Investigating keyword extraction for identifying units of stance in legislative texts

Ersilia Incelli

Dipartimento Studi geoeconomici, linguistici, statistici, storici per l'analisi regionale
Sapienza University of Rome

Abstract

This paper examines the language and legal terms used in EU and UK legislative texts, in a case study on immigration law. The theoretical and methodological frameworks combine discourse analysis and corpus linguistic techniques, applied to a small corpora of 400.000 words (1988-2009), subsequently divided into two sub-corpora. The analysis investigates the semantic relations around keywords and keyword clusters (ngrams) by carrying out concordance analyses of collocations and observing the frequencies of specific terms, such as *deportation*, *leave to enter*. Arguably, the lexico-semantic/grammatical patterns produced reveal the underlying discourse of a government's stance towards immigration. The analysis also makes use of the related notions of semantic prosody (Louw, 1993), semantic preference and discourse prosody (Stubbs, 2002), for example through an analysis of cultural keywords and the collocates of *alien*, *asylum* and *illegal*, and other node words around the nexus immigrant/crime/terrorism.

Keywords: collocation, immigration, keywords, legal terms, lexico-semantic and grammatical patterns, ngrams

1. Introduction

This paper presents a corpus-driven approach to the study of keywords and multi-word units through collocation and concordance analyses, in order to identify items of meaning which reflect stance in legislative discourse, more specifically in statutory texts from UK and EU immigration law over the period 1999-2009. Arguably, in this case study, the lexico-semantic and grammatical patterns produced can reveal the underlying discourse of a government's stance towards immigration policy, and highlight a government's priorities according to national and cultural interests; these can often stand in the way of the harmonization of EU laws. Although the exploration begins with single keywords and has the study of frequency lists as its driving force, the paper shows that the single unit despite indicating the 'aboutness' of the text (i.e. the main topic of the discourse), it may not uncover all the information needed to retrieve indicators of stance and ideology. Whereas moving from the single word to a phraseology – driven procedure, can reveal how multi-word units tell us more about the specialized discourse as well as the overall ideological undertones, particularly when applying the notions of semantic/ discourse prosody (Louw,1993; Stubbs, 2002) and semantic preference (Stubbs, 2002), i.e. concepts dealing with the semantic connotations of 'node' words and their collocates. As regards the approach, although the findings are the outcome of a corpus-driven exploration, with no previous pre-set notions to the text, once the lists were established and the word units retrieved, the study became more 'corpus-based' in that the exploration moved from the micro-text to the macro-text and back again, exploring previous intuitions, reviewing and confirming them. In this way the analysis can be said to combine 'corpus driven' and 'corpus based' methodological procedures as described by Tognini-Bonelli (2001: 84-87) and Biber (2009: 276).

The work combines corpus linguistic retrieval techniques with well-known theoretical frameworks from discourse analysis. More specifically, the analysis investigates the communicative intent of legal speech acts and the semantic relations around keywords and keyword clusters (ngrams/ congrams) by carrying out concordance analyses of collocations and the frequencies of specific terms, such as: *alien*, *failed asylum-seeker*, *illegal immigrant*, *leave to enter*. In other words, the paper presents a pragmatic study of a highly specialized discourse, and the socio-pragmatic occurrences and pragma-linguistic realization of regulative acts are considered from the point of view of pragmatics.

It is important to specify here that as immigration law and policy is an intensely political issue in the UK and throughout Europe, the approach aims to be as neutral and unbiased as possible in the interpretation of the discourse and data findings; nevertheless it is to an extent from a 'critical' perspective of discourse analysis because of the asymmetrical relationship between the legislative body and the receiver.

The paper will proceed by first a brief explanation of the context and a description of the documents which compile the corpora, followed by the applied theoretical frameworks and the methodological procedures. Section 4 will discuss sample illustrative results, followed by concluding remarks on the pedagogical implications.

