When lexical semantics and word sense disambiguation meet lexicography

A response to issues in lexicographical resources caused by lexical ambiguity

Chinh Ngan Nguyen Le and Julia Miller
Adelaide University

Lexical ambiguity, where a word has multiple related (polysemous) and/or unrelated (homographic) meanings, causes challenges in learning and teaching semi-technical medical vocabulary owing to its lack of consideration in current lexicographical resources, including wordlists and dictionaries. Academic wordlists fail to indicate polysemes and homographs, while conventional dictionaries number multiple meanings of a semi-technical medical word in a vertical format that is unlikely to showcase polysemous and homographic relations. This study focuses on a new lexicographical resource, named SemiMed, which addresses issues in wordlists and dictionaries arising from lexical ambiguity. Hsu's (2013) Medical Word List (MWL), an academic list of semi-technical medical vocabulary, was selected as a starting point. A qualitative analysis underpinned by lexical semantic theories was conducted to analyse polysemes and homographs of MWL words. Then, a quantitative analysis that employed word sense disambiguation practices calculated MWL word meaning frequency. These analyses resulted in two key features, radial networks of meanings and technicality levels, that are expected to help SemiMed resolve issues in academic wordlists and conventional dictionaries caused by lexical ambiguity.

Keywords: wordlists, dictionaries, semi-technical vocabulary, lexical semantics, word sense disambiguation

1. Introduction

A word with multiple meanings is viewed as a common phenomenon (Palmer, 1995). This phenomenon contributes to lexical ambiguity (Cruse, 1986; Kempson,

https://doi.org/10.1075/aral.24138.ngu | Published online: 8 April 2026
Australian Review of Applied Linguistics ISSN 0155-0640 | E-ISSN 1833-7139

1977; Lyons, 1977), which is often characterized by polysemy and homography. Polysemy refers to a word having multiple related meanings, while homography refers to two words with different meanings sharing the same written form. In English for Medical Purposes (EMP), lexical ambiguity is manifested through semi-technical medical (STM) vocabulary, a lexical category lying between technical and general vocabulary, making this type of vocabulary challenging to learn and teach (Le & Miller, 2023; Li & Pemberton, 1994; Thurston & Candlin, 1998). These challenges arise because semi-technical vocabulary usually has more than one meaning and is interpreted differently depending on context. For example, *orbit* has two polysemes — "the path something in space follows round something bigger" and "to follow a path in space round something bigger" — that are often used in the general context. These two general meanings are distantly related to its homograph "the space within the skull that contains the eye," which mostly appears in the medical context.

Although lexical ambiguity has been extensively studied within lexical semantics, with several studies (Cantos & Sanchez, 2001; Lakoff, 1987; Tyler & Evans, 2003a, 2003b) offering insights that may facilitate the learning and teaching of multiple meanings of STM words, existing lexicographical resources do not appear to address this multi-meaning phenomenon comprehensively from a lexical semantic perspective. Dictionaries, which may be the first option for learners and teachers, can create confusion due to their conventional approach of distinguishing multiple meanings using vertically numbered word senses, which does not fully depict polysemous and homographic relations (Aitchison, 2003; Ostermann, 2015). Additionally, this format may challenge learners to locate a correct sub-entry when they look up a headword with multiple meanings (Nesi & Haill, 2002). Often, learners refer to the first sub-entry and neglect the others (Boonmoh et al., 2006). Medical meanings, however, such as those used in STM vocabulary, tend to be more unusual (Nesi & Haill, 2002), and thus do not always appear in the first sub-entry. A learner consulting a dictionary could thus fail to locate medical meanings of a STM word.

Academic wordlists such as Hsu's (2013) Medical Word List (MWL) are considered an alternative pathway to learning and teaching STM vocabulary. The MWL is useful because the number of STM words is limited to a manageable size (595 words). The idea of narrowing the focus of learning and teaching down to a definite set of frequently used STM words is worth acknowledging. However, the absence of word meanings offered by the MWL, due to a loose connection between the wordlist's development and word sense disambiguation (WSD), may restrict its usefulness in resolving the challenges of words with multiple meanings. This study thus considers the issues caused by lexical ambiguity in learning and teaching STM vocabulary, which are not adequately addressed in current

resources such as conventional dictionaries and academic wordlists. It aims to develop a new lexicographical resource (SemiMed) that takes lexical semantics and word sense disambiguation into account to minimise the confusion caused by the vertical format of sense enumeration in conventional dictionaries and to compensate for the lack of word meanings in current wordlists. The study is guided by two research questions:

1. How do lexical semantic and word sense disambiguation principles inform the development of SemiMed?
2. What pedagogical implications might be considered for the use of SemiMed?

This study is structured as follows. The *Literature review* will (1) discuss the complexity of lexical ambiguity as it emerges in lexical semantics; (2) present a lexicographic perspective on lexical ambiguity that distances itself from the lexical semantic view, emphasizing the consequences of the conventional format for distinguishing multiple meanings using vertically numbered word senses; and (3) highlight issues related to lexical ambiguity in wordlists, particularly in relation to WSD. The literature review provides the theoretical background for the *Methods* section, which outlines a step-by-step methodological procedure through an exemplar. The *findings and discussion* section then presents SemiMed and explores its potential contributions to learning and teaching STM vocabulary.

2. Literature review

2.1 Lexical ambiguity in lexical semantics

2.1.1 *Sources of homography and polysemy*

A review of current literature has highlighted several sources of homography and polysemy, particularly including language change, lexical borrowing and semantic change (Béjoint, 1990; Bréal, 1900; Carston, 2021; Cowie, 1988; Murphy, 2010; Ullmann, 1962; Vicente & Falkum, 2017).

Homography often results from historical changes in English, which accidentally cause unrelated words to come closer together in form. For instance, Murphy (2010, p. 87–88) discusses the words *sole* (noun: "bottom of a shoe") and *sole* (adjective: "only"), which evolved independently from Latin roots *solea* and *solus*, respectively, but eventually converged in form due to phonological simplification over time. Lexical borrowing also contributes to homography (Jackson, 2013), as seen in *yen* (currency) from Japanese and *yen* (desire), which are unrelated in origin but share a form in English (Murphy, 2010, p. 94).

Unlike homography, polysemy is a product of semantic change, whereby existing words acquire new but related meanings over time (Bréal, 1900; Ullmann, 1962). Language users tend to extend old meanings in predictable or conventional ways (Murphy, 2010; Vicente & Falkum, 2017). For instance, *coat* originally referred to "an outer garment" but now also denotes "animal fur" or "paint layers." These related meanings coexist and thus considered polysemous.

Nonetheless, semantic change does not always lead to polysemy. In some cases, extended meanings overshadow the original, as in *undertaker*, which once meant "anyone who undertakes a task" but now primarily refers to "someone who undertakes funeral preparations for others" and the word is no longer considered polysemous (Murphy, 2010, p. 95). In other cases, extended meanings may drift so far apart that users no longer perceive them as related. Returning to *sole* as a noun, it has another meaning in addition to "the bottom surface of a shoe," which is "a type of flatfish." The "fish" meaning comes from the same root (*solea*) as the "sandal" meaning because of its resemblance to a flat shoe (Murphy, 2010; Online Etymology Dictionary) but is seen as a separate homograph by language users today.

2.1.2 *Distinctions between homography and polysemy*

Both homography and polysemy emerge through various mechanisms, making the distinction between them challenging. The etymology of the word and the (un)relatedness of the word meanings stand out as two predominant approaches to homography and polysemy distinctions (Carston, 2021; Klepousniotou, 2002; Leech, 1974; Lehrer, 1974; Lyons, 1968, 1977; Panman, 1982). The former approach traces the word origin to distinguish polysemy from homography: homographs are derived from distinct roots, while polysemous words are not. This kind of distinction is observable in virtually all historical dictionaries, where polysemous meanings are listed (and usually numbered) under a single entry as different meanings of a single word. In contrast, homographs are treated as separate words and thus given separate entries. Relying on the etymological derivation of words may be helpful, because etymologically related rules, in theory, would be applicable to recognize the homography and/or polysemy derived from coincidence and semantic change. However, in reality, decisions are not always straightforward (Klepousniotou, 2002; Lyons, 1968, 1977; Panman, 1982) because "there are many words about whose historical derivation we are uncertain" (Panman, 1982, p. 118) and "it is not always very clear how far back we should go in tracing the history of words" (Klepousniotou, 2002, p. 206). This can be seen again in the case of *sole* (fish and sandal). Even though the meanings are derived from the same lexical source, the source per se is still hard to trace back and not apparent to every present-day English L1 user.

The latter approach rests on the native speaker's intuition to judge the relatedness of meaning and then determine whether meanings are closely related enough to be polysemous. This approach might be an alternative, especially, as mentioned earlier, when the word's etymology is hardly traceable. Looking again at the example of *sole* (fish and sandal), present-day English L1 users unaware of the etymological connection between the "fish" and "sandal" meanings could possibly feel that "fish" and "sandal" are not related to each other and decide that the two meanings are in a relation of homography (Murphy, 2010). The possibility that some people may perceive "fish" and "sandal" as homographs, however, does not necessarily exclude the possibility that others may see a similarity in the shape of "fish" and "sandal" from which they conclude that these two meanings are related to each other and thus should be deemed polysemous. The native speaker-based distinction between relatedness and unrelatedness is therefore not as apparent as it may seem.

