, effects, environmental, high, impacts, low, management, medium, mitigation, ocean, regional, risk, section, temperature, vulnerability, warming, water	181	1,047,222
NGOs	climate change, energy, action, advanced, billion, "combined heat and power", coal, evolution, figure, hydrogen, international, lignite, non, nuclear, OECD, oil, plants, reference, sustainable, waste	197	1,063,213
Renewable Energy Sector	climate change, energy, access, deployment, for, grid, gw, local, market, Mexico, mw, potential, remap, rural, savings, sustainable, UK, USD, use, utility	230	1,063,986

2.2. Comparative Analysis of Results.

2.2.1. Semantic Categories.

41 As can be seen in Table 15 below, when it comes to semantic categories, the BootCat corpus reveals a far greater extent of similarity across the three sub corpora. New semantic categories common to all the three sub-corpora emerge in the BootCat data, such as “Cause and Effect Connection”, “Mental object”, “Wanted”, “Giving”. Semantic categories such as “Numbers”, “Substances and Materials”, “Unmatched” (in italics), which were previously common only to NGO and UN corpora, are identified as common to all the three BootCat sub-corpora thus suggesting greater uniformity across our corpus.

42 Conversely, the BootCat results are much scarcer when it comes to semantic categories common to NGO and UN sub-corpora, with only three results as compared to nine in the DIY corpus. Only one semantic category, highlighted in bold, is identified in both corpora. This tendency is further confirmed while analysing the semantic categories common to NGO and RES sub-corpora, nearly twice as many results emerge in the DIY corpus. Finally, while the DIY corpus revealed no results common only to UN and RES corpora, “Places” is identified as common to both.

Table 15. Comparative Analysis of DIY and BootCat Results for the Semantic Categories

Sub-corpora	DIY Corpus	BootCat Corpus
NGO+UN+RES	- **Weather** - **Change** - Places - **Science and Technology in general** - Green Issues	- *Numbers* - *Substances and materials: Gas* - *Unmatched* - **Science and technology in general** - **Change** - Cause & Effect/Connection - **Weather** - Giving - Mental Object: Means, method - Wanted - Using
NGO+UN	- Numbers - Unmatched - Geographical Terms - Geographical Names - **Substances and Materials: Liquid** - Measurement: Size - Temperature: Hot/on fire - Quantities/Quantities: Little	- **Substances and Materials: Liquid** - Green Issues - Temperature
NGO+RES	- **Interested/excited/energetic** - **Electricity and electrical equipment** - **Substances & Materials generally** - Time: new and young - Money generally - **Money and pay** - The Universe - Giving - Using	- **Interested/excited/energetic** - **Substances & Materials generally** - **Electricity and electrical equipment** - Quantities - Money: Cost and price - **Money and pay**
UN+RES	**[no results]**	**- Places**

43 Notwithstanding the varying number of identified semantic categories common to different sub-corpora, it can be noted that our DIY and BootCat corpora are coherent in terms of results inasmuch as the semantic categories identified in one were more often than not found in the other corpus too.

2.2.2. Lexical Units in categories shared by the three sub-corpora

44 To go into a more detailed analysis of results, we will consider the lexical units included in semantic categories common to the three sub-corpora. In a comparative purpose, we only present those which were also present in the three DIY corpora that is to say "Weather", "Change" and "Science and Technology in general".

45 While examining the BootCat results of the most used lexical items in the "Weather" semantic category, two features seem to be noteworthy. On the one hand, there is a striking coherence with 60 per cent of the lexical units (18 out of 30) being shared by all the three sub-corpora, while on the other, very few lexical units are common in combinations of two. Thus only two are shared by the B-NGO and B-UN corpora, HURRICANE and HURRICANES, which being the singular and plural forms of the same noun further reduce the specificity which could be attributed to this combination of sub-corpora. The same is true for the UN and RES, as well as the NGO and RES combinations, as they merely share the lexical units of INUNDATION and CLOUD, CLOUDS, weather conditions respectively. As a result, it can be concluded that while there is a strong degree of similarity across all three sub-corpora, each sub-corpus remains distinct enough. As for the features characteristic of the way each discourse community treats the issue of the "Weather", with lexical units such as HAZE, FOG, HUMID, DRAUGHT, CHOPPY, the B-NGO sub-corpus seems to place the focus on the quality of air, and possibly on the

negative impacts of atmospheric pollution. The major concern of the UN discourse, on the other hand, seems to be the impact of weather on local populations and agriculture with lexical units like FLOODPLAIN(S), FLOODED, HAILSTORM, HEATWAVES, AVALANCHE, INUNDATIONS, CHINOOK and RAINY specific to its corpus. Finally, the lexical units exclusively present in the RES list of the 30 most used units point to concerns with the development of renewable sources of energy (wind and solar) with MONSOON(S), SUNNY, WIND MARKET, WIND OUTPUT, WINDY, TORNADO, RAINS.

