Crowdsourcing ratings for single lexical items: a core vocabulary perspective

Elena VOLODINA
University of Gothenburg, Sweden

David ALFTER
University of Gothenburg, Sweden; Université Catholique de Louvain, Belgium

Therese LINDSTRÖM TIEDEMANN
University of Helsinki, Finland

In this study, we investigate theoretical and practical issues connected to differentiating between core and peripheral vocabulary at different levels of linguistic proficiency using statistical approaches combined with crowdsourcing. We also investigate whether crowdsourcing second language learners' rankings can be used for assigning levels to unseen vocabulary. The study is performed on Swedish single-word items.

The four hypotheses we examine are: (1) there is core vocabulary for each proficiency level, but this is only true until CEFR level B2 (upper-intermediate); (2) core vocabulary shows more systematicity in its behavior and usage, whereas peripheral items have more idiosyncratic behavior; (3) given that we have truly core items (aka anchor items) for each level, we can place any new unseen item in relation to the identified core items by using a series of comparative judgment tasks, this way assigning a "target" level for a previously unseen item; and (4) non-experts will perform on par with experts

Volodina, E., Alfter, D., Lindström Tiedemann, T.: Crowdsourcing ratings for single lexical items: a core vocabulary perspective. Slovenščina 2.0, 10(2): 5–61.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2022.2.5-61

in a comparative judgment setting. The hypotheses have been largely confirmed: In relation to (1) and (2), our results show that there seems to be some systematicity in core vocabulary for early to mid-levels (A1-B1) while we find less systematicity for higher levels (B2-C1). In relation to (3), we suggest crowdsourcing word rankings using comparative judgment with known anchor words as a method to assign a "target" level to unseen words. With regard to (4), we confirm the previous findings that non-experts, in our case language learners, can be effectively used for the linguistic annotation tasks in a comparative judgment setting.

Keywords: core vocabulary and language learning, non-expert crowdsourcing, single lexical items, CEFR levels, comparative judgment

1 Introduction

We set out to explore two broader questions in this study, both in the context of second language acquisition: The *first question* concerns theoretical and practical issues connected to differentiating between core and peripheral vocabulary at different levels of linguistic proficiency – that is, which vocabulary is critical for learners to know at a particular level (i.e. learners *need to know* it) versus which vocabulary is *good to know*. In other words, is there common core vocabulary for learners at different levels, and does it behave differently from peripheral vocabulary? In connection to this, we apply statistics and crowdsourcing to examine whether there are any particular word behavior patterns that can help us differentiate between core and peripheral vocabulary, which we study through hypotheses 1–3, as introduced in Section 3.2.

The *second question* concerns theoretical and practical aspects of using second language learners as crowdsourcers for the task of linguistic annotation. In particular, can we use second language learners to rank vocabulary according to difficulty? We experiment with crowdsourcing as a method to identify the receptive proficiency level of previously unseen vocabulary items (henceforth called *unknown items*) in relation to confirmed core (and peripheral) items, and compare teachers' and learners' votes (hypothesis 4 in Section 3.2).

In essence, we ask the following overarching question: Can we use crowdsourcing to identify core and peripheral vocabulary for a certain level? The study is partly motivated by the practical need to classify unseen vocabulary by target proficiency levels as necessary input for the automatic generation of learning materials, and/or for automatic assessment of learner production. We start from a simple assumption that if we ask crowdsourcers to explicitly compare two items at a time, of which one is *core* (with a confirmed level) and the other is a new item (i.e. with an unknown level), then the latter will end up having a rank close to the core items of the level of proficiency which it belongs to. Thus, if we have good *anchor words* (i.e. established core words per level), *unknown* words should appear in relative proximity to the anchor words of the corresponding level after a round of comparisons and votes (Example 1). The current study is designed to investigate how true this assumption is.

Example 1: *Illustration of relative ranking of an unknown item*

For example, given the following nouns with known "target" levels (core/anchor items)

`A1 – party; A2 – view; B1 – variety; B2 – purchase; C1 – reliability`

we need to place the noun `pillar` relative to the vocabulary above.

To do that, we compare `pillar` to each of the words/or groups of words (i.e. Is `party` more difficult than `pillar` or vice versa? Is `view` more difficult than `pillar` or vice versa?) and collect votes from crowdsourcers. Based on the votes, we assign "difficulty scores" to each item in each comparison task. After collecting three to five votes for each possible mini-task, we can see where the collected scores point us. For example, in this hypothetical case it might have pointed to the proximity of scores between `reliability` and `pillar`, and hence the appropriateness of `pillar` at C1 level.

Two broader theoretical questions arise in connection with such an endeavor. One is the well-known issue of *what core vocabulary actually is* (e.g. Stein, 2017; Carter, 1982). The other is a relatively new topic connected to the *reliability of crowdsourcing non-expert judgments* as a method of producing linguistic annotations (e.g. Paquot et al., 2022; Alfter et al., 2021).

In short, we assume that **core** vocabulary at a certain proficiency level is *vocabulary known by all learners of **that** target language at **that** particular level*. In our current experiment we focus on items known receptively, i.e. items which can be understood but the learners do not need to be able to use them productively yet. We focus on lexical word classes (nouns, verbs, adjectives and adverbs) and further assume that all items that do not belong to the "core" vocabulary at a particular level but occur in texts aimed at learners of these levels are **peripheral** vocabulary (i.e. good-to-know).

The design of the study is inherited from Alfter et al. (2021), where best-worst scaling was used to crowdsource the relative difficulty of multiword expressions and compare annotations from second language professionals (*experts*), on the one hand, and from second language learners (*non-experts*), on the other. While the main focus of the study by Alfter et al. (2021) was to see how the design of a crowdsourcing task may influence the reliability of the linguistic annotation by experts and non-experts, the main task of the current study is to see whether anchor words (single lexical items) per level will be ranked consistently close together, and thus may serve as anchors to derive the levels of unseen words. If confirmed, such a property can be exploited by other languages for the (inexpensive) creation of similar resources. The secondary task of the current study is to confirm findings by Alfter et al. (2021) about experts and non-experts being able to produce comparable annotations in comparative judgment settings, this time tested on *single lexical items* instead of multiword expressions.

The study is performed on Swedish, but the methodology presented here is applicable to any language. In Section 2 we start with a short note on the notion of core vocabulary, how it has been applied to language learning, as well as a short introduction to crowdsourcing non-expert judgments. Sections 3, 4 and 5 introduce the experimental setup, item selection and practical issues. The results and analyses are presented in Section 6, followed by the discussion and conclusions in Sections 7 and 8.

2 Related work

In this section, we present some of the earlier work related to the focus the current study. There are three main axes that we explore: core vocabulary from a theoretical perspective (2.1), core vocabularies for language learning (2.2), and crowdsourcing linguistic annotations using non-experts (2.3).

2.1 Core vocabulary – a theoretical perspective

Core vocabulary may be assumed to comprise lexical items that are known to all users of a language and thus form a shared vocabulary that all users would be able to use and understand, which echoes the *Basic Language Cognition* theory of Hulstijn (2019). Therefore, it is useful, both theoretically and practically, to understand what makes a lexical item a core item and which properties are characteristic of these.

Several paradigms have been proposed for testing language vocabulary for lexical coreness, e.g. Lehmann (1991) or Bell (2013). Carter (1982) lists the following properties of core vocabulary (presented here in a significantly shortened form):

- Collocational span, i.e. core items will collocate with a wide number of other items, e.g. *fat* book, *fat* cat, etc.
- Semantic neutrality, i.e. core vocabulary will exhibit less stylistically colored and/or less specific meaning than other items with shared semantics, e.g. *thin* versus *skinny*, *undersized*, *scraggy*.
- Definitional power, i.e. core vocabulary tends to be used to explain other vocabulary, e.g. *smile* being used to explain *grin*, *smirk*, *beam*. Here core vocabulary will enter syntactic constructions to explain non-core items, e.g. *non-core noun (individual)=adjective+core noun (a single person);* e.g. *non-core verb (stroll)=core verb+adverbial (walk in a relaxed way),* etc.
- High placement in semantic networks, i.e. core vocabulary items tend to be hypernyms to a number of hyponyms, e.g. *flower* to *tulip*, *rose*, etc.
- Antonymy, i.e. core vocabulary often has an antonymous counterpart, which is less common in non-core vocabulary.

- A cognitive basis reflecting the (semantic and sociolinguistic) norms of the usage, i.e. more normative (unmarked) use is characteristic of core vocabulary, e.g. male species versus female: *lion* vs *lioness*.

The above criteria suggest that core vocabulary is *useful* in many fields, and the concept has been researched and applied in fields like lexicography (West, 1953; Brezina and Gablasova, 2015), language learning (Carter, 1987), comparative historical linguistics (Swadesh, 1971), diachronic lexicostatistics (Márquez, 2007), speech pathology (Crosbie et al., 2006) and other areas. Stein (2017, p.760) argues that *usefulness* is "a *function* of core vocabulary" and not vice versa (i.e. not all useful vocabulary can qualify to be part of core vocabulary). Similarly, the *high frequency* of the core vocabulary is a reflection of the *usefulness* of core vocabulary. Stein (2017) warns against using frequency and usefulness as defining characteristics of core vocabulary. These properties may be used as a proxy for identifying core items, but one needs to keep in mind that not all frequent or useful items belong to the core vocabulary; and not all core items are equally frequent or equally useful (cf. *zip-code*, *bread* or *toothbrush*).

Besides, lexis shows resistance to systematization (Carter, 1982), which implies fuzziness in the definition of the core vocabulary. Some items could exhibit two of the six properties above, and yet be considered core, while others may exhibit all six, altogether leading to different degrees of coreness. Dixon (1971) also claimed that adjectives and adverbs are harder to categorize in this respect.

2.2 Core vocabularies for language learning

Numerous attempts to identify the core, or common, vocabulary for language learners have been made, some prominent examples being the General Service List (West, 1953; Brezina and Gablasova, 2015), the English Vocabulary Profile (Capel, 2015), the Routledge series[1] of most frequent core vocabulary for learners (e.g. Familiar, 2021; Lonsdale and Le Bras, 2009), and a series of Kelly lists for several languages

1 https://www.routledge.com/Routledge-Frequency-Dictionaries/book-series/RFD?a=1

(Kilgarriff et al., 2014). Strategies for the selection of lexical items for inclusion have been different in different resources – *from* strict frequency indications based on various types of corpora *to* combinations of intuitions, judgments of importance, frequency indications and overlaps between concepts in the different languages (see Kilgarriff et al. (2014) for the latter). All of the lists claim that the identified vocabulary is useful for learners. The connection between the *objective token frequency* and the notion of *usefulness*, however, is not always clear (cf. Stein, 2017). Nonetheless, even though such lists will never be beyond criticism at a theoretical level, they make it possible, with a certain degree of objectivity, to address some central assumptions about vocabulary and its hypothetical importance to language as a system and to language learners in particular.

Several lists have been compiled for Swedish with language learners in mind, such as SVALex (François et al., 2016), SweLLex (Volodina et al., 2016), NyLLex (Holmer and Rennes, 2022), the Kelly-list (Volodina and Johansson Kokkinakis, 2012a, 2012b) and SweVoc (Mühlenbock and Johansson Kokkinakis, 2012), each of which has been compiled on different corpora. The unifying lexical unit for these lists is the *lemgram*,[2] i.e. a combination of lemma, its part-of-speech and its inflectional paradigm, e.g. lemgrams *can, verb* (can-could) and *can, verb* (can-canned) will have two separate entries in a list.

Holmer and Rennes (2022) compared two lists, SVALex and SweVoc (both generated from reading materials), with NyLLex (also based on reading comprehension texts) for overlaps, and identified that they have approximately 52–68% overlap. There was a 40% overlap between SweLLex (based on learner essays) and NyLLex. This suggests that the overlapping 40–50–60% of vocabulary definitely belongs to the core vocabulary that is useful for learners. This is not to say that other, non-overlapping, items do not belong to core vocabulary. In the non-overlapping cases, there are other characteristics that would qualify vocabulary to be included in the core, as outlined in Section 2.1. Holmer and Rennes (2022) further correlated the indications of proficiency levels in SVALex, where CEFR level indications are inherited

2 For better readability we use the shortened term lemma in the rest of the article to refer to lemgrams.

from the texts used for teaching at these levels, and the readability levels (1–6) used in the NyLLex resource, identifying that approximately 20% of the vocabulary items per level overlap at exactly the same levels (i.e. CEFR A1 in SVALex with Level 1 in NyLLex).

Lexical resources like the ones described here are very valuable for language teaching and for the development of teaching materials. However, there will always be some items that have not been included in the lists, or have not been marked for appropriateness at certain levels of proficiency (or readability). Teachers, test developers, and assessors alike will thus need a method that would allow them to place new lexical items in relation to the items on the list. We are experimenting with ways to address this issue in this study, where we use experts and non-experts to classify unseen (unknown) vocabulary in relation to items of known levels in a crowdsourcing experiment.