Each approach has its own shortcomings, including uncertainty about the historical derivation of words (Klepousniotou, 2002) and the undesirably high level of subjectivity resulting from the existence of arbitrariness (Atkins, 1991). Combining the two approaches may remedy the shortcomings; for example, intuitive judgment can be informed by etymology-based evidence to minimize subjectivity. However, such a combination has not been observed in recent lexical studies.

2.1.3 *Approaches polysemy*

There have been several approaches to explaining polysemy, among which *monosemy* and *polysemy* are two fundamental positions that diverge theories on polysemy into separate directions (Cruse, 1992; Janssen, 2003; Murphy, 2010). The monosemy approach treats polysemy as a surface phenomenon. It posits that lexical entries are abstract and underspecified, with meaning derived from context or generative rules (Carston, 2021; Pustejovsky, 1995; Ruhl, 1989). This view supports a simplified mental lexicon where each word has one core meaning. However, critics argue it overlooks the complexity and irregularity of natural language (Murphy, 2010).

In contrast, the polysemy approach views words as complex, conceptual categories with distinct, related senses stored in the mind. Brugman and Lakoff (1988), building on Rosch's (1975) prototype theory, suggest that meanings of polysemous words are stored as categories centered around a prototypical sense. This framework, aligning with cognitive semantics, shows a word's multiple meanings in a radial category, with more typical senses located near the center and less typical ones further out.

Rosch's prototype theory posits that categorization is based on resemblance to the best example or prototype. For instance, in the category *bird*, a *robin* is identified more readily than an *ostrich* due to its prototypical features (e.g., flying, singing, and feathered). Lakoff (1987) applies this idea to lexical semantics, proposing that word meanings are organized similarly. For example, *mother* can mean "biological parent," "nurturer" or even metaphorical sources like "mother lode." These meanings are connected through shared conceptual features and organized around the prototype of *mother* and they are not generated by rules but understood through shared conventions among speakers.

As Lakoff (1987) emphasizes that word meanings in each conceptual category are conventionalized, critics question the ability of his theory to provide objective results (Dominiek, 1998; Tyler & Evans, 2003a). Evans and Green (2006, p. 342) point out that Lakoff's model of radial category is a result of "intuitions (and perhaps also the imagination)" of involved analysts rather than of lexical representations that language users actually store in their mind. Moreover, they argue that semantic analysts do not always agree about the central meaning (the prototype) from which other meanings in a category are derived.

The development of Lakoff's proposed radial category was later refined by Tyler and Evans (2003a, 2003b) through what is known as the Principled Polysemy approach, in which they proposed a set of criteria to determine the prototypical meaning (see Step 3, Section 3.1 for more details). However, in essence, a radial category does not change its nature. It remains "a conceptual category in which the range of concepts are organized relative to a central or prototypical concept" (Evans & Green, 2006, p. 331). In other words, the radial category visualizes how different meanings of a word interact vis-à-vis a central meaning, the one that typically presents mutual semantic properties of other meanings. Polysemy under this perspective, unlike the monosemic view, is structured in a "highly complex" way (Evans & Green, 2006, p. 328) and has "multidimensional structural relations" (Geeraerts, 2006, p. 351).

Cantos and Sanchez (2001), in agreement with Evans and Green (2006) and Geeraerts (2006), observe intricate relations of word meanings and thus develop a Lexical Constellation (LC) model, in which word meanings interact in a multidimensional rather than a linear way. The term LC, which is often used in astronomy, is interpreted within the field of lexical semantics, according to Cantos and Sanchez (2001), as a visualized network of word meanings. In this sense, Cantos and Sanchez (2001, p. 109) perceive a word as "a hierarchical structure whereby each element [each meaning] is directly or indirectly dependent on other elements [meanings]," which is likened to a constellation of stars.

The generic LC model has a core meaning placed at the centre and surrounded by multiple, related meanings located in outer layers which showcase the

degree of interconnectivity. The distance and proximity between meanings indicate how close their relation is. As an illustration, Figure 1 presents an LC of *heart* in which the three meanings in the first layer (central part; thoughts, emotions, feelings; shape of a heart) are more directly related to the core meaning (physical organ in persons/animals) than the meanings on the second layer (core/centre/essence; lover, devotion, sympathy; courage; card with figure of heart). The intersection between meaning clusters in the outer layers indicates their interconnection. For example, "central part of anything" closely links with "core/centre/essence" as they intersect with each other.

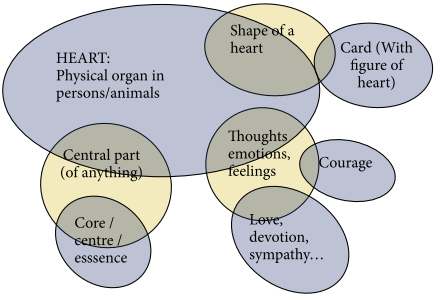

Figure 1. An LC of *heart* (Adapted from Rizzo & Sanchez, 2010, p. 112)

Cantos and Sanchez (2001) claim that their model is capable of describing the intricacies of meanings that have evolved throughout a word's history (i.e., how (new) meanings are rooted in core semantic properties and take additional semantic features over time regardless of their unchanged written form) which is seen as "a permanent source of ambiguity, hence of possible misunderstanding" (Rizzo & Sanchez, 2010, p. 110). Furthermore, Cantos and Sanchez's (2001) LC model has, according to Perez (2013), not only the advantage of analysing semantic complexity, but also the potential to explain semi-technical words. In her study, Perez (2013) examines semi-technical words in law and emphasizes a need to visualize complex semantic relations within multi-meaning words through clear visualizations of interconnectivity existing among usual (generally used) and unusual (discipline-specific) meanings of these words to better understand them across various contexts.

2.2 Lexical ambiguity in lexicography

2.2.1 *Dictionary word sense distinctions*

Since dictionaries mostly feature common words, the core vocabulary of a language, which usually harbour more than one meaning (Jorgensen, 1990), a sub-

stantial part of lexicographic work involves dealing with lexical ambiguity. Although both semanticists and lexicographers deal with lexical ambiguity, they approach it in different ways. Lexicographers tend to generalize specific instances of a word in different contexts into a definite number of distinct word senses under a headword that provides language users with idealized descriptions of a word (Atkins & Rundell, 2008; Hanks, 2002; Kilgarriff, 2007; Mel'čuk, 1988). As Atkins and Rundell (2008, p. 273) explain "from the infinite number of individual situations in which a word appears, lexicographers derive a finite set of [lexical units] which collectively explain how that word contributes to the meaning of all of the individual events" and "which instantiate a one-to-many relationship (where one dictionary sense matches many language events)." This sense distinction process consists of several steps, as follows:

> (1) analyzing instances of usage, typically in concordances or lexical profiles, and
> (2) provisionally identifying different word senses (this is the subjective, intuitive part), then
> (3) collecting good, typical corpus examples for each of these provisional senses. As long as you have plenty of data, one-off oddities can usually be ignored, but ambiguous cases (the examples that you cannot confidently assign to one or other of your provisional senses) should be stored for further analysis (step 5);
> (4) analysing each cluster of examples in turn, the lexicographer identifies the features that are typically associated with it (and that distinguish it from all the other clusters);
> (5) finally, our inventory of senses is refined if necessary (which may involve further splitting, or conversely, lumping of closely related clusters) so that all uses of the word that occur frequently in text are fully accounted for.
>
> (Atkins & Rundell, 2008, p. 312–313)

Although sorting sentence citations sourced from corpora to distinguish word senses sounds straightforward, lexicographers in fact may not arrive at consistent sets of discrete, non-overlapping senses as might be expected (Atkins et al., 2003; Grefenstette & Hanks, 2023; Kilgarriff, 1997). Kilgarriff (1997, 2007, p. 8) and Hanks (2000) agree that lexicographers often find corpus-derived citations contain "loose and overlapping word meanings, and standard or conventional meanings extended [...] and exploited in a bewildering variety of ways." This gives rise to a challenge related to the complexity of polysemy, as addressed in lexical semantics.

Kilgarriff (2007, p. 8) takes the classic example of *bank* to present this challenge. When it comes to a coarse-grained distinction between the word's unrelated meanings (its homographs) "financial institution" and "edge of river," *bank* has two clearly distinct senses. However, it becomes "bewildering" (Kilgarriff, 2007, p. 8) to make fine-grained distinctions among polysemous meanings of *bank*

as "financial institution," e.g., whether "a fund or reserve of money" and "a supply of something held in reserve (blood bank)," to name just a few, should be treated as two distinct senses. A distinction between these two is harder to make than the one between "financial institution" and "river side" because they share overlapping semantic properties, causing blurry boundaries around individual meanings. Not surprisingly, lexicographers, when distinguishing senses of polysemous words, have to "give a sharply delineated presentation of something that is in fact fuzzy" (Hanks, 1990, p. 32) and that, according to Van der Meer (2004), tortures lexicographic practices.

