Modelling semantic acquisition in second language learning

Ekaterina Kochmar
The ALTA Institute
University of Cambridge
ek358@cam.ac.uk

Ekaterina Shutova
Computer Laboratory
University of Cambridge
es407@cam.ac.uk

Abstract

Using methods of statistical analysis, we investigate how semantic knowledge is acquired in English as a second language and evaluate the pace of development across a number of predicate types and content word combinations, as well as across the levels of language proficiency and native languages. Our exploratory study helps identify the most problematic areas for language learners with different backgrounds and at different stages of learning.

1 Introduction

Acquisition of semantic knowledge and vocabulary of a second language (L2), including appropriate word choice and awareness of selectional preference restrictions, are widely recognised as important aspects of L2 learning by native speakers, language teachers and learners themselves. Previous research demonstrated strong correlation between semantic knowledge and proficiency level (Shei and Pain, 2000; Alderson, 2005), and argued that the use of collocations makes one's speech more native-like (Kjellmer, 1991; Aston, 1995; Granger and Bestgen, 2014). James (1998) noted that learners often equate L2 mastery with mastery of L2 vocabulary, and Leacock et al. (2014) mention an experiment in which teachers of English ranked word choice errors among the most serious errors in L2 writing. At the same time, it has also been argued that acquisition of semantic knowledge proceeds on a word-by-word basis with each word being acquired as a separate construct (Gyllstad et al., 2015), and acquisition of content word combinations knowledge is slow and uneven, presenting challenges even at high proficiency levels (Bahns and Eldaw, 1993; Laufer and Waldman, 2011; Thewissen, 2013).

Native speakers are believed to be experts in their own language (James, 1998), and the language norm is usually set based on their preferences (Wulff and Gries, 2011). Apart from errors, learner English is often characterised by differences in the probabilistic distribution of lexical items which are expressed in under- or overuse of certain constructions (De Cock, 2004; Durrant and Schmitt, 2009; Laufer and Waldman, 2011; Wulff and Gries, 2011). In this paper, we adopt statistical approach and assume that native and learner language are characterised by different distributions. We investigate how non-native use of language develops and how closely it approximates native use at different levels of proficiency.

The native language distribution is modelled using a combination of the *British National Corpus* (*BNC*) and *ukWaC*, while learner language distributions are modelled using *Cambridge Learner Corpus* (*CLC*). CLC covers various L1 backgrounds as well as 6 language proficiency levels defined by the *Common European Framework of Reference for Languages* (*CEFR*) (Council of Europe, 2011a), ranging from "basic" (A1-A2) to "independent" (B1-B2) to "proficient" (C1-C2). In contrast to much of previous research, we run the experiments both on a wider scale, using a large corpus of learner English, and to finer level of granularity, exploring learner development across proficiency levels. Table 1 defines the amount and range of linguistic constructions that the learners are expected to be familiar with at different levels. Specifically, we explore:

(1) the pace of semantic knowledge and vocabulary acquisition across levels;

(2) the influence of one's L1 on the development of semantic knowledge;

(3) acquisition and development of selectional preference patterns across levels.

Level	Descriptor
A1	Has a very *basic repertoire of words* and *simple phrases* related to *personal details* and *particular concrete situations*.
A2	Uses basic sentence patterns with *memorised phrases*, *groups of a few words* and formulae in order to communicate limited information in *simple everyday situations*.
B1	Has enough language to get by, with *sufficient vocabulary* to express him/herself with some hesitation and circumlocutions on topics such as *family, hobbies and interests, work, travel, and current events*.
B2	Has a *sufficient range of language* to be able to give clear descriptions, express viewpoints on *most general topics*, without much conspicuous searching for words, using some complex sentence forms to do so.
C1	Has *a good command of a broad range of language* allowing him/her to select a formulation to express him/herself clearly in an appropriate style on *a wide range of general, academic, professional or leisure topics* without having to restrict what he/she wants to say.
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey *finer shades of meaning* precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a *good command of idiomatic expressions and colloquialisms*.

Table 1: CEFR descriptors of general linguistic and vocabulary range (Council of Europe, 2011b)

2 Previous research

Within NLP, it is more typical to explore learner language from the perspective of automated assessment or error detection and correction (Leacock et al., 2014) which focus on the contrast between learner and native language in terms of errors in L2, rather than from a language development perspective. The latter was studied more extensively by Second Language Acquisition (SLA) researchers. Previous research looked into vocabulary acquisition and language development assessing passive, or *receptive*, vocabulary knowledge (Gyllstad et al., 2015) and trying to estimate the vocabulary that the learners might understand at different proficiency levels (Nation, 2006; Bergsma and Yarowsky, 2013). The vocabulary size tests of the type proposed by Nation (2012) were shown to not be appropriate to test *productive* vocabulary knowledge as they suffer from overestimation of the vocabulary size (Gyllstad et al., 2015). Using learner writing to estimate the productive vocabulary size provides more reliable results, but previous studies in this area were performed on a smaller scale, either focusing on a limited number of proficiency levels (Gilquin and Granger, 2011; Granger and Bestgen, 2014), L1s (Gilquin and Granger, 2011; Granger and Bestgen, 2014; Siyanova-Chanturia, 2015), or on overall smaller datasets (Grant and Ginther, 2000; Granger and Bestgen, 2014).

