Comparing Native and Learner Englishes Using a Large Pre-trained Language Model

Tatsuya Aoyama
Georgetown University
ta571@georgetown.edu

Abstract

The use of lexical items by L2 speakers of English has been analyzed through a variety of methods; however, they are either (i) infeasible for a large-scale learner corpus study or (ii) designed to measure vocabulary breadth, rather than depth. This paper presents the preliminary results of an ongoing work to utilize contextualized word embeddings (CWEs) obtained from a large pre-trained language model to measure the depth of L2 speakers' vocabulary knowledge, operationalized as how similar L2 speakers' use of a given word is to that of L1 speakers'. We find that (i) the mean distance between L1 CWEs and L2 CWEs of a given word tends to decrease as the proficiency level becomes higher, and that (ii) while words that have similar CWEs in the L1 corpus and L2 corpus tend to reveal interesting properties about the word use, words that have dissimilar CWEs in the two corpora often suffer from domain effects.

1 Introduction

Characterizing learner language has been a major task in second language acquisition (SLA) literature. Various aspects of texts produced by second language (L2) English speakers have been shown to deviate from first language (L1) English speakers, such as syntactic (Pienemann, 1998) and morphological aspects (Goldschneider and DeKeyser, 2001). Among these, deviation in word use is particularly difficult to characterize as the "native-likeness" of word use is not as clear-cut as that of syntactic or morphological knowledge, where the correct and incorrect use is rather clearly defined. Hence, a number of methods to measure L2 speakers' word use have been proposed, and they are roughly categorized as capturing either the breadth (how *many*) or depth (how *well*) of vocabulary knowledge (Wesche and Paribakht, 1996). While a number of automatic measurements for vocabulary breadth have been proposed (e.g., Lu, 2012), measurements for vocabulary depth largely remain infeasible for a large-scale learner corpus study. As such, we propose a new approach to measure vocabulary depth by leveraging the word embeddings obtained from a large language model.

In fact, the idea of using word embeddings to compare word use across different populations is not new; for example, using this approach, Del Tredici and Fernández (2017) investigated semantic variation across different communities of practice, and Hamilton et al. (2016) studied diachronic semantic change. Since this approach is applicable to any comparative analysis as long as it involves multiple populations that speak the same language, the current study aims to extend this approach to measure the depth of vocabulary knowledge, operationalized as how similar L2 speakers' use of a given word is to that of L1 speakers' (i.e., a comparative analysis of L1 and L2 Englishes).

In the subsequent sections, we will first briefly review existing approaches to measure learners' word use in SLA literature (§2.1) and show how methods from distributional semantics can be employed to tackle this problem (§2.2). We then describe the data, model, and experiments (§3), followed by the results (§4) and discussion (§5), including implications and limitations, as well as future directions.

2 Relevant work

2.1 Vocabulary acquisition

Vocabulary acquisition garners a considerable attention in SLA literature, and various operationalizations and measurements have been proposed. For example, within the widely used proficiency measurement framework of complexity, fluency, and accuracy (CAF; Skehan et al., 1998), lexical complexity is measured by several indices, such as type-token ratio and lexical sophistication of L2 writing (Norris and Ortega, 2009). The former represents the (absence of) repetition in vocabulary,

This work is licensed under a Creative Commons Attribution 4.0 International Licence.
Tatsuya Aoyama. Comparing Native and Learner Englishes Using a Large Pre-trained Language Model. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022)*.
Linköping Electronic Conference Proceedings 190: 1–9.

and the latter captures the use of low-frequency lexical items.

Other approaches suggest the importance of breadth-depth distinction: the former represents how many vocabulary items a learner knows, and the latter represents how well a learner knows a certain vocabulary item (Nation, 2001). Multiple choice questions are among the most common measures of vocabulary breadth, whereas a variety of tests are used to measure vocabulary depth, such as completing idiomatic expressions, filling in the blank using collocation knowledge, and writing down synonyms of a given vocabulary item (Milton, 2009).

