Sylviane Granger* and Yves Bestgen

The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study

Abstract: Phraseological competence, the use of (semi-)prefabricated expressions in language, is a major component of second language acquisition. Recent research focused on lexical bundles, i.e. recurrent contiguous strings of words, has highlighted quantitative and qualitative differences between native and non-native speaker use of these strings. Few studies, however, have investigated the development of phraseological competence as a function of degree of proficiency in L2. Relying on a methodology used by Durrant and Schmitt (2009: *IRAL 47, 157–177*) to compare native and non-native speakers, the present study identifies significant differences in the way in which intermediate and advanced learners use collocations. In particular, the intermediate learners tend to overuse high frequency collocations (such as *hard work*) and underuse lower-frequency, but strongly associated, collocations (such as *immortal souls*). The concluding section addresses the limits of the study and points to possible applications in foreign language teaching and automated scoring.

Keywords: phraseology, collocation, bigram, L2 learner corpus, proficiency level, association scores

DOI 10.1515/iral-2014-0011

1 Introduction

Phraseology is one of the aspects that unmistakably distinguishes native speakers of a language from L2 learners, even those that have attained a high level of proficiency. The role played by prefabricated patterns in nativelike fluency has been most convincingly argued by Pawley and Syder in their seminal (1983) article: “Coming ready-made, the memorized sequences need little encoding work.

***Corresponding author: Sylviane Granger:** Université catholique de Louvain.
E-mail: sylviane.granger@uclouvain.be
Yves Bestgen: Research Associate of the F.R.S-FNRS. E-mail: yves.bestgen@uclouvain.be

Freed from the task of composing such sequences word-by-word, so to speak, the speaker can channel his energies into other activities" (p. 208). L2 learners, on the other hand, tend to compose their utterances or sentences word by word, relying mainly on what Sinclair (1991: 109) called the "open-choice principle". While studies focusing on learners' phraseological difficulties are not new, the emergence of native and learner corpora and the design of sophisticated techniques to extract phraseological patterns has resulted in an explosion of studies in the field in recent years. While some studies still rely on the traditional linguistic criteria of semantic non-compositionality and syntactic fixedness to identify phraseological units, a lively strand of research now takes as a starting point multiword sequences identified on the basis of purely quantitative criteria. Among the techniques used, the extraction of n-grams, i.e. recurrent sequences of contiguous words (2 words, 3 words, etc.), is growing increasingly popular (De Cock 2000, 2007; Allen 2009; Groom 2009; Ishikawa 2009; Juknevičienė 2009; Ping 2009; Chen and Baker 2010; Götz and Schilk 2011). Most of the studies using this technique focus on the overall frequency of n-grams in a corpus of learner writing and/or speech and compare it to a comparable native corpus, thereby uncovering interesting patterns of over- and underuse and misuse (for a survey see Paquot and Granger 2012).

In a recent study, Durrant and Schmitt (2009) have used a new approach which differs in two major respects from most learner-corpus-based studies to date. First, their study takes into account the variability inherent in learner language by considering the learner corpus as a series of individual texts rather than one long text. Second, they analyse word combinations on the basis of their association scores in a large native reference corpus using two different statistical measures: Mutual Information (MI), which tends to highlight word combinations made up of low frequency words (such as *immortal souls* or *tectonic plates*) and t-score, which brings out those composed of high frequency words (such as *good example* or *hard work*). They show that learners tend to overuse collocations with high t-scores and underuse those with high MI scores and conclude: "Advanced non-native phraseology differs from that of natives not because it avoids formulaic language altogether but because it overuses high-frequency collocations and underuses the lower-frequency, but strongly-associated, pairs characterised by high mutual information scores" (2009: 175).

Like most studies of the L2 phrasicon, Durrant and Schmitt's study relies on learner data that is not stratified for proficiency. Although their suggestion that "learners are quick to pick up highly frequent collocations, but less common, strongly associated items (e.g. *densely populated, bated breath, preconceived notions*) take longer to acquire" (2009: 174) seems quite reasonable in view of the results of the native vs. non-native comparison, it can only be verified on the basis

of proficiency-stratified learner corpora. The objective of this article is to check Durrant and Schmitt's assumptions by comparing learner corpora representing two different proficiency levels: intermediate and advanced. We wish to contribute to the ongoing research (Green 2008; Galaczi and Khalifa 2009; Thewissen 2013) aimed at discriminating effectively between learners at different proficiency levels, with a particular focus on the highest levels which have been described in less detail in the literature.

Based on the results of Durrant and Schmitt's study we expect to find that:

- the proportion of lower-frequency, but strongly-associated, collocations [attested by MI] should be smaller in the intermediate learner texts than in the advanced learner texts;
- the proportion of high-frequency collocations [attested by t-score] should be larger in the intermediate learner texts than in the advanced learner texts.

In Section 2 we provide a description of the corpus data and the methodology used to extract and categorize the bigrams. Section 3 presents the results of the quantitative comparison of intermediate and advanced learner data while Section 4 takes a more qualitative look at the types of units that distinguish between the two groups of learners. Section 5 sums up the results and makes suggestions for future research in the field.

2 Methodology

2.1 Corpus of learner texts

Our study is based on a sample of 223 texts written by learners of English as a Foreign Language (EFL). The texts are part of a large computer learner corpus, the *International Corpus of Learner English (ICLE)*, which contains 3.7 million words of essays written by intermediate to advanced learners from 16 mother tongue backgrounds (Granger et al. 2009). The sample was extracted from three *ICLE* subcorpora, i.e. 74 essays were taken from the French (FR) component, 71 from the German (GE) component and 78 from the Spanish (SP) component of the learner corpus. The texts display a high degree of homogeneity: they represent the same text type (argumentative essays) and are of similar length (500–900 words).

Each of the 223 learner essays was assigned a grade according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) by two professional raters. The raters were asked to give a holistic score for

Table 1: Holistic CEFR scores: 11-point numerical scale

Holistic CEFR score	B1–	B1	B1+	B2–	B2	B2+
Numerical value	0.67	1	1.33	1.67	2	2.33
Holistic CEFR score	C1–	C1	C1+	C2–	C2	
Numerical value	2.67	3	3.33	3.67	4	

each text as well as a specific grade for each of the five following linguistic competences: vocabulary accuracy, grammatical accuracy, orthographic control, vocabulary range and coherence/cohesion.[1] They were given clear rating guidelines and carried out the rating procedure completely independently, i.e. they never met to discuss results. The essays were presented to the raters in a random order (the same random order for both raters).