2.3 Crowdsourcing linguistic annotation from experts versus non-experts

Kullenberg and Kasperowski (2016) have shown that use of non-experts for scientific projects has been increasing drastically since 2010, primarily in the fields related to natural sciences and medicine. Their analysis demonstrates that non-experts are successfully used for data collection and classification, and are able to perform expert tasks on par with experts. However, the use of non-experts for *linguistic analysis/annotation* is much less researched and continues to pose methodological questions. Below follows a small overview of studies involving crowdsourcing non-expert judgments for linguistic annotation at different levels of linguistic analysis.

Kosem et al. (2018) used a crowd for the task of *sense disambiguation of collocations* in a dictionary project. Their results show that the crowd agreed in 83% of cases, and that the benefits of using a crowd for linguistic annotation are much higher than the costs of employing experts. Lau et al. (2014) employed crowdsourcing for *grammaticality judgments on a sentence level* (binary judgments and gradual ones). The users were filtered through the use of five control items which the authors knew the answers to. Annotations from users who failed to

pass the test were not considered in the analysis. The rest of the (filtered) crowd demonstrated consistency in annotations. Unfortunately, the study did not explicitly compare the output from experts and non-experts. De Clercq et al. (2014) designed an experiment involving experts and non-experts for the task of *ranking documents by readability* using crowdsourcing for the non-experts. Experts annotated the documents for readability directly, while non-experts were given a relative ranking task, i.e., determine which one of two texts is more readable. They found that the non-experts and experts agreed to a large extent (with a Pearson correlation coefficient of 0.90).

Similarly, Alfter et al. (2021) and Lindström Tiedemann et al. (2022) compared the judgments of experts and non-experts on the task of *ranking* Swedish *multiword expressions by difficulty*, explicitly studying the reliability of second language learners (non-experts) as annotators. The experts performed a direct annotation with CEFR levels, in addition to the crowdsourcing experiment, while the non-experts (learners) only participated in the crowdsourcing experiment, in which they were asked to indicate the "easiest" and "hardest" of four multiword expressions, a technique called *best-worst scaling* (Louvière et al., 2015). The study found that experts and non-experts agreed to a large extent (with Pearson correlation coefficients between 0.81 and 0.93). Alfter et al. (2022) adopted the same methodology as in Alfter et al. (2021) – albeit for French, and on *word senses*. They arrived at the same conclusions: non-native speakers (non-experts) and native speakers (experts) largely agree about the difficulty of word senses. This again lines up with previous research investigating the reliability of non-experts in tasks normally requiring expert knowledge. Paquot et al. (2022) set essay assessment into a comparative judgment paradigm, employing both trained assessors (experts) and non-trained academics (non-experts). The results clearly show that the two groups exhibit high similarity in their assessments, thus demonstrating that an untrained crowd can be used reliably for essay assessment tasks.

The short overview presented above demonstrates the use of non-experts for annotation tasks on different linguistic levels: multiword expressions and collocations (Alfter et al., 2021; Kosem et al., 2018), sentences (Lau et al., 2014; Alfter et al., 2022) and texts

(De Clerq et al., 2014; Paquot et al., 2022). A number of these studies have contrasted the use of non-experts and experts, demonstrating that the task design has a crucial impact on the reliability of the annotation results, and in particular that the setting of comparative judgments (e.g. easier–more difficult) yields a high correlation between experts and non-expert annotations. Only one study to date has explicitly tested the use of second language learners for annotation tasks (Alfter et al., 2021), although there were also non-native annotators in Alfter et al. (2022). In the current study we replicate the experimental setting from Alfter et al. (2021), using second language learners alongside second language professionals, but for annotation on the relative difficulty of single lexical items, looking for additional proof to support the findings in Alfter et al. (2021) and Lindström Tiedemann et al. (2022).

3 Methodology and experimental setup

To perform the experiment we need to separate vocabulary items we can observe at each particular CEFR level[3] in our data into core and peripheral (i.e. non-core) vocabulary for that level.

3.1 Core vocabulary for each level of proficiency

The attempts at identifying core vocabulary useful for language learning leads us to asking probably the most intriguing question in connection to this study: **Is there a core vocabulary for each level of proficiency?**

Stubbs (2001, p. 41) defines core words as "…known to all native speakers of the language […] that portion of the vocabulary which speakers could simply not do without". We adapt Stubbs' definition of core vocabulary to our context as follows: **core** vocabulary at a certain proficiency level is *vocabulary known by all learners of **that** target language at **that** particular level*. In our current experiment we focus on items known receptively, i.e. items which can be understood, but the learners do not need to be able to use productively yet.

3 Levels here are represented by the scale used in the Common European Framework of Reference (CEFR, COE 2001), representing 6 levels: A1 (beginner), A2, B1, B2, C1, and C2 (near native).

Given this definition of core vocabulary, we assume that most items from closed (functional) word classes should by default belong to the core vocabulary. Therefore, we focus on lexical word classes which do not demonstrate a similar stability historically, due to the tendency to develop new senses, new collocations, and include new members through borrowings or word formation mechanisms.

Further, we assume that all items that do not belong to the "core" vocabulary at a particular level but occur in texts aimed at learners of these levels, are, by the definition above, **peripheral** vocabulary (i.e. good-to-know) or maybe even incidental (i.e. appearing at a level prematurely due to some "non-pedagogical" reasons or needs, e.g. archaic forms in poetry).

Previous research on core vocabulary indicates that items at proficiency levels above B1 do not belong to the common language core vocabulary. Hulstijn's (2019) theory of *Basic Language Cognition* suggests that all speakers of a certain language, even first language speakers, could manage with the vocabulary and grammatical structures that roughly correspond to what second language learners can be expected to have acquired by the time they complete B1 level. This fact suggests that it might be more challenging to identify items that *all* learners at B2–C2 levels would need to know. The idiosyncrasy of the vocabulary which learners at these levels acquire depends on the interests of the learners, professional specialization, and many other aspects, since at these levels lexical variety, lexical sophistication and specialized vocabulary become the most dominant vocabulary features. In fact, the attitude to core vocabulary becomes explicitly negative in the research on advanced language learning and academic writing, where general purpose language is no longer a focus (e.g. Granger and Larsson, 2021). At this level general vocabulary (often also high-frequency vocabulary) is expected to be replaced by more formal specialized alternatives.

One way or another, we see a strong incentive:

(1) To identify core vocabulary that we may expect all learners at that level to acquire – as an input to the theoretical discussions on the nature of core vocabulary, as well as an input to practical applications within ICALL and assessment. In search of strategies to achieve this,

we work with a set of tools – CEFR-tools – and apply personal judgments to separate vocabulary into core and periphery (see Section 4).

(2) To identify such items among core items for each level that could function as reliable *anchors* or *anchor words* when developing a method for placement of unknown items on a scalar vocabulary list. We prefer in this connection to use the term *anchor* instead of *core* for these items (see Section 4.2). *Anchor* is less charged and avoids the unnecessary confusion between practical exercises like the one we perform in this study and the ongoing theoretical discussions about the nature of core vocabulary in general.

All items that have been observed in our corpora at particular levels but which cannot be classified as core are classified by default as non-core, i.e. peripheral – which optionally may be further classified into more subclasses, e.g. as incidental vocabulary.

Words that have not been observed in our corpora have *unknown* status, and eventually need to be classified into either one of the CEFR levels, or outside the CEFR scope.

3.2 Hypotheses and study overview

Based on the short overview of the related work presented above, we have formulated four *hypotheses* for the current experiment:

1. There is a common *core* vocabulary at *A1–B1 levels*; there is less systematicity at *B2–C2* levels, a hypothesis that is based on assumptions in the Basic Language Cognition theory of Hulstijn (2019).
2. Some systematicity can be observed in the behavior of the *core* items, but less so in the *peripheral* items.
3. Through crowdsourced comparative judgments, *unknown*[4] vocabulary items will demonstrate a perceived difficulty (expressed in numerical scores) equal or comparable to the perceived difficulty of *anchor* items of a particular level (see Example 1).
4. *Non-experts* will perform on par with *experts* in a comparative judgment setting, similar to the results in Alfter et al. (2021).

4 *Unknown* in this context means that the item is not represented in the CEFR-graded lexical resource we have at hand.

The overarching procedure for the experiment is straightforward, even though its implementation – technically and theoretically speaking – is challenging, as will become clear from the text that follows:

- From textbooks, select five (5) *core/anchor* items and five (5) *peripheral* ones per level of proficiency (A1–C1).
- From general language corpora, select two (2) *unknown* items per five (5) frequency bands – i.e. items not represented in the textbooks, but that represent different frequency bands (e.g. 1–1,000 most frequent items; 1,001–2,000; etc.) in other resources.
- Mix all item types in tasks for comparison of *perceived difficulty* of items against each other using best-worst scaling (Louviere et al., 2015).
- Collect votes separately for *experts* (second language professionals of the language in question, Swedish in our case) and *non-experts* (second language learners of the language in question, Swedish in our case).
- Analyze the resulting order of items, focusing on the behavior of the *core/anchor* items, *peripheral* items and *unknown* items using linear scales, and clustering as means of visualization.
- Analyze the resulting order comparing *experts* and *non-experts* as annotators, calculating correlations between the two groups.

An overview of the study setup is shown in Figure 1. The sections below expand on each of the steps of the study.

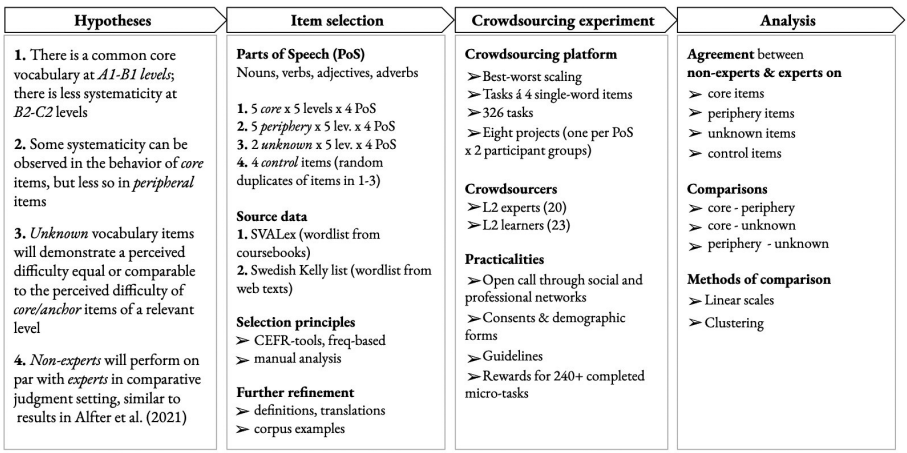


Figure 1: Overview of the study.

4 Item selection

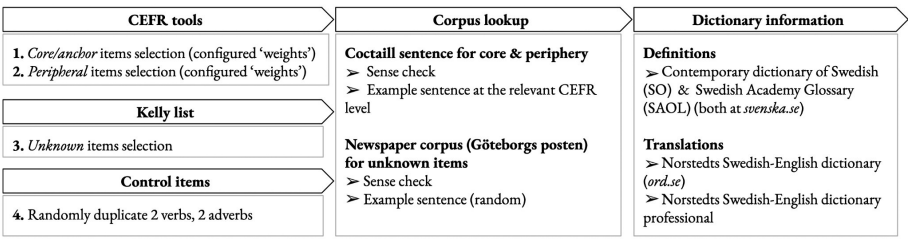


Figure 2: Overview of item selection procedure.

Figure 2 graphically represents the process of item selection for the experiment. For each lexical part-of-speech (PoS) – noun, verb, adjective, adverb – we selected 12 items per CEFR level, split into three different groups:

a) Core/anchor items of a certain CEFR level in the coursebook data in Coctaill (five items) (4.2);
b) Peripheral items of a certain CEFR level in the coursebook data (five items) (4.3);
c) Unknown items which should not appear in the coursebook data at all, with some exceptions (two items) (4.4).

Among the selected items, we also randomly selected two control items in two of the parts-of-speech (i.e. four items in total) to control for the systematicity and reliability of the annotations. These items are duplicates of the already included verbs (*underskatta* 'underestimate'; *förebygga* 'prevent') and adverbs (*således* 'consequently'; *sammanfattningsvis* 'summing up'). The hypothesis with control items is that if the annotations are chaotic, then these items will end up far from each other on the resulting linear scale. If, on the contrary, they end up close to each other, we can assume that the ranking is replicable even with new participants and therefore reliable. We suspected that peripheral items may be more unsystematically annotated and therefore included three control items for periphery and only one for core (*således* 'consequently').