Some approaches (e.g., type-token ratio, vocabulary sophistication) can be applied to written texts automatically (e.g., Lu, 2012) and hence scalable to a large-scale learner corpus study; however, they are not capable of measuring anything beyond vocabulary breadth (i.e., how *many* words) or how advanced those known words are. Other approaches (e.g., idiom, collocation, or synonym) are better proxies for measuring vocabulary depth, yet the reliance on the carefully crafted tests and the need for learners to take them make these approaches expensive. Hence, we argue for the use of methods from distributional semantics to obtain a richer representation of word usage, rather than relying on the *counts* of word use in a given text.

2.2 Distributional semantics

The idea that the meaning of a given word is captured in the distribution of the word (i.e, co-occurring words) in a given corpus is called distributional hypothesis (Harris, 1954). Building on this hypothesis, Salton (1971) proposed vector space model, where a word can be represented as a point in high-dimensional vector space based on the count of neighboring words. This count-based vector space model has spawned a number of studies that investigate its linguistic implications (Erk, 2012).

Replacing this count-based approach, Mikolov et al. (2013a) proposed a prediction-based approach called word2vec, enabling a given word to be represented as a low-dimensional dense vector, instead of the traditional term-to-term sparse vector. Subsequent studies find that surprising amount of linguistic information is captured in this dense representation. For example, Mikolov et al. (2013b) find that word vectors can be added and subtracted to derive another word vector. An often-cited example is that the vector for *Queen* can be approximated by $King - Man + Woman$.

Both count-based and prediction-based approaches described above generate type-based word vectors, meaning that each word type receives a single word vector. The advent of language models capable of taking contexts into consideration, such as ELMo (Peters et al., 2018), enables each *word token*, rather than *word type*, to receive a separate word vector, often referred to as contextualized word embeddings (CWEs). Of such language models, BERT (Devlin et al., 2019) is perhaps the most widely used and extensively studied (see Rogers et al. 2020 for an overview of the studies that investigate BERT's internal representation).

While much work has been devoted to applying word embeddings to downstream NLP tasks, and their usefulness has been widely recognized (e.g., Devlin et al., 2019; Liu et al., 2019), others utilized them for more theoretical investigations, such as semantic variation across communities of practice (Del Tredici and Fernández, 2017), diachronic semantic shift (Del Tredici et al., 2019; Hamilton et al., 2016), and variation in semantic frames across different languages (Sikos and Padó, 2018). Most, if not all, studies of the latter kind (theoretical investigation) adopt type-based word embeddings; hence, this study is arguably the first of its kind to apply CWEs to perform a comparative analysis of language use by multiple populations (see §3.2).

In light of all this, we ask the following two questions:

1. Do CWEs capture the depth of L2 speakers' vocabulary knowledge? In other words, can we infer how well L2 speakers know a given vocabulary item by comparing its CWE from that of L1 speakers'(§4.1)?
2. How does word use differ across the two populations (L1 and L2 speakers), and what are the words that diverge the most/least? (§4.2)?

3 Method

3.1 Data

Since few large-scale learner corpora are readily available, a learner corpus was selected first to ensure that an appropriate native English corpus could be selected based on the nature of the chosen learner corpus. For learner corpus, we use the EF-Cambridge Open Language Database (EFCAMDAT; Huang et al., 2017; Geertzen et al., 2013),

a collection of essay assignments written by non-native English speakers of various first languages. EFCAMDAT contains more than 83,000,000 word tokens in the essays written by more than 170,000 learners of English, and each essay is accompanied by metadata, including the proficiency level that ranges from 1 to 16.[1] For this study, only the essays written by Japanese learners of English are used, amounting to 1,602,328 words from 21,374 essays written by 3,441 learners.

For native English corpus, we use the Louvain Corpus of Native English Essays (LOCNESS; Granger, 2014). LOCNESS is a corpus of argumentative and literary essays written by university students in the U.S. and in the U.K. To make the speaker profile homogeneous and to ensure the comparability of the texts to EFCAMDAT, only the argumentative essays written by university students in the U.S. were included in this study. This resulted in 176 essays and 149,574 words in total.