The scores the raters could choose from ranged from the lower intermediate to the most advanced CEFR levels, i.e. B1, B2, C1 or C2[2]. The scores thus obtained were recorded using an 11-point numerical scale (i.e. B1– = 0.67, B1 = 1, B1+ = 1.33 etc until C2 = 4, cf. Table 1) so as to calculate the degree of correlation between the grades given by the two raters. The correlation between both raters on this 11-point scale reached $r = 0.69$. Overall, both raters completely agreed on a total of 102 texts (46%); they reached near agreement (i.e. a maximal difference of one band score as in B2–C1 or C1–C2) on 87 texts (39%) and disagreed by more than one band score (e.g. B1–C1) on 34 texts (15%). The 34 texts on which the two raters substantially disagreed were submitted to a third rater. We then gave a numerical value to each holistic score (cf. Table 1), computed the mean of these scores and re-interpreted them in terms of CEFR level. The cut-off between the intermediate (B) and advanced (C) level was set at 2.5, scores lower or equal to the cut-off being classified as B. The details of the two CEFR-graded corpora are provided in Table 2.

1 We have only used the holistic scores for our study as the individual linguistic competences are highly correlated (the smallest Pearson's *r* between two scales is larger than 0.915) and display strong correlation with the holistic scores (the smallest *r* between an individual scale and the holistic score is larger than 0.95).

2 The CEF includes a total of six proficiency levels which can be broken down into three groups of two, viz. A1 and A2 (= basic users; elementary proficiency learners), B1 and B2 (= independent users; intermediate proficiency learners), C1 and C2 (= proficient users; advanced proficiency learners). We started the rating procedure at level B1 because the CEF specifies that learners need a B1 proficiency level to attempt essay writing. It was made possible for raters to use + or – signs to further specify quality within each proficiency level.

Table 2: Details of the two CEFR-graded corpora

	B	C
Number of texts	128	95
Total number of words	84,828	66,620
Mean number of words per text	662.7	701.3
Standard deviation	91.9	95.1
Minimum	504	502
Maximum	888	890

2.2 Bigram extraction and categorization

Like Durrant and Schmitt we focus on bigrams, i.e. directly adjacent word pairs. However, we have used a different extraction method. Durrant and Schmitt extracted word pairs manually from a raw corpus, while we extracted them automatically from a part-of-speech (POS) tagged corpus. We used CLAWS 7[3], which has a high degree of accuracy overall (96–97%) and has proved to perform better than other POS-taggers when handling learner data (Van Rooy and Schäfer 2003). Spelling errors were removed from the texts before POS-tagging not only to improve the accuracy of the tagger but also to be able to group instances of the same word pair that differed only in one or two minor spelling errors (e.g. *private correspondance* and *private correspondence* were counted as two occurrences of the bigram *private correspondence*).[4] Likewise we decided to group contracted forms with their corresponding full forms (*I've* > *I have; he's* > *he is*). Ambiguous contracted forms were easily disambiguated thanks to their POS tags, e.g. enclitic *'s* is tagged VBZ (*is*), VHZ (*has*) or POS (genitive).

All bigrams were then extracted from each learner essay. The search algorithm rejected bigrams that were interrupted by punctuation marks or any sequence of characters that does not correspond to a word. For example, the sequence *down_this* in the excerpt *"The Wall is Down!!!" This headline . . .* was not extracted.

3 http://ucrel.lancs.ac.uk/claws/

4 We only normalized words when there was no doubt about the targeted form. This involves changes such as erroneous doubling of consonants (*appartment*), omission of a letter (*completly*) or addition of a letter (*ridicoulous*). No attempt was made to normalize words like *documentals* which could in principle stand for *document*, *documentation* or *documentary*.

To ensure comparability we extracted word pairs that displayed the same syntactic configuration as those selected by Durrant and Schmitt (2009), i.e. Premodifier Noun sequences (MN), which are made up of a noun premodified by another noun (NN) or an adjective (JN). However, we also analysed the two subcategories separately as they are phraseologically different: NN sequences are likely to consist largely of compounds like *ozone layer* or *traffic jam* while JN sequences will include both compounds (*instant coffee, prime minister*) and restricted collocations (*warm welcome, wide range*), a difference which may have an impact on learner use. We also added bigrams made up of an adjective premodified by an adverb (AJ) (*very good, incurably ill*), a category which has been the subject of extensive research in learner corpus studies (Granger 1998; Lorenz 1998; Lee 2006; Wei and Lei 2011).

In addition to these four categories of syntactically-defined word pairs (MN, NN, JN, AJ), we also investigated the role played by all bigrams (henceforth referred to as All)[5]. This category provides a comprehensive picture of the learner phrasicon, from the most frequent bigrams like *of the* and *it is* to the most sophisticated ones like *vested interests* or *conscientious objectors*. It has the added advantage of being very easy to extract. If the All category proved to correlate as efficiently with proficiency level as the POS-based categories, it would be an ideal candidate in the framework of applications like automated scoring.

By way of illustration we provide in Table 3 the top 20 sequences in each category (except MN which subsumes NN and JN) in the whole learner corpus (intermediate and advanced) and their frequency of occurrence.

2.3 Measures of collocational strength and statistical test

Rather than basing the analysis solely on the raw frequency of bigrams, we decided to follow Durrant and Schmitt (2009) and use methods that measure the collocational strength holding between the two words in the bigram[6]. Each word sequence extracted from the learner corpus was searched for in the *British National Corpus (BNC),* a 100 million word collection of samples of written and spoken language designed to represent a wide cross-section of British English from the later part of the 20th century[7]. The word sequences that reached the

5 The bigrams which were made up of two proper nouns were excluded. Thus, *John Huston* was excluded but *European country* was kept.

6 The only difference with Durrant and Schmitt is that they computed association scores on the basis of sequences of words while we based our analysis on sequences of word + POS-tag.

