http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

An Investigation of Corpus Contributions to Lexicographic Challenges over the Past Ten Years

Esra M. Abdelzaher, *University of Debrecen, Doctoral School of Linguistics; and University of Debrecen, Institute of English and American Studies, Hungary (esra.abdelzaher@gmail.com)*

Abstract: This study explores the role of corpus linguistics in addressing lexicographic challenges over the past ten years. It examines 25 studies published from 2012 to 2021 that employed corpus solutions to solve lexicographic problems. Challenging tasks are either relevant to the macrostructure or the microstructure of dictionaries. In the past decade, lexicographers made extensive use of corpus tools to create dictionaries and improve existing ones. This effort included compiling general and specialized headword lists, identifying idiom variations, detecting collocational patterns, identifying and ordering words senses and differentiating polysemous words and near-synonyms. Frequency, keyword extraction, and word sketches are among the most effective aids for lexicographers. According to the investigated studies, almost all dictionaries may benefit from corpus tools at the macro and microstructure levels.

Keywords: ACCESS STRUCTURE, CORPUS TOOLS, COLLOCATIONAL ANALYSIS, LEXICOGRAPHIC CHALLENGES, MICROSTRUCTURE, MACROSTRUCTURE, SENSE DELINEATION, SPECIALIZED DICTIONARIES, TERM EXTRACTION

Opsomming: 'n Ondersoek na die bydrae wat korpusse oor die afgelope tien jaar tot leksikografiese uitdagings gemaak het. Hierdie studie ondersoek die rol van die korpuslinguistiek in die benadering tot leksikografiese uitdagings oor die afgelope tien jaar. Dit ondersoek 25 studies wat vanaf 2012 tot 2021 gepubliseer is. Uitdagende take is óf relevant tot die makrostruktuur óf tot die mikrostruktuur van woordeboeke. In die afgelope dekade het leksikograwe uitgebreid gebruik gemaak van korpusgereedskap om woordeboeke te skep en bestaandes te verbeter. Dit het die samestelling van algemene en gespesialiseerde lemmalyste, die identifisering van idioomvariasies, die bepaling van kollokasionele patrone, die identifisering en ordening van woordbetekenisse en die onderskeiding van polisemiese woorde en ampersinomieme ingesluit. Frekwensie, sleutelwoordonttrekking en woordbeskrywings is van die mees effektiewe hulpmiddels vir leksikograwe. Volgens die studies wat ondersoek is, kan byna alle woordeboeke baat by korpusgereedskap op makro- sowel as mikrostruktuurvlakke.

Sleutelwoorde: TOEGANGSTRUKTUUR, KORPUSGEREEDSKAP, KOLLOKASIONELE ANALISE, LEKSIKOGRAFIESE UITDAGINGS, MIKROSTRUKTUUR, MAKROSTRUKTUUR, BETEKENISOMSKRYWING, GESPESIALISEERDE WOORDEBOEKE, TERM-ONTTREKKING

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

1. Introduction

The field of lexicography has witnessed significant changes over the years. In his description of the significant influences in the field over 2000 years, Hanks (2013) referred to computers and corpora as the second most influential factor, after the invention of printing, in lexicography. The new computational and corpus technology provided both lexicographers and users with innovative tools to compile and consult dictionaries.

The significant contributions of corpus tools in lexicography were first and most salient in the context of Monolingual Learner's Dictionaries (MLDs). As Sinclair (1992) stated, *Collins COBUILD English Language Dictionary* was the first to rely on corpus evidence in the compilation process. The compilers aimed to improve the understandability of the dictionary and provide a representative picture of modern English. Corpus evidence facilitated the inclusion of common senses and the exclusion of obsolete word forms and word senses. Also, the frequency of use was the main criterion in organizing information in the entry. The dictionary replaced the authoritative artificial examples with corpus-based examples of word use. It used complete actual sentences to explain the meaning and the typical use of a word.

The innovative corpus-driven approaches in the multiple editions of *COBUILD* were further discussed by lexicographers. For instance, Heuberger (2018) pointed to presenting the frequency of words as an initiative launched in the second edition of *COBUILD.* This initiative helped learners identify the words they should remember because they are likely to encounter them. Unlike most dictionaries, the online version of *COBUILD* presents frequency information for a large number of words. Moreover, it offers information about the frequency of using the word over the years. Distributional and frequency information is one of the major advantages of corpus methods.

Nowadays, the role of corpus tools is almost conventional in dictionary-making. The continuous update of corpora and their instant accessibility made them unmatched valuable resources for dictionary-makers (Abecassis 2007). Basically, corpus-analysis helped lexicographers take several decisions about the macro- and microstructure of dictionaries. The macrostructure of a dictionary refers to all entries of the lemmas in the wordlist. At the macrostructure level, lexicographers have to decide which words will be included in the dictionary based on the purpose and the target user. They may include or exclude specialized terms, jargon or loan words. Also, they should decide upon the lexical items that will be granted headword status, e.g., lemmas, inflected words or phrases (Atkins 2008). In this regard, corpus-based frequency, user needs and dictionary types have a significant role in the choice of words for lexicographical treatment. Although frequency lists are now essential to compile wordlists for dictionaries, user needs and dictionary types have more significant roles. For instance, a synonymy dictionary would discard a frequent word if it does not have a synonym (Bogaards 2013).

The microstructure level is form and sense-related. It is relevant to the

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

selection and presentation of information in the entry. Lexicographers may choose to include or discard specific pieces of etymological, external and internal information about a word. Internal word information is relevant to its morphology, orthography, semantics and phonological features. Internal information embraces both facts relevant to the word's form and sense. External information refers to the relation between a word and other words. It includes paradigmatic relations (e.g., POS and synonyms), syntagmatic relations (e.g., collocations), relational links (e.g., cross-references to derivational forms) and usage information (e.g., genre, dialect). Lexicographers need to find senses, split or lump and order them, and find lexicographically relevant information for each sense. They also have to select elaborative examples for each sense (Atkins 2008, Atkins and Rundell 2008).

