http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Corpus-based Headword Selection Procedures for LSP Word Lists and LSP Dictionaries

Milica Vuković-Stamatović, *Faculty of Philology, University of Montenegro, Nikšić, Montenegro (vmilica@ucg.ac.me)*
and
Branka Živković, *Faculty of Philology, University of Montenegro, Nikšić, Montenegro (brankaz@ucg.ac.me)*

Abstract: In compiling both Language for Specific Purposes (LSP) word lists for foreign language learners and LSP dictionaries, the headword-selection process is of paramount importance. LSP word lists and LSP dictionaries will function effectively if they contain appropriate terms and register items, i.e. the lexical items that end users need. In this paper, we first present corpus-based LSP word lists, with special emphasis on how they were compiled. In the process, the make-up and size of the specialised corpus are important, as is the choice of the headword selection methods used. Among the possible criteria are word frequency, keyness, specialised occurrence, range, and dispersion, as well as some non-corpus linguistic methods that are more rarely applied. A greater variety of methods is used for compiling headword lists for LSP dictionaries, and of the corpus linguistic methods, frequency is typically solely applied. The article compares headword selection procedures for LSP word lists and LSP dictionaries before discussing how they can mutually inform one another.

Keywords: LANGUAGE FOR SPECIFIC PURPOSES, LSP WORD LIST, LSP DICTIONARY, CORPUS LINGUISTICS, HEADWORD LIST, TERMS, HEADWORD SELECTION

Opsomming: Korpusgebaseerde lemmaseleksieprosedures vir TSD-woordelyste en -woordeboeke. In die samestelling van beide Taal vir Spesifieke Doeleindes-(TSD-)woordelyste vir vreemdetaalleerders en TSD-woordeboeke is die lemmaseleksieproses van kardinale belang. TSD-woordelyste en TSD-woordeboeke sal effektief funksioneer indien hulle toepaslike terme en registeritems, m.a.w. die leksikale items wat eindgebruikers benodig, bevat. In hierdie artikel word korpusgebaseerde TSD-woordelyste eerste bespreek, met besondere klem op hul samestelling. In hierdie proses is die samestelling en grootte van die gespesialiseerde korpus, asook die keuse van die lemmaseleksiemetodes wat gebruik word, belangrik. Onder die moontlike kriteria is woordfrekwensie, sleutelstatus, gespesialiseerde voorkoms, omvang en verspreiding, asook enkele nie-korpus-linguistiese metodes wat minder gereeld toegepas word. 'n Groter verskeidenheid metodes is gebruik vir die samestelling van lemmalyste vir TSD-woordeboeke, en van die korpuslinguistiese metodes is slegs frekwensie tipies toegepas. Lemmaseleksieprosedures vir TSD-woordelyste en TSD-woordeboeke word in die artikel vergelyk voordat daar bespreek word hoe

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

hulle mekaar wedersyds van inligting kan voorsien.

Sleutelwoorde: TAAL VIR SPESIFIEKE DOELEINDES, TSD-WOORDELYS, TSD-WOORDEBOEK, KORPUSLINGUISTIEK, LEMMALYS, TERME, LEMMASELEKSIE

1. Introduction

Word lists have many purposes in the process of teaching and learning a foreign language: they can be used as resources for vocabulary learning (Khani and Tazik 2013; Yang 2015), guidelines for designing curricula and courses, as well as for selecting reading and listening materials (Wang, Liang and Ge 2008; Jin et al. 2013), and guidelines for teachers in organising their explicit vocabulary teaching (Khani and Tazik 2013). The selection of headwords for inclusion in certain word lists has become an important strand of applied research in the field of foreign language teaching and learning in general, and language for specific purposes (LSP) in particular. As vocabulary sizes attained by native speakers are never attained by a vast majority of foreign language learners, the rationale guiding this type of research is to produce word lists of the sizes which are manageable for them to learn from. Word lists should provide language learners with the most useful words they need for a particular language function they are pursuing, for instance, attending university studies in a foreign language or reading research articles from a particular specialist field in a foreign language. Some of these functions are related to LSP contexts and for them, consequently, LSP word lists are produced. Most of them are, in fact, English for Specific Purposes (ESP) word lists, given that English is the language which is most widely taught as a foreign language around the world.

In the past, both general and LSP word lists used to be compiled manually, typically relying on the compiler's intuition and, more rarely, on an authentic corpus of a very limited size by today's standards (West 1953; cf. Gilner 2011). However, over the past two decades, they have principally been derived from vast authentic corpora of general or specialised texts, which are carefully constructed having particular types of foreign language learners in mind, and then scanned for words meeting certain criteria or a combination of criteria, such as the frequency of occurrence, distribution, range, or keyness (Coxhead 2000; Coxhead and Hirsh 2007; Brezina and Gablasova 2013; Browne et al. 2013a, Gardner and Davies 2014, etc.). The choice of the criteria and the related "cut-off" points (for instance, how frequent a word has to be to be included in a certain word list) are informed by the target users' needs and involve a number of decisions during the compilation of the list. As corpora and software solutions evolve, so do the different methods for selecting those words. In this paper we will discuss various word lists intended for LSP learning, with a focus on how they were compiled.

Selection of headwords for any dictionary, including specialised dictionaries, is also governed by the needs of its end users (Fuertes-Olivera and Arribas-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Baño 2008), i.e. what should be taken into account are different types of users, user situations and user needs (Tarp 2008), according to the theory of lexicographic functions (Bergenholtz and Tarp 1995; Tarp 2008). In principle, there are four main methods of selecting headwords for dictionaries — these assume relying on the existing dictionaries, grammar and etymology, canonical literary texts, or corpora (Esandi-Baztan and Fuertes-Olivera 2020). The fourth method, compiling headword lists based on corpora, has been an option for the past few decades and is now widely used in the process of making general dictionaries. However, as Bowker (2010: 166) notes, the use of corpus linguistic methods has been rather slow to take hold in the creation of specialised dictionaries. When it comes to the methods and procedures of compiling corpora for the purpose of creating LSP dictionaries as a type of specialised dictionaries, one may only rarely find detailed accounts regarding this issue (cf. Khumalo 2015; Đurović 2021; Kruse and Heid 2021). Also, typically, few details are also presented relating to the corpus-linguistic procedures employed as part of the process of selecting headwords from specialised corpora — most studies only briefly note that it is the frequency criterion that was applied (cf. Rundell and Kilgarriff 2011), without delving into the type of details that are provided by various specialised word-list compilers (cf. Lei and Liu 2016; Todd 2017; Dang 2018, etc.). In addition, in these accounts, further corpus-linguistic procedures for headword selection beyond simple frequency are only sometimes mentioned in LSP dictionary research and projects (cf. Khumalo 2015; Đurović 2021; Kruse and Heid 2021).