It is widely accepted that vocabulary develops over time, and richer vocabulary is characteristic of better language knowledge (Laufer and Waldman, 1995; Grant and Ginther, 2000). Moreover, as students become more proficient writers, they do not only start operating with an overall larger vocabulary, but also become more precise in their word choice which is reflected in the increase of the type-token ratio (TTR) (Ferris, 1994; Engber, 1995; Frase et al., 1999; Grant and Ginther, 2000). However, the methodology of tagging the word choice and measuring TTR similar to that adopted in Grant and Ginther (2000) fails taking the *omissions* into account, while the method proposed in this paper helps alleviate this problem.

With respect to the development of selectional preference patterns and phraseological knowledge, Siyanova-Chanturia (2015) show that L2 learners even at lower levels do not just focus on single words acquisition but also attend to combinatorial linguistic mechanisms. The studies of Durrant and Schmitt (2009) and Granger and Bestgen (2014) suggest that intermediate learners tend to overuse high frequency collocations (such as *hard work*) and underuse lower-frequency collocations (such as *immortal souls*), while as proficiency in the language increases, this balance changes. Durrant and Schmitt (2009) argue that learners at the lower proficiency levels seem to over-rely on forms which are common in the language, and Paquot and Granger (2012) note that this might be related to the fact that learners feel confident using such common forms.

An interesting observation concerns the pace of semantic knowledge development: for instance, Laufer and Waldman (1995) observed that advanced learners' vocabulary is too varied to remain stable across different samples of writing. Laufer and Waldman (2011) and Nesselhauf (2005) investigated the development of collocational knowledge and came to a somewhat counterintuitive conclusion that more proficient learners produce more deviant collocations than their

less proficient counterparts. Thewissen (2008) argue that higher-level learners attempt a much wider range of lexical phrases which are not always error-free, and produce a large number of near-hits as compared to their lower intermediate counterparts. Paquot and Granger (2012) conclude that at an advanced level, learners take more risks, try out more complex lexical phrases and as a result, produce errors, but those are of a different, more 'advanced' nature than the basic errors typical of earlier stages.

A number of studies looked into L1 influence on L2 development (Siyanova-Chanturia, 2015; Paquot and Granger, 2012). Typically, researchers report negative effects of L1 transfer (Lorenz, 1999; Gilquin, 2007; Nesselhauf, 2005; Laufer and Waldman, 2011; Paquot and Granger, 2012), but some research also suggests that the learners whose L1 belongs to the same language family as English are more likely to make fewer mistakes than the learners from other L1 backgrounds (Waibel, 2008; Alejo Gonzalez, 2010; Gilquin and Granger, 2011).

3 Experimental setup

We focus on three types of content word combinations that are some of the most frequent in learner writing and have previously been found challenging for language learners (Lorenz, 1999; Paquot and Granger, 2012): adjective–noun (AN), verb–direct object (VO) and subject–verb (SV). We (1) investigate how the use of the predicating words (*adjectives* and *verbs*) within these combinations develops over time,[1] and (2) look into how their selectional preference patterns change across levels of language proficiency. We do not focus on *collocations* specifically for two reasons: firstly, there is a lot of disagreement in defining collocations (cf. Foster (2010), Nesselhauf (2005), Hoey (1991)), and secondly, learners have been shown to have difficulties with all types of content word combinations, including those that are referred to as 'free' (Paquot and Granger, 2012).

3.1 Data

Learner data: We have extracted the data for our experiments from the Cambridge Learner Corpus (CLC), which is a 52.5 million-word corpus of learner English collected by Cambridge University Press and Cambridge English Language Assessment (Nicholls, 2003). It comprises essays written during examinations in English by language learners with over 80 L1s and representing all 6 CEFR levels (Council of Europe, 2011a). Since the learners are not restricted in the word choice,[2] we believe that the range of vocabulary used in the essays is representative of what is in learners' active lexicon and, therefore, reflects semantic knowledge internalised at this point.

	Lvl	Types	Tokens	TTR	#Preds
AN	A1	7,053	41,502	0.1699	720
	A2	12,365	69,161	0.1788	1,010
	B1	37,198	179,791	0.2069	2,198
	B2	54,782	250,807	0.2184	2,699
	C1	59,965	250,263	0.2396	2,832
	C2	63,937	209,984	0.3045	3,664
VO	A1	9,690	58,399	0.1659	761
	A2	19,413	104,123	0.1864	1,238
	B1	45,826	217,100	0.2111	2,133
	B2	66,621	288,129	0.2312	2,499
	C1	67,235	247,842	0.2713	2,607
	C2	63,223	200,038	0.3161	2,764
SV	A1	7,553	40,657	0.1858	776
	A2	15,825	75,749	0.2089	1,323
	B1	49,282	187,378	0.2630	2,370
	B2	75,109	281,490	0.2668	2,867
	C1	83,832	293,654	0.2855	3,070
	C2	80,779	232,702	0.3471	3,283

Table 2: Overall statistics

We have extracted the word combinations from the full CLC parsed with the RASP (Briscoe et al., 2006). Table 2 summarises learner data: we include the number of *types* (unique combinations), *tokens* (overall number of combinations), *type-token ratio* (TTR) as well as the number of predicates for each level. Table 2 demonstrates that the overall number of the combinations and predicates as well as TTR constantly increase from A1 through to C2, with the largest increase between levels A2 and B1,[3] when the learners transfer from *beginners* to *intermediate* and start using the vocabulary beyond *basic* and *simple*, and between levels C1 and C2, when learners are expected to master *idiomatic expressions* and *colloquialisms*.

