Selecting query terms to build a specialised corpus from a restricted-access database*

Costas Gabrielatos
Lancaster University

Abstract

This paper proposes an accessible measure of the relevance of additional terms to a given query, describes and comments on the steps leading to its development, and discusses its utility. The measure, termed relative query term relevance (RQTR), draws on techniques used in information retrieval, and can be combined with a technique used in creating corpora from the world wide web, namely keyword analysis. It is independent of reference corpora, and does not require knowledge of the number of (relevant) documents in the database. Although it does not make use of user/expert judgements of document relevance, it does allow for subjective decisions. However, subjective decisions are triangulated against two objective indicators: keyness and, mainly, RQTR.

1 Motivation and central issues

The primary motivation for examining issues related to query formulation and expansion was the need to compile a corpus for the ESRC funded project entitled *Discourses of refugees and asylum seekers in the UK Press 1996–2006*, which aims to explore the discourses surrounding these groups and to account for the construction of their identities in the UK press.[1] Further motivation was provided by the idiosyncrasies of the online database from which the texts would be retrieved. Specifically, the database interface imposed certain access limitations, such as the number of documents returned for each query, and information regarding the number of database documents matching a given query (more details follow later in this section). Although the discussion will draw on the work carried out as part of the project, the technique presented in this paper can be employed in a wider set of circumstances, for example, in instances when the idiosyncrasies and restrictions outlined in this section do not apply.

When compiling a specialised corpus from a text database by use of a query, there is a trade-off between precision and recall (e.g. Chowdhury 2004: 170). That is, there is a tension between, on the one hand, 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. Seen from a different perspective, the trade-off is between a corpus that can be deemed incomplete, and one which contains noise (i.e. irrelevant texts). In the former case, some aspects of the use of, or relations between, terms and/or concepts may be underrepresented or missed – depending on the size of the corpus in relation to the body of relevant available data. In the latter case, statistical results may be skewed (notably keyness),[2] and the corpus building, as well as any mark-up and annotation, can become unduly time consuming. It would be helpful, therefore, to use objective indicators of the degree to which a candidate query term is expected to return relevant documents, or, to be more precise, the degree to which the addition of a term to the query results in the addition of relevant documents. Such indicators would then inform decisions regarding the terms to be included in the query.

In order for the term ‘relevant document’ to have any meaning, the compilers of a specialised corpus need to define what the corpus would ideally contain, and then “adjust [their] parameters” according to what is feasible under the particular circumstances (Sinclair 2004: 81). An obvious starting point for the compilation of a query is lexis denoting the entities, concepts, states, relations or processes that are to be investigated (e.g. Chowdhury 2004: 169). With regard to the particular project, the best starting point seemed to be the title and description of aims – which also settled the question of the source of the texts. In this light, two core query terms seemed to suggest themselves, *refugee(s)* and *asylum seeker(s)*, leading to the following core query: ‘*refugee* OR asylum seeker**’. The decision to use these two terms as the core query may be considered subjective; however, given the clearly defined purpose for the corpus compilation, their selection was at least inescapable, and arguably objective within the project parameters (issues of subjectivity/objectivity are revisited in sections 2 and 3). Of course, a corpus built using only this core query would yield very useful insights (see Baker and McEnery 2005), particularly given the ten-year span of the texts comprising the corpus. It is estimated that the core query alone, used on a database of twelve UK national newspapers from 1996 to 2005, would yield a corpus of 35–40 million words. However, since one of the aims of the project is to build on existing research, it seems appropriate to examine the feasibility of compiling a richer corpus.

One argument for a richer corpus is that some terms may have overlapping uses. Baker and McEnery (2005: 201) report that although the website of the Office of the United Nations High Commissioner for Refugees "is focussed around refugees … there were still a number of references to asylum seekers …, suggesting that the two identities share a common ground". This observation seems to be supported by an aspect of measuring query term relevance (see section 2 and note 9). Terms may also be related in terms of sequence or change of state. For instance, the status of persons may change from *asylum seekers* to *refugees*, or vice versa, according to the definition adopted. Dictionary definitions present an asylum seeker as a refugee who has applied for asylum, and so imply the sequence '*refugee* → *asylum seeker*', whereas the definitions of the Refugee Council[3] imply the opposite sequence (see Table 1:).

Table 1: Definitions of *refugee* and *asylum seeker*

	refugee	***asylum seeker***
Longman dictionary of contemporary English on CD-ROM (2003)	Someone who has been forced to leave their country, especially during a war, or for political or religious reasons.	Someone who leaves their own country because they are in danger, especially for political reasons, and who asks the government of another country to allow them to live there.
Refugee Council	Someone whose asylum application has been successful and who is allowed to stay in another country having proved they would face persecution back home.	Someone who has fled persecution in their homeland, has arrived in another country, made themselves known to the authorities and exercised the legal right to apply for asylum.

Conversely, terms with almost identical dictionary definitions may be used less interchangeably than expected, as is the case of *immigrant* and *migrant*. More importantly, the terms *refugee(s)* and *asylum seeker(s)* are frequently used interchangeably with the terms *immigrant(s)* and, less so, *migrant(s)* (e.g. Greenslade 2005: 5). Thus, one newspaper may describe a person as an asylum seeker, whereas another may refer to him/her as an (illegal) immigrant. The subsequent collocational analysis has showed considerable overlap of the collocates of *refugees/asylum seekers* and those of *immigrants/migrants*, which indicates an overlap in usage (Gabrielatos and Baker 2006). Wilson (2006: 13), employing critical discourse analysis to examine 242 articles from Scottish newspapers, came to the same conclusion. In that light, it seems worthwhile to add such related terms to the query.

The addition of query terms relevant to the core terms seems to also be supported by the observed tendency for representations of groups in the press to "include or exclude social actors to suit their interests and purposes in relation to the readers for whom they are intended" (van Leeuwen 1996: 38). In other words, even if an article reports on or discusses issues related, directly or indirectly, to refugees or asylum seekers, these two groups may not necessarily be referred to explicitly. If, however, the query string includes as many other terms as possible referring to the same or similar groups, then it is expected to capture a large proportion of those articles in which the groups in question are not mentioned explicitly.

Further support for the addition of relevant query terms comes from the methodology to be used in the data analysis. The analysis involves the examination of the collocations and resulting lexical networks of the core terms *refugee(s)* and *asylum seeker(s)*, and the interrelations in meaning/use that they may reveal. Arguably, these interrelations will potentially become clearer if the study could also take into account the collocational patterns and lexical networks of the related terms. For example, *terrorism* registers as a very strong key word when two sample corpora drawn from the database using the core query are compared to the written BNC Sampler.[4] That is, *terrorism* seems to be strongly associated with topics related to the terms *refugee(s)* or *asylum seeker(s)*, or, at the very least, to be present in texts containing one or both of these core query terms. It would be helpful, therefore, to examine what other terms (i.e. entities, concepts, states or processes) *terrorism* tends to be associated with in the corpus. Another example is the case of *asylum*. As one of the groups in focus is those who seek asylum, it seems beneficial to examine its collocational networks in the corpus to be constructed, in order to examine possible links between its different uses. These relations can, of course, also be examined in a representative general corpus, but there are also arguments for examining such associations within the same corpus. The collocational relations established within the specialised corpus can yield additional insights, as they would reveal the use of the term *terrorism* not in a diverse (albeit representative) range of genres and text types, but in the same clearly specified range of texts in which the associations of the core terms themselves were also established (see McEnery 2006). To put it simply, the associations would be compared against the same background.

In sum, additional query terms would ideally return articles which do not contain the core query terms, but are either about the groups denoted by the core terms, or about groups, processes, etc. which are treated as being related to them. However, if such related terms also return a disproportionate number of articles irrelevant to the core query terms, then their addition to the query would render the compilation of a specialised corpus unnecessarily time consuming, or, in the case of the present project, impracticable. For example, the addition of *terrorism* alone to the core query results in a six-fold increase in the size of a sample corpus spanning thirty days, which translates into a 50–100 per cent increase in the time needed to collect the documents (see also section 2). It seems clear, then, that, desirable as it may be, compiling a corpus containing all terms related, to any degree, to the core query is impossible under the circumstances. This brings us back to the issue of the principled selection of query terms, to which we will now turn.