Core/anchor and peripheral lexical items for the experiment were selected from **Coctaill** (Volodina et al., 2014), a corpus of coursebooks

for Swedish as a second language published in Sweden between 1997 and 2014, and intended for adult learners. The coursebooks in Coctaill have been linked to CEFR levels with the help of teachers and these levels have been projected to the texts in each book and consequently to lexical items in those texts. This way we can see on which levels lexical items occur in the coursebooks. As a means of filtering the items in Coctaill and helping us select the best candidates for core and periphery we used *CEFR tools*[5] (see 4.1).

The unknown lexical items were picked from the **Swedish Kelly-list** (Volodina and Johansson Kokkinakis, 2012a, 2012b). The Kelly list includes CEFR level indications which are based on frequency bands of approximately 1,500 items (cf. Kilgarriff et al., 2014) and we picked two items per level and PoS from A1–C1.

CEFR proficiency levels focus on communicative abilities and should primarily be thought of as a continuum (COE, 2018, p. 34, cf. Ortega, 2012). Communicative skills can often be achieved through different grammatical and lexical means, and hence it can be difficult to link specific lexical items (single or multiword) to a particular proficiency level. Still, lexical control (COE, 2001, p. 112) and vocabulary size are clearly part of the linguistic competences which a learner has to acquire (COE, 2001, p. 108). The original CEFR publication claimed that detailed lists of vocabulary should be possible to specify for each language (e.g. *Threshold level* 1990) (COE, 2001, p. 30) and encouraged attempts to link communicative tasks to specific vocabulary (COE, 2001, p. 33). The authors note: "Users of the Framework may wish to consider and where appropriate state:

- which lexical elements (fixed expressions and single word forms) the learner will need/be equipped/be required to recognise and/or to use;
- how they are selected and ordered." (COE, 2001, p. 112).

In Alfter et al. (2021) and Lindström Tiedemann et al. (2022) items were linked to their first level of occurrence in the coursebooks, regardless of how many books they appeared in or whether they recurred at later levels. This method of level assignment may be too simplistic,

5 https://spraakbanken.gu.se/larkalabb/cefrtools (publication under preparation).

which is one of our reasons for investigating both core and periphery vocabulary in this study.

4.1 CEFR Tools

To select items we used output from *CEFR tools* (Alfter, 2021), as illustrated in Figure 3, which shows how lemmas (i.e. base forms) are used at different levels in the coursebook corpus Coctaill (Volodina et al., 2014) and in learner essays (the SweLL pilot corpus, Volodina et al., 2016). It also predicts the level of unseen items with the Coctaill LM-score and the SiWoCo-score (see below for more information).

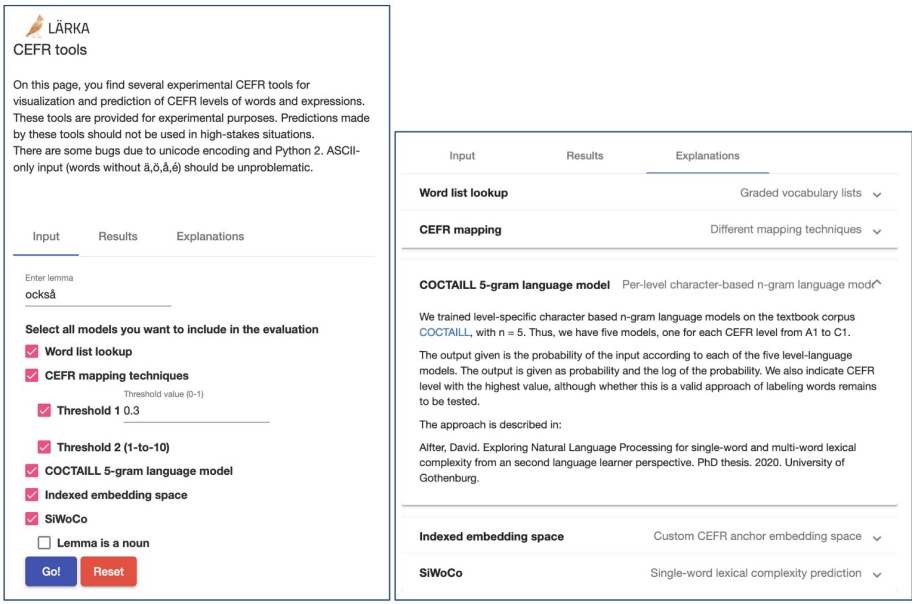


Figure 3: CEFR tools in Lärka (Alfter et al., 2019).

CEFR tools use various algorithms and techniques to indicate (or predict, depending on the algorithm) CEFR levels for Swedish words, both for known and unknown vocabulary (Alfter, 2021).

Word list lookup returns the level of first occurrence in SVALex (François et al., 2016) for receptive vocabulary and SweLLex (Volodina et al., 2016) for productive vocabulary.

CEFR mapping techniques uses two threshold techniques to derive a level from the underlying distribution across levels, one based on a

variable threshold (Threshold 1 in Figure 3, fixed at 0.3[6]; Alfter et al., 2016) and one based on a fixed threshold (Threshold 2 (1-to-10) in Figure 3; Hawkins and Filipovic, 2012). The first threshold technique assigns as the level that at which a word occurs 30% more often that at the previous level. The 1-to-10 threshold technique assigns as level that at which a word occurs at least ten times as often at the previous level. CEFR-mappings do not produce predictions for unseen words, being based on observed frequencies, but may deviate from the first level of occurrence.

COCTAILL 5-gram language model (Coctaill LM for short) uses character-based n-gram language models trained on subparts of the COCTAILL corpus, with one language model per CEFR level. For prediction, each of the language models calculates the probability of the word belonging to the language model and the highest scoring model is used as a prediction. This model can also predict levels for unseen words, i.e. words not included in the coursebook data (*Coctaill*).

Indexed embedding space uses two models that are trained by injecting the CEFR levels as *words* into the embedding space (cf. Alfter et al., 2016, 2021; Wang et al., 2018): first of all, a linear model in which the training data is used as-is, and secondly, a shuffled model, in which the training data was shuffled prior to training. The results show that the shuffled model seems to generalize better.

Finally, *SiWoCo* (Single Word Complexity) automatically extracts numerous word-level features to predict both a receptive and a productive level at which the word should be possible to understand and produce, respectively, and can predict levels for unseen words as well (Alfter and Volodina, 2018).

In addition to the CEFR tool scores, we calculated the following metrics based on the automatic predictions: homogeneity, majority level and percentage agreement.

We define *homogeneity* as a weighted score that takes into account the divergence in levels from the majority level. *The majority level* is defined as the level that most methods agreed upon. In cases of a tie,

6 The threshold value is fixed at 0.3 (a value which can be adapted) but the underlying frequency distributions are transformed to fit into the interval [0,1], thus in effect this is a variable threshold, even if the value is fixed.

the majority level is taken as the first level (alphabetically). Example: If out of eight predictions, seven methods resulted in a prediction of A1 and one method in A2, the homogeneity would be greater than if seven methods produced A1 and one method produced B1, since the difference from A1 to A2 is smaller than the difference from A1 to B1. The maximum value is 1, while the minimum value is bound by the number of predictions and can be negative (with a minimum of around -1.8 for eight predictors, depending on the predictions). Homogeneity does not contain information about which level the agreement was on, and distances between levels are treated as equal, as the purpose of this measure is to measure the proximity of predictions (i.e. the homogeneity score is the same if seven predictions say A1 and one says B1, or if seven say B2 and one produces C2).

More formally, homogeneity *H* for item *x* with CEFR predictions $p \in P$ is calculated as shown in Equation 1:

$$H(x) = \frac{c(m(P))}{c(P)} - \Sigma_{p \in P, p \neq m(P)} \left(\frac{|p - m(P)|}{c(P)} * c(p)\right) \quad (1)$$

where *c(P)* is the count of predictions for item *x*, *m(P)* is the majority level in *P* (the first item alphabetically in case of ties), *c(m(P))* is the count of the majority level, and *c(p)* is the number of times *p* was predicted by different methods.

We also use *percentage agreement* (Scott, 1955) as a score to indicate the agreement of the different predictions. An agreement of 1 means that all predictions agree. Note that this score does not contain information about which level the agreement was on, and is calculated as the count of the *majority level* divided by the number of predictions. For example, if out of eight predictions seven methods produce A1 and one method B1 (or for that matter A2, B2 or C1), the agreement would be ⅞.

Finally, we supplemented the data with frequency information from SVALex, namely the relative frequency overall, per CEFR level and the number of documents which contained the item.

4.2 Selecting core items

To pick core items we started by selecting a *receptive majority level* in CEFR-tools (e.g. A1), since we were primarily interested in the receptive proficiency of the learners. We then selected the highest possible level of *agreement* among all items at that level. If this did not give enough items to work with, we added the next agreement level and carried on like this until we had enough items. Below we give some examples to demonstrate this approach, see also Table 1.

(1) *Också* 'also' was picked as an A1 core adverb, and hence the receptive majority level was A1 in CEFR tools. The *agreement score* was filtered to be as high as possible and for this item the receptive agreement score is 0.875 which is basically as high as it is possible to get without the agreement being complete (1). In addition, we note that the receptive *homogeneity score* for this item is 0.75 and hence also very high.

(2) *Ledsen* 'sad' was picked as a core A2 adjective. The receptive majority level was estimated to be A2 in CEFR tools. The agreement score was filtered as high as possible and for this item it is also 0.875, while the homogeneity score is 0.625. Both scores are thus very high, but homogeneity is slightly lower than for the adverb *också*. The item was picked anyway since there was a lack of (appropriate) items with higher homogeneity – higher homogeneity was observed for rather international words that we considered inappropriate for our experiment, such as *modern* 'modern', *obligatorisk* 'obligatory', and *ironisk* 'ironic'.

We tried to keep the level in *Coctaill LM* (Coctaill-based Language Model) and *SiWoCo* (Single Word Complexity prediction model) the same as the level under selection whenever possible. We excluded any items where the scores were more than one level out from the receptive majority level we had selected.

(1) The core item *också* (adverb A1 core) has Coctaill LM A1 and SiWoCo A2 and was hence within the range of what we allowed in these measurements.

(2) The adjectival core item *ledsen* (A2) has a Coctaill LM measure of A2 and SiWoCo A2.

If there were still items with a level prediction from one of the measurements in CEFR tools that was more than one level above the receptive majority level we excluded those as well.

Table 1: *Scores from CEFR tools for the selection methods including four examples, and their corpus frequencies*

CEFR tools and Corpus frequency look up	core range (rule of thumb)	också.adv 'also' A1.core	ledsen.adv 'sad' A2.core	granska.vb 'inspect' B2.per	olycka.nn 'accident' A1.per
Word list lookup					
first occurrence (receptive) SVALex	actual lev.	A1	A2	B2	A1
first occurrence (receptive) SenSVALex	actual lev.	A1	A2	B2	A1
CEFR mapping					
threshold 0.3 (receptive)	actual lev.	A1	A2	C1	B1
threshold 1-to-10 (receptive)	actual lev.	A1	A2	B2	A1
Coctaill language model					
5-gram (prediction)	±1 level	A1	A2	A2	A2
Indexed embedding space					
linear model	±1 level	A1	A2	A2	A2
shuffled model	disregarded	A1	B2	C1	B2
SiWoCo prediction					
receptive (prediction)	actual lev. (±1)	A2	A2	A1	A1
Majority level					
receptive	actual lev.	A1	A2	B2	A1
productive	> actual lev.	A1	A2	C1	B2
Homogeneity					
receptive	0.6 - 1.0	0.75	0.625	-0.75	-0.375
productive	disregarded	0.2	1	-0.2	0.2
Agreement					
receptive	0.7 - 1.0	0.875	0.875	0.375	0.5
SVALex frequency look up					
receptive freq (total-relative)	top freq	32,751,295	714,988	78,403	646,122
receptive freq (total-docs)	top freq	422	18	5	20
receptive freq (level-relative)	top freq	44,397,093	1,556,716	61,296	31,368
receptive freq (level-docs)	top freq	27	5	1	1
Coctaill coursebook inclusion					
# books at the current level (max.≈4/lev.)	> 1 book	3	3	1	1
# books at the next level up (max.≈4/lev.)	≥ 1 book	4	3	2	1

Furthermore, we excluded *productive majority levels* that were lower than the selected receptive majority level, i.e. if we had selected B1 as the receptive majority level the productive majority level based on the SweLL pilot corpus should not be A1 or A2, but could be B1–C2.

(1) For *också* (adverb A1 core) the productive majority level was A1.
(2) *Ledsen* (adjective A2 core) had A2 as the productive majority level.

For adjectives we actually sidestepped our guideline regarding the productive majority level for some core items, and selected some items which have a lower productive majority level than the receptive majority level, e.g. *duktig* 'clever, capable' (A2 core adjective) which has A1 as the productive majority level, most probably due to the word being very frequent in spoken language as a form of praise.

