A New Yardstick of Comparison for Contrastive Interlanguage Analysis
A Study on the ICNALE Global Rating Archives

Shin'ichiro Ishikawa(✉)

Kobe University, Kobe, Japan
iskwshin@gmail.com

Abstract. In the field of learner corpus research, there has been a discussion about what data should be adopted as a reference or a yardstick for contrastive interlanguage analysis (CIA). Utilising the data taken from the ICNALE Global Rating Archives, the author identified a set of Asian learner essays whose quality is guaranteed by a great number of raters having different L1, regional, and occupational backgrounds, and he also examined the possibility of using them as an alternative to the conventional native speaker reference data. This paper discussed three research questions and revealed that (i) among four ENS essays, only one was rated as Level A (70% +), (ii) among 136 learner essays, nine were rated as Level A; and (iii) the choice of a different yardstick (i.e., the essays of native speaker students, the essays of native speaker working adults, and Level A learner essays) leads to the identification of different sets of overused/underused words. These findings will require us to reconsider the type of a yardstick to be adopted in CIA and also the type of an L2 output model to be presented in L2 classrooms from a new angle.

Keywords: ICNALE · Learner Corpus Research · Contrastive Interlanguage Analysis · Reference/Yardstick

1 Introduction

1.1 CIA

Many of the studies based on learner corpora adopt an analytical method called contrastive interlanguage analysis (CIA). CIA is a practical technique to effectively analyse the features of learner outputs. CIA was proposed by Granger (1996), and then it was revised by Granger (2015), which we call CIA1 and CIA2, respectively.

As surveyed in Ishikawa (forthcoming), CIA1 consists of a comparison between a native language (NL) and an interlanguage (IL) and an additional comparison between different interlanguages. The former helps researchers identify the divergence seen in the interlanguage use of the target learner group, and the latter helps them examine whether it is specific to that group (in this case, the divergence is presumably caused by the transfer from the learners' L1), or rather common to a variety of learners (in this case, the divergence is presumably caused by a general developmental problem).

For example, by comparing the essays of Chinese learners of English with those of L1 English native speakers (ENS) as a yardstick, we can identify the divergent features of Chinese learners. Then, by comparing the essays of Chinese learners with those of Japanese or Korean learners, one can investigate whether the features observed in the first comparison are unique to Chinese learners or not, in other words, whether they are the results of the transfer from L1 Chinese or not. These comparisons are usually conducted with a keyword identification function incorporated in major corpus analytic tools, which automatically count the frequency of all the words appearing in a target text and a reference text and identify the words occurring significantly more or less in the target text, using the statistics such as chi-squared values and log-likelihood ratios (LLR). These are usually interpreted as "overuse" or "underuse" by a target group. Such CIA-based examination of overuse/underuse is "a real eye-opener" for learner corpus researchers and also for L2 teachers (Gilquin, 2020, p. 291).

1.2   A Yardstick for CIA

Many of the researchers who adopted CIA1 as their research framework used ENS outputs as a reference or a yardstick for comparison. Surveying representative learner corpus studies, Gilquin (2022) reports that 56% of them involve direct comparisons of learner and ENS data and 26% of them refer to ENS norms in some manner.

Though using ENS outputs as a yardstick seems to be reasonable and even inevitable to some extent, some linguists have criticised CIA researchers' easy dependence on ENS outputs. Granger (2015) admits that "[o]ver the years, CIA has been subjected to a range of criticism, most targeted at the L1/L2 branch." Most of the criticisms are levelled against CIA's dependence on a comparison, its use of a narrow range of ENS outputs, and its negative backwash effects.

First, when comparing learner outputs with something else, learner corpus researchers naturally see them indirectly and with decreased attention. In one sense, the comparison makes learners and their outputs marginalised, which Bley-Vroman (1983) calls comparative fallacy. Selinker (2014) also suggests that "interlanguages must be described in their own terms" rather than in comparison to an external reference.