2 Query term selection

There are simple formulas which calculate the degree of precision and recall of a query, as Figures 1 and 2 show (Baeza-Yates and Ribeiro-Neto 1999: 75). |Ra| is the set of retrieved relevant documents, |A| is the set of retrieved documents, and |R| is the set of relevant documents in the database.

$$\boxed{\textbf{Precision} = \frac{|\boldsymbol{Ra}|}{|\boldsymbol{A}|}}$$

Figure 1: Calculation of Precision

$$\boxed{\textbf{Recall} = \frac{|\boldsymbol{Ra}|}{|\boldsymbol{R}|}}$$

Figure 2: Calculation of Recall

However, these formulas are not applicable to the present case, as the number of relevant documents in the database is unknown. The same applies to more complex models, such as *best match searching* and *relevance feedback* (see Chowdhury 2004: 180–182). Also, establishing the relevance of the additional documents retrieved by the candidate terms is exactly what is sought here. Assessing the relevance of each candidate term by reading (a sample of) the documents returned by the addition of each candidate term to the core query, as in the case of *user relevance feedback* (e.g. Buckly, Salton and Allan 1994: 292;

Baeza-Yates and Ribeiro-Neto 1999: 75) introduces more subjective decisions, irrespective of whether a number of judges are involved (e.g. Belew and Hutton 1996), or, as in the *vector processing model*, the documents are returned in order of relevance, either based on the number of query terms in the returned documents, or on the indexing of documents in the database (Chowdhury 2004: 176–180). In fact, reliance on indexing can exclude relevant documents, not only because it is unlikely that the indexing was carried out with the particular project in mind, but also because even metaphorical uses of the core terms, which may not be indexed as relevant, are considered relevant for the purposes of the present project. For example, although *refugees* is a database index term, and documents are returned with a weight on the index term relevance, not all documents containing the word 'refugees' are so indexed, presumably because it was decided that this group was not one of the main topics in the document. Similarly, approaches to establishing the probability of relevance of query terms also rely on knowledge which, in the present case, was unavailable, or would be prohibitively time-consuming to acquire, such as the number of documents in the database, the number of words in the collection, the number of relevant documents for a given query term, or the frequency of each term in each document examined for relevance (e.g. Roberstson and Sparck Jones 1976; Boughanem et al. 2006). The approaches outlined above would also be impractical in view of the interconnected project-specific constraints relating to the number of candidate terms (more than 100), as well as the available time, finances and human resources (see also Baroni and Bernardini 2003) – particularly as the corpus was a means to an end.

Another reason why techniques developed within the field of information retrieval are not entirely helpful in this case may lie in the need to use Boolean queries, which the database interface operates with, as "the Boolean model is in reality more a data (instead of information) retrieval model" (Baeza-Yates and Ribeiro-Neto 1999: 26). The distinction between data and information seems pertinent to the project, as it is data that is sought, data which will be analysed for the information described in the project aims. That is, the corpus needs to contain articles relevant to refugees and asylum seekers (and related groups) without any selection bias regarding the content of articles (i.e. the information given or the stance adopted in them). Attempting to retrieve documents containing specific information can impose bias on the data collection and, consequently, on the study outcomes. A further reported drawback of Boolean queries is that they do not allow for relevance ranking of the retrieved documents (Chowdhury 2004: 174). However, this does not pose a problem for our purposes; on the contrary, it simplifies matters. Once a term is deemed relevant, any article containing it is also deemed relevant, and a document containing a single relevant query term is considered as relevant as one containing two or more (another reason why indexing is not helpful in this case). This is because even metaphorical or humorous uses of relevant terms are desirable, in that they can provide insights into the representation of the two groups.[5] At this juncture, the literature on compiling corpora from the web seems worth investigating.

The compilation of corpora from the world wide web involves the use of an API (Application Programming Interface), that is, a service which allows third-party software access to a search engine's index of web pages. The first step is to decide on an initial set of terms (or *seeds*) which are expected to return relevant texts, irrespective of whether relevance is defined in terms of genre, topic or language (e.g. Ghani et al. 2001; Baroni and Bernardini 2004; Baroni and Sharoff 2005). These initial terms are combined randomly in equal sets (e.g. pairs or triplets) to be used as queries (Baroni and Bernardini 2004: 1314). The documents retrieved from each query (or a portion of them) are used to compile a pilot corpus. The corpus derived thus is compared to a reference corpus to establish keywords in the pilot corpus – Baroni and Sharoff (2005) suggest using the 40 top keywords. A random sub-set of these keywords is used to form new sets of queries. This procedure is repeated as required, although Baroni and Bernardini (2004: 1314) report that they did not have to repeat the procedure more than two or three times. This technique seems to be an adaptation of relevance feedback. Instead of users or experts reading (a sample of) the documents to assign a relevance score, the decision is largely reached through successive keyword comparisons. However, Baroni and Bernardini (2003: 4, 2004: 1314) acknowledge that the number of initial terms, the cut-off point for the use of key words as interim query terms, and the selection of documents for each pilot corpus are subjective decisions, sometimes based on trial and error. In sum, the procedure may not be entirely objective; it is, however, free from the decisions of human readers. Given the availability of a set of software tools (*BootCaT*) which would automate the procedure (Baroni and Bernardini 2003; Baroni et al. 2006), this technique would be considered promising for our purposes. However, the tools cannot be applied to the particular database.

The techniques used here (see section 3) adapt and combine elements of the procedures outlined above. Candidate terms are selected through a keyword comparison of a pilot corpus of database documents returned by the core query and a representative corpus of British English;[6] however, introspectively selected candidate terms were also tested for relevance. As regards relevance, the focus was shifted from the relevance of documents to the relevance of additional query terms, that is, the degree to which they are found in the same documents containing one or more of the core query terms. In other words, reading-based decisions were replaced by an indicator (RQTR) reflecting the number of additional documents returned by the addition of each candidate term to the core query.

The use of pilot corpora, rather than a corpus compiled by applying the core query to the whole sub-section of the database required for the project (in terms of newspapers and time span), was dictated by the constraints imposed by the database interface, which do not appear to be unique to the database used to compile the corpus, and are not dissimilar to those imposed by search engine APIs.[7] The database interface imposed restrictions on the number of documents returned for each query, as well as the number of documents that could be downloaded at a time, and gave no information about the number of documents that would be retrieved in absence of the restrictions. Combined, these restrictions make it impossible to establish the number of documents when a query returns more documents than the limit, without breaking down the query time span into smaller units (a hit-and-miss affair). As an indication of the time investment that working under those restrictions would entail, consider the case of using only the core query, *refugee* OR asylum seeker**, as initial seeds. Extrapolating from the frequency of the two terms and the number of documents in the corpus, it was estimated that repeating the procedure with combinations of the 40 top keywords even only twice, which is the minimum number of repetitions that Baroni and Bernardini (2004: 1314) report, would take up at least one-third of the time available for the project – clearly, an inordinate amount of time.

Due to these restrictions, the pilot corpora used in the process of determining the query to be used for the corpus compilation were not based on all the texts available in the database over the period in question. The first pilot corpus (henceforth UK1) contained articles published between 11 September and 10 October 2005 (342,590 words); the second (henceforth UK6) contains articles published during six random months spanning the duration of the intended corpus: October 1996, December 1998, February 2000, April 2002, June 2004, August 2005 (2,658,184 words). Both corpora comprised texts from twelve UK national newspapers returned using the core query string. UK6 is balanced more towards the present in order to preserve the balance between broadsheets and tabloids, as the database contains a higher proportion of broadsheets before 2000. The reference corpora used for the keyword comparison were derived from the BNC (Aston and Burnard 1998): the written BNC Sampler (henceforth BNC-S; 1,082,171 words), and the newspaper sub-corpus of the written BNC (henceforth BNC-N; 9,670,226 words).

A second reason for adapting the process used to build corpora from the world wide web is the ephemeral nature of many entities, events, etc. referred to in newspaper articles. For many of the top forty keywords it is not apparent that they have registered statistical significance because of their relation to the perception of the nature/status of refugees and asylum seekers. Rather, these terms seem to have been important during the time period covered in the pilot corpus (e.g. *darfur, hurricane, iraq, orleans, wolfgang*), referred to groups with a long-standing relation to the core terms (e.g. *jewish, palestinian*), referred to political entities relevant to the UK (e.g. *blair, eu, labour, tony*), or were not specifically related to the core terms (e.g. *killed, police, war*).[8] In the same vein, keyword comparisons with a reference corpus that is not contemporary with the pilot corpus are bound to favour words referring to entities, concepts, etc. which were not current in the period represented by the reference corpus. In our case, the two corpora do not even overlap: the BNC contains documents up to 1994, whereas UK6 spans 1996–2005.

We also need to consider whether, irrespective of their time-specificity, some words registered keyness not because they are related to refugees or asylum seekers, but because UK6 comprises newspaper texts, whereas BNC-S is a general corpus. To establish whether keywords are news-specific rather than specific to the core query terms, UK6 was also compared to BNC-N. The comparison showed considerable overlap between the two keyword lists: 60 per cent in the top 100 content keywords, rising to 80 per cent in the top 40. This seems to indicate that keyness is more the result of the key terms' relation to the core terms than their specificity to newspaper articles. However, this does not diminish the possibility that the keyness of a large number of words was mainly due to the different time spans.