In this paper we compare corpus-based headword selection procedures used for producing LSP word lists and LSP dictionaries, bearing in mind that there are some similarities (although, also, important differences) between these two types of lexicographic products. We focus on the steps in headword selection that are based on corpus bearing in mind the important place that corpora currently have in their creation. The premise from which we depart is that the two fields can mutually inform and contribute to one another in terms of the corpus-based headword selection procedures.

We will first present an overview of word lists, with a special focus on LSP word lists and how they are produced (section 2), after which we discuss LSP dictionaries and how headwords are selected for them (section 3). Section 4 compares headword selection for LSP world lists and LSP dictionaries.

2. Word lists

This section first provides a brief overview of general and academic word lists, after which the focus is narrowed down to discipline-specific or LSP word lists.

Reviews of word lists used for the purposes of foreign language teaching and learning typically start by presenting West's General Service List (GSL) (1953) (cf. Coxhead 2000; Coxhead and Hirsh 2007; Gardner and Davies 2014; Dang and Webb 2016; Dang, Coxhead and Webb 2017; Dang 2018; McQuillan 2020, etc.).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Although West's list was not generated using computer software, it was based on an authentic word corpus of 5 million words representing General English. About 2,000 word families[1] were manually extracted and suggested to be the first words to be learned by any English language learner (they were mostly chosen according to the frequency criterion). This word list was very influential in English Language Teaching (ELT) and was used widely for decades (Nation 2013; Coxhead 2018). The emergence of the computer solutions providing data on a word's frequency and coverage in a corpus showed why — it turned out that West's list covered about 80% of the words used in most general English texts, or 4 in every 5 words. As English has about 70,000 word families (Nagy and Anderson 1984; Nation 2013), this word list proved to be a very useful resource (Coxhead 2000; Nation 2013).

In the ensuing decades, other English words were built too (for instance, Campion and Elley 1971; Praninskas 1972; Lynn 1973; Ghadessy 1979; Xue and Nation 1984, etc.), however, the next word list which can match the influence of the GSL, the Academic Word List (AWL), came only in 2000 (Coxhead 2000). Its influence lies not only in how widely it was used in ELT, but the methodology of its compilation also set standards for many of the ensuing word lists (among them, Fraser 2007; Konstantikis 2007; Wang, Liang and Ge 2008; Khani and Tazik 2013; Valipouri and Nassaji 2013; Hsu 2013; Minshall 2013; Hsu 2014; Liu and Han 2015; Yang 2015; Lei and Liu 2016, etc.). The AWL contains 570 word families which are common in academic writing. To produce the list, Coxhead compiled a corpus of 3.5 million words of academic texts. The words were extracted according to the following criteria: (1) specialised occurrence (the words had to be outside high-frequency general words (outside the GSL in this case)), (2) frequency, (3) dispersion (the words had to occur in all the corpus's subsections while featuring a certain frequency in all of them, and they also had to occur in at least half of the academic disciplines involved in the corpus) (Coxhead 2000). These carefully weighed and strict criteria ensured that the word list would have a substantial coverage in any academic corpus, not just in the one it was derived from (Coxhead 2000). Indeed, the AWL's coverage of 10% in the corpus of its origin held strongly in many other academic corpora compiled later — for instance, it featured 10.07% in the academic medical corpus (Chen and Ge 2007), 11.17% in the academic applied linguistics corpus (Vongpumivitch et al. 2009), 9.96% in academic chemistry corpus (Valipouri and Nassaji 2013), etc. These impressive results confirmed that any future word list would have to be carefully made, so as to be as useful as possible in a variety of similar language contexts.

One of the rare issues that may be contended against the AWL is the relatively small corpus it was derived from taking into account that it aims to be a general academic word list, an issue which the ensuing general lists have been trying to overcome. The dated GSL needed to be replaced and two new GSLs were offered to both research and instructional purposes in 2013. Brezina and Gablasova (2013) based their New GSL, containing about 2,500 lemmas, on a combined corpus of samples from 4 different corpora, together making 12 bil-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

lion words. The lemmas from each of the 4 corpora were selected based on the criterion of the Average Reduced Frequency (this measure is obtained from the absolute frequency of the word and its distribution in the corpus (Savický and Hlavácová 2002)), and then the 4 lists were compared for overlaps — the shared items entered the New GSL. The same year, Browne, Culligan and Phillips (2013a) used a 273-million-word section of the Cambridge English Corpus to derive their list of about 2,800 lemmas based on the frequency criterion. Both lists outperform the old GSL in modern corpora, typically by a few percentage points.

Browne, Culligan and Phillips (2013b) also created the New AWL, containing 963 lemmas, by excluding the words already contained in the NGSL. Another replacement for the AWL was offered by Gardner and Davies (2014), who used a 120-million-word corpus (an academic subsection from COCA), to produce a list of about 3,000 lemmas (the Academic Vocabulary List, or the AVL). They did not exclude any group of words, but employed the keyness criterion solely: the authors took into account the ratio of words in their academic corpus, compared to a non-academic corpus. Newman (2016) and Hernandez (2017) found that the AVL outperforms the old AWL, while not much data is available on how the NAWL performs against other similar lists.

Other researchers have investigated whether lists such as the AWL might be created for other languages. Cobb and Horst (2004) studied the vocabulary profile of French and determined that the high-frequency vocabulary of this language is in fact more frequent than the high-frequency vocabulary of English (2,000 most frequent French words reach a 90% coverage in most texts they examined), which excludes the need for creating additional lists for learners as these would reach very small coverages. Such results for French did not discourage other researchers to pursue creating corpus-based academic word lists for other languages, however. A Nordic joint-research project resulted in the creation of the academic word lists for Swedish, Norwegian and Danish (Kokkinakis et al. 2012; Jansson et al. 2012; Ribeck et al. 2014; Johannessen et al. 2016). Two more independent lists have also been created for Danish — a word list of general, high-frequency items (2,000 words), as well as a word list of academic vocabulary (402 words) (Jakobsen et al. 2018). An Academic Vocabulary List in Russian has also been compiled recently (Talalakina et al. 2020). The development of all these word lists heavily relied on the English word-list projects presented above.

The word lists mentioned so far include general and non-discipline specific academic word lists. Unlike these, other word lists are much more specialised and these are the focus of this paper. They and the methods used for compiling them will be presented in the following section.

2.1 Corpus-based headword selection procedures for LSP word lists

Realising the importance of the role of the communicative contexts in which certain foreign language learners will typically find themselves (Miller 2014: 305),

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

teaching LSP began to be strongly differentiated from teaching General Foreign Language in the 1960's. LSP teachers and researchers realised that taking the learners' specific needs into account, particularly their vocabulary needs, led to more effective teaching of the specialised language that they needed. With the rise of the ITC industry, corpus-based discipline-specific word lists, produced with the use of computers and from vast corpora, began to emerge at the turn of this century.