2. The context and corpora

2.1. The data and corpora

The study set out to examine and compare the legislative discourse of UK and EU immigration laws (1999-2009). To do this a corpora was built up and divided into two sub-corpora, one consisting of major UK immigration legislation and one consisting of twenty-six EU Regulations and Directives covering various aspects of immigration law for European Member States and citizens (Tab. 1)**.** The laws were accessible via EUR-Lex and N-Lex [1]. The UK corpus, (UK.immig.txt) totaling 201.560 words with a type/token ratio of 3.613/201.560, was made up of 7 large Acts, namely the *Immigration and Asylum Act*, 1999, *Nationality, Immigration, Asylum Act*, 2002, *Asylum and Immigration (Treatment of Claimants) Act*, 2004, *Immigration, Asylum and Nationality*, 2006, *UK Borders Act*, 2007, *the Counter-Terrorism Act*, 2008, and the *Criminal Justice and Immigration Act*, 2008. The EU corpus (EU. immig. txt.) consisted of 26 documents (2000-2009), totalling 178.747 words with a type/token ratio of 6.575/ 178.747, compiled mainly of EU secondary legislation, i.e. Regulations and Directives, (regulations being more binding in law than Directives, which pragmatically affects the communicative speech act and the way the law is implemented by each member state).

UK corpus (1999-2008)	**EU corpus (2000-2008)**
7 documents	26 documents
Total words = 201.560	Total words = 178.747
Type/Token ratio=0,018	Type/Token ratio=0,037
Source N-Lex	Source EUR-Lex

Table 1: Sub-corpora details

The EU Treaty and Charters (primary legislations) were reduced to avoid irrelevant text. Owing to space constraints I cannot list the EU documents here but they included for example: *Council Directive 2001/40/EC on the mutual recognition of decisions on the expulsion of third*

[1] http://eur-lex.europa.eu/it/index.htm.

country nationals [2]. The UK laws were fewer but longer in length than the EU laws (some Acts over a 100 pages), whereas the EU documents were shorter in length, (from 2 to 10 pages). I also reduced the length of some of the Acts in the UK corpus, in particular the *Crime and Immigration Act 2008* and the *Counter-Terrorism Act 2008*, to avoid an overuse of words and terms on Crime and Terrorism. I used the lexis-nexus search terms *immigration*, *immigrant*, *asylum-seeker*, *refugee*, to retrieve the sections of the legislative texts relevant to immigration. However, there is a clear lexical-nexus created by the juxtaposition of 'crime/immigration/terrorism', which draws attention and warrants further investigation.

I can immediately begin forming an hypothesis from the type/token ratio of the corpora. Whilst the UK corpus is larger than the EU corpus, it has a much lower type/token ratio (0,018 v. 0,037 respectively), meaning the UK texts had half the number of variety of words than the EU texts. In effect, this is indicative of the language of legal discourse, in that it reflects the basic nature of legal English and confirms the fact that legal English prefers repetitive language. A change of the traditional wording and traditional formulaic language may increase the risk of unknowingly changing the law, so legal drafters prefer to repeat the same term (sometimes centuries' old) to avoid ambiguity [3]. So the hypothesis that the EU documents may be richer in lexical variety than the UK laws needed to be verified.

2.2. Aims and research questions

As legislative texts are mainly neutral in tone and highly formulaic in nature, it would be interesting to identify the legal forms that reflect a government's attitude on an issue as highly charged as immigration policy. So the overall research aim is to empirically observe the specialized terminology, and identify lexico-semantic patterns which express the socio-pragmatic functions of the law, that is, the regulative and directive speech acts, and consequently 'commit behaviour' (Trosborg, 1995: 31); at the same time identifying the negative or positive connotations of specialized terminology. Further research questions can be defined as: do these units of meaning also construct the identity of refugees and asylum seekers in the UK and EU legislation? What do these word choices reveal about the underlying ideology? Moreover, for our pedagogical purposes, can we relate these features to established theories of lexico-grammatical patterns, such as Sinclair's 'idiom principle' (1987) (i.e. the phraseological tendency, whereby words are co-selected by speakers) or Michael Hoey's 'Lexical Priming claim' (2005) [4].

In actual fact, quite a lot of data was retrieved from the corpora; for the purpose of this paper I will point out only a few illustrative examples which here include the following node words and their collocates: *asylum*, *refugee*, *immigrant*, *illegal*, *alien*; high frequency ngrams including peculiar specialized terminology such as**:** *leave to enter*, *enter or remain*, *withdrawal of support*, *liable to deportation*, *entry clearance*, *third-country national*, *stateless person.*

Finally, the paper attempts to demonstrate how a corpus-driven approach to the retrieval of information from corpora (without preset concepts) can be a source of 'rich repositories of social information and offer considerable potential for research in socio-linguistics and discourse analysis' (Mautner, 2007: 51). We will now turn to the theories of these fields of research in the next section.