Note that the selected subcorpus from EFCAMDAT has the mean length of 74.96 words per essay, whereas the LOCNESS counterpart has the mean length of 849.85 words per essay. This is partly due to the fact that EFCAMDAT consists of essays written by learners of varying proficiency levels, and the ones written by lower proficiency learners are much shorter. Although the two corpora are not perfectly comparable, they share the same genre (i.e., essay assignment) and are considered at least minimally appropriate for the purpose of this study. Implications and limitations of the difference between the two corpora will be further discussed in §5.2.

3.2 Model

To obtain CWEs, we used `flair` implementation of BERT (Akbik et al., 2019). BERT has rarely been, if ever, used to perform a comparative analysis of language use by multiple populations, and most studies of this kind use type-based word embeddings, as discussed in §2.2. Although a similar approach could have been taken in this study as well, BERT was preferred for a few reasons.

First, language models like BERT are pre-trained on a large amount of data; therefore, we can obtain a CWE that represents the meaning of the word *given the context* by simply feeding a word with its context (e.g., sentence) to the model. Training a language model from scratch, as is the case with word2vec, often requires a large amount of data, and this is an important advantage in favor of BERT given the limited amount of accessible L2 English texts.

Second, BERT consists of 12 (`bert-base`) or 24 (`bert-large`) layers, and a number of studies have suggested that each layer encodes different linguistic information, such as surface, syntactic, and semantic information (e.g., Tenney et al., 2019; Jawahar et al., 2019). With these insights about BERT's internal structure, different aspects of word use could potentially be elucidated by analyzing CWEs obtained from each of the BERT's internal layers. We will leave this to future studies (see §5.2) and focus on the CWEs obtained from the final layer in this paper.

Lastly, BERT's attention mechanism (Vaswani et al., 2017) allows the model to learn how much attention to pay for each word (i.e. how much to *weigh* each word) in a given context. Therefore, CWEs obtained from BERT is, in a sense, a weighted sum of all the words' embeddings in a given context. This may allow us to capture the difference in word use more fully, compared to models that only use its k neighboring words, where k is a hyperparameter, during its training phase (e.g., word2vec).

3.3 Experiment

To answer the two research questions, we need to define what it means to *compare* the word use between two populations. Here, we base our comparative analyses on the two centroids of a given word, one for each population. More specifically, the following steps were taken to obtain CWEs for each of the native and learner corpora described in §3.1.

1. Create a vocabulary list of 1,000 most frequent lexical items.
2. Obtain CWEs for each occurrence of each lexical item in the vocabulary list created in 1.
3. Calculate the centroid of the embeddings for each lexical item.

In 1, we only use the top 1,000 frequent words to ensure that the centroids obtained in 3 is a reliable representation of the word usage by the population (i.e., L1 or L2). For 2, an entire sentence was fed into BERT (Devlin et al., 2019) to obtain a CWE of each occurrence of a given word. Once the above steps are completed for each of the corpora,

[1] Full details available at: `https://corpus.mml.cam.ac.uk/faq/EFCamDat-Intro_release2.pdf`

the similarity score between the two centroids was calculated for each lexical item that appears in both of the two vocabulary lists. For example, if a word *argument* appears 15 and 30 times and in native and learner corpus, respectively, and qualifies as the 1,000 most frequent items in both corpora, the centroid of those 15 CWEs of *argument* from the native corpus and the centroid of those 30 CWEs of *argument* from the learner corpus will be compared using Euclidean distance. Since we are interested in the difference (or similarity) in the word use between the two corpora, rather than among the individuals, we calculate the ratio of inter-corpus distance to intra-corpus average distance. Formally, we define the metrics as following:

$$Score = \frac{Dist_{inter-corpus}}{Dist_{intra-corpus}} \quad (1)$$

$\text{Dist}_{\text{inter-corpus}}$ is a simple Euclidean distance between the two centroids. $\text{Dist}_{\text{intra-corpus}}$ is an average distance of each CWE from the centroid of a given word w in a given corpus c:

$$\frac{\sum_{c\in\{L1,L2\}} \sum_{i=1}^{N_{w,c}} |Centroid_{w,c} - \vec{w_{i,c}}|}{\sum_{c\in\{L1,L2\}} N_{w,c}}, \quad (2)$$

where $N_{w,c}$ represents the total number of occurrences of the word w in a given corpus c.