46 Comparing the above BootCat results with those of the DIY corpus (discussed in section 1), we could state that the thematic coherence of discourses dealing with the issue of climate change (as represented by the semantic tag "Weather") is largely confirmed with 56% (17 lexical units out of 30) and 60% of lexical units being shared by all three sub-corpora both in the DIY and the BootCat respectively. The links previously observed between the NGO corpus and the two others are further reinforced, whereas the discrepancy between the UN and RES corpora is scaled down by the presence of the lexical unit INUNDATION in both BootCat sub-corpora. More interestingly, while the term MONSOON appears exclusively in the DIY UN corpus, it does so in the BootCat RES corpus. Therefore, the discrepancy between the two sub-corpora is to be questioned through further analyses. Another feature specific to the NGO and RES corpora, namely the focus on the energy market, and in particular on the wind and solar energy, is largely confirmed by the BootCat results.

Table 16. Comparative analysis of discourse community specificities (Weather)

Sub-corpora	DIY Corpus	BootCat Corpus
NGO	climate-; rainwater; cloudy; climate.; flooded; windy; cloud_companies; rains	Zaman; non-commercial; haze; humid; fog ; draught ; choppy
UN	monsoon; inundation; hurricane; hurricanes; snowfall; humid; **rainy**; climate_factors; **avalanche**; weather_conditions	floodplain; chinook; hailstorm; floodplains; flooded; heatwaves; **avalanche**; inundations; **rainy**
RES	inclement; draught; sunnier; wind_companies; wind-farm; climate-windstorms; rainfalls; wind_mills; wind_sector; humid; mist	monsoon; sunny; wind_market; monsoons; tornado; rains; wind_output; windy

47 A comparative analysis of the specificities identified by the DIY and BootCat are not only incongruent, they further mitigate the specificities of each discourse by establishing new links within combinations of sub-corpora, and most importantly by drawing a link between the UN and RES sub-corpora with the lexical unit MONSOON. A new unit 'HUMID' is identified as common to all the three.

48 A slightly lower degree of uniformity (13 out 30, i.e. 43%) is established in the 'Change' semantic category, which confirms the DIY results (12 out of 30). The strong link within the combination of NGO and RES sub-corpora is reinforced with 8 lexical units being common to both, in line with the DIY results (7/30 common lexical units). However, in this semantic category the link between the NGO and UN sub-corpora is weakened with only 3 shared lexical units in the BootCat and 4 in the DIY corpus. Finally, the emerging link in the BootCat results between the UN and RES sub-corpora is confirmed with 4 lexical units being shared by both, while there were none common to this combination of sub-corpora in the DIY results.

Table 17. Comparative analysis of discourse community specificities (Change)

Sub-corpora	DIY Corpus	BootCat Corpus
NGO	fluctuating; conversion; happen; occur; substitution; transformation; converter; becomes	reforming; reformer; revision; replaced; formation; modifications
UN	**adaptive;** shifts; adapting; **occurrence;** adaptations; **restoration; occurred;** evolutionary; altered; alter; **changed**; affecting; oscillation; **occurring**	**adaptive; occurred; occurring;** raster; **restoration;** get; affects; experienced; **changed; occurrence**
RES	transformed; experienced; replacement; switch; momentum; adjustments; **switching**; adapted; adjusted; replaced; replacing	substitution; becoming; replace; **switching**; transform

49 Unlike the above discussed two categories, the comparative analyses in the "Science and Technology" semantic category point to some incongruence between DIY and BootCat results, with merely one third (10 out of 30) of co-occurrences across the three sub-corpora in the former and almost half (14/30) of lexical items common to the three sub-corpora in the latter. As for the possible links between sub-corpora, the DIY results established no combinations within this semantic category, while BootCat results confirmed the links already established in other categories: namely, the strong links of NGO sub-corpus with the other two (4/30 lexical units common in both combinations, NGO-UN and NGO-RES), and the tenuous link between the UN and RES corpora (1/30).