2.2.2 *Dictionary word sense presentations*

Given the lexicographic approach to distinguishing dictionary word senses, dictionaries conventionally list senses in a vertical structure, reflecting the intended outcome of the five-step sense identification process: sets of discrete, non-overlapping senses. A dictionary entry is usually "a list of neatly separated, consecutively numbered lexical meanings" (Geeraerts, 1990, p. 198).

There are several ways to order polysemous and homographic meanings within a dictionary entry. Two internal structures commonly used in conventional dictionaries are linearization and hierarchy. In a linear structure, "all [meanings] have equal status … [and] are presented on one level" (Moerdijk, 2003, p. 285). A hierarchical structure, on the other hand, has "two or more levels on which related [meanings] are grouped" (Moerdijk, 2003, p. 286). Similar to the problems in dictionary word sense distinctions, dictionary word sense presentations that aim to respect discrete, non-overlapping senses cause issues related to polysemy (and homography).

Given that meanings are all listed on the same level, linearization may not imply semantic inter-relatedness; thus, dictionary users, when they look at linearly organized meanings, may tend to treat each meaning as a discrete element that has no relation to remaining meanings. For example, Cambridge Dictionary (https://dictionary.cambridge.org/) does not flag (a) the distinction between homography and polysemy, and (b) the relation among polysemous meanings, due to its linear structure of presenting homographs (e.g., *colon*) and polysemous meanings of words (e.g., *benign*) at the same level (Figure 2). Moreover, users who consult Cambridge Dictionary to understand, for example, what *benign* means in the medical context, may encounter the challenge raised by Nesi and Haill (2002) and Boonmoh et al. (2006) of failing to locate the medical meaning as it is not the first meaning presented in the entry.

In a hierarchical structure, meaning groupings determined from their relatedness seem to provide more straightforward indications than the linear structure does. The distinction between homography and polysemy is drawn because

colon	benign
n. (body part) the lower and bigger half of the bowels in which water is removed from solid waste n. (sign) the symbol: used in writing, especially to introduce a list of things or a sentence or phrase taken from somewhere else	adj. (person) pleasant and kind adj. (disease) a benign growth is not cancer and is not likely to be harmful

Figure 2. The linear structure of *colon* and *benign*. Definitions from Cambridge Dictionary online (https://dictionary.cambridge.org/) in the order in which they appear

homographs and polysemous words are grouped in separate entries. Polysemous meanings are grouped within an entry in a hierarchical order (Figure 3). Nevertheless, although the hierarchy of polysemous meanings establishes their relationship, how the different meanings relate to each other is not explicitly showcased. In other words, from the standpoint of lexical semantics, particularly the multidimensional structural relations of polysemy (Evans & Green, 2006; Geeraerts, 2006; Lakoff, 1987), a hierarchical structure still has a minimal capacity for showcasing polysemous relations.

colon (n)	benign (adj)
Entry 1: the part of the large intestine that extends from the cecum to the rectum Entry 2 1 plural cola: a rhythmical unit of an utterance 2 plural colons: a: a punctuation mark b: the sign Entry 3: a colonial farmer or plantation owner	1a: of a mild type or character that does not threaten health or life b: having no significant effect 2: of a gentle disposition 3a: showing kindness and gentleness b: favourable, wholesome

Figure 3. The hierarchical structure of *colon* and *benign*. Definitions from Merriam-Webster Dictionary online (https://www.merriam-webster.com/) in the order in which they appear

Although lexicographers and semanticists take different approaches to describe words, word meanings presented in dictionaries are ultimately subject to human perception. That means dictionaries are expected to offer "access points" to a bigger picture of the language in one's mind (Ostermann, 2015, p. 65), given that word sense presentations in dictionaries should follow (cognitive) semantic principles in accordance with those that govern how word meanings are represented mentally. However, Aitchison (2003, p. 13) claims that the "fluidity and flexibility of the mental lexicon [...] contrasts strongly with the fixed vocabulary of any [dictionary]," revealing an issue around the vertical listing of word meanings

in dictionaries. This problem is articulated through "the fact that lexicographers [...] have to project a multidimensional clustered semantic structure onto the linear order of the dictionary" (Geeraerts, 2001, p. 14). A consequence of the contrast between presentations of dictionary word senses and mental representations of word meanings is that with such a desire for neatness, dictionaries follow the default structure of word meanings vertically numbered under entries, i.e., the two internal structures discussed, so strictly that semantic considerations, such as polysemy and homography, are consequently overlooked (Aitchison, 2003; Ostermann, 2015).

2.3 Lexical ambiguity in wordlists and word sense disambiguation

2.3.1 *Wordlists*

A wordlist is a list that indicates a finite number of frequent words learners need to master for their particular learning purposes. It is thought that frequently used words should receive more learning (and teaching) time because they will be more useful to learners (Garnier & Schmitt, 2015). An example of STM wordlists is Hsu's (2013) MWL, which contains 595 semi-technical words that appear so frequently in the medical context that learners of EMP are advised to spend their time learning the list to gain adequate comprehension of what they hear or read.

Wordlists created on the basis of frequency like the MWL have an undeniable pedagogical potential, which is to set an explicit, attainable goal for vocabulary learning and teaching. However, although wordlists help inform teachers how many words learners should know (i.e., the breadth of their vocabulary knowledge), it has been argued that "vocabulary learning is not simply remembering a list of words but rather a complex process" (Yu & Trainin, 2022, p. 235). Yamamoto (2014) states that there is another equally important aspect of vocabulary knowledge, that is, the depth of vocabulary knowledge, which should be considered simultaneously with the breadth of vocabulary knowledge. However, the depth of vocabulary knowledge, e.g., multiple meanings of semi-technical words in the MWL, is only superficially treated. This is because wordlist creation typically rests upon the automatic corpus-based distinction of word forms rather than word meanings, thereby disregarding the multiplicity of meaning (Watson-Todd, 2017).

Having the frequency of word forms as an underlying basis for the selection of candidate words, the MWL comes at a price, that is, failing to include polysemy and homography in the presentation. More specifically, the MWL only provides information about the number of times each word occurs without indicating word meanings, i.e., polysemes and homographs. There has been a growing concern over the absence of semantic explanation in the MWL, because words in this list

are semi-technical and thus tend to have more than one meaning. Le and Miller (2023) revisit MWL and reveal that 51% of words are polysemes or homographs or both. This finding suggests that multiple meanings should be considered in word form frequency-based wordlists like MWL.

2.3.2 *Word sense disambiguation*

To fully address the phenomenon of multiple meanings in word form frequency-based wordlists, extensive work is necessary to "[disambiguate] a word that can have many senses based on its usage context [e.g., a corpus]" (Vidhu Bhala & Abirami, 2014, p. 159). This kind of work is situated in the area of word sense disambiguation (WSD), which has arisen since computers have been involved in building solutions for human language problems. In word form frequency-based wordlists, disambiguating meanings of polysemes and homographs has been overlooked because, as mentioned earlier,

> each word form is counted as a single word, though in reality, each word form may represent a number of distinct meanings, some of which depend strongly on the reading context, and some of which are quite different from each other in meaning. (Grabe, 1991, p. 392)

This is exemplified in the case of *bear*. Gardner (2007) notes that when appearing in different parts of speech, the word *bear* can be perceived as two homographs (*bear* as a noun meaning "an animal" and *bear* as a verb meaning "to carry"). *To bear* is also a polyseme having 13 meanings across contexts. Gardner (2007, p. 253) anticipates that machine-based frequency counts of the written form of *bear* link all of its meanings together and thus incur some risks for learners:

> (a) they will overestimate the true coverage of the word forms; (b) they will underestimate the actual user knowledge required to negotiate the word forms; and/or (c) they will underestimate the actual number of meanings inherent in the word forms.

The fact that one or more of these risks may exist in word form frequency-based lists raises questions about their usefulness, especially when teaching is based on word form frequency without taking account of multiple meanings (Gardner, 2007). From a practical standpoint, Biemiller and Slonim (2001, as cited in Gardner, 2007, p. 252) state that "general print frequency of word forms is a poor predictor of learners' root word knowledge". Moreover, since Ravin and Leacock (2000, p. 1) remind us that "the most commonly used words tend to be the most polysemous," the construct of the word in corpus-based lists of frequently used words might need to be re-operationalized.

From this premise, Knowles and Mohd Don (2004) suggest that individual word meanings should be considered as the basis for frequency-based analyses. Biemiller and Slonim (2001, p. 510) advocate this by further reasoning that "frequencies of word meanings rather than word forms might lead to better predictions of learners' root word knowledge." Although Gardner (2007, p. 253) comments that "such frequencies would be very hard to produce", the task may not be impossible. There have been several approaches to WSD that are believed to facilitate the count of word meaning frequency. For reasons of space, we only elaborate on only one method (Yarowsky, 1993) within the scope of this paper.

The one sense per collocation method is named after the one sense per collocation heuristic, which was first introduced by Yarowsky (1993). This heuristic places the focus on word collocations, since words in close proximity offer reliable and consistent hints about the meaning of a target word. In regards to polysemy, Yarowsky's (1993) one sense per collocation hypothesis resonates closely with Hoey's (2012, p. 13) Lexical Priming, which postulates that "when a word is polysemous, the collocations ... of one sense of the word differ from those of its other senses."