An overview of recent LSP word lists, along with the details of the corpora from which they were derived and the methods used for their creation, is given in the Appendix (while not entirely exhaustive, the table presents most of the word lists which have been described in scholarly papers). As was the case with general and academic word lists, the field of researching and compiling LSP word lists is almost exclusively related to the English language and, consequently, English word lists dominate the literature (as can be seen in the Appendix). Many of these lists follow in the AWL's footsteps given that they rely or build on the criteria used by Coxhead (2000) (see Section 2). Here we will provide a generalised description of the corpora and methods typically used to create LSP word lists.

The texts for LSP corpora are chosen bearing the LSP word list's target users in mind. The corpora from which word lists are produced are typically custom-made, which makes their creation challenging and time-consuming. They also need to be of a relevant size. The corpora from which the LSP word lists were made vary widely in terms of their size — most of the word lists were developed from a specialised corpus of 1–2 million words (Mudraya 2006; Coxhead and Hirsh 2007; Wang, Liang and Ge 2008; Vongpumivitch, Huang and Chang 2009; Khani and Tazik 2013; Yang 2015; Todd 2017; Kwary and Artha 2017; Tongpoon-Patanasorn 2018; Đurović 2021). However, a recent tendency is to use larger corpora — most of the corpora from the last decade featured 4 or more million words (Valipouri and Nassaji 2013; Hsu 2013; Hsu 2014; Lei and Liu 2016; Moini and Islamizadeh 2016; Dang 2018; Khany and Kalantari 2021; Kamrotov et al. 2022). The biggest corpus used is the most recent one — a corpus of almost 30 million words of accounting research articles, which was used to obtain a list of the most frequent 658 accounting words (Khany and Kalantari 2021).

The LSP word-list compilers who intend to apply the word selection criteria of range and dispersion need to think carefully about the make-up of their corpora as they generally need to have equal subsections of texts from various subfields. These corpora thus need to be well-structured and balanced; even though this is a challenging task, some researchers were able to produce significantly large and at the same time well-structured corpora — for instance, such is the English Hard Science Spoken Corpus of 6.5 million words, produced by Dang (2018), which features 12 subsections representing 12 hard science disciplines. This size is all the more impressive bearing in mind that this is a corpus of spoken language.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

The sizes of LSP word lists also vary widely — from 92 (Martínez, Beck and Panza 2009) to 1,595 headwords (Dang 2018) and, again, the needs of the end users are taken into account when determining the list's size, as is the case with dictionaries.

The criteria used for the selection of words for various recent LSP word lists can be summarised as follows:

1. frequency (the number is set depending on how large a list is wanted),
2. specialised occurrence (being outside the most frequent 2,000 or 3,000 words, so as to avoid general high-frequency words; additionally, being outside the most frequent academic words (as represented by a chosen academic word list); finally, this also assumes the exclusion of proper nouns, symbols, abbreviations, numbers, non-words, etc.),
3. dispersion (typically, occurrence in at least half the disciplines/subsections which make the corpus, or being below some dispersion value (different methods for determining these are available)),
4. keyness (being found in the specialised corpus more frequently than in a reference corpus),
5. expert opinion (experts use rating scales and assign more points to more technical words),
6. cross-comparison with specialised dictionaries.

The first four are purely corpus-linguistic methods and assume automatic extraction of words based on the word-list compiler's decisions regarding the thresholds applied, while the last two depend on consulting either experts or specialised dictionaries, and are much more time-consuming. The final two steps have been generally avoided in developing most LSP word lists; having applied several corpus-linguistic filters, the word-list compilers found them unnecessary. Experts and dictionaries were consulted in the creation of just four out twenty-four LSP word lists presented in the Appendix (Wang, Liang and Ge 2008; Valipouri and Nassaji 2013; Jin et al. 2013; Tongpoon-Patanasorn 2018).

It should be added that the finalised LSP word lists are also typically validated in one or several independent corpora (following Coxhead 2000) and, if their expected coverages hold in new corpora, such word lists are assumed to be truly representative.

Few studies, typically those early ones or those using a vast corpus, used just one word-selection criterion (typically, frequency or keyness) (Mudraya 2006), while most of the studies employed a combined approach by using several of the methods — most often, following Coxhead's method (2000) (the first three steps above). None of the studies applied all the six methods combined.

As can be seen, the field of producing and investigating word lists developed as part of applied linguistics by Anglo-Saxon scholars, who, despite the fact that there are now many authors in it who are not Anglo-Saxon, still dominate it to a large extent. Most of the word lists are in fact English word lists. The creation of word lists is guided by pragmatic principles and the field remains

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

atheoretical. So far, in the literature, there have not been any proposals to introduce a theory which would support the field.

3. LSP dictionaries

As Bowker (2010) explains, LSP dictionaries belong to specialised dictionaries, i.e. dictionaries which treat specialised fields. They are also seen as a type of restricted dictionaries (Burkhanov 1998), where the term *restricted* does not imply their smaller size but reflects the fact that they focus on specific and precise vocabulary (Mihindou 2004). LSP dictionaries exist in many fields of knowledge (Landau 2001), while developing the metalexicography related to them is in full swing (Fuertes-Olivera and Arribas-Baño 2008).

While the Anglo-Saxon strand in lexicography is mostly atheoretical (as was the case with the field of compiling word lists), the strand influenced by German and Nordic scholars advocates for developing lexicographical theories for guiding dictionary research and compilation (Fuertes-Olivera et al. 2013). As mentioned earlier, what is taken into account in the process of compiling any dictionary, including a specialised one, are the different types of users, user situations and user needs related to them, in line with the theory of lexicographic functions (Bergenholtz and Tarp 1995; Tarp 2008). This is one of the lexicographic theories which is very influential in pedagogical lexicography, including specialised pedagogical lexicography.

As for users, specialised dictionaries have a more limited target audience than general dictionaries. According to Bergenholtz and Tarp (1995), their user type is decided based on user's mother language, level of encyclopedic knowledge, and native- and foreign-language competence. Applying these criteria, the authors identify four major user types for specialised dictionaries: experts with a high level of encyclopedic and foreign language competence, experts with a high level of encyclopedic competence and low level of foreign language competence, laypersons with a low level of encyclopedic competence and foreign language competence, and layperson with a low level of encyclopedic competence and a high level of foreign language competence. Some more types are added by Fuertes-Olivera and Arribas-Baño (2008), who, among these user types, identify the following: experts from the specific field, semi-experts, experts from related or other fields, interested laypeople who would like to read some books or periodicals from the field, LSP students, translators, interpreters, etc.