Table 18. Comparative analysis of discourse community specificities (Science & Technology in general)

Sub-corpora	DIY Corpus	BootCat Corpus
NGO	neodymium; tellurium; **radioactive**; reactors; scientist; observatory; NPS; se; state_of_the_art; tes; reactor; state-of-the-art;	biotechnology; **radioactive**; experiments; formulae; experiment; NRC; atomic_energy; instrumentation
UN	biology; **hydrology**; **oceanography**; envorinmenal_science; ph; climatology; s.e.; experiments; biogeography; epidemiology; chemistry; meteorology; T.E; **psychology**; physiology; marine_biology; SES; biota	**hydrology;** ecologists; lab; geology topography; **oceanography**; aeronautics; physics; **psychology**; analyser ; wavelength
RES	engineers; evs; lab; formula; LR; **laboratories**; electronic_engineers; **technicians**; nd; **engineer**; **po**; n.d.	Nd; NRCS; **laboratories**; technology_transfer; reactors; **engineer**; **po**; **technicians**; tech; tes; Neuroscience

50 As Table 18 shows, unlike the above two categories, there is greater coherence (demonstrated in bold type) between DIY and BootCat results in the semantic category of "Science and Technology in General". Concerning the specificities of each discourse community, it can be noted that there are more lexical units with negative connotation in the NGO sub-corpus which suggests a particular concern with noxious effects of chemicals and nuclear energy in general: NEODYMIUM, TELLURIUM, RADIOACTIVE. In contrast, the most used lexical units in the UN discourse seem to point to advances of the science and technology, and are positively connoted. Finally, the lexical units in the RES sub-corpus point to technical and engineering solutions. The most used lexical units in this "Science and Technology" semantic category are positive, as is the case with the UN corpus. Overall, while the latter two shape their discourse mostly pointing to possible solutions, NGOs point to problems that could possibly stem from scientific and technological advances.

2.2.3. Lexical Units on Justice

51 To consider the place of issues of justice in our corpora, we began by referring to semantic tags with *WMatrix*. As in the DIY corpora, we did not find semantic tags related to justice and ethics among the most represented in our corpora. In the B-NGO corpus, "Ethical" appears only 179th in terms of Log-Likelihood, in the B-UN corpus "General Ethics" appears 206th while in the B-RES corpus no related semantic tag appears. As was the case with the DIY corpus, the *Antconc* word count feature was used in a second stage to find out the number of occurrences of our lexical units of interest in the Bootstrapped corpus. The results are presented in Table 19. The striking absence of our three main terms of interest in the UN and RES sub-corpora in the DIY results are not only confirmed in the BootCat, but amplified with only a few occurrences in NGO and UN sub-corpora. However, further comparative analyses of results reveal a significant incongruence as there is not the same distribution of terms across sub-corpora.

Table 19. Frequency counts of terms of interest in the Bootstrapped corpus with *Antconc*

Term of interest	UN	NGO	RES
	W.f. (wpm)	W.f. (wpm)	W.f. (wpm)
Environmental Justice	5(4)	9(8)	0(0)
Climate Justice	0(0)	0(0)	0(0)
Energy Justice	0(0)	0(0)	0(0)
Human Rights	18(17)	11(10)	6(5)
Equity	25(23)	45(42)	75(70)
Inequity	0(0)	1(0)	0(0)
Equality	8(7)	27(25)	9(8)
Inequality	6(5)	15(14)	3(2)
Fairness	1(0)	2(1)	0(0)
Accessibility	28(26)	15(14)	18(16)
Safety	237(226)	2,391(2,248)	119(111)
Distribution	238(227)	332(312)	559(525)
People-centered	3(2)	0(0)	0(0)
Representation	28(26)	11(10)	22(20)

52 Opposite tendencies are observed when it comes to the lexical unit DISTRIBUTION, which had the highest frequency of occurrence in the DIY UN sub-corpus, and the lowest in the RES sub-corpus. In the BootCat data, the occurrence of the lexical unit is the most frequent in the RES corpus and the least frequent in the UN sub-corpus. The same is true for the term EQUITY. Furthermore, while CLIMATE JUSTICE has a high frequency of occurrence in the DIY NGO corpus, it does not appear at all in the BootCat results. All these differences seem to suggest that the DIY corpus allows finer distinctions, which, however, requires further research. In order to look for such differences, the third

point we are going to investigate is whether significant clusters appear distinctively in our three BootCat corpora.