Native data: To estimate the general linguistic and vocabulary range of a native speaker, we have extracted the statistics on the use of ANs, VOs and SVs and the predicates from a combination of the BNC (Burnard, 2007) and ukWaC (Ferraresi et al.,

[1] We combine adjectives in AN and verbs in VO and SV combinations under the term of *predicating words* because we assume that they impose the selectional restrictions on the arguments (nouns) within the corresponding combinations.

[2] It can be argued that vocabulary selection is restricted by essay prompts; we address this issue in §5.

[3] The increase is statistically significant at 0.05 with t-test.

2008), which together amount to more than 2 billion words. For consistency, the native data has also been parsed with RASP (Briscoe et al., 2006).

3.2 Statistical methods

Distribution similarity: We measure the similarity between two distributions using Kullback-Leibler (KL) divergence (MacKay, 2003) which for distributions Q and P is defined as:

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (1)$$

In our experiments, P is the distribution in the learner data and Q is the distribution in the native data. The closer the two distributions are, the lower the value of D_{KL}. To support the results, we additionally measure the Pearson correlation coefficient (PCC) between the predicates and content word combinations in the learner and native data. PCC is higher for the more similar distributions.

Argument clustering: To address the issue of data sparsity, we estimate selectional preferences (*SP*) over *argument classes* as well as *individual arguments*. We obtain SP classes using spectral clustering of nouns with lexico-syntactic features, which has been shown effective in previous lexical classification tasks (Brew and Schulte im Walde, 2002; Sun and Korhonen, 2009). Spectral clustering partitions the data relying on a matrix that records similarities between all pairs of data points. We use *Jensen-Shannon divergence* to measure the similarity between feature vectors for nouns w_i and w_j as follows:

$$d_{\mathrm{JS}}(w_i, w_j) = \frac{1}{2} d_{\mathrm{KL}}(w_i||m) + \frac{1}{2} d_{\mathrm{KL}}(w_j||m), \quad (2)$$

where d_{KL} is the KL divergence, and m is the average of w_i and w_j. We construct the similarity matrix S computing similarities S_{ij} as $S_{ij} = \exp(-d_{\mathrm{JS}}(w_i, w_j))$. The matrix S encodes a similarity graph G over the nouns, where S_{ij} are the adjacency weights. The clustering problem can then be defined as identifying the optimal partition, or *cut*, of the graph into clusters, such that the intra-cluster weights are high and the inter-cluster weights are low. We cluster $2,000$ most frequent nouns in the BNC, using their grammatical relations as features. The features consist of verb lemmas occurring in the subject, direct object and indirect object relations with the given nouns in the RASP-parsed BNC. The feature vectors are constructed from the corpus counts and normalized by the sum of the feature values.

Selectional preference model: Once the SP classes are obtained, we quantify the strength of association between a given predicate and each of the classes. We adopt an information theoretic measure proposed by Resnik (1993) for this purpose. Resnik first measures *selectional preference strength* (*SPS*) of a predicate in terms of KL divergence between the distribution of noun classes occurring as arguments of the predicate, $p(c|v)$, and the prior distribution of the noun classes, $p(c)$:

$$\mathrm{SPS}_R(v) = \sum_c p(c|v) \log \frac{p(c|v)}{p(c)}, \quad (3)$$

where R is the grammatical relation for which SPs are computed. SPS measures how strongly the predicate constrains its arguments. Selectional association with a particular argument class is then defined as a relative contribution of that argument class to the overall SPS of the predicate:

$$\mathrm{Ass}_R(v, c) = \frac{1}{\mathrm{SPS}_R(v)} p(c|v) \log \frac{p(c|v)}{p(c)} \quad (4)$$

We extract VO and SV relations, map the argument heads to SP classes and quantify selectional association of a given predicate with each SP class.

4 Experimental results

We run a series of experiments to test the aspects of semantic knowledge acquisition outlined in §1.

4.1 Pace of semantic knowledge acquisition

Table 2 shows that at the lower levels learners operate with quite a small vocabulary. Many previous studies argued that learners at lower levels tend to overuse high frequency lexical items, whereas over time they expand their vocabulary with less frequent lexical items. It has also been argued that semantic knowledge acquisition is an unsteady process (see §2). First, we explore how exactly the semantic knowledge develops across proficiency levels, and investigate whether content word choice error rates – the proportion of word combinations where the predicate in chosen inappropriately as, for example, in **choose decision* instead of *make decision*, or **actual room* instead of *current room* – decrease over time.

For that, we identify 10 frequency bands for predicating words within each combination type using native English data. Each band covers from 363 (within band 1 of the most frequent predicates) up to $7,672$ (within band 10 of the least

frequent ones) unique adjectives in ANs, and similarly from 281 to 3,676 verbs in VOs, and 297 to 3,367 verbs in SVs. For instance, band 1 contains such adjectives as *big* and *good*, and verbs *give*, *go* and *see*, while band 10 contains adjectives *behaviouristic* and *decipherable*, and verbs *factor*, *garnish* and *mesmerise*. It is reasonable to expect that learners are familiar with the "simpler" words from band 1 even at the lower proficiency levels, while they might find words from band 10 much more challenging. In order to quantitatively assess this, we measure the proportion of new predicating words used at each level and map it to the identified frequency bands. Next, we estimate the error rates for each level and for each frequency band.