Next, when conducting an L1/L2 comparison, learner corpus researchers often rely on a relatively limited variety of ENS output samples whose language quality has not been necessarily guaranteed. Regarding the former, Leech (1998) notes, "Yet which native speakers? American, Australian, British, or Caribbean? Highly educated or less so? Old or young?" (p. xix). Then, regarding the latter, he suggests that native-speaking "students do not necessarily provide models that everyone would want to imitate", especially in speeches. In addition, some researchers show doubt about the appropriateness of the construct of "native speakers." For example, Prodromou (1998) ironically mentions that what is real for ENS may also be real for the learners studying in Britain, but it may be unreal for EFL learners in Greece and surreal for ESL learners in Calcutta (p. 266). Thus as Callies (2015) suggests, learner corpus researchers often face the question, "Should this [a yardstick for comparison] be only corpora representing the language of (monolingual) native speakers? And if so, what variety should serve as the comparative basis? And should researchers compare learner data to L1 peers, e.g., novice writers of similar academic standing (students) or expert writers (professionals)?" (p. 40).

Finally, when identifying learners' overuse, underuse, and misuse from the comparison of learner outputs with the ENS outputs, learner corpus researchers may unconsciously accept and even strengthen a set of biased views that ENS are superior to learners; learners need to get closer to ENS; English is a sole referential language in the world; English should be taught only by ENS teachers, and learners need to accept the Western identity. First, Cook (1997) showcases how SLA researchers have accepted and reproduced a view that "L2 learners are failures compared to native speakers." Larsen-Freeman (2014) comments that "[b]y continuing to equate identity with idealised native speaker production as a definition of success, it is difficult to avoid seeing the learner's IL [interlanguage] as anything but deficient" (p. 217). Second, Leech (1998) explicitly casts doubt on the view that "the goal of foreign language learning is to approximate closer and closer to the performance of native speakers" (p. xix). Third, Phillipson (1992) regards "the dominance of English worldwide, and efforts to promote the language" as English linguistic imperialism, which he notes "permeates all the other types of imperialism, since language is the means used to mediate and express them" (p. 65). Phillipson (2012) adds that linguistic imperialism entails "exploitation, injustice, inequality, and hierarchy that privileges those able to use the dominant language." Fourth, Holliday (2006) points out that there exists "the belief that 'native-speaker' teachers represent a 'Western culture' from which spring the ideals both of the English language and of English language teaching methodology," which he calls native speakerism. Such an attitude undoubtedly marginalises and stigmatises non-native English teachers. Holliday also adds that "the 'native speaker' ideal plays a widespread and complex iconic role outside as well as inside the English-speaking West." Prodromou (2008) points out that such a bias may be inherent in CL itself, which he says has continuously advocated a dogma of "real English" in foreign language teaching (pp. 5–7). Fifth, Prodromou (2008) warns that teaching the features of ENS' authentic outputs would inevitably involve "imposing a false identity on the learners" (p. 23).

Among these, the criticism of comparative fallacy may be somewhat extreme in that such a position finally leads to the abolition of any type of comparison in scientific research, which very few would support (Granger, 2015). However, the remaining two criticisms levelled at CIA's tendency to depend on the ENS norm sound considerably convincing. Learner corpus researchers would be required to reconsider the appropriate yardstick to be adopted for CIA.

Carefully scrutinising these criticisms, Granger (2015) presented a revised CIA model. CIA2 replaces the term "native language" with "reference language" and emphasises its diatypic (i.e., register-related) and dialectal (i.e., dialect-related) variability. Also, it suggests the possible task and learner variabilities in interlanguages. CIA2 has come to be "more in line with the current state of foreign language theory and practice" (Callies, 2015, p. 40).

1.3 A New Yardstick

Gilquin (2022) surveys the discussions about a matter of a yardstick in learner corpus research, and she emphasises the need for a shift from "one norm to rule them all" toward diversified "corpus-derived rules." Though she does not necessarily protest the use of ENS yardstick data if it is chosen appropriately, she mentions that learner corpus

researchers can use non-native expert outputs (e.g., academic papers) or institutionalised "new Englishes" as an alternative reference, and they can also choose to focus only on a comparison of the outputs of learners with different L1 backgrounds, with low/high L2 exposure, at different L2 acquisition stages, or those produced in different situations (e.g., timed and untimed).

Among these alternative references, the use of non-native expert outputs as a yard-stick seems to be a promising direction, especially when considering the current status of the English language in the world. However, comparing students' essays with authentic academic papers written by non-native professional scholars would be problematic because they are completely different in terms of the contents, structures, and expected readers.

