Maristella GATTO
Università di degli Studi di Bari "Aldo Moro"

maristella.gatto@uniba.it

LANGUAGE SECTION

'Sustainable' corpora for transnational subjects: methods and tools

Abstract

This paper presents semi-automated methods for corpus compilation developed over the past few years in the context of research on the "web as/for corpus". These methods and tools have already proved extremely useful for the quick compilation of *ad hoc* monolingual/multilingual corpora for terminology extraction (Baroni and Bernardini 2004; Baroni et al. 2009; Bernardini and Ferraresi 2013), but have recently started to attract attention also in the context of corpus-based Critical Discourse Analysis, where flexible tools for the compilation and exploration of corpora might be promising allies in the effort to join forces between corpus linguistics and critical studies (Gabrielatos 2007; Baker 2008; Wild et al. 2013).

For their characteristics, these quick *ad hoc* corpora might prove particularly useful in research and teaching contexts dealing with issues whose topicality requires continuous updating of the resources, as is the case with the corpora for *immigration* and *sustainable tourism* discussed in the present paper. In the first case study, a corpus consisting of the complete debate on the Immigration Bill 2014 was compiled automatically, to provide a comprehensive overview of the parliamentary debate in the context of classroom activities with postgraduate students in Modern Languages for International Cooperation. In the second case study a corpus of texts taken from the official website of the World Tourism Organization (WTO) was compiled through the automatic extraction of keywords from a small pilot corpus representative of 'discourse' within this specific organization. In both cases the corpora were meant to provide a snapshot of ongoing discourse, as a complement to more focused qualitative research carried out with other methods or to prompt classroom discussion. The examples reported are indicative of the possible benefits of integrating corpora built 'on demand' in the context of research or classroom activities not necessarily centred on corpus linguistics alone.

1. Background and aims

In recent years several tools and methods for corpus compilation have been developed in the context of research on the "web as/for corpus", which have already proved extremely useful for the creation of large general purpose reference corpora for a variety of languages as well as for the creation of monolingual/multilingual corpora for terminology extraction (Baroni and Bernardini 2004; Baroni et al. 2009; Bernardini and Ferraresi 2013). More recently these corpora have attracted attention in the context of corpus-based Critical Discourse Analysis, where flexible tools for the compilation and exploration of corpora (e.g. WebBootCaT and Sketch Engine) seem to be promising allies in the effort to join forces between corpus linguistics and critical studies (Gabrielatos 2007; Baker 2008; Wild et al. 2013).

Without questioning the validity of the established practice of building carefully compiled traditional corpora, (especially for the purposes of Critical Discourse Analysis) this paper aims to show whether the possibility of creating *ad hoc* corpora in a few minutes for a variety of domains and genres can contribute to spread the use of quantitative evidence to support, validate and stimulate the work of researchers primarily engaged in qualitative analysis of language data. For their characteristics, these quick *ad hoc* corpora could be defined as 'renewable corpora', since they are easily and rapidly created and recreated on the basis of customized criteria and variable parameters, as well as 'sustainable corpora', since they can be maintained, updated and regenerated at an extremely favorable cost-effectiveness ratio. These characteristics make them particularly useful in research and teaching contexts dealing with issues whose topicality requires continuous updating of the resources. The examples reported below in Section 3 are indicative of the possible benefits of using corpora quickly built 'on demand' in the context of research or classroom activities, where they performed quite well even as single use corpora to prompt classroom discussion or to provide subsidiary quantitative evidence to complement research carried out with other methods

2. Methods and tools

The tools and methods presented in this paper have now become a standard in the creation of specialized corpora for translation purposes as a development of the practice of creating Do-It-Yourself, 'quick-and-dirty', disposable corpora (Zanettin 2002). More specifically, the paper deals with corpora compiled through a semi-automated process using either BootCaT (Baroni and Bernardini 2004), a stand-alone software, or WebBootCaT, a service available through the Sketch Engine website[1]. The key feature of these tools is that they take as a starting point for corpus compilation just a number key words or phrases which the user considers likely to occur in the domain for which a corpus is going to be built These words are called 'seeds' and are transformed by the system into a set of automated queries submitted to an ordinary search engine. Thus, if the intention is to create a corpus on alternative forms of tourism based on issues regarding sustainability and responsibility, one can input the words "tourism", "sustainable", "responsible", "green", "nature", "environment", "eco-friendly" and let the software perform the search, download and clean the text, and compile the corpus (see Gatto 2014: 140ff). An alternative procedure allowed by the tool is to compile a corpus using a number of known URLs: in this case the system automatically performs the crawl, downloads and cleans the texts, and returns them to the user in the form of a corpus in text-only (.txt) format for analysis with the most common concordancers.

Whether used in the standalone version or through the Sketch Engine, these systems pose a common challenge in terms of strategies to 'bootstrap' the process of corpus compilation. In particular, choosing the seed terms has been recognized as a particularly sensitive area, as discussed in Gabrielatos (2007: 6), because of the tension between "creating a corpus in which all the texts are relevant, but which does not contain all relevant texts available in the database, and, [on the other], creating a corpus which does contain all available relevant texts, albeit at the expense of irrelevant texts also being included". For this reason, it is important not to rely simply on intuition when choosing the seeds, and search for alternative ways to make sure that seed terms are not arbitrarily chosen. Several solutions have been described in works by Gabrielatos (2007), Zanettin (2012) and Bernardini (2013), depending on the research aims.

[1] BootCaT can be downloaded for free from the developers' website http://bootcat.dipintra.it/. WebBootCat can be accessed by registered users from https://www.sketchengine.co.uk/

This is the case with the two corpora for *immigration* and *sustainable tourism* discussed in the present paper. Since the corpora presented in this paper were aimed at providing a snapshot of discourse within specific discourse communities, crawl form URLs was prioritized in the first case study, whereas the compilation of a small 'pilot' corpus from known URLs in order to extract more controlled keywords to use as seeds was opted for in the second case study. In the first case study, a corpus was compiled automatically to provide a comprehensive overview of the parliamentary debate on the Immigration Bill 2014, in the context of classroom activities with postgraduate students in Modern Languages for International Cooperation. The corpus was queried for key words like MIGRANT or IMMIGRANT to investigate patterns of usage for these words in this specific context, and proved extremely useful in foregrounding the role played by specific lexico-grammar patterns in the creation of what could be termed in Van Dijk's terminology as in-groups and out-groups (Van Dijk 2006: 126). In the second case study a corpus of texts taken from the official website of the World Tourism Organization (WTO) was compiled through the automatic extraction of keywords obtained from a small pilot corpus obtained by crawling the web starting from the URLs derived from links in the "UNWTO A to Z" in the website home page.