Tarp (2010) argues that there are many situations in which learners can benefit from specialised dictionaries — cognitive situations include systematic study of the specialised subject field and of problems related to the translation of specialised texts; communicative situations include reception and production of specialised texts in the mother tongue and in a foreign language, as well translation of specialised texts, while practical situations refer to various operative and interpretive situations.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

The mentioned user types have different needs in the mentioned different types of situations. These needs can be primary or function-related needs, which are the needs for information necessary to gain knowledge or solve a problem through using a dictionary, or they can be secondary or usage-related needs, which includes the need to know something about a specific dictionary and to know how to use it (Tarp 2008).

There are different classifications of LSP dictionaries but we will briefly mention two which are relevant for our paper. Based on their size, there are two basic types — *maximising* LSP dictionaries, which attempt at covering as much of a field's terminology as possible, and *minimising* LSP dictionaries, in which a portion of the terminology is covered, typically only the most frequent items (Bergenholtz and Tarp 1995). Another possible classification recognises LSP dictionaries containing field-specific *terms* only, as opposed to general words, and *hybrid* LSP dictionaries, which combine both specialist and general words (Campoy Camillo 2002; Bowker 2010).

LSP dictionaries for learners are a subtype of specialised dictionaries which are intended to assist users in learning about the terms and concepts used in a specific field, in one or more languages (Bowker 2010). Their purpose is to serve as auxiliary tools in the process of teaching and learning the language for specific purposes (Fuertes-Olivera and Arribas-Baño 2008). According to the mentioned theory of lexicographic functions, they are utility tools which assist learners in the process of learning LSP.

3.1 Corpus-based headword selection procedures for LSP dictionaries

The process of headword selection is central in learner's lexicography (Xue and Tarp 2018), given that "dictionaries only function if they contain appropriate data," Nielsen (2018: 79). In this process, the three main questions that need to be posed refer to the size of the headword list, criteria and principles guiding their selection, and the empirical basis that their selection relies on (Tarp 2008). Tarp (2008) further suggests that headwords can be selected based on three sources, i.e. by means of introspection, using available descriptions in various publications (dictionaries, textbooks, etc.), and based on corpora. Building corpora as part of the preparatory stage for headword selection for LSP dictionaries is significant (Nkomo 2008: 105). Having compared corpus-based and intuition-based approaches, Verlinde and Selva (2001: 597) argue that it is the corpus-based lexicography that gives the "strong and necessary empirical evidence to the lexicographer's personal intuition", but they also note that intuition still remains helpful in filling in the gaps in cases when corpora are not balanced.

As said earlier, Bowker (2010: 166) argues that the use of corpus linguistic methods has been rather slow to take hold in the creation of specialised dictionaries, on account of the fact that not so many specialised corpora are available. Specialised corpora used for making dictionaries also tend to be relatively small, especially in comparison with the mega-corpora used for producing

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

general dictionaries. Bowker (2010) cites the example of the specialised dictionary *Dictionnaire d'apprentissage du français des affaires* (DAFA) as a commendable example, given that it was based on a corpus of 25 million words. Taking into account the latest technological developments, recently, the compilation of such, relatively large, corpora has become much less of an issue.

The mentioned theory of lexicographic functions (Bergenholtz and Tarp 1995; Tarp 2008) suggests that headwords should be selected according to user's needs. When selecting headwords based on corpora, this, among other things, practically means that it is the user needs which govern the selection of texts which will enter such corpora. To illustrate how this can work in practice, we will briefly note how headwords for a Spanish accounting dictionary were selected (Fuertes-Olivera et al. 2013). Thus, following the mentioned function theory and the principle of *relevance*, the authors created a list of around 6,000 accounting texts, based on which three experts in accounting and one lexicographer derived a stock of around 3,000 terms. Of the corpus-linguistic methods applied in this processing of the corpus, the authors calculated the word *frequencies* in their corpus, to inform their decisions of which terms to include in their specialised dictionary. They also used the Internet as a corpus and performed Google searches using particular word strings to extract additional 1,000 terms. Finally, 2,000 more terms were added through intensive reading of basic accounting texts. Such a hybrid approach was applied so as to ensure that the principle of relevance is adhered to. The authors argue and add that future updates of the term stock will be done by additionally analysing the log-files related to the online use of this dictionary (Fuertes-Olivera et al. 2013).

Other authors, too, mention applying the principle of frequency as one of the key steps taken in the process of selecting headwords for dictionaries (cf. Campoy Cubillo 2002; Hanks 2012; Rundell and Kilgarriff 2011). This criterion provides "solid empirical evidence for the occurrence of a word in actual language" (Xue and Tarp 2018). At the same time, they also argue that frequency may be misleading in some specialised fields which are updated constantly, such as accounting (Fuertes-Olivera and Nielsen 2011). Rundell and Kilgarriff (2011) rightly mention the fact that frequency is not a good selection criterion for extracting multiword items as candidates for headword lists. Likewise, Nielsen (2018: 81-82) suggests that frequency solely cannot guarantee that all relevant words will be selected, but that it should be used as a basis for the further selection process.

In some LSP dictionary compilation projects, similar to the methodology used in the production of LSP word lists, frequency is combined with additional corpus-linguistic methods — thus, for instance, Khumalo (2015) and Đurović (2021) also use keyness; however, they do not ensure that the corpus contains equal shares of various subdisciplines of the field which it represents and, consequently, they do not apply the range filter. Some LSP dictionary compilers additionally use a more innovative, pattern-based approach (Kruse and Heid 2021).

Frequency and relevance are suggested as two major criteria in Xue and

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Tarp too (2018). However, Tarp (2008) warns against the exalted status given to corpora and corpus-linguistic methods by certain lexicographers, arguing that corpora, however large they may be, can still be unrepresentative, and that the criteria of *relevance* and *systematicity* also need to be taken into account.

What may be deduced from these various accounts is that corpora play an important role when selecting headwords for specialised dictionaries, and that word frequencies in a corpus can significantly inform the process of headword selection.

4. Comparison of corpus-based headword selection procedures for LSP word lists and LSP dictionaries

As we have seen, headword selection procedures for both LSP word lists and LSP dictionaries are guided by the needs of their users. The chief users of LSP word lists are LSP learners. LSP word lists are also used by LSP teachers and LSP material developers but, again, to the benefit of their end users — LSP learners. When it comes to the users of LSP dictionaries, as noted earlier, LSP learners make up an important category among them, however, many more categories of users are possible as well (e.g. translators, semi-experts, experts from other fields, etc.). This basic distinction in the types of users of the two products — LSP word lists and LSP dictionaries, has implications for how headwords are selected as part of their compilation procedures.

When comparing corpus-based headword selection procedures for LSP word lists and LSP dictionaries, we can see that the former are compiled using corpus-linguistics methods almost exclusively, whereas a greater complexity of methods is used for the latter. A significant part of this difference may be explained by the respective homogeneity and heterogeneity of the end users of the two products, as explained above.