The underlying mechanism of this method used to disambiguate polysemous meanings is demonstrated in Cantos et al.'s (2009, p. 79) diagram (see Figure 4), where three meanings of a polysemous word are discriminated by their sets of collocations.

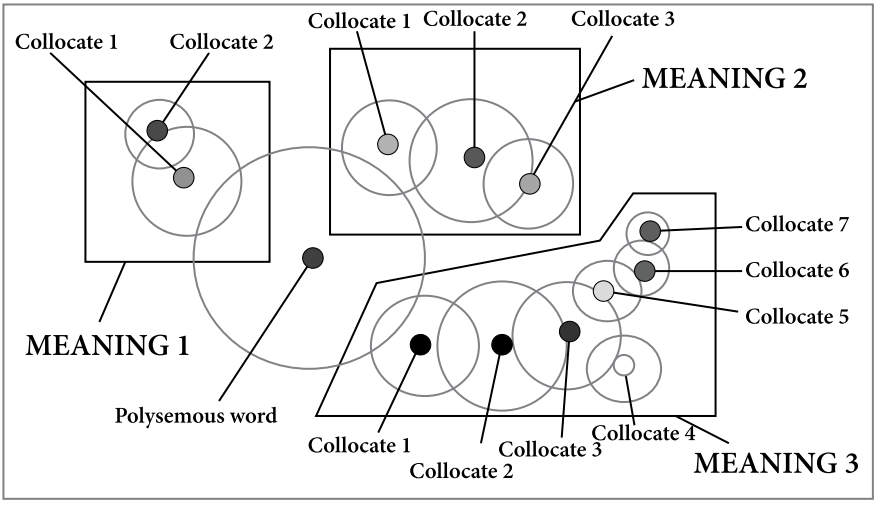

Figure 4. Cantos et al.'s (2009) illustration of word meaning disambiguation based on collocations

It is widely agreed by corpus linguists that "the meaning of a word is dependent on the other words associated with it in a particular text (cotext), and that words are only ambiguous when isolated from their cotext" (Sinclair, 2004, as cited in Gardner, 2007, p. 252). Conceivably, putting multi-meaning words in wordlists back into a particular context (a corpus) and investigating their collocations may provide valuable disambiguating clues. Hence, the one sense per collocation method is promising and may be achievable as corpus analysis software now allows the automatic export of collocational data.

3. Methods

In response to the concern raised in the *literature review* about the inadequate treatment of polysemy and homography in dictionaries and wordlists due to the disconnection between these resources and lexical semantic views on and WSD practices of polysemy and homography, this study incorporates lexical semantics and WSD in the development of a new resource (named SemiMed) to fully address lexical ambiguity in STM vocabulary.

The MWL (Figure 5) was selected as a starting point because, first, it characterizes features of a list developed by calculating word forms in a corpus. Secondly, as mentioned earlier, Le and Miller's (2023) evaluation of the list revealed that of the 595 words, over half (302 words) have multiple meanings (polysemes, homographs or both). The reliance solely on word form frequency as a basis for list compilation indicates an urgent need to conduct an extensive study to enrich semantic information absent in the MWL.

The methodological approach taken in this study is a mixed method based on the lexical semantic theories and WSD practices explored in the *Literature review*. A qualitative analysis was initially conducted to analyse different meanings of MWL words. A combined approach that considered both etymology and speaker intuition was used to distinguish polysemy from homography. Lakoff's (1987) radial category, Tyler and Evans's (2003a, 2003b) Principled Polysemy and Cantos and Sanchez's (2001) LC shed light on the qualitative analysis to respond to the minimal capacity to capture semantic intricacies in dictionaries. This allows a higher level of the dictionary default structure, where polysemy and homography are not implicit or implied, but explicitly visualized. Rather than vertically listing polysemous and homographic meanings under word entries such as *colon* and *benign* in Figures 2 and 3, the qualitative analysis allowed SemiMed to visualize how these meanings interact vis-à-vis a core meaning. Then, a quantitative analysis followed, with a focus on word meaning frequency to substantiate and validate results from the qualitative analysis. The analytical method was based

MWL	Range	Frequency	BNC
PRESENCE	30	7,635	BNC4
RECUR	30	7,607	BNC6
INFLAME	30	7,554	BNC5
BIOPSY	29	7,487	BNC8
PULMONARY	30	7,301	BNC11
PATHOLOGY	30	6,949	BNC7
VAGINA	24	6,941	BNC8
INHIBIT	29	6,840	BNC5
MUTATE	27	6,780	BNC9
ANAESTHESIA	28	6,662	BNC12
HEPATIC	24	6,660	BNC12
MALIGN	30	6,556	BNC6
ABDOMEN	29	6,507	BNC9
DEFECT	30	6,496	BNC5

Figure 5. An excerpt from Hsu's (2013) Medical Word List

on Yarowsky's (1993) one sense per collocation heuristic to address the automatic corpus-based distinction of word forms in the MWL. The step-by-step methodological procedure is illustrated via the case of the word *defect*.

3.1 Qualitative analysis

Step 1. Look up each word in the Oxford English Dictionary (OED)

The words were looked up in the OED because the MWL provides no semantic information other than range and frequency statistics (Figure 5).

There were two reasons for selecting the OED to look up the MWL's headwords. First, since the MWL contains semi-technical medical words, a type of vocabulary with general and medical meanings, it was necessary to choose a dictionary that presents both meanings in great detail. The OED was a suitable option because, unlike medical dictionaries (which usually focus on presenting only medical meanings), it can be used as a general dictionary that presents word meanings in general contexts. The OED is more suitable than other general dictionaries because it is a historical dictionary comprising a wide range of meanings in different disciplines, including medical ones. This means it does not miss out medical meanings, which are sometimes absent in general dictionaries. Secondly, as a historical dictionary, the OED includes etymological information rarely seen

in learners' (general and medical) dictionaries and offers a complete account of each word's history, i.e., past and present word meanings. These two features of the OED are beneficial in providing rich data for a core meaning-based analysis.

As in the examination of *defect* in the OED (Figure 6), each word was looked up and then the results were refined by removing obsolete and merging overlapping meanings to prepare the word for further analysis.

Defect	***BNC: Band 5***	***Range: 30***	***Frequency: 6,496***

n. Lack or absence of something necessary or desirable; a deficiency, a want. Also: the state or fact of being deficient or falling short.

n. An imperfection in a person or thing; a shortcoming, a failing; a fault, flaw, or abnormality.

v. To abandon or desert a person, party, organization, or cause, esp. in favour of an opposing one.

Figure 6. OED definitions of *defect*

Step 2. Simplify OED definitions

Next, the original definitions from the OED were simplified to ensure learners of English, particularly those at lower levels, could fully understand individual definitions. Moreover, using simply reworded OED definitions optimizes the space of text bubbles in LCs. The simplification of OED definitions was underpinned by two principles (Atkins & Rundell, 2008, p. 433; Mel'čuk, 1988, p. 171–175):

- The decomposition principle: Defining language must be semantically simpler than the content it defines.
- The univocity principle: Each defining term used in defining a word must not be ambiguous.

Figure 7 shows OED definitions of *defect* after the simplification process.

Defect	***BNC: Band 5***	***Range: 30***	***Frequency: 6,496***

(a) n. A condition where there is not enough of something

(b) n. An imperfection in a person or thing

(c) v. To leave (and join the other side)

Figure 7. Simplified OED definitions of *defect*

Step 3. Classify OED definitions into polysemes and homographs based on the core meaning

Once the OED definitions were simplified, they were classified into polysemes and homographs. The Principled Polysemy approach informed the identification of core meaning. A core meaning needs to fulfil at least one, and preferably more than one, of three criteria: "(1) [closely relates to the] historically earliest attested meaning, (2) predominance in the semantic network, [...] (3) predictability regarding other senses" (Evans, 2005, p. 44).

Defect	***BNC: Band 5***	***Range: 30***	***Frequency: 6,496***
(a) n. A condition where there is not enough of something (b) n. An imperfection in a person or thing (c) v. To leave (and join the other side)			
Core meaning(s)	Meaning a)	Meaning b)	Meaning c)
Core meaning 1: A lack	✔	✔	
Core meaning 2: To leave (and join the other side)			✔

Figure 8. Core and other meanings (polysemes and homograph) of *defect*

The first criterion was determined by looking at the etymological reference in the OED. The remaining criteria were based on the intuitive judgment of three evaluators. A word was classified as polysemous if all its senses shared a core meaning. New senses creating new core meanings would be classified as homographs. Figure 8 showcases polysemes and a homograph of *defect* resulting from the core meaning-based analysis. This result was used as an input for the next step involving the visualization of the meanings of *defect* in LCs.

Step 4. Visualize core and other meanings (polysemes and homographs) in LCs

After arriving at the core, polysemous and homographic meanings of *defect* (Figure 9), the process of visualizing the meanings in LCs was carried out. Each LC represents a (polysemous) word, and if a word has homographs, its homographs are presented in separate LCs. *Defect* has two polysemous meanings derived from core meaning 1 and one homograph representing core meaning 2. The word thus has two LCs, which are featured in Figure 10.