---

[2] A full list of the 26 regulations and directives can be obtained on request from the author.

[3] However, as widely known, this sometimes has the opposite effect when 'all inclusiveness' (Bhatia, 1993: 102) creates ambiguity and vagueness.

[4] According to Hoey (2005), all language encountered primes us so that we are likely to use the language in the same way as we encountered it, hence we may be primed to recognize collocations.

3. Theoretical frameworks: a combination of corpus linguistic methodologies and discourse analysis

The paper makes use of a combination of methodologies associated with corpus linguistics (CL) and discourse analysis (DA) and to some extent critical discourse analysis (CDA), owing to the asymmetrical relationship between the legislative body and the audience/receiver of the law (Fairclough, 2001). The work follows a framework similar to that proposed by Baker et al. (2008) in a project on immigration identity in the British media, which in turn follows a corpus-based methodology developed by Partington (2003) in what he defines as a nascent interdisciplinary field of corpus-assisted discourse studies (CADS). The CL methodological approach used in this paper is informed by lexical/pattern grammar (Hunston and Francis, 2000; Sinclair, 2004), and as mentioned above, by the related notions of collocation (Sinclair, 1991), semantic preference (Stubbs, 2001) and semantic/discourse prosody (Louw, 1993; Stubbs, 2001). In brief we can say, semantic preference refers to the semantic rather than evaluative aspects of a word or group of words, *e.g.* 'glass' prefers 'drinks' (Stubbs, 2001: 65). Instead, semantic prosody is an evaluative notion; Louw (1993: 157) defines semantic prosody as «the constant aura of meaning with which a form is imbued by its collocates». For example, in the texts, the verb *constitute* is used predominantly for unpleasant events. Discourse prosody extends over more than one unit in a 'linear string' (Stubbs, 2002: 65). This notion implies that collocates need not to be adjacent to the node for their meaning to influence that node and the whole text (Baker et al., 2008). For example, close collocates of *alien* such as *fingerprint data*, *illegal*, *apprehend*, create a negative connotation of the word *alien*, although the word is not negative in itself.

The DA component of the research is informed by social theory viewing retrieved linguistic data as social practice which reflects and produces ideologies (Habermas, 1967; Fairclough, 2001). However, while this work focuses on combining CL and DA quantitative and qualitative analytical tools, a third theoretical component needs to be mentioned here, if only in brief. The pragmatic element involved in what are essentially *legal speech acts* entails a pragmatic analysis of the relationship between form, utterance function and context starting from the observation of the functions of performatives and directive commands in regulative acts; whereas the CL approach starts from a lexico-grammatical perspective. This paper will focus on the CL strand of the applied theoretical frameworks, to which I will now turn.

3.1. Corpus linguistic theoretical framework and methodology

In actual fact, corpus linguistics makes use of a variety of methods for data retrieval. In this work, the two theoretical notions central to the analysis are *keywords* together with *keyness* and *collocation*. A quantitative and qualitative analysis of the keywords and collocates in the corpora was able to highlight the existence of types of embedded discourse. Examining how such keywords occur in context, their common patterns of co-occurrence and their associated grammatical categories (colligation) contributed to further revelations. Keyness, defined as the statistically significant high frequency keywords or clusters indicating the 'aboutness' (topic) of a text, was useful as a tool for contrastive study of the two sub-corpora (Scott and Tribble, 2006: 55).

Sinclair (1991: 115-117) saw the importance of the collocates of a node as contributing to its meaning, providing 'a semantic analysis of a word', as did Hunston and Francis's (2000) ground breaking work on the use of corpus data to derive abstract, grammatical patterns for the usage of a word. Moreover, Hunston (2002: 109) claimed the context of collocates also 'convey messages implicitly'. In fact, Stubbs (2001) took up the idea of 'cultural keywords' and calls keywords «nodes around which ideological battles are fought» (2001: 188). This can certainly be

the case for highly emotional or value-laden words such as, *immigration, asylum*, *alien*, *illegal*, *third-country national*, *stateless person*, *deportation*, and the collocates around these 'node' words which form a set of multi-word units expressing not only ideology but also metaphorically construed concepts such as, the metaphor of England as a 'container', or Europe as a 'fortress'. This can be seen around words and their (nearby) collocates such as: *borders*, *right/leave to enter or remain*, *leave stamp*, *liable to deportation*, *deportation measures*, *entry clearance.*

4. Methodological procedure and results

I develop a procedure of interpretation that moves from the micro-scale evidence of keywords to key clusters, to beyond the word level, to the retrieval and collection of recurrent word patterns and further beyond the sentence level to a macro-level analysis of multiword units and specialized terminology which serve to identify meaning, subsequently applied to pedagogical settings.