7 http://www.natcorp.ox.ac.uk/corpus/

Table 3: Top 20 word combinations in the whole learner corpus

	NN		JN		AJ		All	
1	university_degrees	30	military_service	80	most_important	43	of_the	1027
2	university_degree	19	other_hand	65	very_important	24	in_the	702
3	prison_system	15	human_beings	55	very_difficult	19	it_is	647
4	capital_punishment	14	human_being	28	more_equal	14	to_the	414
5	death_penalty	14	other_people	26	more_important	14	to_be	407
6	world_war	11	everyday_life	25	more_practical	10	on_the	315
7	ozone_layer	9	national_identity	23	very_good	10	do_not	312
8	television_set	9	european_ countries	22	more_difficult	9	is_not	270
9	university_studies	8	long_time	22	very_common	8	is_the	261
10	public_opinion	7	real_world	22	very_different	8	have_to	252
11	tv_channels	7	european_ community	20	very_dangerous	7	of_a	248
12	speed_limit	6	modern_world	20	very_easy	7	in_a	238
13	animal_farm	5	other_words	18	even_worse	6	and_the	233
14	death_sentence	5	professional_ army	17	more_powerful	6	is_a	227
15	leisure_time	5	fast_food	15	so_important	6	that_the	217
16	life_imprisonment	5	only_children	15	very_expensive	6	for_the	215
17	science_technology	5	other_countries	15	very_hard	6	as_a	204
18	soap_opera	5	european_union	14	completely_ different	5	they_are	203
19	soap_operas	5	different_ countries	13	more_careful	5	there_is	184
20	acid_rain	4	human_life	13	more_ comfortable	5	can_not	183

frequency threshold of 5 occurrences in the *BNC* were assigned two association scores: MI score (also called 'pointwise mutual information') and t-score. The sequences that had less than 5 occurrences were categorized as below threshold (BT), because association measures computed for bigrams below this threshold are thought to be unreliable (Durrant and Schmitt 2009: 168). The two

Table 4: Thresholds used for the association measures

Categories of bigrams	MI	t-score
Non-collocational	<3	<2
Collocational: Low	≥3 and <5	≥2 and <6
Collocational: Medium	≥5 and <7	≥6 and <10
Collocational: High	≥7	≥10

association scores were computed by means of the formulas reported in Evert (2009: 1225).

Based on their collocational strength the bigrams were then divided into a number of categories. Durrant and Schmidt propose 7 bands of t-score and 7 bands of MI, but systematically collapse the bands into broader 'high' vs. 'low' groupings for the statistical tests. We opted for a happy medium between these two solutions and separated the collocates into 4 categories: one category of non-collocational (NC) bigrams and three categories of collocational bigrams of varying association strengths: low (L), medium (M) and high (H). The minimum values for collocational status (2 for t-score and 3 for MI) are standard in the bigram literature (e.g. Stubbs 1995; Hunston 2002; Durrant and Schmitt 2009). Table 4 gives the thresholds used for each category.

The raw data on which our study is based were obtained as follows. For each learner text, the bigrams in each of the five categories (MN, NN, JN, AJ and All) are first divided into two groups: those that appear at least 5 times in the reference corpus and the others. The BT score is the percentage of the first group of bigrams out of the total number of bigrams of that category in a given text. The remaining bigrams, i.e. those that reach the frequency threshold of 5 occurrences, are distributed among the 4 MI bands in function of their association value in the *BNC*. Here too the resulting score represents a percentage out of the total number of bigrams of that category in a given text. The same procedure is used for the 4 t-score bands. By way of illustration we provide in Table 5

Table 5: Raw frequencies and percentages for the bigrams of the All category in a learner's text

	BT	MI				t-score			
	BT	NC	L	M	H	NC	L	M	H
Raw frequency	116	346	148	57	32	64	93	47	379
Percentage	16.60	49.50	21.17	8.15	4.58	9.16	13.30	6.72	54.22

the data submitted to the statistical analyses for one particular category (All) in one learner text containing 599 bigrams. The first line contains the raw data and the second line the percentages that will be used in our study. As clearly appears from the table, the percentages for the four MI bands and the BT set sum to 100% as do the percentages for the four t-score bands and the BT set.

To determine whether the average percentages in each band (NC, L, M and H) for each category of collocations (NN, JN, etc) are different depending on CEFR grade (B or C), we used the Student's t test for independent samples. For each test, an indication is given as to whether the effect of the CEFR grade factor is statistically significant for three conventional thresholds ($p \leq .05$, $p \leq .01$ and $p \leq .001$). A measure of effect size, Cohen's *d*, is also provided. This expresses the difference between the means of the two groups in standard deviation unit. Following Cohen (1988), we consider that a *d* of about 0.20 indicates a small effect, one of about 0.50 a medium effect, and one of about 0.80 indicates a large effect. In this study, we did not expect very large effect sizes because it can be assumed that many factors other than the learners' phraseological competence determine the quality of the essay, factors such as the presence of spelling errors or the syntactic complexity for example. In addition, the classification of the bigrams in the bands is necessarily imperfect because of the limits inherent to any reference corpus. For example, some bigrams classified as High by analyzing the *BNC* may qualify as Medium based on another reference corpus. Similarly, some bigrams that are below threshold in the *BNC* may well have enough occurrences in a different reference corpus to justify their inclusion in the analysis.[8]

3 Intermediate and advanced learners: quantitative approach

Table 6 gives the total number of bigrams[9] for each category in the two CEFR-graded corpora (B and C) and the average frequency per text. As the B-level

8 It would be interesting to replicate the study with a much larger and more recent corpus, such as the Corpus of Contemporary American English (COCA), a very large and balanced corpus of American English (cf. http://corpus.byu.edu/coca/)

9 As explained in Section 2.2, we have only extracted bigrams made up of two contiguous words. This explains that for the category All, we only analysed 890 bigrams for 1,000 words.

Table 6: Number of bigrams in B and C corpora: breakdown per category

	B			C		
	Total number of extracted bigrams	**Average number of bigrams per text**	**Average number of bigrams per 1000 words**	**Total number of extracted bigrams**	**Average number of bigrams per text**	**Average number of bigrams per 1000 words**
MN	4662	36.42	54.4	3901	41.06	58.5
NN	777	6.07	9.0	713	7.51	10.8
JN	3885	30.35	45.4	3188	33.56	47.7
AJ	834	6.52	9.8	653	6.87	9.8
All	75491	589.77	890.5	59378	625.03	891.1

texts are slightly shorter than the C-level texts, the average number of bigrams per 1,000 words is also provided. The latter values were analysed with Student *t* test to assess whether some categories of bigrams are statistically more frequent in one CEFR level than in another. None of the tests were significant with a threshold of 0.05. An analysis exclusively based on frequency fails to reveal any difference between intermediate and advanced learners. The next stage in the analysis consists in investigating whether similar results will be obtained when the collocational strength within bigrams is taken into account.

