A Domain Based Approach to Semantic Lexicon Expansion

Sheryl Prentice: *Lancaster University, Department of Psychology, UK (s.r.prentice1@lancaster.ac.uk)*

Paul Rayson: *Lancaster University, School of Computing and Communications, UK (p.rayson@lancaster.ac.uk)*

Jo Knight: *Lancaster University, Medical School, UK (jo.knight@lancaster.ac.uk)*

Mahmoud El-Haj: *Lancaster University, School of Computing and Communications, UK (m.el-haj@lancaster.ac.uk)*

Solly Elstein: *University of Birmingham, Department of English Language and Linguistics, UK (SXE981@student.bham.ac.uk)*

A Domain Based Approach to Semantic Lexicon Expansion

Abstract

Current approaches to the expansion of semantic lexicons for corpus annotation are somewhat ad hoc in nature and do not generally offer a systematic means of identifying areas for development within one's lexicon. The present paper sets forward a domain based approach to semantic lexicon expansion, targeting UCREL's Semantic Analysis System (USAS). First, an updated version of the lexicon is compared to representative corpora to ascertain areas of underrepresentation in a novel method which we call K-FLUX analysis. Second, an example set of underrepresented types are targeted for development using domain specific corpora. Collectively, the results show that some corpora are more successful than others in supplementing the existing USAS lexicon. The paper discusses the various factors that should be borne in mind when utilising the proposed method before concluding with how findings might inform future developments of the lexicon, and crucially, the semantic system on which it is based.

1. Summarising the USAS lexicon and its uses

UCREL's (University Centre for Computer Corpus Research on Language) Semantic Analysis System (USAS) is a computational framework and taxonomy (Archer et al. 2002, Rayson et al. 2004a) that allows a user to analyse the types (a 'type' being a category or class of object) present within their data set (a list of USAS types and examples is provided in Supplementary Online Material: Background Study, Table 1, p. 2). The system relies on an internal semantic lexicon, i.e., a list of tokens and their associated sense(s), which have been manually coded by linguists. To clarify, by 'token' we mean a specific instance or occurrence of a type and by 'sense' we refer to 'A way in which an expression or a situation [in this case a token] can be interpreted; a meaning' (Oxford University Press 2021).

In some ways the lexicon resembles an annotated corpus, with tokens listed alongside senses. However, rather than tokens being listed with a specific sense (or senses) due to their context of occurrence (e.g. 'He began to wear on her', where the context tells us that *wear* refers to the wearing down of an individual rather than the wearing of a piece of clothing), the tokens are instead listed independently of context, with all their possible senses, thus *wear* would be listed with both the aforementioned senses. When the lexicon is deployed on an unannotated corpus, word sense disambiguation methods are used to determine the most suitable sense(s) and the corpus it has been deployed on becomes a semantically annotated corpus.

The system has been utilised for a number of purposes since its creation, including for the purpose of identifying metaphors of war, game, sport, living organisms, building, physical forces, machine, and journey within the discourse of business publications (Kheovichai 2015). The system has also been used to compare the linguistic content of a set of problem based learning sessions used in the training of clinicians, to identify key questioning, reasoning, explanatory and technical vocabulary introduced at different stages in clinicians' training to be used as a teaching and learning resource (Da Silva and Dennick 2010). In the field of forensics, the system has been used, for example, to compare the semantic differences between a corpus of fraudulent versus genuine scientific publications to identify an overproduction of scientific discourse and a greater use of amplifiers and terms relating to certainty in fraudulent papers (Markowitz and Hancock 2014).

While the system was originally devised with general use in mind, its application in more specialised areas prompts the question of how well the system's current lexicon represents particular domains (defined as 'an area of activity, interest, or knowledge', (Longman 2021) within the general English language. To ensure that the system is effective for analyses across a range of domains and returns a minimal number of unmatched tokens, it is of further importance that its lexicon be continually expanded. This paper therefore presents a method for the domain based expansion of the USAS lexicon, which could also be adopted for lexicons of a similar nature. It is to a review of current methods in this area that the paper now turns.

2. Domain based approaches to lexicon expansion

Researchers have tackled lexicon expansion from a number of different perspectives, including computational linguistics, psycholinguistics, English for Specific Purposes (ESP), English as a Foreign Language (EFL), and lexicographical and terminology extraction. Löfberg (2017), for example, proposes guidelines for domain specific expansion by focussing on such tasks as named entity recognition, internet content monitoring, and psychological profiling.

An event based approach has also been suggested, in which one samples data occurring in response to a particular domain specific event and uses this data to assemble a list of new tokens to add to an existing lexicon (see Downes and Goodman 2014). Olteanu et al. (2014), for example, use the locations and hashtags of six disasters to collect crisis related tweets, and crowdsourcing to filter crisis related tokens from non crisis tokens. To extend the lexicon, crisis tokens are used to obtain 'pseudo-relevant' tweets. This is a method known as pseudo relevance feedback in the field of Information Retrieval, in which the most frequent terms in the highest ranked documents returned by one's initial query, or a comparison between the highest ranked documents and all documents returned by one's query, are used to extend the scope of one's search (Cao et al. 2008). Unigrams and bigrams not already present in the lexicon are amassed. Tokens are then selected for inclusion based on their frequency of occurrence in the pseudo relevant tweets or on their cooccurrence with original query tokens.

An event based approach would not be sufficient to extend specific domains within the USAS lexicon in that tokens might be too tied to events and not general enough in nature. Adopting a pseudo relevance feedback approach that is not tied to specific events and is instead tied to a set of tokens from a specific type in the lexicon could, however, be viable. Nonetheless, the relevance of the tokens returned by this approach has been brought into question and researchers have recommended the use of additional processes to predict the usefulness of its output (Cao et al. 2008). However, there are other methods that may be more suited to the present case.