4.1. Software tools

To access the corpus and retrieve the keyword lists, collocations and n-grams, I used the software TALTAC2, (Bolasco et al., 2005). The software retrieved frequency lists and allowed me to compare the texts through an analysis of the 'relative deviation' keywords, that is words which were key in one corpus and not in the other [5] (Bolasco, 1999: 223). I also made use of a semantic annotation system software called Wmatrix (USAS), produced by Paul Rayson (2003). The semantic tagset used by USAS was first based on the Longman Lexicon of Contemporary English by Tom McArthur (1981). It has currently 21 major discourse fields, with further subdivisions of specific areas [6]. In this way, the software identifies key semantic categories or domains, meaning both high and low frequency words can be categorized within one semantic area. Key domain analysis provides a useful tool for capturing low frequency words that (although not key by themselves) do become 'key' when viewed alongside terms with similar meaning [7]. These low frequency words can be just as effective as high frequency words in creating the 'aboutness' of the text and uncovering the underlying discourse, which may otherwise have gone unnoticed in a quantitative analysis. The top semantic domains in this corpora were Law and Order (tag G2.2) in the UK corpus, and Government (tag G1) in the EU corpus.

4.2. Data findings: from keyword analysis to phraseology patterns

Owing to space constraints, I will pinpoint only a few areas of interest and give a few illustrative examples. Overall the set of patterns revealed expressions that are most frequent and most typical of the legal specialised text type. For example, typical coherent connectors used in legal discourse such as *in accordance with* (EU immig.txt) and *by virtue of* (Uk immig.txt. But there were also unexpected results centred on recurring n-grams in the corpus which also uncovered 'implicit' meaning (Hunston, 2002: 109), *e.g.* multi-word units such as, *liable to deportation*, or *fingerprint data*. The exploration facilitated the isolation of patterns which express or introduced evaluative meanings typical of the law. This confirms Stubbs (2001: 215), who stated evaluative meanings «are conveyed not only by individual words but also by longer phrases and syntactic structures».

---

[5] TALTAC2 calls this tool '*scarto standardizzato*'. The formula for the calculation is $z_i = \frac{f_i - f_i^*}{\sqrt{f_i^*}}$.

[6] The full tagset is available online: http://ucrel.lancs.ac.uk/usas/.

[7] See Rayson (2003: 100-113), for a more detailed exploration of the advantages of the key domains approach.

For example, if we take the reoccurring item *constitute*, the word is not negative in itself but the negativity is assigned by its collocates and nearby collocates, or nearby groups of words which convey the evaluative meaning as negative or positive, *e.g.* 'constitute a danger', 'constitute a crime'.

4.3. Key Keywords

As the corpus spanned the ten year period 1999-2009, this created a problem for extracting high frequency words which really were key and constant throughout the whole corpora, and not just key to a particular year or to a particular law, such as the overuse of the word *terrorism* after Sept.11, 2001. Such words could be concentrated in a portion of the corpus and not really represent the whole corpus. Scott (1998: 97-98) calls keywords in a particular period of time 'seasonal collocates'. One way to counter this problem is to calculate "key keywords" or 'consistent collocates' (Scott, 2004: 115), i.e. words which are constant throughout the whole corpus. TALTAC2 calculates *key keywords* using a dispersion analysis which measures a keyword's diffusion throughout the corpus, locating where the word appears, as well as the number of times it appears in each text (Bolasco, 1999) (Tab. 2). Tab. 3 below reveals the resulting key keywords of the two sub-corpus (excluding prepositions and high frequency names). Once the key keywords were established, this was a good starting point to begin the analysis.