Table 7 sums up the results of the comparison of the association scores – both MI and t-score – of all the categories of bigrams in the intermediate corpus as compared to the advanced corpus. For each category of collocational strength – non-collocational, low, medium and high – and for the 'beyond threshold' category, we provide the mean percentages for the B-level and for the C-level (the highest percentage is printed in bold type). The shaded cells highlight the cases where a significant difference was found between the intermediate and the advanced corpus. Details of significance levels and effect sizes are provided in Appendix 1. The values have been computed on the basis of all the bigrams present in the learner texts (tokens) as well as all the different bigrams present in each text (types). In counts based on bigram types, even if a learner uses a bigram several times, the bigram is only counted once.

The results strongly confirm our main hypothesis, which predicted a smaller proportion of lower-frequency, but strongly-associated, collocations [attested by MI] and a larger proportion of high-frequency collocations [attested by t-score] in the intermediate learner texts than in the advanced learner texts. This is confirmed for both types and tokens, the only exception being the t-scores for MN and

Table 7: Association scores (MI and t-score) for the 5 categories of word sequences (types and tokens)

Bigram type	Type/ Token	CEFR level	MI				t-score				BT
			NC	L	M	H	NC	L	M	H	BT
MN	Tokens	B	**10.5**	**16.5**	19.7	15.9	3.7	20.9	11.7	**26.3**	37.4
		C	7.4	12.4	17.4	**20.8**	2.5	22.5	11.4	21.7	**41.9**
	Types	B	**10.9**	**16.8**	17.9	14.5	**4.0**	21.0	11.8	**23.4**	39.9
		C	7.2	12.7	17.1	**19.5**	2.2	22.7	11.6	20.0	**43.5**
NN	Tokens	B	6.3	7.0	11.8	24.1	2.8	18.6	10.1	17.9	50.7
		C	3.2	6.5	12.4	**33.6**	0.7	22.6	12.4	19.9	44.3
	Types	B	6.0	7.1	10.9	23.3	2.8	18.0	10.2	16.3	52.6
		C	3.3	6.6	11.7	**33.6**	0.8	22.6	12.5	19.2	44.9
JN	Tokens	B	**10.8**	**18.8**	20.7	13.6	3.8	21.4	11.5	**27.3**	36.1
		C	7.9	13.4	18.3	**17.6**	2.6	22.2	10.8	21.6	**42.8**
	Types	B	**11.3**	**18.7**	19.3	12.6	**4.0**	21.7	11.7	**24.5**	38.0
		C	7.6	13.6	18.1	**16.4**	2.2	22.3	11.0	20.2	**44.3**
AJ	Tokens	B	**16.8**	23.4	27.7	11.8	5.7	17.6	14.0	**43.4**	19.3
		C	11.4	21.9	27.1	8.2	4.4	19.6	13.7	30.9	**31.3**
	Types	B	**17.1**	23.8	27.9	11.4	5.5	18.1	13.9	**42.8**	19.7
		C	11.5	21.9	27.0	8.2	4.4	19.7	13.6	30.9	**31.4**
All	Tokens	B	**57.9**	22.8	6.7	1.7	**14.2**	10.4	6.9	**57.6**	10.9
		C	55.1	23.2	6.7	**2.5**	12.9	**11.3**	**7.8**	55.6	**12.5**
	Types	B	**57.9**	21.9	6.2	1.7	**15.5**	11.2	7.4	**53.5**	12.3
		C	54.7	22.6	6.5	**2.4**	13.8	**12.2**	**8.2**	52.0	**13.8**

JN which are highly significant for non collocational (NC) bigram types, but not tokens.

Other interesting trends emerge from the Table 7:

- There are marked differences between the categories of word combinations: effects are particularly strong for *All* and weak for NN and AJ.
- The results for NN and JN are quite different. This confirms our initial assumption (see Section 2.2) and calls into question the validity of the merged category (MN). It would be interesting to investigate whether similar results would be found in a native vs. non-native comparison.
- Some collocational bands appear to discriminate more effectively than others. *High* is nearly always significant both for MI and t-score. NC also

displays a large number of significant scores often with large effect sizes. Non-collocational bigrams are systematically more frequent in B than in C. *Medium i*s rarely significant, which puts into question the relevance of this band.

Unlike the analysis based on frequencies only, the analysis in terms of collocational strength reveals very different behaviour in intermediate and advanced learners. It shows that L2 phraseological competence is characterised by a mixture of high frequency and low frequency collocations. As proficiency in the language increases, the balance between the two types of units changes: the proportion of high-scoring, high frequency sequences tends to decrease and that of high-scoring, but less frequent sequences to increase, but both types of units play an important role in language and continue to co-exist.

The main differences overall concern high-scoring combinations both for MI and t-score. This trend also appears from Durrant and Schmitt's study, which registered the most significant results for collocations from the highest bands ($t \geq 10$; $MI > 7$). The crucial role played by high-scoring combinations is also confirmed by the results of Durrant and Doherty's (2010: 145) study of collocational priming which suggests that "collocational priming may be restricted to word pairs which score very high on association measures (MI and t-score). Priming was only demonstrated here between pairs with MI scores of at least 6 and t-scores of at least 7.5". As suggested by the authors, these results should lead to a revision of the cut-off points traditionally used to identify collocations likely to be of linguistic interest.

The last column in Table 7 presents the results for the 'beyond threshold' (BT) category, i.e. the bigrams that are absent from the *BNC* or too infrequent (less than 5 occurrences) to be assigned a collocational score. The results show that there are significantly more BT bigrams in C than in B, except for NN which displays results pointing in the same direction but without reaching significance. Likewise, Durrant and Schmitt (2009: 174) found more BT bigrams in native speaker texts than those of non-native speakers, a finding for which they provided the following interpretation: "Firstly, non-natives' relative 'underuse' of low frequency and novel combinations would appear to indicate a degree of conservatism in their production – learners seem to over-rely on forms which are (according to *BNC* data) common in the language". This interpretation needs to be carefully assessed, however, as the BT category may contain both creative combinations, which are more likely to be used by advanced learners, and erroneous combinations, which will tend to be produced in greater quantity by less advanced learners. Space limitations do not permit us to analyse this category further in this article.