There are several online and offline tools to conduct corpus analysis. The Sketch Engine software is probably the most effective and widely used by lexicographers. The introduction of word sketches presented lexicographers with sorted lexico-grammatical collocations of the target. Word sketches effectively summarize thousands of concordance lines and provide lexicographers with a report of the word's behavior (Heuberger 2018). This feature was efficiently employed in the compilation of *Macmillan English Dictionary for Advanced Learners* (MEDAL). In most cases, lexicographers could link a specific word sense to a particular collocational pattern (Hanks 2013).

In addition, the software has a random sampling option to extract a sample of the concordance for analysis. It also hosts a large number of corpora in several languages as well as the option of uploading the user's own corpus. The software has recently introduced an option to automatically select examples that better serve the purposes of lexicographers.

Offline tools include ANTConc which can be installed on computer devices and used offline. It processes files stored on the device and offers a variety of content analysis options. In addition to the basic frequency list compilation and n-gram clusters, Key Words In Context (KWIC) can be displayed in different modes through ANTConc. It also has several statistical methods for detecting collocations (Faaß 2018). WordSmith Tools offer, in addition to the standard content analysis options, part of speech concordancing for tagged corpora, manual lemmatizing of wordlists and plotting keywords. Relevantly, Yoshikoder is unique for its dictionary options. This offline software allows the search for nested lists of words collectively or individually. Users can add a dictionary that includes words referring to animals, for instance, subcategorize them into domestic and predators and search for the entire dictionary or the part relevant to domestic animals only. There are no limits on the number of added dictionaries or their sub-categories.

Although corpus tools have revolutionized the field of lexicography, some challenging tasks remain hard to address. Kilgarriff (1998) assessed the most difficult choices lexicographers had to make during the compilation of the *Longman Dictionary of Contemporary English*. In his short report, he considered the most complicated tasks the ones that lack clear rules or guidelines. He categorized

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

the tasks into analysis, i.e., tasks relevant to the pre-writing stage and the analysis of word behavior in context, and synthesis, i.e., tasks relevant to the content that will be finally presented in the dictionary. The most challenging task during the analysis stage was "splitting; identifying senses of a word" (Kilgarriff 1998: 53). There are no instructions or in-dictionary style guides on how to split or lump senses in an entry. Also, at that time, there was no valuable information on how to deal with this problem in books. Therefore, lexicographers used to, and perhaps still, depend on their experience and intuition. Accordingly, the task ranked second in the list of the most challenging tasks after choosing the words of a definition. The problematic tasks consist of the inclusion of headwords, treatment of Multiple Word Expressions (MWEs), selection or invention of examples, and dealing with grammar and inflected words, among others.

This study is concerned with the lexicographic tasks to which the contribution of corpora and corpus tools is most significant. It offers an investigation of the lexicographic studies that highlighted the role of the corpus in solving macro- and microstructure problems in the last decade. It addresses the following questions. In the past ten years,

1. What are the lexicographic challenges that were successfully addressed by corpus tools?
2. Which corpus options are highly contributing to improving the lexicographic practice?
3. Which types of dictionaries did benefit from corpus tools?

2. Methodology

The literature on the use of corpus tools in lexicography is voluminous. The present investigation used the flow diagram of PRISMA 2020 to search the literature in order to answer the previously mentioned questions. The PRISMA 2020 statement provides detailed guidelines for scholars aiming at investigating the literature to find answers for certain questions. It clarifies that scholars should clearly state the sources of data, the search terms and the date on which the databases were accessed. Scholars are required to precisely report the number of the retrieved data items (reflected in the identification step in the diagram) and should clarify if all of the results were examined to reach the set objective or not (screening step). After that, the selected articles should be examined according to the eligibility criteria (i.e., inclusion and exclusion) that best suit each investigation so that the inclusion of the selected articles can be justified (Page et al. 2021).

2.1 Search Procedure

The following databases were the primary sources of data: PubMed, Science Direct, Web of Science, Scopus, Mendeley, ERIH PLUS, PsycINFO, ProQuest, and Crossref. The searched keywords were "Corpus", AND "lexicograph*", OR "lexis" OR "lexicon*"AND "dictionar*". The search excluded "NLP" AND/OR

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

"survey" OR "computation*."

The time range was customized from 2012 to 2021. 358 articles were found. Using Mendeley software, the duplicated articles were deleted. Also, non-English results, book chapters and articles published before the specified date were filtered manually. The unique articles were 98. After excluding the author-focused articles, surveys and NLP-oriented articles, the total number of research articles became 73. Screening the titles and abstracts of all articles excluded studies that presented only theoretical argument, mentioned only lexicographic implications and discussed only the role of corpus-based dictionaries in pedagogical contexts. The screened articles amounted to 53. After reading the full manuscripts to be vetted for eligibility, articles that addressed the role of the corpus with no relevance to a lexicographic problem and articles that provided no analytic examples were excluded. The accepted analytical examples had to involve corpora or corpus tools with relevance to a lexicographic task. The articles that described the construction of a corpus tool without defining a particular lexicographic challenge or used audiovisual corpora are considered out of the scope of this study. The total number of the included article became 25 (Figure 1).

