Words that matter in L2 research and pedagogy: A corpus-linguistics perspective

A considerable amount of current vocabulary research draws on corpus linguistics methods (McLean, 2018). Here, we offer a critical evaluation of how words are conceptualized and word counts operationalised in vocabulary research that uses corpora as evidence. In our reflection, we comment on the main points from Webb (current issue) and discuss them in the light of the current trends in corpus linguistics. We first revisit the discussion of the appropriate unit of measurement in vocabulary research and argue in favour of lemma as the basic unit of analysis. Second, we go beyond the units proposed in the current debate, considering further directions in corpus-based vocabulary research. Third, we briefly review the impact of the software tools used in vocabulary studies.

The four lexical units at the centre of the current debate on measuring and counting words – word families (WFs), flemmas, lemmas and types – differ in how they group words according to their morphological connections. Whether the focus is on research, assessment or pedagogical applications, the key issue in this debate concerns the extent to which it is possible to use the knowledge of one item from a group of related lexical items to predict the knowledge of other items in the group. This may involve, for example, predicting whether a learner who understands the meaning of *watch* (noun) also understands the meaning of *watches* (noun, plural), *watches* (verb, 3rd person singular) and *watchful* (deverbal adjective). Given how central this question is, empirical evidence supporting the choice of different units is surprisingly limited (e.g. McLean, 2018; Brown et al, 2020; Stoeckel et al, 2020) and the discussion often relies on assumptions about i) morphological and semantic distance of lexical items and ii) learners' ability to recognise morphological links between different words and understand the meaning of words with inflectional and/or derivational affixes (e.g. *–(e)s, -ed, -er, -ful*). Needless to say, there is great variation in both. Generally, the larger the unit, the less predictable the semantic connections between individual items are. Learners' ability to use morphological connections for guessing and inferring meanings of more/less semantically related words is not only dependent on learners' proficiency and prior knowledge but varies considerably according to specific lexical items, as learners may be able to make appropriate semantic connections between some words (e.g., *predictable* and *unpredictable*) but not others (*valuable* and *invaluable*).

As far as corpus linguistics is concerned, there is no 'lemma dilemma' (Webb, current issue). A large majority of current corpora are lemmatized, allowing the identification and counting of lemmas with ease, while software tools such as #LancsBox and Sketch Engine also lemmatize corpora created by users, e.g., corpora of student essays compiled by teachers. The advantage of using lemma as a unit is that it is more precise and requires fewer assumptions about the (morphological and semantic) knowledge on the part of learners than other units (with the exception of type, which is, however, far too specific). While from the pedagogical perspective, the more broadly-defined units such as WFs may appear useful, as discussed by Webb (current issue), such expectations require further empirical evidence from different pedagogical contexts, as the usefulness of each unit will likely vary according to learners' needs (e.g. Brezina & Gablasova, 2017). Importantly, lemma can be very easily 'translated' into one of the broader lexical units, flemmas and WFs, for pedagogical or research purposes if required

(e.g. comparison with previous studies); in other words, if we use lemma as a unit of analysis we can easily convert this information into the broader units using a simple many-to-one mapping; the opposite process is not possible. Given these arguments in favour of lemma as a general unit of language analysis, it is somewhat surprising that a strong case is still being made for the use of units, WFs and flemmas, which require greater number of assumptions to justify their use. It seems prudent, given the current state of knowledge, for lemma to be considered the default unit in research and pedagogical applications, while further evidence is being collected.

Our reflection has, so far, focused on the lexical units highlighted in the current debate (Webb, this issue); we should, however, also consider other issues in operationalising words that are key for further progress in vocabulary research. While lemma is arguably the best choice given the current technology for automatic processing of data in corpus linguistics – lemmas can be identified reliably – it also suffers, to some extent, from issues connected with identifying lexical units based on their form rather than function. It lumps together homonyms and polysemous words such as *bank* (river side vs. institution) and *cloud* (meteorological phenomenon vs. network of servers), although to a considerably smaller extent than flemmas or WFs. Distinguishing between word meanings would undoubtedly be relevant for research, assessment and pedagogy (Gardner, 2007). To address this issue, *lexeme* (lemma + sense disambiguation) should be considered and technology for automatic lexeme identification should be further developed. In addition, the research should pay more attention to the role of multi-word units (MWUs) such as phrasal words (e.g., *make up* and *pay off*) and compounds (e.g., *guinea pig*) which function as single lexical items and are very frequent in language. Traditionally, the single- and multi-word items have been kept separate in lexical analyses, assessment as well as corpus-based pedagogical resources (e.g., wordlists). However, this approach largely contradicts theories of language processing and evidence about language use (e.g., Siyanova-Chanturia, & Pellicer-Sanchez, 2018). Moreover, if vocabulary research analyses constituent words of MWUs as individual items, this is likely to skew the estimates of the amount and type of knowledge necessary for comprehension and production. For example, while *go* is an extremely frequent form in language use, when counting its occurrence during language analysis, it is important to distinguish whether it was used as a single word or part of MWUs such as the phrasal verbs *go off (the milk has gone off*) or *go down* (*this decision will not go down well*), as such uses require different lexical knowledge for understanding or producing the word on the part of learners.

Finally, while the discussion about validity of different lexical units appears to judge these units solely on their theoretical merit, often the selection of a specific unit comes down to "researcher preference, computer limitations, or convenience" (Garner, 2007, p. 249). This in turn, has an impact on larger trends and directions in the field. A crucial role in this process is played by software tools currently available in vocabulary research and the units they employ. For example, following the implementation of WFs in RANGE, a number of more recent instruments (VocabProfile, AntWordProfiler and MultiLingProfiler) likewise offer WFs for the analysis of lexical coverage, while other tools (e.g. TAALES, LancsLex) offer lemmas. From the computational perspective, WFs are easiest to implement, which is a possible reason why the coverage tools which do not include morphological analysis necessary for

identification of lemmas opt for WFs. Practical expediency, however, is rarely a good guide in operationalization of complex constructs such as lexical coverage. Software tools have a two-fold impact on the field. On the one hand, they enable progress by making complex quantitative analysis of texts available to a broad range of researchers and practitioners; on the other hand, they potentially limit the progress by implementing specific methodological choices (such as the type of the lexical unit) and influencing the direction of research. For example, Stoekel et al (2020) reported that lemmas tend to estimate learners' knowledge better than flemmas but they still conclude that "for pedagogy, flemma-based lists may remain useful, since current vocabulary profiling software does not distinguish orthographically identical word forms" (p. 605), going on to note that "if such lists are used with learners, they would be of greater value if POS were explicitly noted for each entry". Likewise, the units such as lexeme or MWUs may be underplayed in the current debate because these constructs are more difficult to implement in software tools. However, the field of corpus linguistics is dynamic and always evolving in response to theoretical understanding of language and the need for methodological innovation. Researchers thus should be guided by the theoretical considerations and methodological precision striving for insights which can, in turn, drive further innovation in corpus analysis.

References

Brezina, V., & Gablasova, D. (2017). How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. *Applied Linguistics*, *38*(5), 764-767.

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The Most Appropriate Lexical Unit for L2 Vocabulary Research and Pedagogy: A Brief Review of the Evidence. *Applied Linguistics*.

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

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823-845.

Siyanova-Chanturia, A., & Pellicer-Sanchez, A. (Eds.). (2018). *Understanding formulaic language: A second language acquisition perspective*. Routledge.

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit?. *Applied Linguistics*, *41*(4), 601-606.