A seed token approach, for example, popular in sentiment analysis, entails using a set of known tokens with the same sense to find new tokens with that sense. Vania et al. (2014), for example, take seed tokens with high polarity or subjectivity values from existing sentiment lexicons to assemble three token ngrams from a corpus of user reviews. Seed tokens are replaced with a place holder and structures with a frequency of 50 or above are searched within the remaining corpus data to produce new candidate tokens. One could also plot known tokens with remaining corpus tokens in a multidimensional space to discern groupings (see Pantel et al. 2009).

Another approach would be to utilise existing domain specific resources, as in the field of English for Specific Purposes (ESP). Granger and Paquot (2010), for example, utilise the academic portion of the British National Corpus (BNC) and EFL learner corpora to assemble a list of ~900 academic tokens and phrases to build the Louvain English for Academic Purposes (EAP) dictionary. Where existing resources for a particular domain are not available, these can be built. Rao et al. (2013), for example, compile a corpus of emotion related news consisting of articles and headlines that have been rated according to a set of emotions by human judges. The method of Latent Dirichlet Allocation (LDA), which assumes that documents are generated from a number of latent topics, is used to model emotional types in this corpus.

As with all lexicons, in expanding the USAS lexicon, two key factors must be considered: first, the acquisition of new tokens, and second, the assignment of their sense(s). The domain based approaches to lexicon expansion covered thus far tend to rely on the tools concerned to automatically generate new tokens and assign them to appropriate senses based on their relationship with existing known tokens or their appearance in domain specific resources. However, Makki et al. (2014) have observed that involving users in the assignment of tokens significantly improves the quality of the lexicon generated. Therefore, in Study 2, the present paper aims to produce a semiautomated approach to domain specific lexicon expansion that allows for replicability, without losing the advantages that human judgement affords. In addition, Study 1 sets forward a method to identify areas of weakness within a semantic lexicon, before one deploys a domain specific lexicon expansion method.

3. Study 1: Identifying underrepresented areas of the USAS lexicon

In order to identify underrepresented areas of the USAS lexicon, this paper makes use of a novel adaption to the corpus linguistic method of keyword analysis (a method that establishes tokens 'whose frequency (or infrequency) in a text or corpus is statistically significant, when compared to the standards set by a reference corpus' (Bondi 2010: 3), which we call K-FLUX analysis (see Prentice et al. 2021). K-FLUX analysis allows one to compare three or more corpora and establish how the frequency of a token, type, or domain fluctuates across the corpora.

The approach is particularly useful in terms of establishing similarities and differences between corpora within one's comparison set. It is being used in this case to compare an updated version of the USAS lexicon described in Supplementary Online Material: Background Study with the content of a corpus of BBC news articles (henceforth BBCNews), the British English 2006 (BE06) corpus (Baker 2009) and the American English 2006 (AME06) corpus (Potts and Baker 2012). Please see Supplementary Online Material, Figure 4, which details our rationale for using these corpora. The document Supplementary Online Material: Background Study provides further details on the corpora and their construction.

3.1. *Method*

The Background Study (see Supplementary Online Material) entailed running the BBCNews, BE06 and AME06 corpora through the software tool Wmatrix (Rayson 2021), where each corpus was tagged for parts of speech and USAS' semantic types. While the Background Study focussed on the unmatched tokens produced by this process, this study focusses on the matched tokens. First, a broad sweep frequency of each corpus was run in Wmatrix's interface. A broad sweep frequency lists all possible semantic tags assigned to each token in each corpus. Using the standard semantic tag frequency lists would not have been appropriate in this case, as the standard lists only contain the primary sense assigned to each token, while this study is interested in all senses assigned to tokens.

Note, however, that there are issues with such an approach, in that broad sweep frequency lists include all possible tags assigned to each token by the tagger. As one moves down the list of possible tags for each token, the tagger's uncertainty increases, thus terms that the tagger views as potential instances of a given type (but which are not in fact instances of said type) are included in the counts. This introduces the possibility of inflated counts in the reference corpora and is an important limitation of the methodology used. Nevertheless, employing arbitrary cutoff points within these lists (i.e. attempting to state at which point the tagger is less certain) would risk losing valid senses assigned to tokens that have many possible interpretations. See Supplementary Online Material, Figure 3 for an illustration and discussion of the tradeoff between precision and recall.

The broad sweep frequency lists were downloaded and counts were conducted on the number of occurrences of each tag in each corpus' broad sweep frequency list. Only top level tags were considered (i.e. X4, but not X4.1, etc.). The counts for each corpus were entered into UCREL's loglikelihood calculation spreadsheet (http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx). The same counts were also conducted on the full USAS lexicon list, and its counts entered into the same sheet, creating four comparison corpora. This UCREL resource allows for a comparison between three or more corpora, and calculates the expected frequency of occurrence of each type, its loglikelihood value, and its BIC (Bayesian Information Criterion) score.

Loglikelihood is used in the field of corpus linguistics to measure token or type differences between corpora. The measure divides a token or type observed frequency by its expected frequency and converts this to a base 2 logarithm (further details of how this is calculated can be found at http://ucrel.lancs.ac.uk/llwizard.html). Loglikelihood is used rather than chi squared due to demonstrations of higher accuracy when working with natural language corpora (Rayson et al. 2004b). However, as Rayson et al. (2004b) point out, both chi squared and loglikelihood are vulnerable to observations in low numbers (i.e. expected frequencies of <5) and at the 99th percentile. Therefore, the authors recommend ensuring that expected frequencies are 11 or more to maintain test accuracy. This is not an obstacle in the current case, as all semantic subtypes occur at expected frequencies greater than 11 in all comparison corpora. In cases where loglikelihood values indicate an observable difference between the corpora on a given token or type, BIC scores have been suggested as a supplementary measure to assist the researcher in assessing the quality of the evidence for an observable difference (see Wilson 2013).