Defect
Core meaning 1: A lack
a. (n) A condition where there is not enough of something
b. (n) An imperfection in a person or thing
Core meaning 2: c) (v) To leave (and join the other side)

Figure 9. Description of *defect* resulting from the qualitative analysis

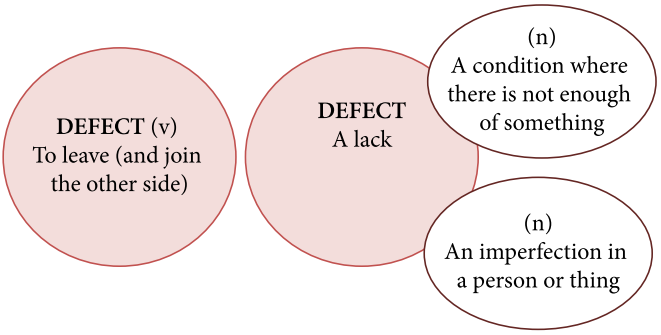

Figure 10. LCs of *defect* resulting from the qualitative analysis

The first LC (on the right) illustrates a polysemous word with core meaning 1 placed at the centre surrounded by meanings *a* and *b*. The remaining LC (on the left) is of a homograph with core meaning 2 placed at the centre. No other related meanings were derived.

3.2 Quantitative analysis

LCs produced from the qualitative analysis were then validated through a corpus-based analysis. At this stage, frequencies of word meanings visualized in qualitatively established LCs were examined in general and specialized corpora to (a) assess whether each particular meaning was frequent enough to keep it in the LCs and, if so, (b) specify in which context(s) the meaning is more likely to appear.

The analysis tool used in the quantitative analysis was Sketch Engine (Word Sketch https://ske.li/ufw). Two corpora, English Web 2020 and Medical Web Corpus, freely accessible for Sketch Engine subscribers, were selected to represent general and specialized corpora. The one sense per collocation method was applied to disambiguate multi-meaning words examined in the two corpora. The unit of analysis was word meaning frequency and the unit of counting was collocate frequency.

Step 5. Examine the top 15 collocates in English Web 2020 and the Medical Web Corpus

Word Sketch in Sketch Engine exported a list of collocates, from which the top 15 most frequent collocates were identified. After multiple trials, a cut-off line of fifteen was found to be an optimal window to retrieve collocates with high frequency of occurrence and practical significance in disambiguating their node words.

Fifteen collocates were selected on the basis of their frequency and typicality scores, both of which were automatically computed by Word Sketch. The primary selection criterion was frequency. Next, typicality was considered, to check whether frequent collocates are strong collocates (i.e., ones that do not often co-occur with many other words). Typicality ensures that frequent collocates are typical of particular meanings. Knowing the typicality score facilitates the follow-up process of assigning meanings to collocates.

After the identification of the top 15 frequent collocates, meanings were assigned to individual collocates. Tables 1 to 3 indicate the results of the examination of the top 15 collocates of *defect* in English Web 2020 and Medical Web Corpus (a search of *defect* as a verb in the Medical Web Corpus returning no results):

Table 1. Top 15 most frequent collocates and meanings of *defect* (n) in English Web 2020

Collocate	Frequency	Typicality	Meaning of *defect* in relation to collocate
1. birth	58,021	10.7	n. Something wrong with part of the body
2. heart	14,079	7.9	n. Something wrong with part of the body
3. congenital	12,062	9.6	n. Something wrong with part of the body
4. genetic	10,309	7.8	n. Something wrong with part of the body
5. neural tube	6,533	8.2	n. Something wrong with part of the body
6. manufacturing	6,359	7.2	n. Something that is not perfect
7. structural	4,923	7.0	n. Something that is not perfect
8. construction	3,058	5.7	n. Something that is not perfect
9. design	2,612	4.8	n. Something that is not perfect
10. physical	2,605	4.3	n. Something wrong with part of the body
11. visual field	2,556	5.4	n. Something wrong with part of the body
12. product	2,497	4.7	n. Something that is not perfect
13.ventricular septal	2,449	7.4	n. Something wrong with part of the body
14. atrial septal	2,176	7.2	n. Something wrong with part of the body
15. surface	2,144	5.5	n. Something that is not perfect

Table 2. Top 15 most frequent collocates and meanings of *defect* (n) in Medical Web Corpus

Collocate	Frequency	Typicality	Meaning of *defect* in relation to collocate
1. birth	275	9.8	n. Something wrong with part of the body
2. congenital	123	9.4	n. Something wrong with part of the body
3. genetic	85	8.6	n. Something wrong with part of the body
4. heart	75	6.3	n. Something wrong with part of the body
5. neural tube	72	9.4	n. Something wrong with part of the body
6. ventricular septal	56	9.3	n. Something wrong with part of the body
7. valvular	37	8.3	n. Something wrong with part of the body
8. visual field	36	7.6	n. Something wrong with part of the body
9. cardiac	32	6.5	n. Something wrong with part of the body
10. mental	31	6.0	n. Something wrong with part of the body
11. physical	27	5.7	n. Something wrong with part of the body
12. afferent pupillary	20	7.7	n. Something wrong with part of the body
13. perfusion	20	7.4	n. Something wrong with part of the body
14. speech	18	6.6	n. Something wrong with part of the body
15. atrial septal	17	7.2	n. Something wrong with part of the body

Step 6. Rate meanings on a 4-level technicality scale

The examination of the top 15 most frequent meanings resulted in five possibilities. A technicality scale (Table 4) was devised from these five possibilities to specify the degree of technicality of a particular meaning. The scale has four levels ranked in ascending order of technicality (from Level 0: purely general to Level 3: highly technical).

As can be seen from Tables 1, 2, and 3, *defect* has two purely general meanings (Level 0): "something that is not perfect" and "to leave (and join the other side)," and one meaning used in both general and medical contexts (Level 2): "something wrong with part of the body".

Table 3. Top 15 most frequent collocates and meanings of *defect* (v) in English Web 2020

Collocate	Frequency	Typicality	Meaning of *defect* in relation to collocate
1. soldier	484	4.3	v. To leave (and join the other side)
2. officer	387	3.2	v. To leave (and join the other side)
3. member	365	1.8	v. To leave (and join the other side)
4. pilot	225	4.3	v. To leave (and join the other side)
5. player	184	1.1	v. To leave (and join the other side)
6. official	159	1.4	v. To leave (and join the other side)
7. customer	158	1.6	v. To leave (and join the other side)
8. voter	155	3.4	v. To leave (and join the other side)
9. general	151	4.3	v. To leave (and join the other side)
10. army	145	3.2	v. To leave (and join the other side)
11. troop	136	3.1	v. To leave (and join the other side)
12. councillor	134	5.5	v. To leave (and join the other side)
13. leader	111	1.2	v. To leave (and join the other side)
14. agent	108	2.2	v. To leave (and join the other side)
15. commander	107	4.6	v. To leave (and join the other side)

Table 4. Descriptors of four technicality levels

Level	Descriptor
Level 0 (Not indicated in LCs)	*This meaning is solely used in general contexts* Possibility 1: Found in the top 15 meanings in English Web 2020, but not found in the Medical Web Corpus
Level 1	*This is the generally used meaning* Possibility 2: Found in the top 15 meanings in English Web 2020, and outside the top 15 meanings in the Medical Web Corpus
Level 2	*This meaning is used in both general and medical contexts* Possibility 3: Found in the top 15 meanings in both English Web 2020 and the Medical Web Corpus
Level 3	*This meaning is used only in medical contexts* Possibility 4: Found in the top 15 meanings in the Medical Web Corpus, and outside the top 15 meanings in English Web 2020; or Possibility 5: Found in the top 15 meanings in the Medical Web Corpus, but not found in English Web 2020

Step 7. Modify LCs resulting from the qualitative analysis (if necessary) and add the technicality level to the LCs

The technicality rating of the meanings of *defect* was incorporated into the word's quantitative analysis results. The analysis of *defect* in the two corpora revealed that meaning *a* was not found in either English Web 2020 or the Medical Web Corpus, leading to its removal from the final LCs. Meaning *b* was broken down into two sub-meanings ranked at two different levels of technicality (Level 0: Something that is not perfect, Level 2: Something wrong with part of the body). Below is the finalized description of *defect*:

Defect
Core meaning 1: A lack
(b) (n) An imperfection
(b.1) (n) Something that is not perfect
(b.2) (n) Something wrong with part of the body (2)
Core meaning 2: (c) (v) To leave (and join the other side)

Figure 11. Description of *defect* resulting from the quantitative analysis

Accordingly, the qualitatively established LCs were modified to exclude meaning *a* and form a cluster of sub-meanings *b.1* and *b.2* (Figure 11). Numbers (1, 2, and 3) were placed at the bottom of each LC text bubble to indicate the technicality levels of each meaning. In the case of *defect*, only one sub-meaning had a technicality level.