2.2.4. Significant Compounds and Clusters

53 We will start by analysing the word clusters with VULNERABLE.

Table 20. Word clusters with VULNERAB*

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	vulnerability assessment	9(8)	vulnerability assessment	688(656)	vulnerability assessment	5(4)
2	vulnerability to climate change	5(4)	vulnerability analysis	147(140)	vulnerable countries	5(4)
3	vulnerable to climate change	5(4)	vulnerability analysis tools	115(109)	vulnerabilities to climate change	3(2)
4	vulnerable countries	4(3)	vulnerability in south east	97(92)	vulnerability assessment and adaptation	3(2)
5	vulnerability of ecosystems	3(2)	vulnerability assessments	69(65)	vulnerability to climate change	3(2)
6	vulnerability of reduction	3(2)	vulnerability question	47(44)		
7	vulnerable employment	3(2)	vulnerability and adaptation	45(42)		
8			vulnerability to climate change	41(39)		
9			vulnerability manager	40(38)		
10			vulnerability index	37(35)		

54 As Table 20 shows, three word clusters are identified as common to all three sub-corpora. However, the BootCat results do not allow for links to be established between the NGO and UN sub-corpora, as was the case in the DIY corpus. In fact, no combinations of corpora are distinguished.

55 Next, we come to the term RESPONSIBLE and its derivatives. The first striking feature of the BootCat corpus is the virtual absence of clusters with the lexeme RESPONSIBLE in the NGO sub-corpus. This contrasts with the DIY results, which revealed a great number of clusters in this semantic category. Furthermore, it is interesting to note that the only cluster that appears in the NGO sub-corpus is shared by the RES sub-corpus.

Table 21. Word clusters with RESPONSIB*

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	responsible business conduct	11(10)	responsible for administering	5(4)	responsible business conduct	10(9)
2			responsible for assessing	4(3)	responsible economies	4(3)
3			responsible for management	4(3)		
4			responsible office	4(3)		
5			responsible agency	3(2)		
6			responsible for implementing	3(2)		

56 The results in the RES sub-corpus are more centred on the economy, and seem to suggest a national, rather than international perspective on the issues of climate and energy justice (IMPLEMENTING NATIONAL STRATEGIES). This confirms the DIY results, where the RES corpus was equally turned to the energy market.

57 Finally, we will examine the results for the "energy" cluster.

58 As in the case of the DIY corpus, we consider the first ten occurrences of two-word units having the structure [ENERGY + NOUN] and the first ten occurrences of two-word units having the structure [ADJECTIVE + ENERGY] in our Bootstrapped corpus.

Table 22. [ENERGY+NOUN] in the three sub-corpora

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	energy efficiency	493(463)	energy efficiency	57(54)	energy efficiency	1004(943)
2	energy sector	233(219)	energy consumption	47(44)	energy storage	395(371)
3	energy agency	203(190)	**energy policy**	35(33)	energy use	375(352)
4	energy demand	194(182)	energy use	24(22)	energy share	288(270)
5	**energy outlook**	186(174)	energy management	23(21)	energy consumption	283(265)
6	energy use	167(157)	energy sources	22(21)	energy supply	237(222)
7	energy source	153(143)	energy harvesting	17(16)	energy savings	228(214)
8	energy consumption	147(138)	energy nexus	14(13)	energy technologies	171(160)
9	energy supply	146(137)	energy sector	14(13)	energy agency	164(154)
10	energy technologies	106(99)	energy demand	12(11)	energy sources	161(151)

59 Though the ENERGY EFFICIENCY cluster is common to all the three sub-corpora, its frequency of occurrence is more than doubled in the BootCat corpus (1004 instead of 407 previously). In this sense, the DIY corpus shows a greater degree of homogeneity in the use of this cluster by the three sub-corpora (418, 484 and 407 for the NGO, UN and RES corpora respectively). Both DIY and BootCat corpora identify four clusters common to all the three sub-corpora, namely ENERGY DEMAND, ENERGY USE, ENERGY EFFICIENCY and ENERGY SYSTEM in the DIY corpus and ENERGY USE, ENERGY CONSUMPTION, ENERGY EFFICIENCY and ENERGY SOURCES in the BootCat. Though the four clusters are not identical in the two corpora, they are semantically close as they refer to the energy demand, energy consumption and sources of energy that constitute the whole energy system. The results in both corpora identify previously observed combinations of sub-corpora and confirm the proximity of the NGO corpus with the other two, and the lack of links between the UN and RES sub-corpora.