The corpora from which LSP word lists are derived are rather large and typically well-structured and balanced, as we have seen. The details regarding their make-up are usually presented very precisely and transparently in the scholarly papers on LSP word lists, as well as given central prominence in them. On the other hand, the descriptions of corpora used for developing headword lists for LSP dictionaries are usually not presented in such details and, typically, in the papers describing these projects relatively little space is devoted to the process of term extraction. In addition, equal representation of various subfields is rarely ensured in them. LSP word lists compilers argue that this is a good practice which allows that the frequencies of the terms obtained to reflect all subfields equally, and we tend to agree here. An implication from this comparison is that LSP lexicographers might invest this type of effort into compiling corpora from which they intend to extract terms. Moreover, given that many useful and balanced corpora have already been produced as part of LSP word-list research, some of these could be used for making LSP dictionaries as well.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Both compilers of LSP word lists and compilers LSP dictionaries use frequency as a major criterion for deciding which words should enter their products. In the process of producing LSP word lists, compilers typically either follow the cut-off points used in seminal research (such as Coxhead 2000) or, more frequently nowadays, the cut-off points are governed by the coverage achieved with the obtained word list, a coverage that allows for a certain threshold of reading or listening comprehension to be met.

As for LSP dictionaries, in the literature we have not encountered detailed arguments around the chosen thresholds. The size of LSP dictionaries, in theory, should be governed by the user needs (even though there are always practical and financial constraints to LSP dictionary projects) (Tarp 2008). However, so far, no method of quantifying them has been developed yet (and might not be, given the complexities involved).

Research and projects involving LSP dictionaries frequently mention that frequency cannot be the sole criterion for selecting headwords, usually citing *relevance* as another major criterion to be applied, which, however, is much more difficult to define and employ. Likewise, as we have seen in the LSP word-list research, the criterion of simple frequency is also never applied as the sole criterion. Additional criteria may be applied as well, although these are also based on frequency to some extent. Thus, an important criterion for selecting headwords for LSP word lists is that of *specialised occurrence*, as presented earlier, applied by excluding words which are highly frequent in general, reference corpora (typically 2,000 to 3,000 most frequent words in the case of English). Academic words can also be excluded, to ensure more technicality. Another criterion is that of *range* — applying this filter ensures that a word appears in a sufficient number of a discipline's subfields, so that it is equally valuable across that discipline, and not more valuable for some subspecialisations and less valuable for others. To apply this criterion, however, one needs a corpus with equal subsections from the various subfields, as argued above. If the required structure of the corpus is not achieved, various *dispersion* thresholds can be applied. These criteria for guiding term extraction are rarely used when compiling headword lists for LSP dictionaries.

One more criterion frequently mentioned when compiling LSP word lists is that of *keyness*, which is relatively easy to apply as no special make-up of the corpus is needed for it. As explained earlier, the frequency of the words in a specialised corpus is compared against that featured in a reference corpus and so the words found to be much more frequent in that specialised corpus are identified as terms. As we have seen, this criterion is sometimes used when extracting terms for LSP dictionaries as well.

Very often, the mentioned additional criteria are used in combination when compiling LSP word lists. LSP word list compilers argue that applying them, in addition to simple frequency, ensures that the headwords selected are indeed *relevant*. The notion of *relevance* is more difficult to define for a product such as an LSP dictionary given its rather heterogeneous target audience; how-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

ever, applying at least some of the forementioned filters could help facilitate and automate that process.

The mentioned filters used for obtaining LSP word lists have been found deficient, however, when it comes to extracting multi-word units and collocations and, in fact, none of the word lists presented here contain such items. This is a major drawback to LSP word lists in general and a limitation that should be borne in mind if one were to apply some of the said methods for selecting preliminary headword lists for LSP dictionaries. Still, the ease with which most of the presented filters can be applied certainly recommends them for use in combination with other methods.

Once an LSP word list is obtained via corpus linguistic methods, the work of the LSP word list compiler is either completed or almost completed in most cases, whereas much more work remains for a lexicographer compiling a headword list for their LSP dictionary.

The principle of *systematicity* is hardly ever applied to the LSP word lists obtained via corpus-linguistic methods. For instance, the Science List (Coxhead and Hirsh 2007) contains names of some common chemical elements (such as *oxygen, potassium*, etc.), while the names of other common elements are not mentioned (such as *sulfur,* for instance); it is debateable whether the word *sulfur* is less useful to a science student learning English than the word *potassium*, for instance. Moreover, the Science List includes the word *chloride*, however, it does not include the name of the chemical element whose negatively charged ionic form it represents — *chlorine*. Thus, in general, word-list makers rely, perhaps too much, on automated procedures and avoid discussing these types of issues. As opposed to that, in LSP dictionary research and projects, systematicity is one of the central principles guiding the creation of headword lists. Observing this principle when developing LSP word lists, we argue, could improve them, as the illogicalities of the types exemplified above typically stem from the imperfections of the corpus (in this case, the over-presence of texts mentioning the names of some particular chemical elements) and ought to be corrected when noticed. We would argue that, however large, well-structured and balanced a corpus may be, it will always suffer from some imperfections and cannot be trusted entirely.

When finalised, LSP word lists are sometimes subjected to validation in additional corpora (not the ones they were derived from), to test how much coverage they would have in new texts. Validation, although effort- and time-consuming, is a commendable step to be taken, in our opinion. The frequency of preliminary, candidate headword lists for LSP dictionaries, could also be checked in additional specialised corpora, so as to, perhaps, rule out some candidate terms which in validation corpora feature significantly lower frequencies as opposed to that from the first corpus.

In developing LSP word lists, experts from the specialist fields are almost never involved, while they are always involved in compiling LSP dictionaries. This step is usually skipped in the making of modern word lists, given that

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

several automatic corpus-linguistic filters have already been applied. Although this is a demanding step, involving experts in the creation of any LSP product is advisable.

5. Conclusion

In this paper, we presented most modern LSP word lists and commented on how they were created. We also discussed corpus-based headword selection procedures for LSP dictionaries. A number of both similarities and differences were found in the two selection procedures and it was noted that both of them could, in some ways, benefit from being mutually informed.

On the one hand, more effort could be invested in the creation of LSP corpora, in terms of their size, make-up and balance, and also more corpus-linguistic selection procedures could be applied when compiling headword lists for LSP dictionaries than is currently typically the case, to facilitate the process. More transparency and precision when reporting on the corpora used and the corpus-linguistic methods applied for compiling headwords lists for LSP dictionaries is also advised. Lists obtained should also be validated in additional corpora, when possible.

