M.A. Radmila Palinkašević[1]
Preschool Teacher Training College "Mihailo Palov"
Vršac

Original scientific paper
UDK: 37.025
DOI: 10.17810/2015.61

==========

SPECIALIZED WORD LISTS – SURVEY OF THE LITERATURE – RESEARCH PERSPECTIVE

Abstract: Word lists present an essential tool in vocabulary teaching. Compilation of specific word lists for various fields is one of the most prominent branches of research in this field at the moment. New methodological changes in word list formation have been proposed because of the appearance of the New-GSL (Brezina & Gablasova, 2013) and AVL (Gardner & Davies, 2013). The aim of this paper is twofold. The first one is to present a word list overview which would reflect how these changes have affected the field so far. The second is to serve as a methodological guide for future researchers interested in specific word list formation. The paper provides an overview of the two most significant word lists the GSL (West, 1953) and AWL (Coxhead, 2000) along with their proposed replacements, detailed information about seven specific word lists, which were published from 2013 to 2016, with a specific focus on corpus formation criteria, word selection criteria and validity and relevance testing.

Key words: specific word lists, methodology, research overwiew, New-GSL, AVL, AWL.

Introduction

The large vocabulary of the English language is one of the most serious challenges which English language learners face (Lessard-Clouston, 2013). It is still impossible to precisely count the number of words in a language. However, to illustrate the huge amount of vocabulary which learners face, we will present the fact that the Webster's 3rd dictionary has a vocabulary of around 54,000 word families (Nation & Waring, 1997). Word lists help learners and teachers orient themselves in the sea of English language vocabulary. Word lists present the most frequently used and significant vocabulary in the language or a specific genre or scientific area. They are useful for establishing vocabulary learning goals, assessing vocabulary knowledge and growth, analyzing text difficulty and richness, creating and modifying reading materials, designing vocabulary learning tools, determining the vocabulary components of academic curricula etc. (Gardner & Davies, 2013).

The most famous general word list is the General Service List GSL (West, 1953) which contains 2000 of the most frequent words in the English language (Bell, 2001). The first 1000 words in the General Word List together with proper nouns cover 78%-81% of written texts, and around 85% of spoken text (Nation & Deweerdt 2001). Although this list is quite old it is still widely

[1] palinkasevic@gmail.com

used (Coxhead 2011, Gilner2011, Wolfe 2015). With the GSL a large amount of texts is covered using a relatively small number of vocabulary, but it is not enough to understand an average text. In order to be able to understand a text 98% of vocabulary needs to be understood (Nation, 2006). For this reason the need for specialized word lists arose. In the 1970-s pioneering scholars in the area of vocabulary formed a number of academic vocabulary word lists (Gardner & Davies, 2013). In an attempt to create a more comprehensive word list Nation combined the results of the previous studies and his own to form the University Word List (UWL - Xue and Nation, 1984). It uses the GSL as its foundation, since the acquisition of the most significant academic vocabulary was a logical next step for language learners who had mastered the GSL and needed English in educational settings. UWL presents 836 word families which are not in the GSL but which are important for academic texts. The UWL covers approximately 8,5% of academic texts. This list was replaced by the Academic Word List (Coxhead, 2000) which contains 570 word families and covers approximately 10% of academic texts. The release of the AWL had a significant impact on the field of English for Academic Purposes but also on the field of English for Specific purposes and vocabulary teaching in general. Two new research waves were started in reaction to the AWL. The first one consists of research studies which test the AWL's applicability in numerous academic corpora, which usually present a specific scientific area (Hyland &Tse 2007; Chen &Ge2007; Konstantakis 2007; Coxhead & Hirsch 2007; Martinez et al. 2009; Vongpumivitch et al. 2009; Li &Qian 2010; Yazhen& Lei 2013; Dang& Webb 2014; Liqin&Xinlu2014; Mozaffari & Moini 2014). The second wave consists of studies which aim to form their own specific word lists, usually in specific academic genres and areas (Coxhead & Hirsch 2007; Lessard-Clouston 2013; Surtees & Horst 2013; Wolfe 2015).

In 2013 the New General Service List (New-GSL, Brezina&Gablasova, 2013) and Academic Vocabulary List (AVL, Gardner & Davies, 2013) were formed to replace the GSL and AWL. These two new lists brought with them changes in word list formation methodology, which are just starting to affect the field. The aim of this study is to present an overview of specific word lists with an accent on word list formation methodology. The focus will be on the specific word lists which were formed in the last three years. The reason for such focus is twofold. Firstly, in order to monitor the changes in the field because the new methodologies were proposed three years ago. Secondly, because an excellent overview of significant word lists was published in 2013 by Lassard-Clouston. Certain standards and important elements which should be implemented for high standards in this type of research will be presented. The target audience of this paper are researchers interested in forming their own specific word lists and those interested in word lists in general.

Theoretical background

In the formation and usage of word lists two stand out as the most influential the GSL (West, 1953) and AWL (Coxhead, 2000). Their influence is reflected in the fact that they have been used as word list formation methodology models (e.g.Minshall2013;Wolfe2015; Lei & Liu 2016) and as a word selection starting point in many word list formation studies (e.g. Coxhead & Hirsch 2007; Chung 2009; Ng et al. 2013, Yang 2015). Recently new word lists, with modernized methodologies have been proposed to replace them –the New-GSL (Brezina & Gablasova, 2013) and AVL (Gardner & Davies, 2013). In order to fully grasp the changes that are taking place in this specific field of vocabulary acquisition, the first step will be to discuss the methodology and characteristics of these four word lists.