60 As for features characterising discourse communities, interestingly, both DIY and BootCat corpora point to ENERGY OUTLOOK and ENERGY POLICY as exclusively characteristic of the NGO and UN sub-corpora correspondingly. Energy technologies that would enable energy savings seem to be the main focus of the RES corpus in both DIY and BootCat.

Table 23. [ADJECTIVE+ENERGY] in the three sub-corpora

	NGO		UN		RES	
	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)	Cluster	W.f. (wpm)
1	renewable energy	388(364)	renewable energy	56(53)	renewable energy	2882(2708)
2	nuclear energy	359(337)	impact energy	36(34)	wind energy	182(171)
3	world energy	227(213)	wind energy	33(31)	sustainable energy	182(171)
4	primary energy	157(147)	international energy	21(20)	national renewable energy	171(160)
5	international energy	129(121)	Mississippi energy	16(15)	clean energy	146(137)
6	global energy	120(112)	building energy	13(12)	electric energy	146(137)
7	clean energy	106(99)	final energy	12(11)	final energy	146(137)
8	sustainable energy	88(82)	residential energy	10(9)	global energy	124(116)
9	atomic energy	72(67)	food-energy	10(9)	primary energy	124(116)
10	thermal energy	63(59)	geothermal energy	9(8)	total energy	114(107)

61 The three clusters INTERNATIONAL ENERGY, WORLD ENERGY, and GLOBAL ENERGY in the B-NGO corpus could reveal a greater focus on energy issues at a global scale. The first two of these clusters are equally present in the DIY corpus.

62 The UN sub-corpus seems to be concerned with long-term solutions to climate issues with clusters, such as IMPACT ENERGY, BUILDING ENERGY and FINAL ENERGY. The cluster NATIONAL RENEWABLE ENERGY seems to confirm a focus on the national scale in the RES sub-corpus. Not surprisingly, the cluster RENEWABLE ENERGY ranks top in the RES sub-corpus with 2 882 occurrences.

63 In the BootCat corpus, the only cluster that appears as common to the three sub-corpora is RENEWABLE ENERGY. Unlike in the DIY corpus, the cluster GLOBAL ENERGY no longer figures. Similarly, the number of clusters common to NGO and UN corpora gets down to one in the BootCat, instead of three in the DIY corpus. INTERNATIONAL ENERGY appears as the fourth most common cluster in the UN corpus in the BootCat and the 3rd most common in the DIY. PRIMARY ENERGY and WORLD ENERGY are no longer identified in the BootCat. Nevertheless, in the comparative analysis of the NGO and RES corpora, the BootCat results identify four clusters, SUSTAINABLE ENERGY, CLEAN ENERGY, GLOBAL ENERGY, PRIMARY ENERGY instead of one, that of CLEAN ENERGY in DIY corpus. More interestingly, what is identified as common to NGO and UN in DIY corpus (PRIMARY ENERGY) is done so for the combination of NGO and RES sub-corpora here.

Conclusion

64 Our comparative analysis of first results obtained in the DIY and BootCat corpora seems to suggest that BootCat can be a very useful and efficient tool to extend existing small corpora. The specialised lexical units and terms can thus be considered in more contexts and the identification of collocations and multi-word units can be enhanced. To get these results, one must be aware of the importance of building a strongly representative corpus in the first stage, as the keywords extracted from the first corpora will be all the more important as they will be used as seeds to constitute the second. Although we believe this experience shows that BootCat and equivalent bootstrapping tools should not be neglected as potential extensions to existing corpora, it has also shown certain limits. There are variations between discourse communities which were visible in the DIY corpora and which are not in the BootCat corpora. The loss of specificity concerning author, date, and genre also means that there are many questions which could be considered in our DIY corpus that would be meaningless in the BootCat corpus. This enables us to highlight how important it is to question your objectives when you adopt a corpus constitution method. Having a large corpus that can itself be divided into small sub-corpora, according to the questions we are aiming to answer, is definitely more useful than having a single, undifferentiated, corpus on the issue. Proceeding through steps and contrasting small corpora that can then be combined into a big one to answer certain questions seems like a good way of building a corpus (Chareaudau 2009). The study presented here is a preliminary work to check the representativity of our corpora. So far, we have mainly presented quantitative data on the corpora but aim to analyse them in more qualitative terms to further the understanding of the concepts of interest.