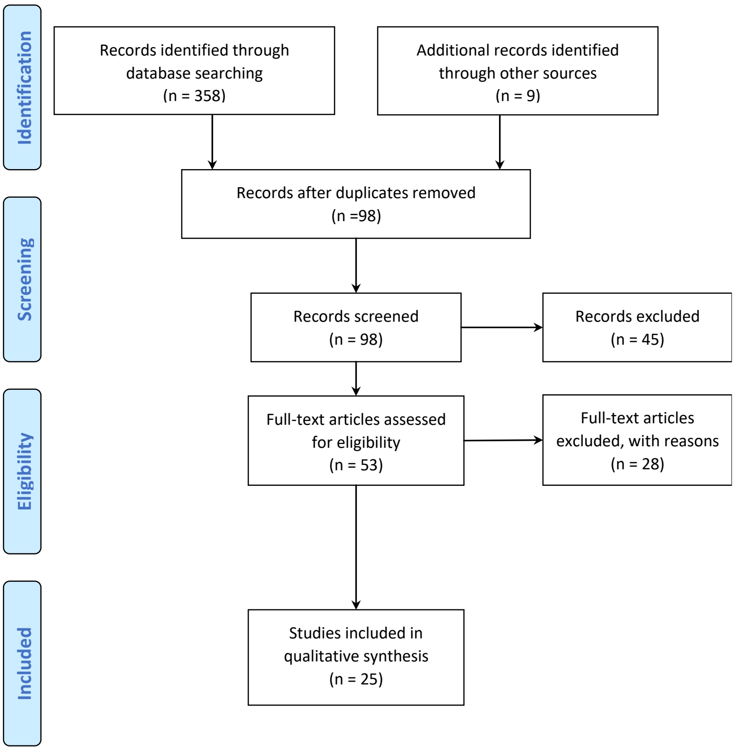
Identification
Records identified through database searching (n = 358)
Additional records identified through other sources (n = 9)
Screening
Records after duplicates removed (n =98)
Records screened (n = 98)
Records excluded (n = 45)
Eligibility
Full-text articles assessed for eligibility (n = 53)
Full-text articles excluded, with reasons (n = 28)
Included
Studies included in qualitative synthesis (n = 25)

Figure 1: PRISMA flow chart showing the search method

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

2.2 Eligibility criteria

2.2.1 Criteria of Inclusion

The inclusion criteria for this study were as follows:

1. Original articles that integrated corpus analysis to address a lexicographic task or challenge.
2. The included studies must analyze at least a single case study relevant to a particular lexicographic challenge.
3. All studies must offer a corpus-based solution, either primary or secondary.
4. All studies must use a lexicographic resource, either primary or secondary.

2.2.2 Criteria of Exclusion

The exclusion criteria for the present study were as follows:

1. Book reviews, review articles, book chapters
2. NLP-oriented research, e.g., report on corpus-tools or algorithms
3. Theoretical arguments only
4. Articles that are written in a non-English language
5. Non-textual corpora

3. Results

Only four out of the 25 analyzed studies used corpus tools to address a task at the macrostructure level, whereas 12 studies were mainly concerned with challenges at the microstructure level. In addition, ten studies used corpus tools to take decisions at both macro- and microstructure levels.

At the macrostructure level, lemma-related problems show considerable diversity. The use of corpus in the compilation of a lemma list as the first step towards a representative headword list is central to the macrostructure tasks. This task, in some cases, was further complicated by the type of language the lexicographer needs a lemma list for. For English, the task is less challenging because of the availability of a large tagged corpus and the poor-morphological system of the language. However, for endangered, less-resourced and agglutinative languages, the compilation of a lemma list is problematic. It usually depends on corpus-driven analysis that retrieves frequency wordlists and complements the process with a corpus-based analysis of the frequency and distribution of other words cited from previous dictionaries or suggested by lexicographers.

At the microstructure level, concordances (especially KWIC), word sketches,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

collocations and frequency were the corpus options employed by scholars to tackle sense-related tasks. The tools were proved to be effective in detecting and describing collocational patterns, identifying, ordering and describing senses, detecting polysemy and differentiating near-synonyms. Nevertheless, this corpus-based analysis was usually triangulated with qualitative and often theoretical analyses. Theories like Frame Semantics, Conceptual Metaphor Theory, Norms and Exploitations were usually present to motivate this corpus-based analysis in order to obtain the best results. These tools also improved the entries of colloquial and regional words in dictionaries.

Moreover, several studies focused on a linguistic aspect or phenomenon and adopted a corpus approach to detect, include and record it in dictionaries. Therefore, they dealt with challenges at the macro- and microstructure levels. For instance, the identification of specialized terms, through frequency and term extraction, and including them in dictionaries should be relevant only to the macrostructure of a dictionary. However, conducting a collocational analysis and relying on corpus citations of construct entries for specialized terms is associated with the microstructure. The same applies to the selection, inclusion and description of MWEs and phrases.

All of the analyzed studies used at least one lexicographic resource at the analysis stage. In this regard, almost all types of dictionaries were used either as the main project reported in the study (13 studies) or as a reference point for comparison with the corpus-based results (12 studies). The studies focused on monolingual, bilingual and multilingual dictionaries that mostly covered Germanic, Romance and Slavic languages. The typological spread of dictionaries included native, learners, specialized, general language, adult and children dictionaries. In addition to traditional dictionaries, lexical databases and innovative dictionaries such as the *Pattern Dictionary of English Verbs* (PDEV) were used (**Table 1** in the Appendix).

4. Discussion

From a psycholinguistic perspective, dictionary makers attempt to reflect the mental lexicon of a native speaker so that learners would benefit the most from dictionary entries, i.e., lemmas and senses (Anshen and Aronoff 1999, Atkins and Rundell 2008, Fillmore and Atkins 1992, Jorgensen 1990). However, compiling a headword list that is comprehensive is an unattainable goal given the varieties and dynamicity of natural languages (Atkins and Rundell 2008, Lew 2013). Therefore, lexicographers aim at collecting the most frequent words in a language. They assume that the most frequent words represent the core vocabulary in the lexicon and speakers are most likely to encounter them in everyday life.