For research question 1, hypothesizing that the depths of L2 learners' vocabulary knowledge increase as they become more proficient in their L2, we would expect the distance score to decrease (i.e., L2 CWEs become more similar to L1 CWEs) for learners with higher proficiency levels. In order to show this, the experimental steps are slightly modified: instead of using the aggregate L2 corpus, we treat each proficiency level as a separate population; hence, the steps defined above were repeated 16 times, once per proficiency level. Each of these sub corpora will be referred to as L2-{level} (e.g., L2-6 for the level 6 sub-corpus taken from the L2 corpus).

For research question 2, no such modifications were necessary, as we are interested in investigating how the word use diverges between the two populations, as well as what words show most/least divergences.

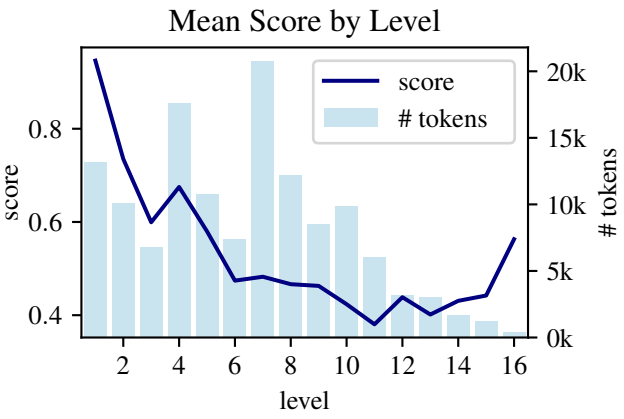

Figure 1: Mean Score by Proficiency Level

4 Results

4.1 Research question 1

A total of 100 words were found to be among the top 1,000 frequent words in all of the corpora (i.e., L1 corpus and 16 proficiency-based L2 subcorpora). To quantify the (dis)similarities in word use between L1 corpus and each of the 16 L2 subcorpora, we calculate the distance score defined in eq. (1) for each of these common 100 word types, and take their mean weighted by the number of tokens per word type for each of the 16 comparisons (i.e. L1 vs L2-1, L1 vs L2-2, ..., L1 vs L2-16). The results are summarized in figure 1, where the line plot represents the mean distance score, and the bar graph represents the total number of occurrences of the 100 common words per proficiency level.

We can observe an overall decreasing tendency, with the trend being particularly pronounced at levels from 1 to 11. This seems to validate the use of CWEs to measure the depth of vocabulary knowledge, hypothesizing that the depth grows (i.e. distance score should decrease) as the proficiency level becomes higher.

However, the slight increase in the mean distance score towards the highest proficiency levels is worth noting. Although a further investigation is necessary to explain this result, a few possibilities are considered here. First, as the bar graph suggests, the start of the increase in the mean distance score at level 12 coincides with the decrease in the sample size. That is to say, the number of tokens at levels 12 to 16 are substantially smaller than those at levels 1 to 11. Hence, this may be the result of small sample sizes, and the result may have been different with larger sample sizes for higher proficiency learners. Another possibility is that,