Apart from the more international words mentioned above, *duktig* was the only A2 adjective with relatively high overall scores that suited our needs. It appears in six coursebook documents at the A2 level and on all higher levels, too.

We selected our *core lexical items* based on the items remaining after filtering according to the above principles. If there were still a lot of items to choose from we inspected the frequency information per book. Core items should occur in more than one book, or at least in more than one text. For nouns we tried to make sure that the items were from different topics, and that the items appeared to be reasonably well related with the core topics of that CEFR level according to the CEFR documentation (COE, 2001; 2018). The same was not possible with all PoS, since there were not always so many items per level that we could choose from or no clear topical domain.

(1) *Också* (adverb A1 core) appeared in 27 documents at A1-level and appeared frequently in documents at all other levels in the coursebooks.
(2) The adjectival core A2 item *ledsen* appeared in 5 documents on the A2-level and was present in documents on all levels above A2, and not at all on A1. Three of four books on A2-level contained the word.

After having selected potential core items we checked all items in Coctaill to see how they were used in actual texts in Coctaill.

4.3 Selecting peripheral items

Peripheral items were similarly picked by first choosing the *receptive majority level*, e.g. A1. We then selected that the lemma should ideally only appear in one (or a maximum of two) documents at that level, and we also checked the number of documents on the next level up that used the item. We deselected the current level in Coctaill LM and selected as low *agreement* as possible. Finally, the *productive majority level* should not be the same as the selected receptive majority level, but rather it should be higher in accordance with the assumption that productive proficiency often comes after receptive proficiency. This approach proved very difficult for C1 adverbs, since there were too few items, and most of them have been used as core.

(1) The verb *granska* 'to inspect/check' was picked as a peripheral B2 item. It fulfilled the requirement of appearing only in one document at that level, and also appearing in rather few documents on the next level (four documents). Coctaill LM was A2, and thus different to the level aimed at, and the receptive agreement score was 0.375, and hence very low. Receptive homogeneity was also very low, -0.75. The majority productive level was C1, and hence higher than the level aimed at.

(2) The noun *olycka* 'accident' was picked as a peripheral A1 item. It occurred in only one document at that level, and also appeared in only one document at A2, after which it became slightly more common appearing in five documents at B1-level, seven at B2, and six at C1. Coctaill LM was A2, and hence slightly higher than the level aimed at. The receptive agreement score was 0.5 and receptive homogeneity was -0.375. The majority productive level was B2, and hence clearly higher than A1.

4.4 Selecting unknown items

Finally, we selected two lemmas that were not in the coursebooks, for each level and part-of-speech, from the Kelly list from each frequency band (1–5) associated with the CEFR levels A1–C1. We had trouble choosing adverbs, especially at the A1-level, since most of the adverbs in Kelly were also included in Coctaill. Sometimes we thus made an

exception and selected items that were also in Coctaill, but from higher CEFR levels and with very low frequency.

4.5 Translations, definitions and examples

All lemmas were selected with a given part-of-speech and in a certain sense, effectively disambiguating polysemous words. For each item, we selected example sentences from Coctaill (*core* and *periphery*) and from the *Göteborgsposten* corpus[7] (for *unknown* items).

We also provided definitions of the Swedish words, based on two dictionaries: *Svensk ordbok* (SO, Contemporary dictionary of the Swedish Academy) or *Svenska akademiens ordlista* (SAOL, The Swedish Academy Glossary), both available at svenska.se, and translations to English, primarily based on Norstedts Swedish-English online dictionary (ord.se), but supplemented with some additional translations from the dictionary *Norstedts svensk-engelska ordbok professionell* (Norstedts Swedish-English dictionary, professional edition).

Example sentences, definitions and translations were included in the crowdsourcing experiment as an extra feature (if you clicked on one of the items, these would appear).[8]

5 Crowdsourcing experiment

The current study is a replication of Alfter et al. (2021) with regard to the use of best-worst scaling for crowdsourcing linguistic annotation (Louviere et al., 2015). The same number of items per project has been selected (60), and the same redundancy-reducing combinatorial algorithm has been used, resulting in the same number of micro-tasks per project (326). Likewise, we deploy the projects on the pyBossa[9] platform based on an open-source customizable framework for crowdsourcing tasks developed by SciFabric. We apply the same strategy to convert votes from the crowdsourcers into linear scales for further exploration.

Three things differ: Unlike the previous study, we (1) focus on the ranking of single items (as opposed to multiword expressions in

7 *Göteborgsposten* – a newspaper published in Gothenburg.
8 All of the corpora are available through Korp (Borin et al., 2012).
9 https://pybossa.com

Alfter et al., 2021), (2) we investigate the behavior of core, peripheral and unknown vocabulary (as opposed to the focus on the effects of design of an annotation task), and (3) we explore clustering as a method for visualizing and disentangling the results of the linear scale approach.

5.1 Practicalities and implementation

We implemented the *best-worst scaling projects* on pyBossa, a crowdsourcing platform. For each participant group (expert and non-expert) we set up separate projects for nouns, verbs, adjectives and adverbs, in total eight pyBossa projects (4 x 2 groups). Each project contained 326 micro-tasks (see Figure 4 for an example of a micro-task). Each micro-task contained four single lexical items and a possibility to mark one of them as the easiest (to the left) and one of them as the most difficult (to the right). Clicking on an item would open a field below the task showing the lexical item's definition, translation and an example of its use in a sentence from Coctaill (core and periphery) or from the *Göteborgsposten* newspaper corpus (unknown). The participants could see how many tasks they had completed, as well as open a feedback form and leave a message for us.

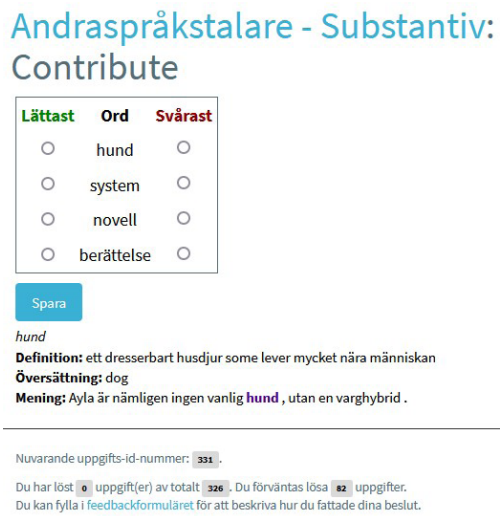


Figure 4: Example of a micro-task for nouns in pyBossa.

Following the traditions of crowdsourcing, we issued an *open call for participation* (Fort, 2016). All participants were recruited using our professional and social networks. A small *reward*[10] was promised to any participant who completed at least 240 micro-tasks, which was estimated to take a maximum of two hours, with an estimated 30 seconds per micro-task, and in fact took 2 hours on average.[11] Our intention was to collect at least three votes per micro-task from L2 learners (from now on 'non-experts') and three votes per micro-task from L2 professionals (i.e. L2 experts), in accordance with the findings in Alfter et al. (2021).

Before starting the projects, participants were asked to give their *consent*[12] to collect some *demographic information* about them. The latter included information about their gender, year of birth, country of residence, highest education level, native language(s), self-assessed level in Swedish, and an email for linking their pyBossa accounts with the demographic profiles as well as for further contact. For those who marked 'L2 professional' (from now on 'experts'), an additional question was asked about teaching experience counted in the number of years and level/type of teaching (elementary school, high school, Swedish for adults, etc). The demographic information was necessary to separate the participants into experts and non-experts and thus pursue our research interests (hypothesis 4).

On completion of the form, participants received an email with links to the relevant pyBossa projects and *guidelines*[13] in Swedish. Swedish was used in the guidelines as a way to filter participants with insufficient knowledge of the language (we aimed at B1 or more advanced speakers of Swedish). The guidelines included information on the purpose of the experiment, instructions on how to c*reate a pyBossa account*, details about the four part-of-speech-based *pyBossa projects* and explanations of *how to complete micro-tasks* (see Table 2 for the exact formulation of the task).

10 All participants who completed 240 micro-tasks received a digital voucher for the Amazon online store.

11 Based on the average time per task as detailed in Section 6.4. However, it should be noted that not every participant completed 240 tasks and that the time per task contains outlier values which skew the actual values.

12 Consents and socio-demographic information were collected via an online form,

13 Guidelines: https://docs.google.com/document/d/1gROsxmo4UPoe-bOPKYKJ6Z58tYKMrSwn-Jgt-Vn1GHL0/edit?usp=sharing

Table 2: *Excerpt of Guidelines with the definition of the task*

Guidelines (in Swedish)	Translation into English
3.1 Beskrivning av uppgifterna	**3.1 Description of the task**
[...] Du får se fyra (4) ord i taget, och din uppgift är att markera vilket ord som är **svårast att förstå** av de fyra, och vilket som är **lättast att förstå** av dessa fyra (relativ svårighetsgrad). Med "förstå" menar vi att *kunna förstå ord i en text som man läser på egen hand*. [...]	[...] You will see four (4) words at a time, and your task is to mark which word is **the most difficult to understand** out of the four, and which one is **the easiest to understand** (relative difficulty). By "understand" we mean *to be able to understand the word in a text that you read on your own*. [...]
Efter vi har samlat in röster (rankningar) från flera deltagare, kan vi analysera ifall **intuitionerna** om ordens svårighetsgrad stämmer mellan andraspråkstalare och lärare / forskare. Du behöver alltså inte fundera mycket på varför du ser ett ord som lättare eller svårare än ett annat utan använd din intuition framför allt. Men om det är något speciellt som du tycker spelar in i din bedömning så får du gärna kommentera det i feedbackformuläret. Du kan lämna återkoppling via formuläret flera gånger. Det är anonymt.	After collecting votes (rankings) from several participants, we can analyze whether **intuitions** about the difficulty of the words coincide between second language speakers and teachers/researchers. You need not think a lot why you see a word as easier or more difficult compared to another, instead please primarily use your intuition. If you feel there is something that influences your judgments, feel free to comment in the feedback form. You can leave comments several times. It is anonymous.

The projects were open for a month, during which we successfully reached the desired *number of votes* for each of the eight projects.

5.2 Demographic information

A total of 43 participants were recruited through the open call, of those 23 were non-experts ('L2 learners') and 20 experts. Tables 3a and 3b (and the graphs in Appendix 1 for better visualization) present the detailed demographic statistics. One learner left all fields blank, and therefore the total counts in the 'L2 learner' column add up to 22 for all rows.

We can see that women were far better represented in both groups than men, as were university-level or higher educated participants. Among learners, Finnish, Dutch and English were the most represented first languages, but even other languages, such as French, German, Polish, Russian and Ukrainian occurred. For L2 experts, we have

a majority with Swedish as their first language, followed by Finnish, and a few other languages, including Bosnian, Dutch, English, German. The participants reported several countries of residence, including Belgium, Finland, France, Germany, Sweden and the UK.

The presence of Finland as a country of residence and Finnish as a mother tongue is not surprising: Swedish is an official language in Finland, is an obligatory school subject for all pupils (either as an L1 or L2), and is required in some occupations. This explains the number of both L2 learners and L2 experts with Finnish as their first language, and the number of residents in Finland. We were positively surprised to see representatives from other countries than Sweden and Finland, and L2 experts who have mother tongues other than Swedish or Finnish.

As we intended, L2 learners below B1 did not participate, but those at advanced levels were well represented (15 participants out of 23). The levels are assessed by the learners themselves based on their experience and our short explanations. We thus need to keep in mind that the votes provided by the non-experts in this study come from advanced language users, and the results might, potentially, differ if we had a majority of non-experts at B1 level. Language experts are predominantly native speakers of Swedish, but also include seven participants indicating that their level is C1 or C2.

The majority of the L2 experts indicated that they are in one way or another involved with teaching Swedish proficiency courses as well, and thus can be assumed to understand what makes vocabulary relevant or difficult for learners. More than half of them teach Swedish proficiency courses for adults, and one third of them to children at secondary school level.

Table 3a: *Demographic information about the participants*

	L2 learners (non-experts)	L2 professionals (experts)
Total	23	20
Gender		
male	9	3
female	12	17
other	1	-
Age		
... -20	1	-
21-30	8	5
31-40	5	6
41-50	6	2
51-60	1	3
60- ...	1	4
Mother tongues (L1)		
Bosnian/Swedish	-	1
Dutch	4	1
Dutch/Flemish	1	-
English	4	1
Finnish	7	4
French	1	-
German	1	1
Polish	1	-
Russian	2	-
Swedish	-	12
Ukrainian/Russian	1	-
Country of residence		
Belgium	4	2
Finland	8	6
France	-	1
Germany	1	1
Sweden	8	10
UK	1	-

Table 3b: *Demographic information about the participants*

	L2 learners (non-experts)	L2 professionals (experts)
Knowledge of Swedish	(self-estimation)	
B1	4	-
B2	3	-
C1	13	4
C2	2	3
Native speaker	-	13
Education (highest level)		
college/upper-secondary school	2	-
university	11	11
PhD	8	9
higher vocational education	1	-
Teaching experience		
1-5 years	-	7
6-10 years	-	5
15-30 years	-	4
31+ years	-	2
None	-	2
Teaching level		
secondary school	-	4
upper-secondary school	-	2
college	-	1
SFI*	-	2
adult education	-	3
university	-	5
none	-	3

*SFI - Swedish for Immigrants (adult education)

6 Results and analysis

6.1 Control items

Four items among the verbs and adverbs were duplicated as control items (one core item and three periphery items), to see whether the crowdsourcers annotated items consistently. If they did, then these items should appear next to each other in the linear rankings.