4.1 Frequency lists and compiling wordlists for general language dictionaries

Collins COBUILD pioneered the use of corpus frequency to compile wordlists for monolingual dictionaries. Sinclair (1992) explained that the dictionary used a star-rated system to indicate the frequency of a headword in spoken and written English (according to corpus frequency). The online version of the dictionary still refers to the frequency of words using a single or five stars for rarest and most frequent words respectively. After this successful attempt, corpora have been widely used to compile dictionary wordlists.

De Schryver and Nabirye (2018a) displayed their successful use of corpora in building a dictionary for Bantu languages and used Lusoga as a case study. The standard case for lexicographers is generating a lemmatized wordlist based on a tagged general reference corpus. For Lusoga and less-resourced languages, tagged corpora and automatic lemmatizers are hardly available. Therefore, the author(s) had to depend only on a corpus-driven frequency list and apply several steps to convert it into a lemmatized one. They used *WordSmith Tools* to retrieve a frequency wordlist from a 1.7-million-word corpus of Lusoga along with their distributions. The list contained 200,000 types that were further processed to obtain a headword list for the dictionary. The author(s) had to rely on frequency again to shorten the wordlist into the most frequent 10,000 words. This step kept all words that had 12 or more instances in the corpus, which are arguably the most common words in Lusoga. After that, the authors used *TLex* to annotate the shortlist according to the POS and the lemma.

Unlike Lusoga, creating a lemma list for Swahili was less problematic given the availability of a larger and tagged corpus. Wójtowicz (2016) described the role corpus played in constructing a learner Swahili–Polish dictionary that was, unlike Swahili dictionaries, not based on previous dictionaries. The lemmas and the dictionary entries were based on the Helsinki Corpus of Swahili, the largest for an African language. The corpus is tagged, and it contains more than 12 million tokens of Standard Swahili. The headword list consists of 10,000 types, including nouns, verbs, adjectives, adverbs and MWEs. It was mainly derived from the corpus and expanded with words from pedagogical Swahili resources.

Similarly, Spence (2021) aimed at using the Hupa corpus as an integral part of compiling the Hupa–English dictionary. However, the task was far more complicated for this endangered and lesser-resourced language. Given its productive polysynthetic morphology, the lexicon of Hupa is more dynamic than most languages. The Hupa lexicon is not a list of content and functional words that are subject to neologism. Instead, it is a set of rules according to which morphological forms combine to form new words. Given this language-specificity, corpus provides authentic examples of the productive processes of word formation and for the created words. However, the available Hupa corpus consists of approximately 55,000 words which is a very small size. Although the corpus consists of the speech of 20 speakers talking about personal experiences, cultural practices and traditional stories, its size is too small to allow effective corpus-driven analysis. On the one hand, it is impossible to include the infinite number of words that can be formed according to the Hupa rules. On the other hand, the Hupa corpus is too small to derive a representa-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

tive frequency-based wordlist. Therefore, the project had to rely on a previous print dictionary and complement the wordlist with words appearing in the Hupa corpus. The corpus helped in creating new dictionary entries and enriching existing ones which were originally based on the print dictionary. It was a solution to bridge the wide gap between the infinite possible words and the ones that have already been produced by Hupa speakers.

Prinsloo (2015) discussed how the size of the corpus might affect the coverage of the lemma list, especially for less-resourced languages. It mainly focused on Afrikaans and Sepedi. The study compared the overlap between frequency lists obtained from one-million and ten-million tokens of the same corpora. There was 77.5% and 72.8% overlap between the two lists for Afrikaans and Sepedi, respectively. However, the loss was significant as it missed words that are part of the core vocabulary, e.g., *seyalemoya/radio* and *kamano/ relationship*. Also, the coverage of idioms was influenced by the size of the corpus. Although essential idioms were present in the one-million corpus, they were not frequent enough to be detected by a lexicographer.

4.2 Keyword extraction and compiling wordlists for specialized dictionaries

Specialized dictionaries may target a specific language phenomenon such as collocations or idioms or specific terminology such as legal or medical terms. For instance, Pimentel (2013) used a term extractor (which is somehow equivalent to keyword extraction) to identify legal terms from a specific corpus of legal judgments called JuriDiCo. The extractor automatically identified words that are specific to JuriDiCo if compared to a general reference corpus. The term extraction function also enabled Boz et al. (2018) to select terms relevant to Turkish lexicography from their specialized corpus of dissertations, textbooks and research articles relevant to lexicography. In the same vein, Cabezas-García and Faber (2018) created an 8-million-word specialized corpus from environmental and eco texts. Then, they used a term extractor in order to identify words that are specific to the ecological discourse automatically.

Aiming at different target users, Wild, Kilgarriff and Tugwell (2013) explained the role of corpus analysis in compiling children's dictionaries. They clarified the specificity of the discourse directed to children and argued that children dictionaries should not be merely simplified versions of adult dictionaries. They compared the keywords of *Oxford Children's Corpus* (30 million tokens) with the *Oxford English Corpus* to make decisions on the headword list for the children's dictionary. The classification of the key lemmas in the two corpora showed the discrepancy in the themes presented for adults and children. For instance, language related to the physical world was highly present in the children corpus, whereas politics and religion were characteristic of the adult corpus. Therefore, the study advocated the use of specific children's corpus in compiling children's dictionaries in order to refine the headword lists.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

Specialized dictionaries can make partial use of corpus tools too. For instance, Gizatova (2018) relied on frequency and collocations to construct a bilingual dictionary of idioms. The list of idioms was not corpus-based; it was based on previous dictionaries of idioms. However, the frequency of the idiom in the corpus was the main criterion for keeping or excluding it from the constructed dictionary (40% of dictionary-based idioms were excluded). Also, collocational analysis of the idioms detected the variant forms of an idiom and its semantic preferences. Moreover, analyzing the occurrences of the idioms in parallel corpus revealed dissimilarities between the supposedly equivalent idioms in English and Russian. It displayed the contexts in which the idioms and their dictionary-recommended translations are not equivalent.

