Neo-classical compounds in student writing: a corpus-based study

Smith, S.

Verbum XXXIV, 2012, nº 2, 277-296

NEO-CLASSICAL COMPOUNDS IN STUDENT WRITING: A CORPUS-BASED STUDY*

Simon SMITH
Coventry University

RÉSUMÉ

Le présent article examine l'utilisation des composés néo-classiques (CNC) au sein du corpus BAWE (British Academic Written English). Sont étudiés des textes rédigés par des étudiants anglophones et non-anglophones dans le cadre de leurs études de trois disciplines (Anglais, Gestion et Ingénierie). La répartition des CNC au sein de ces catégories y sera discutée. Nous démontrerons que la répartition des CNC dans les trois disciplines varie sensiblement et que l'utilisation des CNC parmi les anglophones diffère de celle des locuteurs non natifs, tandis que son utilisation au sein de groupes non anglophones reste assez homogène. Une taxinomie des CNC est proposée dans le but de faciliter l'analyse qui suivra. Certaines implications de l'étude pour l' apprentissage de l'anglais sont également présentées.

ABSTRACT

This article examines the use of neo-classical compounds (NCCs) in the British Academic Written English corpus (BAWE). Academic texts written by students from both English and other linguistic backgrounds, in three academic disciplines (English, Business Studies and Engineering) are studied, and the distribution of NCCs among these categories is discussed. It is shown that NCC distribution in the three disciplines differs considerably, and that native English speaker usage of NCCs is at variance with that of non-native speakers, but that usage among different L1 groups is fairly constant. A typology of NCCs is presented, to aid the analysis which follows. Implications of the study for English teaching are also presented.

1. INTRODUCTION

In this study, we look at the distribution of neo-classical compounds (NCCs) in academic English texts, written by both native and non-native students. We look for patterns in the distribution, and explore the

* The assistance of Coventry University colleagues Dr Nicole Keng and Professor Hilary Nesi in the preparation of this paper is gratefully acknowledged.

significance of NCC knowledge and awareness in language teaching and learning.

Bauer (1988) defines NCCs as words consisting of two or more free morphemes (of Latin or Ancient Greek) which are bound, not free, in the modern language concerned, such as English *biology*. As is widely known, a large proportion of English vocabulary is of Graeco-Latin (G-L) origin. Many native speakers consider frequent use of G-L words, including NCCs, to be indicative of a greater vocabulary and indeed of a higher level of education, or even social class: Corson (1982; 1985) posited the existence of a "lexical bar" in English, whereby members of certain social classes who do not acquire the vocabulary necessary to express more abstract technical and academic thought are denied full access to the curriculum as they go through the school system. Certainly, a high proportion of scientific and technical terms do take the form of NCCs. Lüdeling (2006) presents some of the historical reasons for the arrival of NCCs in modern European languages. During the European Enlightenment, advances in science meant that much new terminology was needed, while at the same time use of Latin for formal or academic purposes gave way to the vernacular tongues of the continent. A common mechanism for generating needed new vernacular terminology was to borrow and piece together elements from the still prestigious classical languages.

In Smith and Keng (2014), we demonstrated that Chinese native speakers' knowledge of G-L words (including NCCs) was weaker across the board than that of French or Finnish students, while Finns had the best knowledge. When only G-L vocabulary was taken into account, the French students improved on their score for items of all origins, while the Finnish students performed slightly worse, and the Chinese performance worsened more significantly. This confirmed our hypothesis that the more G-L words a language attests, the more likely the native speaker of that language is to know English words of G-L origin than ones of unspecified origin.

In this study, we build on Smith and Keng (2014) to compare the performance of English learners with English native speakers, while at the same time confining the investigation to the use of NCCs, rather than of G-L words generally. We examine student academic writing from the British Academic Written English corpus (BAWE[1]). The corpus includes annotation of the native language of each student author. It is clear that one of the features differentiating the work of writers of different language back-

[1] The BAWE data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

grounds will be vocabulary choice; and it is also plausible that density and appropriateness of NCCs will contribute to this differentiation. This study investigates the distribution of NCC use by writers of different linguistic origins.

The analysis is conducted using both a narrow and wide definition of neo-classical compound, by establishing two sets of NCCs, prototypical and extended, and presenting findings on the usage of both sets. This approach was motivated by the fact that there is no straightforward and unambiguous definition of NCC, with Bauer (1998) observing that "neoclassical compounding acts as some kind of prototype, from which actual formations may diverge in unpredictable ways." Some of these divergences are represented as Types within our framework.