The General Service List of English Words (West, 1953) consists of 1,907 main entries and 3,751 orthographically different words (common derivatives and compounds) (Gilner, 2011). Although it was published in 1953 it actually represents a revised version of the *Interim Report on Vocabulary Selection* from 1936 (Brezina & Gablasova, 2013). The goal of the GSL is the selection of a core vocabulary of general application in foreign language instruction (Faucett et al. 1936 as cited in Gilner, 2011). The GSL was formed using a 5 million word corpus. The most relevant characteristics of the GSL are: frequency, universality (words used in many countries), utility (words used to talk about various topics) and usefulness (words that can be used to define or describe other words) (Gilner, 2011).West used both qualitative and quantitative selection criteria. The qualitative criteria were:

1. Ease of learning – words selected on the basis of the similarity of word forms, even if they do not meet the frequency standard;
2. Necessity and cover – the idea here is to include all of the words needed to cover the most important ideas/concepts with few redundancies (e.g. the word *to preserve* (food) was included despite its relatively low frequency because it presented an important concept);
3. Stylistic and emotional neutrality –neutral expression of ideas is the main language function, therefore some stylistically marked high frequency words were excluded (Brezina&Gablasova, 2013).

When the GSL came out it was innovative and groundbreaking. Its value is also clearly confirmed by the fact that it was the most influential general word list for more than 50 years.

Over the years the GSL was compared to other general word lists which had been formed using more modern corpora, however, the coverage differences were never large enough to merit a substitution of the lists (Gilner, 2011). For example Gilner and Morales (2008 as cited in Gilner, 2011) compared the coverage of the first two BNC frequency bands with the GSL in the English Language Teaching Corpus of 1,157,493 words. The coverage of the first and second BNC bands was 80.43% and 7.65% respectively, while the coverage of the first and second 1000 words of the GSL was 80.02% and 6.71%.

The main reason why general word lists resembled each other was that the main criteria for word selection remained the same: frequency rank, range measure, word family structure, token coverage and corpus choice (Gilner, 2011). The reason why the new-GSL is considered as a possible substitute is that the methodology and list presentation form have been slightly changed. These changes resulted in a general service word list which could be more practical for language learners and teachers.

The New-GSL (Brezina&Gablasova, 2013) consists of a total of 2,494 lemmas. It can be divided into the base part consisting of 2,116 lemmas and the current vocabulary part consisting of 378 lemmas. The goal of the list is the same as the GSL, namely, to provide the core high-frequency vocabulary which will aid beginner English language learners. The first main difference between the GSL and New-GSL is the development corpus. The corpus used for the development of the new-GSL consists of four different corpora LOB, BNC, BE06 and EnTenTen12 which together amount to more than 12 billion running words. The LOB and BE06 are both relatively small corpora consisting of 1 million words each. However, they were used because they had been built using the same criteria, but in different time periods. Both consist of carefully selected texts from 15 genres of writing, but LOB was developed in 1961 while BE06 contains texts from the period of 2005 to 2007. The formation time difference also

exists between the BNC corpus which was formed in the 1990s and the EnTenTen12 which was constructed in 2012. EnTenTen12 is the largest of the four corpora, consists of 12 billion words compiled through web-crawling and cleaning of raw data available online. The New-GSL was formed through the following steps:

1. Word lists were created for each of the four corpora – *Sketch Engine* was usedto form lemmatized word lists which included information about word classes. The word selection criteria was the average reduced frequency – ARF. It takes into account both the absolute frequency and the distribution of the lexical item in the corpus. A 3000 lemma word list was compiled for each of the 4 corpora;
2. The four word lists were compared pairwise – the overlap between all of the word lists is high within 78% to 84%. These results indicate that a strong and stable core of common vocabulary exists;
3. The identification of the common lexical core among the 4 wordlists – 2,116lemmas;
4. Identification of lexical items from the BE06 and EnTenTen12 corpora which represent recent vocabulary changes – 378 lemmas which do not appear in LOB and BNC but which appear in both BE06 and EnTenTen12 were identified;
5. The formation of the New-GSL.

After the New-GSL had been formed it was compared to the GSL and AWL. The three lists overlap to a large extent and only 178 lemmas from the New-GSL do not appear in either GSL not AWL.One of the most important distinctions between the GSL and New-GSL is the fact that the former consists of word families and the latter of lemmas. Since word families are more extensive the number of lemmas which form the GSL were calculated to enable objective comparison. The GSL consists of about 4100 lemmas while the New-GSL consists of 2,494 lemmas. The coverage of the GSL and new-GSL was tested on all four of the corpora used in this study. The GSL covered 84.1% of LOB, 82% of BNC, 80.6% of BE06 and 80.1% of EnTenTen12 while the New-GSL covered 81.7% of LOB, 80.3% of BNC, 80.1% of BE06 and 80.4% of EnTenTen12. From these results it can be seen that the lists achieve very similar coverage but the learning work load of the New-GSL is significantly lower.