4.3 Collocations/concordance and finding senses and structuring dictionary entries

The challenge of sense delineation is common between lexical semantics and lexicography. Over time, linguists and lexicographers started to adopt conflicting views on meaning. To elaborate, in the traditional view, words are believed to have several types of meaning, such as lexical and contextual meaning. Lexical meaning is the semantic content of the word regardless of the contexts in which it may be used. In contrast, contextual meaning arises when the word is used in real communicative situations. On the one hand, some linguists rejected the possibility of having a word meaning outside the context of use (cf. Kilgarriff 1992; Hanks 2004). On the other hand, linguists such as Louw (1995) rejected the existence of contextual meaning as part of the word meaning. He equated any context-based interpretation of a word with word use and called for recording only the lexical meaning of a word in a dictionary (Bergenholtz and Gouws 2017).

Although corpus tools present numerous authentic word usages, converting corpus citations into an organized list of senses that appeal to dictionary users is a laborious lexicographic task. The process of discovering senses from corpus citations does not follow a conventional method (Lew 2013). For instance, Kilgarriff (2005) proposed a model that aimed at putting corpus into dictionaries (PCID). The model relied on collocate-to-sense mapping but added a grammatical dimension to the collocational relation (word sketch collocations which are now conventionally used in lexicographic practice). In addition, Hanks (2004) revealed through his *Corpus Pattern Analysis* (CPA) how the meanings of a word could be mapped to patterns of usage. He adopted a corpus-driven approach based on the *Theory of Norms and Exploitations* (TNE) in order to examine word meanings in contexts instead of assuming the existence of meaning in isolation from the context. His project required a massive lexicographic effort to process word uses, find and record usage patterns and associate each pattern with a meaning.

In this regard, the discussed studies relied heavily on word sketches and collocational patterns to identify, split, lump and order senses in dictionary entries. They usually combined corpus-based analysis with theoretical analysis. Dalpanagioti (2018) conducted a frame-based analysis to detect polysemy in a corpus through the co-occurrence patterns. The paper provided the verb *fly* as a case study. Word sketches detected the verb's co-occurrence patterns which were further analyzed (in a sample concordance) according to the principles of the Frame Semantics theory. The analysis of the co-occurrence pattern of *fly* and *flag*, *kite* and *banner* showed that the collocates fill in the theme semantic role (e.g., *the flag flew from the castle's topmost tower*). The meaning of *fly* in this pattern indicates that the theme is fixed at one point while moving in the air. This sense evokes the frame of "Moving in place". When this pattern changes to involve a human, building or a vessel in the agent position (e.g., *the ship was flying a quarantine flag*), another sense is identified. In the second pattern, the meaning of *fly* denotes "raise a flag and make it float in the air". This sense evokes the frame of "Cause motion" (Dalpanagioti 2018: 14). That is to say; polysemy is detected if the word is used in different patterns that represent different arguments (i.e., frame elements which are frame-specific semantic roles). Accordingly, the senses are split into different frames.

Smirnova (2021) used collocational patterns as sense distinguishers and polysemy detectors in another context. The study did not rely on a theoretical linguistic background. It was rather motivated by the literature on psychology which argued that "vastness and accommodation" are typical features of this psychological experience whereas "threat", "beauty", "ability", "virtue" and "supernatural" are non-central to it (Smirnova 2021: 231). Two binary variants (i.e., positive and negative) of *awe* have been the focus of psychological studies. The study analyzed a sample concordance of the noun *awe* cited from a 14-billion-word corpus. The collocations of the target word and the concordance helped the scholar define the multiple senses of *awe* and the evaluative attitude of the experiences expressed by the different uses of the word. Although the results confirmed vastness as a core feature of *awe*, it suggested distinguishing between transcendental *awe* (linguistically reflected in collocates such as *God* or *supernatural*) and mundane *awe* (linguistically reflected in collocates such as *landscape* or *technology*). Moreover, the study argued for classifying the variants of *awe* into transcendental ambivalent, mundane ambivalent, transcendental positive, mundane positive, transcendental negative and mundane negative based on the collocational analysis.

Similarly, De Schryver and Nabirye (2018b) performed manual annotation of a sample concordance of the verb *-v-* in Lusoga. They were able to map the different usage patterns to meaning potential and construct two entries for the verb accordingly. Senses were organized according to their frequency in the analyzed sample.

4.4 Collocations/concordance and accounting for language varieties

From a sociolinguistic perspective, a dictionary should reflect a representative

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

picture of the language and cover its varieties (Dolezal 2020). Whereas scholars such as Siepmann (2015) argued for allowing a better description of the colloquial variety and reliance on spoken corpus in dictionaries (for French), other scholars such as Xia et al. (2016) called for the inclusion of varieties that are specific to the non-native speakers of the language.

Adopting the perspective of "world Englishes", Xia et al. (2016) argued for the inclusion of "China English" in learner's dictionaries. They considered it part of the core vocabulary of English as it represents a variety spoken by an expanding number of users and has its own characteristics. First, they criticized current English dictionaries for the marginal inclusion of China English words. Although they acknowledge frequency as a criterion for inclusion, they called for using English corpora produced by Chinese speakers so that frequency-based wordlists would reflect the centrality of some China English words. They compared the frequency of a sample of common China English words in the China English Corpus and the British National Corpus (BNC) to support their argument. In addition, they further called for a Chinese-oriented description of the English word senses. They used concordance lines and world knowledge to show how the meaning of the noun *house* is associated with two different concepts for Chinese and English speakers. Accordingly, the definitions targeting Chinese learners should consider such differences.