Figure 1 shows the distribution of the new vocabulary acquired at each level mapped against the frequency bands, as well as the distribution of the error rates across the frequency bands at each level.[4] While we observe that, as expected, learners expand their vocabulary acquiring words from lower frequency bands, the following trends are worth noting: most of the verb predicates in VOs and SVs that the learners know at level A1 are covered by frequency band 1. At A2 and B1 they still expand their vocabulary with some verbs from band 1, but starting with level B2 none of the new vocabulary comes from this band. Most new verbs in VOs at level C2 are covered by band 10, and in SVs by band 4. For adjectives, most new vocabulary at A1 and A2 comes from band 1, at B1 – band 3, at B2 – band 5, at C1 – band 8 and at C2 – band 10. Predictably, the error rates decrease towards the higher proficiency levels and within the higher frequency bands. The highest error rates are observed on the bands covering less frequent words: for example, even though the error rates are overall lower for C2 level, the highest error rate for C2 is associated with band 10 for all three types of combinations which confirms that semantic acquisition is challenging even at advanced levels.

While these results corroborate previous findings and show quantitatively how semantic knowledge develops across levels, we look further into how it approximates native English. In particular, it is reasonable to assume that the variety of English used by language learners at the lower proficiency levels is more dissimilar to the native English both for predicates and content word combinations, while it approximates native language distributions at upper levels. To test that, we calculate PCC and KL (see §3.2) and expect that towards C2 level PCC increases and approximates 1.0, while KL decreases and approximates 0.0.

	Lvl	PCC_{pred}	KL_{pred}	PCC_{comb}	KL_{comb}
AN	A1	0.3497	2.8737	0.1052	4.2909
	A2	0.4338	2.5073	0.1382	3.6463
	B1	0.7036	1.3101	0.2785	2.6212
	B2	0.7968	0.9408	0.4627	2.2058
	C1	0.8482	0.7959	0.4896	2.1183
	C2	**0.8188**	**0.7990**	**0.4817**	2.0451
VO	A1	0.6226	1.8469	0.0975	4.5220
	A2	0.7811	1.3115	0.1973	3.5465
	B1	0.8749	0.9080	0.3339	2.5350
	B2	0.9270	0.5965	0.5454	1.9129
	C1	0.9395	0.5541	0.6082	1.7994
	C2	**0.9262**	**0.6106**	**0.5736**	**1.8145**
SV	A1	0.9669	1.2729	0.1660	4.2648
	A2	0.9716	1.0038	0.2336	3.3381
	B1	0.9824	0.6898	0.4758	2.3194
	B2	0.9859	0.5623	0.6306	1.9506
	C1	0.9873	0.5141	0.6637	1.8733
	C2	**0.9870**	**0.5230**	**0.5954**	**1.9079**

Table 3: Predicates (*pred*) and combinations (*comb*) distributions

Table 3 presents the PCC and KL values for the distribution of the adjectives and verbs in columns marked with *pred* for predicating words, and for combinations in columns marked with *comb*. These values show that PCC steadily increases while KL steadily decreases from level A1 through to level C1, with the biggest "jump" between levels A2 and B1 for the adjectives and verbs in SVs, and A1 and A2 for the verbs in VOs. However, we note that at level C2 predicating words distribution is less similar to native English distribution than at level C1 for all types of combinations – we mark these values in the table in bold. We hypothesise that at level C2 the learners are already familiar with the basic vocabulary and start experimenting with the use of novel constructions which might result in a quite distinct variety of English (see Thewissen (2008) and Paquot and Granger (2012) for similar hypotheses). To investigate this further, we identify 10 predicates per combination type such that after removing them from the list of predicates, KL between the learner and native distribution improves (see Table 4).

What makes the use of these predicates by learners different from native use? Column "*#B*" in Table 4 presents the mean of the frequency bands and shows that most of these predicates come from the first two frequency bands, so they

[4] More detailed description is available at www.cl.cam.ac.uk/~ek358/vocab-acquisition.html.