Keyword	*Total frequency*	*Asylum 04*	*Borders act 07*	*Counter terrorism 08*	*Criminal 08*	*Immig. 1999*	*Immigration 06*	*Nationality 02*	*Dispersion*	*Use*
section	3.260	428	234	302	1.049	164	238	845	0,942	3072,5
act	2.700	361	212	224	929	107	232	635	0,913	2465,3
person	1.702	228	88	222	460	156	92	456	0,912	1552,9
subsection	1.431	196	137	91	253	131	146	422	0,875	1203,5
may	1.376	184	76	52	561	69	105	309	0,876	1187,4
immigration	1.158	310	102	14	202	59	151	300	0,763	868,4
paragraph	1.157	159	105	36	86	135	80	362	0,794	764,3
order	1.153	132	11	100	459	18	49	185	0,777	741,4
not	1.092	53	79	223	384	55	59	97	0,766	727,7
shall	945	188	89	24	134	3	123	384	0,723	683,5

Table 2: Keyword plot: number of instances in the text in which a word appears in UK corpus (TALTAC2)

UK				*EU*			
Keyword	*Frequencies*	*Keyword*	*Frequencies*	*Keyword*	*Frequencies*	*Keyword*	*Frequencies*
section	3.260	shall	945	member	2.900	council	658
act	2.700	Schedule	936	shall	2.817	national	635
person	1.702	has	825	article	2.166	asylum	630
subsection	1.431	United	593	be	2.152	accordance	624
may	1.376	if	792	state	1.513	UNION	606
immigration	1.158	Secretary	747	states	1.501	treaty	603
paragraph	1.157	appeal	742	this	1.332	are	598
order	1.153	made	717	EUROPEAN	1.169	community	569
not	1.092	Asylum	713	is	1.038	their	565
any	1.063	Provision	665	not	852	residence	521
offence	1.040	after	631	application	739	any	514
part	978	court	583	may	735	has	496
state	963	criminal	560				

Table 3: Top 25 key keywords in UK and EU corpora

4.3.1. Keyword Overview

Keyword analyses makes it possible to compare the two sub-corpora for differences and similarities. We can begin by observing the overuse of the words describing the type of legal text

and format, for example, *paragraph*, *subsection*, *schedule* are predominant in the UK corpus, whereas the EU texts largely refer to *directives*, *article*, *provisions*. The UK corpus overuses *may* and this raises research interest in the pragmatic function and felicity conditions of *may* and its associated high frequency collocate *if*. Whereas the EU corpus prefers deontic *shall* and its close collocate ngram *to ensure that*. The word *immigration* is key to the UK corpus as also *offence*, *criminal* and *order*. The second highest frequency keyword is *person*, but the EU texts use the word *person* less and prefer to use more of a variety of human referentials (see later in this section).

Another useful tool for comparing two sub-corpora is what TALTAC2 calls the 'relative deviation' of keywords ('*scarto standardizzato*'), which calculates words which are key in one corpus and not in the other. Tab. 4 and 5a/b present a few examples, which is just a fragment of the data.

Relative deviation	*Word Form EU*	*Total Frequency EU*	*Word Form UK*	*Total frequency UK*
334,3	states	1.475	states	12
322,8	treaty	579	treaty	2
248,5	nationals	315	nationals	1
205,9	article	2.157	article	65
194,2	measures	349	measures	2
192,7	directive	602	directive	6

Table 4: Sample relative deviation keywords: words key to one corpus and not to the other

EU v. UK				*EU v. UK*			
Rel. Dev.	*Word*	*Tot.freq. EU*	*UK*	*Rel. Dev.*	*Word*	*Tot. freq. EU*	*UK*
29,3	integration	38	1	-24,0	person	276	1.702
50,6	shall	2.814	945	-6,7	support	23	136
67,9	protection	391	19	-11,1	may	713	1.376
45,5	ensure	219	13	-6,5	deportation	1	85
25,9	host	198	0				

5a. *5b.*

Table 5: a/b. Sample illustration of keyword differences between the two sub- corpora

For example, in Tab. 5a we have the words *host* and *integration* whereas *host* has no hits in the UK texts and *integration* is hardly present. This raises hypotheses and assumptions on the underlying discourse in the legislation of each corpus, which can be more closely inspected through expanded concordance lines and a more manual analysis of the texts. The discourse in fact confirms that as regards immigration policy the EU can afford to be less populist than member state countries and more concerned with meeting international requirements on the *protection* and *integration* of *refugees/immigrants/migrant workers*. Whereas, as we can see, the UK has an overuse of the terms *support* and *deportation* (Tab. 5b). A closer observation of the concordance lines for *support* reveal its meaning and use in the UK corpus (Tab. 6). In fact, the UK is traditionally concerned with welfare assistance and support, and the media often emphasizes the economic aspects of immigration (sometimes as an 'economic burden').