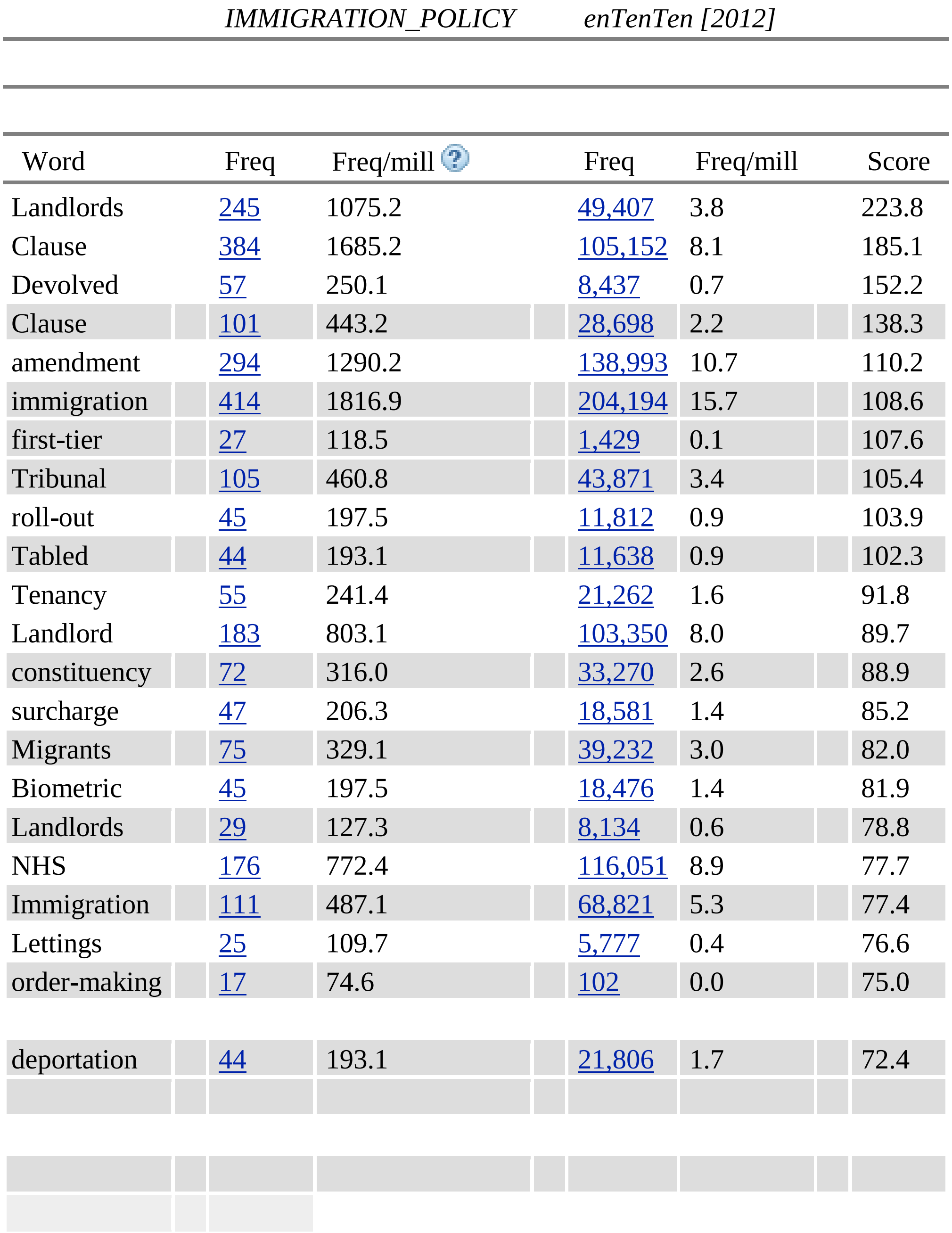
3. Case studies

3.1. "People who have no right to be here" in the Debate on the Immigration Bill 2014

The first case study concerns a preliminary investigation of the 2014 Immigration Bill, performed in classroom activities with a group of postgraduate students in a Modern Languages for International Cooperation MA Programme at the University of Bari[2]. Corpus-based investigations of the representation of migrants, immigrants, refugees and asylum-seeker are among the seminal and most influential research projects aimed at evaluating the "useful synergy" between corpus linguistics and Critical Discourse Analysis (Baker et. al 2007; Gabrielatos and Baker 2008). Also relevant is the dissemination of results from research by "The Migration Observatory" at the University of Oxford[3]. In the present case study the aim was simply to have a comprehensive overview of the parliamentary debate on the Immigration Bill 2014, and a corpus consisting of the complete debate was compiled by performing an automatic download of the texts from the URLs corresponding to each stage of the debate (as available at the British Parliament website)[4]. The resulting data set was a 200.026 word corpus, compiled and Part-of-Speech tagged, which was ready to be queried online through the Sketch Engine in less than 10 minutes.

The starting point for the investigation was the list of keywords extracted by the system using enTenTen 2012 as a reference corpus[5]. Apart from the list of proper nouns of MPs and the presence of abbreviations like Hon., or of personal address markers as Friend and Mr Speaker, which clearly reflect the specific genre of the Parliamentary debate, the list of keywords retrieved basically outlined the main concerns of the debate in terms of both discourse situation (clause, amendment, …) and issues at stake (landlords lettings, devolved, NHS surcharge etc.…), as shown in Figure 1. below:

[2] The compilation of the 2014 Immigration Bill corpus was part of classroom activities in the English Language and Translation course in the a.y. 2014-2015. The students accessed WebBootCat through a one-month free trial registration to the Sketch Engine.
[3] http://www.migrationobservatory.ox.ac.uk/
[4] http://services.parliament.uk/bills/2013-14/immigration/stages.html
[5] enTenTen 2012 is a member of a family of web corpora made available through the Sketch Engine. See https://www.sketchengine.co.uk/ententen-corpus/



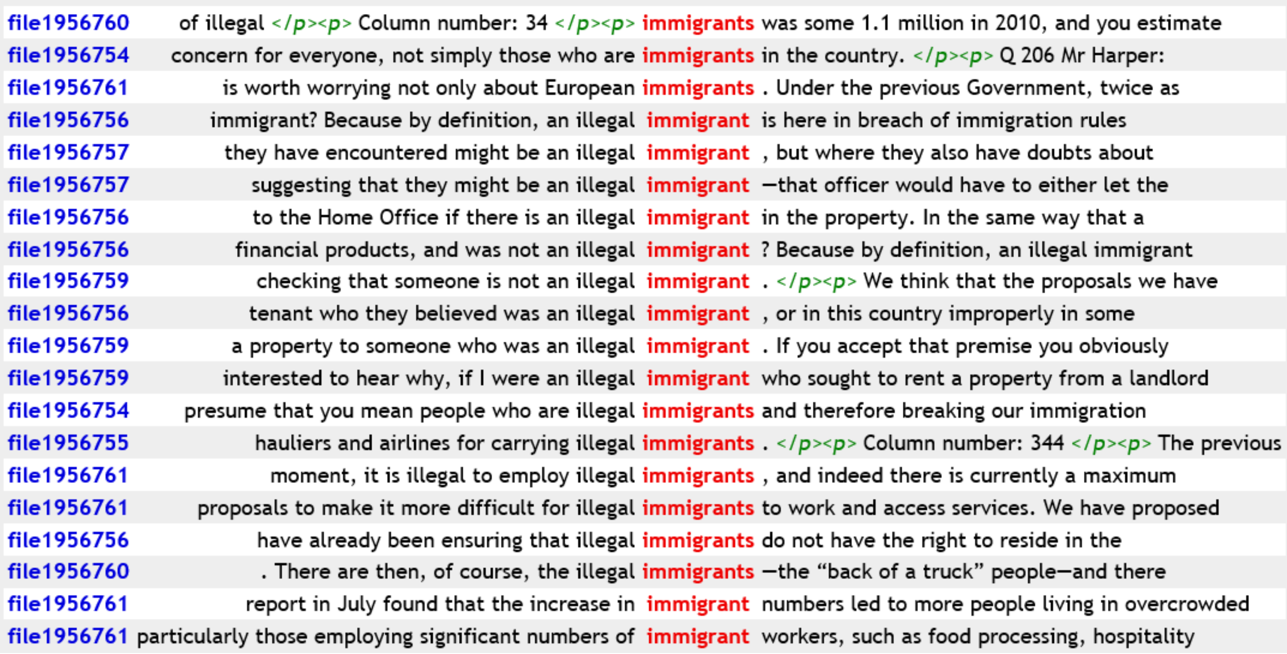
	IMMIGRATION_POLICY		enTenTen [2012]		
Word	Freq	Freq/mill	Freq	Freq/mill	Score
Landlords	245	1075.2	49,407	3.8	223.8
Clause	384	1685.2	105,152	8.1	185.1
Devolved	57	250.1	8,437	0.7	152.2
Clause	101	443.2	28,698	2.2	138.3
amendment	294	1290.2	138,993	10.7	110.2
immigration	414	1816.9	204,194	15.7	108.6
first-tier	27	118.5	1,429	0.1	107.6
Tribunal	105	460.8	43,871	3.4	105.4
roll-out	45	197.5	11,812	0.9	103.9
Tabled	44	193.1	11,638	0.9	102.3
Tenancy	55	241.4	21,262	1.6	91.8
Landlord	183	803.1	103,350	8.0	89.7
constituency	72	316.0	33,270	2.6	88.9
surcharge	47	206.3	18,581	1.4	85.2
Migrants	75	329.1	39,232	3.0	82.0
Biometric	45	197.5	18,476	1.4	81.9
Landlords	29	127.3	8,134	0.6	78.8
NHS	176	772.4	116,051	8.9	77.7
Immigration	111	487.1	68,821	5.3	77.4
Lettings	25	109.7	5,777	0.4	76.6
order-making	17	74.6	102	0.0	75.0
deportation	44	193.1	21,806	1.7	72.4

Figure 1. A sample from the list of keywords form the Immigration Bill Corpus (lines including names and abbreviations have been removed)

The corpus was then queried to explore the behaviour of specific words. Most students focused their attention on the words “immigrant” and “migrant” and noticed in the first place that as a debate on immigration, counting the lemma IMMIGRATION itself as a keyword with 414 occurrences (see Figure 1. above), the texts did not contain a high number of occurrences of the lemma IMMIGRANT itself, featuring instead the more ‘politically correct’ form MIGRANT as one of the keywords. Going back to the frequency list they noticed indeed that the lemma IMMIGRANT occurs only 21 times in the whole corpus, almost invariably in the collocation with “illegal” (see Figure 2. below), whereas the 105 occurrences of MIGRANT mostly referred to “economic migrants” and “temporary migrants”, with fewer examples for “illegal migrant” or “illegal migrants” (a datum which perfectly matches findings about usage of the word “migrant” in a recent corpus-based study by the Migration Observatory[6]). This triggered a question on whether there might be other

[6] The Migration Observatory, “Migrants in the newspapers: An influx of illegal, failed, economic terrorists?”,, http://www.migrationobservatory.ox.ac.uk/press/migrants-in-the-newspapers-an-influx-of-illegal-failed-economic-terrorists/