This may suggest the need for us to look for other types of non-native outputs that can be used as a yardstick. In the following sections, we would like to discuss the possibility of identifying a set of non-native student essays and speeches whose quality is guaranteed, which we call a benchmark sample, and utilising them as a new model in L2 teaching and a new yardstick for CIA.

2   The ICNALE GRA

The author of this paper has long been engaged in the compilation of the International Corpus Network of Asian Learners of English (ICNALE), which is one of the largest collections of Asian learners' speeches and essays (Ishikawa, 2013; Ishikawa, forthcoming). It consists of the modules such as the ICNALE Written Essays, the ICNALE Edited Essays, the ICNALE Spoken Monologues, and the ICNALE Spoken Dialogues.

These modules include not only Asian learners' L2 output samples but also the metadata about their L2 proficiency levels, which are decided from their scores in the English proficiency tests such as TOEIC and TOEFL or in the common vocabulary size tests. Thus, all the learners are classified into four bands linked to the Common European Framework of Reference for Languages (CEFR): A2, B1_1 (B1 Lower), B1_2 (B1 Upper), and B2 +. However, these modules include no data about the linguistic quality of each output.

Therefore, the author recently began a new project. First, we chose 140 learner speeches and the same number of learner essays from the existing ICNALE modules. Then, we recruited more than 100 raters with varied L1, regional, and occupational backgrounds (Ishikawa, 2020) and asked them to assess the whole speeches or the whole essays. The data collected to date has been released as the ICNALE Global Rating Archives versions 0.1 and 0.2.

2.1   Raters

In this project, we first decided to recruit more than 50 raters for each of the essay and speech assessments. Considering that the number of raters in most of the high-stake performance tests is one or two, 50 is a considerably overwhelming number. We thought that the true quality of a learner output could be judged only by the collective knowledge of a large group of raters.

Next, we aimed to recruit raters with varied backgrounds. In TESOL, many believe that a learner output can be assessed in a reliable manner only by experienced ENS teachers. Such a belief, however, seems to be bound by two kinds of dogmas: ENS centrism and teacher centrism, both of which have come to be criticised in applied linguistics. For example, Holliday (2006) suggests that English language teaching has been strongly influenced by the belief that "'native-speaker' teachers represent a 'Western culture' from which spring the ideals both of the English language and of English language teaching methodology." However, the deification of ENS teachers would not be appropriate any longer when thinking of the rapid spread of English as a lingua franca (ELF) in Asia and the world and the unstoppable shift from a teacher-centred approach to a learner-centred approach in the communicative language teaching (CLT).

Regarding the regional and mother tongue diversity, we recruited raters from all of the Inner Circle, where English is spoken as a mother tongue, the Outer Circle, where English is used as a second or an official language, and the Expanding Circle, where English is taught at schools as a foreign language. The rater nationalities currently cover three regions in the Inner Circle (US, Australia, and Canada), five regions in the Outer Circle (Hong Kong, India, Malaysia, Pakistan, and the Philippines), and 10 regions in the Expanding Circle (Cambodia, China, Indonesia, Japan, Jordan, Korea, Laos, Taiwan, Thailand, and Vietnam). The raters' L1 backgrounds cover 17 languages (Arabic, Cantonese, Chinese, English, Filipino, Hmong, Indonesian, Japanese, Konkani, Korean, Lao, Malay, Punjabi, Thai, Urdu, Uyghur, and Vietnamese).

Then, regarding occupational diversity, we collected data not only from English teachers but also from graduate students studying economics, engineering, and sciences, academics majoring in psychology and history, and business people in the fields of media/advertisement, software development, accounting, biomedical engineering, and customer service. The inclusion of the assessments by the business people is especially important because they are significant players in global ELF communication. For instance, the Vienna–Oxford International Corpus of English v2.0 (VOICE), which is one of the most widely used ELF corpora, collects only 25.5% of the whole data from the "educational" domain and 10.1% from the "professional research and science" domains, while it collects 54.5% from the business domains ("professional business" and "professional organisational") (from "Statistics VOICE 2.0 Online").

2.2 Rating

In order to encourage the raters to assess the quality of learner outputs from a balanced viewpoint, we prepared a detailed rating rubric. It includes a holistic assessment (100 points) and ten kinds of analytic assessments (10 X 10 = 100 points). The latter is based on the ten criteria, which cover language (intelligibility, complexity, accuracy, fluency), content (comprehensibility, logicality, sophistication, purposefulness), and attitude (willingness to communicate, involvement).