4 Intermediate vs. advanced learners: qualitative approach

It is important to take a close look at the word sequences that hide behind the figures in order to grasp the full meaning of the quantitative results. In this section we attempt to shed additional light on the marked difference between JN and NN (Section 4.1), the largely non-discriminating results for AJ (Section 4.2) and the highly significant results for All (Section 4.3).

4.1 Noun+Noun and Adjective+Noun sequences

The quantitative results in Table 7 show that the only difference between intermediate and advanced learners as regards Noun+Noun sequences is that the intermediate learners use significantly fewer high-scoring, lower-frequency NN sequences (both types and tokens) as attested by the MI scores. There is no significant difference as regards the higher-frequency sequences (attested by t-score). Table 8 lists the top-scoring NN sequences in the advanced learner corpus. The

Table 8: Top-scoring NN sequences in the advanced learner corpus (t-score and MI)

NN	t-score	NN	MI
world_war	61.6	spaghetti_bolognaise	18.6
member_states	40.6	chewing_gum	17.7
health_care	39.9	cayenne_pepper	16.4
police_station	30.3	lunatic_asylum	15.2
city_council	30.3	conveyor_belt	14.8
starting_point	30.7	soap_operas	14.2
health_services	28.4	lemon_juice	13.9
swimming_pool	28.2	ozone_layer	13.8
living_room	26.8	swimming_pool	13.4
phone_call	25.8	refugee_camps	13.3
heart_attack	25.3	steering_wheel	13.2
member_state	24.7	chilli_sauce	13.1
family_life	24.0	sesame_seed	13.1
grammar_school	23.8	fairy_tales	13.0
ice_cream	21.7	inferiority_complex	12.9
life_expectancy	21.6	chocolate_mousse	12.5
ozone_layer	21.3	drug_addict	12.4
education_system	20.8	channel_tunnel	12.3
insurance_company	20.3	ice_cream	12.2
channel_tunnel	20.3	rice_pudding	12.1

sequences in the left-hand column illustrate the kinds of sequences that are used with similar frequency by intermediate and advanced learners, while the sequences to the right illustrate the sequences that intermediate learners prove to underuse.[10]

The differences between the intermediate and advanced learners are much more pronounced as regards the use of Adjective+Noun sequences. Both parts of our hypothesis are confirmed for both types and tokens. Table 9 lists the JN sequences that achieve the highest scores for the two association measures in the intermediate corpus.

Table 9: Top-scoring JN sequences in the intermediate corpus (t-score and MI)

JN	t-score	JN	MI
prime_minister	97.2	nitrous_oxide	17.4
other_hand	73.9	hippocratic_oath	16.4
long_time	64.2	conscientious_objectors	15.9
local_authorities	63.4	juvenile_delinquency	15.1
great_deal	63.3	ultraviolet_radiation	13.8
other_people	61.5	conscientious_objection	12.6
young_people	59.5	fellow_countrymen	12.4
other_words	56.9	vicious_circle	12.2
other_side	54.4	pop_music	12.2
recent_years	52.4	double-edged_weapon	12.1
wide_range	52.3	armed_forces	12.0
young_man	51.3	male_chauvinist	11.9
higher_education	49.9	vested_interests	11.9
little_bit	49.9	human_beings	11.8
human_rights	47.9	technological_advances	11.8
social_services	47.7	flowering_plants	11.7
general_election	46.7	presidential_elections	11.7
social_security	46.4	warm_welcome	11.7
other_things	44.5	nervous_breakdowns	11.7
european_community	43.9	armed_robbery	11.6

10 It is interesting to note that four NN sequences appear in the two columns (*channel tunnel, ice cream, ozone layer* and *swimming pool*), a phenomenon that is not observed in the lists of the top-scoring bigrams for the other categories. The reason is that the frequency of co-occurrence of these four sequences is quite high and represents a sizeable proportion of the overall frequency of either word of the pair. As a result, the sequences meet the criteria to be selected by the two association indices. Another factor that plays a role is that NN sequences usually have relatively weak t-scores. The mean t-score for the 20 NN bigrams in Table 8 is 28.32, while it reaches 56.36 and 41.6 for JN and AJ sequences respectively.

It is interesting to note that JN collocations are often neglected in textbooks. The vocabulary lists tend to include single adjectives (e.g. *lavish, close-knit, noticeable* or *elderly*) even when they are used with a typical noun collocate in the corresponding unit (e.g. *lavish dinner, close-knit family, noticeable change, elderly parents*)[11]. This presentation method is clearly not conducive to successful acquisition of these types of collocations.

4.2 Adverb+Adjective sequences

The main difference in the use of Adverb+Adjective sequences by intermediate and advanced learners is that intermediate learners use more higher-frequency sequences identified by the t-score and non-collocational sequences identified by the MI score. While our hypothesis led us to expect a higher number of lower-frequency sequences in the advanced corpus, there is no significant difference between the two learner populations as regards these sequences. The difference between the two categories of sequence appears clearly from Table 10.

Our results show that learners progressively use fewer sequences that include adverbs displaying wide collocability like *very, more* and *too* but do not make significant progress in the use of more collocationally restricted combinations like *incurably ill* or *virtually impossible*. Like JN, this is a category that would clearly benefit from greater teaching focus.

4.3 All bigrams

The quality of the full range of bigrams turns out to be one of the best discriminators between the two learner populations. Both parts of our hypothesis are confirmed and the effect sizes are often quite large. This result brings support to Ellis et al.'s (2008) view of L2 phraseology as primarily influenced by n-gram frequency as opposed to native speaker phraseology which relies more on the mutual information score of the string:

> It is notable, however, that native speakers and ESL learners are sensitive to different metrics: for native speakers, (. . .) it is the MI of the string, the degree to which the words cohere at levels above those expected by chance, that influences their processing. In contrast, formula processing in the non-natives, despite their many years of ESL instruction, was a

11 All the examples come from *New Headway Intermediate – Student's Book* (Soars and Soars 2009).

Table 10: Top scoring AJ sequences in the intermediate corpus (t-score and MI)