On the other hand, the creation of LSP word lists could be improved by applying additional non-corpus linguistic methods in their compilation, which is necessary to eliminate the illogicalities stemming from imperfectly balanced corpora, as well as to add the necessary multi-word units to them.

Another observation that imposes itself from the comparison made in this paper is that the compilation and study of word lists remain atheoretical, while at least one strand of LSP dictionaries research has strong theoretical foundations. As we conclude this paper, we will ask the reader and ourselves if, perhaps, the moment has arrived that the field of word-list compilation and research be supported by a theory similar to that of the theory of lexicographic functions.

Endnote

1. A word family includes the headword with all its inflected and derived forms (for instance, *suggest, suggests, suggested, suggesting, suggestion, suggestions*).

References

Bergenholtz, H. and S. Tarp (Eds.). 1995. *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries.* Amsterdam: John Benjamins.

Bowker, L. 2010. The Contribution of Corpus Linguistics to the Development of Specialised Dictionaries for Learners. Fuertes-Olivera, P. (Ed.). 2010. *Specialised Dictionaries for Learners*: 155-168. Berlin: De Gruyter.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Brezina, V. and D. Gablasova. 2013. Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics* 36(1): 1-22.

Browne, C., B. Culligan and J. Phillips. 2013a. *New General Service List Project*. http://www.newgeneralservicelist.org/. Accessed 1 September 2019.

Browne, C., B. Culligan and J. Phillips. 2013b. *New Academic World List*. http://www.newgeneralservicelist.org/nawl-new-academic-word-list/. Accessed 1 September 2019.

Burkhanov, I.Y. 1998. *Lexicography: A Dictionary of Basic Terminology*. Rzeszów: Wydawnictwo. Wyższej Szkoły Pedagogicznej w Rzeszowie.

Campion, M.E. and W.B. Elley. 1971. *An Academic Vocabulary List*. Wellington: New Zealand Council for Educational Research.

Campoy Cubillo, M.C. 2002. Dictionary Use and Dictionary Needs of ESP Students: An Experimental Approach. *International Journal of Lexicography* 15(3): 206-228.

Chen, Q. and G. Ge. 2007. A Corpus-based Lexical Study on Frequency and Distribution of Coxhead's AWL Word Families in Medical Research Articles. *English for Specific Purposes* 26: 502-514.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly* 34(2): 213-238.

Coxhead, A. 2018. *Vocabulary and English for Specific Purposes Research: Quantitative and Qualitative Perspectives.* London: Routledge.

Coxhead, A. and D. Hirsch. 2007. A Pilot Science-specific Word List. *Revue Française de Linguistique Appliquée* 12(2): 65-78.

Dang, T.N.Y. 2018. A Hard Science Spoken Word List. *ITL — International Journal of Applied Linguistics* 169(1): 44-71.

Dang, T.N.Y., A. Coxhead and S. Webb. 2017. The Academic Spoken Word List. *Language Learning* 67(4): 959-997.

Dang, T.N.Y. and S. Webb. 2016. Evaluating Lists of High-frequency Words. *ITL — International Journal of Applied Linguistics* 167(2): 132-158.

Đurović, Z. 2021. Corpus Linguistics Methods for Building ESP Word Lists, Glossaries and Dictionaries on the Example of a Marine Engineering Word List. *Lexikos* 31: 259-282.

Esandi-Baztan, M.A. and P.A. Fuertes-Olivera. 2020. Selecting an Initial Lemma List in Specialized Lexicography: A Case Study in the Field of Graphic Engineering. *Lexikos* 30: 57-89.

Fraser, S. 2007. Providing ESP Learners with the Vocabulary They Need: Corpora and the Creation of Specialized Word Lists. *Hiroshima Studies in Language and Language Education* 10: 127-143.

Fuertes-Olivera, P.A. and A. Arribas-Baño. 2008. *Pedagogical Specialised Lexicography: The Representation of Meaning in English and Spanish Business Dictionaries.* Amsterdam/Philadelphia: John Benjamins.

Fuertes-Olivera, P.A. and S. Nielsen. 2011. Online Dictionaries for Assisting Translators of LSP Texts: The Accounting Dictionaries. *International Journal of Lexicography* 25(2): 191-215.

Fuertes-Olivera, P.A., H. Bergenholtz, P. Gordo-Gómez, S. Nielsen, M. Niño-Amo, Á. de los Ríos, A. Sastre-Ruano and M. Velasco-Sacristán. 2013. From Theory to Practice: The Selection of Spanish Lemmas in the Accounting Dictionaries. *Fachsprache: International Journal of Specialized Communication* 35(1–2): 25-42.

Gardner, D. and M. Davies. 2014. A New Academic Vocabulary List. *Applied Linguistics* 35(3): 305-327.

Ghadessy, M. 1979. Frequency Counts, Wordlists, and Material Preparation: A New Approach. *English Teaching Forum* 17(1): 24-27.

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

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Hernandez, M.M. 2017. *Comparing the AWL and AVL in Textbooks from an Intensive English Program.* M.A. thesis. Provo, USA: Brigham Young University.

Hsu, W. 2013. Bridging the Vocabulary Gap for EFL Medical Undergraduates: The Establishment of a Medical Word List. *Language Teaching Research* 17(4): 454-484.

Hsu, W. 2014. Measuring the Vocabulary Load of Engineering Textbooks for EFL Undergraduates. *English for Specific Purposes* 33: 54-65.

Jansson, H., S. Johansson Kokkinakis, J. Ribeck and E. Sköldberg. 2012. A Swedish Academic Word List: Methods and Data. Vatvedt Fjeld, R. and J.M. Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress 7–11 August, 2012, Oslo, Norway:* 955-960. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Jin, N.Y. et al. 2013. Development of the Engineering Technology Word List for Vocational Schools in Malaysia. *International Education Research* 1(1): 43-49.

Johannessen, J.M., A. Saidi and K. Hagen. 2016. Constructing a Norwegian Academic Wordlist. Calzolari, N. et al. (Eds.). 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation, May 23–28, 2016, Portorož, Slovenia (LREC'16):* 1457-1462. Portorož: European Language Resources Association.

Kamrotov, M., E. Talalakina and D. Stukal. 2022. Technical Vocabulary in Languages for Special Purposes: The Corpus-based Russian Economics Word List. *Lingua* 273: 103326.

Khani, R. and K. Tazik. 2013. Towards the Development of an Academic Word List for Applied Linguistics Research Articles. *RELC Journal* 44(2): 209-232.

Khany, R. and B. Kalantari. 2021. Accounting Academic Word List (AAWL): A Corpus-based Study. *Journal of Foreign Language Teaching and Translation Studies* 6(1): 35-58.

Khumalo, L. 2015. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary. *Lexikos* 25: 495-506.