Bayesian Information Criterion (BIC) scores are typically used in statistics to select the best model with which to fit one's data (Neath and Cavanaugh 2012). Within the field of corpus linguistics, this statistic has been adapted to assess the quality of evidence in favour of a difference between one's comparison corpora (Wilson 2013). It is this interpretation that we work with in the present paper. Bayesian Information Criterion (BIC) scores are calculated in

this case by taking the loglikelihood value of a type, subtracting the degrees of freedom (in this case 5), and multiplying the figure by the natural logarithm of the total number of tokens in each corpus, summed. Scores of 2-6 are said to be positive, 6-10 strong and greater than 10 very strong (Wilson 2013).

Having calculated these values for each type, a means of establishing which corpus was responsible for underrepresented or overrepresented types was required. This was crucial in this case, as the aim was to establish which types were underrepresented in the USAS lexicon in particular. Therefore, a K-FLUX analysis was conducted (see Prentice et al. 2021). As indicated at the beginning of Section 3, a K-FLUX analysis is an adaption to the traditional keyword method that entails comparing the observed and expected frequencies for each type in each corpus. If a corpus' observed frequency was greater than its expected frequency for a given semantic type, this was recorded as an instance of overuse and labelled with an 'O'. Conversely, if a corpus' observed frequency was lower than its expected frequency for a given semantic type, this was recorded as an instance of underuse and labelled with a 'U'. In a further development to this method, the actual and percentage difference between observed and expected frequencies were also recorded. All types marked with a U for underuse in the lexicon were extracted and ordered according to their BIC score, from high to low.

Only types with both high loglikelihood values (i.e. 6.63 or above, $p < 0.01$) *and* positive BIC scores were considered, as collectively these suggest (i) types on which marked differences in usage are indicated between the corpora and (ii) types on which we have varying degrees of evidence for the indicated differences. In sum, these are the types that stand out the most when the lexicon is compared with the reference corpora. The full process was repeated with the old USAS lexicon (i.e. before the addition of tokens outlined in Supplementary Online Material: Background Study), to allow for an overview of underrepresented types before and after the addition of new tokens. Finally, similarities and differences between the corpora in terms of their observed overuse or underuse of particular semantic types were examined. This approach presents a swift means of assessing underrepresented and overrepresented areas in a semantic lexicon.

3.2. *Results*

Tables 4a and 4b (see Supplementary Online Material) present the results of the K-FLUX analysis, displaying the underrepresented semantic subtypes in the updated USAS lexicon (U) when compared to one or more of the BBCNews, BE06 and AME06 comparison corpora.

Please note that subtypes listed in Tables 4a and 4b are subtypes of the broader 20 types presented in Supplementary Online Material: Background Study, Table 2 (also see types and subtypes in Table 1 of the same document for further detail). As with broader types, subtypes are drawn (i.e. copied across) from the Longman Lexicon of Contemporary English (McArthur 1981) and were not devised by the paper's authors. It is recognised that, as with all manually devised typologies, there will be some disagreement over the inclusion and labelling of particular types due to the subjective nature with which types were formed. However, this paper concentrates on the application of the typology, rather than on the decision making process behind its original creation.

Results are ordered by BIC score (high to low). Underrepresented subtypes in each corpus are marked with a U and overrepresented subtypes with an O, with size of difference and percentage difference between O and U in brackets. The majority of subtypes are not underrepresented in the lexicon when compared to the remaining corpora. Discounting the subtypes within *Names and Grammatical Words* (i.e. types Z1-Z9), only 26 (or 25%) of 104

top level subtypes are. For the reader's reference, top level subtypes include USAS' A1, A2, etc. types, but not further subtypes, such as A1.1.1, etc.

Tables 4a and 4b separate out underrepresented subtypes into those with arguably more general applicability (i.e. subtypes within the T.Time, N.Numbers and Measurement and A.Abstract and General Tokens semantic types) and those with applicability to more specific areas, such as sociology (i.e. subtypes under the umbrella of S.Social Actions, States and Processes), psychology (i.e. subtypes under the umbrella of X.Psychological Actions, States and Processes), business (i.e. subtypes under the umbrella of I.Money and Commerce), and logistics (i.e. subtypes under the umbrella of M.Movement, Location, Travel, Transport). This process was undertaken manually based on looking at the tokens assigned to each type and making a subjective judgement as to whether or not such tokens might conceivably appear in a wide range of contexts.

When the K-FLUX analysis is repeated on the original USAS lexicon (i.e. the old version, before the additional tokens described in Supplementary Online Material: Background Study were added), the same types are marked as being underused in comparison to one or more of the remaining corpora. The wider domains to which the types are associated, i.e. those of A.General and Abstract Tokens, I.Money and Commerce, M.Movement, Location, Travel Transport, N.Numbers and Measurement, S.Social Actions, States and Processes, T.*Time*, and X.Psychological Actions, States and Processes, remain underrepresented, even once the additional lexicon material has been added in.

3.3. *Discussion*

The subtype splits presented in Tables 4a and 4b suggest that there are a similar number of subtypes with more general applicability and subtypes with applicability to specific areas being underrepresented in the lexicon. One might expect the former to already be sufficiently represented. The reasons why sets that one might think would be more or less complete are showing as underrepresented in the lexicon are twofold. The first reason is the comparison corpora used. The second is the use of 'sweep' lists. As discussed above, these lists capture all possible senses assigned to a token found in the BBC, BE06 or AME06 corpus, including those that the tagger is less certain of. This introduces the possibility of inflated counts in the reference corpora. The results further suggest that using more up to date general English language reference corpora (specifically, those utilised here) to expand the lexicon of the semantic tagger is not sufficient in this case to supplement areas of the lexicon that were already underrepresented.