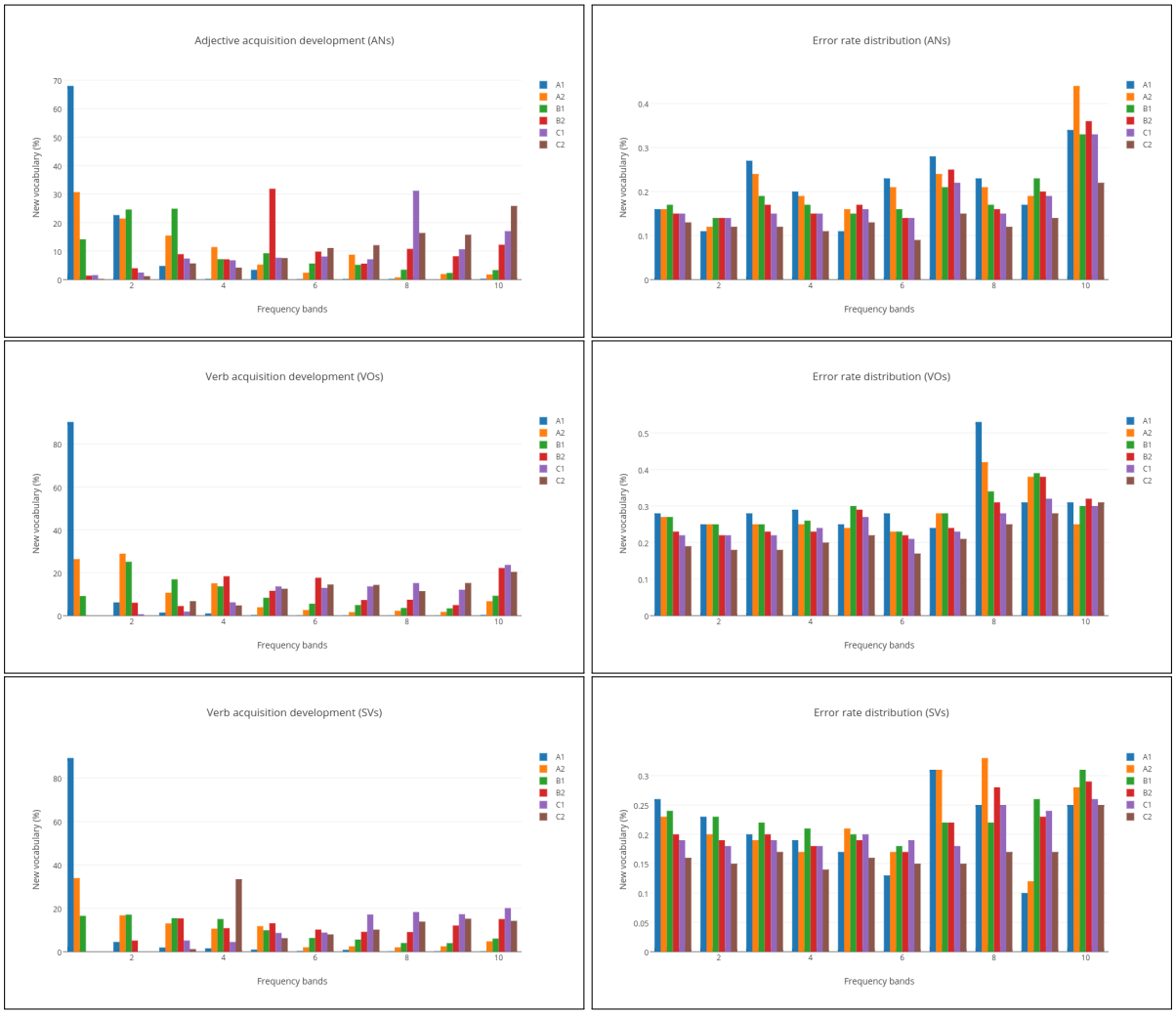

Figure 1: Predicates acquisition and error rate distribution across levels.

represent frequent words that are overused by the learners. We calculate the average error rates for the combinations with these predicates (column "*ErR*") and compare them to the average error rate over all predicates for each level (in parentheses). For adjectives and verbs in VOs the error rates are comparable or below the average error rate at the lower levels, and higher than the average at the upper levels. Verbs in SVs demonstrate an opposite trend: at the lower levels error rates associated with the use of these predicates are higher than average, while at the upper levels they are comparable or lower. We conclude that the differences in the distributions at the lower levels are caused by the overuse of the basic vocabulary, while towards the upper levels it is due to occasionally incorrect use of more diverse vocabulary.

The rightmost columns of Table 3 also compare the distribution of the ANs, VOs and SVs in the learner data to those in the native English data. We note that, similarly to the distribution of the predicates, the use of the content word combinations becomes more similar to native use towards higher levels of language proficiency, and to further confirm our hypothesis about the peculiar use of language at C2, we observe a disruption of this trend at C2 level for VOs and SVs. We also note that the development goes at quicker pace between A1 through to B2, and slows down at the upper levels.

4.2 L1 effects

L1 influence on the word choice has been extensively studied by SLA researchers (Siyanova-Chanturia, 2015; Paquot and Granger, 2012). It seems reasonable to expect that the similarity between one's L1 and L2 should facilitate semantic acquisition in L2: for example, if L1 and L2 belong to the same language group, they can be expected to bear considerable semantic similarities that might help learners acquire semantic knowl-