Kokkinakis, S.J., E. Sköldberg, B. Henriksen, K. Kinn and J.B. Johannessen. 2012. Developing Academic Word Lists for Swedish, Norwegian and Danish — A Joint Research Project. Vatvedt Fjeld, R. and J.M. Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress 7–11 August, 2012, Oslo, Norway:* 563-569. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Konstantakis, N. 2007. Creating a Business Word List for Teaching Business English. *ELIA* 7: 79-102.

Kruse, T. and U. Heid. 2021. Lemma Selection and Microstructure: Definitions and Semantic Relations of a Domain-Specific e-Dictionary of the Mathematical Field of Graph Theory. Gavriilidou, Z., M. Mitsiaki and A. Fliatouras (Eds.). 2021. *Proceedings of the XIX Euralex International Congress, Alexandroupolis, Greece, 7–11 September 2021. Volume 1*: 227-233. Komotini, Greece: European Association for Lexicography.

Kwary, D.A. and A.F. Artha. 2017. The Academic Article Word List for Social Sciences. *MEXTESOL* 41(4): 1-11.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography.* Second edition. New York/Cambridge: Cambridge University Press.

Lei, L. and D. Liu. 2016. A New Medical Academic Word List: A Corpus-based Study with Enhanced Methodology. *English for Academic Purposes* 22: 42-53.

Liu, J. and L. Han. 2015. A Corpus-based Environmental Academic Word List Building and Its Validity Test. *English for Specific Purposes* 39: 1-11.

Lynn, R.W. 1973. Preparing Word Lists: A Suggested Method. *RELC Journal* 4: 25-32.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Martínez, I.A., S.C. Beck and C.B. Panza. 2009. Academic Vocabulary in Agriculture Research Articles: A Corpus-based Study. *English for Specific Purposes* 28: 183-198.

McQuillan, J. 2020. Harry Potter and the Prisoners of Vocabulary Instruction: Acquiring Academic Language at Hogwarts. *Reading in a Foreign Language* 32(2): 122-142.

Mihindou, G.-R. 2004. Some Features of Monolingual LSP Dictionaries. *Lexikos* 14: 118-136.

Miller, L. 2014. English for Science and Technology. Bhatia, V. and S. Bremner (Eds.). 2014. *The Routledge Handbook of Language and Professional Communication*: 332-468. Oxon/New York: Routledge.

Minshall, D.E. 2013. *A Computer Science Word List.* M.A. thesis. Swansea, Wales: Swansea University.

Moini, R. and Z. Islamizadeh. 2016. Do We Need Discipline-specific Academic Word Lists? Linguistics Academic Word List (LAWL). *Journal of Teaching Language Skills* 35(3): 65-90.

Mudraya, O. 2006. Engineering English: A Lexical Frequency Instructional Model. *English for Specific Purposes* 25(2): 235-256.

Nagy, W.E. and R.C. Anderson. 1984. How Many Words are There in Printed School English? *Reading Research Quarterly:* 304-330.

Nation, I.S.P. 2013. *Learning Vocabulary in Another Language.* Second edition. Cambridge: Cambridge University Press.

Newman, J.A. 2016. *A Corpus-based Comparison of the Academic Word List and the Academic Vocabulary List.* M.A. thesis. Provo, USA: Brigham Young University.

Nielsen, S. 2018. LSP Lexicography and Typology of Specialised Dictionaries. Humbley, J., G. Budin and C. Laurén (Eds.). 2018. *Languages for Special Purposes: An International Handbook:* 71-95. Berlin/Boston: De Gruyter.

Nkomo, D. 2008. *Towards a Theoretical Model for LSP Lexicography in Ndebele with Special Reference to a Dictionary of Linguistic and Literary Terms.* M.A. thesis. Stellenbosch: Stellenbosch University.

Praninskas, J. 1972. *American University Word List*. London: Longman.

Ribeck, J., H. Jansson and E. Sköldberg. 2014. Från aspekt till övergripande — en ordlista över svensk akademisk vokabulär. *Nordiske studier i leksikografi* 12: 370-384.

Rundell, M. and A. Kilgarriff. 2011. Automating the Creation of Dictionaries: Where Will it All End? Meunier, F. et al. (Eds.). 2011. *A Taste for Corpora. In Honour of Sylviane Granger*: 257-282. Amsterdam/Philadelphia: John Benjamins.

Savický, P. and J. Hlavácová. 2002. Measures of Word Commonness. *Journal of Quantitative Linguistics* 9(3): 215-231.

Talalakina, E., D. Stukal and M. Kamrotov. 2020. Developing and Validating an Academic Vocabulary List in Russian: A Computational Approach. *Modern Language Journal* 104(3): 618-646.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Berlin/New York: Max Niemeyer.

Tarp, S. 2010. Functions of Specialised Learners' Dictionaries. Fuertes-Olivera, P. (Ed.). 2010. *Specialised Dictionaries for Learners*: 39-53. Berlin/New York: De Gruyter.

Todd, R.W. 2017. An Opaque Engineering Word List: Which Words Should a Teacher Focus on? *English for Specific Purposes* 45: 31-39.

Tongpoon-Patanasorn, A. 2018. Developing a Frequent Technical Words List for Finance: A Hybrid Approach. *English for Specific Purposes* 51: 45-54.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Valipouri, L. and H. Nassaji. 2013. A Corpus-based Study of Academic Vocabulary in Chemistry Research Articles. *Journal of English for Academic Purposes* 12(4): 248-263.

Verlinde, S. and T. Selva. 2001. Corpus-based Versus Intuition-based Lexicography: Defining a Word List for a French Learners' Dictionary. Rayson, P., A. Wilson, T. McEnery, A. Hardie and S. Khoja (Eds.). 2001. *Proceedings of the Corpus Linguistics 2001 Conference*: 594-598. Lancaster: University Centre for Computer Corpus Research on Language, Lancaster University.

Vongpumivitch, V., J.-Y. Huang and Y.-C. Chang. 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: 33-41.

Wang, J., S. Liang and G. Ge. 2008. Establishment of a Medical Academic Word List. *English for Specific Purposes* 27(4): 442-458.

Ward, J. 2009. A Basic Engineering English Word List for Less Proficient Foundation Engineering Undergraduates. *English for Specific Purposes* 28(3): 170-182.

West, M. 1953. *A General Service List of English Words*. London: Longman, Green & Co.

Xue, G. and P. Nation. 1984. A University Wordlist. *Language Learning and Communication* 3(2): 215-229.

Xue, M. and S. Tarp. 2018. Towards Chinese Learner's Dictionaries for Foreigners Living in China: Some Problems Related to Lemma Selection. *Lexikos* 28: 384-404.