Word	Score	D_{Inter}	D_{Intra}	# (L1)	# (L2)	Word	Score	D_{Inter}	D_{Intra}	# (L1)	# (L2)
men	0.21	1.51	7.31	147	351	basketball	4.80	18.60	3.87	17	251
women	0.22	1.35	6.04	320	484	expensive	3.43	16.57	4.83	16	596
since	0.23	2.32	10.16	101	479	communication	3.04	15.42	5.08	14	214
time	0.24	2.12	8.87	288	3039	enjoy	2.73	16.29	5.97	18	831
things	0.24	1.80	7.44	111	666	concern	2.10	8.45	4.02	20	429
according	0.24	2.36	9.84	55	125	wild	1.72	8.10	4.71	37	182
one	0.26	2.55	9.97	595	2798	florida	1.65	6.37	3.85	49	128
less	0.26	1.98	7.75	82	254	name	1.33	7.62	5.72	24	2446
say	0.26	2.09	7.91	108	384	actually	1.25	11.35	9.11	38	300
even	0.27	2.83	10.64	213	594	contact	1.24	7.63	6.17	19	464

Table 1: Top 10 similar words (left) and top 10 dissimilar words (right)

because essay topics vary across proficiency levels, it may simply be the case that the topics at higher proficiency levels happened to be different from the topics L1 essays are written on. Alternatively, and more interestingly, if this U-shaped curve is truly capturing the relationship between the depth of vocabulary knowledge and proficiency level, we may have to revise our hypothesis that the distance score will monotonically decrease as a function of proficiency level. We will return to this point in §5.1.

4.2 Research question 2

A total of 465 words were found to be among the top 1,000 frequent words in both corpora. Of these 465 words, the most similar and dissimilar words were identified based on the distance score defined in eq. (1), and the results are summarized in table 1.

4.2.1 Words with smaller distances

For the 10 most similar words (left), the score ranges from 0.21 to 0.27, meaning that the distance between the L1 centroid and the L2 centroid was about 4-5 times smaller than the average distance between the centroid and each of the word token *within* the corpus. The 2 most similar words, *men* and *women*, are perhaps due to the similar essay topics coincidentally present in the two corpora. A naive comparison of the 10 most frequent words that appear in the same sentence as the word *men* show that *women*, *equal*, *society*, and *children* often co-occur with *men* in L1 corpus, while *women*, *work*, and *equal* are the common neighboring words in L2 corpus. This large overlap suggests that *men* and *women* both occur in the context of an essay prompt about gender equality in both corpora.

More interestingly, other words in table 1 include generic words that could appear in a variety of contexts, such as *since*, *according*, *even*, and *things*.

Word	# (L1)	Word	# (L2)
would	22	though	69
people	21	people	53
one	16	work	29
though	15	one	29
may	14	many	25
time	14	think	25
make	10	get	24
could	10	like	22
many	9	time	22
use	9	go	22

Table 2: Top 10 Co-occurring words with *even* in L1 corpus (left) and in L2 corpus (right)

Since the context in which these words appear is likely to be affected by the domain of the text, it is reasonable to expect these words to have high inter-corpus distances; however, they all have relatively small inter-corpus distance. For example, the 10 most frequently co-occurring items of the word *even* in each of the two corpora are summarized in table 2. In both corpora, *though*, *many*, and *people* seem to co-occur frequently with the word *even*. A possible interpretation is that, in argumentative essays, both L1 and L2 English speakers use *even* as a way to express concession or contrast (as in *even though*), and that the conceding or contrasting proposition, which is secondary to the main proposition, tends to be general.

In a similar vein, *things* frequently co-occur with *people*, *money*, *life*, and *time* in L1 corpus, and with *people*, *time*, and *life* in L2 corpus. This overlapping in the co-occurring words seems to suggest that L1 and L2 English speakers both use the word *things* as a way to describe general facts or truths about the world, pertaining to people, money, and time. This may be due to the genre of the L1 and L2 texts–because they both contain argumentative es-

Word	# (L1)	Word	# (L2)
many	10	like	16
prayer	7	'm	13
would	6	however	11
people	4	people	10
religion	4	work	9
society	3	know	7
public	3	show	5
students	3	much	5
recite	3	really	5
times	3	person	5

Table 3: Top 10 Co-occurring words with *actually* in L1 corpus (left) and in L2 corpus (right)

says,[2] these general statements may be used to set the scene before introducing the main arguments.

