

Developing a Corpus-Based Immigration Keyword and Key Cluster List for ESP Settings

Tharin Pitakpong¹

Supakorn Phoocharoensil²

Piyapong Laosrirattanachai³

Valaya Alongkorn Rajabhat University under the Royal Patronage, Thailand¹

Thammasat University, Thailand²

Kasetsart University, Thailand³

*Corresponding author's email: piyapong.l@ku.th

Received July 29, 2026; revised August 17, 2026; accepted August 17, 2026

Abstract

Vocabulary knowledge plays a central role in English for Specific Purposes, yet immigration remains a professional domain that has received little attention in specialised word list research. This study developed a corpus-based Immigration Keyword and Key Cluster List (IKKCL) to support ESP learners working with immigration forms. Data were collected from 1,670 official immigration forms downloaded from government websites in different countries. Of these, 1,500 forms formed the primary Global Immigration Forms Corpus (2,939,837 tokens), and the remaining 170 formed a secondary validation corpus (179,897 tokens). Candidate items were extracted through keyword analysis, applying Log-likelihood, Log Ratio, and Simple Maths together against the British National Corpus, followed by lexical profiling and, for two-word clusters, compounding identification and Mutual Information filtering. Surviving candidates were then validated against five institutional immigration glossaries and, where needed, examined through expert concordance analysis before being classified into technical, semi-technical, and supportive profiles. The final IKKCL contained 589 items, including 21 technical, 196 semi-technical, and 372 supportive words and clusters. The list covered an average of 19.06% of running words across both corpora, with near-identical coverage on the unseen validation corpus, indicating stable performance within the same sampling frame. Combined with the General Service List and the Academic Word List, total coverage reached approximately 90%, falling short of the threshold typically associated with independent reading comprehension. The IKKCL can be used as a resource for teaching and learning in courses such as English for Immigration and Tourism, as well as related curricula, through approaches such as data-driven learning.

Keywords: word list, immigration, keyword, cluster, corpus-based

In global communication, English plays a crucial role, often being used in professional, academic, and international contexts. The four language skills—listening, speaking, reading, and writing—are central to effective communication (Yalcinkaya et al., 2009). Vocabulary is a fundamental component of language competence as it forms the foundation of learners' ability to use and develop these skills. Wilkins (1972) emphasised the importance of vocabulary, stating that while limited grammar may still allow some meaning to be conveyed, without vocabulary, nothing can be communicated. Vocabulary knowledge is therefore particularly important in specialised professional environments, where the use of precise meaning and context-appropriate language is essential. English for Specific Purposes (ESP) has received sustained attention because it aims to meet the communication needs of learners in specific academic and professional contexts (Hyland, 2019). One important goal of ESP teaching and learning is to provide learners with specialised vocabulary related to their professional contexts and registers (Otto, 2021). Different professional fields rely on different vocabulary to communicate effectively within their occupational contexts. Consequently, the development of specialised vocabulary resources remains a key issue in ESP research and pedagogy.

One professional domain that has received limited lexical attention is immigration. Immigration procedures are closely linked to international mobility, tourism, border control, and governmental administration. In many countries, immigration personnel, tourism professionals, and related service workers need to communicate with international travellers using English. One important communicative task involves explaining and verifying immigration forms, such as visa applications, arrival and departure forms, declarations, permits, and records of identity-related information. These forms contain technical vocabulary that learners in fields related to immigration need to understand and use accurately. The specialised language of immigration documents is not limited to single words; it also includes multi-word units or lexical clusters, such as *legal representative*, *physical address*, or *travel document*. Such word combinations often function as fixed or semi-fixed expressions, which convey meanings that cannot be fully understood through isolated words alone. Therefore, focusing exclusively on single-word vocabulary may not adequately represent the lexical demands in the context of immigration. Creating specialised word lists can help meet the vocabulary needs of learners by providing targeted lexical resources for learning and teaching. Explicit vocabulary instruction supported by carefully designed word lists has been recognised as an effective approach to vocabulary development (Nation, 2016; Schmitt, 2000). This is the rationale behind the current study.

Literature Review

Word lists have a long history of development. A major milestone was the creation of the General Service List (GSL) (West, 1953), which consists of approximately

2,000 high-frequency English words widely used in everyday communication and generally covers around 80% of a general corpus. Another significant turning point was the emergence of specialised word lists such as the Academic Word List (AWL) (Coxhead, 2000), which contains approximately 570 academic vocabulary items and covers around 10% of academic corpora. Since then, the development of other specialised word lists has received considerable attention. Generally, specialised word lists aim to cover approximately 5% to 8% of a corpus, with the ultimate goal of combining the GSL, AWL, and subject-specific technical word lists to enable learners to reach a vocabulary threshold of at least 95% to 98% of running words in that specific corpus (Schmitt et al., 2011), which is believed to help learners achieve effective text comprehension. Regarding research on word list development, one of the most debated issues concerns the approaches and criteria used to develop word lists. In early work, frequency was the most widely accepted metric, stemming from the rationale that the more often a word appears in a corpus, the more important it must be. However, relying solely on frequency can lead to biased results, as high-frequency words often include function words or generic terms that lack specific relevance to a subject field (Schmitt & Schmitt, 2014), directly contradicting the objective of creating specialised word lists. To mitigate this bias, the range criterion was introduced. This metric determines how widely a word is distributed across various texts within a corpus. Without range, a word might meet the threshold simply because it appears repeatedly in a single large source, despite not appearing in other sources at all. Combining both metrics creates a more rigorous filtering system. Later, lexical profiling was proposed as a supplementary criterion (Coxhead, 2000), with the underlying principle that a word should belong to only one lexical profile. For instance, if a candidate word already appears in the profiles of the GSL or AWL, it will be excluded (Meebangsai et al., 2023). Conversely, words that do not appear in these baseline reference lists are retained, as they have a higher probability of being technical vocabulary. Beyond these corpus-based quantitative metrics, researchers also incorporate judgement-based, qualitative methods into the selection process. These methods include expert verification, where words passing the initial corpus filter are compiled into rating-scale questionnaires for subject-matter experts to evaluate, as well as cross-checking with technical dictionaries and the consideration of contextual clues.

Although word list development methods involve rigorous and complex processes to obtain vocabulary that is more specialised and suitable for learners, questions remain, particularly about how frequency and range cut-off points should be determined. Since no empirical studies have established how these cut-off points should be established, there are no fixed criteria or universal cut-off points in word list development. In this regard, past studies tended to consider raw frequency or absolute frequency from previous research and apply the rule of three to their own corpus to determine cut-off points. The criteria therefore remain flexible, indicating that vocabulary filtering criteria vary according to the specific focus of the word list

(Hyland & Tse, 2007) and the characteristics of the database used to compile the corpus (Laosrirattanachai & Laosrirattanachai, 2021). However, a clearer shift in the research paradigm has emerged, as word list developers have begun to consider relative frequency metrics instead (Dang & Webb, 2025). This shift involves using relative frequency, such as keyword analysis, instead of absolute frequency, as well as the use of dispersion measures to replace range.

Keyword analysis functions by comparing the frequency of words between two corpora: a target corpus (which is usually a self-compiled specialised corpus for studying a specific subject) and a reference corpus (typically a much larger general corpus, such as the British National Corpus or BNC). Modern keyword analysis relies on three distinct yet highly complementary statistical metrics to improve accuracy: Log-likelihood (LL), Simple Maths (SM), and Log Ratio (LR). LL tests whether that the difference in word frequency between two corpora is unlikely to have occurred by chance (Rayson & Garside, 2000). The main strength of LL lies in its reliability when dealing with large datasets, as it allows researchers to establish clear critical values based on desired confidence levels. Its primary limitation is its extreme sensitivity to corpus size, meaning that as the volume of data increases, LL values surge rapidly, making almost every word appear statistically significant. For this reason, its current usage is often restricted to serving as a preliminary screening tool to check whether a word is worth further analysis. Meanwhile, SM (Kilgariff, 2009) was designed to assist in ranking keywords. The main advantage of SM is the use of a fixed parameter to weight scores, which prevents extremely rare words from appearing disproportionately near the top of the list. SM thus helps create a balance between frequent words and genuinely specialised vocabulary. Nevertheless, SM does not provide p-values in the way LL does, so SM is primarily used to rank keywords from highest to lowest to present a more natural picture of a specialised corpus. Finally, LR (Hardie, 2014) functions as an effect size metric that quantitatively indicates how prominent a word is by measuring how many times more often a word appears in one corpus compared to another. The strength of LR is its transparency and easy interpretability for human researchers, but a limitation arises when a word does not occur at all in the reference corpus, making the calculation impossible without an adjustment. The systematic integration of these three statistical metrics helps build a robust and highly reliable framework by using each metric as a filter for different analytical dimensions.