Yang, M.-N. 2015. A Nursing Academic Word List. *English for Specific Purposes* 37: 27-38.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Appendix

Author	List	Number of words	Corpus/ Coverage in the corpus	Methodology
Mudraya (2006)	Engineering English word list	1,260 word families	2 mill. (13 text-books) / not given	— frequency of the word of at least 100
Coxhead and Hirsh (2007)	Pilot science list	318 word families	1,761,380 words (14 disciplines) / 3.79%	— exclusion of the GSL & the AWL words — occurrence in at least half the disciplines — frequency of at least 50 in the corpus — the dispersion factor of at least 35 — exclusion of proper nouns, symbols and abbreviations
Fraser (2007)	Pharmacology word list (PWI)	601 word families	185,000 (51 research articles (RAs))	— exclusion of the GSL and the AWL words — exclusion of proper names, nationalities, numbers, abbreviations and acronyms — occurrence in at least 6 RAs — frequency of 10 or more and occurrence in at least 2 RAs
Konstantakis (2007)	Business word list for undergraduates (BWL)	560 word families	600,000 (Published Materials Corpus, compiled by Nelson (2000); 33 course books) / 2.79%	— exclusion of the GSL and the AWL words — occurrence in at least 5 course books — exclusion of proper names, numericals, Latin words, nationalities, acronyms, interjections and abbreviations — frequency of at least 10 in the corpus
Wang, Liang and Ge (2008)	Medical academic word list (MAWL)	623 word families	1,093,000 words (288 research articles from 96 journals) / 12.24%	— exclusion of the GSL words — occurrence in at least half the disciplines — frequency of at least 30 in the corpus — two English for Medicine professors consulted for differentiating between technical and academic vocabulary
Ward (2009)	Basic Engineering list (BEL)	299 word types	271,000 words (25 textbooks in engineering) / 16.4%	— exclusion of function words — definition of word as a word type — frequency of at least 5 in each of the 5 subsections
Martínez, Beck and Panza (2009)	A reduced AWL for agriculture	92 word families	826,416 (218 RAs) / coverage not given	— inclusion of the academic words above the mean for academic words

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Vongpumivitch, Huang and Chang (2009)	Applied linguistics word list	475 AWL word forms and 128 non-AWL content word forms	1.5 mill. (200 RAs from 5 journals) / 8.6% for the 475 AWL words + 2.8% for the non-AWL words	— frequency of at least 50 in the corpus — occurrence of at least 5 times in at least 5 journals — exclusion of the GSL, function words and abbreviations
Jin et al. (2013)	Engineering technology word list (ETWL)	313 word lists	124,584 words (2 textbooks) / 8.7%	— exclusion of the GSL and the AWL words — to be defined as technical, a word had to appear as an entry in an online engineering dictionary — cross-checked by two experts from the field to make sure the words were technical and semi-technical
Khani and Tazik (2013)	Applied linguistics academic word list	773 word families	1,553,450 words (240 RAs from 12 journals) / 12.48%	— exclusion of the GSL words — frequency of at least 50 in the corpus — occurrence of at least 4 times in at least half the journals
Valipouri and Nassaji (2013)	Chemistry Academic word list (CAWL)	1,400 word families	4 mill. (1,185 RAs from 38 journals) / 81.18%	— frequency of at least 114 in the corpus — frequency of at least 10 in all subsections — exclusion of abbreviations and function words — three chemistry professors excluded technical words using a rating scale
Hsu (2013)	Medical word list (MWL)	595 word families	15 mill. (155 medical textbooks) / 10.72%	— exclusion of the BNC 3000 — occurrence in more than half the subsections — frequency of at least 863 in the corpus
Minshall (2013)	Computer science word list	433 word families	3,661,337 tokens (RAs and conference proceedings from 10 subdisciplines) / 6%	— outside the GSL and the AWL — occurrence in at least half the subdisciplines — occurrence of at least 80 in the corpus
Hsu (2014)	Engineering English word list (EEWL)	729 word families	4.57 mill. (100 engineering textbooks) / 14.3%	— exclusion of the BNC 2000 — occurrence in all subsections — occurrence in at least 95 out 100 textbooks — frequency of at least 288 times in the corpus — exclusion of exclamations, interjections and proper names
Liu and Han (2015)	Environmental academic word list (EAWL)	458 word families	862,242 tokens (200 RAs) / 15.43%	— exclusion of the GSL words — frequency of at least 30 in the corpus — occurrence in at least 8 out 10 subsections

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1713 (Article)

Yang (2015)	Nursing academic word list	676 word families	1,006,934 words (252 RAs) / 13.64%	— exclusion of the GSL words range — occurrence in at least half the sub-sections — frequency of at least 33 in the cor-pus
Lei and Liu (2016)	Medical Academic Vocabulary List (MAVL)	819 lem-mas	6.2 mill. (760 medi-cal RAs from 38 journals (MAEC corpus + 1 three-volume textbook (MAET corpus)), 19.44% in MAEC and 20.18% in MAET	— frequency of 28.57 per 1 mill — ratio of at least 1.5 (at least 50% higher frequency in the academic corpus than in a non-academic corpus) — occurrence of 20% of the expected frequency in at least half the sub-sections — dispersion of at least 0.5 (Jullian's D) — no lemma should occur more than 3 times the expected frequency in more than any 3 out 21 subsec-tions — special meaning criterion checked via 2 medical dictionaries
Moini and Islamizadeh (2016)	Linguistics academic word list	1,263 word families	4 mill. (700 RAs from 4 subdisci-plines)/ not given	— a frequency of at least 114 in the corpus — occurrence of at least 10 times in each subdiscipline
Todd (2017)	Opaque engineering word list	186 items	1.15 mill. (27 text-books) / not given	— exclusion of non-words, abbrevia-tions, and function words — occurrence in at least 15 out 27 textbooks — similar word types combined under one entry
Kwary and Artha (2017)	Academic article word list for social sciences	350 word families	1,040,259 tokens (122 RAs)	— outside the GSL — occurrence in all the 6 subdisci-plines — frequency
Dang (2018)	Hard science spoken word list	1,595 word families	6.5 mill. words of spoken language from 12 disciplines	— occurrence in at least half the disciplines and both subsections of the corpus — frequency of at least 175 in the corpus — dispersion (DP value below 0.6)
Tongpoon-Patanasorn (2018)	Frequent technical word list for finance	569 word families	2 mill. words (books, journals, websites and news-papers)	— keyness — rating scale, completed by experts
Khany and Kalantari (2021)	Accounting academic word list	658 word families	29.5 mill. words (2,098 accounting RAs)	— outside the GSL — frequency of at least 839 in the corpus and at least 50 in each journal
Kamrotov, Talalakina and Stukal (2022)	Russian economics word list	507 lemmas	10.5 million words (economics RAs and mass media economics articles)	— keyness (1.5 more frequent in the economics corpus than in other corpora) — degree of dispersion over 0.25 — minimum frequency of 10 in the corpus