‘labels’ used to refer to immigrants, which are to be considered as the main participants in the discourse world represented in the Immigration Bill.



file1956760	of illegal </p><p> Column number: 34 </p><p>	immigrants	was some 1.1 million in 2010, and you estimate
file1956754	concern for everyone, not simply those who are	immigrants	in the country. </p><p> Q 206 Mr Harper:
file1956761	is worth worrying not only about European	immigrants	. Under the previous Government, twice as
file1956756	immigrant? Because by definition, an illegal	immigrant	is here in breach of immigration rules
file1956757	they have encountered might be an illegal	immigrant	, but where they also have doubts about
file1956757	suggesting that they might be an illegal	immigrant	—that officer would have to either let the
file1956756	to the Home Office if there is an illegal	immigrant	in the property. In the same way that a
file1956756	financial products, and was not an illegal	immigrant	? Because by definition, an illegal immigrant
file1956759	checking that someone is not an illegal	immigrant	. </p><p> We think that the proposals we have
file1956756	tenant who they believed was an illegal	immigrant	, or in this country improperly in some
file1956759	a property to someone who was an illegal	immigrant	. If you accept that premise you obviously
file1956759	interested to hear why, if I were an illegal	immigrant	who sought to rent a property from a landlord
file1956754	presume that you mean people who are illegal	immigrants	and therefore breaking our immigration
file1956755	hauliers and airlines for carrying illegal	immigrants	. </p><p> Column number: 344 </p><p> The previous
file1956761	moment, it is illegal to employ illegal	immigrants	, and indeed there is currently a maximum
file1956761	proposals to make it more difficult for illegal	immigrants	to work and access services. We have proposed
file1956756	have already been ensuring that illegal	immigrants	do not have the right to reside in the
file1956760	. There are then, of course, the illegal	immigrants	—the “back of a truck” people—and there
file1956761	report in July found that the increase in	immigrant	numbers led to more people living in overcrowded
file1956761	particularly those employing significant numbers of	immigrant	workers, such as food processing, hospitality

Figure 2. Concordance lines for the lemma IMMIGRANT *form the Immigration Bill Corpus*

With this in mind, a new exploration of the frequency wordlist revealed the special function of the word “people”. As suggested in Mahlberg (2005: 99ff.), “people” belongs to the category of general nouns which may typically have what she calls a “local textual function” in terms of cohesion and coherence. In the case of “people”, as used in this corpus, what seemed noticeable was a repeated collocation with “who”, the strongest collocate computed by the system, which was taken as evidence of a tendency to use the word “people” with a qualifying post-modification. In the specific case of the Immigration Bill, the pattern “people who…” seemed indeed to perform the local textual function of introducing specific categories labelled through a periphrasis rather than directly through a more specific noun or a through a premodifier. One such category appears to be the one labelled as “people who have no right to be here” which, a close reading of excerpts from the debate suggests, can be related to unmodified “migrants” (rather than “illegal immigrants”) through patterns of coherence and cohesion:

> We always prefer **people who have no right or valid leave to be in the United Kingdom** to return home voluntarily. However, if **they** do not do so, it is right that **they can be removed** quickly and easily. The amendment is intended to ensure that **a person** must be given written notice of **their removal**. […] At the moment, **migrants** are told that **they are not allowed to be here**, and we have to tell **them** separately about their removal. (Mr Harper, Public Bill Committee, Tuesday 5 November 2013, Morning) [7]

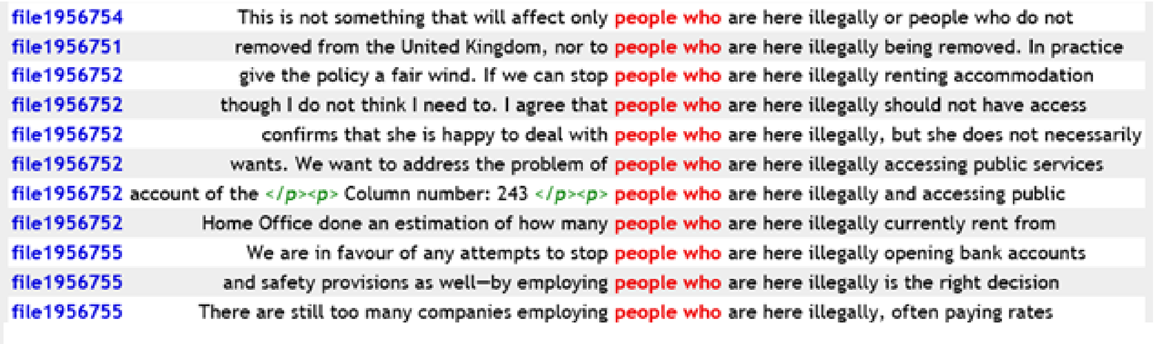
Starting from this example, a more comprehensive exploration of concordance lines for the pattern “people who” revealed a large number of occurrences for such clusters as “people who are here illegally” (11), “people who have no right to be here /in this country” (7), or “people who should not be in this country” and a number of possible variants:

- people who have no right or valid leave to be in the United Kingdom

[7] http://www.publications.parliament.uk/pa/cm201314/cmpublic/immigration/131105/am/131105s01.htm

- people who are told that they have no right to be in the country
- people who have no right to be here,
- people who have no right to be in this country
- people who should not be here
- people who do not have the right to remain in this country
- people who have no right to be here.

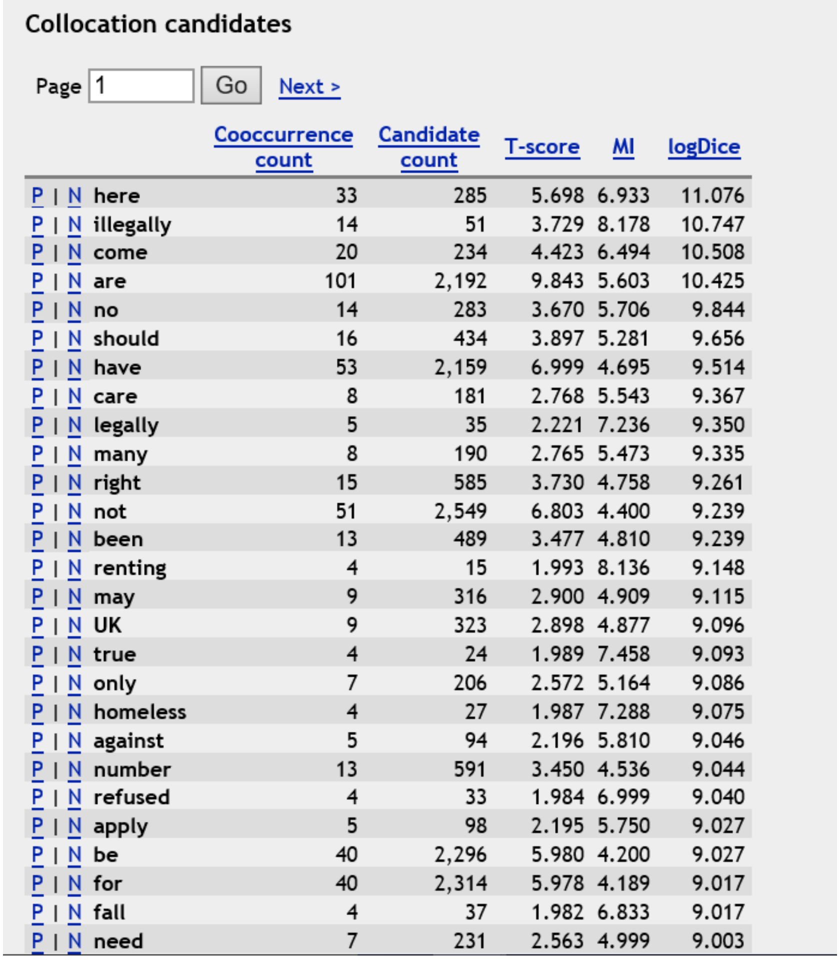
These variants indicate that the category of illegal immigrants is referred to in the texts of the debate in many different ways. Here is, by way of example, the complete list of concordance lines for “people who are here illegally”:



```
file1956754            This is not something that will affect only people who are here illegally or people who do not
file1956751          removed from the United Kingdom, nor to people who are here illegally being removed. In practice
file1956752           give the policy a fair wind. If we can stop people who are here illegally renting accommodation
file1956752        though I do not think I need to. I agree that people who are here illegally should not have access
file1956752              confirms that she is happy to deal with people who are here illegally, but she does not necessarily
file1956752         wants. We want to address the problem of people who are here illegally accessing public services
file1956752 account of the </p><p> Column number: 243 </p><p> people who are here illegally and accessing public
file1956752       Home Office done an estimation of how many people who are here illegally currently rent from
file1956755            We are in favour of any attempts to stop people who are here illegally opening bank accounts
file1956755         and safety provisions as well—by employing people who are here illegally is the right decision
file1956755      There are still too many companies employing people who are here illegally, often paying rates
```