For adverbs, the two occurrences of the core item *således* ('consequently') appear at ranks 55 and 56 for non-experts and 51 and 52

for experts, which shows that they were ranked consistently. The other adverbial control word *sammanfattningsvis* ('summing up') appears at ranks 44 and 46 for non-experts and ranks 34 and 35 for experts, again showing consistent rankings.

The verb *underskatta* ('underestimate') appears at ranks 32 and 33 for non-experts and ranks 39 and 40 for experts, once again showing good ranking consistency in both groups. However, the second control verb, *förebygga* ('prevent'), appears at ranks 35 and 42 for non-experts and at ranks 30 and 35 for experts. This verb shows more variation than the other items, especially for non-experts, with a difference of seven ranks (five for experts) on a scale of 60, amounting to ≈10% (≈8%). This could be due to the fact that it appeared at two levels in the coursebooks, and not only at the level it was picked for but also the level *before* it (B2). Another reason could be the co-occurrence with items against which the two occurrences of *förebygga* (förebygga_1 and förebygga_2) have been compared in the experiment. Both alternative explanations should be tested further, as should items' distributions in the coursebooks (see Appendix 2) to see whether we can find a way to prevent 'förebygga'-cases in the future applications of this method.

To conclude, the use of control items has shown that crowdsourcers generally vote very systematically. Even when there seems to be a difference in perception of items' difficulty, the difference in the resulting scalar ranking is relatively modest. We consider, therefore, the resulting ranking (and the method to generate this type of ranking) reliable for our purposes.

6.2 Linear scale

Here we present the results obtained by aggregating the votes for each item into linear scales. To calculate the linear ranking, each time an item was marked as the easiest within a mini-task it received a score of 1, and if marked as the most difficult one it received a score of 3 (see Figure 4 for an example of a mini-task). The unmarked two items received a score of 2. After all votes were collected, the average scores per item were calculated and used for linear ranking (column

'Linear score' in Table 4). As a result, we can explore the positioning of unknown items relative to core items. Table 4 illustrates an excerpt of the resulting linear scale for the unknown word *likaledes* ('likewise, also') and its four closest core neighbors (periphery and unknown neighbors omitted). This example clearly demonstrates that the most probable level that we can expect *likaledes* to appear at and be understood at is C1, both according to teacher votes and to learner votes.

Table 4: *An excerpt of a linear ranking of the unknown item* likaledes *('likewise, also') with a window of four closest core items around it*

Lemma	Linear score	CEFR	Coreness	Rank
Learners				
sammanfattningsvis (' to sum up')	2.35	C1	core	44
jämförelsevis ('comparatively')	2.38	C1	core	45
likaledes ('likewise, also')	2.45	_	**unknown**	48
därigenom ('in that way')	2.50	C1	core	50
bevisligen ('demonstrably')	2.54	C1	core	51
Teachers				
ytterst ('farthest out')	2.26	B2	core	47
därigenom ('in that way')	2.28	C1	core	48
följaktligen ('consequently, accordingly')	2.42	C1	core	50
likaledes ('likewise, also')	2.73	_	**unknown**	55
bevisligen ('demonstrably')	2.74	C1	core	56

We then calculate the correlation between the expert and non-expert rankings by using the Pearson correlation coefficient. Overall, the rankings are quite correlated, ranging from 0.77 (Pearson correlation coefficient) to 0.94 overall. Table 5 gives an overview of the correlation coefficients by part-of-speech, as well as a more detailed overview by core, peripheral and unknown words. It shows that experts and non-experts agreed most on verbs (0.94) and least on nouns (0.77). Appendix 4 gives further details of the correlations by level.

Table 5: *Pearson correlation coefficients between experts and non-experts by part-of-speech and core, peripheral and unknown*

	Overall	Core	Peripheral	Unknown
Nouns	0.77	0.84	0.60	0.52
Verbs	0.94	0.95	0.90	0.78
Adjectives	0.92	0.92	0.88	0.84
Adverbs	0.91	0.90	0.90	0.92

Note. The grey background indicates where there is a reasonably high correlation ≥0.84.

We see a pattern in the correlations of the core, peripheral and unknown vocabulary: participants agreed most on core vocabulary for three of the parts-of-speech (verbs, adjectives, adverbs), a bit less on peripheral vocabulary, and least on unknown vocabulary. Adverbs are the only category where this trend seems reversed, with most agreement on the unknown vocabulary. However, this category also shows the least variation overall, with coefficients around 0.90. This could be related to the fact that we also had some trouble picking items for this part-of-speech since there were simply fewer adverbs to choose from in the data. It could also be an idiosyncratic result and we would need more data to confirm the reason for this difference between adverbs and the other groups, or if more data would result in the same trend as for the other parts-of-speech.

6.3 Clustering

While the linear scale gives a good overview of the relative difficulty of items, it remains a one-dimensional representation. As we want to explore how to assign levels to words of unknown level based on anchor words of known level, such a representation might not suffice. As illustrated in Figure 5, increasing the dimensionality can uncover relations that are not visible at lower dimensions. In the figure the green stars represent words of unknown level, while the blue dots and brown squares represent anchor words of known level. From looking at the one-dimensional (i.e., linear scale) visualization, it is rather difficult to attach a level to those words of unknown level based on the proximity of anchor words. Even the two-dimensional representation does not show a clear trend. However, the three-dimensional representation

shows clearly that one of the words of unknown level is very close to the blue dot anchor words, while the other one is close to the brown square anchor words. While we acknowledge that vectors based on pairwise comparisons from the linear scales will show high degrees of intercorrelation, we explore this technique in an attempt to untangle the low-dimensional representation and to visualize the results.

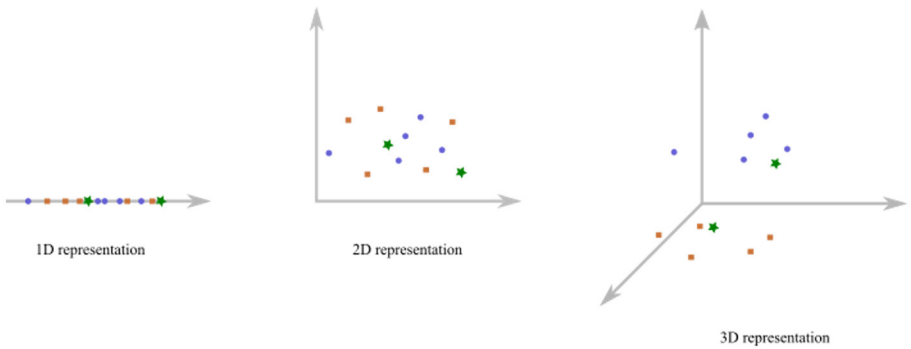


Figure 5: Clustering results in 1-, 2- and 3-dimensional representations.

Following this intuition, we use the distances from the linear scale to represent our words in high dimensional space. To do so, we calculate the distance between each pair of words, so that each word has 59 distances, one for each of the other words, plus a distance of zero to itself. These distances are then interpreted as coordinates in a 60-dimensional space.

By using this high dimensional representation, we want to see whether we can assign levels to words of unknown levels. As a first step, we perform a clustering analysis on the core and peripheral data in order to see whether clustering might be a viable choice for assessing the difficulty of unknown words. As we assume the levels of core and peripheral vocabulary to be known and valid, we can use these labels to see to what extent a clustering algorithm generates the expected results, and for the clustering in this step we use KMeans (McQueen, 1967).

Tables 6a–6d show the overall confusion matrices by group (experts and non-experts) and part-of-speech (nouns, verbs, adjectives and adverbs), excluding words of unknown level, since their true level is not known. Numbers in bold indicate cases where the clustering algorithm assigned most of the elements in a class to the correct class.

Table 6a: *Adverbs, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**6**	3	0	0	0		**8**	3	0	0	0
A2	2	**3**	1	1	0		0	1	0	2	2
B1	2	2	**8**	3	0		2	4	**5**	2	0
B2	0	1	1	3	5		0	1	3	**4**	2
C1	0	1	0	3	**5**		0	1	2	2	**6**

Table 6b: *Adjectives, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**5**	1	0	0	0		**6**	1	0	0	0
A2	5	5	0	4	0		4	4	0	3	1
B1	0	2	**8**	3	1		0	3	**5**	2	0
B2	0	2	2	3	5		0	0	1	1	4
C1	0	0	0	0	**4**		0	2	4	4	**5**

Table 6c: *Verbs, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**5**	2	0	0	0		**4**	2	0	0	0
A2	3	**4**	1	1	0		3	**4**	1	0	0
B1	0	0	**4**	3	2		1	3	**4**	3	0
B2	1	3	3	3	2		0	1	1	**4**	3
C1	1	1	2	3	**6**		2	0	4	3	**7**

Table 6d: *Nouns, L2 speakers (left) and experts (right)*

Predicted→ Gold	A1	A2	B1	B2	C1		A1	A2	B1	B2	C1
A1	**5**	2	0	0	0		**5**	2	0	0	0
A2	4	4	2	2	2		4	4	2	2	2
B1	1	2	**4**	2	1		1	2	**4**	2	1
B2	0	1	3	3	2		0	1	3	3	2
C1	0	1	1	3	**5**		0	1	1	3	**5**

As can be gathered from the confusion matrices, the clustering tends to perform well on the extremes of the scale (levels A1 and C1) but also around B1, with most occurrences of these items correctly clustered.

In the second step, we want to assign levels to words with an unknown level. In order to do so, we use another clustering algorithm, the k-nearest-neighbors (*k-NN*; Fix and Hodges, 1989) algorithm to see which anchor words are closest to the unknown words. We then predict the level of the unknown word as the majority level of its five closest neighbors. Tables 7a–7d present the results of this analysis.

Table 7a: *Clustering results for unknown adverbs*[14]

Adverbs	Cf. Kelly level	Predicted levels	
		L2 speakers	L2 experts
enbart ('solely, only')	A1	B1	B1
således ('consequently')	A1	C1	C1
förvisso ('certainly')	A2	C1	C1
såklart ('absolutely')	A2	B1	B2
mestadels ('mostly')	B1	C1	B1
sedermera ('afterwards')	B1	C1	C1
tillika ('moreover')	B2	C1	C1
fortsättningsvis ('henceforth')	B2	B2	C1
likaledes ('likewise')	C1	C1	C1
massvis ('lots of')	C1	B2	C1

Table 7b: *Clustering results for unknown adjectives*

Adjectives	Cf. Kelly level	Predicted levels	
		L2 speakers	L2 experts
vag ('vague')	A1	B2	B2
uppenbar ('obvious')	A1	B2	B1
facklig ('trade union')	A2	B2	B2
rättslig ('legal')	A2	B2	B2
skeptisk ('skeptical')	B1	B1	B1
nyliberal ('neo-liberal')	B1	B1	B1
medborgerlig ('civil')	B2	B2	B1
byråkratisk ('bureaucratic')	B2	B1	B1
välmående ('healthy')	C1	B1	B1
ovannämnd ('above-mentioned')	C1	B2	B2

14 The grey background in Tables 7a–7d marks agreement between the two groups or between at least one group and the CEFR-level predicted in the Swedish Kelly-list.

Table 7c: *Clustering results for unknown verbs*

Verbs	Cf. Kelly level	Predicted levels	
		L2 speakers	L2 experts
kapa (‘hijack’)	A1	B1	B1
ämna (‘intend to’)	A1	B1	B1
förespråka (‘advocate’)	A2	B1	B1
tillhandahålla (‘supply’)	A2	B1	B1
erinra (‘remind’)	B1	B1	B1
påvisa (‘prove’)	B1	B1	B1
avlägsna (‘remove’)	B2	B1	B1
genomsyra (‘permeate’)	B2	B1	B1
understödja (‘support’)	C1	B1	B1
beskåda (‘regard’)	C1	B1	B1

Table 7d: *Clustering results for unknown nouns*

Nouns	Cf. Kelly level	Predicted level	
		L2 speakers	Experts
medlemsstat (‘member state’)	A1	B1	B1
pelare (‘pillar’)	A1	B1	B1
upphovsrätt (‘copyright’)	A2	B2	B1
penningpolitik (‘monetary policy’)	A2	B2	B1
fildelare (‘file sharer’)	B1	B2	B1
antagande (‘assumption’)	B1	B2	B1
mervärde (‘surplus’)	B2	B2	B1
sökmotor (‘search engine’)	B2	B1	A2
vapenvila (‘cease-fire’)	C1	B1	B1
dotterbolag (‘subsidiary company’)	C1	B1	B1

The predicted levels largely overlap between the two voter groups (Tables 7a–7d), with differences within one level, except for the adverb *mestadels* (‘mostly’), which is predicted to be both B1 and C1. All unknown words are further predicted as at least B1 (except for *sökmotor* ‘search engine’ which is predicted as A2 by the experts), which

confirms our findings for linear scales, where unknown words end up in the middle and the end of the scale.