Notwithstanding, this argument does not sound plausible for various reasons. First, monolingual learner's dictionaries usually target a proficiency level (e.g., beginners, intermediate, advanced). They cannot specify users according to their first language, given the status of English as a lingua franca. Moreover, non-native varieties of English such as Spanish English or Indian English are also expanding and displaying their own distinguishing features. They are admitted in English dictionaries when they are influential enough to be frequently uttered by native speakers and, hence, reflected in native general reference corpora. According to Ooi (2021), there are several cases of admitting Japanese words to the big five dictionaries even without the regional label "in Japan" because they were frequent enough in native general corpora.

5. Conclusion

To conclude, lexicographers made extensive use of corpus tools in the past decade to construct dictionaries and improve existing ones. Compiling headword lists, detecting collocational patterns, identifying word senses and revealing idiom variations have been some of the lexicographic challenges that are successfully met through corpus tools. The frequency either for wordlists or specific lemmas, keyword extraction and word sketches are among the most effective aids for lexicographers. According to the explored studies, almost all dictionaries benefit from corpus tools at the macro and microstructure levels. The need for using these tools became most pressing while compiling specialized dictionaries and landscaping language varieties.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

Worthy mentioning, no study can be comprehensive enough to include every work at the intersection between corpora and lexicography. The scope of the present study aimed at covering some of the corpus contributions to lexicography in the past ten years only. Therefore, early significant contributions were not tackled in the study. Moreover, the discussed studies are cited only from the journals that are indexed in PubMed, Science Direct, Web of Science, Scopus, Mendeley, ERIH PLUS, PsycINFO, ProQuest, and Crossref. Accordingly, studies that were published in conference proceedings and journals that are not indexed in the abovementioned databases were not covered in the study regardless of their relevance.

References

Abecassis, M. 2007. Is Lexicography Making Progress? On Dictionary Use and Language Learners' Needs. *Lexikos* 17: 247-258. https://doi.org/10.5788/17-0-555.

Anshen, F. and M. Aronoff. 1999. Using Dictionaries to Study the Mental Lexicon. *Brain and Language* 68(1–2): 16-26.

Atkins, S. 2008. Theoretical Lexicography and its Relation to Dictionary-making. Fontenelle, T. (Ed.). 2008. *Practical Lexicography: A Reader*: 31-50. Oxford: Oxford University Press.

Atkins, S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Azkarate, M. and D. Lindemann. 2018. Basque Lexicography and Purism. *International Journal of Lexicography* 31(2): 132-150. https://doi.org/10.1093/ijl/ecy003.

Bergenholtz, H. and R.H. Gouws. 2017. Polyseme Selection, Lemma Selection and Article Selection. *Lexikos* 27: 107-131. https://doi.org/10.5788/27-1-1396.

Bogaards, P. 2013. A History of Research in Lexicography. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 19-31. London/New York: Bloomsbury Academic.

Boz, E., F. Bozkurt and F. Doğru. 2018. Corpus-based Research on Terminology of Turkish Lexicography (CBRT-TURKLEX). *Lexikos* 28: 428-439. https://doi.org/10.5788/28-1-1472.

Cabezas-García, M. and P. Faber. 2018. Phraseology in Specialized Resources: An Approach to Complex Nominals. *Lexicography* 5(1): 55-83. https://doi.org/10.1007/s40607-018-0046-x.

Dalpanagioti, T. 2018. A Frame-semantic Approach to Co-occurrence Patterns: A Lexicographic Study of English and Greek Motion Verbs. *International Journal of Lexicography* 31(4): 420-451. https://doi.org/10.1093/ijl/ecy016.

Dalpanagioti, T. 2019. From Corpus Usages to Cognitively Informed Dictionary Senses: Reconstructing an MLD Entry for the Verb 'float'. *Lexicography* 6(2): 75-104. https://doi.org/10.1007/s40607-019-00059-5.

De Schryver, G.-M. and M. Nabirye. 2018a. Corpus-driven Bantu Lexicography. Part 2: Lemmatisation and Rulers for Lusoga. *Lexikos* 28: 79-111. https://doi.org/10.5788/28-1-1458.

De Schryver, G.-M. and M. Nabirye. 2018b. Corpus-driven Bantu Lexicography. Part 3: Mapping Meaning onto Use in Lusoga. *Lexikos* 28: 112-151. https://doi.org/10.5788/28-1-1459.

Dolezal, F.T. 2020. World Englishes and Lexicography. Nelson, C.L. et al. 2019. *The Handbook of World Englishes*: 725-740. Second edition. Hoboken, NJ: Wiley Blackwell.

Faaß, G. 2018. Lexicography and Corpus Linguistics. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 123-137. Abingdon, Oxon: Routledge. https://doi.org/10.4324/9781315104942-9.

Frankenberg-Garcia, A., G.P. Rees and R. Lew. 2021. Slipping through the Cracks of e-Lexicography. *International Journal of Lexicography* 34(2): 206-234. https://doi.org/10.1093/ijl/ecaa022.

Gizatova, G. 2018. A Corpus-based Approach to Lexicography: A New English–Russian Phraseological Dictionary. *International Journal of English Linguistics* 8(3): 357-363. https://doi.org/10.5539/ijel.v8n3p357.

Hanks, P. 2004. Corpus Pattern Analysis. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 87-97. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne-Sud.

Hanks, P. 2013. Lexicography from Earliest Times to the Present. Allan, K. 2013. *The Oxford Handbook of the History of Linguistics*: 503-536. Oxford: Oxford University Press.

Hanks, P. and S. Može. 2019. The Way to Analyze "way": A Case Study in Word-specific Local Grammar. *International Journal of Lexicography* 32(3): 247-269. https://doi.org/10.1093/ijl/ecz005.

Heuberger, R. 2018. Dictionaries to Assist Teaching and Learning. Fuertes-Olivera (Ed.). 2018. *The Routledge Handbook of Lexicography*: 300-316. London: Routledge. https://doi.org/10.4324/9781315104942-20.