Finally, a keyword analysis effectively treats the compared corpora as single texts. As a result, some words may register keyness because they have very high frequencies in a relatively small number of documents, even if this clustering is not representative of the majority of documents in the corpus. In addition, and particularly for this project, this characteristic also tends to boost the keyness of words in broadsheets, as, on average, articles in them are usually much longer than in tabloids (in the corpus, articles in broadsheets are on average 46.6% longer than in tabloids). A technique which can solve these problems is the calculation of *key-keywords* (Scott 2004: 115), that is, words which are key in a number of texts in a corpus, in order to establish *associates*, that is, "key-words associated with a key key-word" (ibid.: 109). This would be a helpful technique if it were not prohibitively time-consuming in the present context, as it would entail downloading one document at a time (UK6 contains almost 4,000 documents), as files were downloaded in batches of up to 200 documents. More importantly, the technique does not bypass the problem of time-specific keywords outlined above.

These considerations suggest that keyness alone may not always be a good indicator of the suitability of candidate query terms. For instance, including any single one of the terms mentioned above (e.g. *blair, hurricane*, *palestinian*) would decrease recall and create an overlarge corpus, without necessarily increasing precision enough for the inclusion to be justified. Furthermore, using the iterative process described above to objectively discard these terms would require an investment in time which does not seem justified within the constraints of this project, particularly when considering the document download restrictions of the database. In this light, it might not seem unreasonable to opt for examining the lists of key n-grams and choosing those which are consistent with our subjective assessment of candidate term relevance. However, since introspection, on its own, is not a reliable indicator, it would be best to introduce a second objective method of measuring candidate query term relevance, which would then be used as a means to triangulate decisions regarding additions to the query.

3 Measuring candidate query term relevance

The procedure applies to decisions on the inclusion of additional query terms, after a core query has been formulated. The objective is to establish whether query terms can be added which will return a sufficient number of relevant documents not containing the core terms, without creating undue noise. The addition of query terms should return the minimum possible number of unrelated documents. The underlying principle is that helpful additional terms are those which can be shown to be associated with the core terms in a sufficient number of contexts; that is, the term should demonstrate preference for texts containing the core query. The first step in quantifying that preference is establishing the ratio of the number of texts returned by the query *'core query AND*[9] *candidate term'* (henceforth, CQ&T) to the number of texts returned by a query containing only the candidate term. The query term relevance score (henceforth, QTR) is calculated as shown in Figure 3:

$$\boxed{\mathbf{QTR} = \frac{\mathbf{CQ\&T}}{\mathbf{T}}}$$

Figure 3: Calculation of Relevance

QTR is, in essence, a *global technique*, in that it examines word co-occurrences in a corpus in order to expand a given query (see Xu and Croft 1996: 4–5), albeit using a sample of the available documents. Also, the nature of QTR is not unlike that of scores in the *vector processing* model, in which "the similarity between two objects [i.e. documents] is computed as a function of the number of properties [i.e. index terms] that are assigned to both objects; in addition, the number of properties that is jointly absent from both the objects may also be taken into account" (Chowdhury 2004: 176). In the QTR score, these properties are the relative co-occurrence of candidate and core-query terms in documents, or, in other words, the relative frequency of the presence or absence of a candidate term in documents containing one or more of the core query terms. QTR also has aspects in common with the *best match searching* model, which is "a term weighing scheme that reflects the importance of a term" (Chowdhury 2004: 180).

It must be clarified that, as it stands, the QTR score means very little on its own, and its main utility is to help establish the baseline score (see below). As will be seen later in this section, QTR scores are sensitive to the make-up of the pilot corpus. Therefore, QTR should be interpreted in relation to three other scores: that of clearly relevant terms, for example, those which are relevant by definition (in our case, the core query terms), that of clearly irrelevant ones (see below), and the baseline. The baseline for an acceptable level of relevance is indicated by the lowest QTR derived for one of the core terms when the rest are used as the core query. In other words, the threshold marking preference is set by the baseline score. At the same time, we should also take into consideration the distance between the scores of candidate terms and clearly unrelated terms. Let us use UK1 to demonstrate how the baseline relevance is calculated. Having established the core query '*refugee* OR asylum seeker**', we will now calculate QTR for each of its constituent terms, treating the other as a candidate term (see Tables 2 and 3).

Table 2: Relevance of *asylum seeker** with *refugee** as the core query

CQ&T (*refugee* AND asylum seeker**)	T (*asylum seeker**)	QTR
39	125	*0.312*

Table 3: Relevance of *refugee** with *asylum seeker** as the core query

CQ&T (*asylum seeker* AND refugee**)	T (*refugee**)	QTR
39	349	*0.112*

Following the premise that terms are deemed good candidates if their relevance score is at least equal to that of the lowest-scoring core term, the baseline for candidate term relevance would be QTR=*0.112*.[10]

If we take into account corpus-based research which strongly indicates that different forms of a lemma may enter into different collocational patterns and demonstrate different semantic prosodies/preferences[11] (e.g. Sinclair 1991: 53–65, 154–156), then it seems appropriate to calculate the relevance score of the different forms of a candidate term separately, rather than only the **stem** (stem plus affix wildcards), as different forms may yield different scores. On the other hand, it may be desirable to treat synonymous terms as a single query item. For example, although *emigrant* shows QTR below the baseline, its relevance increases if we calculate QTR for the query '**migrant*' (i.e. *emigrant OR immigrant OR migrant*), which almost equals the baseline (QTR=*0.108*). This seems to be consistent with the findings of Baker (2004), who suggests, in relation to keywords, that corpus-based research would be wise to also examine the keyness of groups of notionally related low-frequency keywords.

Table 4 shows the result of the initial examination, which compared the keyness[12] (or lack of it) and the QTR score of candidate terms, as well three check terms, that is, clearly irrelevant terms: *dvd, guitar, lemon*.[13] Keywords were derived from the comparison of UK1 and the written BNC Sampler. Some candidate terms were selected because they were among the strongest keywords, others because they were introspectively deemed to be closely related to the core terms (some of the latter are lower-ranking keywords, others are non-key). In the LL column, bold indicates that the term is one of the top 40 keywords in the comparison of unigrams, bigrams and trigrams (as appropriate), or, in the case of wildcarded terms, that it has a score which would place it among the top 40; the symbol × indicates that a term is not key. In the QTR column, bold indicates that the relevance score is above the baseline. Candidate terms are listed in alphabetical order.

Table 4: Keyness and relevance of candidate and check terms in UK1

Candidate Terms	**LL**	***QTR***
abuse	28.1	*0.015*
abused	22.4	*0.017*
blair	**396.3**	*0.015*
deportation	90.6	***0.208***
deported	68.9	***0.216***
deportees	30.4	***0.292***
deporting	19.7	***0.182***
*deport**[14]	**147.5**	***0.137***
displace(s)	✘	*0*
displaced	✘	***0.113***
displacement	✘	*0.034*
displacing	✘	***0.154***
*displac**	✘	*0.076*
dvd	34.2	*0.008*
emigrant	✘	*0.037*
emigrated	✘	*0.081*
emigration	✘	*0.063*
*emigr**	✘	*0.071*
*ethnic minorit**	✘	*0.064*
evacuate	22.8	*0.062*
evacuated	23.6	*0.051*
evacuating	✘	*0.059*
evacuation(s)	✘	*0.049*
evacuee(s)	62.7	*0.082*
*evacu**	80.5	*0.045*
expelled	✘	*0.048*
expulsion	✘	*0.083*
extradition OR expulsion	36.8	*0.040*
extradition	18.5	*0.024*
firm but fair	✘	*0*
fugitive(s)	✘	*0.014*
genocide	**125.4**	*0.087*
guitar	✘	*0.052*

Candidate Terms	**LL**	***QTR***
human rights	**123.6**	*0.057*
hurricane	**217.2**	*0.017*
illegal alien(s)	✘	*0*
illegal entry	✘	***0.154***
illegal immigrant(s)	**60.8**	***0.150***
*immigr**	**457.7**	*0.098*
immigr OR emigr**	**442.0**	*0.097*
immigr OR emigr* OR migrant*	**532.9**	*0.095*
**migrant*	**281.2**	*0.108*
immigrant(s)	**191.6**	***0.114***
*immigrate**	✘	*0*
immigration	**265.4**	***0.132***
leave to remain	✘	***0.143***
lemon	✘	*0.006*
migrant(s)	92.6	***0.156***
policy	✘	*0.018*
settler(s)	✘	***0.160***
stranded	22.4	*0.028*
terrorism	**160.3**	*0.025*
threat	16.1	*0.013*
unemployed	✘	*0.024*
unemployment	✘	*0.023*

An initial observation is that keyness does not tend to coincide with relevance. Fewer than one-third of the top-40 key terms, and of all the key terms examined, establish relevance above the baseline (30.8% and 31% respectively), whereas almost one in five (19.2%) of non-key terms have relevance higher than B. This is quite interesting given the very low baseline score (B = 0.112). Conversely, lack of keyness does seem to coincide with lack of relevance (80.8% of the cases examined). What is more, there are cases when keywords have a QTR score very close to that of clearly irrelevant terms, a discrepancy that becomes more striking when such terms are among the top-40 keywords (*blair, hurricane, genocide, terrorism*).