Figure 3. A sample of concordance lines for the pattern “people who” from the Immigration Bill Corpus

It is also interesting to note that the pattern “people who” has precisely the words “here”, “come”, and “illegally” as its most salient collocates (see Figure 4. below):



Collocation candidates

		Cooccurrence count	Candidate count	T-score	MI	logDice
P | N	here	33	285	5.698	6.933	11.076
P | N	illegally	14	51	3.729	8.178	10.747
P | N	come	20	234	4.423	6.494	10.508
P | N	are	101	2,192	9.843	5.603	10.425
P | N	no	14	283	3.670	5.706	9.844
P | N	should	16	434	3.897	5.281	9.656
P | N	have	53	2,159	6.999	4.695	9.514
P | N	care	8	181	2.768	5.543	9.367
P | N	legally	5	35	2.221	7.236	9.350
P | N	many	8	190	2.765	5.473	9.335
P | N	right	15	585	3.730	4.758	9.261
P | N	not	51	2,549	6.803	4.400	9.239
P | N	been	13	489	3.477	4.810	9.239
P | N	renting	4	15	1.993	8.136	9.148
P | N	may	9	316	2.900	4.909	9.115
P | N	UK	9	323	2.898	4.877	9.096
P | N	true	4	24	1.989	7.458	9.093
P | N	only	7	206	2.572	5.164	9.086
P | N	homeless	4	27	1.987	7.288	9.075
P | N	against	5	94	2.196	5.810	9.046
P | N	number	13	591	3.450	4.536	9.044
P | N	refused	4	33	1.984	6.999	9.040
P | N	apply	5	98	2.195	5.750	9.027
P | N	be	40	2,296	5.980	4.200	9.027
P | N	for	40	2,314	5.978	4.189	9.017
P | N	fall	4	37	1.982	6.833	9.017
P | N	need	7	231	2.563	4.999	9.003

Figure 4. A sample from the list of collocates for "people who" form the Immigration Bill Corpus

Taking further advantage from the fact that in corpora compiled using the tools discussed in the present paper each word still retains its link to the original web text, the post-modifying relative clause "who have no right to be here" was searched in the whole debate to highlight all possible forms of discursive formation of what could be well termed with van Dijk as an out-group (2006), consisting of people whose 'negative' representation is crucially grounded in this specific lack of the right of abode. Examples of postmodifying "who have no right to be here" were found to occur in particular in speech by the Minister for Immigration (Mr Harper), as in the following extract from the debate:

> 9.15 am
> **Mr Harper:** The right hon. Gentleman makes a good point about those **who have no right to be here** and the mechanism by which *they are removed from the country*. There are approximately 14,000 enforced removals a year whereby *people are arrested, detained and then removed*. Sometimes it involves hiring escorts and is an expensive process. About 29,000 people depart voluntarily. There are different levels of voluntariness; some go completely voluntarily, others we assist in their departure from the United Kingdom but without having to enforce it.
> The right hon. Gentleman is right. It could be argued that the first thing we should do with *all those whose extension of leave is refused* or **who have no right to be here** is arrest them, detain them and remove them. I would argue that that would not be a sensible use of taxpayer resources, because an enforced removal can cost about £15,000. That does not include the incredibly expensive cases where escorts have to be hired.
> When someone is refused leave to be in the United Kingdom or their leave is curtailed and they are told that they have no right to be here, the first option should be for them to leave voluntarily. A significant number do, and we have seen quite a lot of success in encouraging more people to leave the United Kingdom voluntarily. That is much better for them. It saves them having to go through the process of being arrested, detained and removed. It also means that they are much more likely in future to be able to return to the United Kingdom legally. If we have to use taxpayer resources to enforce their removal, we will put in place a 10-year re-entry ban. The limits are much lower if they remove themselves voluntarily.

> The right hon. Gentleman makes a good point. Part of what we are trying to do in the Bill is to make it more difficult to remain in the UK voluntarily, so that people **who have no right to be here**, of whom a significant number come here lawfully and then overstay, choose to leave voluntarily and we do not have to use enormous sums of money, which we get from hard-working families, to remove them. (Mr Harper, Public Bill Committee, Tuesday 5 November 2013, Morning)[8]

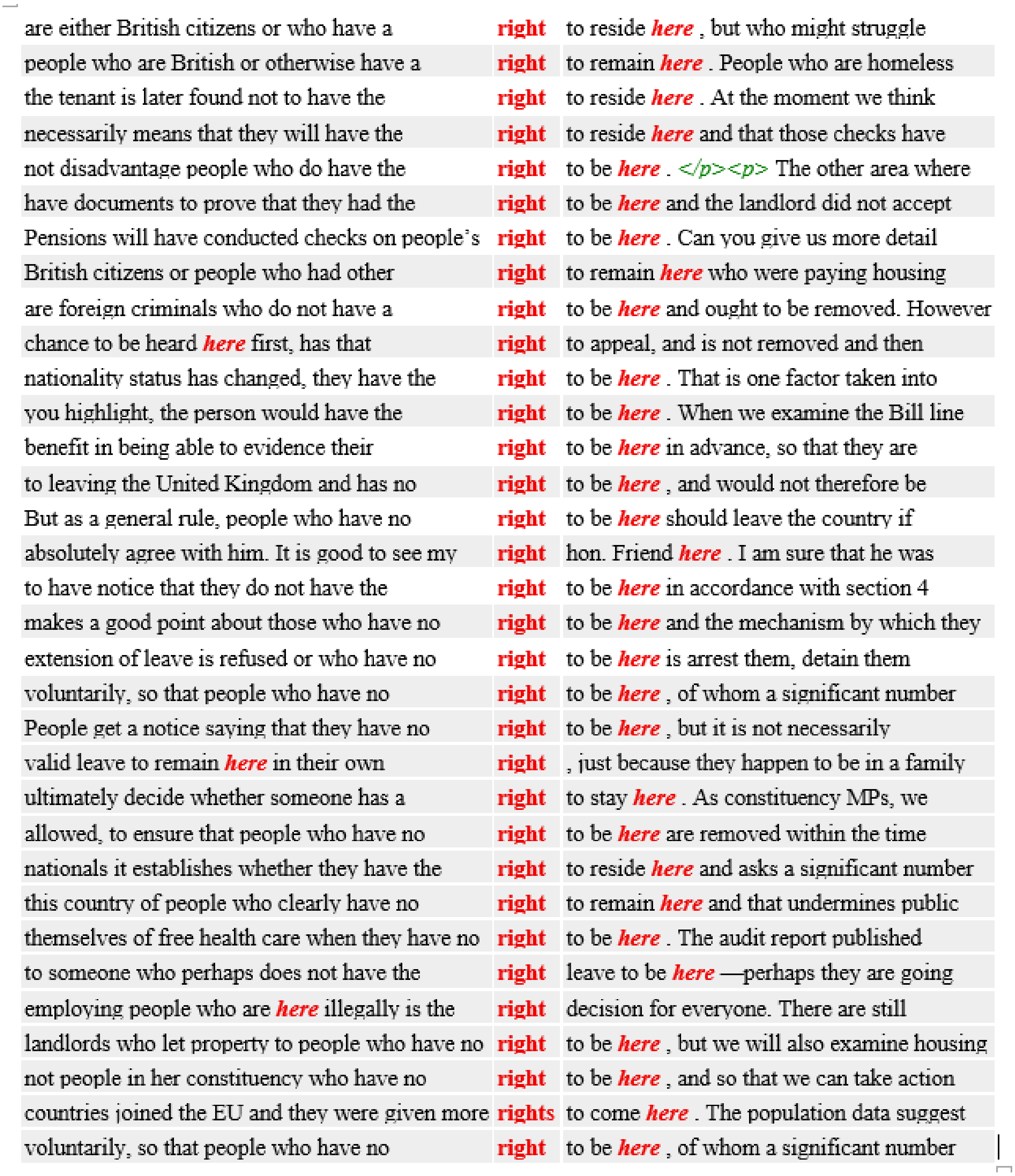
The repetition of the clause "who have no right to be here" and its variants is indeed worth further investigation, as it appears to have become a 'formula' which may have come to the Bill from the language of ordinary citizens, newspapers, speeches by politicians, and may well be interpreted in terms of intertextuality/interdiscursivity (Fairclough 1992) as well as in terms of the *vox populi* strategy whereby other voices are incorporated in the language of politics and find there new resonance (Van Dijk 1993). It comes certainly as no surprise that Home Secretary Teresa May closes her speech in the House of Commons at the second reading of the Immigration Bill on 22nd October 2013 with the words:

> Fixing the immigration system is not something that can be done overnight. There were too many problems with the system that we inherited for that to be possible. However, this Bill will help us further along that road. It is frankly ridiculous that the Government has to operate such a complex system to deal with **foreigners who fail to abide by our laws**. It is ridiculous that the odds are stacked in favour of **illegal migrants**. It is unacceptable that hard working taxpayers have to compete with **people who have no right to be here**.[9]

In more general terms it could be argued that the preliminary cursory investigation of the corpus consisting of the complete Immigration Bill 2014 debate allowed a sort of "distant reading" (Moretti 2013), to borrow a highly evocative recently coined expression in the context of the digital humanities, which can complement and enhance other forms of linguistic investigation. This was for instance the case of the exploration of the loose pattern based on the repeated co-occurrence of the two words "right" and "here", observed in the examples above. The data retrieved from the corpus suggested that this pattern mostly corresponds to a general strategy of (de)legitimation whereby a discursive juxtaposition of two categories was created, i.e the ingroup of British citizens or people who otherwise have a right to be here, and the outgroup of those "who have no right to be here":

[8] http://www.publications.parliament.uk/pa/cm201314/cmpublic/immigration/131105/am/131105s01.htm

[9] https://www.gov.uk/government/speeches/speech-by-home-secretary-on-second-reading-of-immigration-bill



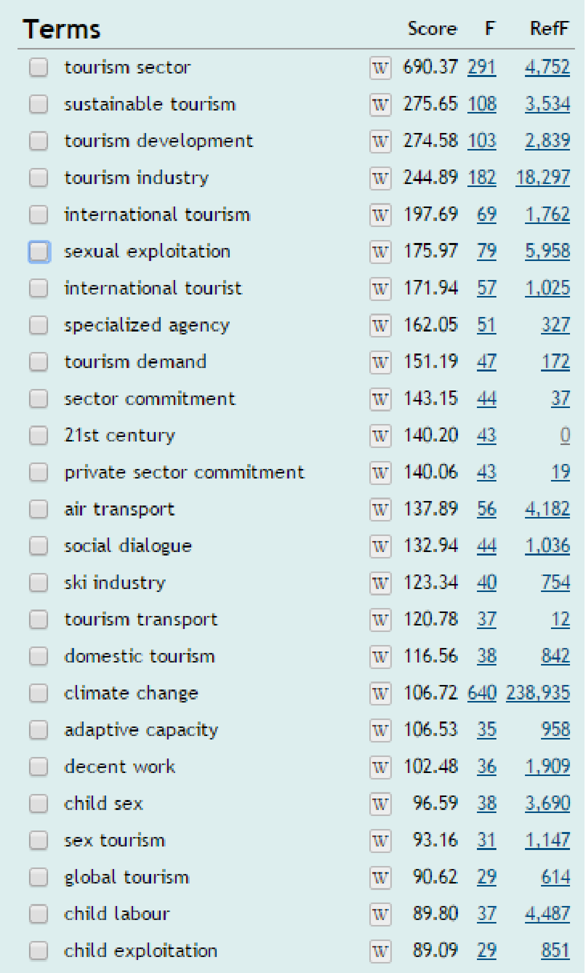
are either British citizens or who have a	right	to reside here , but who might struggle
people who are British or otherwise have a	right	to remain here . People who are homeless
the tenant is later found not to have the	right	to reside here . At the moment we think
necessarily means that they will have the	right	to reside here and that those checks have
not disadvantage people who do have the	right	to be here . </p><p> The other area where
have documents to prove that they had the	right	to be here and the landlord did not accept
Pensions will have conducted checks on people's	right	to be here . Can you give us more detail
British citizens or people who had other	right	to remain here who were paying housing
are foreign criminals who do not have a	right	to be here and ought to be removed. However
chance to be heard here first, has that	right	to appeal, and is not removed and then
nationality status has changed, they have the	right	to be here . That is one factor taken into
you highlight, the person would have the	right	to be here . When we examine the Bill line
benefit in being able to evidence their	right	to be here in advance, so that they are
to leaving the United Kingdom and has no	right	to be here , and would not therefore be
But as a general rule, people who have no	right	to be here should leave the country if
absolutely agree with him. It is good to see my	right	hon. Friend here . I am sure that he was
to have notice that they do not have the	right	to be here in accordance with section 4
makes a good point about those who have no	right	to be here and the mechanism by which they
extension of leave is refused or who have no	right	to be here is arrest them, detain them
voluntarily, so that people who have no	right	to be here , of whom a significant number
People get a notice saying that they have no	right	to be here , but it is not necessarily
valid leave to remain here in their own	right	, just because they happen to be in a family
ultimately decide whether someone has a	right	to stay here . As constituency MPs, we
allowed, to ensure that people who have no	right	to be here are removed within the time
nationals it establishes whether they have the	right	to reside here and asks a significant number
this country of people who clearly have no	right	to remain here and that undermines public
themselves of free health care when they have no	right	to be here . The audit report published
to someone who perhaps does not have the	right	leave to be here —perhaps they are going
employing people who are here illegally is the	right	decision for everyone. There are still
landlords who let property to people who have no	right	to be here , but we will also examine housing
not people in her constituency who have no	right	to be here , and so that we can take action
countries joined the EU and they were given more	rights	to come here . The population data suggest
voluntarily, so that people who have no	right	to be here , of whom a significant number

Figure 5. Concordance lines for the pattern "right" + "here" from the Immigration Bill Corpus

Examples like those briefly discussed above rest on the assumption that there is such a thing as the right of being "somewhere" which is being strongly re-asserted. Thanks to corpus-based evidence obtained from this quickly compiled corpus, "right" and "here" could thus be interpreted in the context of the Immigration Bill as two key lexical items which - by virtue of strong ideological implications ("right") and pragmatic force ("here") – significantly shape ongoing discourse on immigration in the UK.

3.2. Sustainability in the World Tourism Organization

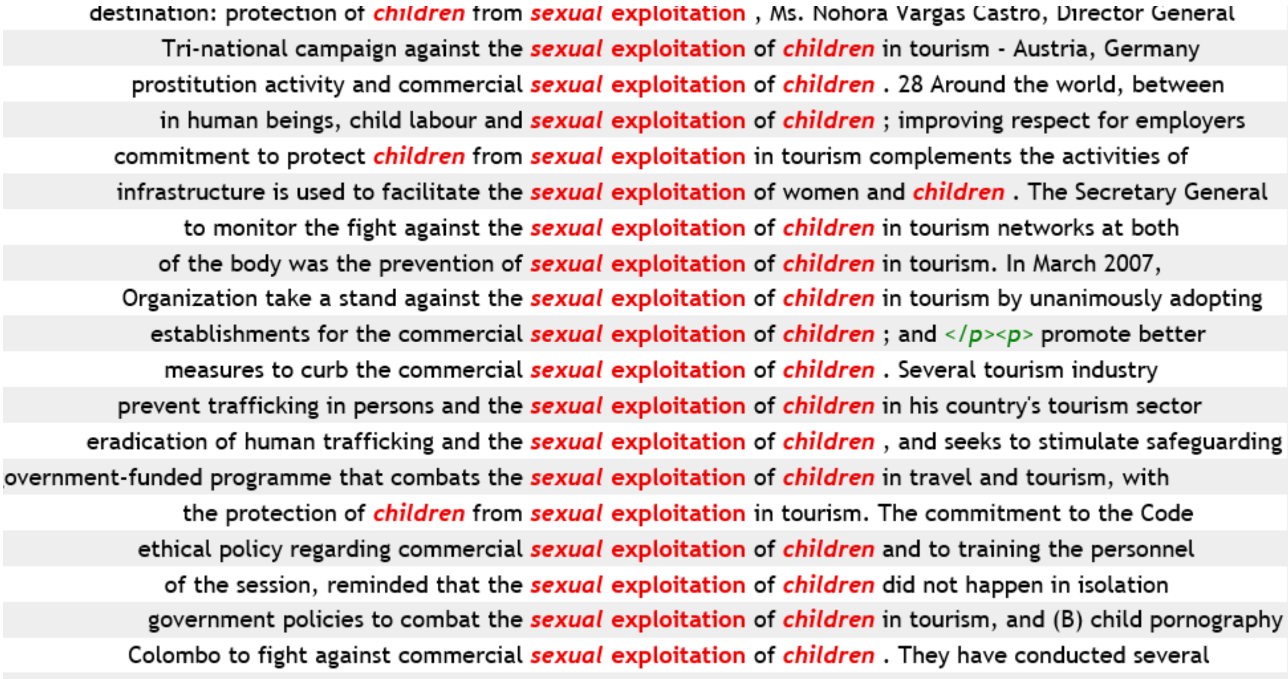
The second case study concerns the compilation of a corpus of texts taken from the official website of the World Tourism Organization (WTO), the UN specialized agency for tourism. The corpus was compiled through the automatic download of a limited number of webpages using the URLs of each single page link available in the “UNWTO A to Z” section in the home page, as a heuristic method to get a small pilot corpus representative of ‘discourse’ within this specific organization. From this small pilot corpus, keywords were automatically extracted by comparison with a very large English reference corpus of English (enTenTen 2012). Using these keywords as seeds, and limiting the crawl to the WTO website itself, the process was run a second time to obtain a larger corpus. From this larger corpus (240.510 words) a number of key terms and patterns were extracted as reported in Figure 7. below:



Terms		Score	F	RefF
□ tourism sector	W	690.37	291	4,752
□ sustainable tourism	W	275.65	108	3,534
□ tourism development	W	274.58	103	2,839
□ tourism industry	W	244.89	182	18,297
□ international tourism	W	197.69	69	1,762
□ sexual exploitation	W	175.97	79	5,958
□ international tourist	W	171.94	57	1,025
□ specialized agency	W	162.05	51	327
□ tourism demand	W	151.19	47	172
□ sector commitment	W	143.15	44	37
□ 21st century	W	140.20	43	0
□ private sector commitment	W	140.06	43	19
□ air transport	W	137.89	56	4,182
□ social dialogue	W	132.94	44	1,036
□ ski industry	W	123.34	40	754
□ tourism transport	W	120.78	37	12
□ domestic tourism	W	116.56	38	842
□ climate change	W	106.72	640	238,935
□ adaptive capacity	W	106.53	35	958
□ decent work	W	102.48	36	1,909
□ child sex	W	96.59	38	3,690
□ sex tourism	W	93.16	31	1,147
□ global tourism	W	90.62	29	614
□ child labour	W	89.80	37	4,487
□ child exploitation	W	89.09	29	851

Figure 6. Key terms extracted from the WTO Corpus

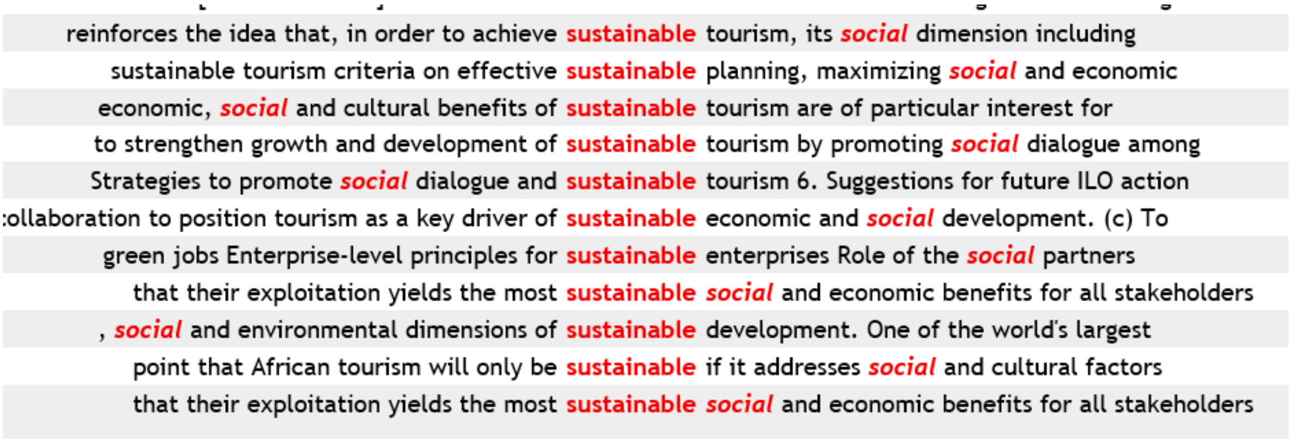
At first glance the key phrases extracted appear to be indicative of the global approach of the WTO to tourism discourse, as the list does not only contain obvious two or three word clusters like “tourism sector” “sustainable tourism”, “tourism development”, “tourism industry”, “private sector commitment”, but also includes a significant number of occurrences for “sexual exploitation”, “social dialogue”, “climate change”, “decent work”, “child labour”, “child exploitation”, which are all together evocative of the wider spectrum of urgent issues which the WTO is to address at global level in the 21st century. For instance, a sample of concordance lines for the phrase “sexual exploitation” immediately suggests that this is a really crucial concern for the WTO, as it is mostly related to the exploitation of children in the sex tourism industry:



```
destination: protection of children from sexual exploitation , Ms. Nohora Vargas Castro, Director General
Tri-national campaign against the sexual exploitation of children in tourism - Austria, Germany
prostitution activity and commercial sexual exploitation of children . 28 Around the world, between
in human beings, child labour and sexual exploitation of children ; improving respect for employers
commitment to protect children from sexual exploitation in tourism complements the activities of
infrastructure is used to facilitate the sexual exploitation of women and children . The Secretary General
to monitor the fight against the sexual exploitation of children in tourism networks at both
of the body was the prevention of sexual exploitation of children in tourism. In March 2007,
Organization take a stand against the sexual exploitation of children in tourism by unanimously adopting
establishments for the commercial sexual exploitation of children ; and </p><p> promote better
measures to curb the commercial sexual exploitation of children . Several tourism industry
prevent trafficking in persons and the sexual exploitation of children in his country's tourism sector
eradication of human trafficking and the sexual exploitation of children , and seeks to stimulate safeguarding
overnment-funded programme that combats the sexual exploitation of children in travel and tourism, with
the protection of children from sexual exploitation in tourism. The commitment to the Code
ethical policy regarding commercial sexual exploitation of children and to training the personnel
of the session, reminded that the sexual exploitation of children did not happen in isolation
government policies to combat the sexual exploitation of children in tourism, and (B) child pornography
Colombo to fight against commercial sexual exploitation of children . They have conducted several
```

Figure 7. Sample of concordance lines for "sexual exploitation" from the WTO corpus

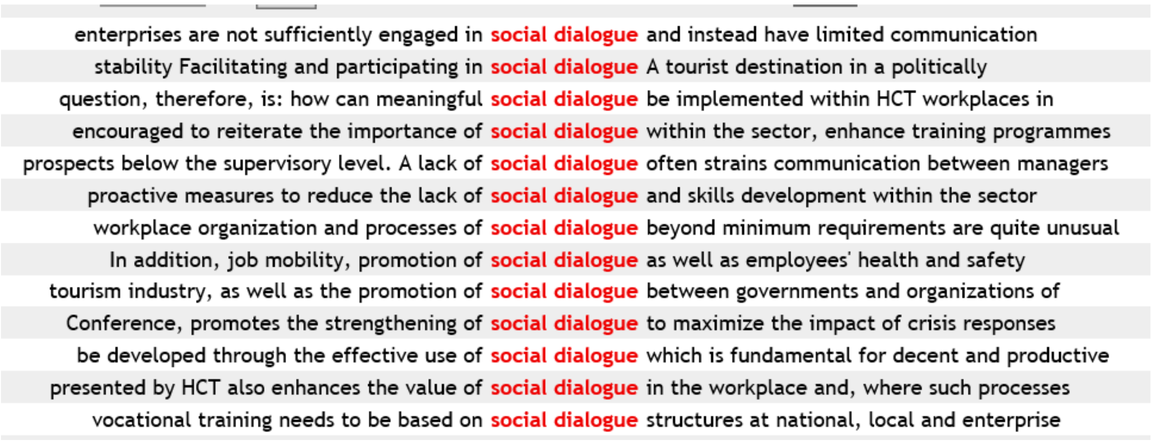
Similarly, such phrases as "social dialogue and "decent work" suggest a strong commitment by the WTO to implement the social dimension of sustainability, which is reflected in its discourse. As a matter of fact, one notices such emphasis on the social dimension of sustainability in a number of patterns of co-occurrence of "sustainable" words relating to the socio-economic domain. Here is a sample of co-occurrence with "social":



```
reinforces the idea that, in order to achieve sustainable tourism, its social dimension including
sustainable tourism criteria on effective sustainable planning, maximizing social and economic
economic, social and cultural benefits of sustainable tourism are of particular interest for
to strengthen growth and development of sustainable tourism by promoting social dialogue among
Strategies to promote social dialogue and sustainable tourism 6. Suggestions for future ILO action
:ollaboration to position tourism as a key driver of sustainable economic and social development. (c) To
green jobs Enterprise-level principles for sustainable enterprises Role of the social partners
that their exploitation yields the most sustainable social and economic benefits for all stakeholders
, social and environmental dimensions of sustainable development. One of the world's largest
point that African tourism will only be sustainable if it addresses social and cultural factors
that their exploitation yields the most sustainable social and economic benefits for all stakeholders
```

Figure 8. Sample of concordance lines for "sustainable" + "social"