The AWL (Coxhead, 2000) consists of 570 word families. The target audience of this list were first year university students who needed English for their studies. The goal of the list was to provide these learners with a word list which would enable them to understand first year teaching materials with a manageable vocabulary learning load. The corpus used for the development of the list consisted of 3.5 million running words and covered 28 subject areas.It was divided into four discipline areas: arts, commerce, law and science.The corpus contained 414 texts balanced for length and taken from textbooks, articles, book chapters and laboratory manuals, which were used in the first year of university study. The criteria for word selection were:

1. The GSL words were excluded from the list – this will later be also known as the specialized occurrence criteria (Minshall, 2013);
2. Frequency – the word had to occur 100 times or more in the corpus;
3. Range – the word had to occur in 15 or more of the subject areas;
4. Uniformity – the word had to occur over 10 times in the four disciplines;

The 570 word families selected in this way were divided into ten sub-lists divided according to frequency. The AWL covers about 10% of academic texts in various academic disciplines (Coxhead, 2011). It has been used for the construction of numerous EAP textbooks and teaching materials.

The Academic Vocabulary List (AVL Gardner & Davies, 2013) was designed with a purpose to replace the AWL. Two main reasons were given why the AWL needed to be replaced:

1. Word families used for initial AWL counts
2. Relationship of AWL and GSL – AWL is a layered list which uses GSL as a starting point. GSL is firstly criticized for its age. Secondly AWL is criticized for containing words which are listed in the high-frequency BNC lists. On the basis of the fact that a total of 451 word families are found in the first 4000 most frequent word families of BNC, the authors state that AWL is merely a subset of the high frequency words of English. They state that GSL is no longer an accurate reflection of high-frequency English. However, the authors did not provide any studies or data to confirm this claim.

In order to avoid the aforementioned shortcomings of the AWL a new word list formation methodology was proposed and implemented in the formation of the AVL.

The list was created using a 120 million academic word corpus, which was taken from a 425 million word Corpus of Contemporary American English - COCA. The corpus consists of nine disciplines (Education, Humanities, History, Social science, Philosophy, religion and psychology, Law and political science, Science and technology, Medicine and health, Business and finance). It consists of academic journals (85 million), academically oriented magazines (31.5 million) and finance sections of newspapers (7.5 million). The corpus is tagged for grammatical parts of speech by the CLAWS 7 tagger from Lanchester University. Four criteria were used for the selection of lemmas which form the AVL:

1. Ratio – the frequency of the lemma must be at least 50% higher in the academic corpus than in the non academicprotion of the COCA. Therefore the 1.5 Ratio was selected.
2. Range – the lemma must occur at least 20% of the expected frequency in at least seven of the nine academic disciplines;
3. Dispersion – lemmas must have a dispersion of at least 0.80. This measure was developed by Julliard & Chang-Rodriguez (1964) and it shows how evenly a word is spread across the corpus – 1.0 means that the word is perfectly evenly spread across the corpus.
4. Discipline Measure – the word cannot occur more than three times in the expected frequency in any of the nine disciplines.

The completed AVL list consists of 3000 lemmas or 2000 word families. Some of the advantages of the list are the following:

- The families are ordered by frequency;
- The frequency of each lemma is given;
- The words are grouped by lemma;
- The lemmas are seperated by parts of speech which gives insight into the word meaning;
- For technical/discipline specific words the discipline is indicated;

The coverage of the AVL was tested on Academic, Newspaper and Fiction texts in the COCA and BNC corpora. In COCA it covered 13.8% and in BNC it cvered 13.7% while in the other two corpuses it covered a smaller amount of the corpus as expected (8.0% and 7.0% in newspaper sections and 3.4% in the fiction section of borh corpuses).

The coverage of the first 570 word families of the AVL was compared to the coverage of AWL in both the COCA and BNC. AWL covered 7.2% of COCA and 6.9% of BNC while AVL covered 13.8% of COCA and 13.7% of BNC.

It should be mentioned that the described replacements for the GSL and AWL are not the only ones offered. For example, in the same year when the New-GSL by Gardner and Davies was published, another improved version of the GSL also called the New-GSL was published by Culligan, Phillips and Browne just a few months before (Browne, 2014). However, these lists included a methodology extremely similar to the original list methodology and were therefore not discussed in this paper.

Overview of significant word lists in the last three years

The word lists which will be described in this section are specialized word lists. This means that they offer a list of the most significant vocabulary in a specific scientific area or genre. Depending on the methodology of list formation, specific word lists either fall into the category of layered word lists or corpus comparison word lists.

The layered approach is modeled after the AWL formation methodology. In this approach the more specific word lists builds upon a general word list and use it as a starting point in the word selection process. The layered approach is intended for intermediate to advanced learners and assumes that the population for whom the list is created is already familiar with the vocabulary of word lists used as the starting basis (Surtees & Horst, 2013). In the corpus comparison approach "word types or families are included in the word list if they are significantly more frequent in a specialized corpus than in a corpus of more general texts or a list generated from a general corpus (Coxhead & Hirsh, 2007). In this approach, all specialized words, including those in the first 2000 most frequent families, are identified using electronic 'term extractors' (Chung & Nation, 2004) that use statistical measures to calculate relative frequency." (Surtees & Horst, 2013 p. 58).

Lei and Liu (2016) believe that the widespread preference of the layered approach might be influenced by Nation's (2001) classification of words into high frequency words, academic words, technical words and low-frequency words. They note however: "Such a classification and the underlying learning-order assumption have, however, been questioned because several studies (e.g. Cobb, 2009; Neufeld, Hancioglu, & Eldridge, 2011; Gardner & Davies, 2014) have found that some AWL items were among the most frequent words in the British National Corpus (BNC) and Davies's (2008) Corpus of Contemporary American English (COCA), challenging a clear-cut division between the high-frequency words and the academic words based only on frequency" (Lei & Liu, 2016 p. 43).