Subtypes with applicability to specific subjects (such as those listed in Table 4b) offer the opportunity for an alternative approach to lexicon expansion. Rather than using general English language corpora, targeted corpora could be built to supplement their existing vocabulary. The outline and evaluation of such an approach will be the focus of the remainder of this paper. Subtypes with more general applicability (i.e. those listed in Table 4a) may require a different approach, which will not be explored in the present paper, but which represents an avenue for further work and improvement of the lexicon.

Before moving on, it is worth discussing some limitations that are brought to light via the results presented in Tables 4a and 4b. The first is the potential limitations to using a corpus of news texts for the exercise of determining lexicon overuse or underuse. Types such as T1.Time and N1.Numbers, for example, are overrepresented in the BBCNews corpus relative to the lexicon, BE06 and AME06 seemingly due to the report style nature of its content. An investigation of concordances reveals that type T1 includes frequent references to the timing of events (e.g. '23:42 UK time'), reporting on the magnitude of a situation (e.g.

'African women are four times less likely to get the disease'), references to time periods or points in time (e.g. 'the Cypriot president announced in mid March of this year', 'During that time', 'tough/hard times', 'it is time to X'), and references to repeated events (e.g. 'for a second time'). Meanwhile, N1 includes frequent references to the ages of individuals (e.g. 'Kelly, 17, has cared for her mother'), numbers of entities (e.g. 'between three and eight people'), or statistics (e.g. 'unemployment falls by 35,000').

A second point surrounds types with low kappa values, particularly those with no agreement (i.e. types with negative kappa values, such as A4.Classification) and types in the fair to slight agreement range (i.e. 0.01 – 0.40). Broadly speaking, specific types in Table 4b present lower kappa values than the general types presented in Table 4a, with the exception of types such as I4.Industry and S4.Kin. The results suggest that a number of types, particularly specific types, are unclear to participants. Such types require further investigation, as they present a potential obstacle to reliably applying the lexicon within particular domains. To this end, Study 2 will investigate a set of specific types in further detail. One potential solution would be to write descriptors for each type in order to make their boundaries more clearly defined for raters, rather than the current reliance on intuitive assignment based on the steps outlined in Supplementary Online Material: Background Study. This solution will be trialled in Study 2, which follows.

4. Study 2: Expanding underrepresented areas of the USAS lexicon

Having established underrepresented types in the USAS lexicon, an approach was required to supplement these areas. As Study 1 indicated that English language reference corpora, or more specifically the reference corpora featured in the present paper, would not be sufficient for this task, a more targeted approach was needed. In order to illustrate the process, this study considers semantic types in one of the underrepresented areas/domains in the USAS lexicon, namely, X.Psychological Actions, States and Processes.

4.1. *Method*

To begin, the lexicon was searched for lists of lexical tokens assigned to one or more of the underrepresented types within the X.Psychological Actions, States and Processes domain, namely, X1.Psychological actions, states and processes, X2.Mental actions and processes, X4.Mental object, and X6.Deciding. Tokens assigned the X1.Psychological actions, states and processes tag included *hallucinate*, *depression*, and *agoraphobia*. Tokens given the X2.Mental actions and processes tag included *association*, *cognition* and *hypnotism*. Tokens with a X4.Mental objects tag included *ideology*, *hypothesis* and *stereotype*, while tokens with a X6.Deciding tag included *unresolved*, *predetermine*, and *indecisiveness*.

The tokens within these categories generally refer to psychological or psychiatric conditions, treatments and symptoms. Given the nature of the tokens contained within the selected underrepresented types, three psychiatric and psychological reference corpora were obtained. The first of these is an existing corpus of 1,000 psychiatric association study abstracts, which was used as a reference corpus from which to draw a set of suitable seed tokens. The corpus consists of 500 genetic psychiatric association abstracts and 500 parallel general psychiatric association abstracts drawn from PubMed. The corpus amounts to 20,592 tokens. Abstracts in both the genetic and general sub corpora cover the same date range (from 1st July 2016 and 30th December 2018). For the purpose of this study, this corpus shall be referred to as the PsychStudy corpus.

The second psychiatric and psychological reference corpus was obtained from MTSamples (2020), which contains a collection of transcribed, open source, medical

transcriptions and medical reports. All transcribed reports contained in the 'Psychiatry/Psychology' category were downloaded and combined into one corpus containing 53 reports and 52,470 tokens. This will be referred to as the MTPsych corpus. The third psychiatric and psychological reference corpus was created using the Medical Web Corpus, which is available via SketchEngine (Lexical Computing 2020). The corpus consists of 42,054,011 tokens and 526 documents collected from the internet, which have a focus on the medical domain (for further details, see Kilgarriff et al. 2010). This corpus was searched for the query 'psych.*', which produced a set of concordance lines containing the query. These concordances were randomised and the first 10,000 concordances downloaded using the SketchEngine interface. This resulted in a corpus of 379,789 words, which will be referred to as the PsychMed corpus.

The PsychStudy, MTSamples and PsychMed corpora were uploaded to Wmatrix, where they were run through the CLAWS part of speech tagger and the USAS semantic tagger. As with the BBCNews, BE06 and AME06 corpora utilised in Study 1, a broad sweep frequency list was downloaded for each corpus, which contained all possible senses assigned to each token in each corpus. Tokens assigned to any of the four underrepresented subtypes were extracted and ordered in frequency from high to low, with four frequency lists assembled (one for each subtype).