At this point, we need to consider whether these discrepancies are due to the fact that the pilot corpus only spanned one month, as the relevance score need not be static, but may well be dynamic. That is, a candidate term may not have been closely related to the core terms towards the beginning of the period in question (i.e. 1996–2005), but it may have been increasingly treated as relevant in recent years (or vice versa). A further indicator, then, of the relevance of a candidate term is the consistency of its score over time. Similarly, the keyness of terms can also be expected to change over time, particularly the further the publication date of newspaper articles is removed from the period covered in the reference corpus. For this reason, a second pilot corpus (UK6) was compiled, which included articles spanning the period 1996–2005. The objective was to establish whether, and to what extent, the keyness or relevance of candidate terms would be affected by the different composition of the two pilot corpora in terms of the publication dates of the texts they contained. It has to be clarified that what is of interest in this comparison is not so much the strength of keyness, but whether the keyness of a term, or the relation of QTR to the baseline, would be consistent in the two pilot corpora. As Tables 5 and 6 show, the calculation of baseline relevance in UK6 confirms that the QTR score is indeed sensitive to the make-up of the corpus, as the baseline score (B) is different from the one calculated for UK1.

Table 5: Relevance of *asylum seeker** with *refugee** as the core query in UK6

CQ&T (*refugee* AND asylum seeker**)	T (*asylum seeker**)	*QTR*
593	1403	*0.423*

Table 6: Relevance of *refugee** with *asylum seeker** as the core query in UK6

CQ&T (*asylum seeker* AND refugee**)	T (*refugee**)	*QTR*
593	2596	*0.228*

It is clear that the QTR score does not lend itself to comparisons between two corpora, and, consequently, it does not allow for reliability checks, as the baseline score changes with the corpus make-up. However, this can be easily remedied if instead of the absolute relevance we calculate the relative relevance (RQTR) of a term. The RQTR score measures the relative distance of the QTR score of a candidate term from the baseline (Figure 4). The introduction of the baseline score in the calculation of RQTR seems compatible with the *best match* *searching* model, in that “a best match search matches a set of query words against the set of words corresponding to each item in the database, calculates a measure of similarity between the query and the item, and then sorts the retrieved items in order of decreasing similarity” (Chowdhury 2004: 180). However, best match searching requires the database documents to be indexed, which would be undesirable, even if the database to be used did include pertinent index terms (see section 2). When comparing the RQTR scores of a candidate query term in two pilot corpora, we are examining whether, and to what extent, the term scores are higher or lower than B, as established in each corpus – effectively neutralising inter-corpus fluctuations in the baseline score.

$$\boxed{\mathrm{RQTR} = \frac{(\mathrm{QTR}-\mathrm{B}) \ast 100}{\mathrm{B}}}$$

Figure 4: Calculating RQTR from QTR and B

The utility of relative relevance is twofold. It allows comparisons of relevance between corpora, and it facilitates the comparison between the relevance of different candidate terms in a given corpus. However, as it stands, RQTR is only helpful for comparisons in cases of negative RQTR scores, as the minimum possible RQTR is always the same (*-100*). The minimum possible RQTR is calculated when QTR is zero, that is, when the candidate term is never found in the same database texts with the core query terms. In the case of positive scores, the maximum possible RQTR score depends on the baseline score (it is inversely proportionate to it), and can fluctuate widely. The maximum RQTR is derived when QTR is 1, that is, when the candidate term is always found in the same database texts with the core query terms. For example, with B=0.228, the maximum RQTR score is *338.6*, whereas with B=0.112, the maximum RQTR score is *792.8*. Therefore, in order to be able to compare the distance (higher or lower) from the point where QTR=B, we need to normalise positive RQTR scores. We derive the normalised positive RQTR score (RQTRn) by calculating positive RQTR values as if the maximum possible RQTR were 100 (Figure 5):

$$\boxed{\mathrm{RQTRn} = \frac{\mathrm{RQTR} \ast 100}{\mathrm{max.\ RQTR}}}$$

Figure 5: Calculating RQTRn

If we substitute $\frac{(QTR - B) * 100}{B}$ for RQTR, and $\frac{(1-B) * 100}{B}$ for max.RQTR (as the maximum possible QTR value is 1) in the above formula, then we have a formula for calculating RQTRn which makes use of the QTR and B scores, so there is no need to calculate RQTR and max.RQTR (Figure 6:).

$$\boxed{RQTRn = \frac{(QTR - B) * 100}{1-B}}$$

Figure 6: Calculating RQTRn without RQTR

For ease of reference, RQTR will denote both negative and normalised positive scores, that is, when the RQTR score is positive it should be understood to have been normalised. The RQTR score indicates relevance on a bi-directional scale, by treating the baseline as the zero point: when QTR=B, then RQTR=0. The scores show whether the relevance of a candidate term is higher (positive) or lower (negative) than the baseline relevance, and also indicate the extent of the distance from the baseline. See Table 7 for details:

Table 7: Interpreting RQTR scores.

RQTR	**Interpretation**
+100	Full relevance: the candidate term is always found in database texts containing one or more of the core query terms.
0	Baseline relevance: the candidate term has the same level of relevance as that set as the minimum for inclusion to the final query.[15]
–100	No relevance: the candidate term is never found in database texts containing any of the core query terms.

The RQTR score is useful in two ways. First, it makes explicit the distance from the baseline score (either positive or negative), and thus facilitates the comparison of term relevance scores within the same corpus. Second, it enables the comparison of relevance between corpora derived from different time periods.[16] Interestingly, candidate terms with RQTR of +100 need not be added to the query as they will be returned by the core query alone.

Now we are able to carry out a more systematic comparison of keyness and query term relevance, particularly as the previous comparison only used a some-how arbitrary list of items. The top-40 keywords in the comparisons between the two pilot corpora (UK1 and UK6) and the two reference corpora (BNC-S and BNC-N) were combined, producing a total of 74 distinct keywords. In order to create the most helpful list possible, certain words are not included: the core terms (*refugee**, *asylum*, *seeker**), function words, and frequent verbs (e.g. *say*). Table 8 compares the keyness (LL) and relative relevance (RQTR) of these terms. For ease of comparison, and since the highest *p* value in all four comparisons is as small as 10^{-14} (see Table 8 for details), in Table 9, top-40 keyness is indicated by the symbol '✓' and non-keyness by '✗'; additionally, top-40 keyness and positive RQTR are indicated by shading.

Table 8: Lowest top-40 LL scores and highest *p* values in the keyword comparisons

	Lowest top-40 LL	**Highest top-40 *p***
UK1*BNC-S	107.8	$p<10^{-14}$
UK1*BNC-N	171.1	$p<10^{-15}$
UK6*BNC-S	323.6	$p<10^{-16}$
UK6*BNC-N	1203.4	$p<10^{-18}$

Table 9: Comparison of LL and RQTR scores of top-40 keywords in UK1 and UK2

CQ: refugee OR asylum seeker**	***Top-40 Keyness***		***RQTR***	
Top-40 keywords (*n* = 74)	**UK1**	**UK6**	**UK1**	**UK6**
afghan	✗	✓	*-33.0*	***+3.5***
afghanistan	✗	✓	*-59.8*	*-38.2*
al	✓	✓	*-72.3*	*-81.6*
arafat	✗	✓	*-17.0*	***+8.7***
ariel	✗	✓	***+1.9***	***+5.3***
army	✗	✓	*-72.3*	*-69.7*
attacks	✓	✓	*-90.2*	*-87.3*