Another important advancement is using dispersion as a filtering criterion to replace traditional range. Dispersion measured using Deviation of Proportions (DP) indicates how evenly a word is distributed across a corpus. Values range from 0 to 1, where 0 indicates a completely even distribution and 1 represents an uneven distribution (Gries, 2008). In addition to relative frequency and dispersion, other qualitative methods are also used in word list development, such as analysing concordance lines. Using this data-driven corpus-based approach allows researchers to examine the surrounding linguistic context of a word, which helps them determine whether a

candidate term qualifies as technical vocabulary and assess the level of technicality of that term (Laosrirattanachai & Laosrirattanachai, 2025).

Previous studies have proposed specialised word lists for various fields (Coxhead, 2000; Drašler & Kavalir, 2026; Hanks et al., 2024; Laosrirattanachai et al., 2026a; Phoocharoensil, 2026), and such resources have been used to support vocabulary learning in EFL and ESP contexts (Lessard-Clouston, 2013). However, although specialised lists have been developed in many domains, there appears to be a lack of corpus-based lexical resources specifically for the immigration context. This study therefore aimed to develop a corpus-based Immigration Keyword and Key Cluster List (IKKCL) that combines single-word keywords and two-word clusters, using data from immigration-related forms from government immigration websites in various countries. The compiled word list aims to support learners in ESP environments, especially those studying English for tourism, immigration, and related fields, and to serve as a practical resource for instructional materials development, classroom teaching, and self-directed vocabulary learning. To achieve the objectives of this study, the following research questions were formulated.

1. What are the specific single-word keywords and multi-word key clusters that constitute the Immigration Keyword and Key Cluster List (IKKCL)?
2. To what extent does the IKKCL cover the running texts in both the primary Global Immigration Forms Corpus (GIFC) and the secondary validation corpus (2nd-GIFC)?

Research methods

Corpus Compilation

To develop the IKKCL, data were collected from the official government websites responsible for immigration affairs in countries and regions worldwide by downloading all publicly available immigration forms. Eligible documents were official immigration forms written entirely in English, available in PDF format, and compatible with corpus analysis software. The search yielded 2,711 forms. Duplicate files were removed using AllDup software, and the remaining files were manually screened to exclude non-English documents and scanned image-based PDFs. After screening, 1,670 forms from 153 countries and regions were retained. Because responsibility for immigration varies across countries, agency-level classification was not recorded. Instead, the forms were grouped into six categories based on form type, as shown in Figure 1. These categories were not evenly represented. Legal, monitoring, business and finance forms and visa and residence permit forms together accounted for over two-thirds of the full dataset, while health, medical and biometrics forms and consent, information and compliance forms made up a comparatively small share.

Figure 1

Composition of the Corpus by Form Type

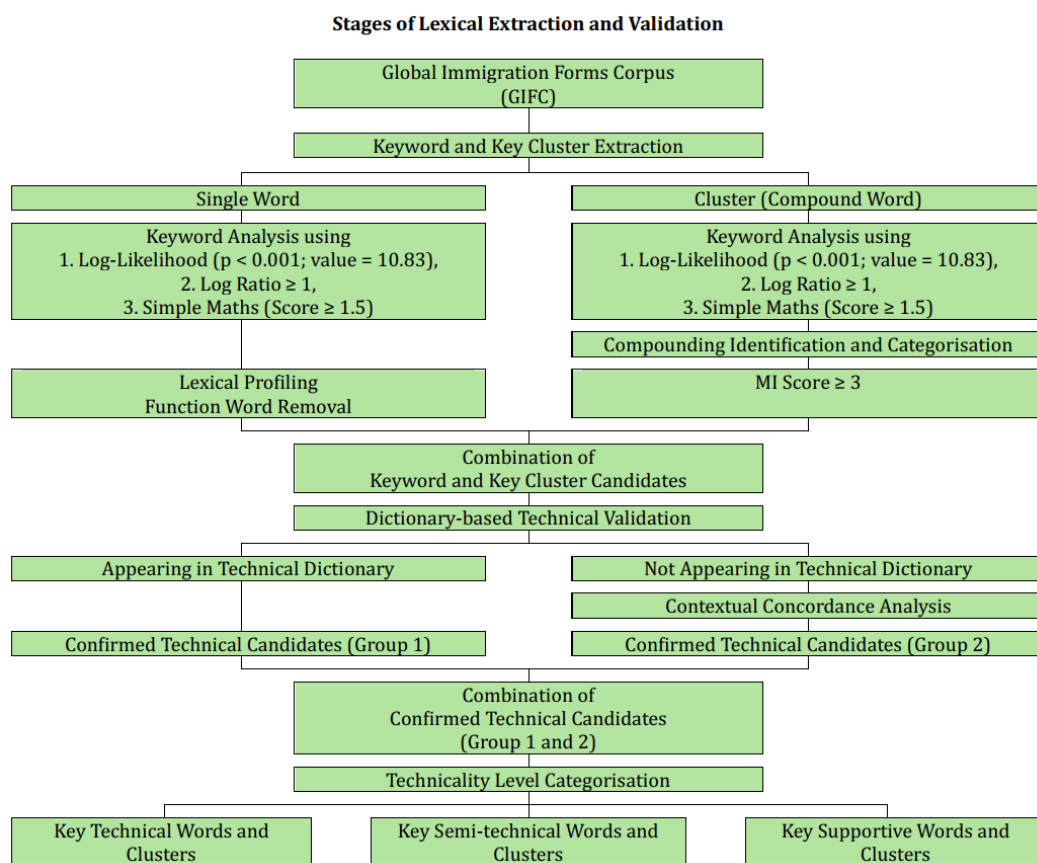
Form type	All forms	GIFC	2nd-GIFC
1 Legal, Monitoring, Business & Finance	576	517	59
2 Visa & Residence Permits	571	512	59
3 General Administration & Personal Records	203	183	20
4 Citizenship & National Identity	191	172	19
5 Health, Medical & Biometrics	75	67	8
6 Consent, Information & Compliance	54	49	5
Total	1,670	1,500	170

A stratified random sample of 170 forms was drawn with approximately 10% selected from each form type. These forms were compiled into the 2nd-GIFC, containing 179,897 tokens (14,013 types), and were used to evaluate the text coverage of the IKKCL. The remaining 1,500 forms were compiled into the Global Immigration Forms Corpus (GIFC), containing 2,939,837 tokens (44,445 types).

Development of the IKKCL

Corpus research requires analysing large amounts of data, which necessitates appropriate procedures and methods. A preliminary trial analysis was therefore conducted. The research team initially designed a framework for developing the IKKCL and trialled it on approximately 30% of the GIFC. The trial demonstrated that the dispersion criterion should be removed, as immigration documents are diverse and divided into distinct types. Because each form has a completely different focus, the vocabulary used changes dramatically from one document to another. This was empirically confirmed during the trial phase, where the DP calculation revealed that several domain-relevant terms had DP values near 1 (e.g., *migrant* = 0.87, *dual citizenship* = 0.91) and would thus be incorrectly excluded. Consequently, dispersion was deemed unsuitable for a topically heterogeneous corpus and was omitted from the analysis. The development process of the IKKCL is shown in Figure 2.

Figure 2
Stages of IKKCL Development



As illustrated in Figure 2, the development of the IKKCL followed a sequential process divided into two main phases, a corpus-based approach and a judgement-based approach. A corpus-based approach was chosen initially because of its efficiency in processing large volumes of textual data. Because the target list contains both single words and two-word clusters, the selection process had to be separated. Extracting single words and extracting multi-word units require distinct corpus linguistic techniques. Therefore, the development of these two lists proceeded along separate pathways before being integrated.