State may (a) provide financial	**support**	to an international organisation
on (1) ; (b) provide financial	**support**	to an organisation in the United
on and Asylum Act 1999 (c. 33) (	**support**	: payment to local authority :

Table 6: Concordance lines for 'support'. UK immig.txt

Another key issue associated with immigration is the risk of terrorism and criminal offence which is evident from the reoccurring lexical patterns around collocates associated with *deportation*. A concordance analysis of the keyword *deportation* reveals a highly negative semantic prosody and preference (Tab. 7). Close collocates of *deportation* include: *removal*, *liable to*, *measures*, *subject to*, *criminals*, *pending.*

sentence or order of a court. 35	**Deportation**	or removal : cooperation (1) Th
ment will facilitate the persons	**deportation**	or removal from the United Kingdo
ke provision about the automatic	**deportation**	of criminals under the UK Borders
ion 131 , and (b) is liable to	**deportation**	, but can not be removed from the

Table 7: Concordance lines for 'deportation'- UK immig.txt

4.3.2. Human referentials and 'alien'

Interestingly, UK laws have an overuse of the item *person/persons* (17.827 hits UK v. 276 hits EU). I used Wmatrix's semantic annotation system (Tag S2) to identify the human references in order to see how EU legislation refer to people. The EU corpus uses more of a variety of nominalizations for people, such as: *asylum-seeker*, *refugee*, *third-country national*, *stateless person*,*immigrant*,*migrant*,*member/s*,*host*,*men*,*entrant*,*child*,*children*,*women*,*people*,*individuals*, *population*, *individual*, *human_beings*, *Mr.*, *alien/s.* The word *immigrant* is absent in the UK texts and *migrant* has only 4 hits in the UK corpus. A surprising occurrence was the word 'alien'; although it was not a high frequency word it nevertheless warranted investigation. *Alien* (noun and adjective) appears 40 times in the EU texts and is completely absent from the UK texts.

I used TALTAC2 to carry out a keyword plot of the occurrence of the word *alien* which showed me exactly where and when it appears in the EU directives. I examined the collocates in the concordance lines for its positive and negative semantic prosody. I also investigated the word outside the corpus. The word disappeared from UK legislation in the 1980s and was replaced with *person* or *individua*l [8]. This may have something to do with concern for 'political correctness', and is also related to Brown and Levison's (1987) 'politeness theory'. However, the lemma *alien* continues to appear in European laws. I checked the main connotations of the word *alien* in the BNC (British National Corpus) and apart from the meaning 'something from out of space', and its use as a legal term, the word is frequently used negatively (for example, 'feeling alien and lost'). If we check how it is used in the EU directives we can see the collocation and concordance analyses point to negative connotations. (Tab. 8).

Member State to check whether an	alien	found illegally present on its territory
applicant for asylum and of every	alien	who is apprehended in connection with
fingerprints of all fingers of every	alien	of at least 14 years of age who is

Table 8: Excerpt of concordance lines for 'alien'

If we look at extended text we can see the semantic network creating highly negative discourse prosody; although the collocates of *alien* are not adjacent to the node word, the feel and tone

[8] The 'Aliens Act 1905' was the first immigration law in England regulating 'undesirable' immigrants.

of the text (directive) becomes negative. Nearby collocates such as: *fingers*, *irregular crossing*, *control*, *external border, apprehended*, *different categories of aliens*, are highly evaluative. This also raises the issue of whether the drafters of the EU directive are aware of the negative connotations of *alien*.

Nevertheless, although the UK law is careful to use referentials which imply a more neutral stance in order to produce a calmer tone for a highly charged issue such as immigration policy, close inspection of the collocates around nominalizations such as *asylum- seeker* and *refugee*, have further undertones. As we can see (Tab. 9) the strongest collocate of a*sylum* is *seeker* and the strongest collocates of *asylum- seeker* are the evaluative adjectives *destitute*, *desperate* and *failed*, together with the high frequency multi-word units *withdrawal of support*, *enter or remain*, and specialized terminology such as: *deportation*, *leave stamp* and *entry clearance.*

Third class of ineligible person: failed asylum-seeker 6 (1) Paragraph 1 applies to
12 applicants 43 Accommodation 44 Failed asylum-seekers: withdrawal of support 45
) may confer a discretion." 44 Failed asylum-seekers: withdrawal of support (1)

Table 9: Concordance lines for 'failed asylum seeker' and 'withdrawal of support'