CQ: refugee* OR asylum seeker*	**Top-40 Keyness**		**RQTR**	
Top-40 keywords (*n* = 74)	**UK1**	**UK6**	**UK1**	**UK6**
bethlehem	✗	✓	*-100*	***+12.7***
blair('s)	✓	✓	*-86.6*	*-85.5*
blunkett	✗	✓	*-91.1*	*-47.4*
bondi	✓	✗	*-55.4*	*-84.6*
britain	✓	✓	*-83.9*	*-82.0*
camp(s)	✓	✓	*-40.2*	*-43.9*
civilians	✗	✓	*-41.1*	*-29.4*
congo	✓	✗	*-52.7*	*-19.3*
country	✗	✓	*-85.7*	*-85.1*
darfur	✓	✗	***+12.2***	***+15.4***
eu	✓	✓	*-88.4*	*-81.6*
family	✓	✗	*-85.7*	*-89.5*
gaza	✓	✓	***+4.7***	***0***
gbp	✗	✓	*-88.4*	*-94.3*
genocide	✓	✗	*-22.3*	*-39.9*
hamas	✓	✗	***+15.5***	***+11.4***
home	✓	✓	*-93.8*	*-88.2*
human	✓	✓	*-72.3*	*-76.8*
hurricane	✓	✗	*-84.8*	*-88.1*
immigrant(s)	✓	✓	***+0.2***	***+3.2***
immigration	✓	✓	***+2.2***	***+11.5***
iraq	✓	✓	*-66.1*	*-83.8*
israel('s)	✓	✓	*-48.2*	*-35.1*
israeli(s)	✓	✓	*-28.6*	*-18.0*

CQ: refugee OR asylum seeker**	***Top-40 Keyness***		***RQTR***	
Top-40 keywords (*n* = 74)	**UK1**	**UK6**	**UK1**	**UK6**
jack	✓	✘	*-81.3*	*-84.6*
jenin	✘	✓	*-100*	***+63.2***
jerusalem	✘	✓	*-44.6*	*-15.4*
jewish	✓	✓	*-7.1*	*-45.6*
jew(s)	✓	✘	*-34.8*	*-44.3*
katrina	✓	✘	*-84.8*	*-79.8*
killed	✓	✓	*-70.5*	*-79.8*
kosovo	✘	✓	*-6.3*	*-22.8*
labour	✓	✘	*-75.8*	*-86.8*
libeskind	✓	✘	***+4.7***	*-100*
louisiana	✓	✘	*-55.4*	*-80.7*
migrant(s)	✓	✘	***+4.9***	***+7.9***
nazi	✓	✘	*-20.5*	*-62.7*
office	✘	✓	*-82.1*	*-84.6*
orleans	✓	✘	*-63.4*	*-86.0*
palestinian(s)	✓	✓	*-6.3*	*-14.5*
peace	✘	✓	*-72.3*	*-75.0*
police	✓	✓	*-83.9*	*-88.6*
pounds	✓	✓	*-94.6*	*-95.2*
powell	✘	✓	*-92.0*	*-70.2*
ramallah	✘	✓	*-57.1*	***+21.6***
ransome	✓	✘	*-61.6*	*-100*
rwanda	✓	✘	*-12.5*	***+5.3***
rwandan	✓	✘	***+1.9***	***+19.4***

CQ: refugee OR asylum seeker**	***Top-40 Keyness***		***RQTR***	
Top-40 keywords (*n* = 74)	**UK1**	**UK6**	**UK1**	**UK6**
saddam	✓	✗	*-83.0*	*-85.5*
samir	✓	✗	***+9.9***	*-34.2*
secretary	✗	✓	*-80.4*	*-82.5*
seth	✓	✗	*-67.0*	*-97.8*
sharon	✗	✓	*-81.3*	*-45.2*
soldiers	✗	✓	*-68.8*	*-63.2*
suicide	✗	✓	*-20.5*	*-68.4*
sudan	✓	✗	*-12.5*	*-33.8*
taliban	✗	✓	*65.2*	*-30.3*
terror	✗	✓	*-75.9*	*-73.7*
terrorism	✓	✓	*-77.7*	*-70.6*
tony	✓	✗	*-88.4*	*-90.8*
un	✗	✓	*-76.8*	*-47.8*
walter	✓	✗	*-64.3*	*96.5*
war	✓	✓	*-73.2*	*-77.6*
wiesenthal	✓	✗	*-33.9*	*-47.4*
wolfgang	✓	✗	*-21.4*	*-87.7*
zarqawi	✓	✗	*-73.2*	*-94.7*
zimbabwe	✓	✗	*-53.6*	*-86.8*

The RQTR score shows a significantly higher consistency between different corpora than top-40 keyness. When scores in the two pilot corpora are compared, RQTR polarity coincides for the vast majority of terms (89.2%), whereas top-40 keyness only coincides in just over a quarter of cases (28.4%). Also, in both pilot corpora, top-40 keyness and positive RQTR coincide in only

12.2 per cent of the cases; that is, most of the top-40 keywords would be expected to return documents largely unrelated to the core query terms. At this point, we need to revisit the introspectively selected terms examined above (Table 3) and examine the LL and RQTR scores in both UK1 and UK6 (see Table 10; bold indicates positive RQTR scores and keyness at top-40 level for each comparison; LL scores that are lower than 15.3 are indicated by ✗; parentheses before ✗ indicate keyness if the threshold is lowered to LL≥6.63, $p \leq 10^{-2}$ – for an explanation of why a lower keyness threshold was also examined, see page 29).

Table 10: Comparison of LL and RQTR scores of introspectively selected terms in UK1 and UK2

Candidate Terms	UK1* BNC-S	UK1* BNC-N	UK1 ***RQTR***	UK6* BNC-S	UK6* BNC-N	UK6 RQTR
abuse(s)	28.0	(6.63) ✗	*-86.6*	103.7	103.4	*-87.3*
abused	22.3	✗	*-84.8*	32.0	(8.3) ✗	*-89.0*
deportation(s)	90.3	**203.7**	***+10.8***	126.2	637.9	*+7.5*
deported	68.7	**129.5**	***+11.7***	132.6	552.4	*+1.7*
deportee(s)	30.3	32.4	***+20.3***	(12.3) ✗	(11.4) ✗	*+1.0*
deporting	19.6	39.3	***+7.9***	21.0	89.2	*+10.2*
*deport**	**147.5**	**418.8**	***+2.8***	**381.3**	**1594.5**	***+5.4***
displace(s)	✗	✗	*-100*	✗	✗	*-87.7*
displaced	(14.0) ✗	31.4	***+0.1***	59.6	279.9	*-9.2*
displacement	✗	✗	*-69.6*	✗	(9.8) ✗	*-70.6*
displacing	✗	✗	***+4.7***	✗	✗	*-83.3*
*displac**	(9.1) ✗	28.3	*-32.1*	44.0	256.6	*-44.7*
dvd(s)(‘s)	48.5	114.8	*-92.8*	24.6	110.5	*-96.5*
emigrant(s)	✗	✗	*-66.9*	✗	✗	*-56.1*
emigrated	(11.9) ✗	(8.1) ✗	*-27.7*	✗	✗	*-83.8*
emigration	✗	✗	*-43.8*	✗	✗	*-50.9*
*emigr**	(7.3) ✗	(7.7) ✗	*-36.6*	16.5	39.8	*-69.7*
*ethnic minorit**	(12.8) ✗	27.6	*-42.8*	20.3	102.0	*-61.4*

Candidate Terms	**UK1* BNC-S**	**UK1* BNC-N**	**UK1 *RCTR***	**UK6* BNC-S**	**UK6* BNC-N**	**UK16 *RTQR***
evacuate(s)	22.8	17.8	*-44.6*	34.8	67.4	*-4.5*
evacuated	23.6	25.6	*-54.4*	(11.6) ×	17.5	*-66.7*
evacuating	×	×	*-47.3*	×	×	*-46.9*
evacuation(s)	(12.3) ×	24.2	*-56.2*	(12.8) ×	61.1	*-56.6*
evacuee(s)	62.7	149.9	*-26.8*	×	(8.4) ×	*-59.6*
*evacu**	80.5	158.1	*-59.8*	53.4	138.7	*-55.2*
expelled	(11.8) ×	(9.4) ×	*-57.1*	27.2	55.3	*-75.0*
expulsion(s)	18.6	25.3	*-25.9*	17.6	89.7	*-56.6*
extradition(s)	18.5	(11.3) ×	*-78.6*	75.9	170.3	*-69.7*
extradition(s) OR expulsion(s)	36.8	34.5	*-64.3*	103.8	259.8	*-66.7*
firm but fair	×	×	*-100*	×	(7.8) ×	*-39.5*
fugitive(s)	×	×	*-87.5*	(7.3) ×	23.6	*-75.0*
guitar(s)(‘s)	×	×	*-53.6*	×	×	*-96.9*
hijack(s)	×	×	*-80.2*	187.9	721.3	***+6.6***
hijacker(s)	×	×	*-74.5*	135.9	1149.2	***+10.2***
*hijack**	×	×	*-91.9*	**487.4**	**2240.4**	*-27.19*
human rights	**123.6**	**282.5**	*-50.0*	**358.9**	**1809.6**	*-48.2*
illegal alien(s)	×	×	*-100*	×	(7.8) ×	***+44.4***
illegal entry(ies)	×	×	***+4.7***	×	(13.5) ×	***+13.6***
illegal immigrant(s)	60.8	132.8	***+4.3***	17.1	911.3	***+16.2***
*immigrate**	×	×	*-100*	×	×	*-100*
*immigr**	**457.7**	**862.2**	*-12.5*	**1528.2**	**6481.6**	***+2.5***
immigr OR emigr**	**442.0**	**772.2**	*-13.4*	**1502.4**	**5964.8**	*-4.8*
immigr OR emigr* OR migrant**	**532.9**	**1000.7**	*-15.2*	**1569.1**	**6640.3**	*-7.0*
**migrant(s)*	**281.5**	**579.7**	*-3.6*	**648.2**	**2901.6**	***+1.0***
leave to remain	×	(6.8) ×	*+3.5*	37.6	162.6	***+30.6***