4.2.2 Words with larger distances

Of the 10 most dissimilar words (right), *basketball* had the highest score of 4.80, meaning that the distance between the two centroids was almost 5 times larger than the average distance within each of the corpora. The co-occurring words in the L1 corpus, such as *respect*, *men's*, *women*, *coach*, and *colleges* suggest that L1 English speakers tend to use *basketball* in the context of collegiate athletics. In the L2 corpus, on the other hand, the frequent co-occurring words, such as *afternoon*, *every*, *games*, and *computer* seem to suggest that the word is used in the context of hobbies or daily routines. This may not be so much of a difference in the word use as a difference in the domain, since EFCAMDAT contains non-argumentative essay assignments, such as describing routines.

Table 3 lists the top 10 co-occurring words of another dissimilar word *actually* in L1 corpus (left) and in L2 corpus (right). Apart from the effect of domain difference similar to above observation on *basketball*, table 3 reveals interesting ways in which its usage differs between the two population.

First, it is worth noting the contrast between *even* and *actually*. That is to say, although both of them are adverb and carry less "content" compared to more strongly content words, such as verbs and nouns, *even* is robust to the domain difference, whereas *actually* is not. This may be explained by their use in discourse. On the one hand, *actually* is commonly used to introduce a piece of information expected to be surprising to the audience, and the proposition is often specific to the topic or domain of the text. *Even*, on the other hand, can be used to mark concession or contrast (as in *even though*) as discussed above, and the subordinate clause introduced by *even* tends to be a general statement to which the main clause (often the main proposition) is antithetical.

Second, in L2 corpus, *however* is frequently used in combination with *actually*. A manual inspection of these 11 sentences where *actually* and *however* co-occurred shows that these two words occur within the same clause in 4 of these 11 sentences, meaning that they are used to modify the same proposition as shown in an example below:

(1) However, actually it's still difficult for women to continue their work after they get married and have children. *(writing id = 1064583)*

Notably, L1 corpus contains only 1 co-occurrence of *actually* and *however*, and it is not within the same clause. This difference may reasonably be attributed to the difference in the size of the two corpora; however, it may be the case that the meaning of *however* is construed slightly differently by the two populations. However, whether this is a difference in meaning (e.g., *actually* containing contrastive meaning or not) or a mere collocation knowledge (e.g., *actually* and *however* are simply not used together conventionally) remains inconclusive.

5 Discussion

5.1 Implication

This paper argued for the use of CWEs obtained from a large pre-trained language model to analyze the word use of L1 and L2 speakers of English and presented the preliminary results. We found that (i) the mean distance between L1 CWEs and L2 CWEs of a given word tended to decrease as a function of proficiency level, and that (ii) while similar uses of a given word by the two population are due to either a domain effect (e.g., *men* and *women*) or a particular function the word plays in the discourse (e.g., *things* and *even*), dissimilar uses of a word, on the other hand, were mostly the result of domain differences (e.g., *basketball*).

For (i), an exception to the overall decreasing trend was observed at levels 12 to 16, where the mean distance score slightly increased. This may be due to methodological reasons, such as imbal-

[2]Texts in EFCAMDAT are all essay assignments, but they include non-argumentative ones as well.

anced sample sizes and essay topics (see §4.1). Alternatively, U-shaped curve may actually be representative of the true relationship between the depth of vocabulary knowledge and proficiency level, rather than caused by some methodological limitations. This phenomenon, where a certain aspect of linguistic knowledge appears to *regress* in the process of L2 learning, is referred to as *backsliding* or *regression* (Selinker, 1972;Selinker and Lamendella, 1981; Lantolf and Aljaafreh, 1995). If this was the case, it may be an oversimplification to operationalize the depth of vocabulary knowledge as how similar an L2 learner's use of a given word is to that of L1 speakers'. Although this point remains inconclusive in our preliminary results, it is an important question to address in the future research.

5.2 Limitations and future directions

Although this study contributes to the existing body of literature by arguing for the use of CWEs obtained from a large pre-trained language model to investigate L1 and L2 Englishes, some limitations and future directions were identified.