Whereas the strongest collocates for *refugee* are: *Convention* (68 hits) and *status* (14) in the UK texts; and *status* (85), *protection* (6) and *means* (7) in the EU texts. These collocates reveal that *refugee* has more legal status. The law knows the difference between *asylum-seeker* and r*efugee*, (whereas often the terms are confused or used interchangeably in the media). *Migrant* in both corpus has economic status, whereas the item *immigrant* only appears in the EU texts, What is more, *immigrant* in the EU texts appears with *illega*l, as well as with other metaphors usually associated with immigration, such as: *flow*, *influx*, *curb*, *combat, human trafficking*, *smuggling*, *clandestine.*

4.5. Lexical patterns and ngrams

Space allows me to point out only a few illustrative examples. Legislative statutory texts are particularly well suited to congram and ngram extraction (lexical patterns) because of the highly formulaic language of legal discourse. The fixed and formulated terms which make up the specialized discourse of legal English make it easy to identify a high density of reoccurring legal lexical patterns. For example, we can see in Tab. 10 the high number of connectors typically used in legal discourse such as the high frequency ngram *for the purpose of* and *in respect of*, in the UK legal texts; whereas EU directives extensively use the connector *in accordance with.* Unfortunately it is not the purpose of this paper to observe lexical cohesion patterns for textual mapping but this is a possible area for future research in legislative discourse.

As an illustrative example of a reoccurring lexical pattern, we can observe the high frequency ngram *to enter or remain* in the UK texts. As we can see from the title of the laws *Border Act*, *Counter terrorism Act*, Britain is concerned with its 'borders' and this explains the high frequency terms, *entry clearance*, *entry/entrance*, *to enter or remain* and other lexical items related to the pragmatic function of permission to enter or obligation to leave (*deportation*). If we inspect the collocates of *to enter or remain* by expanding the concordance lines and calculating the strongest collocates of this ngram, we get a number of patterns concerned with agents, doers and directives to be carried out (Fig. 1).

the secretary of state	713	immigration and asylum act	325	for the purposes of	303
secretary of state may	182	is amended as follows	105	as follows in	104
in respect of the	96	to enter or remain	89		

Table 10: Top key ngrams with no. of hits in UK immig.txt

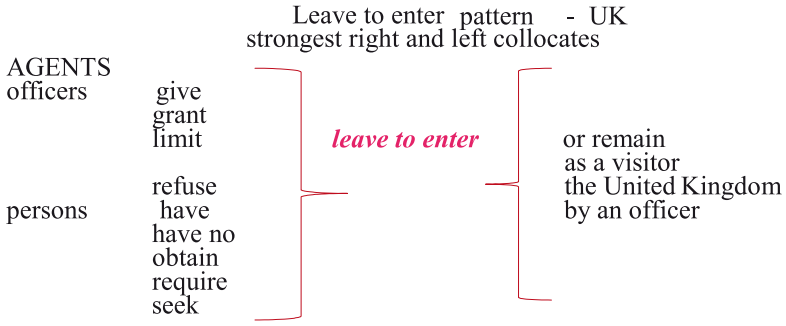

Figure 1: Investigate leave to enter pattern

The example above shows how by following the evidence of corpora, as opposed to intuition, it is possible to discover multi-word expressions that are not yet listed in handbooks of phraseology (in a specific field) and to detect new usages and forms of multi-word units in specialized discourse and this implies pedagogical applications.

5. Concluding remarks

The paper has attempted to show how CL tools used for text mining and the retrieval of information can supplement and inform traditional methods in discourse analysis. In particular, we see how keywords, identified for evaluative purposes, such as in this study the lexical items *constitute*, *deportation*, *failed asylum*, can reveal a great deal about reoccurring patterns and uncover underlying discourses and meaning, which may go unnoticed using only a manual procedure. This paper has also tried to show how semantic key domains can capture words that, because of their low frequency, would not be identified as keywords in their own right, but contribute to the stance of a text, as we saw with the example of alien and its near-by collocates.

The analysis also confirmed the hypothesis that EU institutional legislative language uses more of a variety of lexical items, in a way similar to the lexis found in the media or political rhetoric. In fact, there is a trend towards standard Plain English usage in EU legislative texts, to facilitate the interpretation of legislative meaning; nevertheless this may run into the problem of rendering the correct semantic meanings of lexical items, avoiding the negative connotations of words.