The top five most frequently occurring tokens annotated for each subtype in each corpus were compared. Tokens appearing in at least two of the three psychiatric and psychological reference corpora were used to create a set of 'seed' tokens. The selected tokens were then used to form the following queries, which were searched in PubMed to create specialised corpora centred around particular subtypes. For the reader's reference, queries are formatted according to PubMed conventions. Tokens in square brackets represent searches for tokens within titles or abstracts of papers.

(1) X1.Psychological actions, states and processes: ((((psychiatr*[Title/Abstract]) OR psycholog*[Title/Abstract]) OR psychos*[Title/Abstract]) OR psychotic*[Title/Abstract]) OR bipolar[Title/Abstract]
(2) X2.Mental actions and processes: (((mental*[Title/Abstract]) OR study*[Title/Abstract]) OR studi*[Title/Abstract]
(3) X4.Mental objects: (((treatment*[Title/Abstract]) OR problem*[Title/Abstract]) OR diagnos*[Title/Abstract]
(4) X6.Deciding: (((((((mak*[Title/Abstract]) OR made[Title/Abstract]) OR find*[Title/Abstract]) OR found[Title/Abstract]) consider*[Title/Abstract]) estimat*[Title/Abstract]) conclu*[Title/Abstract]

As with previous lexicon expansion studies, this approach was based on the rationale that assembling corpora around tokens in underrepresented types might assist in drawing out synonymous tokens that do not yet exist in the USAS lexicon. Each search was filtered to include the first 200 hits, sorted according to best match results. The abstracts associated with these results were then downloaded and saved to text format before being uploaded to Wmatrix for part of speech and semantic tagging.

As with the psychiatry and psychology reference corpora, the seed corpora were run through part of speech and USAS semantic tagging in Wmatrix. For all three reference corpora and the seed corpora, a list of domain specific unmatched tokens was downloaded. Tokens were then assigned to appropriate semantic subtypes using a combination of the methods outlined in Supplementary Online Material: Background Study, Section 3.2 and a

description of each domain specific subtype. Specifically, the subtype X1.Psychological Actions, States and Processes captures tokens broadly relating to human psychology. The subtype X4.Mental Objects collectively refers to conceptual objects that have been devised by humans to assist their understanding or the understanding of others. X6.Deciding refers to mental processes involved in decision making, while X2.Mental Actions and Processes refers to brain based lexis and processes of thought.

The process outlined in Study 1 was then repeated for each of the reference corpora and the seed corpora. This entailed conducting counts of tokens assigned to each of the four previously underrepresented X.*Psychological Actions, States and Processes* subtypes and entering updated figures into the loglikelihood calculation spreadsheet, alongside the existing figures for the BBCNews, BE06 and AME06 corpora. A new K-FLUX analysis (described in detail in Study 1) was conducted on the updated figures to ascertain whether any of the corpora had been successful in sufficiently representing the X.Psychological Actions, States and Processes domain.

This being the case, one would expect to observe that its previously underrepresented subtypes of X1.Psychological actions, states and processes*,* X4.Mental object, X6.Deciding and X2.Mental Actions and Processes would now show as being overused in the USAS lexicon. However, note that the keyness method used favours difference. Therefore, a previously underused subtype in the lexicon may become overused relatively easily if the difference between its observed and expected frequencies was relatively small to begin with. For this reason, the actual and percentage difference between observed and expected frequencies was recorded. In addition, the percentage increase on previously recorded observed frequencies was established.

Finally, a member of the project team with expertise in psychiatric studies (a naïve annotator, who had not used the system before) was shown a randomly selected 10% sample of the newly annotated tokens produced by each corpus and asked to select which of the four semantic subtypes (X1, X2, X4, X6, or none) they would primarily assign each token to. These annotations were compared to the annotations given by the initial rater (a linguist from the team with prior USAS experience and knowledge of psychology). Fleiss kappa values were computed for each of the four semantic subtypes in each corpus to ascertain the degree to which individuals with knowledge of this domain agreed that a) particular tokens belonged to particular psychological types and b) which corpus or corpora produced tokens that raters most agreed on.

To further assess the value of tokens added by each corpus, a list of 14,404 tokens from the APA dictionary (American Psychological Association 2020) was extracted (please note the analysis excludes MWEs in this dictionary). The APA dictionary list is assumed to represent useful key tokens within the domain of psychology, which one working within this discipline might expect to find in the USAS lexicon. The existing lexicon and its addition described in Supplementary Online Material: Background Study were first compared to the APA list to establish the lexicon's existing coverage of this list. The unmatched items from each of the psychology corpora were compared with the APA list in turn to ascertain what (if any) additional coverage each corpus offered. As the APA potentially contains tokens that have low frequency in general use, the frequencies of APA tokens found and not found in the extended lexicon and the featured psychology corpora were looked up in the wordlists of the BNC and a 20,639,864 token corpus of 30 psychology textbooks (provided by Xodabande 2020).

In addition to the above, the word list from the psychology textbook corpus was used to derive a set of core vocabulary items in the psychology domain. The word list from this

corpus was organised in order of frequency and a cumulative frequency from the addition of each token in the list was calculated. Tokens representing 90% of the content of the corpus were extracted to create the core vocabulary list. This list was then compared to the lexicon, the additional lexicon items, and each of the psychology reference corpora in turn to provide an idea of the extent of coverage of core psychology tokens present in general use.

4.2. *Results*

The process of running the seed corpora and domain reference corpora through part of speech and USAS semantic type tagging is summarized in Table 5 (see Supplementary Online Material), which provides the number of unmatched tokens in each corpus and the number of those tokens subsequently assigned to each of the four subtypes: X1.Psychological Actions, States and Processes, X2.Mental Actions and Processes, X4.Mental Objects and X6.Deciding. Table 6 (see Supplementary Online Material) shows the results of adding the tokens from each corpus to the K-FLUX analysis, coupled with interrater agreement scores (Fleiss kappa values) for each semantic subtype in each corpus.