First, as has been mentioned in §3.1, the mean sentence length of the selected L2 subcorpus is much shorter (74.96 words) than that of the L1 corpus (849.85). Since longer essays may contain more anaphoric expressions such as pronouns, it may affect the contexts in which words occur. However, since low proficiency L2 speakers tend to write shorter sentences, it is challenging to balance the mean sentence length across the two populations.

Second, we opted for BERT as a way to obtain CWEs because of its use of the entire sentence to contextualize the word embeddings. However, this might have amplified the domain effect (i.e., differences in topic, prompt). Hence, using a model that leverages more immediately neighboring words (by adjusting the hyperparamter of n-gram size), such as word2vec (Mikolov et al., 2013a), separately for each of the two corpora may enable a more domain-agnostic comparison of the word use. In fact, Sikos and Padó (2018) used English and German corpora of different domains to train separate frame embeddings using word2vec, yet the results yielded meaningful comparisons and implications. However, training a model from scratch is not an viable option for the data used in this study, since some sub-copora, especially the ones with a higher proficiency, had substantially smaller sample sizes.

Third, BERT's layers have been shown to encode distinct linguistic information (Rogers et al., 2020). For example, middle layers encode syntactic information (Hewitt and Manning, 2019), whereas higher (closer to the final) layers encode more abstract semantic information (Jawahar et al., 2019; Tenney et al., 2019). Although this study used the outputs from the final layer, future studies could obtain different insights by obtaining outputs from each of the 12 layers.

Lastly, once the above limitations are resolved and more meaningful differences in word use between L1 and L2 Englishes can be reliably obtained, it may be promising to investigate the embedding space in multilingual BERT (Devlin et al., 2019). For example, hypothesizing that the different word use is the result of L1 transfer, the deviation of the centroid vector may be in the direction of the vector of an equivalent word in the learners' L1 (e.g., Japanese speakers' use of an English word *love* may be slightly shifted towards its Japanese counterpart *ai*, compared to native English speakers' use of the same word).

Acknowledgements

We thank the two anonymous reviewers for their insightful and constructive feedback.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254. Citeseer.

Jennifer M Goldschneider and Robert M DeKeyser. 2001. Explaining the "natural order of l2 morpheme acquisition" in english: A meta-analysis of multiple determinants. *Language learning*, 51(1):1–50.

Sylviane Granger. 2014. The computer learner corpus: a versatile new source of data for sla research. In *Learner English on computer*, pages 3–18. Routledge.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Yan Huang, Jeroen Geertzen, Rachel Baker, Anna Korhonen, Theodora Alexopoulou, and EF Education First. 2017. The ef cambridge open language database (efcamdat): Information for users.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

James P Lantolf and Ali Aljaafreh. 1995. Second language learning in the zone of proximal development: A revolutionary experience. *International Journal of Educational Research*, 23(7):619–632.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

James Milton. 2009. Measuring second language vocabulary acquisition. In *Measuring Second Language Vocabulary Acquisition*. Multilingual Matters.

Ian SP Nation. 2001. *Learning vocabulary in another language*. Cambridge university press.

John M Norris and Lourdes Ortega. 2009. Towards an organic approach to investigating caf in instructed sla: The case of complexity. *Applied linguistics*, 30(4):555–578.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Manfred Pienemann. 1998. *Language processing and second language development: Processability theory*, volume 15. John Benjamins Publishing.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

G Salton. 1971. The smart system. *Retrieval Results and Future Plans*, 260.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–231.

Larry Selinker and John T Lamendella. 1981. Updating the interlanguage hypothesis. *Studies in Second Language Acquisition*, 3(2):201–220.

Jennifer Sikos and Sebastian Padó. 2018. Using embeddings to compare FrameNet frames across languages. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 91–101, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Peter Skehan et al. 1998. *A cognitive approach to language learning*. Oxford University Press.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Marjorie Wesche and T Sima Paribakht. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53(1):13–40.