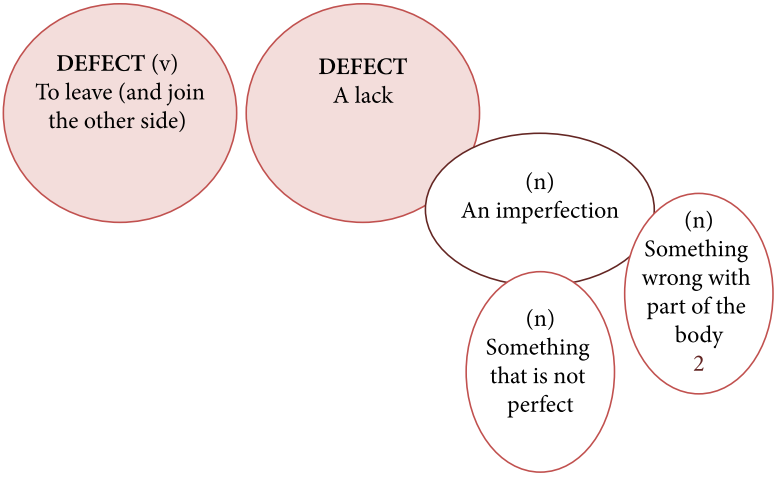

Figure 12. LCs of *defect* resulting from the quantitative analysis

4. Findings and discussion

4.1 SemiMed template

A generic template of SemiMed featuring three types of words (single-meaning words, multi-meaning words with one core meaning and multi-meaning words with more than one core meaning) was designed with key elements:

- Headword
- Core meaning
  - Homograph: Meaning of a homograph
  - Polysemous word: Shared meaning from which other polysemous meanings are derived
- PoS: Part of speech
- Meaning cluster
  - Meaning 1: A polysemous meaning
  - Meaning 1.1, 1.2: Sub-meanings of meaning 1
- Technicality level
  - Level 1: This is the generally used meaning
  - Level 2: This meaning is used in both general and medical contexts
  - Level 3: This meaning is used only in medical contexts

Generic LC designs accompanied by specific examples are shown in Figures 13 to 18.

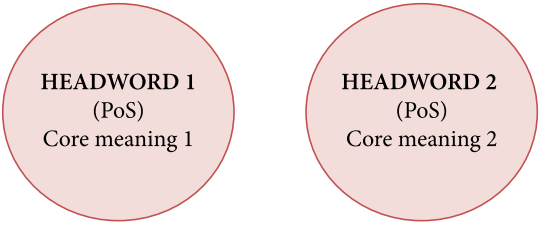

Figure 13. A generic LC for single-meaning words

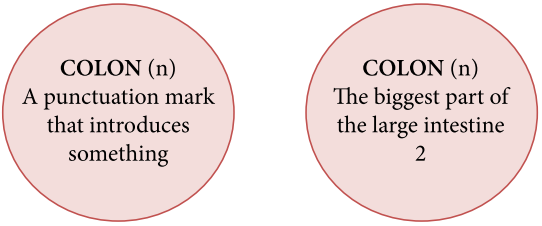

Figure 14. An example LC of a single-meaning word

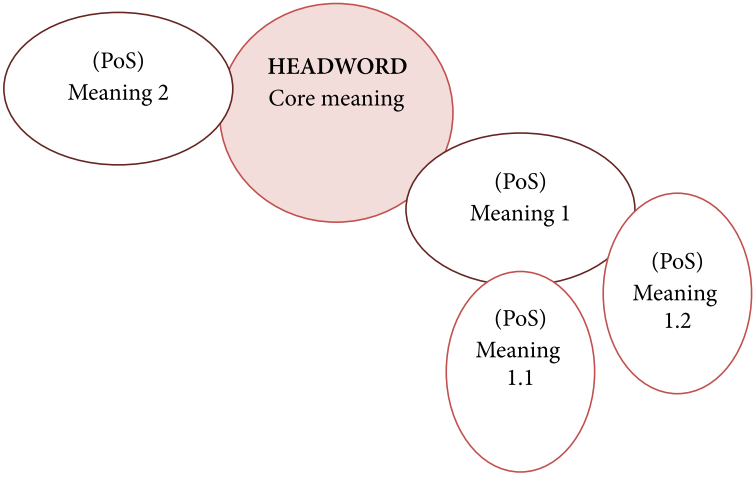

Figure 15. A generic LC for multi-meaning words with a single core meaning

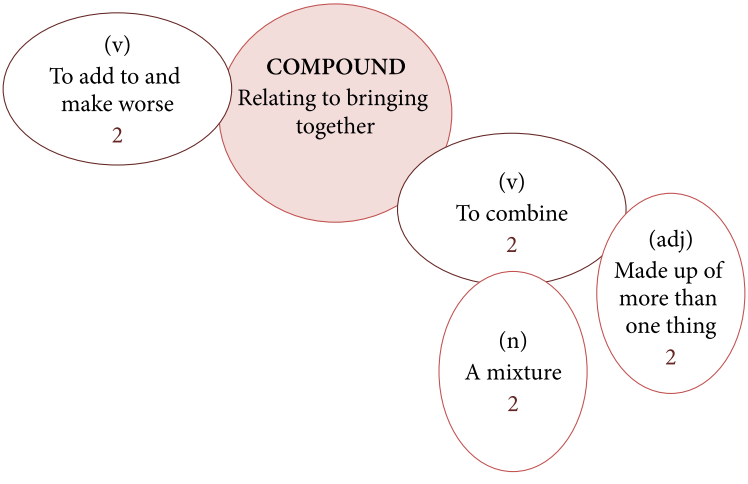

Figure 16. An example LC of a multi-meaning word with a single core meaning

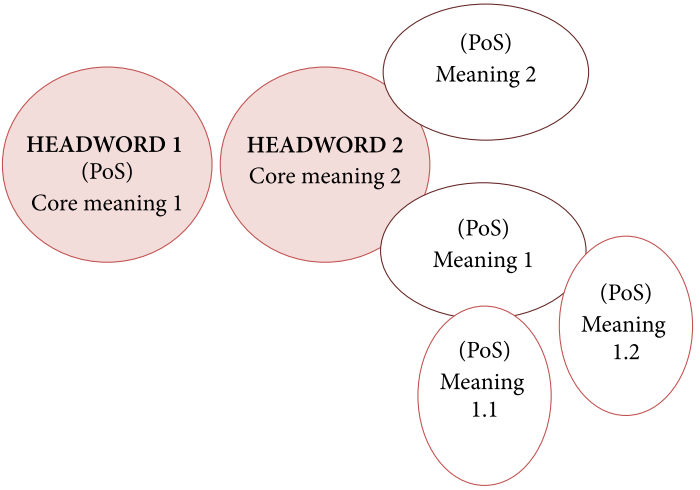

Figure 17. A generic LC for multi-meaning words with more than one core meaning

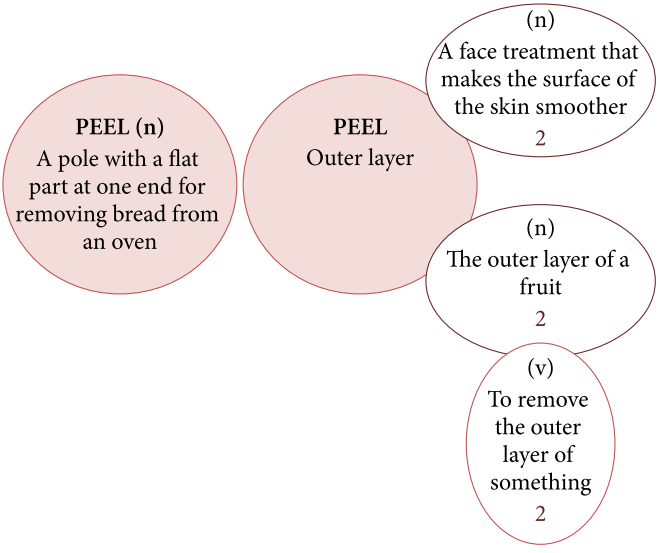

Figure 18. An example LC of a multi-meaning word with more than one core meaning

4.2 Pedagogical implications

The design of SemiMed, which visualizes semantic relations existing in the MWL's multi-meaning words, provides an extended version of Hsu's MWL that improves the pedagogical usefulness of this resource from simply informing learners and teachers of how many words to focus on to guiding how these words can be learned and taught (Le & Miller, 2025). SemiMed increases the semantic input in MWL by detailing the meanings of words in addition to their range and frequency. Word meanings are not vertically listed as in conventional dictionaries. Rather, they are placed in a radiate structure that permits an explicit illustration of how generally used meanings interact with discipline-specific ones (Le & Miller, 2025).