The BAWE corpus is also annotated by broad discipline group (life sciences, physical sciences, social sciences and humanities) and by individual major (Engineering, Business and the like). The study investigates the relative density of NCCs in different disciplines, and seeks to establish whether NCCs, or certain classes of NCC, are more likely to occur in some disciplines than in others. With this in mind, the following research questions are addressed in the present study.

RQ1. Is there a significant difference between native and non-native student writers' work, in terms of NCC density?

RQ2. Do students of different non-English L1s use NCCs differently?

RQ3. Are any NCCs limited to specific disciplines or discipline groups in the BAWE corpus? If so, do they share any distinctive formal characteristics?

2. ORGANIZATION OF THE PAPER

The next section of the paper begins with a survey of the relevant literature. Before any attempt to answer the research questions can be made, it is necessary to determine exactly what lexemes are admitted to the NCC class. The explanatory approaches of various authors will be set out.

Following this, and drawing on the various approaches found in the literature, different types of NCCs are consolidated into a typology of prototypical and extended set NCCs. The typology will facilitate the subsequent analysis of academic texts, in terms of density and distribution of NCCs, that is the main business of this paper.

In the Methodology section, we describe the two-stage procedure by which NCCs are identified in the corpus and other aspects of the data collection. The procedure for distinguishing "difficult" NCCs from those which are commonly known is also explained.

The Results section presents NCC density data at the two stages of data collection, and at the two difficulty levels, by various participant groupings or subcorpora, including native language and academic discipline. The distributions of NCCs in the different subcorpora are then compared in a Discussion section. The paper concludes with an account of the limitations of the study, and of its implications for teaching and learning.

3. BACKGROUND AND LITERATURE REVIEW

Bauer (1988), Amiot & Dal (2007) and Lüdeling (2006) characterize and define NCCs, with examples from a range of modern European languages. Amiot & Dal (2007) offer one of the most constraining definitions of NCC. For them, an item has to meet all of four criteria to be considered an NCC. Both of the item's components must have existed as lexemes in Latin or Greek. Equally, the components must **not** be free morphemes in the modern language, so an item such as *sociolinguistics* would be disallowed. There must be a linking vowel {o} or {i} between the components, and "generally [the words should] belong to the learned vocabulary of scientific or technical fields." In their study, they treat French lexical items which appear to break some of the constraints, such as *ludothèque*, which is informal rather than learned, and *cassettothèque*, which is also informal and includes a non-classical component. It is precisely this sort of definitional difficulty that prompted us to set up our Typology of NCCs, presented below.

As noted in the introduction, Bauer (1988) only insisted that the first and second of Amiot & Dal's constraints must apply. Ten years later, however, Bauer (1998) presents arguments for **and** against allowing *sociolinguistics* as an NCC. Booij (2005) extended Bauer's earlier definition to include words of which one component morpheme can have an independent existence as a lexeme in the modern language, but with a different or more restricted meaning; this allows for *telegraph*, for example, to be admitted to the class of NCCs, because its component morpheme {graph} does not carry the same meaning (in the original Greek) as the modern word *graph*. Baeskow (2006) does not distinguish between the *biology* and *telegraph* types, while admitting words such as *insecticide* and *nanoplankton* to a separate class of NCCs, in which one of the morphemes may exist in the language as an independent word. The English words *insect* and *plankton* have, in fact, undergone a substantial meaning shift away from the classical words from which they are derived. Words such as *microchip* and *biofeedback*, which include a non-classical free morpheme, however, are not treated by Baeskow (2006) as NCCs.

Bauer (1983) allows *hypertrophy* to stand as an NCC, as the second morpheme (clearly unrelated to the independent word *trophy*) is bound. *Hyperactive*, however, Bauer analyses as an affixed form, since *active* exists as an independent word with substantially the same meaning. In this study,

we include all these word types. In our typology, presented in the next section, we would classify *microchip* and *hyperactive* as a Type 2 NCC (a free form and a bound form), and *hypertrophy* or *telegraph* as Type 1 (two bound forms). Both are **prototypical** NCCs, by our account.

Like Amiot & Dal (2007), Quirk *et al.* (1985) state that NCCs can be characterized by the presence of a binding vowel or empty morph, such as the {o} in *psychotherapy*, (or the {i} in *insecticide*) and some papers in the present volume confine their investigation to words of this type. In our data, which is from academic writing, there are some examples of NCCs which do not contain a standard binding vowel, for example *archaebacterium* and *cryptanalysis*, as well as more widely known examples such as *telephone*. In many other cases, the vowel appears to be an integral part of the initial component, as with words beginning *bio-*, or to serve no binding function, as with *microorganism*. In this paper, therefore, we do not insist on its presence.