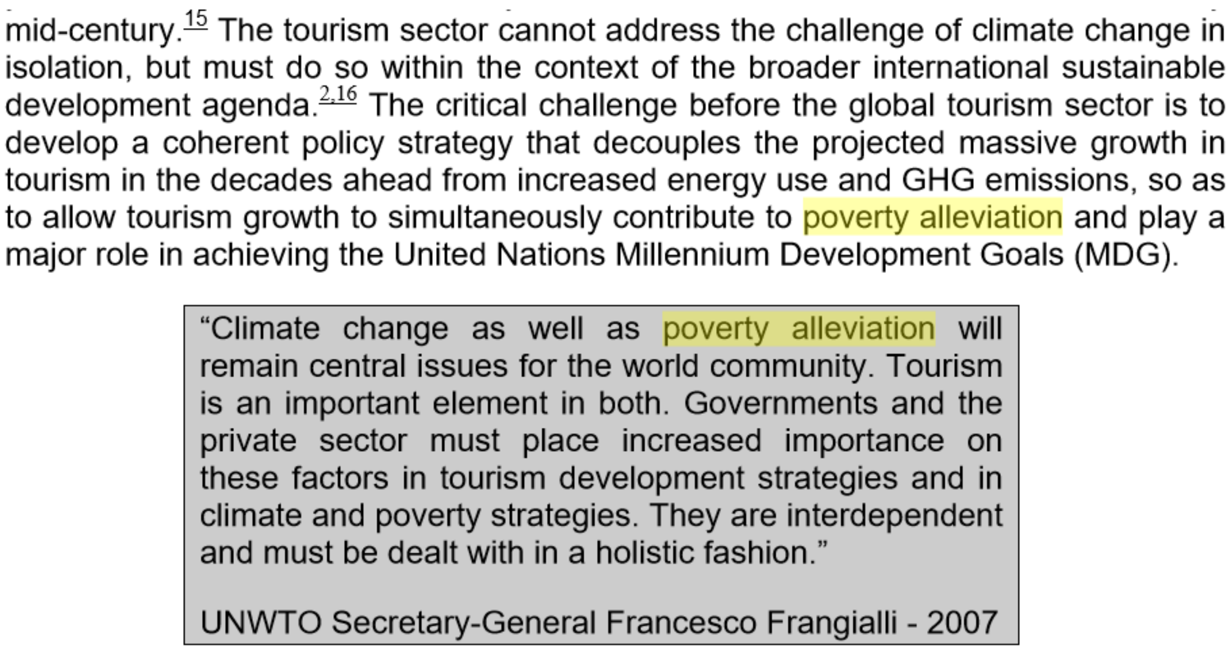
More specifically, concordance lines for "social dialogue", one of the key phrases computed by the system, suggests that this is something which discourse within the WTO is striving to promote, strengthen, or simply call attention to by stressing a lack:



enterprises are not sufficiently engaged in **social dialogue** and instead have limited communication
stability Facilitating and participating in **social dialogue** A tourist destination in a politically
question, therefore, is: how can meaningful **social dialogue** be implemented within HCT workplaces in
encouraged to reiterate the importance of **social dialogue** within the sector, enhance training programmes
prospects below the supervisory level. A lack of **social dialogue** often strains communication between managers
proactive measures to reduce the lack of **social dialogue** and skills development within the sector
workplace organization and processes of **social dialogue** beyond minimum requirements are quite unusual
In addition, job mobility, promotion of **social dialogue** as well as employees' health and safety
tourism industry, as well as the promotion of **social dialogue** between governments and organizations of
Conference, promotes the strengthening of **social dialogue** to maximize the impact of crisis responses
be developed through the effective use of **social dialogue** which is fundamental for decent and productive
presented by HCT also enhances the value of **social dialogue** in the workplace and, where such processes
vocational training needs to be based on **social dialogue** structures at national, local and enterprise

Figure 9. Sample of concordance lines for "social dialogue" from the WTO Corpus

Finally, even the relatively obvious focus on "climate change" in WTO discourse shows that this key concern is not to be seen in isolation but as part of a more comprehensive policy, as in the following excerpt from a publication by the WTO on climate change and tourism, which provides evidence in context of patterns of co-occurrence for such key phrases as "climate change" and "poverty alleviation":



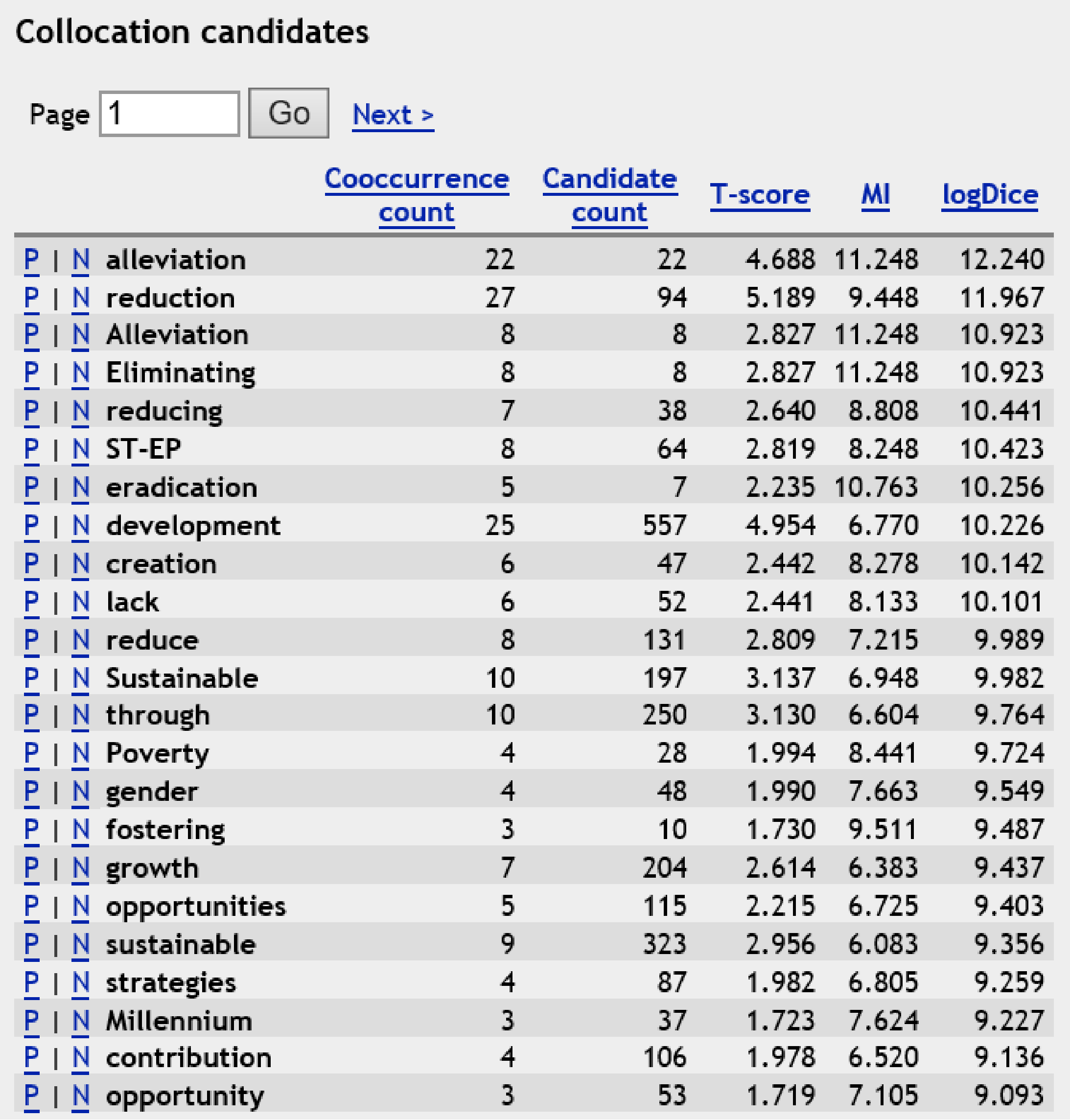
mid-century.[15] The tourism sector cannot address the challenge of climate change in isolation, but must do so within the context of the broader international sustainable development agenda.[2,16] The critical challenge before the global tourism sector is to develop a coherent policy strategy that decouples the projected massive growth in tourism in the decades ahead from increased energy use and GHG emissions, so as to allow tourism growth to simultaneously contribute to poverty alleviation and play a major role in achieving the United Nations Millennium Development Goals (MDG).

"Climate change as well as poverty alleviation will remain central issues for the world community. Tourism is an important element in both. Governments and the private sector must place increased importance on these factors in tourism development strategies and in climate and poverty strategies. They are interdependent and must be dealt with in a holistic fashion."

UNWTO Secretary-General Francesco Frangialli - 2007

Figure 10. A screenshot from the WTO publication Climate Change and Tourism: Responding to Global Challenges (2007) as accessed from the WTO corpus

The presence of the phrase "poverty alleviation" in the context of a publication entitled Climate *Change and Tourism: Responding to Global Challenges (2007)* reveals that poverty is a really key concern in WTO discourse on sustainable tourism development. Indeed, the most frequent collocates for "poverty" in the WTO corpus are "alleviation", "reduction", "eliminating", "eradication", but there is also clear reference to precise projects for sustainable tourism and the eradication of poverty ("ST-EP") and to the Millennium Development Goals ("Millennium") as shown in Figure 12. below.