AJ	t-score	AJ	MI
very_good	67.7	incurably_ill	11.8
most_important	67.6	mentally_ill	11.8
more_likely	59.4	wide_open	11.1
more_important	50.2	physically_handicapped	11.0
very_different	47.1	fully_justified	10.0
more_difficult	46.3	potentially_dangerous	9.8
very_important	45.3	much_nicer	9.4
very_difficult	44.7	virtually_impossible	9.3
too_late	42.6	highly_developed	8.7
very_small	38.0	highly_specialized	8.7
very_nice	37.3	absolutely_imperative	8.6
very_high	35.6	highly_probable	8.5
much_better	33.7	practically_impossible	8.5
more_complex	32.8	somewhat_ambiguous	8.5
most_common	32.6	terribly_sad	8.4
most_popular	31.5	totally_dependent	8.4
more_effective	30.8	relatively_cheap	8.4
very_close	30.6	extremely_difficult	8.2
very_similar	29.5	much_harder	8.2
more_general	28.7	most_influential	8.0

> result of the frequency of the string rather than its coherence. For learners at this stage of development, it is the number of times the string appears in the input that determines fluency. (Ellis et al. 2008: 384)

Table 11 lists the top 100 bigrams in the advanced corpus classified in decreasing order of t-score and MI score[12]. The top-scoring bigrams identified on the basis of t-score show how this type of phraseology is progressively acquired: many of the sequences get top scores partly because the words that compose them are very frequent grammatical words. In many studies, these bigrams are rejected as 'uninteresting'. We think, however, that they should be included as they play a role in the gradual acquisition of the syntagmatic axis of language. Besides purely grammatical sequences like *of the* and *to be* which are not phraseological in the traditional sense, the list of the bigrams that are assigned top t-scores in Table 10 also contains sequences such as *there is, the same, you know, I think, out of, more*

12 The association scores may be slightly different when computed on the basis of words (as is the case for *All*) or words + POS-tags (as is the case for the other categories of bigrams).

Table 11: List of the top 100 bigrams in the advanced corpus classified in decreasing order of t-score and MI score

All bigrams	t-score	All bigrams	MI
of_the	665.4	woe_betide	19.4
in_the	562.1	homo_sapiens	18.9
it_is	486.6	spaghetti_bolognaise	18.3
to_be	393.4	alma_mater	17.7
on_the	381.7	self-raising_flour	16.3
it_was	344.6	cayenne_pepper	16.2
do_not	335.5	conscientious_objectors	15.9
at_the	310.1	chewing_gum	15.8
i_am	287.4	conscientious_objector	15.4
there_is	285.8	tabasco_sauce	15.1
for_the	282.9	barbed_wire	14.9
from_the	282.8	juvenile_delinquents	14.8
by_the	278.7	conveyor_belt	14.8
will_be	277.3	baked_beans	14.6
have_been	274.4	soap_operas	14.2
did_not	262.8	mentally_handicapped	14.2
that_is	258.1	lunatic_asylum	14.1
had_been	256.3	dear_sirs	14.0
to_the	252.9	doorbell_rang	14.0
he_was	252.5	old-age_pensioners	14.0
with_the	251.0	arabian_peninsula	13.9
has_been	249.8	global_warming	13.9
as_a	248.8	lemon_juice	13.9
is_a	247.1	ozone_layer	13.8
is_not	245.4	high-heeled_shoes	13.7
i_have	244.0	pink_floyd	13.5
they_are	243.0	corporal_punishment	13.4
would_be	235.6	universal_suffrage	13.3
with_a	227.5	zebra_crossing	13.3
you_are	226.6	transformational_grammar	13.3
can_be	224.4	refugee_camps	13.3
the_same	223.6	chilli_sauce	13.1
he_had	220.3	baltic_republics	13.1
can_not	218.4	fairy_tales	13.0
if_you	217.8	swimming_pool	12.8
the_first	216.1	khaki_shorts	12.8
one_of	215.9	second-class_citizens	12.8
i_do	211.6	comfortably_furnished	12.7
in_a	211.2	steering_wheel	12.7
there_was	209.7	tropical_forests	12.6
we_have	205.8	sesame_seed	12.6
going_to	202.9	armani_suits	12.6

Table 11 (cont.)

All bigrams	t-score	All bigrams	MI
you_know	202.3	chocolate_mousse	12.5
i_think	198.9	racial_hatred	12.5
for_a	197.6	drug_addict	12.4
you_have	197.1	second-class_citizen	12.3
this_is	195.7	star_trek	12.3
they_were	192.7	sore_throat	12.2
there_are	191.7	tiny_tots	12.2
out_of	191.5	channel_tunnel	12.2
should_be	191.1	ice_cream	12.2
we_are	190.4	martial_arts	12.1
into_the	190.0	vicious_circle	12.1
may_be	189.8	silk_stockings	12.0
does_not	189.0	rice_pudding	12.0
you_can	187.3	poached_eggs	12.0
that_the	185.5	raw_materials	11.9
was_not	181.0	mountain_bikes	11.9
she_was	180.7	silk_scarf	11.9
he_said	180.3	nineteenth_century	11.9
she_had	180.0	civil_servants	11.9
more_than	179.9	ice-cream_parlour	11.8
was_a	179.1	leather_jackets	11.8
number_of	177.7	tsarist_regime	11.8
to_make	177.2	import_tariffs	11.8
to_do	176.6	roam_freely	11.8
part_of	176.1	twentieth_century	11.8
i_was	175.8	human_beings	11.8
such_as	174.8	apple_pie	11.8
i_will	173.9	silk_blouse	11.7
as_well	170.7	god_bless	11.7
and_i	170.4	mentally_ill	11.7
would_have	170.0	taxi_driver	11.7
and_then	169.6	tomato_soup	11.7
have_got	169.3	straw_hat	11.7
want_to	167.7	artificial_intelligence	11.6
not_be	166.8	armed_forces	11.6
to_get	166.5	sandy_beach	11.6
that_he	165.5	junk_mail	11.6
he_is	164.2	favourite_pastime	11.6
could_not	163.1	prime_minister	11.5
able_to	162.8	sun_shines	11.5
of_course	162.3	industrialized_countries	11.4
i_would	162.3	alarm_clocks	11.4
do_you	161.4	video_recorder	11.4

Table 11 (cont.)