Candidate Terms	**UK1* BNC-S**	**UK1* BNC-N**	**UK1 *RQTR***	**UK6* BNC-S**	**UK6* BNC-N**	**UK6 *RQTR***
lemon(s)(‘s)	×	×	*-94.6*	×	×	*-96.0*
massacre(s)	27.4	56.5	*-48.9*	170.6	754.9	*-19.3*
persecution(s)	24.2	62.1	***+8.2***	97.7	549.1	*+11.4*
*persecut**	29.6	60.6	***+4.0***	149.5	665.8	*+4.1*
policy(ies)	×	26.1	*-83.9*	(11.6) ×	156.0	*-85.5*
racism	39.3	32.2	*-55.6*	186.3	561.9	*-57.5*
*racis**	95.4	96.9	*-58.3*	**470.6**	**1546.8**	*-57.0*
settler(s)	(14.7) ×	69.9	***+5.4***	96.2	646.3	*-44.7*
stranded	22.4	(7.3) ×	*-75.0*	26.8	(10.7) ×	*-87.7*
threat(s)	18.1	×	*-88.4*	99.9	74.4	*-89.5*
unemployed	×	×	*-78.6*	×	×	*-87.3*
unemployment	×	×	*-79.5*	×	×	*-88.6*

Again, keyness does not tend to correlate with relevance. Terms being key in both comparisons have positive RQTR in just above one-third of the cases: 34.8 per cent in UK1 and 35.3 per cent in UK6. The correlation is much lower when we consider terms only being key in one comparison, among which keyness coincides with positive RQTR in 18.2 per cent of the cases in UK1 and never in UK6. Overall, keyness in at least one comparison corresponds to relevance in just over a quarter of instances: 29.4 per cent (UK1) and 28.2 per cent (UK6). The discrepancy is rendered more striking if we consider that the baseline scores are rather low (0.112 for UK1 and 0.228 for UK6). That is, terms need to be present in only 11.2 or 22.8 per cent of the documents containing one or more core query terms in order to register relevance. Consequently, it might be reasonably expected that high-ranking keywords would also register positive RQTR, or, in other words, that keyness would be better able to discriminate between relevant and non-relevant terms, but this is not the case here (explanations follow later in this section). Conversely, lack of keyness correlates highly with lack of relevance. From the terms being non-key in both comparisons, 85.7 per cent (UK1) and 94.1 per cent (UK6) have negative RQTR. However, this also suggests that if keyness were the sole criterion for further examining the suitability of query terms, then a not insignificant proportion of relevant terms would be left out (14.6% in UK1, 5.9% in UK6). This observation points towards a further utility of RQTR, namely the ability to test the relevance of introspectively selected non-key terms. Regarding consistency between the two sample corpora, both keyness and RQTR show similar results, with RQTR being relatively more consistent. RQTR polarity is the same in the two sample corpora for 85.7 per cent of terms.[17] Keyness (or its lack) corresponds in both pilot corpora in 75 per cent and 78.6 per cent of the cases, in the keyword comparisons with BNC-S and BNC-N.

However, we also need to examine whether the low correspondence between keyness and relevance is due to the threshold set for keyness ($\mathrm{LL} \geq 15.13$, $p \leq 10^{-4}$), as opposed to the more frequently used lower threshold of $\mathrm{LL} \geq 6.63$, $p \leq 10^{-2}$ (McEnery 2006: 233, nn. 32). To this end, the same comparisons discussed above were carried out with the lower keyness value, in order to establish whether the lower threshold would increase the correspondence of keyness with relevance. As Tables 11 to 15 show, this does not seem to be the case. For candidate terms being key in both, or at least one, comparison (Tables 11 and 13 respectively), the correspondence seems to mostly decline with the lower keyness threshold. It is significantly higher for terms being key in only one comparison (Table 12); however, the increase from 18.2 per cent to 50 per cent in UK1 only reflects the correspondence in two terms. Also, in no instance does the correlation go above half of the cases, and is overall no more than one-third (Table 13). Predictably, the correspondence between lack of keyness and lack of relevance increases with a lower threshold for keyness (Table 14). Consistency of keyness between different reference corpora shows some increase in only one corpus, and is never higher than the consistency of RQTR (Table 15).

Table 11: Correlations: keyness in both comparisons and positive RQTR

	UK1	**UK6**
LL ≥ 15.13	34.8%	35.3%
LL ≥ 6.63	30.3%	32.5%

Table 12: Correlations: keyness in one comparison and positive RQTR

	UK1	**UK6**
LL ≥ 15.13	18.2%	0%
LL ≥ 6.63	50%	40%

Table 13: Correlations: keyness in at least one comparison and positive RQTR

	UK1	**UK6**
LL ≥ 15.13	29.4%	28.2%
LL ≥ 6.63	28.9%	33.3%

Table 14: Correlations: lack of keyness and negative RQTR

	UK1	**UK6**
LL ≥ 15.13	85.7%	94.1%
LL ≥ 6.63	88.9%	100%

Table 15: Consistency of keyness and relevance respectively

	Keyness		**Relevance**
	BNC-S	**BNC-N**	
LL ≥ 15.13	75.0%	78.6%	85.7%
LL ≥ 6.63	85.7%	78.6%	

One reason for the discrepancy between the keyness and relevance of the terms examined is, arguably, that texts in the reference corpora predate those in the sample corpora, and, as a result, some words (e.g. names of politicians) will establish keyness irrespective of their relevance to the core query terms. Another possible reason is that some terms, although clearly associated with the core query (e.g. Palestinians), are not central to the perception/presentation of the groups in focus. More generally, discrepancies may also be the result of the nature of the two measures. Keyword analysis regards the pilot and reference corpora as single documents, whereas RQTR looks at co-occurrence within individual texts in the pilot corpus. In that respect, calculating RQTR is not unlike calculating key-keywords (see section 2). However, RQTR bypasses the problem of time- or genre-specific keywords, as it does not depend on a reference corpus. For most of the terms with negative RQTR, it may be argued that their unsuitability for the target corpus was self-evident, and that there was little justification in investing time in examining the merits of their inclusion in the query. However, if these terms are indeed patently irrelevant, then, given the fact that some of them were (strong) keywords, this should be regarded as clear testimony to RQTR being more successful than keyness as an objective indicator of the suitability of candidate terms. In the light of the above, the RQTR score seems to be more reliable than keyness for purposes of query expansion. Nevertheless, the combination of the two measures seems advisable for two reasons. Neither measure is in itself entirely consistent when applied to different corpora. More importantly, as the process of calculating RQTR is itself an investment in time, and given the high correlation of non-keyness to non-relevance, keyword analyses can be employed to limit the number of candidate query terms, without excluding the possibility of also considering non-key candidate terms.