As regards pedagogical applications, the implications are self-evident. *Ngramming* and *congramming* (lexical, semantic/pragmatic, grammatical patterns) can serve as a tool for textual analysis, and can be used to help raise student awareness of the 'idiom principle' (Sinclair, 1987; 2004; Cheng et al., 2006), in that it helps students to find co-occurring words and 'chunks' in general as a suitable tool to enable students to master the discourses and genres of their specific disciplines (Mauranen, 2006; Bhatia, 2004). It is important that students are made aware of the semantic meaning and connotations of a specialized term and just the isolated denotation of a word or word groups. Students need to be exposed to specialized legislative texts and language learning exercises which raise their awareness of specific phraseology and highlight the use and connotative meaning of a specialized term. Corpus linguistics is potentially useful as a pedagogical tool for showing and unravelling how and in what ways selected 'strings of words' serve specific discourse purposes and perform 'implicit or explicit' verbal acts in specialized contexts. According to Römer and Schulze (2008), scrutiny of patterns in specialised discourse 'is still in its infancy', and the specialised meaning and the relationships expressed and encoded by patterns or phraseological items are inseparably intertwined with the particular domains in which they are produced. All in all, the paper has tried to provide a view on the way specialized

knowledge is encoded and expressed and organized in specialized corpora through computer assisted language analysis as a way into texts through mining texts.

References

Baker P., Gabrielatos C., Khosravinik M., Krzyzanowski M, McEnery A. and Wodak R. (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 and Society*, 19 (3): 273-30.

Bhatia V. (1993). *Analysing Genre- Language Use in Professional Settings*. London: Longman.

Bhatia V.K. (2004). *Worlds of Written Discourse. A Genre-Based View*. London: Continuum.

Biber D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 37: 275-311.

Bolasco S. (1999). *L'analisi multidimensionale dei dati.* Roma: Carocci.

Bolasco S., Baiocchi F. and Morrone A. (2005). *TaLTac2*. Roma: CISU.

Brown P. and Levinson C.S. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Cheng W., Greaves C. and Warren M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11, 4: 411-33.

Fairclough N. (2001). *Language and Power*. (2nd edition). London: Longman.

Habermas J. (1967). *Theorie und Praxis*. Neuwied am Rhein: Luchterland.

Hoey M. (2005). *Lexcial Priming: A new theory of words and language.* London: Routledge.

Hunston S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston S. and Francis G. (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam-Philadelphia: John Benjamins.

Louw B. (1993). The diagnostic potential of semantic prosodies. In Baker, M., Francis, G. and Tognini-Bonelli, E., editors, *Text and Technology: In honour of John Sinclair*, Amsterdam: John Benjamins, pp. 157-176.

Mauranen A. (2004). Spoken - general: Spoken corpus for an ordinary learner. In Sinclair, J.McH, editor, *How to Use Corpora in Language Teaching*, Amsterdam: John Benjamin, pp. 89-105.

Mautner G. (2007). Mining large corpora for social information: The case of elderly. *Language in Society*, 36: 51-72.

McArthur T. (1981). *Longman Lexicon of Contemporary English*. London: Longman.

Partington A. (2003). *The Linguistics of Political Argumentation: The Spin-doctor and the Wolf-pack at the White House*. London: Routledge.

Rayson P.(2003). *WMatrix/USAS – Semantic Annotation System*. Lancaster: University of Lancaster.

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

Römer U. and Schulze. R. (editors) (2008). Patterns, Meaningful Units and Specialized Discourses. Special issue 13(3) of the *International Journal of Corpus Linguistics*. Amsterdam: John Benjamins.

Scott M. (1998). *WordSmith Tools Help Manual, Version 3.0*. Oxford: Mike Scott-Oxford University Press.

Scott M. (2004). *Wordsmith Tools version 4*. Oxford: Oxford University Press.

Scott M. and Tribble C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education.* Philadelphia: John Benjamins.

Sinclair J. (1987). The nature of the evidence. In J. McH. Sinclair, editor, *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins, pp. 150-159.

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair J. (2004). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stubbs M. (2002). *Words and Phrases*. *Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Tognini-Bonelli E. (2001). *Corpus linguistics at work*. Amsterdam-Philadelphia: John Benjamins.

Trosborg A. (1995). Statutes and contracts: An analysis of legal speech acts in the English language of the law. *Journal of Pragmatics*, 23: 31-53.