All bigrams	t-score	All bigrams	MI
are_not	160.9	19th_century	11.4
of_a	160.8	curly_hair	11.4
but_i	158.0	mitigating_circumstances	11.4
over_the	157.4	manifests_itself	11.3
at_least	157.1	alarm_clock	11.3
a_few	156.8	taken_aback	11.3
could_be	156.3	fifteenth_century	11.2
not_know	155.8	plastic_beaker	11.2
but_it	155.6	20th_century	11.2
what_is	153.9	poole_harbour	11.2
of_his	153.7	elder_brother	11.2
that_it	153.5	male_chauvinists	11.2
would_not	153.4	neatly_stacked	11.1
i_mean	152.9	shaven_heads	11.1

than, number of, part of, such as, and then, as well, want to, able to, of course, at least, a few and *I mean*, that are more clearly phraseological.

The top-scoring bigrams identified on the basis of MI are of a completely different order. They show that one of the hallmarks of advanced students is that they have at their disposal a vast lexical repertoire including sequences as varied as *baked beans, mentally handicapped, junk mail, doorbell rang, manifests itself, god bless* and *taken aback* which cover the whole gamut of multiword units.

5 Conclusion

The overall results of the current study are consistent with the conclusions obtained by Durrant and Schmitt from a comparison of native and non-native speakers. They show that intermediate and advanced learners differ significantly in their use of phraseology manifested in the use of bigrams. Our initial hypothesis, that intermediate learner texts should be characterized by a smaller proportion of lower-frequency, but strongly-associated, collocations [attested by MI] than advanced learner texts and a higher proportion of high-frequency collocations [attested by t-score], is confirmed. Our study extends the scope of Durrant and Schmitt's study by investigating a larger number of collocations. Results show that it is useful to make a distinction between Noun+Noun and Adjective+Noun combinations as only the latter distinguish the B and C levels. The category

All also proves to be particularly effective in distinguishing B from C. The fact that it requires no pre-processing of the data is also a clear advantage as the accuracy rate of POS-tagging is likely to decrease at lower proficiency levels. One general lesson that can be drawn from our study is that it is useful to use two measures of collocational strength rather than just one, as is usually the case in studies of this type. The concurrent use of MI and t-score makes it possible to highlight two aspects of phraseology, each of which appears to play a part in foreign language acquisition.

Our research findings have significant implications for foreign language teaching. The phraseological coverage of pedagogical materials tends to focus on word-like units like compounds and phrasal verbs and allocate only limited space to the types of collocationally-restricted combinations that students need to master if they are to achieve an advanced level of proficiency. If we are to help learners attain a high level of phraseological competence, a key ingredient to fluency, a change of teaching priority is clearly in order.

Besides foreign language teaching, the main application of our study is to automated scoring, a research field that has grown considerably in the last 15 years and holds great potential for the teaching and testing of foreign languages. The standard approach to automated scoring relies on linguistic indices that are more or less strongly correlated with essay quality (Dikli 2006). The prototype of this type of system, the e-Rater developed by Educational Testing Services (ETS), relies on a set of automatically detected spelling and grammar errors, measures of lexical richness and text length (Burstein 2009; Burstein et al. 2004). Our results suggest that collocational strength could contribute to the prediction of essay scores, at least to distinguish between intermediate and advanced levels (for preliminary results see Vidakovic and Barker 2010). Our option of categorizing bigrams on the basis of POS-tagged corpora rather than manually takes on full significance when seen in the perspective of automated scoring. Admittedly the procedure results in some dubious pairs made up of words that are not syntactically linked (e.g. *centuries man* in *During centuries man has been subject to the law of nature*). However, these cases are few and far between. In addition, the fact that our results are compatible with those obtained manually suggests that the benefits outweigh the disadvantages. With the fast development of robust parsers it is possible to clean up the data, at least partly, so as to keep only the syntactically linked pairs, which Seretan et al. (2004) refer to as 'syntactic bigrams'.

Our research is preliminary and only tackles a limited number of issues in a vast and complex research field. It therefore leaves the ground open for a wide range of follow-up studies. Two directions would appear to be particularly worthy of investigation. First, it would be useful to carry out a similar study with data covering the full proficiency range, not just the B and C levels. The aim would be

to find out whether the collocational strength of bigrams is capable of distinguishing beginner levels (A levels in the CEFR) from intermediate levels. Will A level learners tend to display a higher proportion of high t-score combinations than intermediate learners or do these combinations emerge at a later stage? A second useful direction of further investigation might well be to focus on a different type of learner population. Our study only involves learners of English as a foreign language who usually learn the language in an input-poor environment. It would be extremely interesting to examine whether learners of English as a second language, who are immersed in the target language, tend to acquire the high MI combinations quicker. And these are but two examples. The field has only just begun to reveal its potential.

References

Allen, David. 2009. Lexical bundles in learner writing: an analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education*, 1. 105–127. http://park.itc.u-tokyo.ac.jp/eigo/KJEE/001/105-127.pdf (accessed 23 November 2013).

Burstein, Jill. 2009. Opportunities for Natural Language Processing Research in Education. Lecture Notes in Computer Science, 5449. 6–27.

Burstein, Jill, Martin Chodorow & Claudia Leacock. 2004. Automated essay evaluation: The Criterion Online Writing Service. *AI Magazine* 25. 27–36.

Chen, Yu-Hua & Paul Baker. 2010. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology* 14(2). 30–49.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

De Cock, Sylvie. 2000. Repetitive phrase chunkiness and advanced EFL speech and writing. In Christian Mair & Marianne Hundt (eds.), *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora*, 51–68. Amsterdam: Rodopi.

De Cock, Sylvie. 2007. Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight. In Mari Carmen Campoy & Maria José Luzón.(eds.), *Spoken Corpora in Applied Linguistics*, 217–233. Bern: Peter Lang.

Dikli, Semire. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment* 5(1). ejournals.bc.edu/ojs/index.php/jtla/article/download/1640/1489 (accessed 18 November 2012)

Durrant, Philip & Alice Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2). 125–155.

Durrant, Philip & Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *International Journal of Applied Linguistics in Language Teaching* 47. 157–177.

Ellis, Nick, Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly* 42(3). 375–396.

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*, 1211–1248. Berlin: Mouton de Gruyter.

Galaczi, Evelina & Hanan Khalifa. 2009. Cambridge ESOL's CEFR DVD of speaking performances: What's the story? *Cambridge ESOL: Research Notes* 37. 23–29.