Ideally, then, a successful candidate term would be a high-ranking keyword with a high positive RQTR, while also being introspectively plausible (i.e. consistent with our knowledge and experience). However, if an entirely suitable reference corpus is not available, then in cases of discrepancy (i.e. lack of keyness but positive RQTR, and vice versa) the relevance score carries more weight. In the same vein, we may also add introspectively plausible terms with a positive RQTR, irrespective of their keyness, such as, *highjack(s), hijacker(s), illegal alien(s), illegal entry, leave to remain*. Also, it seems reasonable to add all forms of a lemma or word family[18] if a good proportion of the forms are key words with a positive RQTR score: *deport** (instead of only *deportation, deported, deporting*), *immigr** (instead of only *immigrant(s), immigration*), as well as *emigr**, because of its semantic similarity to *immigr**. Finally, part of a compound or, more generally, a meaningful n-gram may substitute for the whole compound/n-gram if it has a positive RQTR score and is found in a large number of meaningful n-grams, preferably if a good number of them are key. For example, *asylum* can substitute for *asylum seeker(s)* (see Appendices 1 and 2). Table 16 summarises the main steps involved in formulating the final query:

Table 16: Summary of main steps taken for query formulation

- Selection of a minimum of two core query terms based on a clear definition of the content of the corpus to be compiled.
- Creation of a (sample) corpus using the core query terms linked by the Boolean operator 'OR'.[19]
- Calculation of the baseline score using QTR.
- Keyword analysis using an appropriate reference corpus (if available).
- Selection of candidate query terms among the (high-ranking) keywords, as well as through introspection. Selection of clearly irrelevant terms.
- Calculation of RQTR for the candidate and irrelevant terms.
- Examination of RQTR scores for final decision.

It is recognised that a query built using the techniques discussed here may exclude some relevant texts, while including some irrelevant ones. However, it must also be stressed that the compilation of a corpus containing all and only the relevant texts in the database would require reading the retrieved documents, which, given the scope of the intended corpus, would be unrealistic. It is also recognised that the procedure is not entirely objective. What can be argued is that any subjective decisions are guided, if not constrained, by objective indicators. Also, the involvement of subjectivity is a characteristic shared with a large number of other techniques. For instance, selecting initial seeds, deciding on the number of top keywords to include in subsequent queries, defining and assigning document index terms, weighing the relevance of a document to a query by means of user/expert reading, or setting the *p* value that marks statistical significance are all largely subjective decisions. In the light of this, it must be clarified that neither the baseline nor the RQTR score are necessarily binding. Corpus compilers may choose to set a higher or lower baseline score to suit their purposes; for example, they may select as the baseline the highest rather than the lowest QTR score among core query terms. Similarly, they may decide to exclude candidate terms with (low) positive RQTR scores, or include terms with (low) negative scores, depending on their circumstances and aims.

4 Conclusion

The introduction of RQTR was intended as a means of triangulating decisions on query expansion by supplementing keyness as an objective indicator of candidate query term relevance, as well as providing a way of evaluating the relevance of introspectively selected candidate terms. It must be reiterated that RQTR requires that at least two clearly relevant terms can be selected, so that a baseline can be established. However, all query-expansion procedures have similar requirements. Furthermore, it seems unlikely that a good definition of the content of a specialised corpus will not suggest the requisite minimum of two clearly relevant terms. Therefore, it can be argued that the procedure discussed here is both principled, in that objective indicators are used, and conscious, in that the process is not fully automatic. To be more precise, the term selection conforms to one or both of the specified objective requirements, while at the same time having introspective plausibility.

An important consideration when constructing/expanding a query is that any additional terms should not add undue noise. Ideally, then, suitable candidate terms would be strong key words with positive RQTR. However, it seems reasonable to add to the query other forms of the lemma or word family that a relevant term belongs to if a good proportion of these forms, or their combinations, have positive RQTR. Also, in the same way that keyness is not an absolute criterion, but depends on the maximum *p* value that is considered acceptable for statistical significance, the baseline score can be adjusted according to the corpus compilers' needs.

RQTR will also be a suitable technique on its own in other instances, particularly when an appropriate reference corpus (for the calculation of keywords) is not available. The RQTR score may be independent of reference corpora, but depends, to some extent, on the sample corpus; however, it is more consistent than keyness in that respect. Also, it disposes of the need to know the total number of documents in the database, and the need to manually examine retrieved documents. While RQTR bypasses the restrictions usually posed by database interfaces, the use of the technique is not limited to restricted-access text databases; on the contrary, the reliability of the RQTR score should increase as the access restrictions decrease. Finally, the procedures and calculations involved are expected to be accessible to all linguists or language educators who might want to build a specialised corpus drawing texts from a database.

Notes

* I am grateful to Paul Baker (Lancaster University) for his comments on a previous draft. I would also like to thank Sebastian Hoffmann (Lancaster University) for the newspaper sub-corpus of the BNC, and the participants of the meeting of the Lancaster University Corpus Research Group on 30 October 2006 for their comments on an earlier version of this paper.

1. Funded by the ESRC (RES-000-22-1381); principal investigator: Paul Baker.
2. Keywords are those words which are statistically significantly more frequent in the corpus under analysis when compared to another corpus (Rayson and Garside 2000; Scott 2001).
3. http://www.refugeecouncil.org.uk/practice/basics/truth.htm
4. In fact, the term *terrorism* is one of the highest ranking keywords with LL scores of 160.3 (p<10^{-15}) and 339.8 (p<10^{-16}) in two pilot sub-corpora (see section 2 for details).
5. Examples of metaphorical uses of *refugee(s)*:

   “After another half-hour, they packed us into a smaller train that crawled back almost to Falkirk before halting for another hour, because the station ahead was overrun by refugees from a London express unable to reach Edinburgh.” (*The Daily Mail*, 24 July 1998).

   “But Spurs – hopeless, hapless and complacent beyond belief – defended with all the savvy of refugees from a greasy spoon café” (*The Mirror*, 2 December 1999).

   “Ironically, extending the minimum wage to 16- and 17-year-olds may well keep them out of the workforce. Many employers will decide that illiterate refugees from our comprehensive schools give very poor value in comparison with Kurds, Poles and Africans” (*The Daily Telegraph*, 3 January 2004).
6. The keywords analysis was carried out using WordSmith Tools 4 (Scott 2004). Significance was calculated using the Log Likelihood statistic, with the minimum statistical significance set at $p \leq 10^{-4}$, LL≥15.13 (see Rayson et al. 2004).
7. For example, both Google and Yahoo APIs allow a maximum of 1,000 and 5,000 queries per day respectively, and return up to 1,000 pages per query. It is also possible that the current largely open access to databases of web pages compiled by search engines may be restricted in the future. A case in point is the recent announcement by Google that, as of 5 December 2006,

they have stopped issuing new accounts for their SOAP search API. For details, see:
http://code.google.com/apis/soapsearch/index.html (Google), and
http://www.informit.com/articles/article.asp?p=382421&seqNum=2&rl=1
(Yahoo).

8. Forty out of the top hundred keywords are proper nouns or adjectives denoting ethnicity.
9. This is the Boolean ‘AND’.
10. The comparison of QTR scores seems to also have the potential to contribute to the semantic analysis of lexis, particularly when their meanings overlap. For example, the difference in the relevance scores of *refugee(s)* and *asylum seekers(s)* may be interpreted as indicating that people labelled as asylum seekers tend to also be presented as, or conflated with, refugees (and, arguably, perceived as such) more often than people labelled as ‘refugees’ tend to be presented as, or conflated with, asylum seekers. More tentatively, it could be argued that the notion of ‘refugee’ is a semantic component of ‘asylum seeker’. This interpretation is supported by either of the two sets of definitions mentioned previously, as well as by the results of the collocational analysis of the two terms in the corpus (Gabrielatos and Baker 2006).
11. For a discussion of *semantic prosody/preference*, see also Louw (1993) and Stubbs (2002: 65–66).
12. The LL scores for word forms have been derived from WordSmith Tools, those of lemmas or groups of word forms (e.g. *immigrant** and *emigrant**) have been calculated manually using Paul Rayson’s online Log Likelihood Calculator (http://ucrel.lancs.ac.uk/llwizard.html).
13. Since the interest is in terms rather than word forms, and since the database interface also returns the genitive forms of nouns whichever the form used in the query (and this cannot be remedied through the use of Boolean operators), RQTR and LL scores of nouns reflect the relevance/keyness of singular, plural and genitive forms taken together. This does not influence the results, as all the combined forms registered appropriate keyness. Also, this allowed for the inclusion of more key terms.
14. The query excluded the words *Deportivo* (a football team) and *deportment*.
15. In this case, the baseline is the lowest-scoring core query term.
16. Provided, of course, that calculations refer to the same core query and database.
17. The majority (78.4%) of terms also show comparable RQTR scores.