Also, in the rubric, we explicitly mentioned that the rating standard should be "a professional and proficient ELF speaker who regularly uses English for their business/research purposes" rather than an ideal ENS speaker.

Raters were asked to carefully read the rating guide, which explains the rating principles, the content of the ten criteria, how to decide a score, and so on. Then, they were

asked to take a check test. Only those who passed the test were allowed to begin the rating, which guarantees that all the raters understood the rating principles appropriately and did the assessment job on common ground.

When collecting assessment data, we often realise that some raters are stricter than others, and some raters tend to assign almost the same scores to all the outputs. This undoubtedly deteriorates the quality of rating data. Therefore, in this project, we required all the raters to ensure that the mean of their rating scores falls within the range of 45–55 for a holistic score (/100) and 4–6 for a category score (/10) and the standard deviation (SD) falls within the range of 20–30 for a holistic score (/100) and 2–3 for each category score (/10).

3 Research Design

3.1 Aim and RQs

The aim of the current study is to reconsider the linguistic quality of the ENS outputs as a conventional yardstick, identify a set of learners' benchmark samples that can be used as a new yardstick for CIA, and examine the possible effects of choosing a different yardstick on the results of a keyword analysis as a major CIA technique. This study focuses only on essay rating data.

Our research questions are shown below:

RQ1 Are all the ENS essays rated highly?
RQ2 Are there any learner essays rated highly?
RQ3 How does the choice in yardstick data influence the results of the keyword analysis?

3.2 Data

In this study, we use the rating data that 60 raters assigned to 140 essays, which are taken from the ICNALE GRA Version 0.2. The essays are all written about the topic of "a part-time job for college students" in 200–300 words, in 20–40 min, and without the help of a dictionary.

We chose 20 essays from each of the six regions in the Expanding Circle, four essays from each of the four regions in the Outer Circle, and four ENS essays. All the writers are college students. Thus, each of the 60 raters assessed the whole of the outputs of 20 Chinese students (A2:5, B1_1:5. B1_2:5, B2 +:5), 20 Indonesian students (A2:5, B1_1:6. B1_2:6, B2 +:3), 20 Japanese students (A2:5, B1_1:5. B1_2:5, B2 +:5), 20 Korean students (A2:5, B1_1:5. B1_2:5, B2 +:5), 20 Thai students (A2:6, B1_1:6. B1_2:6, B2 +:2), 20 Taiwanese students (A2:5, B1_1:5. B1_2:5, B2 +:5), 4 Hong Kong students (B2 +:4), 4 Pakistan students (B1_2:1, B2 +:3), 4 Philippine students (B2 + :4), 4 Singaporean students (B2 +:4), and 4 ENS.

3.3 Method

First, we calculate the overall rating score (ORS), which is the mean of the 100-point holistic score and the 100-point analytical score sum, both of which are based on the

Table 1. The ORS values of the ENS essays

Code	Region	ORS	SD	CV	Level
ENS_001	AUS	68.69	22.05	0.32	B
ENS_002	UK	57.03	26.35	0.46	C
ENS_003	NZ	72.57	18.31	0.25	A
ENS_006	US	60.35	20.70	0.34	B
Mean	---	64.66	21.85	0.34	---

Notes: AUS- Australia, UK: the United Kingdom, NZ: New Zealand, and US: the United States of America

scores assigned by 60 raters. Also, we calculate the standard deviation (SD) and the coefficient of variance (CV). CV, which is obtained by dividing SD by the mean values, is regarded as a measure of the variance in the rating scores. We then classify the essays into six levels: Level A (ORS: 70% +), Level B (60% +), Level C (50% +), Level D (40% +), Level E (30% +), and Level F (< 30%).

Then, for RQ1, we examine the ORS values of four ENS essays to see whether all of them are classified as Level A. For RQ2, we identify a set of learner samples whose ORS are higher than 70%, which are regarded as a candidate for a new yardstick for CIA. Finally, for RQ3, we compare all the essays written by novice learners at A2 level in six regions in the Expanding Circle with three kinds of reference data: the essays of all the ENS students, the essays of all the ENS working adults (including both English teachers and business persons), and the Level A learner essays. Then, we conduct keyword analyses to identify the three sets of the top 10 keywords overused or underused by learners at A2 level on the basis of the log-likelihood ratios. Our attention is paid to how the choice of a different yardstick may or may not influence the results of the keyword analyses.