	Lvl	Predicates	#B	ErR
adj	A1	dear, mobile, favorite, other, national, blue, nice, pink, international, young	1.5	0.15 (0.16)
	A2	dear, mobile, favorite, local, national, nice, social, blue, pink, young	1.5	0.14 (0.16)
	B1	dear, best, nice, national, wealthy, beautiful, big, good, english, funny	1.4	0.15 (0.17)
	B2	dear, good, british, nice, wealthy, best, national, wonderful, important, big	1.3	0.15 (0.16)
	C1	dear, national, upward, wealthy, british, english, negative, bad, full, important	1.8	0.17 (0.15)
	C2	wealthy, national, dear, full, british, important, further, current, serial, european	1.4	0.15 (0.13)
v $_{VO}$	A1	buy, paint, like, watch, go, wear, bring, play, provide, make	1.1	0.28 (0.28)
	A2	buy, paint, provide, like, go, watch, attend, wear, book, confirm	1.2	0.26 (0.27)
	B1	buy, watch, include, provide, like, go, spend, offer, film, love	1.2	0.22 (0.27)
	B2	include, provide, spend, rent, support, contain, follow, raise, create, cover	1.1	0.26 (0.24)
	C1	include, excel, improve, concern, provide, solve, show, reach, spend, allow	1.4	0.21 (0.22)
	C2	spend, broaden, offer, earn, solve, allow, require, cover, use, enable	1.3	0.21 (0.19)
v $_{SV}$	A1	cost, make, use, park, have, show, find, say, wish, take	1.1	0.31 (0.25)
	A2	cost, use, include, make, provide, park, say, attend, find, show	1.1	0.27 (0.22)
	B1	include, like, wish, watch, provide, require, spend, set, decrease, amaze	1.4	0.22 (0.23)
	B2	increase, like, include, reward, decrease, interest, spend, provide, require, involve	1.4	0.20 (0.20)
	C1	increase, decrease, include, spend, like, change, say, show, improve, apply	1.2	0.20 (0.19)
	C2	include, increase, spend, frame, live, like, require, provide, set, base	1.3	0.16 (0.16)

Table 4: Top 10 predicates contributing to the difference between learner and native language distribution

edge in L2, while one may expect to observe slower learning pace for speakers of more distant L1s (Gilquin and Granger, 2011).

To test to what extent L1 exerts influence on L2 semantic knowledge acquisition, we consider three language groups – Germanic L1s (GE) that belong to the same group as English (EN), Romance L1s (RM) that represent a different group within the same family of the Indo-European languages, and Asian L1s (AS) representing a group of languages most distant from English among the three.[5] We measure KL divergence for the three pairs, GE–EN, RM–EN and AS–EN, on the distribution of the predicates.

The results reported in Table 5 contradict our original assumption as we observe that the variety of English used by speakers of Romance L1s is closer to native English than the variety used by speakers of Germanic L1s. Furthermore, the variety of English used by speakers of Asian L1s, especially at the lower levels, is more similar to native English than the variety used by Germanic L1

[5] GE include Danish, Dutch, German, Norwegian and Swedish; RM include French, Italian, Portuguese, Romanian and Spanish; AS include Thai, Vietnamese and different varieties of Chinese.

	Lvl	GE	RM	AS
adj	A1	4.3318	3.5133	3.8219
	A2	3.3723	3.2955	3.2837
	B1	2.3309	2.3874	1.7002
	B2	1.4971	1.4109	1.3849
	C1	1.1840	1.1088	1.2562
	C2	**1.2880**	1.0543	**1.3716**
v_{VO}	A1	2.1994	2.0347	2.0446
	A2	1.6371	1.6478	1.6204
	B1	1.3751	1.2139	0.9772
	B2	0.9280	0.7363	0.8622
	C1	**0.9389**	0.7050	0.8164
	C2	**0.9806**	**0.7512**	**0.9465**
v_{SV}	A1	2.2841	1.3059	1.3300
	A2	1.6275	1.1930	1.2918
	B1	1.1583	0.9629	0.8604
	B2	0.8576	0.6862	**0.8636**
	C1	**0.8631**	0.6326	0.8158
	C2	**0.8818**	**0.7098**	**0.9283**

Table 5: Predicate distributions per language groups (*KL*)

speakers. We hypothesise that since Asian L1s are very different from English, the speakers of these languages may prefer to use prefabricated phrases more often than speakers of Germanic L1s, which makes their language more native-like. Similar hypotheses have been formulated earlier: for example, Gilquin and Granger (2011) noted that learners, especially at the lower levels, are likely to repeat expressions that are familiar to them and appear to be safe, and Hulstijn and Marchena (1989) noted that learners tend to rely on "play-it-safe" strategy rather than experiment unless they are confident in their vocabulary knowledge. We assume that speakers of Germanic L1s might feel more confident in their semantic knowledge and as a result be more "adventurous" in their use of English than speakers of Asian L1s. Our experiments on the individual L1s within each group show same trends as observed for L1 groups.

4.3 Selectional preference patterns

Finally, we investigate how selectional preference patterns develop across proficiency levels and whether they approximate native English patterns. For each predicate in learner and native data, we form argument clusters using the methodology described in §3.2, estimate SP strength for the predicates at each level using eq. 3, and then apply KL divergence and PCC to measure the difference.