A disadvantage of using GSL and AWL in the layered approach is that it presupposes definite boundaries based on the assumption that words in the GSL and AWL are just general and academic but not technical. However, words used infrequently in everyday language may have one meaning in the general language and a different meaning in specialized communicative settings. Furthermore, frequency counts reveal that many topic-related words in a specialized corpus are actually general words which acquire a specialized meaning in a particular field (Muñoz, 2015). On the other hand, an advantage of the corpus comparison

approach is that it enables researchers to identify words within the GSL and AWL that have acquired technical meanings in specific disciplines (Muñoz, 2015).

The method of word list formation depends on two key factors: intended audience of the list and the overall aim of the list. Depending on the audience the corpus formation criteria and word selection criteria will be chosen. If the aim of the word list is to provide ample text coverage through the identification of the most frequent vocabulary then frequency will be the most significant factor, while if the goal is the identification of vocabulary of a specific field then corpus comparison is more suitable (Surtees & Horst, 2013).

Another distinctive feature of word lists is whether they use word families or lemmas as key elements. The researchers who adopt the lemma approach consider that each lemma needs to be taught separately for adequate comprehension while those who select word family approach believe that learners will also recognize the derived meanings if they know the family headword meaning (Surtees & Horst, 2013). Members of extensive word families may not share the same core meanings (e.g. *react* - respond, *reactionary* - strongly opposed to social and political change, *reactivation* - to make something happen again and *reactor* - device) and these meaning differences are accentuated further as members of word families cross over the various academic disciplines (Hyland and Tse, 2007 as cited in Gardner & Davies, 2013). When using word families it is also possible that only one of the core meanings is highly frequent in the specific field while the rest are rarely used. This would indicate that learners would waste their learning effort on low frequency words they do not need. In the past, the usage of word families was justified because the corpus searching tools could not identify parts of speech and therefore lemmas, but now is the time to take advantage of the possibilities that technological development has made available.

In this section seven specific word lists published from 2013 to 2016 will be explored. The word lists are:

1. International Student Word List - ISWL (Surtees & Horst, 2013)
2. Computer Science Word List - CSWL (Minshall, 2013)
3. Chemistry Academic Word List - CAWL (Valipouri&Nassaji, 2013)
4. TED Word List - TWL (Wolfe, 2015)
5. Environmental Academic Word List - EAWL (Liu & Han, 2015)
6. Nursing Academic Word List - NAWL (Yang, 2015)
7. Medical Academic Vocabulary List - MAVL (Lei & Liu, 2016)

We will present their basic information, corpus compilation criteria, word selection criteria and word list quality test procedures. The basic information about the word lists is given in *Table 1*.

Table 1. Basic word list information

Word list	Size	Corpus	Corpus coverage of the word list	Layered approach or corpus comparison	Target audience and goal
International Student Word List (ISWL) 2013	226 lemmas	147,000 word corpus of Canadian university website	4.4%	Layered approach – first and second BNC frequency	Target audience are international students. The goal is to help them achieve 95% known vocabulary of the university website literature

		literature		bands	about admission, program requirements, insurance, immigration student life etc.
Computer Science Word List (CSWL) 2013	433 word families	Computer Science Corus (CSC) 3.661.337 token corpus	6%	Layered approach – GSL and AWL	The goal of CSWL is to be a pedagogical tool in the instruction of non-native English speakers who are studying computer science in UK universities.
Chemistry Academic Word List (CAWL) 2013	1400 word families	Chemistry research article corpus 4 million words	81.18%	Corpus comparison	The aim of the CAWL construction was to provide chemistry students with a manageable vocabulary load which would help them understand research articles. Target audience are chemistry students in Iran with a low level of English language knowledge.
TED Word List (TWL) 2015	421 word families – 2502 words in total	TED corpus 3.868.390 tokens	2.7%	Layered approach – GSL and AWL	The aim is to increase TED talk usability in the ESL classrooms.
Environmental Academic Word List (EAWL) 2015	458 word families	Environmental science corpus 862,242 tokens	15.43%	Layered approach - GSL	The goal of the EAWL is to provide the vocabulary which environmental students need in order to understand texts in this particular subject area.
Nursing Academic Word List (NAWL) 2015	676 word families	Nursing Research Article Corpus (NARC) 1,006,934 words	13.64%	Layered approach - GSL	The aim of the NAWL is to provide nursing graduate students with the vocabulary which will help the read and write academic papers.
Medical Academic Vocabulary List (MAVL) 2016	891 lemmas	Medical Academic English Corpus (MAEC) 2.7 million words and Medical Textbook English Corpus – MTEC consisting of 3.5 million	19.44% of MAEC 20.18% of MTEC	Corpus comparison	The MAVL was formed to serve the needs of medical students and non-native English speaking medical professionals and researchers who want to read medical research in English.

As can be seen in *Table 1.*two of the word lists implement the corpus comparison approach, while five implement the layered approach. The word lists used as a base lists for the layered approach were: GSL for two lists, GSL and AWL for two lists and the first and second BNC frequency bands for one list. Although the MAVL uses the corpus comparison approach they identified the words within the New-GSL that have acquired technical meanings in specific disciplines and included them in the list.

Only two of the seven word lists use lemmas, the others use word families. This would indicate that the changes recommended by the new methodology pioneers have not been widely accepted yet.