3.1. Exclusion of affixed items from the study

Not requiring the binding vowel as a criterial feature means that we must distinguish between NCC components and affixes in some other way; otherwise there is a risk that we will end up classifying all polymorphemic words of Graeco-Latin (G-L) origin as NCCs. One issue is productivity: we ignore morphemes which can be freely affixed to a large proportion of words of a given class. We exclude completely the privative prefix {de}, as in *demotivate,* repetitive {re}, as in *rebuild*, as well as the Latinate negating prefixes {in}, {il} and {ir} and similar semi-grammatical formatives. Bauer (1998) treats {trans} as an affix, and in a 1983 publication he comments that NCC components carry more "lexical information" than affixes, with {pre} for example contributing only grammatical information. The implication that classical prepositional forms cannot serve as NCC components, only affixes, seems at odds with the assertion noted above that *hypertrophy* begins with a NCC component (*hyper* is the Greek word for "over").

In our data, forms similar to almost all words beginning with prepositions {sub} (under) and {com}/{con}/{col} (with) were found to have already existed in ancient Greek or Latin, so that they could not in any event be considered **neo**-classical. Often, as say with *compulsory* or *subjunctive*, it is not possible to distil the semantic contribution of the first part without detailed etymological knowledge. These were therefore treated as ordinary prefixes and excluded from our analysis completely. Forms which were already **compounds** in antiquity, rather than affixed forms, are however incorporated into the extended set of NCCs, as Type 6 in our typology.

For certain other classical elements, such as {inter}, {pre} and {trans}, the semantic contribution is relatively transparent; words formed with these elements since classical times were also admitted to the extended set, as

Types 7 or 8. Items such as {hyper}, {hypo}, {intra} and {supra} have rather more specialist roles in English word formation processes than their plain prepositional roles in Greek and Latin would suggest, so these are treated alongside {bio} and {psycho} as prototypical NCCs.

We exclude Latin-origin suffixes whose only role is to change word class, such as the nominalising {tio}, rendered in English as {tion} in words such as *hydration.* However, we do include in Type 1 words ending in {-itis}, {-ase}, and {-ate}, which could plausibly be treated as suffixes. We opted for Type 1 because these components have quite specific semantics in the fields of medicine, biology and chemistry respectively, and are not simple grammatical forms.

3.2. NCC typology design

Our typology of NCCs is set out first, and described in the following paragraphs.

Type	Description	Example
1.	2 bound forms, first seen in a modern language	schizophrenia
2.	Bound form + free form, first seen in modern language	nanotechnology
3.	2 bound forms, or 1 bound + 1 free; first seen in post-classical Latin (3rd century onwards)	psychological
4.	Form first appearing in classical Latin or Greek, with substantially different meaning in modern English	technological
5.	Backformation or clipped form	contraception
6.	Form first appearing in classical Latin or Greek	physiological
7.	Bound form which has prefix-like productivity (e.g. *inter-* or *trans-* followed by free form)	postmodernism
8.	Bound form which has prefix-like productivity followed by bound form	intertextuality

Table 1. – Typology of NCCs used in the study

Type 1 included a number of terms from linguistics, as well as many medical terms, such as *encephalopathy* and *bronchiectasis*, and enzyme names such as *dehydrogenase*. The {pathy} and {ase} of these terms may be productive in medicine and biology, and could be considered suffixes in those fields. However, following McCray, Browne & Moore (1988) and Namer & Zweigenbaum (2004), we assume that they are compound components. Many words ending in {logy} and {graphy} are also categorized here, including *historiography*, *chromatography* and *immunological*.

The second type of NCC includes copulative forms such as *musculo-skeletal* and *cardiovascular*, as well as the more typical headed compounds,

such as *nanotechnology*. Although the compounds did not exist in classical times, some (for example *electrophoresis*) are composed of the almost exact juxtaposition of two classical elements. Others incorporate a free form which itself is derived from Latin or Greek, but has undergone changes of meaning or form. Examples are *phytochemicals* and *telecommunications.* Finally, contra Baeskow (2006), we include here compounds whose free components are not classical at all, but which do not really differ from the latter subtype in any formal way, such as *microwave* and *electrowetting*. As Baeskow herself points out, these compounds differ from nonce lexical items such as *queenomania* and *hamburgerology,* or the *bobologue* and *déclinologue* of French (Amiot & Dal 2007), or the Dutch *netwerkcratie* (rule by net-working; Beelen 2004). The reason is that these nonce words to have been coined in order to draw attention to the phenomenon by the oddness of the word – clearly, that does not apply to a word like *microwave.* Beelen does actually refer to his *netwerkcratie* and *kengetallencratie* (rule by market indicators) as 'neoklassieke composite' (NCCs). Since our BAWE analysis did not contain any nonce compounds, we were excused the decision on what to do with them.