Table 6 overviews the similarity between the SP models in learner and native data for the arguments and argument clusters (see columns with *cl*). As before, we observe that the SP models in

	Lvl	PCC	KL	PCC_{cl}	KL_{cl}
AN	A1	0.1661	0.1481	0.2835	0.1980
	A2	0.4375	0.0843	0.5449	0.1149
	B1	0.5808	0.0494	0.5597	0.0897
	B2	0.6133	0.0395	0.5940	0.0765
	C1	0.6526	0.0372	0.6408	0.0729
	C2	**0.6428**	0.0364	**0.5866**	**0.0762**
VO	A1	0.4959	0.0966	0.5976	0.1533
	A2	**0.3893**	0.0917	**0.5414**	0.1430
	B1	0.6181	0.0579	0.7429	0.0810
	B2	0.6759	0.0412	**0.6987**	0.0749
	C1	0.7172	0.0354	0.7576	0.0634
	C2	**0.7168**	**0.0379**	0.7609	**0.0645**
SV	A1	0.6069	0.1254	0.4475	0.1722
	A2	**0.6061**	0.0956	0.4934	0.1604
	B1	**0.6053**	0.0837	0.5008	0.1538
	B2	0.6500	0.0612	**0.4248**	0.1515
	C1	0.6539	0.0553	0.4972	0.1306
	C2	0.6599	**0.0595**	0.5164	**0.1418**

Table 6: Selectional preference distribution

	Predicate	Learner language	Native language
AN	additional kind	worker, teacher, staff girl, woman, person	information, item, detail consent, permission, approval
VO	reserve stipulate	bathroom, hall, room price, rent, salary	privilege, right, status rule, need, norm
SV	bind reflect	treaty, contract, deal gear, clothes, mask	gene, tissue, cell rise, change, improvement

Table 7: Examples of the most strongly associated arguments

learner data become more similar to those in native language towards upper levels. Both ANs and VOs show the biggest improvements between A2 and B1, and we observe the disruption in this trend at the levels A2 and C2 (we mark those in bold).

Next, we look into the set of predicates that have the most different SP patterns in the learner and native language, and using eq. 4, identify the argument cluster that is most strongly associated with each of these predicates in the learner and native data. For the sake of space, in Table 7 we present only some illustrative examples from different levels and combination types.[6] The experiments suggest that the difference between the learner and native SP models might be due to the learners' use of concrete nouns with the adjectives and verbs where native speakers prefer abstract nouns.

To further investigate this hypothesis, we identify 10 predicates per combination type and proficiency level with the most distinct selectional preference patterns. Using the MRC Psycholinguistic Database (Wilson, 1988), we calculate the average concreteness score for the arguments clusters in learner and native data. Our results show that at the lower levels learners use more concrete arguments than native speakers, with the difference statistically significant at 0.05 with t-test, while the difference becomes less pronounced towards C1-C2 levels. Our results for productive vocabulary knowledge corroborate previous findings on the relation between receptive vocabulary knowledge and acquisition of abstract concepts (Tanaka et al., 2013; Vajjala and Meurers, 2014).

The results show that the difference in selectional preference patterns between the learner and native language is due to the concreteness of the selected arguments. This may reflect (1) the difficulty in acquiring semantics of abstract concepts in L2, or, alternatively, (2) L1-based instructional practices that may focus first on teaching concrete concepts before abstract concepts. The awareness of this discrepancy can serve as further guidance for language instructors and learners, and help make one's language use more native-like.

5 Discussion and conclusions

This paper reports the results of a large-scale corpus-based exploratory study of semantic knowledge acquisition by L2 learners. In contrast to previous work, we ran experiments on a wider scale, using a large learner corpus, and at finer granularity, investigating L2 development across 6 CEFR proficiency levels. We show that (1) the learners tend to overuse highly frequent English words across all proficiency levels, although towards the higher levels the lexical distributions in learner and in native language become more similar; (2) the two peaks of vocabulary acquisition are associated with the transition between beginner and intermediate levels (A2-B1), and between the two proficient levels (C1-C2); (3) lexical distribution at upper proficient level (C2) is less similar to native distribution than at lower proficient level (C1) which may be due to the more creative language use at C2; (4) the variety of English used by speakers of more distant L1s at lower levels of proficiency is closer to native English than the variety used by speakers of closer L1s, which might be an effect of "play-it-safe" strategy adopted by learners; (5) concrete nouns tend to be more strongly associated with the predicates in learner language than abstract nouns. The methodology presented in this paper can help identify the gaps in learner vocabulary knowledge and tailor vocabulary acquisition exercises to the needs of learners at dif-

[6] Full lists are available at www.cl.cam.ac.uk/~ek358/vocab-acquisition.html.

ferent proficiency levels.

We admit that potential topic and genre bias of learner exams data is a limitation of our corpus-based approach. We believe that corpus-based studies of the type presented in this paper will facilitate further research into semantic knowledge development, although it is possible that learner corpora provide only limited access to productive learner vocabulary. As Siyanova-Chanturia (2015) notes "in an ideal world, one would use the same topic across and within all tested levels, but in a language classroom, this is hardly possible". The future work will investigate possible solutions for this problem such as (1) augmentation of the data with other learner corpora, (2) use of fill-in-the-gaps exercises that test vocabulary knowledge directly, and (3) sampling of the native data to more closely reflect the selection of topics in the learner data.

Acknowledgments

We are grateful to the BEA reviewers for their helpful and instructive feedback. Ekaterina Kochmar's research is supported by Cambridge English Language Assessment via the ALTA Institute. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

References

J. C. Alderson, editor. 2005. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London; New York: Continuum.

R. A. Alejo Gonzalez. 2010. L2 Spanish acquisition of English phrasal verbs: a cognitive linguistic analysis of L1 influence. In *M. C. Campoy-Cubillo, B. Belles-Fortuno, & M. L. Gea-Valor (eds.), Corpus-based approaches to English language teaching*, London, UK: Continuum, pages 149–166.