The size of the word lists varies to a great extent, which is only to be expected given the different word formation approaches and the use of different base lists as starting points. The largest word lists is the CAWL with 1400 word families. Its size is explained by the fact that it used a corpus comparison approach because of the target audience's low level of English proficiency. Following this logic, the word lists which use only a general word list as a starting point should be medium sized and the smallest word lists in size should be the ones that build upon both a general word list and the AWL. Therefore, EAWL – 458 word families, NAWL – 676 word families and ISWL – 226 lemmas and to a certain extent MAVL – 891 lemmas (since it uses the New-GSL to some degree in its construction) should constitute the medium sized word lists, while the CSWL – 433 word families and the TWL – 421 word families should constitute the small lists. However, this rule of thumb does not apply to the explored lists since the smallest list of 226 lemmas uses only the general word list as the starting basis.

Regarding the text coverage data, the situation is similarly varied as the word list size data. TWL provides the smallest coverage of 2.7% while the CAWL provides the largest 81.18%. The base word lists influence the expected coverage of the specialized word lists, just like the word list size. However, the goal of all of the word lists is to achieve 95%-98% of corpus coverage (with the base word lists, if they were used) so as to enable undisturbed text understanding.

The corpus size variation is not as drastic as the other characteristics. The smallest corpus is the one used for the ISWL and consists of 147,000 words, while the largest corpus was used for the CSWL and consists of 4 million words. However, for the construction of MAVL two word lists were used which together consist of 5.2 million words.

For the studies which deal with specific word list formation, the corpus is of key importance since the majority of data analysis is done precisely on it. For the word list to be representative of a specific field the corpus must also be representative. We also gave information on corpora division into sub-corpora, since it is needed for the implementation of the word selection criteria *range*. The corpus construction criteria for each of the word lists will be presented.

International Student Word List (ISWL) - The texts were sampled from the websites of English Canadian Universities in four provinces (Concordia University, Dalhousie University, University of Toronto and University of British Columbia). An even text sample was gathered from each University website. All postal and email addresses, telephone numbers and lists of organizations and store names were removed from the corpus. All acronyms, place names, institution names, names of products stores and websites were placed in the 1000 word band because of their low learning burden.

Computer Science Word List (CSWL) - During the construction of the CSC the author paid special attention to the properties of size, balance and representativeness. Since AWL was used as a model for the study a corpus size of 3.661.337 tokens was judged to be large enough. The corpus was built using journal articles, special interest group newsletters and

conference proceedings. Ten sub-disciplines of computer science as defined by the Association of Computing Machinery were included in the corpus. The sub-disciplines are: computer system organization, computing methodologies, hardware, human-centered computing, information systems, mathematics and computing, networks, security and privacy, software and its engineering, theory of computation. The CSC consists of twenty different sub corpora (two primary text types across 10 sub-disciplines). For the property of representativeness to be fulfilled each sub-corpora contained about 180,000 tokens. The CSC is partitioned into two corpora of equal size: Computer Science Journal Article Corpus and Computer Science Conference Proceedings Corpus. In the construction of the corpus 408 texts were used by more than a 1000 authors. Lists of references, appendixes, page titles, authors' names, keywords, content pages, copyright information, publication names, abbreviations, acronyms, all non-alphabetical data and tabular data were removed from the texts which went into the corpus.

Chemistry Academic Word List (CAWL) - The Chemistry research article corpus consists of 4 million words. The following criteria were used for the text selection: 1. Scientific papers were selected from four areas in chemistry: analytical chemistry, inorganic chemistry, organic chemistry and physical/theoretical chemistry; 2. Ten scientific journals were selected from each subject area and 10 papers were randomly chosen from each journal; 3. The scientific papers were divided according to length to short, medium and long, so that an even sample was gathered.

TED Word List (TWL) - The corpus consisted of 1790 TED talk transcriptions from June 2006 to December 2014. No validation tests were conducted for this study. However, for this study it would be impossible to form a new corpus using the same formation criteria since all of the TED talks available at the time went into the formation of the main corpus. The only alternative which could be a solution is to wait and compare the text coverage of the TWL on a smaller corpus of newer TED talks which were published after December 2014.

Environmental Academic Word List (EAWL) - The environmental academic word list consists of 862,242 tokens from 200 research papers from 10 areas of ecological studies. One journal was selected for each area of ecological studies. Charts, diagrams, numbers, appendices, bibliographies, equations and other textual components which could not be processed by computer software programs were removed.

Nursing Academic Word List (NAWL) - During the creation of the NARC four criteria were used for the text selection process: 1. Only research articles focusing on empirical studies (with the *Introduction, Method, Result, Discussion* sections) were included; 2. The authors needed to be native English language speakers;3. That the research articles were published from 1995 to 2011; 4.That their length was between 2,000 and 10,000 words. Twenty one sub-disciplines of nursing are represented in the NARC.

Medical Academic Vocabulary List (MAVL) - The fact that two custom corpuses are used is specific for this word list. The MAEC consists of 760 articles taken from 38 academic journals in medicine. Twenty articles were taken from each journal. The journals were randomly selected from 176 SCI-indexed medical journals and they cover 21 specialist areas i.e. sub-corpora. However, a limitation is the fact that the number of journals included in each area was not identical. The tables, figures, notes, endnotes and footnotes were removed. Medical English Textbook Corpus – METC consisting of 3.5 million words, it was designed as a cross-

check reference corpus. The corpus is comprised of the 3-volume *Oxford Textbook of Medicine* (Warrell et al. 2003).