Software settings and thresholds applied throughout the IKKCL development are summarised in Table 1. Case-insensitive matching was applied at every stage involving automated word or cluster identification, so that spelling variants differing only in capitalisation were treated as instances of the same item.

Table 1
Software Settings and Selection Thresholds

Stage	Software (version)	Setting / Threshold
Keyword analysis (single word & cluster)	#LancsBox X (v6.0.1)	Reference corpus: BNC2014 (bundled with #LancsBox X); Log-likelihood ≥ 10.83 ($p < 0.001$); Log Ratio ≥ 1 ; Simple Maths ≥ 1.5 ; case-insensitive matching
Trial dispersion check	#LancsBox X (v6.0.1)	Deviation of Proportions (DP; Gries, 2008)
Lexical profiling (function word removal)	AntWordProfiler (v2.2.1)	Reference list: Function Word List (Nation, 2022); case-insensitive matching
Compound identification	Manual (research team)	Two independent raters; disagreement resolved by a third rater
MI score filtering (clusters)	AntConc (v4.2.0)	Mutual Information (MI) statistic; window span L1–R1; threshold MI ≥ 3
Single-word coverage	AntWordProfiler (v2.2.1)	Reference lists: GSL, AWL (default), IKKCL (added)
Cluster coverage	AntConc (v4.2.0)	Frequency count of each cluster (word list / concordance search)

Corpus-Based Approaches

Single-Word Candidate Selection. The extraction of single-word candidates began with a keyword analysis using #LancsBox X (Brezina & Platt, 2026) to compare the GIFC with the BNC. As outlined above, the three measures were applied together so each could offset the weaknesses of the others. LL was used at a critical value of 10.83 ($p < 0.001$), well above the standard baseline of 6.63 ($p < 0.01$) (Gabrielatos & Marchi, 2011; Rayson & Garside, 2000). The LR threshold was set at ≥ 1 , meaning a candidate had to occur at least twice as often in the immigration corpus to be selected (Hardie, 2014). SM was applied with a threshold of ≥ 1.5 to rank the surviving candidates (Kilgariff, 2009). A word had to satisfy all three criteria at once to move forward. Candidates that passed this stage but still functioned as general function words were then removed through lexical profiling in AntWordProfiler (Anthony, 2026a), using the Function Word List (Nation, 2022) as the reference. It should be noted that GSL items were not removed at this stage, unlike the exclusion procedure commonly used in specialised word list research. This reflects the nature of immigration vocabulary, where familiar everyday words often take on narrower or legally defined meanings in official texts. For example, *act*, *issue*, and *period* are common in general English but carry more specific meanings in immigration forms. Excluding them based solely on the GSL would therefore have removed much of the semi-technical vocabulary targeted by this study. This decision is further supported by word list developers who argue against excluding GSL words, proposing that specialised

vocabulary should cut across high- (e.g., GSL), mid-, and low-frequency words (Gardner & Davies, 2014; Nation, 2016). Instead, function words were filtered because they carry little lexical meaning and do not show this type of semantic shift. Words found on this reference list were discarded, while the remaining words advanced to the integration stage.

Cluster Candidate Selection. The filtering of two-word clusters also began with a keyword analysis in #LancsBox X, using the same statistical criteria as the single words. In corpus linguistics, multi-word combinations extracted through statistics are generally called collocations. To refine these collocations into meaningful vocabulary entries, this study focused exclusively on compound words. Crucially, this cluster pathway only selected open-form compounds, which are words written with a space between them but functioning as a single concept, such as *permanent resident*. Closed-form compounds written as a single word, like *passport*, and hyphenated compounds, like *re-entry*, were automatically captured during the single-word filtering phase.

Open-form compounds were identified manually because distinguishing genuine compounds like *permanent resident* from non-compound fragments requires linguistic judgment. Two research team members independently reviewed all candidate clusters and achieved 100% agreement, making resolution by a third member unnecessary. Finally, the remaining clusters were analysed using AntConc (Anthony, 2026b) to calculate their MI score. The threshold criterion of $MI \geq 3$ was used as a filter to ensure that the selected word pairs have a strong semantic bond and co-occur intentionally rather than by chance. At the end of this phase, the selected candidate terms, both single-word and cluster types, were combined, completing the corpus-based stage.

Judgement-based approaches

Dictionary-Based Technical Validation. As immigration is a specialised field, using general standard dictionaries to verify whether words are technical vocabulary is insufficient. Furthermore, the search identified no specialised immigration dictionary. Instead, five institutional glossaries and online dictionaries were located: Glossary on Migration (IOM, 2019), USCIS Glossary, Immigration Equality Asylum Manual Glossary, SimpleCitizen Immigration Term Dictionary, and NAFSA Immigration Glossary. All candidate terms combined from the corpus-based approach stage were cross-checked against these five sources. Candidates appearing in at least one source proceeded to the next stage.

Contextual Concordance Analysis. Candidates that did not appear in any glossary were not discarded immediately but underwent a contextual concordance analysis. Three experts with more than five years of experience in teaching English for tourism, aviation, and immigration were invited to validate the list, as these three fields regularly intersect around immigration documentation. Aviation and tourism personnel routinely

handle arrival, departure, and declaration forms alongside immigration officers, and English instruction in these fields commonly draws on the same document types examined in this study. Previous studies developing specialised word lists have typically relied on questionnaires containing candidate lexical items and four-point rating scales, requiring experts to judge each item based on their professional knowledge and experience. In the present study, however, to facilitate a more accurate and contextually appropriate evaluation, the three experts were first trained to use concordance lines to examine and interpret the meanings of lexical items within their surrounding contexts. Subsequently, they independently analysed the concordance lines for each word in the GIFC using #LancsBox X, drawing on their professional knowledge and experience to determine whether its contextual meaning was relevant to the immigration domain (Yes/No). The three experts initially achieved 98.32% raw percent agreement on word relevance, and inter-rater reliability was further confirmed by Fleiss' kappa ($\kappa = .977$), indicating almost perfect agreement (Landis & Koch, 1977). The remaining disputed items were discussed until full consensus was reached. Ultimately, only words unanimously deemed relevant to the immigration domain proceeded to the technicality level categorisation.

Technicality Level Categorisation. The validated candidates were then categorised into three distinct vocabulary profiles based on operational definitions adapted from previous corpus-based word list studies (Laosrirattanachai et al., 2026b).

The technical profile comprises vocabulary that is used, and carries meaning, exclusively within the immigration domain. These words do not occur with comparable meaning in general or other specialised contexts, making their presence in a text a strong and unambiguous marker of immigration-related content.

The semi-technical profile consists of polysemous words that carry an established meaning in general or other specialised contexts, but which acquire an additional, domain-specific meaning when used within immigration texts. Their interpretation therefore depends on the surrounding immigration context, distinguishing them from words whose meaning remains stable across contexts.

The supportive profile refers to vocabulary whose meaning remains essentially the same whether it occurs in immigration texts or in general and other specialised contexts. These words do not acquire a new, context-specific sense within the immigration domain. Their significance lies instead in their high frequency of occurrence in immigration documents, where they act as a rhetorical scaffold that helps readers follow the structure and logic of a document, rather than conveying subject-specific content.

The same three experts who participated in the contextual concordance analysis stage were asked to independently classify the words into these three categories using the same concordance lines examined in the previous stage. The three experts achieved 93.38% initial raw percent agreement on word classification, corresponding to a Fleiss' kappa of $\kappa = .906$, which likewise reflects almost perfect agreement (Landis &

Koch, 1977). Words with complete consensus were immediately categorised, while disputed items were discussed until full agreement was reached.