For adverbs, the non-expert clustering perfectly aligns with their ranking: A2 words (according to the clustering) have ranks lower than B1 words (according to the clustering), which in turn have ranks lower than B2 words (according to the clustering). For the expert clustering, this is almost true: *såklart* ('absolutely') is at rank 16 in the ranking while *enbart* ('solely, only') is found at rank 27. However, *enbart* is predicted as B1 and *såklart* B2.

For adjectives, the non-expert clustering also perfectly aligns with their ranking, as all words predicted as B1 come before words predicted as B2 in the ranking. This also holds true for the expert clustering.

For verbs, clustering is perfectly identical for both groups, and all words are predicted to be of level B1 in both groups.

For nouns, the non-expert clustering again perfectly aligns with their ranking, as B1 words (according to the clustering) are all found before B2 words (according to the clustering) in the ranking. This is not the case for the expert clustering, as all words are predicted as B1 except for *sökmotor* ('search engine') which is predicted as A2, yet is found at rank 44 according to the linear scale, while *pelare* ('pillar') – predicted as B1 – is ranked at 27.

Finally, if we compare the predicted levels to the levels assigned to these words in Kelly, we can see a rather large discrepancy in most of the cases, especially concerning the lowest levels, which could indicate that frequency alone may not be sufficient as a predictor of CEFR levels.

6.4 Time investment

Figure 6 shows the average time needed per task for non-experts and experts in seconds.[15] The box shows the first and third quartiles (lower and upper lines of the box), the orange line dividing the box indicates the median, the whiskers show the minimum and maximum values (outliers not counted), and the dots show outlier values. We hypothesize that

15 Outliers of more than 100 seconds are excluded in order to improve the readability of the figure. Extreme outliers go up to 3,500 seconds.

outliers indicate moments when participants were interrupted during the experiment (e.g. went to get coffee or answered the phone) without closing the program. As we see, despite some obvious outliers, the average projected time of 30 seconds per task is well met.

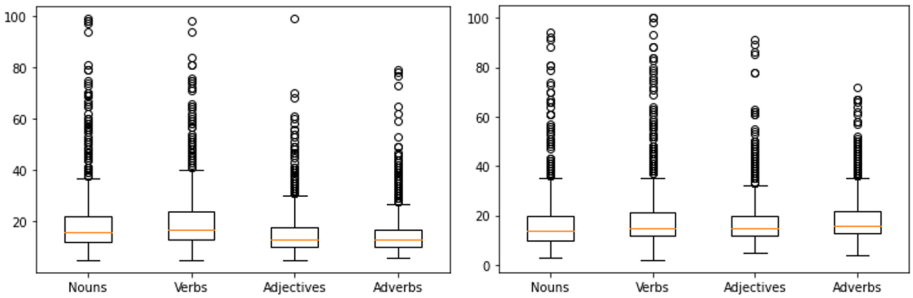


Figure 6: Boxplots of time taken per task in seconds, for learners (left) and experts (right).

6.5 Qualitative analysis

The easiest 10 items (Table 8) in all four parts-of-speech were mainly core items, according to both learners and experts. For nouns and verbs, 90–100% of the easiest ten items were core items, but they ranged from A1–B1 in the case of the learners' ranking of nouns whereas the experts picked A1–A2 core items as the easiest ten nouns, and learners and experts picked A1–A2 core items as 90% of the ten easiest verbs (see Appendix 3). For adjectives and adverbs, 80% of the easiest ten items were picked from the core items by learners and 60–70% by the experts. All of the core items picked among the easiest ten adjectives and adverbs were from A1–A2 in the coursebooks.

Even among the 20 easiest items, it was mainly core items that were picked by both learners and experts in all four parts-of-speech. However, these items range from A1–C1 and there are more peripheral items compared to the easiest 10. Among the easiest 20 adverbs experts even include one of the unknown items (*såklart* 'absolutely'), an item picked from A2 in the Kelly list and hence from the second frequency band.

Looking instead at how the core items were ranked, we see that they appear primarily among the 20 easiest items in all four parts-of-speech

Table 8: *Core items among the items ranked as the 10 easiest items by learners and experts*

	Learners	Experts
Nouns	A1: *pappa, kaffe, klocka, dag, frukost* (daddy, coffee, clock, day, breakfast)	A1: *pappa, kaffe, klocka, frukost, dag*
	A2: *kilo, doktor, mage, kött, flygplan* (kilo, doctor, belly, meat, airplane)	A2: *doktor, flygplan, mage, kött, kilo*
Verbs	A1: *äta, gå, titta, stå, heta* (eat, go, look, stand, be named)	A1: *gå, titta, äta, heta, stå*
	A2: *cykla, kontakta, trycka, meddela* (bike, contact, push, inform)	A2: *cykla, kontakta, meddela, trycka*
Adjectives	A1: *liten, glad, stor, bra, halv* (small, glad, big, good, half)	A1: *liten, bra, stor, glad, halv*
	A2: *grön, lycklig, vuxen* (green, happy, grown-up)	*A2: grön*
Adverbs	A1: *också, hemma, där, ibland* (also, at home, there, sometimes)	A1: *sedan, också, ibland, hemma, där* (since, ...)
	A2: *ej, dit, inne, var* (not, there, inside, where)	*A2: dit, ej*

Note. The grey background indicates that learners and experts agreed on these items as being among the 10 easiest core items.

and for both learners and experts (Table 9). Conversely, we find very few core items among the most difficult items (20–30%) in all four parts-of-speech and both for learners and experts. Furthermore, the core items which are among the most difficult are from B1–C1 and never A1–A2, whereas the easiest core items range from A1–B2 for learners, A1–C1 for experts.

Table 9: *Dispersion of the core items in the ranking experiment according to the number of core items among items number 1–20, 21–40 and 41–60, including the levels predicted for those levels based on CEFR tools*

	1–20	21–40	41–60
Noun core items (learners)	14 (A1–B2)	7 (B1–C1)	4 (B2–C1)
Noun core items (experts)	11 (A1–B1)	9 (B1–C1)	5 (B1–C1)
Verb core items (learners)	12 (A1–B1)	6 (B1–C1)	6 (B1–C1)
Verb core items (experts)	12 (A1–B1)	7 (B1–C1)	5 (B1–C1)
Adjective core items (learners)	13 (A1–B2)	8 (B1–C1)	4 (B2–C1)
Adjective core items (experts)	13 (A1–C1)	8 (A2–C1)	4 (B2–C1)
Adverb core items (learners)	13 (A1–B2)	5 (B1–B2)	6 (B2–C1)
Adverb core items (experts)	12 (A1–B2)	6 (A2–C1)	6 (B1–C1)

The items which were ranked as the ten most difficult contain very few core items, and the core items which appear here are usually C1 (the learners chose eight of these, the experts five, see Table 10). These words do not contain any clearly international items, and both learners and experts agree to a large extent, although the former include three more items than the latter. There is also one item that is only included by the experts (*anvisning* 'directions, instructions'), whereas learners include four items which the experts did not include: *bedra, förebygga, värdesätta* and *följaktligen.*

Table 10: *Core items among the items ranked as the 10 most difficult by learners and experts*

	Learners	Experts
Nouns	*utformning* ('design')	*utformning* *anvisning* ('directions')
Verbs	*bedra* ('deceive') *förebygga* ('prevent') *värdesätta* ('value')	-
Adjectives	*övergripande* ('comprehensive') *nedlåtande* ('condescending')	*övergripande* *nedlåtande*
Adverbs	*bevisligen* ('demonstrably') *följaktligen* ('consequently')	*bevisligen*

Note. The grey background indicates that the same words were seen as among the 10 most difficult by learners and experts.

7 Discussion

The way we designed the experiment shows that frequency-based statistical measures and predictions offered by CEFR-tools can, indeed, help stratify vocabulary into (so-called) core and periphery items. The items we have picked as core based on the ranges in Table 1 behaved differently from those that we chose as periphery. It is important to bear in mind that, apart from pure frequency ranges, we also applied the principles of topicality and/or usefulness where many candidate items were available or, conversely, where too few were available. In general, we have seen that the core/anchor items per level have been confirmed as such on a linear scale by voters with different linguistic

backgrounds. These are the items that could be effectively used for further experiments on a method of assigning unseen items to a proficiency level.

We can thus claim that we have developed a strategy to identify words capable of being reliable anchors, namely, using *CEFR tools* by applying various statistical measures. The ranges that we have experimented with, have helped us capture the coreness of certain vocabulary items for each level, as confirmed by both expert and non-expert ratings. However, we have also realized that cognates and internationally recognizable items give a false sense of simplicity, can easily mislead and should not be used for experiments of this type (cf. Lindström Tiedemann et al., 2022). The item selection and crowdsourcing experiment have enriched us with a list of items that we recommend using in the future for assigning unknown vocabulary to the target levels. The core items used in our experiment are listed in Appendix 5, both in Swedish and translated into English. It would be most interesting to see whether the same items represent coreness in other languages for the corresponding levels.

Empirical analysis of unknown items in relation to our anchor words has shed new light on **how frequencies, usefulness/topicality, coreness and language learning may be related** (see 6.3). Frequency has been claimed to be a consequence of being core, and not vice versa (Stein, 2017), although frequency is often taken as a proxy of coreness. The problem with this type of simplification is that not all core items are frequent (e.g. the frequency of *Tuesday* compared to *Friday*; *brown* and *white*) thus frequency may lead to contentious results.

To demonstrate the last claim, the unknown items that we selected from the Swedish Kelly list have been related to CEFR levels based on frequency bands. Our results, however, show that these levels very rarely coincide with the levels predicted through positions on a linear scale, even though the learners and experts in this experiment were in high agreement about their relative difficulty. We take this to mean that although frequency is important in learning new vocabulary, CEFR levels (and 'coreness' of the items) cannot simply be related to frequency bands. This finding is highly relevant to second language acquisition, since vocabulary assessment tends to rely on testing vocabulary in

relation to frequency bands. Most importantly, frequencies in general corpora may be irrelevant for L2 contexts, whereas L2-relevant corpora are likely to contain more reliable frequency indications.

Unknown items coming from the Kelly list were found only once among the easiest 20 items and only among the expert judgments. Unknown items were, instead, very highly represented among those ranked as the most difficult 20, or even the most difficult 10 (see Appendix 3), which also serves as an indication that at least levels below B1 were poorly predicted by the Kelly list, based on frequencies from general corpora.

The results of the study suggest that the same setup, but limited to two–three anchor items per level (e.g. 10 anchor items in total) and one–two unknown words (e.g. 12 items for a "project" in total), could help resolve the question of an unknown word and its placement on a CEFR scale for learning and assessment purposes. Since we already know the relations (easier–more difficult) between the anchor words, the number of micro-tasks would be dramatically reduced by only testing these relations for the unknown words. A suggestion for item placement could thus be achieved in a very limited time. Moreover, our findings indicate that we can let any user with sufficient knowledge of Swedish vote, without controlling for their background (i.e. native speakers, trained experts or non-native speakers). Testing this approach as a quick method for resolving level assignment for previously unseen vocabulary items is planned in future work. A number of questions need to be addressed in this context, for example:

- How many micro-tasks are optimal? How much time will it take to place one new item?
- How many votes do we need? (cf. Alfter et al., 2021)
- How stable is the ranking? Does decreasing the number of votes affect placement reliability?

Carter (1982, p. 46) summarizes his theoretical analysis of the nature of core vocabulary by saying that "…no single criterion can be taken to produce definitely a core vocabulary item. Rather some combination can help define the strength of the 'coreness' but it will also, to some extent, be affected by the purposes for which a definition of a core lexis

is sought." Based on Carter (1982) we may stipulate that the testing paradigm for core vocabulary (Section 2.1) will not be applicable in its entirety to all types of core vocabularies. For example, a cognitive basis will be less critical for pedagogical uses of core vocabulary; for general lexicography, definitional power and semantic network placement would figure most prominently; while for diachronic lexicostatistics, semantic neutrality and frequency will be the most important properties.