G. Aston. 1995. *Corpora in language pedagogy: Matching theory and practice. In G. Cook & B. Seidlhofer (eds.), Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*, Oxford: Oxford University Press, pages 257–270.

J. Bahns and M. Eldaw. 1993. Should we teach EFL students collocations? *System* 21:101–114.

S. Bergsma and D. Yarowsky. 2013. Learning Domain-Specific, L1-Specific Measures of Word Readability. *TAL* 54(1):203–226.

C. Brew and S. Schulte im Walde. 2002. Spectral clustering for German verbs. In *Proceedings of EMNLP*. pages 117–124.

E. Briscoe, J. Carroll, and R. Watson. 2006. The Second Release of the RASP System. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006) Interactive Presentation Sessions*. pages 59–68.

L. Burnard, editor. 2007. *Reference Guide for the British National Corpus*. Research Technologies Service at Oxford University Computing Services.

The Council of Europe. 2011a. Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

The Council of Europe. 2011b. Forms for detailed analysis of examinations or tests.

S. De Cock. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literature (BELL), New Series* 2:225–246.

P. Durrant and N. Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics* 47:157–177.

C. Engber. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4:139–155.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*. pages 47–54.

D. Ferris. 1994. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly* 28:414–420.

P. Foster. 2010. Rules and routines: a consideration of their role in the task-based language production of native and non-native speakers. In *M. Bygate, P. Skehan, & M. Swain (eds.), Researching pedagogic tasks: Second language learning, teaching and testing*, Harlow, UK: Longman, pages 75–93.

L. Frase, J. Faletti, A. Ginther, and L. Grant. 1999. Computer Analysis of the TOEFL Test of Written English. Technical report, Princeton, NJ: Educational Testing Service.

G. Gilquin. 2007. To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift fur Anglistik und Amerikanistik* 55:273–291.

G. Gilquin and S. Granger. 2011. From EFL to ESL: Evidence from the International Corpus of Learner English. In *Mukherjee J., Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, John Benjamins Publishing Company: Amsterdam and Philadelphia, pages 55–78.

S. Granger and Y. Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching (IRAL)* 52:229–252.

L. Grant and A. Ginther. 2000. Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing* 9(2):123–145.

H. Gyllstad, L. Vilkait/e, and N. Schmitt. 2015. Assessing vocabulary size through multiple-choice formats: issues with guessing and sampling rates. *International Journal of Applied Linguistics (ITL)* 166(2):278–306.

M. Hoey. 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.

J. Hulstijn and E. Marchena. 1989. Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition* 11(3):241–255.

C. James. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. London: Longman.

G. Kjellmer. 1991. *A mint of phrases. In K. Aijmer & B. Altenberg (eds.), English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Harlow, Essex: Longman, pages 111–127.

B. Laufer and T. Waldman. 1995. Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics* 16(3):307–322.

B. Laufer and T. Waldman. 2011. Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning* 61:647–672.

C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers, second edition.

G. Lorenz. 1999. *Adjective intensification e Learners versus native speakers. A corpus study of argumentative writing*. Rodopi, Amsterdam.

D. J. C. MacKay. 2003. *Information Theory, Inference, and Learning Algorithms (First ed.)*. Cambridge University Press.

I. S. P. Nation. 2006. How Large a Vocabulary Is Needed For Reading and Listening? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes* 63(1):59–82.

I. S. P. Nation. 2012. Vocabulary Size Test Instructions and Description. https://www.victoria.ac.nz/lals/about/staff/paul-nation.

N. Nesselhauf. 2005. *Collocations in a learner corpus*. John Benjamins, Amsterdam.

D. Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*. pages 572–581.

M. Paquot and S. Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32:130–149.

P. Resnik. 1993. Selection and Information: A Class-based Approach to Lexical Relationships. Technical report, University of Pennsylvania.

C. C. Shei and H. Pain. 2000. An ESL Writer's Collocation Aid. *Computer Assisted Language Learning* 13(2):167–182.

A. Siyanova-Chanturia. 2015. Collocation in beginner learner writing: A longitudinal study. *System* 53:148–160.

L. Sun and A. Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proceedings of EMNLP*. pages 638–647.

S. Tanaka, A. Jatowt, M. P. Kato, and K. Tanaka. 2013. Estimating Content Concreteness for Finding Comprehensible Documents. In *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM'13)*. pages 475–484.

J. Thewissen. 2008. The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. In *Proceedings from the 8th Teaching and Language Corpora Conference (TaLC8)*. pages 300–306.

J. Thewissen. 2013. Capturing L2 Development Through Learner Corpus Analysis. *Modern Language Journal (special issue on capturing the dynamics of L2 development through learner corpus analysis)* 97:77–101.

S. Vajjala and D. Meurers. 2014. Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification* pages 1–23.

B. Waibel. 2008. *Phrasal verbs: German and Italian learners of English compared*. VDM, Saarbrucken, Germany.

M. D. Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioral Research Methods, Instruments and Computers* 20:6–11.

S. Wulff and S. T. Gries. 2011. Corpus-driven methods for assessing accuracy in learner production. In *Second Language Task Complexity: Researching the Cognition Hypothesis of language learning and performance*, Peter Robinson (eds.): John Benjamins Publishing, pages 61–87.