Lexical Unit Selection: The Type

In word list development studies, researchers generally choose among three units of counting (Dang, 2020; Nation, 2016), namely *types*, where every spelling variation is counted separately; *lemmas*, which group base forms with inflections; or *word families*, which include prefixes and suffixes like *application* or *inapplicable*. For a specialised domain word list, types are more suitable than the other two options (Dang, 2020). Using word families or lemmas can be misleading in specialised contexts because changing a word from one part of speech to another through suffixation, or using different inflected forms, can completely alter its technical meaning. For instance, the word *alien* as a noun has a strict legal definition in immigration law, whereas the adjective *alienate* belongs to general academic or psychological language. Therefore, types provide an appropriate balance of grammatical precision and semantic clarity for language learners. Accordingly, to develop the IKKCL, the type was selected as the basic counting unit. This means that every word with a different spelling was counted separately, even if it was derived from the same headword. For example, *apply*, *applies*, *applied*, and *applying* were treated as separate items.

The Validation of the IKKCL

The final step involves validating the IKKCL by considering its text coverage. First, the IKKCL was tested against the main corpus, the GIFC, to examine the percentage of running words it covers. Additionally, the combined percentage of GSL, AWL, and IKKCL was considered to determine whether this combined vocabulary reaches the threshold criterion of 95%, which is necessary for learners to achieve independent reading comprehension of these documents (Schmitt et al., 2011). Subsequently, in the second step, the IKKCL was tested against the second corpus, the 2nd-GIFC. As the 170 forms in the 2nd-GIFC were not used at all in developing the IKKCL, testing the IKKCL on this independent dataset serves as a measure of the predictive power of the IKKCL. If the coverage on the 2nd-GIFC is close to the coverage on the GIFC, it indicates stable performance within the same sampling frame. AntWordProfiler calculates text coverage only at the single-word level and cannot process multi-word units. Coverage for the two-word clusters in the IKKCL was therefore calculated separately from single-word coverage. Cluster coverage was calculated first, using the original GIFC and 2nd-GIFC. The frequency of each cluster was obtained using AntConc, and its token coverage was calculated as its frequency multiplied by two, divided by the total number of tokens in the corpus, and multiplied by 100 to express the result as a percentage. This calculation counts the number of words that make up each cluster and multiplies that figure by its frequency, since a single occurrence of a two-word cluster occupies two token positions within the running text rather than one. However, a further methodological

issue arises when two clusters in the IKKCL overlap within the same string, sharing a common word. To identify such cases, all cluster pairs in which the final word of one cluster matched the initial word of another were first listed in a table. Concordance checking was subsequently limited to these potentially overlapping pairs rather than being performed exhaustively across every pair of clusters. For example, the string “*visa application form*” contains both the cluster “*visa application*” and the cluster “*application form*”, so the word *application* could therefore potentially be counted twice.

Since cluster frequencies were searched individually, each instance of an overlapping cluster was cross-checked against its concordance lines in AntConc to determine whether it occurred as part of a larger overlapping sequence. If an instance of “*application form*” was found to be embedded within “*visa application form*”, meaning *application* had already been counted as part of the co-occurring cluster “*visa application*”, that particular instance was counted as contributing only one additional token, *form*, rather than two, to avoid counting *application* a second time. Conversely, if an instance of “*application form*” occurred independently of “*visa application*”, the standard procedure was applied, and both words were counted in full, meaning its frequency was multiplied by two. The token counts obtained from this overlap-adjusted procedure were then combined with those of the remaining clusters to calculate the total cluster coverage. The coverage values of all clusters were summed to obtain the total cluster coverage. All instances of these clusters in the GIFIC and 2nd-GIFIC were then replaced with a placeholder string (“xxx xxx”), producing modified-GIFIC and modified-2nd-GIFIC. This step prevented single words that also form part of a cluster, such as *act* in *aliens act*, from being counted a second time when the containing cluster had already been counted separately.

Single-word coverage for the GSL, AWL, and IKKCL was then calculated from these modified corpora using AntWordProfiler. As the placeholder tokens matched none of the reference lists, they were automatically categorised as OWL items. The total cluster coverage obtained in the previous step was therefore subtracted from the OWL percentage and added to the IKKCL percentage. Further, to avoid double-counting, AntWordProfiler automatically attributes tokens appearing in multiple lists (e.g., *act* in both GSL and IKKCL) exclusively to the newly added list (in this case, the IKKCL). This software setting aligns with the study’s objective to prioritise immigration-specific vocabulary, ensuring that coverage figures in Table 3 summed to 100% without overlap.

Findings

RQ1: Single-Word Keywords and Multi-Word Key Clusters Constituting the IKKCL

Following the two-phase procedure outlined in the Method, candidate items were narrowed down through a corpus-based approach before being subjected to

judgement-based validation. Table 2 summarises the number of candidates retained at each stage of this process.

Table 2

Number of Candidates Retained at Each Stage of IKKCL Development

Stage	Criterion	Candidates retained
1. Key single-word extraction	Log-likelihood	31,881
	Log Ratio	31,899
	Simple Maths	1,016
	All three criteria combined	993
	Lexical profiling	881
2. Key cluster extraction	Log-likelihood	170,527
	Log Ratio	354,661
	Simple Maths	1,757
	All three criteria combined	1,750
	Compounding identification	186
	MI score filtering ($MI \geq 3$)	179
3. Single words + clusters combined	881+179	1,060
4. Dictionary-based technical validation	Found in ≥ 1 glossary	48
5. Contextual concordance analysis	Unanimous expert agreement	541
6. Glossary + concordance combined	-	589
7. Technicality level categorisation	Technical	21
	Semi-technical	196
	Supportive	372

Corpus-Based Approach

In the single-word extraction stage, LL, which was used here as a preliminary screening tool against a critical value of 10.83, retained a large pool of candidates of 31,881 words, while LR, with its threshold set at ≥ 1 , yielded a nearly identical figure of 31,899 words. This near equality is not surprising, as both LL and LR address the same underlying frequency differences between the GIFIC and the BNC, and a large corpus such as the GIFIC inevitably allows many words to pass lenient significance thresholds. SM produced substantially different results. As this metric incorporates a fixed parameter to counterbalance the influence of rare or extreme-frequency items, it functioned as a much stricter filter, reducing the candidate pool to only 1,016 words. In total, 993 words met the criteria for all three statistical measures, such as *prescribed*, *passports*, *settlement*, *approval*, *divorced*, and *adoption*. The subsequent lexical profiling stage, responsible for eliminating function words, reduced the qualifying vocabulary further to 881 words, such as *application*, *residence*, *immigration*, *nonimmigrant*, *comply*, and *Schengen*. The magnitude of this reduction, from a

statistically significant list of nearly 32,000 items down to fewer than 900 candidate words, clearly demonstrates why relying on a single statistical metric is problematic. Each filter targets a different type of noise, namely LL screens out chance occurrences, LR measures effect size, SM manages disproportionate rarity, and lexical profiling removes words that, despite their statistical prominence, are grammatical rather than within the immigration domain.

Despite the larger figures, cluster extraction produced a similar pattern. LL identified 170,527 candidate clusters, and LR, somewhat surprisingly, yielded an even larger total of 354,661 items. This divergence between single-word and cluster numbers reflects the nature of two-word combinations, as the resulting clusters stem from a much broader combinatorial space than single words. Even minor frequency differences between the GFC and the BNC can generate vast numbers of statistically distinct word pairs. SM once again acted as the most rigorous filter, reducing the candidate pool to 1,757 items, and the combination of all three statistical criteria produced 1,750 clusters, such as *residence permit*, *your application*, *this form*, *you must*, and *personal data*. Notably, many of the surviving cluster items, such as *this form* and *you must*, are not compound words in the strict sense, but rather frequent functional fragments. This issue highlights the importance of the compounding identification stage. After filtering specifically for open-form compounds, 186 clusters were retained, such as *personal data*, *application form*, *immigration service*, *travel document*, and *birth certificate*. Every cluster surviving this stage was a compound noun, which reflects the nominal, naming-oriented nature of immigration form language, as these documents are dominated by expressions that name people, documents, procedures, and legal categories, leaving comparatively little room for descriptive or adverbial phrasing. Subsequently, to confirm that these clusters result from word combinations with strong semantic bonds rather than accidental co-occurrences, an MI threshold of ≥ 3 was applied, leaving 179 candidate clusters, such as *alien registration*, *biometric features*, *criminal offence*, *employment contract*, and *naturalisation service*. This sharp reduction from 1,750 down to 179 candidates underlines the crucial importance of this final statistical verification in distinguishing genuine lexical clusters from loosely connected word sequences that merely passed initial frequency thresholds.