We would like to round off this discussion by quoting Borin (2012, p. 63): "It is perhaps not surprising that there should be so little overlap among different kinds of 'core vocabularies', since they aim at capturing different aspects of 'coreness'". We thus do not propose that the suggested anchor words that we claim to be core for the second language of learners will be universal in all settings.

8 Conclusions

Returning to our **hypotheses**, we can now confirm the following:

1. There is a common core vocabulary at *A1–B1 levels*; there is less systematicity at *B2–C1* levels.

Analyzing the correlation between learners and experts we could see that the correlation was generally higher in the core items for A1–B1 than for B2–C1. In fact, we even noticed that the correlation was sometimes higher for the whole A1–B1 group than within a particular level (cf. Appendix 4). We believe this to be an indication that there is a core which relates quite well to the A1–B1 levels in coursebooks, but that the precise order in which these items occur in the coursebooks and how they would be ranked by learners or experts might not coincide as well for each level. This is also related to the fact that we assume that proficiency is a continuum rather than something that can be clearly divided into discrete levels.

2. Some systematicity can be observed in the behavior of the *core* items, but less so in *peripheral* items.

Core items very clearly appear mainly among the items which are ranked among the easiest. This is likely to be because both peripheral

items and unknown items in this experiment do not belong to the core vocabulary, and that different learners vary in their knowledge of these words and hence variations in the ranking of these words.

3. Through crowdsourced comparative judgments, *unknown* vocabulary items will demonstrate a perceived difficulty (expressed in numerical scores) equal or comparable to the perceived difficulty of *anchor* items of a particular level.

The crowdsourcing experiment has shown that sensible levels can be assigned to words of unknown level based on comparative judgments of unknown words against anchor words, as illustrated in Table 4. This seems to confirm that we can use the perceived difficulty of unknown vocabulary items for the assignment of levels based on the closeness of the difficulty of nearby words. However, while this methodology allows for the identification of levels for words included in the experiment, it is not possible to easily calculate actual levels for new words, as the linear ranking and clustering are performed on the distance of each word to every other word (and thus voting to calculate distances between all anchor words and the unknown item is a prerequisite of this approach). The experiments presented in this work also show that **learners** are perfectly aligned with regard to their assessments of the difficulty of words with an unknown level and the subsequent linear scale projection and clustering.

4. *Non-experts* will perform on par with *experts* in a comparative judgment setting.

This question yields a convincing "yes" in response. The study has confirmed previous findings that L2 learners can be used in the same way as experts, given a carefully designed comparative judgment setting.

We found that our method of selecting core items worked well in establishing anchors. To ensure their reliability it is particularly important to make sure that their meaning cannot be derived from international words which appear in many European languages or cognates in English, such as the Swedish *doktor* ('doctor').

While we do not yet have an inexpensive cheap method for ranking items in relation to explicit CEFR levels, there is a good chance that

with the knowledge we have gained one will be soon available. We have seen that clustering can indeed be used to derive sensible levels for words of unknown level; in the future it would be interesting to calculate, for example, the distances between the unknown words in a cluster and the cluster center to see whether this gives any hints as to the coreness of these items.

It is especially encouraging that we can use learners for this type of linguistic annotation, and in fact our results indicate that learners *might* be more attuned to the relative difficulty of words than experts are, since their rankings more often coincide with the coursebook levels. This may be due to the fact that we operationalized *difficulty* in terms of the CEFR, and the CEFR levels specifically target learners. It would thus be logical for learners to be more sensitive to these levels than native speakers and language professionals, a finding also hinted at in Alfter et al. (2022).

Acknowledgments

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1, and by *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We also wish to thank the anonymous reviewers for their valuable comments on a previous version.

References

Alfter, D. (2021). Exploring natural language processing for single-word and multiword lexical complexity from a second language learner perspective. PhD thesis. University of Gothenburg.

Alfter, D., Bizzoni, Y., Agebjörn, A., Volodina, E., & Pilán, I. (2016). From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 1–7).

Alfter, D., Cardon., R., & François, T. (2022). A Dictionary-based Study of Word Sense Difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, (pp. 17–24).

Alfter, D., & Volodina, E. (2018). Towards single word lexical complexity prediction. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 79–88).

Alfter, D., Lindström Tiedemann, T., & Volodina, E. (2021). Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts. *Northern European Journal of Language Technology* (Vol. 1). doi: 10.3384/nejlt.2000-1533.2021.3128

Alfter, D., Borin, L., Pilán, I., Lindström Tiedemann, T., & Volodina, E. (2019). Lärka: from language learning platform to infrastructure for research on language learning. In *Selected papers from the CLARIN Annual Conference 2018, 8–10 October 2018*, *Pisa* (pp. 1–14). Linköping University Electronic Press.

Bell, H. (2013). Core Vocabulary. In C. Chapelle (Ed.) *The encyclopedia of applied linguistics*. Malden, MA: Wiley-Blackwell.

Borin, L. (2012). Core vocabulary: A useful but mystical concept in some kinds of linguistics. In *Shall We Play the Festschrift Game?* (pp. 53–65). Springer, Berlin, Heidelberg.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In: *Proceedings of LREC 2012* (pp. 474–478). Istanbul: ELRA.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1–22.

Capel, A. (2015). The English Vocabulary Profile. *English profile in practice, 5*, 9–27.

Carter, R. (1982). A note on core vocabulary. In Stubbs M. and Carter R. (Eds.), *Nottingham Linguistic Circular, 11*(2), 39–51.

Carter, R. (1987). Is there a core vocabulary? Some implications for language teaching. *Applied linguistics, 8*(2), 178–193.

Council of Europe [COE]. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe [COE]. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (19. 10. 2021)

Crosbie, S., Pine, C., Holm, A., & Dodd, B. (2006). Treating Jarrod: A core vocabulary approach. *Advances in Speech Language Pathology, 8*(3), 316–321.

De Clercq, Orphée, Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering, 20*(3), 293–325.

Dixon, Robert MW. (1971). A method of semantic description. *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, 436–471.

Familiar, L. (2021). *A frequency dictionary of contemporary Arabic fiction: core vocabulary for learners and material developers*. Routledge.

Fix, E., & Lawson Hodges, J. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/ Revue Internationale de Statistique, 57*(3), 238–247.

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

François, T., Volodina, E., Pilán, I., & Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 213–219).

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of THING. *Journal of English for Academic Purposes, 52*, 100999.

Hawkins, J. A., & Filipović, L. (2012). Criterial Features in L2 English. In *English Profile Studies 1*. Cambridge: Cambridge University Press.

Holmer, D., & Rennes, E. (2022). NyLLex: A Novel Resource of Swedish Words Annotated with Reading Proficiency Level. In *Proceedings of the 13th Language Resources and Evaluation Conference* (*LREC), 20 – 25 June 2022, Marseille* (pp. 1326–1331). Retrieved from http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.141.pdf

Hulstijn, J. H. (2019). An individual differences framework for comparing nonnative with native speakers: Perspectives from BLC theory. *Language Learning, 69*, 157–183.

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Bondi Johannessen, J., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., & Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation, 48*(1), 121–163.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018). Collocations dictionary of modern Slovene. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 989–997). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Retrieved from https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/118/211/2939

Kullenberg, C., & Kasperowski, D. (2016). What is citizen science? – a scientometric meta-analysis. PloS one, *11*(1), e0147152.

Lau, J. H., Clark, A., & Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the annual meeting of the cognitive science society, 36*(36).

Lehmann, H. (1991). Towards a core vocabulary for a natural language system. In *Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Retrieved from https://aclanthology.org/E91-1053.pdf

Lindström Tiedemann, T., Alfter, D., & Volodina, E. (2022). CEFR-nivåer och svenska flerordsuttryck [= CEFR levels and Swedish Multiword Expressions]. In S. Björklund, B. Haagensen, M. Nordman & A. Westerlund (Eds.), *Svenskan i Finland 19,* Vaasa:: Svensk-Österbottniska Samfundet r.f. (pp. 218–233). Retrieved from https://www.doria.fi/handle/10024/185549

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge.

Louviere, J. J., N. Flynn, T., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*.

Márquez, M. F. (2007). Renewal of core English vocabulary: A study based on the BNC. *English Studies, 88*(6), 699–723.

Mühlenbock, K., H., & Johansson Kokkinakis, S. (2012). SweVoc-a Swedish vocabulary resource for CALL. In *Proceedings of the SLTC 2012 workshop on NLP for CALL, 25th October 2012, Lund* (pp. 28–34). Linköping University Electronic Press.

Ortega, L. (2012). Interlanguage complexity. In Kortmann, B. & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). De Gruyter.

Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced Adaptive Comparative Judgment: A Community-Based Solution for Proficiency Rating. *Language Learning*.

Scott, William A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325. doi: 10.1086/266577

Stein, G. (2017). Some thoughts on the issue of core vocabularies: A response to Vaclav Brezina and Dana Gablasova: Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 38*(5), 759–763.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell publishers.

Swadesh, M. (2017). *The origin and diversification of language*. Routledge.

Volodina, E., & Johansson Kokkinakis, S. (2012a). Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1040–1046).

Volodina, E., & Johansson Kokkinakis, S. (2012b). Swedish Kelly: Technical Report. GU-ISS-2012-01. The Swedish Language Bank, Gothenburg University.

Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016). SweLLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 76–84).

Volodina, E., Pilán, I., Rødven Eide, S., & Heidarsson, H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning* (pp. 128–144).

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., Carin, L. (2018). Joint Embedding of Words and Labels for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 2321–2331).

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman.

Ocene posameznih leksikalnih elementov, pridobljene z množičenjem: perspektiva osrednjega besedišča

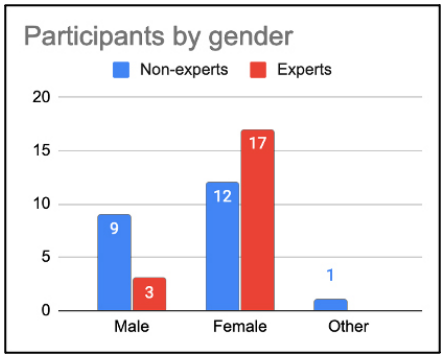
V raziskavi preučujemo teoretična in praktična vprašanja, povezana z razlikovanjem med osrednjim in obrobnim besediščem na različnih ravneh jezikovnega znanja z uporabo statističnih pristopov v kombinaciji z množičenjem. Obenem ugotavljamo, ali je mogoče razvrstitve oseb, ki se učijo drugega jezika, uporabiti za določanje ravni nepoznanega besedišča. Raziskava je izvedena na enobesednih enotah v švedščini.

Preučujemo štiri hipoteze: (1) za vsako raven znanja obstaja osrednje besedišče, vendar to velja le do ravni B2 po CEFR (višja srednja raven); (2) osrednje besedišče kaže večjo sistematičnost v rabi, medtem ko se robni elementi obnašajo bolj idiosinkratično; (3) glede na to, da imamo za vsako raven na voljo ključne elemente (t. i. sidrne elemente), lahko vsako novo nepoznano

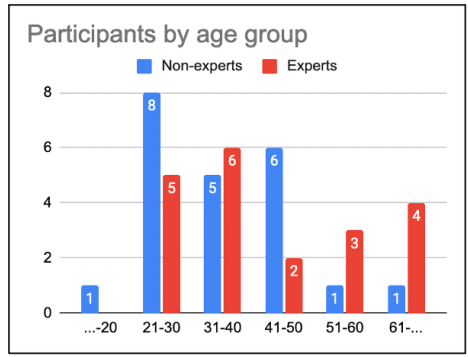
besedo postavimo ob bok omenjenim ključnim elementom z vrsto primerjalnih ocenjevalnih nalog in tako določimo "ciljno" raven za prej nepoznano besedo; in (4) osebe s pomanjkljivim znanjem se bodo v primerjalnem ocenjevanju odrezale enakovredno osebam z dobrim znanjem. Hipoteze smo v veliki meri potrdili: V povezavi z (1) in (2) naši rezultati kažejo, da obstaja določena sistematičnost pri jedrnem besedišču za začetne in srednje ravni (A1-B1), medtem ko smo pri višjih ravneh (B2-C1) opazili manj sistematičnosti. Pri točki (3) predlagamo, da se kot metoda za dodelitev "ciljne" ravni nepoznanim besedam uporabi množičenje ocen besed z uporabo primerjalne presoje in s pomočjo poznanih sidrnih besed. Glede (4) potrjujemo predhodne ugotovitve, da je mogoče za naloge jezikovnega označevanja v okviru primerjalne presoje učinkovito uporabiti nestrokovnjake, v našem primeru učence jezika.