Of the 10% sample tokens assessed (109 in total), 34 were compounds and 46 were morphological variants. Please note, these are not necessarily variants or compounds of existing tokens in the lexicon. Adding the base forms of variants and any associated forms will extend lexicon additions further.

The results of comparing the existing lexicon and the new lexicon tokens (both described in Supplementary Online Material: Background Study), and the various psychological corpora with the APA dictionary list and the core vocabulary of the Psychology Textbook Corpus are presented in Table 7 (see Supplementary Online Material). The second and fourth columns present the number of overlapping tokens provided by the lexicon and the unmatched lists from each corpus (and the percentage of each list that overlap with the APA dictionary or core Psychology Textbook vocabulary). The third and fifth columns add the number of matching tokens from each corpus in turn to the number of matching tokens from the lexicon to show the extent to which coverage changes with the addition of tokens from each corpus.

The frequency thresholds of APA tokens matched and not matched in the expanded lexicon and psychology corpora in both general (BNC) and domain specific (Psychology Textbook) language use are presented in Table 8 (see Supplementary Online Material). Of the 7,482 APA tokens not found in the collected Psych corpora or the expanded lexicon, the BNC matches 2,688 tokens (35.93%) and the Psychology Textbook Corpus matches 2,701 tokens (36.10%). Of the 1,742 APA tokens added by the Psych corpora, the BNC matches 1,257 tokens (72.16%) and the Psychology Textbook Corpus matches 1,311 tokens (75.26%). Finally, of the 5,180 APA tokens found in the expanded lexicon, the BNC matches 5,035 (97.20%) and the Psychology Textbook Corpus matches 4,913 (94.85%).

4.3. *Discussion*

The results presented in Table 6 (Supplementary Online Material) show that 70 of the 109 tokens assessed (64.22%) receive complete agreement between the two consulted raters. The results further suggest that higher rates of interrater reliability are received by subtype X4.Mental Object (with kappa values in the region of 0.64 to 1.00), indicating substantial to near perfect agreement, and subtype X1.Psychological Actions and Processes (with kappa values in the region of 0.51 to 0.70), indicating moderate to substantial agreement. The results for subtype X2.Mental Actions and Processes are more variable, and for X6.Deciding, results typically indicate no agreement, with the exception of result for the PsychStudy corpus. However, the latter result is based on just one token and therefore is not particularly

informative. The findings suggest that raters have difficulty in discerning these subtypes and therefore that such subtypes in the coding scheme need to be revisited.

The results further indicate that care should be taken when interpreting the output of the K-FLUX approach. While some subtypes remain underused (U) with additions from the varying corpora and some now indicate overuse (O), the percentage difference between observed and expected frequencies is often small. One might confidently argue that the subtype X6.Deciding is still underused in the lexicon, despite focused additions, given that the differences in observed and expected frequencies across the board lie in the range of ~60-70%. One might have further confidence in the overuse indicated by the PsychMed corpus at 48.64%. However, the underuse and overuse cases for the remaining corpora and subtypes are less compelling. Size of difference should therefore be borne in mind when making decisions on the basis of the approach.

In terms of the tokens contributed by each of the sources, perhaps unsurprisingly, the PsychMed corpus, the largest of the corpora used, accounts for the greatest percentage increases across subtypes X1, X2 and X4. However, slightly higher kappa values are reflected in the items assessed from other corpora in relation to subtypes X2 and X4. While the Seed corpus contributes the greatest percentage increase for subtype X6, additions from this category across the board are not generally agreed upon by raters. The findings indicate the importance of assessing lexicon additions. A greater number of additional tokens does not necessarily equate to a greater number of quality lexicon additions.

The results of the APA dictionary list coverage (see Supplementary Online Material – Table 7) show that, in terms of raw numbers, the psychologically orientated corpora provide more additional coverage of the APA dictionary than the general approach to lexicon expansion detailed in Supplementary Online Material: Background Study (depicted by 'new lexicon items'). However, when considered as a percentage, the number of APA matching tokens relative to the size of the full list of new tokens returned by the seed and PsychMed corpora is not vastly different to the percentage returned by the new lexicon items. This may be because the majority of seed and PsychMed APA matching tokens are already present in the lexicon. The percentage of PsychStudy and MTPsych unmatched items that overlap with the APA dictionary is twice to three times that of the seed and PsychMed corpora.

Interestingly, neither the new lexicon items nor the psychologically orientated corpora add significantly to the lexicon's coverage of the APA list. There are a number of potential reasons for this. The corpora represent language in use, rather than tokens deemed by experts to be worthy of definition. Some tokens within the APA dictionary may refer to rare phenomena that are not picked up by the corpora. In addition, the corpora represent current usage, meaning outdated or old tokens featured in the APA dictionary will not be picked up in modern usage.

Nevertheless, this paper is primarily concerned with coverage of core or common vocabulary in frequent use within the psychological domain. As the results in Table 8 (Supplementary Online Material) demonstrate, the extended lexicon covers the most frequently used APA tokens both in general and domain specific use. The psychology corpora add APA tokens that tend to occur between 10 and 100 times, while those APA tokens that remain unmatched tend to be in the lowest frequency range of general and domain specific use (i.e. less than 10 occurrences). In addition, as shown by the results of the coverage of core psychology textbook vocabulary in Table 8, the lexicon and its additions result in coverage of more than 84% of core vocabulary within the target domain.