When combining the candidate terms from both single-word and cluster types selected via the corpus-based approach, a combined total of 1,060 candidates was obtained, comprising 881 single words plus 179 clusters, which served as input for the subsequent judgement-based stage.

Judgement-based approach

To conduct a dictionary-based technical validation, 1,060 candidate words were checked against five institutional glossaries. Only 48 items appeared in at least one glossary. Examples included *withdrawal*, *foreign national*, *human trafficking*, *beneficiary* and *petitioner*.

This low number was not unexpected. As mentioned earlier, no specialised immigration dictionary currently exists. Institutional glossaries also tend to focus on formal legal terms rather than the broader operational words found in immigration forms. The remaining 1,012 words therefore required contextual analysis using concordance lines to determine their relevance to immigration. Three experts with experience in teaching immigration, tourism, and aviation evaluated the vocabulary set and agreed that 541 items (including *statement*, *residency*, *aliens act*, *antecedents certificate*, and *employment authorization*) were contextually relevant to the field. Combining these words with those validated by the glossaries produced a final set of 589 items. These items formed the IKKCL before further categorisation.

The same three experts then classified the technicality levels of these 589 items with the same concordances at hand. They separated the words into three profiles. The technical profile includes words with fixed meanings related to this field, which are not found outside immigration. It includes 21 items, such as *aliens act*, *deportation*, *entry ban*, *INA*, and *work visa*. The semi-technical profile contains words that have general meanings but that, in an immigration setting, have a more specific meaning. This profile is considerably larger, with 196 items. Among these are *adjustment*, *approval*, *bridging*, *clearance*, *fingerprints*, and *naturalisation service*. The supportive profile is made up of terms that help connect texts rather than immigration concepts. It has 372 items, including *application form*, *biometric features*, *contact information*, *financial support*, *privacy act*, and *widowed*.

Several observations can be made from the distribution of the words in the three profiles. The semi-technical profile contains about nine times as many items as the technical profile. This result is consistent with the suggestion made in the introduction that the language of immigration is made up of general words whose meanings in official documents take on particularised senses. Learners may already know some common words such as *bridging*, which in everyday use suggests a physical or temporal link but in immigration forms refers to a specific temporary visa category that allows a person to remain lawfully while a further application is being decided. Such terms can therefore be difficult for learners to fully grasp when they appear on immigration forms. This means that the teaching of vocabulary in this area should not be restricted to the new technical terms alone. It also needs to show how ordinary words are reframed in this context when they take on new meanings. The supportive profile is the largest group with 372 items. It plays an important role in facilitating comprehension within the immigration context. This shows that, even without particularly specific meanings, many words are still necessary to understand a specialised domain such as immigration and deserve equal attention.

RQ2: Text coverage of the IKKCL on the GIFC and the 2nd-GIFC

Two corpora were used for this validation stage precisely because coverage on the GIFC alone would only demonstrate how well the IKKCL fits the very data from which it was derived. The 170 forms comprising the 2nd-GIFC were withheld from the entire list-building process described in RQ1. Testing the IKKCL against this independent dataset therefore offers a genuine test of whether the list's coverage remains stable when applied to forms beyond those used in its construction, within the same overall sampling frame. Table 3 presents the coverage figures for each vocabulary profile across both corpora.

Table 3

Text coverage of the GSL, AWL, IKKCL, and OWL in the GIFC and 2nd-GIFC

Profile	GIFC (%)	2nd-GIFC (%)	Average (%)
GSL	68.02	67.88	67.95
AWL	4.91	4.64	4.77
IKKCL (Total)	16.96	16.40	16.68
Technical	0.25	0.01	0.13
Semi-technical	5.93	6.38	6.16
Supportive	10.78	10.01	10.40
OWL	10.11	11.08	10.59
Total	100	100	100

In the GIFC, the GSL accounted for 68.02% of the running words, and a closely comparable figure of 67.88% was recorded in the 2nd-GIFC. The similarity between these two figures already points toward a consistent pattern rather than a corpus-specific anomaly. Averaged across both corpora, the GSL reached 67.95% coverage, in support of this observation. This figure sits noticeably below the roughly 80% coverage typically reported for general corpora (West, 1953). This gap can be explained by a defining feature of immigration forms rather than by a simple shortfall in high-frequency vocabulary. Immigration forms contain a substantial number of polysemous words, many of which carry a familiar, everyday meaning captured within the GSL profile. Within immigration texts, however, the very same word forms undergo a semantic shift and take on specialised, administrative, or legal meanings instead. Because this study classified such dual-meaning items according to their function within the immigration domain, rather than their general meaning, these words were incorporated into the semi-technical profile of the IKKCL. Their tokens were therefore credited toward IKKCL coverage, rather than toward GSL coverage. This redistribution of polysemous vocabulary away from the GSL profile offers a plausible explanation for the below-benchmark figure observed here.

The AWL covered 4.91% of the GIFIC and a slightly lower 4.64% of the 2nd-GIFIC, with an average of 4.77% across both corpora. All three figures fall below the approximately 10% benchmark (Coxhead, 2000) for general academic writing. Immigration forms are administrative documents, not academic texts, and the explanatory prose and argumentative structure typically found in academic writing are largely absent from this register. A smaller role for academic vocabulary within these documents therefore follows naturally from their communicative purpose.

The IKKCL covered 16.96% of the GIFIC and 16.40% of the 2nd-GIFIC, with an average of 16.68% across both. Within each corpus, the supportive profile consistently made the largest individual contribution, reaching 10.78% in the GIFIC and 10.01% in the 2nd-GIFIC. The supportive profile followed at 5.93% and 6.38%, respectively, while the technical profile contributed 0.25% and 0.01%. This ordering held steady across both datasets and was mirrored in the combined average of 6.38% for semi-technical vocabulary, 10.01% for supportive vocabulary, and 0.13% for technical vocabulary. The consistency of this pattern across two independently sampled corpora reinforces the interpretation offered in RQ1, namely that words that shift meaning according to context recur more frequently across running text than those carrying narrowly fixed technical meanings. Highly technical terms such as *entry ban* or *work visa* tend to be tied to specific form types, whereas semi-technical and supportive vocabulary recurs across a much broader range of documents, regardless of form type.

The Outside Word List (OWL), referring to running words falling outside the three reference profiles of GSL, AWL, and IKKCL, accounted for 10.11% of the GIFIC and 11.08% of the 2nd-GIFIC, averaging 10.59% overall. Combined coverage from the GSL, AWL, and IKKCL reached 89.89% in the GIFIC and 88.92% in the 2nd-GIFIC. Both figures fall short of the 95% to 98% threshold generally considered necessary for independent reading comprehension. Regarding the remaining OWL items, scanning and observing these words during the coverage analysis revealed that many consist of proper nouns, numerical expressions, and highly form-specific terminology tied to individual countries or institutions, rather than to immigration as a broader discipline.

The near-identical performance of the IKKCL across the GIFIC and the 2nd-GIFIC (16.96% against 16.40%, a difference of less than one percentage point) indicates stable performance within the same sampling frame. Since the 2nd-GIFIC played no role in the construction of the IKKCL, this consistency demonstrates that the list is not merely an artefact of the corpus it was built from. It instead reflects genuine lexical patterns shared across official the sampled official immigration documentation, which supports the practical value of the IKKCL as a resource capable of assisting learners with immigration forms beyond those specifically sampled for this study.

Discussion

The results have implications that relate not only to the lexical profile of immigration documents, but also to broader methodological questions currently debated in specialised word list development. The following sections discuss these findings in relation to the lexical nature of immigration vocabulary, the methodological choices underpinning the construction of the IKKCL, its text coverage efficiency, and its coverage consistency across corpora.