Hudeček, L. and M. Mihaljević. 2020. The Croatian Web Dictionary — mrežnik project — Goals and Achievements. *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje* 46(2): 645-667. https://doi.org/10.20344/AMP.12295.

Jorgensen, J.C. 1990. The Psychological Reality of Word Senses. *Journal of Psycholinguistic Research* 19(3): 167-190.

Kilgarriff, A. 1992. Dictionary Word Sense Distinctions: An Enquiry into their Nature. *Computers and the Humanities* 26(5–6): 365-387.

Kilgarriff, A. 1998. The Hard Parts of Lexicography. *International Journal of Lexicography* 11(1): 51-54. https://doi.org/10.1093/ijl/11.1.51.

Kilgarriff, A. 2005. Putting the Corpus into the Dictionary. *Proceedings of the Second MEANING Workshop, Trento, Italy, 3–4 February 2005*.

Kochová, P. 2019. Frequency in Corpora as a Signal of Lexicalization (On the Absolute Usage of Comparative and Superlative Adjectives). *Jazykovedný Časopis* 70(2): 148-157. https://doi.org/10.2478/jazcas-2019-0046.

Moon, R. 2013. Braving Synonymy: From Data to Dictionary. *International Journal of Lexicography* 26(3): 260-278. https://doi.org/10.1093/ijl/ect022.

Nelson, K. 2020. Informing Lexicographic Choices through Corpus and Perceptual Data. *International Journal of Lexicography* 33(3): 251-268. https://doi.org/10.1093/ijl/ecz030.

Ooi, V.B.Y. 2021. Issues and Prospects for Incorporating English Use in Japan into the Dictionary. *Asian Englishes* 23(1): 62-78. https://doi.org/10.1080/13488678.2021.1876952.

Page, M.J. et al. 2021. PRISMA 2020. Explanation and Elaboration: Updated Guidance and Exemplars for Reporting Systematic Reviews. *BMJ* 372(71).

Park, M. 2020. New Loan Words in the Neologismenwörterbuch: Corpus-based Development of Lexicographic Information for an Online Dictionary of Contemporary German. *Lexicography* 7(1–2): 5-23. https://doi.org/10.1007/s40607-020-00070-1.

Pimentel, J. 2013. Methodological Bases for Assigning Terminological Equivalents. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 19(2): 237-257. https://doi.org/10.1075/term.19.2.04pim.

Prinsloo, D.J. 2015. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus. *Lexikos* 25: 285-300. https://doi.org/10.5788/25-1-1300.

Renau, I. and A. Alonso Campo. 2016. Systematising Corpus-based Definitions in Second Language Lexicography. *DELTA Documentacao de Estudos Em Linguistica Teorica e Aplicada* 32(4): 953-979. https://doi.org/10.1590/0102-44500700542375362.

Siepmann, D. 2015. Dictionaries and Spoken Language: A Corpus-based Review of French Dictionaries. *International Journal of Lexicography* 28(2): 139-168. https://doi.org/10.1093/ijl/ecv006.

Sinclair, J. 1992. *Collins COBUILD English Language Dictionary*. London: HarperCollins.

Smirnova, A. 2021. In Awe of God, Nature and Technology: A Lexical Approach to the Differentiation of Emotional Responses. *3L: Language, Linguistics, Literature* 27(4): 230-243. https://doi.org/10.17576/3L-2021-2704-16.

Spence, J. 2021. A Corpus Too Small: Uses of Text Data in a Hupa–English Bilingual Dictionary. *International Journal of Lexicography* 34(4): 413-436. https://doi.org/10.1093/ijl/ecab006.

Wild, K., A. Kilgarriff and D. Tugwell. 2013. The Oxford Children's Corpus: Using a Children's Corpus in Lexicography. *International Journal of Lexicography* 26(2): 190-218 https://doi.org/10.1093/ijl/ecs017.

Wójtowicz, B. 2016. Learner Features in a New Corpus-based Swahili Dictionary. *Lexikos* 26: 402-415. https://doi.org/10.5788/26-1-1343.

Xia, L., Y. Xia, Y. Zhang and H. Nesi. 2016. The Corpora of China English: Implications for an EFL Dictionary for Chinese Learners of English. *Lexikos* 26: 416-435. https://doi.org/10.5788/26-1-1342.

Appendix: Table 1

Author(s)	Lexicographic task	Type of challenge	Relevance to dictionary structure	Relevance to other linguistic fields	Analyzed language	Type of the lexicographic resource	Role of the lexicographic resource	Used corpus	Corpus solution	Type of analysis
Wild et al. (2013)	Compiling lemma lists, defining senses, providing examples	Lexicographic	Macrostructure/ Microstructure	Psycholinguistics [children discourse]	English	Children/school dictionaries	Comparison/ reference point	Oxford Children Corpus; Oxford English Corpus	Keywords; lemma frequency; word sketches; concordance	Corpus-based
Moon (2013)	Differentiating near synonyms	Lexicographic	Microstructure	Lexical semantics	English	Monolingual learner's dictionaries	Comparison/ reference point	English general reference corpus	Frequency; distribution; concordance; collocations	Corpus-based
Pimentel (2013)	Compiling a specialized lemma list; identifying senses	Lexicographic	Macrostructure/ Microstructure	Terminology	Portuguese–English	Bilingual Portuguese–English language resource	Main project	Comparable specialized corpus	Keyword extraction; Concordance	Corpus-driven; corpus-based; frame-based

Prinsloo (2015)	Compiling a lemma list for less resourced language; collocations; idioms detection; citations	Lexicographic; language-specific	Macrostructure/ Microstructure	Lexical semantics	Sepedi, Afrikaans; English	Monolingual English dictionary	Comparison/ reference point	Pretoria English Internet Corpus;
								Media 24;
								Sepedi corpus
									Frequency wordlist; Word sketch; Concordance	Corpus-driven; corpus-based