BIBLIOGRAPHY

Anthony L. (2014). *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/.

Baker P. (2004). "Querying keywords: questions of difference, frequency and sense in keywords analysis", Journal of English Linguistics 32(4): 346-359.

Baroni M. & Bernardini S. (2004). "BootCaT: Bootstrapping corpora and terms from the web", *Proceedings of LREC.*

Bernardini S. (2006). "Corpora for translator education and translation practice: achievements and challenges", *Proceedings of the L4Trans Workshop at LREC.*

Bullard R. D. (1990). *Dumping in Dixie: Race, class, and environmental quality*. Boulder, CO: Westview.

Charaudeau P. (2009). "Dis-moi quel est ton corpus, je te dirai quelle est ta problématique", *Corpus* 8 : 37-66.

Dahan A. & Aykut S. (2015). *Gouverner le climat ? Vingt ans de négociations internationales*. Paris: Presses de Sciences Po.

Davies M. (2010). "The Corpus of Contemporary American English as the first reliable monitor corpus of English", *Literary and Linguistic Computing* 25(4): 447-464.

Heffron R. J., McCauley D. & Sovacool B.K. (2015). "Resolving society's energy trilemma through the Energy Justice Metric", *Energy Policy* 87: 168-176.

Heffron R. J. & MacCauley D. (2017). "The concept of energy justice across the disciplines", *Energy Policy* 105: 658-667.

Kilgarriff A., Baisa Vit, Busta J. et al. (2014). "The Sketch Engine: ten years on", Lexicography 1: 7-36.

Kilgarriff A. & Grefenstette G. (2003). "Introduction to the special issue on the web as corpus", *Comput. Linguist*. 29(3): 333-347.

L'Hôte E. & Lemmens M. (2009). "Reframing treason: metaphors of change and progress in new Labour discourse", *CogniTextes* [Online], Volume 3.

Belinda M. (1997). "Do-it-yourself corpora... with a little bit of help from your friends!", *in* B. Lewandowska-Tomaszczyk & P. J. Melia (eds.) *PALC '97 Practical Applications in Language Corpora*. Lodz: Lodz University Press, 403-410.

Mary Robinson Foundation – Climate Justice. (2013). "Climate Justice Baseline" report, published July 2013. http://www.mrfcj.org/resources/climate-justice-baseline/.

Nicholson S. & Chong D. (2011). "Jumping on the Human Rights Bandwagon: How Rights-Based Linkages can Refocus Climate Politics", *Global Environmental Politics* 11(3): 121-136.

Rayson P. (2008). "From key words to key semantic domains", *International Journal of Corpus Linguistics* 13(4): 519-549.

Rhaman M. (2016). "Climate Justice Framing in Bangladeshi Newspapers, 2007-2011", *South Asia Research* 36(2): 186-205.

Rühlemann C. & Aijmer K. (2014). "Introduction. Corpus pragmatics: laying the foundations", *in* K. Aijmer & C. Rühlemann (eds.) *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press, 1-26.

Sovacool B.K., Burke M., Baker L. *et. al.* (2017). "New frontiers and conceptual frameworks for energy justice", *Energy Policy* 105: 677-691.

Swales, J. (1990). "The Concept of Discourse Community", *in Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press, 21-32.

Varantola K. (2003). "Translators and Disposable Corpora", *in* F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in Translator Education.* Manchester: St. Jerome, 55-70.

Zanettin F. (2002). "Corpora in translation practice", *Proceedings of the First International Workshop on Language Resources* (LR) *for Translation Work and Research*, 10-14.