5. General discussion and conclusions

Collectively, the studies presented in this paper have sought to put forward a domain based method to the identification and supplementation of underrepresented areas in a semantic lexicon. It is worth pointing out that one notable limitation of the current work is that it considers only single tokens and not multi token expressions. Future work will be required to extend the lexicon's multi token expression list in a similar fashion.

The paper found that, while general English language reference corpora are suitable to compare with one's lexicon to identify underrepresented areas, they do not necessarily provide one with a means of successfully supplementing all domains. While one might argue that larger reference corpora would combat this issue, a counter argument to this is that there are particular domains that will tend to feature less in everyday language use regardless of the size of one's general language sample.

Further, it would appear that general representative corpora that contain not only an array of domains, but also a variety of text types, would be best suited as a point of comparison with one's lexicon to identify areas of weakness. Despite being around twice the size of the BE06 and AME06 corpora, the BBCNews corpus, which contained only one text type (i.e. news articles) was found to cause the underrepresentation of particular types in both the lexicon and structured English language corpora, due to its large overrepresentation of those types. Indeed, the argument for including a variety of text types to produce a well rounded reference corpus informed the original creation of the BROWN corpus of American English (see Francis and Kučera 1964).

When expanding at the level of a particular domain, however, size had its more obvious advantages, with the largest psychologically orientated corpus producing the greatest number of lexicon additions, as one might expect, and moving three out of four previously underused subtypes into overuse. Nevertheless, it is also worth mentioning that this was the only psychologically orientated corpus used to feature a variety of text types, with all others featuring only one form of text (e.g. research articles or case notes).

The proposed K-FLUX method provided a means of identifying areas in need of supplementation within the USAS lexicon. However, three points require consideration when utilizing this approach: (i) small changes in frequency can move a type from underuse to overuse, or vice versa, and therefore, one must consider the size of the difference in one's evaluations, (ii) deriving a percentage increase on observed frequencies before and after supplementation helps one to establish whether one's source is producing marginal or considerable gains, and (iii) as discussed above, the nature of the comparison corpora has a bearing on one's results and therefore should be carefully considered. Taking this paper's results as an example, while one might reasonably conclude that additions from the PsychMed corpus have provided reasonable improvements to the subtype X1.Psychological Actions and Processes, all further subtypes require further work, as any gains made appear marginal and therefore less reliable.

This leads one to the subject of the subtypes themselves. Indeed, in their work on a comparable domain based hierarchy, Bentivogli et al. (2004: 94) observe that the first incarnation of the WordNet Domains Hierarchy (or WDH) had problems due, in part, to 'the lack of clear semantics of domain labels'. The results from the studies presented in the current paper suggest that certain subtypes are more discernible than others, as indicated by high versus low inter annotator agreement values. Those subtypes with low values require further investigation and revision, for example, incorporation into existing categories. For this, a larger scale evaluation exercise is required, taking larger samples and subjecting them to a wider range of human judgements. The human judgement matrix (described in Supplementary

Online Material: Background Study and presented in Supplementary Online Material: Figure 2) could be used as a guide to which subtypes are typically confused and therefore may logically be combined.

Evaluation and development work could be further informed by Bentivogli et al.'s (2004) aforementioned revisions to the WordNet Domains Hierarchy (WDH), in which the authors mapped WDH subtypes onto the Dewey Decimal Classification (DDC) system and used this as a gold standard to ensure that their subtypes had explicit, unambiguous semantics, that types did not overlap with one another, that types represented all human knowledge and that types had comparable degrees of granularity.

Other future work should include the expansion of other underrepresented areas of the lexicon, using the lessons learnt from the present paper as a guide to best practice. In addition, while human judgements have been investigated in the current study, future work might consider the automated judgements of the semantic tagger itself. While the number of times the tagger erroneously assigns its first sense has been assessed against a corpus of 124,900 words, producing an error rate of 8.95% (Rayson et al. 2004a), recreating the supplementary matrix used for human judgements could provide further useful insights into the discernibility of subtypes from a machine based perspective.

Finally, and for the first time, along with the lexicon additions detailed in this paper and its supplementary material, we are releasing the full English semantic lexicon for academic use under a CC-BY-SA-NC licence, from https://github.com/UCREL/Multilingual-USAS.

6. Acknowledgements

The authors would like to thank Ismail Xodabande of Kharazmi University for supplying the wordlist of psychology textbooks used in Study 2 of this paper. This research was funded, in whole or in part, by the Wellcome Trust, 204475/Z/16/Z. A CC BY or equivalent licence is applied to the AAM arising from this submission, in accordance with the grant's open access conditions.

References

A. Dictionaries

American Psychological Association. 2020. *APA Dictionary of Psychology*. Washington, DC: American Psychological Association (APA). Accessed on 25 June 2021. https://dictionary.apa.org/

Downes, J. and J. E. Goodman. 2014. *Dictionary of Finance and Investment Words*. Hauppauge, NY: Barron's Educational Series.

Longman. 2021. 'Domain'. *Longman Dictionary of Contemporary English Online*. London: Pearson. Accessed on 28 November 2019. https://www.ldoceonline.com/dictionary/domain.

Oxford University Press. 2021. 'Sense'. *Lexico.com*. Oxford: Lexico.com. Accessed on 28 September 2020. https://www.lexico.com/en/definition/sense.

McArthur, T. 1981. *Longman Lexicon of Contemporary English*. Harlow: Longman.

B. Other literature

Baker, P. 2009. 'The BE06 Corpus of British English and Recent Language Change.' *International Journal of Corpus Linguistics* 14.3: 312–337.