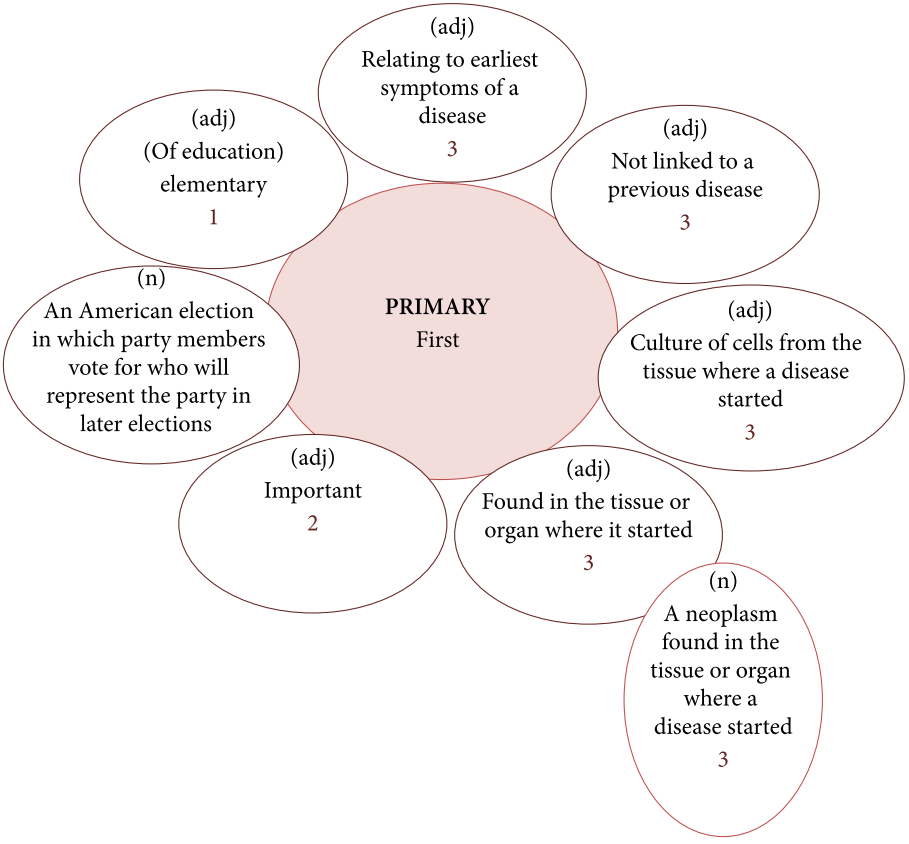

Figure 19. An LC of *primary*

For example, *primary*, illustrated in Figure 19, has five meanings (relating to earliest symptoms of a disease, not linked to a previous disease, culture of cells from the tissue where a disease started, found in the tissue or organ where it started and a neoplasm found in the tissue or organ where a disease started) that are used in a restricted way in the medical context and are closely related to the remaining meanings via the core meaning of "first." SemiMed users can understand general and medical meaning interactions after a first quick look at the LC of *primary* (Le, 2022). This wider understanding is much less achievable in other resources like dictionaries. Cambridge Dictionary, for example, treats each meaning of *primary* as a separate entity having no relation with other meanings listed (Figure 20). SemiMed users would also identify medical meanings with ease because the radiate structure does not have the concept of ranked sub-entries that are often misleading for dictionary users (Le & Miller, 2025). In other words, the comprehension problems raised in previous studies (Boonmoh et al., 2006; Nesi & Haill, 2002) might be eliminated while using SemiMed to look up multi-meaning words.

Primary
(adj) more important than anything else; main
(adj) (education) of or for the teaching of young children, especially those between five and eleven years old
(adj) happening first
(n) in the US, an election in which people choose who will represent a particular party in an election for political office

Figure 20. *Primary* in Cambridge Dictionary (http://dictionary.cambridge.org)

Furthermore, SemiMed LCs not only showcase interactions between general and medical meanings but also specify these semantic interactions, whether general and medical meanings are in a relation of polysemy, homography or both. This feature is expected to resolve the issues around the depth of vocabulary knowledge in Hsu's (2013) MWL and the minimal capacity to showcase semantic intricacies in dictionaries. As Perez (2013) argues, visualizing the complex semantic relations is essential for understanding semi-technical vocabulary. In line with this, a grasp of polysemy and homography gained through LCs may help deepen and facilitate the understanding and interpretation of word meanings across different contexts.

Returning to the word *primary*, for example, it would be pedagogically useful to know if the word is polysemous. Let us assume, for the sake of argument, that SemiMed users already know the general meanings of *primary*. They would be

able to link their prior knowledge of the general meanings with their new knowledge of the medical meanings once they knew that the general and medical meanings are derived from the core meaning of "first." Such links may assist users in consolidating known general meanings and retaining newly acquired medical meanings. Furthermore, SemiMed offers a shortcut to learning multi-meaning words like *primary*. Instead of trying to remember all the meanings listed under *primary* in the Cambridge Dictionary, SemiMed users, after understanding the word's general and medical meanings, can remember the word's core meaning and interpret other meanings depending on future contexts.

Another example of SemiMed's polysemy-and homography-related benefits is *peel*. *Peel* is more complex than *primary* because it has polysemous meanings used in both contexts (a face treatment that makes the surface of the skin smoother, the outer layer of a fruit, and to remove the outer layer of something) and a homograph whose meaning is purely general (a pole with a flat part at one end for removing bread from an oven). SemiMed users are expected, after examining the LC of *peel* (Figure 21), to perceive links between the polysemous meanings via the core meaning of "outer layer" and become aware that "a tool to remove bread from an oven" is the meaning of a word that only shares the same written form. Such interpretations of SemiMed LCs can inform a new vocabulary learning strategy that involves storing core meanings of polysemes and homographs separately and relying on a context to decide which core meaning should be activated and which meaning from the activated core meaning should be used.

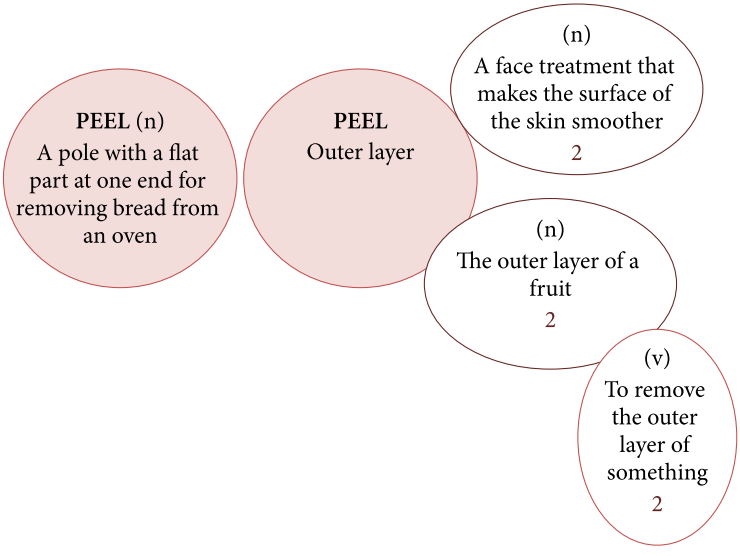

Figure 21. An LC of *peel*

The incorporation of word meaning frequency information into SemiMed LCs consolidates the word form frequency findings inherited from the MWL. The calculation of word meaning frequency in the two corpora proved that it would be unnecessary in some cases to know all the meanings of an MWL word. This was seen in the case of *cardiac*. In the OED, *cardiac* has two meanings: "relating to the heart" and "relating to part of the stomach." However, after the corpus-based calculation, the second meaning was found to be much less frequent in both general and medical contexts and was thus discarded from the LC. In other words, SemiMed filters frequently encountered meanings and increases the pedagogical usefulness of the MWL by highlighting not only which words but also which word meanings are worth learning and teaching.

The technicality levels also enhance the pedagogical potential of SemiMed by providing contextual clues for users to arrive at appropriate interpretations of word meanings, especially unusual ones. Together with the radiate structure, the technicality levels may make the process of looking for unusual meanings in SemiMed more straightforward than in dictionaries because users can effortlessly identify which meanings are medical by scanning LCs to find meanings ranked at Levels 2 and 3 (Le & Miller, 2025). The presence of technicality levels in SemiMed suggests that WSD outcomes, i.e., word meaning frequency, could be transferred to a teachable element in a lexicographical resource and this element is pedagogically practical in use.

5. Conclusion

The present study was designed to address issues of polysemy and homography in STM vocabulary in current resources such as conventional dictionaries and academic wordlists. A novel approach combining theories and practices in lexical semantics and WSD was implemented to remove (a) the confusion the vertically numbering format of dictionary entries may create in understanding polysemy and homography, and (b) the root cause of the lack of semantic depth in the MWL. These issues were uncovered from the review of current literature, which observed that (a) dictionary entries fail to capture polysemous and homographic relations, and (b) the development of wordlists still relies heavily on word form frequency to deal with multi-meaning words. A primary outcome of this study is SemiMed's LC template. Two significant features of SemiMed that make this resource an improvement over word form frequency-based lists like the MWL, and other resources, like conventional dictionaries, are its radiate structure and technicality levels. The visualization of general and medical meaning interrelations in the LCs is believed to deepen the knowledge of vocabulary and elim-

inate the confusion caused by sense-numbered dictionaries, which both have significant effects on learning and teaching multi-meaning words, especially those with unusual meanings found in discipline-specific contexts like STM vocabulary. With its four levels of technicality, SemiMed is one of the very few resources that considers and transfers word meaning frequency into a learnable and teachable aspect of vocabulary learning and teaching. This finding suggests that considering theories of polysemy and homography in lexical semantics can pave the way for non-conventional formats that may become an alternative or even a replacement for the sense enumeration format in conventional dictionaries. Additionally, it proves that the calculation of word meaning frequency is an achievable WSD task that may remedy the shortcomings of academic wordlists that rely entirely on word form frequency.