Lexical Characteristics of the IKKCL

The 589 items in the IKKCL are not distributed evenly across the three technicality profiles, with semi-technical vocabulary forming a group of 196 items, which is nine times the size of the technical profile (21 items). This distribution aligns with the nature of immigration documents, as most of the vocabulary here consists not of rare or unusual words, but rather of general English that acquires a formal and narrower meaning within forms. Words such as *bridging*, and *civil status* convey one meaning in everyday conversation and a specialised meaning when appearing in immigration forms. Chung and Nation (2003) found a comparable pattern when evaluating technical vocabulary levels on a graded scale rather than through a simple binary split between technical and non-technical, and the results of this study support this perspective. This pattern is also closely related to the concept of lexical opacity (Watson Todd, 2017), which describes familiar words that resist straightforward understanding because their meaning shifts within a specialised or institutional context, even though the word form itself is not unfamiliar to learners. Learners who know the everyday meaning of a word like *duration* may still overlook its specific meaning on a residency form, as these two meanings are rarely taught together. This suggests that vocabulary instruction for bureaucratic texts cannot rely solely on teaching unfamiliar words, as it is essential to train learners to notice when familiar words are conveying unfamiliar meanings.

The IKKCL also retains two-word clusters rather than limiting the list to single words. Items such as *travel document*, *employment contract*, and *family reunification* were retained because their meanings are not fully predictable from those of their component words. Martinez and Schmitt (2012) made a similar case when building the PHRASE List, arguing that formulaic multi-word expressions carry meaning that single-word lists cannot show. Immigration forms use such expressions frequently, so a list restricted to single words would omit part of what learners need to recognise. Cluster analysis also reveals shifts in technicality that keyword analysis alone cannot show. The single word *immigrant*, for example, was classified as semi-technical, carrying a general sense of official permission that narrows somewhat in immigration contexts. Once *immigrant* combines with *visa* to form *immigrant visa*, however, the resulting cluster denotes a specific, fixed document type and was classified as technical. This kind of shift would remain invisible in a

single-word list, and it illustrates why examining recurring word combinations is necessary to capture the full phraseological pattern of immigration discourse.

The list also illustrates why type was chosen as the counting unit over lemma or word family. The word family built around *reside* spreads across two technicality profiles rather than staying in one. *Resident*, *residence* and *residency* were classified as semi-technical, while *reside*, *resided*, *resides*, *residing* and *residential* fell under supportive. A family-based count would have merged these forms under one headword despite their different roles in immigration texts. Nation (2016) identifies this exact trade-off between economy and precision when choosing a counting unit for word list development, and the *reside* example shows why sacrificing economy was the more defensible choice here.

Methodological Considerations in Developing the IKKCL

The reduction in the number of candidates across both phases highlights a clear methodological lesson. LL alone retained candidate items in the tens of thousands for both single words and clusters, and LR yielded similarly large figures. Only SM, dictionary validation, and expert concordance analysis narrowed these numbers down to a manageable and genuinely domain-specific set of words. Relying on any single statistical metric would produce a word list dominated by chance co-occurrences or general function words rather than immigration-specific vocabulary. This supports the notion that conducting keyword analysis requires combining statistics with different sensitivities rather than selecting any single metric as sufficient on its own (Hardie, 2014; Kilgarriff, 2009; Rayson & Garside, 2000). Furthermore, it supports the rationale for using both corpus-based and judgement-based approaches in combination, as even after passing three statistical filters and the lexical profiling stage, the judgement-based stage remained essential before this IKKCL could be considered valid.

The decision to remove the dispersion criterion after the trial analysis is a methodological choice that runs against a current trend in the field. Recent work has moved toward dispersion measures as a standard filtering step, treating even distribution across a corpus as a marker of genuine domain relevance (Gablasova et al., 2017; Gries, 2008). The GIFC, however, is not a single register in the way a general corpus is. It is made up of distinct immigration form types, including employment visas, permanent residency applications, and asylum requests, each with its own specific vocabulary. Biber (1993) argued that internal register variation within a corpus should be expected to shape how linguistic features are distributed across its texts, and this appears to be exactly what happened here. A term tied closely to one form type will naturally cluster within that subset of the corpus and score poorly on a dispersion measure, even though it may be highly frequent and functionally important within its own register. Removing dispersion therefore prevented the loss of terms such as those found in the technical profile, which tend to be tied to specific form types rather than

spread evenly across the whole corpus. It is worth noting that uneven sample sizes across form types (as can be seen in Figure 1) may also have contributed to these low dispersion scores, and the present data cannot fully separate this possibility from genuine register variation. Even so, this choice highlights a tension in specialised word list development between corpus-wide distributional evenness and within-register salience, and it suggests that dispersion criteria may need to be applied with more caution when the target corpus itself spans multiple sub-registers.

Text Coverage of the IKKCL and Its Implications

The coverage results in Table 3 raise several points worth discussing. The lower-than-benchmark GSL and AWL figures reported in RQ1, measured against the roughly 80% benchmark for general corpora (West, 1953) and the roughly 10% benchmark for academic corpora (Coxhead, 2000), reflect the redistribution of polysemous, form-specific vocabulary into the IKKCL rather than any weakness in either reference list. The IKKCL itself reached an average coverage of 16.68%, noticeably higher than the 5% to 8% range typically reported for specialised word lists in other fields. Two factors plausibly explain this. First, the IKKCL absorbed a share of vocabulary that would otherwise have counted toward the GSL. Second, immigration forms are a highly repetitive, template-driven genre, so a comparatively small set of words and clusters can recur across a large proportion of running text. This is consistent with the finding that supportive vocabulary, much of it built from standard form-filling language such as *full name*, *signed*, and *reference number*, alone accounted for over 10% of coverage.

The shortfall in combined GSL, AWL, and IKKCL coverage against the 95% to 98% threshold considered necessary for independent comprehension (Schmitt et al., 2011) matters for how the IKKCL should be used in teaching. A learner equipped with all three lists would still encounter roughly one word in ten that falls outside them, likely proper nouns, numbers, and country-specific institutional terms that cannot easily be captured in a general-purpose word list. This suggests that IKKCL instruction should be paired with strategies for handling unlisted vocabulary, such as guessing from context or recognising the structural position of names and figures on a form, rather than being treated as a complete solution to reading comprehension on its own.

Coverage Consistency of the IKKCL across Corpora

The near-identical coverage figures across the two corpora, discussed in RQ2, stand out as one of the clearest strengths of the IKKCL. Because the 170 forms making up the 2nd-GIFC played no role in building the list, the fact that coverage still landed within about one percentage point of the GIFC figure is consistent with the list capturing recurring lexical patterns of immigration documentation rather than characteristics specific to the sample it was drawn from.

This finding should, however, be read within its proper scope. Both corpora were built from the same kind of source material, official government immigration forms collected through the same criteria. The consistency demonstrated here therefore reflects stable performance within this specific genre, rather than evidence of generalisation across immigration discourse as a whole. Immigration-related communication also occurs in spoken interactions at border crossings and visa interviews, in advisory materials produced by immigration lawyers, and in news coverage of migration policy, none of which were sampled in either corpus. These registers are likely to draw on a different mix of vocabulary, since spoken interactional data in particular tends to rely less on the fixed, template-like phrasing that characterises official forms. Testing the list against these other immigration-related text types remains a task for future research and would help establish whether the IKKCL functions as a resource for the immigration domain generally or specifically for the genre of official documentation.

As noted in the Methodology, form types were unevenly sampled. Legal, monitoring, business/finance, and visa/residence forms comprised over two-thirds of the corpus, while health/medical/biometrics and consent/compliance forms were underrepresented. Consequently, the IKKCL primarily reflects the dominant categories, and technical terms specific to smaller categories may be missing. Educators and learners focused on these less-represented forms should view the IKKCL as a starting point rather than an exhaustive resource.