As it can be deduced from the information presented above, all of the authors tried to follow the factors of balance and representativeness. Balance was achieved by controlling the size of the texts so that each sub-corpora is equally represented. The representativeness criterion was reflected in the careful selection of text sources and controlling the age of the texts.

It could be noted that the material for the environmental science corpus was not ideal, considering the EAWL construction goal. It would have been useful for the corpus to include textbooks used at universities. Just like Coxhead did for the construction of the AWL. One explanation for not including textbooks is that textbooks are not ideal for corpus building due to issues such as author bias. "As textbooks are large texts, often with only one author, including them in a corpus can skew the results due to an author's preference for particular words and other idiosyncrasies" (Atkins et al. 1992 as cited in Minshall, 2013 p. 21). This problem was dealt with in two different ways amongst the word lists in this study. In CSWL the validity test corpus included textbooks, hence, if a significant difference in vocabulary usage between textbooks and research articles existed, it would have been noticed. The MAVL on the other hand solved this problem by forming their word list on the basis of two corpora – one comprised of scientific articles written by a plethora of authors and one comprised of textbooks. It should also be noted that, the criteria of only including texts written by native speakers is unfounded. With its implementation significant studies in the field would be excluded. Furthermore, all of the published materials went through the process of review which would indicate their language adequacy.

After the corpus formation the next step is determining the word selection criteria. Information about the word selection criteria for each examined word list is given in *Table 2.*

Table 2. Word Selection Criteria

Word list	**Removal of words from a base list**	**Frequency**	**Range**	**Distribution/ Dispersion**	**Ratio**	**Other**	**Analysis Tool**
International Student Word List (ISWL) 2013	First and second BNC word bands – but also GSL and AWL for easier comparison with other word lists	7 occurrences in the whole corpus	The lemma had to occur in at least 3 of the four sub-corpora	/	/	/	Web Vocabprofiler (Cobb, 2012), Range (Cobb,2009);

Computer Science Word List (CSWL) 2013	GSL and AWL (together they cover 89.22%)	80 occurrences in the whole corpus	The word needed to occur in half of the sub-corpuses	/	/	/	AntWord Profiler (Anrhony, 2008), AntConc (Anthony, 2002)
Chemistry Academic Word List (CAWL) 2013	*Grammar words and acronyms were removed.	The word must occur at least 114 ties in the whole corpus	The word must occur at least 10 ties in all of the four sub-corpora	/	/	3 chemistry experts excluded the terms which are specific for one area of chemistry science;	Range (Heatley et al. 2002)
TED Word List (TWL) 2015	GSL and AWL (together they cover 92%) but also for comparison NGSL and NAWL	100 occurrences in the whole corpus	/	In this research called a TED number. It goes from 0 to 1 but was not used as a word selection factor.	/	/	Vocabulary profiler, TextMate, AntWord Profiler (Anthony, 2014), AntConc;
Environmental Academic Word List (EAWL) 2015	GSL	/	The word needs to occur in 8 form 10 subject areas	/	/	Usage – as defined by Juilland & Chang-Rodriguez, 1964 – value of 30 was the cut off point	Range (Cobb,2009);
Nursing Academic Word List (NAWL) 2015	GSL	The word must occur at least 33 times in the whole corpus	The word must occur in at least 11 of 21 subject areas	/	/	/	Range (Heatly et al. 2002)
Medical Academic Vocabular	New-GSL however the words	71 occurrences in the whole	A lemma should occur	The Julliard's D value of lemma must be at	1.5 frequency	Discipline measur	Stanford CoreNLP – for

y List (MAVL) 2016	were compared to two medical dictionaries to see whether these general words had specific medical if they didn't they were removed	corpus	minimally with 20% of the expected frequency in at least 12 of the 21 sub-corpora	least 0.5.	ratio was used in comparison to the non-academic sub-corpora of the NBC	e- No lemma should occur more than 3 times the expected frequency in more than 3 sub-corpora	lemmatizing and POS-tag words; The processing tasks were done by unnamed programs;

The AWL is frequently used as a word list formation model, most of the studies in this paper are no exception. Three of the word lists (the ISWL, CSWL and NAWL) used the same word selection criteria as Coxhead in her AWL (removal of words from a base list, frequency, range) except for the uniformity criterion. The TWL follows the AWL criteria of removal of words from a base list and frequency but because of the nature of the corpus which is not divided into sub-corpora the range and uniformity criteria could not be implemented. The CAWL also follows the first three steps in word selection as proposed by Coxhead, omits the criteria of uniformity, but also adds a step in the word selection process. Namely, three chemistry experts were selected to exclude the terms which are specific for only one area of chemistry science. The EAWL uses the criteria of removal of words from a base list and frequency, but also the criterion of *usage*, as defined by Juilland & Chang-Rodriguez (1964). What is specific about MAVL is the fact that it combines methods and procedures from Coxhead (2000) and Gardner and Davies (2013). For the word selection process the authors adopt AVL standards of ratio, range, dispersion and discipline measure but also AWL's standard of minimum frequency. This is because the frequency ratio ensures that the given word occurs more frequently in the academic texts than in general texts, but this does not mean that these words are high frequency words in academic texts (Lei & Liu, 2016). The sixth and final step for the inclusion of a lemma in the MAVL was the special medical meaning check. Namely, the lemmas gathered through the previous 5 steps were compared to the AVL. Those which appeared in the AVL needed to have a special medical meaning to be included in the MAVL, which was checked by seeing if the lemma occurred in the Merrian-Webster's medical English dictionary, new edition and Taber's Cyclopedic Medical Dictionary. However, because two custom corpora were used in the MAVL formation, two additional phases need to be implemented. The first phase is the selection of words in the MAEC using the above mentioned criteria. The second phase is checking if the selected words fulfill the frequency criteria in a second corpus - the MTEC.