Funding

This research was funded by the A.S. Hornby Educational Trust under the A.S. Hornby Dictionary Research Award.
Open Access publication of this article was funded through a Transformative Agreement with The University of Adelaide.

References

A. *Dictionaries*

Cambridge Dictionary. https://dictionary.cambridge.org/
Merriam-Webster Medical Dictionary. https://www.merriam-webster.com/medical
Online Etymology Dictionary. https://www.etymonline.com/
Oxford English Dictionary. https://www.oed.com

B. *Other literature*

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Blackwell.
Atkins, B. T. S. (1991). Building a lexicon: The contribution of lexicography. *International Journal of Lexicography*, *4*(3), 167–204.
Atkins, S., Rundell, M., & Sato, H. (2003). The contribution of FrameNet to practical lexicography. *International Journal of Lexicography*, *16*(3), 333–357.
Atkins, B. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
Béjoint, H. (1990). Monosemy and the Dictionary. In *BudaLEX'88 Proceedings of the 3rd International EURALEX Congress* (pp. 11–26). EURALEX.
Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*(3), 498–520.

Boonmoh, A., Singhasiri, W., & Hull, J. (2006). Problems using electronic dictionaries to translate Thai written essays into English. *rEFLections*, *8*, 8–21.

Bréal, M. (1900). *Semantics: Studies in the science of meaning*. Heinemann.

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small, G. Cottrell & M. Tanenhaus (Eds.), *Lexical ambiguity resolution* (pp. 477–508). Morgan Kaufmann.

Cantos, P., & Sanchez, A. (2001). Lexical constellations: What collocates fail to tell. *International Journal of Corpus Linguistics*, *6*(2), 199–228.

Cantos, P., Sanchez, A., & Almela, M. (2009). An attempt to formalize word sense disambiguation: Maximizing efficiency by minimizing computational cost. *RESLA*, *22*, 77–88.

Carston, R. (2021). Polysemy: Pragmatics and sense conventions. *Mind & Language*, *36*(1), 108–133.

Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 126–139). Longman.

Cruse, A. (1986). *Lexical semantics*. Cambridge University Press.

Cruse, A. (1992). Monosemy vs. polysemy. Review of Ruhl (1989). *Linguistics*, *30*, 577–599.

Dominiek, S. (1998). What linguists can and can't tell you about the human mind: A reply to Croft. *Cognitive Linguistics*, *9*, 361–478.

Evans, V. (2005). The meaning of time: Polysemy, the lexicon and conceptual structure. *Journal of Linguistics*, *41*(1), 33–75.

Evans, V., & Green, M. (2006). *Cognitive linguistics: An introduction*. Edinburgh University Press.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241–265.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645–666.

Geeraerts, D. (1990). The lexicographical treatment of prototypical polysemy. In S. Tsohatzidis (Ed.), *Meanings and prototypes: Studies in linguistic categorization* (pp. 195–210). Routledge.

Geeraerts, D. (2001). The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica*, *17*, 6–21.

Geeraerts, D. (2006). *Words and other wonders: Papers on lexical and semantic topics*. De Gruyter.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375–406.

Grefenstette, G., & Hanks, P. (2023). Competing Views of Word Meaning: Word Embeddings and Word Senses. *International Journal of Lexicography*, 1–9.

Hanks, P. (1990). Evidence and intuition in lexicography. In J. Tomaszczyk & B. Lewandowska-Tomaszczyk (Eds.), *Meaning and lexicography* (pp. 31–42). John Benjamins.

Hanks, P. (2000). Do word meanings exist? *Computers and the Humanities*, *34*(1/2), 205–215.

Hanks, P. (2002). Mapping meaning onto use. In M. Corréard, *Lexicography and Natural Language Proceeding: A festschrift in honour of B. T. S. Atkins* (pp. 156–198). EURALEX.

Hoey, M. (2012). *Lexical priming: A new theory of words and language*. Routledge.

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.

Jackson, H. (2013). *Words and their meaning*. Routledge.

Janssen, A. J. M. (2003). Monosemy versus polysemy. In H. Cuyckens, R. Dirven & J. Taylor (Eds.), *Cognitive approaches to lexical semantics* (pp. 93–122). De Gruyter.

Jorgensen, J. C. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research*, *19*(3), 167–190.

Kempson, R. M. (1977). *Semantic theory*. Cambridge University Press.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, *31*(2), 91–113.

Kilgarriff, A. (2007). Word senses. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation: Algorithms and application* (pp. 29–46). Springer.

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, *81*(1–3), 205–223.

Knowles, G., & Mohd Don, Z. (2004). The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, *9*(1), 69–81.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. University of Chicago.

Le, C. N. N., & Miller, J. (2025). Developing and piloting SemiMed — A resource for semi-technical medical vocabulary. *Research Methods in Applied Linguistics*. *4*(3), 1–18.

Le, C. N. N., & Miller, J. (2023). A core meaning-based analysis of English semi-technical vocabulary in the medical field. *English for Specific Purposes*, *70*, 252–266.

Le, C. N. N. (2022). Pilot SemiMed — a mini semantic visualization dictionary of semi-technical medical vocabulary: A response to semantic deficiencies in a medicine-related wordlist (Report on A.S. Hornby Dictionary Research Award Project). https://www.hornby-trust.org.uk/wp-content/uploads/2024/10/ASHDRA-report-Chinh-Ngan-Nguyen-Le-2022.pdf

Leech, G. (1974). *Semantics*. Penguin.

Lehrer, A. (1974). *Semantic fields and lexical structure*. North-Holland.

Li, E. S.-L., & Pemberton, R. (1994). An investigation of students' knowledge of academic and subtechnical vocabulary. In J. Flowerdew & A. K. K. Tong (Eds.), *Entering text* (pp. 183–196). The Hong Kong University of Science and Technology.

Lyons, J. (1968). *Introduction to theoretical linguistics* (Vol. *510*). Cambridge University Press.

Lyon, J. (1977). *Semantics*. Cambridge University Press.

Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, *1*(3), 165–188.

Moerdijk, F. (2003). The codification of semantic information. In P. Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 273–296). John Benjamins.

Murphy, M. L. (2010). *Lexical meaning*. Cambridge University Press.

Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography*, *15*(4), 277–305.

Ostermann, C. (2015). *Cognitive lexicography: A new approach to lexicography making use of cognitive semantics*. De Gruyter.

Palmer, F. R. (1995). *Semantics*. Cambridge University Press.

Panman, O. (1982). Homonymy and polysemy. *Lingua*, *58*(1–2), 105–136.

Perez, M. (2013). Identification and analysis of the specialized vocabulary of British Law Reports: A corpus-driven study of this legal genre at the core of common law legal systems [Doctoral dissertation, University of Murcia]. https://www.tdx.cat/handle/10803/128621?fbclid=IwAR29vxLl5QEpAcg19TKI6ExdP_9GJ92NunxxFLyTHcJYvz4aOKaKcyv2i10#page=1

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Ravin, Y., & Leacock, C. (2000). *Polysemy: Theoretical and computational approaches*. Oxford University Press.

Rizzo, C., & Sanchez, A. (2010). Building new meanings in technical English from the perspective of the lexical constellation model. *Ibérica*, *20*, 107–126.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233.

Ruhl, C. (1989). *On monosemy: A study in linguistic semantics*. State University of New York Press.

Thurston, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, *17*(3), 267–280.

Tyler, A., & Evans, V. (2003a). *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.

Tyler, A., & Evans, V. (2003b). Reconsidering prepositional polysemy networks: The case of over. *Trends in Linguistics Studies and Monographs*, *142*, 95–160.

Ullmann, S. (1962). *Semantics: An introduction to the science of meaning*. Basil Blackwell.

Van der Meer, G. (2004). On defining: Polysemy, core meanings and 'great simplicity'. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 807–815). EURALEX.

Vicente, A., & Falkum, I. L. (2017). Polysemy. *Oxford research encyclopedia of linguistics*.

Vidhu Bhala, R. & Abirami, S. (2014). Trends in word sense disambiguation. *Artificial Intelligence Review*, *42*(2), 159–171.

Watson-Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, *45*, 31–39.

Yamamoto, Y. (2014). Multidimensional vocabulary acquisition through deliberate vocabulary list learning. *System*, *42*, 232–243.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop* (pp. 266–271). ARPA.

Yu, A., & Trainin, G. (2022). A meta-analysis examining technology-assisted L2 vocabulary learning. *ReCALL*, *34*(2), 235–252.

Address for correspondence

Chinh Ngan Nguyen Le
School of Education
Adelaide University
Australia
chinhngan.nguyenle@adelaide.edu.au

Co-author information

Julia Miller
School of Education
Adelaide University
julia.miller@adelaide.edu.au

Publication history

Date received: 7 November 2024
Date accepted: 22 March 2026
Published online: 8 April 2026