Ključne besede: osrednje besedišče in učenje jezika, množičenje pri nestrokovnjakih, posamični leksikalni elementi, ravni CEFR, primerjalna presoja

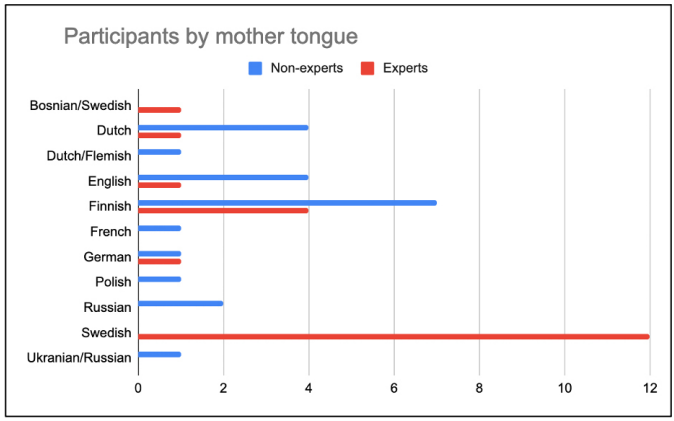
Appendix 1: Demographic information about the participants shown in graphs (1a–h)



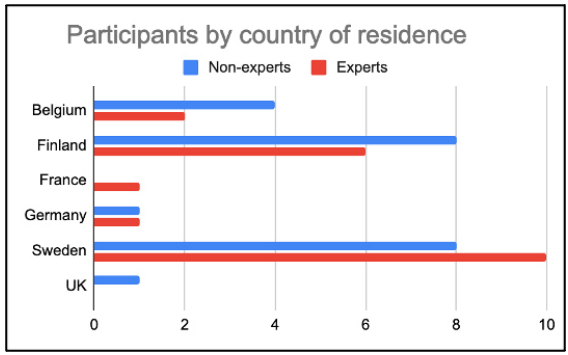
Appendix 1a.



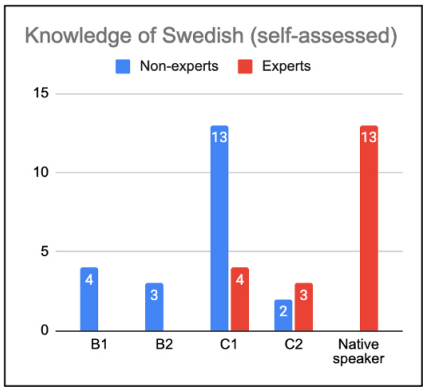
Appendix 1b.



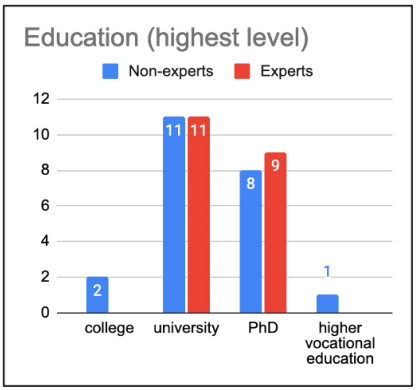
Appendix 1c.



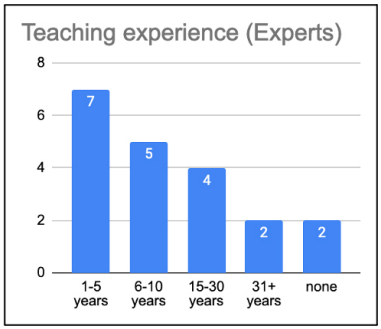
Appendix 1d.



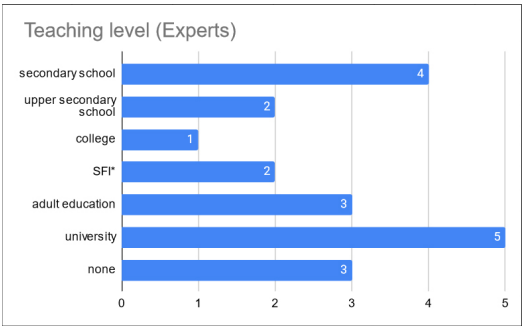
Appendix 1e.



Appendix 1f.



Appendix 1g.



Appendix 1h.

Appendix 2: Control items

	Level	Core / periph.	Agree-ment	Homoge-neity	Product. majority level	Coctaill LM	Docu-ments	Books
således *'consequently'*	B2	core	0.5	-0.5	C1	B2	B2: 2 C1: 2	B2: 1 C1: 2
sammanfattningsvis *'summing up'*	C1	periphery	0.5	-0.25	B1	B2	C1: 2	C1: 1
underskatta *'underestimate'*	C1	periphery	0.75	0.375	C1	C1	C1: 1	C1: 1
förebygga *'prevent'*	C1	periphery	0.5	-0.75	C1	C1	B2: 1 C1: 1	B2: 1 C1: 1

Note. Information about control items.

Appendix 3: Statistics for the included items

	Easiest 10 (1-10)	Easiest 20 (1-20)	Hardest 10 (51-60)	Hardest 20 (41-60)
Nouns (learners)	**90% core (A1–B1)** 10% periphery (A1)	**70% core (A1–B2)** 30% periphery (A1–C1)	10% core (C1) **50% periphery (A2–C1)** 40% unknown	20% core (B2–C1) **50% periphery (A2–C1)** 30% unknown (A1–B2)
Nouns (experts)	**100% core (A1–A2)**	**55% core (A1–B1)** 45% periphery (A1–C1)	20% core (C1) **50% periphery (A2–C1)** 30% unknown (A2–B2)	25% core (B1–C1) **40% periphery (A2–C1)** 35% unknown (A2–C1)
Verbs (learners)	**90% core (A1–A2)** 10% periphery (A1)	**60% core (A1–B1)** 40% periphery (A1–B2)	10% core (B2) 30% periphery (B2–C1) **60% unknown (A1–C1)**	30% core (B1–C1) 30% periphery (B1–C1) **40% unknown (A1–C1)**
Verbs (experts)	**90% core (A1–A2)** 10% periphery (A1)	**60% core (A1–B1)** 40% periphery (A1–B2)	0% core **50% periphery (A2–C1)** **50% unknown (A2–C1)**	25% core (B1–C1) 30% periphery (A2–C1) **45% unknown (A1–C1)**
Adjectives (learners)	**80% core (A1–A2)** 20% periphery (A1–A2)	**65% core (A1–B2)** 35% periphery (A1–B2)	20% core (C1) 30% periphery (B2–C1) **50% unknown (A1–C1)**	20% core (B2–C1) **50% periphery (A2–C1)** 30% unknown (A1–C1)
Adjectives (experts)	**60% core (A1–A2)** 40% periphery (A1–A2)	**65% core (A1–C1)** 35% periphery (A1–B1)	20% core (C1) **40% periphery (B1–C1)** **40% unknown (A1–C1)**	20% core (B2–C1) **50% periphery (A2–C1)** 30% unknown (A1–C1)
Adverbs (learners)	**80% core (A1–A2)** 20% periphery (A1)	**65% core (A1–B2)** 35% periphery (A1–B2)	20% core (C1) 30% periphery (A2–C1) **50% unknown (A1–B2)**	30% core (B2–C1) **35% periphery (A2–C1)** **35% unknown (A2–C1 + AB)**
Adverbs (experts)	**70% core (A1–A2)** 30% periphery (A1)	**60% core (A1–B2)** 35% periphery (A1–A2) 5% unknown (A2)	10% core (C1) 30% periphery (A2–C1) **60% unknown (A2–C1 + AB)**	30% core (B1–C1) 30% periphery (A2–C1) **40% unknown (A2–C1 + AB)**

Note. Core, periphery and unknown items by percentage in the 10 and 20 easiest items and the 10 and 20 most difficult items.

Appendix 4: Correlations

		Core	Peripheral
Nouns	A1	0.88	0.67
	A2	-0.36	0.84
	B1	0.72	0.55
	B2	**0.14**	**0.20**
	C1	0.75	**0.13**
	A1–B1	0.84	0.82
	B2–C1	0.58	**0.16**
Verbs	A1	0.66	0.93
	A2	0.91	0.95
	B1	0.99	0.77
	B2	0.85	0.93
	C1	0.39	0.94
	A1–B1	0.99	0.83
	B2–C1	0.73	0.92
Adjectives	A1	0.55	0.57
	A2	0.52	0.86
	B1	0.86	0.66
	B2	0.94	0.91
	C1	0.89	0.71
	A1–B1	0.93	0.88
	B2–C1	0.95	0.75
Adverbs	A1	-0.57	0.82
	A2	0.67	0.98
	B1	0.68	0.43
	B2	0.63	0.90
	C1	0.74	0.76
	A1–B1	0.81	0.89
	B2–C1	0.83	0.82

Note. Pearson correlation coefficients by part-of-speech and level, core, peripheral and unknown (correlation of learner and expert rankings).

In comparing the correlation scores between levels and between core and periphery we have marked the reasonably high (≥(-)0.72) correlations with a grey background, but a maximum of one per row (i.e. either core or periphery) unless the two values are both >0.9, to make it easier to see which correlations were highest. Correlations marked in bold are particularly low and to be seen as negligible, at ≤(-)0.30.

Appendix 5: Swedish core items with English translations

Level	Swedish core word	English translation
ADVERB		
A1	också	also
A1	där*	there
A1	ibland	sometimes
A1	sedan	then
A1	hemma	at home
A2	var*	where
A2	ej	not
A2	verkligen	really
A2	inne	in; indoors
A2	dit	there
B1	gradvis	gradually
B1	troligen	very likely, probably
B1	alltför	far too, much too
B1	näst	last (but one), second (best)
B1	vanligtvis	usually
B2	ytterst	farthest out
B2	ingenstans	nowhere
B2	möjligen	possibly
B2	således	consequently
B2	säkerligen	certainly
C1	därigenom	in that way
C1	jämförelsevis	comparatively
C1	sammanfattningsvis	to sum up
C1	följaktligen	consequently; accordingly
C1	bevisligen	demonstrably
ADJECTIVE		
A1	liten	small; little
A1	halv*	half
A1	stor	large
A1	glad	happy
A1	bra	well, alright
A2	duktig	capable
A2	ledsen	sad
A2	lycklig	happy

A2	vuxen	adult, grown-up
A2	grön*	green
B1	hemsk	ghastly, terrible
B1	offentlig	public
B1	besviken	disappointed
B1	van	used, accustomed
B1	stilla	calm
B2	kollektiv*	collective, *here*: public (as in public transport)
B2	orättvis	unjust; unfair
B2	tacksam	grateful
B2	relevant*	relevant
B2	främst	foremost
C1	nonchalant*	nonchalant, careless, negligent
C1	förstådd	understood
C1	nedlåtande	condescending, patronizing
C1	acceptabel*	acceptable
C1	övergripande	comprehensive
NOUN		
A1	kaffe*	coffee
A1	dag*	day
A1	pappa*	father, dad
A1	klocka*	watch; clock; at x o'clock
A1	frukost	Breakfast
A2	flygplan	airplane
A2	doktor*	doctor
A2	mage	stomach
A2	kilo*	kilo; kilogram
A2	kött	meat; (flesh)
B1	samarbete	co-operation
B1	ledare*	leader; head; chief
B1	distans*	distance
B1	djurliv	animal life; wildlife
B1	matvana (nb. singular)	eating habits
B2	resurs*	resource; means
B2	avsked	dismissal; goodbye
B2	existens*	existence; livelihood
B2	tillit	confidence; reliance
B2	folkgrupp	ethnic group

C1	anvisning	directions; instructions
C1	flexibilitet*	flexibility
C1	klyfta	gap
C1	utformning	design; shaping
C1	enhet	unit; unity
VERB		
A1	gå	walk
A1	heta	be called, be named
A1	stå	stand
A1	titta	look, glance
A1	äta*	eat
A2	trycka	press; squeeze, oppress sb
A2	meddela	inform sb; let sb know
A2	lägga	put, place; lay
A2	kontakta*	contact, get in touch with
A2	cykla*	cycle; (informal) bike
B1	föreslå	propose, suggest
B1	fokusera*	focus
B1	utvidga	widen; extend; expand; enlarge
B1	stärka	strengthen, confirm, starch
B1	anordna	organize, arrange
B2	brottas	wrestle, fight
B2	tillgodose	meet, satisfy; supply
B2	bedriva	carry on, pursue
B2	klistra	paste, stick
B2	bifoga	enclose, attach
C1	underskatta	underrate, underestimate
C1	värdesätta	value, estimate
C1	bedra	deceive, cheat, be unfaithful to
C1	belysa	light up, illuminate
C1	förebygga	prevent, forestall

Note. Swedish core items for each level and part-of-speech with their translations into English. These are all of our core items, selected as specified in Section 4.2.

Items marked with an asterisk (*) either have cognates in English or the same international loanword is present in English. This is likely to affect how easy English speakers find these items, and hence maybe they should not be seen as anchor items. Cells with a dark grey background are words which learners and experts agreed were among the 10 easiest items, and light grey background marks items which were among the 10 most difficult items by both learners and experts (see Section 6.5).