Pedagogical Implication

In ESP settings, learners are expected to acquire both disciplinary content knowledge and the technical vocabulary of their field simultaneously. Beyond learning technical terms, they must also develop the ability to use them accurately and appropriately according to their intended meanings and disciplinary conventions (Coxhead et al., 2019). The IKKCL developed in this study provides a valuable resource for supporting effective communication, comprising both single-word keywords and two-word lexical clusters. To help learners engage with authentic language use, we recommend incorporating data-driven learning (DDL) into ESP instruction. DDL is an approach that uses corpus data, such as word lists and concordances, to support language learning (Cobb, 2025). Implementing DDL activities in the classroom encourages learners to become more autonomous by exploring linguistic patterns using corpus-based resources. For instance, concordances are regarded as useful sources for helping learners acquire collocations and the different senses of words (Dang & Webb, 2025). This approach is particularly useful for opaque words such as *duration* or *bridging*, where the challenge lies not in recognising the word itself, but in noticing its narrower, specialised sense. Drawing learners' attention explicitly to such words within authentic concordance lines, rather than teaching them as isolated vocabulary items, can help learners connect the familiar and specialised meanings of the same word. Furthermore,

using the IKKCL, which contains both single words and two-word clusters, allows materials developers to show learners how a familiar word can take on a more fixed, technical meaning once it appears in a specific combination, such as *permit* alongside *residence permit* or *work permit*. This aligns with the notion of DDL for learning vocabulary, which aims to maximise the benefits of both word lists and concordances by combining the breadth and depth of vocabulary learning (Cobb, 2025).

More specifically, we suggest two methods: hands-on and hands-off DDL. First, hands-on DDL involves the direct use of corpus data in the teaching process by introducing learners to corpora as well as training them to investigate linguistic patterns independently (Lusta et al., 2025). In this method, the teacher can begin by introducing target words (e.g., *asylum*, *deportation*, *immigrant*) and illustrating how to search for them in corpus analysis software. Subsequently, the learners can examine the target words via concordance lines and observe the common collocations or typical contexts of use while the teacher acts as a facilitator. Second, hands-off DDL concerns the teacher-selected corpus data of the most frequent words or patterns of a target structure prepared by the teacher. Usually, the data are simplified and modified before presenting to the learners as a handout (Saeedakhtar et al., 2020). For example, the teacher can select concordance lines for a target word from the GFC and distribute them to learners as a worksheet. Then, learners can do pair work or group work to complete the worksheet by identifying word meanings, common contexts, or collocations. In the subsequent stage, once learners become familiar with the authentic use of language through DDL, it is possible to assign meaningful tasks based on task-based language teaching (TBLT) such as completing an immigration arrival form, organising travel documents or checking a visa application. It is important to note that hands-on and hands-off DDL are suitable for different language proficiency levels. Most studies have indicated that hands-on DDL is appropriate for advanced and intermediate learners, whereas hands-off DDL works well for learners at an early stage who are not familiar with inductive learning (Boulton, 2009; Daskalovska, 2015; Lee et al., 2017; Saeedakhtar et al., 2020). Therefore, teachers should carefully consider learners' language proficiency levels when designing DDL activities and select either hands-on or hands-off DDL accordingly.

Beyond the classroom, the IKKCL may also serve professional and workplace purposes. Immigration officers and other government personnel who review these forms, as well as tourism and aviation staff who assist travellers in completing them, could use the list as a quick reference for the vocabulary most likely to cause confusion. Translators and interpreters working with immigration documentation may similarly benefit from a ready inventory of terms whose meanings shift between general and institutional use. For ESP instructors and materials developers, the IKKCL and its technicality profiles offer a starting point for designing workplace training materials, onboarding resources, or short courses for professionals preparing to work in immigration-related fields, rather than requiring these materials to be built from scratch.

Conclusion

The IKKCL developed in this study offers a corpus-based vocabulary resource for a domain that had received little lexical attention until now. Drawing on 1,670 official immigration forms collected from government websites across different countries, the two-phase development process, combining statistical filtering with dictionary validation and expert judgement, produced a final list of 589 items divided into technical, semi-technical, and supportive profiles.

The findings showed that supportive vocabulary made up the largest share of the IKKCL, pointing to a domain where ordinary words are frequently used and play a crucial role, rather than one built primarily around unfamiliar technical terms. The list also demonstrated meaningful text coverage, reaching an average of 16.68% across both the primary and secondary corpora, and this coverage remained stable when tested on forms that played no part in the list's development. This consistency supports the claim that the IKKCL reflects genuine lexical patterns in official immigration documentation rather than characteristics specific to the sample corpus.

At the same time, combined coverage from the GSL, AWL, and IKKCL fell short of the threshold generally associated with independent reading comprehension, indicating that the IKKCL works best as one part of a broader vocabulary teaching approach rather than a standalone solution. The pedagogical suggestions offered in this paper, particularly the use of hands-on and hands-off DDL, are intended to help learners engage with the list through authentic immigration texts rather than through memorisation alone.

This study contributes both a practical vocabulary resource for ESP learners in tourism, immigration, and related fields, and a methodological example of how corpus-based and judgement-based approaches can be combined to build a specialised word list for a bureaucratic, form-based genre. Future research extending this work to spoken immigration discourse, testing the list directly with learners, and refining its methodology in the ways discussed above would help build on the foundation established here.

Limitations and Future Recommendations

Several limitations of this study should be acknowledged. First, the GIFIC and 2nd-GIFIC were built entirely from official written forms, which meant the IKKCL was developed and validated within a single register of immigration discourse. As already noted, spoken interactions at border crossings and visa interviews, advisory texts produced by immigration practitioners, and news coverage of migration policy were not represented in either corpus. Whether the IKKCL applies to these other forms of immigration communication remains untested.

Second, the removal of the dispersion criterion, while justified by the heterogeneity of form types within the GIFIC, means that some retained items may be concentrated in a narrow subset of documents rather than spread across the corpus as a whole.

This trade-off was necessary to avoid losing genuinely domain-specific vocabulary, but it also means the IKKCL may include a small number of items with limited applicability outside their specific form type. Future studies could examine whether supplementing the list with form-type metadata, indicating which items are broadly useful across form types, and which are concentrated in a particular one, would improve its pedagogical usefulness.

A further limitation concerns the scope of the unit of analysis. This study extracted single words and two-word clusters only, as a deliberate scope decision for this project. However, immigration forms also make frequent use of three-word fixed expressions, such as *date of birth*, *place of birth*, *next of kin*, and *power of attorney*, which arguably constitute core vocabulary for this genre. Because these longer sequences were not captured, the IKKCL should not be regarded as a complete inventory of fixed multi-word vocabulary in immigration forms. Future studies could extend the current methodology to three-word (and longer) clusters to provide fuller coverage of the fixed expressions characteristic of this domain.

An additional limitation concerns the composition of the reference glossaries used for validation. The five glossaries consulted (IOM, USCIS, Immigration Equality, SimpleCitizen, and NAFSA) are predominantly associated with United States or international organisation usage, whereas the GIFIC was compiled from immigration forms across 153 countries. As a result, the validation process may be more sensitive to terminology conventions common in US or internationally standardised immigration contexts than to country-specific usage elsewhere in the corpus. This imbalance should be borne in mind when interpreting the technical and semi-technical classifications, and future work could incorporate a more geographically diverse set of reference glossaries to reduce this potential bias.

A further limitation was the expert panel's composition. Although experienced in teaching immigration-related English, the three experts lacked formal training in law or legal English. Future studies validating similar word lists should include legal and immigration law experts alongside ESP specialists to enhance legal accuracy.

Finally, this study validated the IKKCL through text coverage figures rather than through direct testing with learners. While coverage provides a useful measure of how much of a text a word list can account for, it does not by itself confirm that learners who know the IKKCL can comprehend immigration forms more effectively, or that the technicality categories assigned by the expert panel match how learners actually process these words in context. Future research involving reading comprehension tests, vocabulary knowledge assessments, or classroom-based DDL interventions using the IKKCL would help establish its pedagogical validity beyond corpus-internal measures. Such studies would also help determine whether the three technicality profiles proposed here, technical, semi-technical, and supportive, correspond to meaningful differences in learning difficulty, or whether a different instructional sequencing would better serve learners in ESP settings for tourism and immigration.

Ethics Statement

This study involved the analysis of publicly available immigration form documents and did not involve interaction with or intervention on human subjects requiring informed consent. The expert panel's role was limited to linguistic classification of vocabulary items and did not involve the collection of personal or sensitive data about the experts themselves. As such, this research was exempt from formal ethics committee review, consistent with institutional guidelines for corpus-based linguistic research not involving human subjects research.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare no conflict of interest.

Data Availability Statement

The datasets generated and analysed during the current study, including the Global Immigration Forms Corpus (GIFC) and the secondary validation corpus (2nd-GIFC), are available from the corresponding author upon reasonable request.