Bentivogli, L., P. Forner, B. Magnini and E. Pianta. 2004. 'Revising the WordNet Domains Hierarchy: Semantics, Coverage and Balancing' In Sérasset, G., S. Armstrong, C. Boitet, A. Popescu-Belis and D. Tufis (eds), *Proceedings of the Workshop on Multilingual Linguistic Resources MLR2004*. Geneva: COLING, 94-101.

Bondi, M. 2010. 'Perspectives on Keywords and Keyness' In Bondi, M. and M. Scott (eds), *Keyness in Texts*. Amsterdam: John Benjamins, 1-18.

Cao, G., J-Y. Nie, J. Gao and S. Robertson. 2008. 'Selecting Good Expansion Terms for Pseudo-Relevance Feedback' In Chua, T-S., M-K. Leong, S. H. Myaeng, D. W. Oard, and F. Sebastiani (eds), *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, July 20-24, 2008.* New York, NY: Association for Computing Machinery, 243-250.

Da Silva, A. L. and R. Dennick. 2010. 'Corpus Analysis of Problem-Based Learning Transcripts: An Exploratory Study.' *Medical Education* 44: 280-288.

Francis, W. N. and H. Kučera. 1964. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, RI: Department of Linguistics, Brown University.

Granger, S. and M. Paquot. 2010. 'The Louvain EAP Dictionary' In Dykstra, A. and T. Schoonheim (eds), *Proceedings of the XIV EURALEX International Congress, Leeuwarden*, *Netherlands, July 6-10, 2010*. Ljouwert: Fryske Akademy, 321-326.

Kheovichai, B. 2015. 'Metaphorical Scenarios in Business Science Discourse.' *Iberica* 29: 155-178.

Kilgarriff, A., S. Reddy, J. Pomikálek, and P. V. S. Avinesh. 2010. 'A Corpus Factory for Many Languages' In Calzolari, N., K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds), *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10, May 19-21, 2010*. Valletta: European Language Resources Association (ELRA), 904-910.

Lexical Computing. (2020). *SketchEngine.* Accessed on 14 July 2020. https://www.sketchengine.eu

Löfberg, L. 2017. 'Creating large semantic lexical resources for the Finnish language', PhD, Lancaster University. https://doi.org/10.17635/lancaster/thesis/3

Makki, R., S. Brookes and E. E. Milios. 2014. 'Context-Specific Sentiment Lexicon Expansion via Minimal User Interaction' In Laramee, B., A. Kerren and J. Braz (eds), *Proceedings of the International Conference on Information Visualization Theory and Applications, IVAPP 2014, Lisbon, Portugal, January 5-8, 2014*. Setúbal: SciTePress, 178-186.

Markowitz, D. M. and J. T. Hancock. 2014. 'Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel.' *PLoS ONE* 9.8: e105937.

MTSamples. 2020. *Transcribed Medical Transcription Sample Reports and Examples*. Accessed on 15 July 2020. https://www.mtsamples.com

Neath, A. A. and J. E. Cavanaugh. 2012. 'The Bayesian Information Criterion: Background, Derivation and Applications.' *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2: 199-203.

Olteanu, A., C. Castillo, F. Diaz and S. Vieweg. 2014. 'CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises' In Adar, E. and P. Resnick (eds), *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, ICWSM-14, Michigan, USA, June 1-4, 2014*. Palo Alto, CA: AAAI Press, 376-385.

Pantel, P., E. Crestan, A. Borkovsky, A. M Popescu and V. Vyas. 2009. 'Web-Scale Distributional Similarity and Entity Set Expansion' In Koehn, P. and R. Mihalcea

(eds), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, Singapore, August 6-7, 2009, Vol. 2*. Stroudsburg, PA: Association for Computational Linguistics, 938-947.

Potts, A. and P. Baker. 2012. 'Does Semantic Tagging Identify Cultural Change in British and American English?' *International Journal of Corpus Linguistics* 17.3: 295–324.

Prentice, S., J. Knight, P. Rayson, M. El Haj and N. Rutherford. 2021. 'Problematising Characteristicness: A Biomedical Association Case Study.' *International Journal of Corpus Linguistics* 26.3: 305-335.

Rao, Y., J. Lei, L. Wenyin, Q. Li and M. Chen. 2014. 'Building Emotional Dictionary for Sentiment Analysis of Online News.' *World Wide Web* 17.4: 723-742.

Rayson, P. 2021. *Wmatrix: A Web-Based Corpus Processing Environment*. Lancaster: Lancaster University.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. 2004a. 'The UCREL Semantic Analysis System' In Guthrie, L., R. Basili , E. Hajicova and F. Jelinek (eds), *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in Association with 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, May 25, 2004*. Lisbon: LREC, 7-12.

Rayson, P., D. Berridge and B. Francis. 2004b. 'Extending the Cochran Rule for the Comparison of Word Frequencies Between Corpora' In Fairon, C., A. Dister, G. Purnelle and J. Denooz (eds), *7th International Conference on Statistical Analysis of Textual Data, JADT 2004, Louvain, Belgium, March 10-12, 2004, Vol. 2*. Louvain: Lexicometrica, 926-936.

Vania, C., M. Ibrahim and M. Adriani. 2014. 'Sentiment Lexicon Generation for an Under-Resourced Language.' *International Journal of Computational Linguistics and Applications* 5.1: 59-72.

Wilson, A. 2013. 'Embracing Bayes Factors for Key Item Analysis in Corpus Linguistics.' In Bieswanger, M. and A. Koll-Stobbe (eds), *New Approaches to the Study of Linguistic Variability*. Frankfurt: Peter Lang, 3-11.

Xodabande, I. 2020. *Investigating the Vocabulary Load of Psychology Textbooks: A Corpus-Based Study*. Working Paper. Accessed on 14 July 2021. http://www.cambridge.org/engage/coe/article-details/5eb1cd112f52fa0019c41968