4   Results and Discussions

4.1   RQ1 ENS Essay Quality

The ORS values and the SD/CV values of four ENS essays are summarised in Table 1.

The analysis shows that among the four essays, one (ENS_003) is classified as Level A, two (ENS_001, ENS_006) as Level B, and one (ENS_002) as Level C. As shown in the fact that the mean ORS value does not reach 70%, our assessment data suggests that ENS outputs are not always a good model to follow.

Then, what difference can be seen between the ENS essays at Level C and Level A? Here we compare the essays of ENS_002 (Female, ORS = 57.03) and ENS_003 (Female, ORS = 72.57).

(1)   It's not. But it never ceases to amaze me how many ordinary people believe that it is and then they regurgitate an awful lot of these boring old wives tales like; character building, responsibility, work experience etc. What a load of bollocks. It is quite

simply a case by case scenario really and that's all it is. Those that need a few extra quid for university fees, college fees, books, party cash or living expenses, work part-time jobs and those who are cashed up, don't. That's it in a nutshell really and end of story… If they just want work experience, then they choose to work in a career related company during the holidays. I honestly can't understand what all the kerfuffel is about when they rave on about character building either. I mean, if I want character building, then I'll join the military or a religious sect. I am not going to build much character stacking shelves in the evenings or delivering newspapers at 4 AM in the morning when I should be sleeping, studying or both. Where I come from, if I was in a job interview and told the interviewer that I delivered newspapers in the mornings they would probably say, straight off the bat, "Oh I see, no cash hey?" It's not important to have a part-time job at all for college purposes unless you need the cash. Period! (ENS_002)

(2)   I am a student completing a full-time degree and working a part-time job out of necessity because my parents do not have enough money to pay for all of my tuition and I wasn't fortunate enough to receive a scholarship although, I applied for several. Since starting work, I can honestly say that I am more confident, I manage my time a lot better than any of my classmates and I seem to have gained a certain level of respect from people that I wasn't aware of before. It is difficult at times, especially during exams, but I have been pretty lucky because I have a very understanding boss who goes out of his way to help me in any way that he can. I know my money situation has improved, no surprises there but I also don't stress out as much as some of my mates do when the pressure is on. I think that the way I think about and approach tasks has improved a lot and I make use of a few mini systems that my boss helped me create which allow me to break my tasks down into smaller chunks. I can only say that it has helped me and I feel a lot better within myself. So, I really do think that all students should get a part-time job if not for the money, then just for the extra skills that they can get while being paid to do so. (ENS_003)

ENS_002 seems to be a passionate writer. Her essay is unique, and her sense of irony stands out. Her claim is interesting and also reasonable, but her consistent use of too much colloquial vocabulary ("an awful lot of…," "these boring old wives tales," "What a load of bollocks," "Period!") and conversational grammar ("It's not""A part-time job is not important for college students," "It is quite simply a case by case scenario really…"  "It is (quite) simply a case-by-case scenario," "Those that…"  "Those who,…" "work part-time jobs"  "work part-time" or "do/have a part-time job") in writing makes her text less readable and less academic, which presumably explains why relatively lower rating scores are assigned to this ENS essay.

Meanwhile, ENS_003's claim seems to be rather banal, but she describes her own experiences in great detail, and she tries to show both the merits and demerits of working part-time to discuss the topic from a more balanced viewpoint. Her stance control and sincere attitude as a writer of academic writing lead to the greater persuasiveness of her essay, which raters seem to have evaluated highly.

As mentioned before, some of the L2 researchers and teachers have tended to believe in the quality of the ENS outputs rather blindly, but our analysis shows that it might not

Table 2. Level A learner essays

Circles	Samples (ORS)
Outer Circle	SIN_006 (79.87), PHL_064 (77.42), SIN_005 (77.13), SIN_014 (76.98), HKG_008 (75.04)
Expanding Circle	CHN_009 (73.87), IDN_177 (70.68), CHN_160 (70.38), KOR_004 (70.24)

Notes: SIN: Singapore, PHL: the Philippines, HKG: Hong Kong, CHN: China, IDN: Indonesia, KOR: Korea

be appropriate. As Leech (1998) notes, ENS' L1 output samples are not always a good model that learners would like to follow.