Götz, Sandra & Marco Schilk. 2011. Formulaic sequences in spoken ENL, ESL and EFL. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, 79–100. Amsterdam: John Benjamins.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. In Anthony Cowie (ed.), *Phraseology: theory, analysis and applications*, 145–160. Oxford: Oxford University Press.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier & Magali Paquot. 2009. *The International Corpus of Learner English. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve: Presses universitaires de Louvain.

Green, Anthony. 2008. English Profile: Functional progression in materials for ELT. *Cambridge ESOL: Research Notes* 33. 19–25.

Groom, Nicholas. 2009. Effects of second language immersion on second language collocational development. In Andy Barfield & Henrick Gyllstad (eds.), *Researching Collocations in Another Language*, 21–33. Houndmills: Palgrave Macmillan.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Ishikawa, Shinichiro. 2009. Phraseology overused and underused by Japanese learners of English: A contrastive interlanguage analysis. In Katsumasa Yagi & Takaaki Kanzaki (eds.), *Phraseology, Corpus Linguistics and Lexicography: Papers from Phraseology 2009 in Japan*, 87–100. Nishinomiya: Kwansei Gakuin University Press.

Juknevičienė, Rita. 2009. Lexical bundles in learner language: Lithuanian learners vs. native speakers. *KaLBOTYRa* 61(3). 61–72.

Lee, Soyoung. 2006. A corpus-based analysis of Korean EFL learners' use of amplifier collocations. *English Teaching* 61(1). 3–17.

Lorenz, Gunter. 1998. Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In Sylviane Granger (ed.), *Learner English on Computer*, 53–66. London & New York: Addison Wesley Longman.

Paquot, Magali & Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32. 130–149.

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt (eds.), *Language and Communication*, 191–226. London: Longman.

Ping, Pang. 2009. A study on the use of four-word lexical bundles in argumentative essays by Chinese English majors: A comparative study based on WECCL and LOCNESS. *Chinese Journal of Applied Linguistics* 32(3). 25–45. http://www.celea.org.cn/teic/85/85-25.pdf (accessed 23 November 2013).

Seretan, Violeta, Luka Nerima & Eric Wehrli. 2004. Using the web as a corpus for the syntactic-based collocation identification. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon. 1871–1874.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Soars, Liz & John Soars. 2009. *New Headway Intermediate Student's Book*. Oxford: Oxford University Press.
Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of language* 2(1). 23–55.
Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *Modern Language Journal* 97. 77–101.
Van Rooy, Bertus & Lande Schäfer. 2003. Automatic POS tagging of a learner corpus: the influence of learner errors on tagger accuracy. In Dawn Archer, Paul Rayson, Andrew Wilson & Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (Technical Papers 16), 835–844. Lancaster University: University Centre for Computer Corpus Research on Language.
Vidakovic, Ivana & Fiona Barker. 2010. Use of words and multi-word units in Skills for Life Writing Examinations. *Cambridge ESOL: Research Notes* 41. 7–14.
Wei, Yaoyu & Lei Lei. 2011. The use of amplifiers in the doctoral dissertations of Chinese EFL learners. *Chinese Journal of Applied Linguistics* 34(1). 47–61.

Appendix 1

			MI				t-score				BT
			NC	**L**	**M**	**H**	**NC**	**L**	**M**	**H**	**BT**
MN	Tokens	p	**	***		***				**	*
		d	.42	.61		.51				.43	.34
		B	**10.5**	**16.5**	19.7	15.9	3.7	20.9	11.7	**26.3**	37.4
		C	7.4	12.4	17.4	**20.8**	2.5	22.5	11.4	21.7	**41.9**
	Types	p	***	***		***	***			**	*
		d	.55	.60		.60	.49			.38	.29
		B	**10.9**	**16.8**	17.9	14.5	**4.0**	21.0	11.8	**23.4**	39.9
		C	7.2	12.7	17.1	**19.5**	2.2	22.7	11.6	20.0	**43.5**
NN	Tokens	p				**					
		d				.43					
		B	6.3	7.0	11.8	24.1	2.8	18.6	10.1	17.9	50.7
		C	3.2	6.5	12.4	**33.6**	0.7	22.6	12.4	19.9	44.3
	Types	p				**					
		d				.48					
		B	6.0	7.1	10.9	23.3	2.8	18.0	10.2	16.3	52.6
		C	3.3	6.6	11.7	**33.6**	0.8	22.6	12.5	19.2	44.9

			MI				t-score				BT
			NC	L	M	H	NC	L	M	H	BT
JN	Tokens	p	**	***		**				***	***
		d	.38	.66		.45				.50	.49
		B	**10.8**	**18.8**	20.7	13.6	3.8	21.4	11.5	**27.3**	36.1
		C	7.9	13.4	18.3	**17.6**	2.6	22.2	10.8	21.6	**42.8**
	Types	p	***	***		***	***			**	**
		d	.51	.65		.47	.48			.44	.46
		B	**11.3**	**18.7**	19.3	12.6	**4.0**	21.7	11.7	**24.5**	38.0
		C	7.6	13.6	18.1	**16.4**	2.2	22.3	11.0	20.2	**44.3**
AJ	Tokens	p	*							***	***
		d	.27							.47	.62
		B	**16.8**	23.4	27.7	11.8	5.7	17.6	14.0	**43.4**	19.3
		C	11.4	21.9	27.1	8.2	4.4	19.6	13.7	30.9	**31.3**
	Types	p	*							***	***
		d	.29							.45	.60
		B	**17.1**	23.8	27.9	11.4	5.5	18.1	13.9	**42.8**	19.7
		C	11.5	21.9	27.0	8.2	4.4	19.7	13.6	30.9	**31.4**
All	Tokens	p	***			***	***	***	***	**	**
		d	.88			.84	.71	.56	.67	.43	.45
		B	**57.9**	22.8	6.7	1.7	**14.2**	10.4	6.9	**57.6**	10.9
		C	55.1	23.2	6.7	**2.5**	12.9	**11.3**	**7.8**	55.6	**12.5**
	Types	p	***			***	***	***	***	*	**
		d	1.01			.89	.88	.58	.64	.33	.39
		B	**57.9**	21.9	6.2	1.7	**15.5**	11.2	7.4	**53.5**	12.3
		C	54.7	22.6	6.5	**2.4**	13.8	**12.2**	**8.2**	52.0	**13.8**

Significance level (* = $p \leq .05$, ** = $p \leq .01$, *** $p \leq .001$)
Size of the effect (*d*)