References

- Anthony, L. (2026a). *AntWordProfiler* (Version 2.2.1) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software/AntWordProfiler>
- Anthony, L. (2026b). *AntConc* (Version 4.2.0) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257. <https://doi.org/10.1093/lc/8.4.243>
- Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. *ReCALL*, 21(1), 37–54. <https://doi.org/10.1017/S0958344009000068>
- Brezina, V., & Platt, W. (2026). *#LancsBox X* [Software]. Lancaster University. <http://lancsbox.lancs.ac.uk>
- Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, 15(2), 103–116. <https://doi.org/10.64152/10125/66770>
- Cobb, T. (2025). Wordlists and data-driven learning. In D. Tafazoli (Ed.), *The Palgrave Encyclopedia of Computer-Assisted Language Learning* (pp. 1–8). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-51447-0_155-1

- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- Coxhead, A., McLaughlin, E., & Reid, A. (2019). The development and application of a specialised word list: The case of fabrication. *Journal of Vocational Education & Training*, 71(2), 175–200. <https://doi.org/10.1080/13636820.2018.1471094>
- Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 288–304). Routledge. <https://doi.org/10.4324/9780429291586>
- Dang, T. N. Y., & Webb, S. (2025). Applications of word lists in second language learning and teaching. *Language Teaching*, 58(3), 291–311. <https://doi.org/10.1017/S0261444825000059>
- Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2), 130–144. <https://doi.org/10.1080/09588221.2013.803982>
- Drašler, A., & Kavalir, M. (2026). Compiling a custom corpus and word list for ESAP: The case of English for Geographers. *English for Specific Purposes*, 81, 92–102. <https://doi.org/10.1016/j.esp.2025.09.004>
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67(S1), 155–179. <https://doi.org/10.1111/lang.12225>
- Gabrielatos, C., & Marchi, A. (2011). *Keyness: Matching metrics to definitions*. University of Portsmouth.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305–327. <https://doi.org/10.1093/applin/amt015>
- Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437. <https://doi.org/10.1075/ijcl.13.4.02gri>
- Hanks, E., Hashimoto, B., & Egbert, J. (2024). The contracts word list: Integral vocabulary for reading and writing English contracts. *English for Specific Purposes*, 75, 37–48. <https://doi.org/10.1016/j.esp.2024.03.002>
- Hardie, A. (2014). *Log Ratio – an informal introduction*. Centre for Corpus Approaches to Social Science (CASS), Lancaster University. <https://cass.lancs.ac.uk/log-ratio-an-informal-introduction/>
- Hyland, K. (2019). English for specific purposes: Some influences and impacts. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 337–353). Springer. https://doi.org/10.1007/978-3-030-02899-2_19
- Hyland, K., & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly*, 41(2), 235–253. <https://doi.org/10.1002/j.1545-7249.2007.tb00058.x>
- IOM. (2019). *Glossary on migration*. International Organization for Migration.

- Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference (CL2009)* (Article 171). University of Liverpool.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Laosrirattanachai, P., & Laosrirattanachai, P. (2021). Applying lexical profiling to construct technical word lists for Thai tourist guides. *PASAA*, 62(1), 61–91. <https://doi.org/10.58837/CHULA.PASAA.62.1.3>
- Laosrirattanachai, P., & Laosrirattanachai, P. (2025). Tracing tourism business research trends in Scopus-indexed journals using corpus-based and judgement-based approaches. *Humanities, Arts and Social Sciences Studies*, 25(1), 32–53. <https://doi.org/10.69598/hasss.25.1.268122>
- Laosrirattanachai, P., Tangnaitammakun, Y., Tinhuatoy, K., Sutthongkong, A., & Laosrirattanachai, P. (2026a). Developing a global stock market terminology list for ESP learners in business English: A corpus-based approach. *The New English Teacher*, 20(2), 125–140. <https://doi.org/10.59865/t.v20i2.9688>
- Laosrirattanachai, P., Rittikote, N., & Laosrirattanachai, P. (2026b). Developing a nature tourism word list for L2 learners in the tourism sector. *Australian Review of Applied Linguistics*, Advance online publication. <https://doi.org/10.1075/aral.25109.lao>
- Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, 21(2), 32–51. <https://doi.org/10.64152/10125/44610>
- Lessard-Clouston, M. (2013). Word lists for vocabulary learning and teaching. *The CATESOL Journal*, 24(1), 287–304. <https://doi.org/10.5070/B5.36167>
- Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2025). Breaking the stereotype of data driven learning: Students' experiences of mobile data-driven learning in a developing country. *SAGE Open*, 15(4), 1–20. <https://doi.org/10.1177/21582440251386904>
- Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, 33(3), 299–320. <https://doi.org/10.1093/applin/ams010>
- Meebangsai, D., Pongtin, P., Kitipoontanakorn, P., & Laosrirattanachai, P. (2023). Investigating proficiency of academic English in student writing: A comparative case study on vocabulary utilization in student research article writing vis-à-vis national and international research. *PASAA*, 67, 66–100. <https://doi.org/10.58837/CHULA.PASAA.67.1.3>
- Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.
- Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.

- Otto, P. (2021). Choosing specialized vocabulary to teach with data-driven learning: An example from civil engineering. *English for Specific Purposes*, 61, 32–46. <https://doi.org/10.1016/j.esp.2020.08.003>
- Phoocharoensil, S. (2026). Constructing the LERWL: A corpus-based approach to identifying technical vocabulary in language education research. *rEFLections*, 33(2), 739–759. <https://doi.org/10.61508/refl.v33i2.290763>
- Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1–6). Association for Computational Linguistics. <https://doi.org/10.3115/1117729.1117730>
- Saeedakhtar, A., Bagerin, M., & Abdi, R. (2020). The effect of hands-on and hands-off data-driven learning on low-intermediate learners' verb-preposition collocations. *System*, 91, Article 102268. <https://doi.org/10.1016/j.system.2020.102268>
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484–503. <https://doi.org/10.1017/S0261444812000018>
- Watson Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, 45, 31–39. <https://doi.org/10.1016/j.esp.2016.08.003>
- West, M. (1953). *A general service list of English words*. Longman.
- Wilkins, D. (1972). *Linguistics in language teaching*. Edward Arnold.
- Yalcinkaya, F., Muluk, N., & Sahin, S. (2009). Effects of listening ability on speaking, writing and reading skills of children who were suspected of auditory processing difficulty. *International Journal of Pediatric Otorhinolaryngology*, 73(8), 1137–1142. <https://doi.org/10.1016/j.ijporl.2009.04.022>

Appendix

The Immigration Key Keyword and Cluster List (IKKCL): Due to the length of the complete list, the full IKKCL, including all single words and two-word clusters is provided as a supplementary spreadsheet accessible at the following link https://docs.google.com/spreadsheets/d/1N0sS_pgeWVVQR32BKtOe-uS9ANkA-0nNtxkUOp4bo8M/edit?usp=sharing. A sample of the list is presented in Figure 3 below for illustration.

Figure 3

Sample Entries from the Immigration Key Keyword and Cluster List (IKKCL)

Immigration Keyword and Key Cluster List (IKKCL) (589 types)					
TECHNICAL IKKCL (21 types)		SEMI-TECHNICAL IKKCL (196 types)		SUPPORTIVE IKKCL (372 types)	
1	alien registration	1	accredited	1	abroad
2	aliens act	2	adjustment	2	access
3	antecedents certificate	3	admission	3	accommodation
4	BRP	4	adoption	4	accompanying
5	CBP	5	adoptive	5	act
6	deportation	6	affidavit	6	address
7	EAD	7	agent	7	addresses
8	employment authorization	8	alien	8	adheres
9	entry ban	9	aliens	9	affairs
10	immigrant visa	10	appeal	10	agencies
11	immigration adviser	11	applicant	11	agency
12	immigration advisers	12	applicants	12	agreement
13	immigration benefit	13	application	13	allowance
14	INA	14	application fee	14	amended
15	INA section	15	application notification	15	annex
16	inadmissibility	16	applications	16	antecedents
17	inadmissible	17	approval	17	apostille
18	nonimmigrant	18	approved	18	appendices
19	nonimmigrant status	19	asylum	19	appendix