The popular usage of the Range program (ISWL, CAWL, EAWL and NAWL) is yet another fact which attests to the strong influence of the AWL methodology process on the formation of specific word lists. The AntWordProfiler is another popular tool and is used in two of the word lists – CSWL and TWL.

After the word lists are completed they are compared to the AWL and other relevant word lists if such exist (e.g. the NAWL was compared to the Medical Academic Word List (Wang et al. 2008).

The word list validity and relevance should also be tested once the word list is complete. The authors of the EAWL proposed validity test criteria of specialized academic word lists:

1. establishment of a validation corpora using the same criteria for text selection and data processing as were used for the main corpus;
2. coverage comparison of a specific word list to the AWL;
3. the usage of a paired t-test for the statistical analysis of the coverage of the specific word list in comparison to the AWL;

Specialized word list relevance is tested by checking the specialized word list coverage against a general corpus, such as the BNC. Chung and Nation (2003) recommend comparison of specific word lists to technical dictionaries in the specific scientific area so that the relevance of the word list could be further confirmed (Minshall, 2013).

Information about the validity and relevance tests conducted in the word lists explored in this paper are given in *Table 3*.

Table 3. Validity and relevance test information

Word list	Validity test - corpus	Relevance test - corpus	Technical dictionary relevance test
International Student Word List (ISWL) 2013	/	/	/
Computer Science Word List (CSWL) 2013	The second corpus for the validity test was made using the same criteria as the CSC except for the fact that textbooks were also used as a source. It consisted of 693,551 tokens from 23 different texts. The coverage of the CSWL was 4.68%. Together with the GSL and AWL the total text coverage of the test corpus was 94.41%.	The CSWL was compared to a fiction corpus in order to test its relevance. It was compiled with the resources of Project Gutenberg and consists of 3,671,673 tokens from 26 different texts. The text coverage of CSWL was 0.39%.	Oxford Dictionary of Computing – ODOC (2008) was used for the comparison. A total of 70.2% of all word families from the CSWL had an entry in the ODOC. These results support the notion that the word list contains relevant subject specific technical lexis.
Chemistry Academic Word List (CAWL) 2013	/	/	/
TED Word List (TWL) 2015	/	/	/
Environmental Academic Word List (EAWL) 2015	Two smaller ecology corpora were formed using the same criteria that were used in the formation of the main ecology corpus. The EAWL coverage was 14.92% and 15.59%. This data proves its relevance.	The coverage of EAWL was tested on the fiction, magazine and newspaper sub-section of the BNC. It covered only 1%, 3.2% and 2.8% respectively, of these sub-sections which	/

		proves its relevance.	
Nursing Academic Word List (NAWL) 2015	/	/	/
Medical Academic Vocabulary List (MAVL) 2016	The MAVL was formed using two medical corpuses, but it was not compared to a third.	The MAVL was tested on the BNC and covered 3.69% of the corpus, which is significantly lower than its coverage of the BNC Academic corpora 6.63% and medical corpora 19.44% and 20.18%.	The comparison to two medical dictionaries was part of the word list compilation process.

An additional analysis to test the practicality of the CSWL was conducted. Namely, the author tested which frequency bands from the BNC and COCA needed to be learned in order to reach the same 95% coverage reached through the combination of GSL, AWL and CSWL (2.992 word families). With the 14k band of the BNC corpus 96.44% of the corpus is covered, and only 93.53% of the corpus is covered with the 25k coverage of the COCA (Minshall, 2013).

As shown in Table 3.only three of the seven word lists implemented validity and relevance checks. Furthermore, no strict criteria regarding the test conditions seem to be followed.

Discussion

Although the majority of the word lists explored in this study use the AWL as a methodology model, significant variation still exists in the specific word list formation methodology. It can be concluded that the methodological changes proposed by the New-GSL and AVL are still not widely implemented. Further research is needed to test if these new methodologies really are superior, as they seem to be. It can be suggested for future researchers to use the methodology proposed by MAVL (if it is in line with the specific word list goal and target audience) which implements the AVL methodology but also enhances it by incorporating some of the features of the AWL. The MAVL methodology had two disadvantages which should be avoided. The first is that the criteria of corpus balance was not controlled as much as it could have been. The second is that the authors did not conduct word list validity and relevance tests. Guidelines for avoiding these disadvantages are covered in the main body of the text.

Conclusion

In this paper an overview of specific word lists which were published from 2013 to 2016 has been given. The overview presented in this paper is not comprehensive; access to a certain number of specific word list papers was not readily available. Furthermore, some papers on specific word lists might have been overlooked in the review of the relevant literature. The main aim was to give information based on which we could observe the changes happening in the field as a result of new methodologies proposed. The second aim was to give a methodological overview and guidelines which would help future researchers, interested in specific word list formation, select the most appropriate methodological methods. As it can be seen from the results, the word list formation process is rather varied at the moment.

Further research is needed to test the usefulness and validity of the new methodological processes.