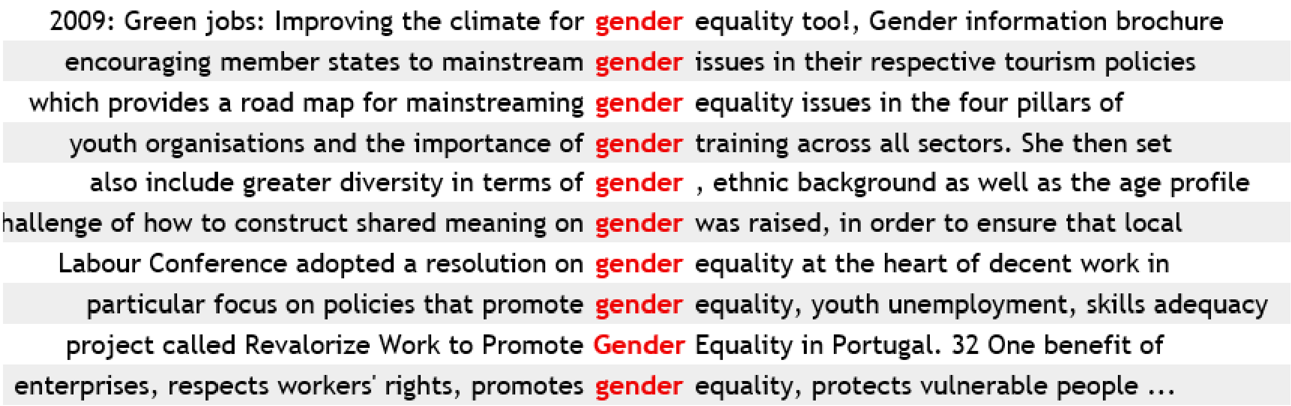
Collocation candidates

Page 1 Go Next >

		Cooccurrence count	Candidate count	T-score	MI	logDice
P | N	alleviation	22	22	4.688	11.248	12.240
P | N	reduction	27	94	5.189	9.448	11.967
P | N	Alleviation	8	8	2.827	11.248	10.923
P | N	Eliminating	8	8	2.827	11.248	10.923
P | N	reducing	7	38	2.640	8.808	10.441
P | N	ST-EP	8	64	2.819	8.248	10.423
P | N	eradication	5	7	2.235	10.763	10.256
P | N	development	25	557	4.954	6.770	10.226
P | N	creation	6	47	2.442	8.278	10.142
P | N	lack	6	52	2.441	8.133	10.101
P | N	reduce	8	131	2.809	7.215	9.989
P | N	Sustainable	10	197	3.137	6.948	9.982
P | N	through	10	250	3.130	6.604	9.764
P | N	Poverty	4	28	1.994	8.441	9.724
P | N	gender	4	48	1.990	7.663	9.549
P | N	fostering	3	10	1.730	9.511	9.487
P | N	growth	7	204	2.614	6.383	9.437
P | N	opportunities	5	115	2.215	6.725	9.403
P | N	sustainable	9	323	2.956	6.083	9.356
P | N	strategies	4	87	1.982	6.805	9.259
P | N	Millennium	3	37	1.723	7.624	9.227
P | N	contribution	4	106	1.978	6.520	9.136
P | N	opportunity	3	53	1.719	7.105	9.093

Figure 11. Sample of collocates for "poverty" from the WTO Corpus

The collocations of "poverty" also indicate patterns of co-occurrence with words which concern issues apparently less obviously related to tourism, which the WTO nonetheless includes and foregrounds. For instance, one of the collocates for "poverty" is "gender", which in turns refers to the question of "gender equality/inequality". And a look at concordance lines for "gender" in this corpus suggests that a concern for gender issues in the tourism industry is part of the organization's overall mainstreaming policy and inclusion strategies, which are connected to all other aspects, including climate change and decent work, as shown in some of the instances reported below:



2009: Green jobs: Improving the climate for	gender	equality too!, Gender information brochure
encouraging member states to mainstream	gender	issues in their respective tourism policies
which provides a road map for mainstreaming	gender	equality issues in the four pillars of
youth organisations and the importance of	gender	training across all sectors. She then set
also include greater diversity in terms of	gender	, ethnic background as well as the age profile
hallenge of how to construct shared meaning on	gender	was raised, in order to ensure that local
Labour Conference adopted a resolution on	gender	equality at the heart of decent work in
particular focus on policies that promote	gender	equality, youth unemployment, skills adequacy
project called Revalorize Work to Promote	Gender	Equality in Portugal. 32 One benefit of
enterprises, respects workers' rights, promotes	gender	equality, protects vulnerable people ...

Figure 12. Sample of concordance lines for "gender" from the WTO Corpus

The preliminary, cursory, exploration of corpus data from the WTO website thus contributes evidence of the holistic approach to tourism which is one of the main goals of the organization. Most of the terms and phrases computed by the system as 'keywords' can be subsumed under the commitment of the WTO to the implementation of sustainability within the tourism industry, which is by no means limited to environmental questions, but includes social and economic aspects too, as indicated by the three pillars of sustainability (social, economic, environmental) acknowledged worldwide.

4. Conclusions

The case studies reported are indicative of the possible benefits of investigating corpus data in the context of research or classroom activities not necessarily centred on corpus linguistics alone. The peculiarity of the corpora focused on in this paper is that they were built through semi-automated methods which reduce to a minimum the time spent in building the corpus. This suggests that these could perform well even as single use corpora to prompt classroom discussion or to provide subsidiary quantitative evidence to complement research carried out with other methods. In the case of the Immigration Bill 2014, the corpus consisting of the complete debate triggered questions about lexical choices in the representation of the category of migrants/immigrants in this specific context, providing evidence of the discursive creation of the opposed categories of people through recurring lexico-grammatical patterns, which could be further investigated with other methods. In the case of the WTO corpus, the simultaneous reading of a large number of texts taken from the organization's official website provided evidence to support a view of a comprehensive commitment to sustainability issues in the context of WTO discourse. It is finally important to stress that - given the topicality of the issues discussed - the two corpora actually created a 'snapshot' of discourse at a certain point in time which could be easily reproduced (using the same criteria) at a different point in time in order to replace/integrate the data, with close to no effort. It is this last feature that, more than anything else, makes compiling and investigating corpora in this way a definitely 'sustainable' approach.

References

Baker et. al. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. «Discourse & Society». 19. 273

Baroni, M. and Bernardini, S. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*. Lisbon: ELDA, 1313-1316.

Baroni, M. and Bernardini, S. 2006. *Wacky! Working Papers on the Web as Corpus*. Bologna: Gedit

Bernardini. S. and Ferraresi, A. 2013. Old needs, new solutions: Comparable corpora for language professionals. In *Building and using comparable corpora,* edited by Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P. Berlin - Heidelberg: Springer, 303 – 319

Fairclough, N. 1992. *Discourse and Social Change.* London: Polity Press

Gabrielatos, C. 2007. Selecting query terms to build a specialised corpus from a restricted-access database. «ICAME Journal», 31: 5-44

Gabrielatos, C. and Baker, P. 2008. Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. «Journal of English Linguistics». 36 (1), 5-38

Gatto, M. 2014. *Web as Corpus. Theory and Practice*. London: Bloomsbury

Kilgarriff, A. et al. 2004. The Sketch Engine. In *Proceedings Euralex*. Lorient, France, 105-116

Moretti, F., 2013. *Distant Reading.* London: Verso

Van Dijk, A.T. 1993. *Elite Discourse and Racism.* Newbury Park, CA: Sage, 1993

Van Dijk, A.T. 2006. *Ideology and Discourse Analysis.* «Journal of Political Ideologies», 11(2), 115-140

Wild, K. et al. 2013. Quantifying Lexical Usage: Vocabulary pertaining to Ecosystems and the Environment. «*Corpora*». 8: 53-79

Zanettin, F. 2002. DIY Corpora: The WWW and the Translator. In *Training the Language Services Provider for the New Millennium*, edited by Maia, B., Haller, J., Ulrych, M. Porto: Facultade de Letras, Universidade do Porto, 239-248

Zanettin, F. 2012.*Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome Publishing

Websites:

Bill stages — Immigration Act 2014 (Dates for all stages of the passage of the Bill, including links to the debates), http://services.parliament.uk/bills/2013-14/immigration/stages.html

Migration in the news interactive chart, http://www.migrationobservatory.ox.ac.uk/data-and-resources/charts/migration-news-interactive-chart
World Tourism Organization, www.unwto.org

Corpus tools:

Sketch Engine, https://the.sketchengine.co.uk
BootCaT, http://bootcat.dipintra.it/