4.2   RQ2 Learner Essay Quality

Then, are there any learner essays classified as Level A? The analysis identified nine essays at that level, which are shown in Table 2.

Table 2 shows that Asian learners not only in the Outer Circle but also in the Expanding Circle can produce high-quality essays, which can be a candidate for a new model in teaching as well as a new yardstick for CIA. Eight of the nine writers are at the B2 + level, but one (CHN_009) is at the B1_2 level, suggesting that there may exist a gap between the learners' proficiency levels estimated from their external test scores and the quality of their essays assessed directly by raters.

Here we like to examine the essay of CHN_009 (Female). This is a noteworthy sample in that it is written by a learner in the Outer Circle and at the B1_2 level.

(3) Nowadays, quite a number of college students are taking part-time jobs. But is it necessary for them to do so? In my opinion, the answer can be various, depending on the students themselves. Undoubtedly, taking a part-time job is a special experience which provides youths with the chance to get close to society as well as to gain more practical skills. In the process, the students learn how to communicate with different kinds of people and how to deal with challenging situations. Besides, the money they get from the job will certainly reduce the burden of their parents more or less. However, it can't be ignored that some students fail more important things in order to take part-time jobs. They spend time which should be for learning to work, thus leading to decline of grades. But the fact is that the first thing for students is still study. Also, some young people fail to spend their money reasonably. Instead, they fall into the habit of wasting money. So, it can be concluded that taking a part-time job is rewarding and eye-opening for those who have good self-regulation but it may have some side effects if the student doesn't have a good plan. In common with many things, part-time job is a double-edged sword. If you want to take it, do be careful. (CHN_009)

Here we like to discuss three outstanding features seen in this essay. First, it is highly involving and engaging. CHN_009 often refers to the readers by asking a question ("But

is it necessary for them to do so?") and using a second-person pronoun ("If you want to take it…"), which helps make readers naturally involved in the text. Second, regarding the topic of whether a college student should work part-time or not, the author tries to present the third position. She introduces both the merits and demerits of working part-time with concrete and ample examples and insists that it is good for the students who can regulate themselves appropriately but may not be good for the other students. Then, she insists that "the answer can be various" because the experience of having a part-time job is "a double-edged sword." Her claim is well-balanced and therefore accepted by a variety of readers. Third, the author ends her essay with a concise but impressive remark ("If you want to take it, do be careful"), which effectively summarises the whole discussion and makes her text convincing and memorable.

It is true that there remain several minor language problems (e.g., "fail more important things"  "fail *in* more important things," "the first thing for students is still study"  "the first thing for the students to do is…" or "the most important thing for the students (to do) is…," "do be careful"  "be careful"), but in reality, they hardly influence the readability of the text. When a teacher uses this sample as one of the models in a classroom, s/he could point out the problems and show how they should be corrected.

As exemplified above, the Level A learner essays, including the one of CHN_009, could be utilised as one of the alternatives to the ENS outputs as a model in teaching and a yardstick for CIA.

4.3   RQ3 Effects of Changing a Yardstick

In the last section, we identified a set of high-quality learner essays that may be used as a new yardstick. Then, when adopting such a new reference for CIA-based keyword analysis, what difference can be seen?

Table 3 shows three sets of the top 10 words overused or underused by Asian learners at A2 level when compared to three referential yardsticks: the essays of ENS students (ENS_ST), those of ENS working adults (ENS_WA), and Level A leaner essays (LEV_A). The overlapping words appear in bold italics. Then, Table 4 summarises the ratios of the overlapping words.

Regarding the overused words, the overlapping ratios are 50–60%. The choice of a different yardstick leads to the identification of partly different sets of keywords, but there also exists overlapping. The data shows that Level A learner essays as a new yardstick can yield keywords considerably similar to those identified by the two ENS yardsticks. Then, overlapping items suggest that novice learners in Asia tend to choose a group viewpoint ("we"), explicitly mention the reasons for their claims ("because"), discuss the ability ("can"), adopt hedging expressions ([I] "think"), and use the words directly related to the essay topic ("money," "society," "job").

Meanwhile, regarding the underused words, the overlapping ratios are only 0–10%. In comparison to overused words, underused words seem to be less stable, less replicable, and less reliable. It is true that Level A learner essays as a new yardstick hardly yield the overlapping keywords, but this also applies to the two ENS yardsticks.