NOTES

1. Log-Likelihood ratios are used within statistical tests to assess differences. Here the comparison is between normalised frequency distributions (see L'Hôte and Lemmens 2009 for more detailed explanations).
2. See http://ucrel.lancs.ac.uk/usas/semtags.txt for a complete list of tags.
3. To identify key semantic categories in a corpus the *WMatrix* software compares it to a reference corpus already integrated in the software. We used the one named British English 2006.
4. Pollutant associated to conventional energy production.
5. Seeds are starting points for automatic retrieval. Web-crawlers typically use URLs as seeds, and BootCat is an original tool in that it takes word combinations to generate those URLs.
6. The term refers to an ordered sequence of a number of items: in BootCat, the user chooses "tuple length", i.e. whether they want to use sequences of two, three, four or more seeds.
7. While BootCat outputs just one text file, the SketchEngine creates a table associating each retrieved file to its word count and enables users to delete specific files before compiling the corpus.

ABSTRACTS

This article offers a descriptive and analytic view of the different stages leading to the constitution of a corpus that is representative of the issues of climate and energy justice. Overall, the corpus contains over five million words and gathers reports, newsletters and web-pages dealing with the most equitable ways of moving to a low-carbon future in the aim of limiting climate change. It can be divided into six sub-corpora, according to types of discourse communities, and methods of constitution. We begin by presenting the small Do It Yourself (DIY) corpora which were used as a starting point. Three discourse communities were selected to observe possible variation in their treatment of the issue: Non-Governmental Organisations (NGOs), United-Nation institutions, and the Renewable Energy Sector (RES). The sources are selected according to author, date, keywords in title. Using the concordance *Antconc* and *WMatrix* software we test the reliability of the corpora for their thematic content, terminology and lexical

unit classification. Our first results enable us to confirm variation between the discourse communities. The discrepancy in sizes and the time-consuming nature of the initial DIY corpus constitution lead us to use BootCat to extend them, using keywords from the corpora as seeds to retrieve and download webpages. We thus contrast a more traditional approach to corpus building to web-as-corpus data gathering methods. We compare the results found in the BootCat corpora to test if they are as specific as those in the DIY corpora. This enables us to draw conclusions on the possible uses and advantages of relatively small corpora for the study of specialised discourse.

Cet article décrit et analyse les différentes étapes de constitution d'un corpus représentatif des questions de justice climatique et énergétique. Le corpus contient cinq millions de mots en tout et rassemble des rapports, des lettres d'information et pages web traitant des solutions équitables à faible empreinte carbone pour limiter le changement climatique. Il est divisé en six sous-corpus selon les types de communautés de discours et de méthodes de constitution. Nous commençons par la présentation du petit corpus fait maison que nous utilisons comme point de départ. Trois communautés de discours ont été sélectionnées afin d'observer d'éventuelles variations dans leur traitement de ces questions : Organisations Non Gouvernementales, institutions onusiennes et organisations du secteur de l'énergie renouvelable. Les sources ont été sélectionnées en fonction des auteurs, dates et mots clés présents dans les titres. Grâce aux logiciels de concordance AntConc et WMatrix, nous avons testé la comparabilité de ces corpus du point de vue de leur contenu thématique, de leur terminologie et de la classification de leurs unités lexicales. Nos premiers résultats nous permettent de confirmer l'existence de variations entre communautés de discours. Le caractère chronophage de notre démarche de constitution d'un corpus « maison », ainsi que le déséquilibre entre le nombre de mots obtenus pour chaque sous-corpus nous conduisent à utiliser BootCat afin de constituer un corpus plus fourni. L'outil utilise des mots clés comme « semences » pour la récupération et le téléchargement automatiques de pages web. Nous pouvons ainsi comparer une méthodologie traditionnelle de constitution de corpus à une méthodologie qui utilise le web en tant que corpus. Nos résultats BootCat sont confrontés à ceux du corpus maison pour voir s'ils révèlent aussi bien les spécificités des sous-corpus. Cette démarche aboutit à des conclusions sur les possibles utilisations de corpus relativement petits, et d'en souligner la pertinence pour l'étude de discours spécialisés.

INDEX

Keywords: Corpus building, BootCat, Small DIY Corpora, Climate Justice, Energy Justice, Discourse Communities, Specialised Discourse
Mots-clés: Constitution de corpus, BootCat, Petits corpus maison, justice climatique, justice énergétique, communautés de discours, discours spécialisés

AUTHORS

CAMILLE BIROS

Université de Grenoble

CAROLINE ROSSI

Université de Grenoble

INESA SAHAKYAN

Université de Grenoble