Siepmann (2015)	Describing colloquial words	Lexicographic	Microstructure	Sociolinguistics	French	Bilingual and monolingual French dictionaries	Comparison/ reference point	French general reference corpus	Frequency, collocational analysis	Corpus-based
Renau and Alonso Campo (2016)	Providing systematic definitions	Lexicographic	Microstructure	Lexical semantics	Spanish	Monolingual learner's dictionary	Main project	Spanish corpus	Concordance	Corpus-based
Wójtowicz (2016)	Compiling lemma list; providing examples	Lexicographic	Macrostructure	Psycholinguistics [mental lexicon theories]	Swahili	Bilingual Swahili–Polish dictionary	Main project	Helsinki Corpus of Swahili	Frequency wordlist; concordance	Corpus-driven
Xia et al. (2016)	Inclusion and description of non-native varieties	Lexicographic	Macrostructure/ Microstructure	Sociolinguistics	English	Monolingual English dictionaries	Comparison/ reference point	China English Corpus; British English corpus	Frequency; concordance	Corpus-based

Author(s)	Lexicographic task	Type of challenge	Relevance to dictionary structure	Relevance to other linguistic fields	Analyzed language	Type of the lexicographic resource	Role of the lexicographic resource	Used corpus	Corpus solution	Type of analysis
Cabezas-garcía and Faber (2018)	Inclusion and description of specialized phrases	Lexicographic	Macrostructure/ Microstructure	Phraseology	English–Spanish	Terminological knowledge base	Main project	Specialized English–Spanish corpus	Term extraction; word sketches; concordance; frequency	Corpus-driven; corpus-based
Dalpanagioti (2018)	Identifying senses of polysemous words	Lexicographic	Microstructure	Lexical semantics; cognitive semantics	English–Greek	Bilingual lexical database	Main project	General reference corpora in English and Greek	Frequency; collocations; concordance	Corpus-based; frame-based
De Schryver and Nabirye (2018a)	Compiling a lemma list	Lexicographic; language-specific	Macrostructure	Psycholinguistics [mental lexicon theories]	Lusoga	Bantu language dictionary	Main project	Lusoga corpus	Frequency wordlist	Corpus-driven
De Schryver and Nabirye (2018b)	Identifying, splitting and lumping senses	Lexicographic	Microstructure	Lexical semantics	Lusoga	Lusoga dictionary	Main project	Bantu language corpus	Concordance; collocations	Corpus-based
Gizatova (2018)	Inclusion of MWEs, recording variations in MWEs; detecting false friends	Lexicographic	Macrostructure/ Microstructure	Applied linguistics [translation equivalence]	English–Russian	Bilingual phraseological dictionary	Main project	English–Russian parallel corpora; general reference corpora in English and Russian	Frequency, concordance; collocation	Corpus-based; contrastive
Dalpanagioti (2019)	Identifying, ordering and linking senses of polysemous lemmas	Lexicographic	Microstructure	Lexical semantics; cognitive semantics	English	Monolingual learners' dictionaries	Comparison/ reference point	English general reference corpus	Word sketches; concordance	Corpus-based; frame-based
Hanks and Može (2019)	Identifying the phraseology of words	Lexicographic	Microstructure	Lexis-grammar interface	English	Pattern dictionary of English verbs	Comparison/ reference point	English general reference corpus	Frequency; word sketches	Corpus Pattern Analysis
Kochová (2019)	Inclusion of inflected forms as lexical units	Lexicographic	Microstructure	Morphology	Czech	Monolingual Czech dictionaries	Comparison/ reference point	Czech corpus	Frequency, collocational analysis	Corpus-based
Nelson (2020)	Defining and ordering word senses	Lexicographic	Microstructure	Lexical semantics	English	Monolingual English dictionary	Comparison/ reference point	English general reference corpus	Concordance	Corpus-based
Frankenberg-Garcia et al. (2021)	Detection and description of collocations	Lexicographic	Microstructure	Lexicology	English	Lexical database for collocations	Main project	Academic and general reference corpora	Keyword extraction; concordance; word sketches; frequency	Corpus-driven; corpus-based

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-1-1714 (Article)

Author(s)	Lexicographic task	Type of challenge	Relevance to dictionary structure	Relevance to other linguistic fields	Analyzed language	Type of the lexicographic resource	Role of the lexicographic resource	Used corpus	Corpus solution	Type of analysis
Hudeček and Mihaljević (2020)	Compiling a lemma list; providing examples	Lexicographic	Macrostructure/ Microstructure	Psycholinguistics [mental lexicon theories]	Croatian	Croatian web dictionary	Main project	Croatian corpus	Frequency wordlist; word sketches; concordance	Corpus-based
Park (2020)	Standardizing the description loan words	Lexicographic	Microstructure	Morphology; phonology; orthography	German	Monolingual German general and specialized dictionaries; specialized online dictionary	Comparison/ reference point	German reference corpus	Frequency	Corpus-based; diachronic
Ooi (2021)	Inclusion and description of regional words	Lexicographic	Microstructure	Sociolinguistics	Japan English	Monolingual English dictionaries	Comparison/ reference point	Japanese English news corpus; GloWbE corpus	Concordance; collocational analysis	Corpus-based; comparative
Smirnova (2021)	Detection of polysemy	Lexicographic	Microstructure	Lexical semantics	English	Monolingual dictionaries	Comparison/ reference point	General reference corpus	Concordance; collocational analysis	Corpus-based
Spence (2021)	Inclusion of words of polysynthetic language; recording inflected forms; providing examples	Language-specific	Macrostructure/ Microstructure	Sociolinguistics	Hupa*–English	Bilingual dictionary	Main project	Hupa text corpus	Frequency wordlist, concordance	Corpus-based