Bibliography:

Anthony, L. (2002). AntConc [Computer software]. Retrieved June 1st, 2013, from http://www.antlab.sci.waseda.ac.jp/software.html

Anthony, L. (2008). AntWordProfiler [Computer software]. Retrieved June 1st, 2013, from http://www.antlab.sci.waseda.ac.jp/software.html

Anthony, L. (2014). AntWorldProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available at:http://www.laurenceanthony.net/

Bell, T. (2001). Extensive reading: speed and comprehension. *The Reading Matrix. Vol. 1, No.1* [electronic version] http://www.readingmatrix.com/articles/bell/article.pdf

Brezina, V. & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, amt018.

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction*, *3*(2).1-10.

Chen, Q. & Ge, G. C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*(4).502-514.

Chung, T. M.& Nation, P. (2003).Technical vocabulary in specialised texts.*Reading in a foreign language*, *15*(2).103-116.

Chung, M. (2009). The Newspaper Word List: A Specialised Vocabulary for Reading Newspapers. *Jalt Journal*, *31*(2).159-182.

Cobb, T. (2009). Range (Version 2) [Web-based corpus analysis software adapted from Heatley & Nation, 1994]. Available at:http://www.lextutor.ca/range/

Cobb, T. (2012).WebVocabprofile (Vresion 3). [Web-based corpus analysis software adapted from Heatley& Nation, 1994]. Available at:http://lextutor.ca/tuples/eng/

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238.

Coxhead, A. & Hirsch, D. (2007). A pilot science-specific word list. *Revue française de linguistiqueappliquée*, *12*(2).65-78.

Coxhead, A. (2011). The Academic Word List 10 Years On: Research and Teaching Implications. *TESOL Quarterly*, 45(2). 355-362.

Dang, T. N. Y. & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*.66-76.

Gardner, D.& Davies, M. (2013).A new academic vocabulary list. *Applied Linguistics*, doi:10.1093/applin/amt015.1-24

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language*, *23*(1), 65-83.

Heatly, A., Nation, I. S. P. &Coxhead, A. (2002).RANGE and FREQUENCY programs. Retrieved from: http://www.vuw.ac.nz/lals/staff/Paul_Nation

Hyland, K. & Tse, P. (2007). Is there an" academic vocabulary"?. *TESOL quarterly*, 41(2).235-253.

Juilland, A.& Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. California: Moulton.

Konstantakis, N. (2007). Creating a business word for teaching Business English. *Elia: Estudios de lingüísticainglesaaplicada*, (7), 79-102.

Lessard-Clouston, M. (2013). Word lists for vocabulary learning and teaching, *The CATESOL Journal,* 24(1). 287-304.

Lei, L. & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, 22.42-53.

Li, Y. & Qian, D. D. (2010).Profiling the Academic Word List (AWL) in a financial corpus. *System*, *38*(3).402-411.

Liu, J. & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, 39.1-11.

Liqin, Y. & Xinlu, G. (2014). Word saliency and frequency of academic words in textbooks: A case study in the new standard college english.*International Education Studies*, *7*(4).14-27.

Martinez, I.A., Beck, S.C. & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3).183-198.

Minshall, D. E. (2013). A Computer Science Word List. *Unpublished MA dissertation, University of Swansea.Available at DE Minshall.*Available at:*www.baleap.org*.

Mozaffari, A.& Moini, R. (2014). Academic Words in Education Research Articles: A Corpus Study. *Procedia-Social and Behavioral Sciences*, *98*.1290-1296.

Muñoz, V.L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, 39, 26-44.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, 14, 6-19.

Nation, P. & Deweerdt, J. (2001). A defence of simplification. *Prospect.*16(3). 55-65 [electronic version] http://www.ameprc.mq.edu.au/docs/prospect_journal/volume_16_no_3/Prospect_16,3_Article_5.pdf

Nation, P. (2006). How Large a Vocabulary Is Needed For Reading and Listening? *The Canadian Modern Language Review*, 63(1).59-82.

Ng, Y. J., Lee, Y. L., Chong S. T., Nurhanis, S., Philip, A., Noor, H.N.A. &Mohd, A.A.T. (2013) Development of Engineering Technology Word List for Vocational Schools in Malaysia. *International Education Research*, 1(1).43-59.

Surtees, V. & Horst, M. (2013). An Alternate Academic Vocabulary: A Word List for Canadian University Websites.

Valipouri, L. & Nassaji, H. (2013).A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, 12(4).248-263.

Vongpumivitch, V. Huang, J. &Chang, Y. (2009) Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers.*English for Specific Purposes.*28(1).33-41.

Wang, J., Liang, S. L.&Ge, G. C. (2008).Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4).442-458.

West, M. (1953).A general service list of English words. London: Longman, Green.

Warrell, D. A., Cox, T. M.& Firth, J. D. (2003). Oxford textbook of medicine (Vols. 1e3). Oxford: Oxford University Press.

Wolfe, J. D. (2015). *The TED word list: an analysis of TED talks to benefit ESL teachers and learners*. Doctoral dissertation.Royal Roads University.

Yang, M. (2015).A nursing academic word list. *English For Specific Purposes*, *37*. 27-38.

Yazhen, S., & Lei, L. E. I. (2013). Academic Lexical Items in Economics Journal Papers. *Chinese Journal of Applied Linguistics, 36(3)*.354-368.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language learning and communication*, *3*(2).215-229.

***Biographical notes*:**

Radmila Palinkašević is a teaching assistant at the Preschool Teacher Training College "Mihailo Palov". She has finished her BA and MA studies at the Faculty of Philology University of Belgrade where she is currently studying for her PhD. She researches and publishes in the fields of English language teaching methodology and English linguistics, with a focus on vocabulary acquisition, word lists, learner anxiety etc.