18. “A word family consists of a base word and all its derived and inflected forms. … [T]he meaning of the base in the derived word must be closely related to the meaning of the base when it stands alone or occurs in other derived forms, for example, *hard* and *hardly* would not be members of the same word family” (Bauer and Nation 1993: 253).
19. The core query, and the queries used in establishing the baseline score, can be more complex than those appropriate for this paper. That is, what was treated as a term in the core query can itself be a Boolean query, as a wildcard is a shorthand for Boolean disjunctions. For instance, ‘*refugee**’ is a shorthand for the query ‘*refugee* OR *refugees* OR *refugee’s* OR *refugees*’. In this light, the core query ‘*refugee** OR *asylum seeker**’ can be more analytically written as follows: (‘*refugee* OR *refugees* OR *refugee’s* OR *refugees*’) OR (*asylum seeker* OR *asylum seekers* OR *asylum seeker’s* OR *asylum seekers*’). Furthermore, the brackets can contain not only disjunction (OR), but also conjunction (AND) or negation (NOT). For example, let us assume that the focus of examination was the representation of women refugees and asylum seekers. A possible core query (using wildcards for brevity) might be the following: (*refugee** AND *wom*n*) OR (*asylum seeker** AND *wom*n*), which can be more simply formulated as: ‘*refugee** OR *asylum seeker** AND *wom*n*’.

References

Aston, Guy and Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. London: Addison Wesley.

Baker, Paul. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32(4): 346–359.

Baker, Paul and Tony McEnery. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics* 4(2): 197–226.

Baroni, Marco and Silvia Bernardini. 2003. The BootCaT toolkit: Simple utilities for bootstrapping corpora and terms from the web, version 0.1.2. Available online: http://sslmit.unibo.it/~baroni/Readme.BootCaT-0.1.2.

Baroni, Marco and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *LREC 2004 Proceedings*, 1313–1316.

Baroni, Marco and Serge Sharoff. 2005. Creating specialized and general corpora using automated search engine queries. Paper presented at *Corpus Linguistics 2005*, Birmingham University, 14–17 July 2005. Available online: http://sslmit.unibo.it/~baroni/wac/serge_marco_wac_talk.slides.pdf.

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek and Pavel Rychlý. 2006. WebBootCaT: Instant domain-specific corpora to support human translators. *Proceedings of EAMT 2006*, 247–252. Available online: http://corpora.fi.muni.cz/bootcat/publications/webbootcat_eamt2006.pdf.

Bauer, Laurie and Paul Nation. 1993. Word families. *International Journal of Lexicography* 6(4): 253–279.

Belew, Richard K. and John Hatton. 1996. RAVE reviews: Acquiring relevance assessments from multiple users. In M. Hearst and H. Hirsh (eds.). *Working notes of the AAAI Spring Symposium on Machine Learning in Information Access*. Menlo Park, CA: AAAI Press.

Boughanem, Mohand, Yannick Loiseau and Henri Prade. 2006. Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In M. Detyniecki, J. M. Jose, A. Nürnberger and C. J. van Rijsbergen (eds.). *Adaptive multimedia retrieval: User, context, and feedback*. Third International Workshop, AMR 2005, Glasgow, UK, July 28–29, 2005: Revised selected papers, 44–54. Berlin: Springer. Also online:
http://www.irit.fr/recherches/RPDMP/persos/Prade/Papers/BougLoiP_AMR.pdf.

Buckley, Chris, Gerard Salton and James Allan. 1994. The effect of adding relevance information in a relevance feedback environment. In W.B. Croft and C.J. van Rijsbergen (eds.). *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland, 3–6 July 1994), 292–300. New York: Springer-Verlag.

Chowdhury, G.G. 2004 (2nd ed.) *Introduction to modern information retrieval*. London: Facet Publishing.

Gabrielatos, Costas and Paul Baker. 2006. Representation of refugees and asylum seekers in UK newspapers: Towards a corpus-based analysis. *Joint Annual Meeting of the British Association for Applied Linguistics and the Irish Association for Applied Linguistics* (BAAL/IRAAL 2006): *From Applied Linguistics to Linguistics Applied: Issues, Practices, Trends*, University College, Cork, Ireland, 7–9 September 2006. Available online: http://eprints.lancs.ac.uk/265.

Ghani, Rayid, Rosie Jones and Dunja Mladeni. 2001. Mining the web to create minority language corpora. *CIKM 2001*, 279–286.

Greenslade, Roy. 2005. Seeking scapegoats: The coverage of asylum in the UK press. Asylum and immigration working paper 5. London: Institute for Public Policy Research. Also available online: http://www.ippr.org/members/download.asp?f=%2Fecomm%2Ffiles%2Fwp5%5Fscapegoats%2Epdf.

Longman dictionary of contemporary English on CD-ROM. 2003. London: Longman.

Louw, William. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.). *Text and technology: In honour of John Sinclair,* 157–176. Philadelphia and Amsterdam: John Benjamins.

McEnery, Tony. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present*. London: Routledge.

Rayson, Paul and Roger Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of Workshop on Comparing Corpora* (at ACL 2000), 1–6.

Rayson, Paul, Damon Berridge and Brian Francis. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Fairon and A. Dister (eds.). *Le Poids des Mots. Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*, Vol. 2, Louvain-la-Neuve, Belgium (March 10–12, 2004), 926–936. Louvain: Presses Universitaires de Louvain.

Robertson, Steven and Karen Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27(3): 129–146.

Scott, Mike. 2001. Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry and R.L. Roseberry (eds.). *Small corpus studies and ELT: Theory and practice,* 47–67. Amsterdam: Benjamins.

Scott, Mike. 2004. *Oxford WordSmith Tools version 4*. Oxford: Oxford University Press. Available online: http://www.lexically.net/downloads/version4/wordsmith.pdf.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John. 2004. Appendix: How to build a corpus. In M. Wynne (ed.). *Developing linguistic corpora: A guide to good practice*, 79–83. Oxford: Oxbow Books. Also online: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/appendix.htm.

Stubbs, Michael. 2002. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

van Leeuven, Theo. 1996. The representation of social actors. In C-R. Caldas-Coulthard and M. Coulthard (eds.). *Texts and practices*. *Readings in Critical Discourse Analysis,* 32–70. London: Routledge.

Wilson, David. 2006. Asylum and the media in Scotland. A report on the portrayal of asylum in the Scottish media undertaken by the Oxfam Asylum Positive Images Network and Glasgow Caledonian University. Available online:
http://oxfamgb.org/ukpp/resources/downloads/asylum_media_scotland.pdf.

Xu, Jinxi and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (SIGIR ‘96), Zurich, Switzerland (August 18–22, 1996). ACM Press, New York, NY, 4–11. DOI: http://doi.acm.org/10.1145/243199.243202.

Appendix 1: Meaningful asylum+noun bigrams in UK6 (in alphabetical order)

asylum+noun bigrams	Freq.
asylum abuses	4
asylum act	16
asylum advice	2
asylum appeals	9
asylum applicant	2
asylum applicants	18
asylum application	40
asylum applications	139
asylum backlog	6
asylum based	2
asylum bid	4
asylum bids	6
asylum bill	62
asylum camp	2
asylum campaign	2
asylum case	12
asylum cases	27
asylum centre	6
asylum centres	7
asylum chaos	2
asylum charities	2
asylum cheats	2
asylum children	2
asylum claim	35
asylum claimant	4
asylum claimants	3

asylum claims	79
asylum clampdown	2
asylum concerns	2
asylum control	2
asylum crisis	8
asylum debacle	2
asylum debate	3
asylum decisions	10
asylum detention	8
asylum door	2
asylum figures	7
asylum fraud	5
asylum fury	3
asylum hearing	2
asylum hearings	5
asylum incidents	3
asylum interview	2
asylum issue	7
asylum issues	3
asylum law	11
asylum laws	26
asylum lawyers	3
asylum league	2
asylum legislation	11
asylum myths	2
asylum option	3
asylum overhaul	2
asylum payouts	3
asylum plans	2

asylum pleas	3
asylum plot	2
asylum policies	11
asylum policy	47
asylum practices	2
asylum problem	5
asylum procedures	4
asylum process	8
asylum queue	2
asylum regime	4
asylum removal	6
asylum requests	13
asylum rights	3
asylum riot	3
asylum row	4
asylum rules	10
asylum scam	8
asylum seeking	4
asylum service	3
asylum shopping	4
asylum spongers	2
asylum statistics	2
asylum status	7
asylum support	16
asylum system	84
asylum system's	2
asylum tradition	3
asylum voucher	2
asylum-shopping	3

Appendix 2: Intransitive verb+asylum bigrams (in alphabetical order)

verb+*asylum* bigrams	Freq.
awaiting asylum	2
claimed asylum	58
claiming asylum	48
denied asylum	10
deny asylum	2
gain asylum	6
gaining asylum	2
give asylum	9
given asylum	8
gives asylum	3
grant asylum	9
granted asylum	45
granting asylum	2
grants asylum	3
have asylum	3
having asylum	2
obtained asylum	2
offer asylum	2
refuse asylum	4
refused asylum	23
requested asylum	9
seek asylum	38
seeking asylum	102
seeks asylum	2
sought asylum	23
want asylum	10
wins asylum	3