Table 3. Key overused/underused words for A2 learners in Asia identified on the basis of three yardsticks

	ENS_ST	ENS_WA	LEV_A
Words overused by A2 learners in Asia			
1	*we*	*we*	*we*
2	*money*	*money*	I
3	*society*	part	*money*
4	study	you	*think*
5	must	time	*can*
6	*job*	*can*	lot
7	free	*job*	know
8	agree	our	don
9	earn	*society*	*because*
10	*because*	*think*	t
Words underused by A2 learners in Asia			
1	groups	would	workforce
2	found	that	studies
3	ve	believe	*their*
4	gain	*their*	few
5	finding	skills	these
6	basically	financial	jobs
7	cut	then	education
8	expectations	just	results
9	requirements	they	thus
10	shows	employers	on

Table 4. Overlapping ratios

	ENS_ST	ENS_WA	LEV_A
Overuse	50% (5/10)	60% (6/10)	50% (5/10)
Underuse	0% (0/10)	10% (1/10)	10% (1/10)

Thus, our analyses revealed that choosing a different yardstick yields different keywords, and Level A learner essays as a yardstick are not necessarily inferior to the conventional ENS yardsticks. This strongly suggests that Level A learner essays are qualified to be used as an additional rather than alternative yardstick for CIA.

5   Conclusions

The current study discussed three RQs. First, regarding RQ1 (ENS essay quality), we revealed that among four ENS essays, one was rated as Level A (70% +), two as Level B (60% +), and one as Level C (50% +). This finding requires us to reconsider the conventional beliefs that ENS' L1 outputs are always a good model that learners should follow and they are a stable and reliable yardstick for CIA.

Then, regarding RQ2 (Learner essay quality), we revealed that among 136 learner essays, nine were rated as Level A. Also, it was shown that not only the Asian learners in the Outer Circle but also those in the Expanding Circle could produce top-quality essays whose values were acknowledged by a great number of raters.

Finally, regarding RQ3 (Effects of changing a yardstick), it was shown that the choice of a different yardstick (ENS students' essays, ENS working adults' essays, and Level A learners' essays) leads to the identification of different sets of overused/underused words. It was also shown that Level A learner essays as a yardstick can yield the overused keywords considerably similar to those identified by the two ENS yardsticks.

These findings shed new light on the discussion about the matter of choosing a model for L2 teaching and a yardstick for CIA. In order to avoid the risk of imposing ENS centrism on L2 learners and teachers, learner corpus researchers now need to explore the possibility of using high-quality learner outputs as a new model/yardstick.

References

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*, 1–17.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp.35–55). Cambridge University Press.

Cook, V. (1997). Monolingual bias in second language acquisition research. *Revista Canaria de Estudios Ingleses*, *34*, 35–50.

Gilquin, G. (2020). Learner corpora. In M. Paquot & S. Th. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 283–304). Springer

Gilquin, G. (2022) One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, *55*(1), 1–13.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerised bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies* (pp.37–51) Lund University Press.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24.

Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*(4), 385–387.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91–118.

Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of Asian learners' L2 English essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World*, *5*, 121–144.

Ishikawa, S. (forthcoming). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English.* Routledge.

Larsen-Freeman, D. (2014). Another step to be taken: Rethinking the end point of the interlanguage continuum. In Z. Han & E. Tarone (Eds.), *Interlanguage. Forty years later* (pp. 203–220). John Benjamins.

Leech, G. (1998). Preface. In S. Granger (Ed.),*Learner English on computer* (pp. xiv–xx). Longman

Phillipson, R. (1992). *Linguistic imperialism.* Oxford University Press.

Phillipson, R. (2012). Linguistic imperialism. In C. A. Chapelle (Ed.), *The encyclopaedia of applied linguistics* (online). Blackwell.

Prodromou. L. (1998). ELT Journal correspondence. *ELT Journal*, *52*(3), 266–267.

Prodromou, L. (2008). *English as a lingua franca: A corpus-based analysis.* Continuum.

Selinker, L. (2014). Interlanguage 40 years on. Three themes from here. In Z. Han & E. Tarone (Eds.), *Interlanguage: Forty years later* (pp. 221–246). John Benjamins.