Of the third type, *psychology* probably entered English from a Modern Latin word *psychologia* which was coined in Germany in the 16th century. *Manufacture* was a mediaeval Latin word, which entered English, via French, around the same time, according to Harper (2013). While it may appear trivially true that an item should have entered English or another modern language directly, and not have existed as a compound in the classical language, the question of compounds entering varieties of Latin that evolved subsequent to the classical era does not seem to appear in the literature; this motivated our decision to include this third item in the typology.

Jackson (2002) writes that NCCs are "for the most part" not known in the classical languages from which they are derived. It is not quite clear what he means by this: perhaps a reference to forms which evolved in post-classical era Latin, or possibly a reference to what we refer to as Type 4. These are compounds which already existed in classical Latin or Greek, with only minor differences in spelling, but with a substantially different meaning. Often, the modern compound represents a concept which was unknown or unavailable in ancient times. *Meteorological*, for example, did not acquire its current meaning until the Middle Ages (Harper, 2013), and in ancient Greek meant a "discussion of high or celestial matters". *Superficialis* meant "relating to the surface" of a physical body to the Romans, with the modern meaning of *superficial* only emerging in the Middle Ages.

Type 5 combined forms, backformations and clipped forms, are of mainly incidental interest, as only a small number of instances were found in our data. They include *contraception* (and *contraceptive*), which according to Harper (2013) was coined in the 19th century from {contra} and (a clipped

form of) *conception*. A few years later, the word *proprioception* was likewise coined from Latin *proprius* and a clipped form of *reception.* The meanings are, of course, entirely unrelated. *Horticultural* and related forms appear 15 times in the corpus, coined (on the analogy of *agriculture*) in the 17th century, while *extrapolating* appears to be a backformation from *extrapolation*, itself coined on the analogy of *interpolation. Ergonomically* and related forms are based on the Greek word for "work" and a clipped form of *economics*.

Biodiversity is another example which incorporates a clipped form (since the origin of the word is clearly "biological diversity"). This is rather similar to the phenomenon described by Iacobini (2004), whereby a component such as {tele}, which in *telephone* and *television* means "distant", yet in many cases has itself taken on the meaning of "television" or "telephone" in Italian, as seen in *telespettatore* (*téléspectateur* in French, "TV viewer"), or *telesoccorso* ("telephone assistance") respectively. *Teletext* and *telemarketing* appear to be similar English examples, although it could be claimed that here the {tele} still actually means "distant"; and they do not, in any case, occur in our corpus.

Types 6 NCCs, for example *physiological* and *philosophical*, existed in classical Latin and Ancient Greek in almost identical forms and with very similar meanings to the equivalent modern words. Therefore, they cannot therefore strictly be said to be neo-classical compounds. Nonetheless, they are included in the extended set because it is clear to the educated speaker that they are compounds, and what their components are.

Morphemes such as {inter} and {trans} are very productive in English and other modern languages, and yet they make a clear semantic contribution: they are not merely grammatical. Type 7 and 8 are, therefore, included in the extended set, as they seem to be on the borderline between affixes and NCC components.

In an earlier phase of the research, a ninth category was set up in the typology: NCCs that had entered English via French, including such items as *infrastructure* and *stereotype*. This type was subsequently abandoned, as the era (classical, post-classical, or modern) that the NCC came into use was deemed of more interest than the exact provenance. Normally, one would expect that many of these words would have arrived via a Romance language. It is of interest to note, therefore, that many NCCs were in fact loaned to English from German, largely because of the contribution of German scientists to learning in the latter part of the 19th century, and early 20th century; they were accustomed to name new concepts and inventions using Latinate coinages. The following is a small sample of this set of words, found in the BAWE corpus: *chemotherapeutic*, *chromosomally*, *phylogenetically*, *phenomenologists, psychopathology, prosopagnosia*, *terminological*, *mitochondrial*, *macromolecules*.

To summarize the typology of NCCs given in Table 1, Types 1 and 2 are deemed to be **prototypical** NCCs. Types 1-8, taken together, constitute the **extended** set of NCCs; Types 3 and 4 did not originate in a modern language, unlike Types 1 and 2, but neither did they originate in a language of antiquity, at least not with the same sense as they have today. Types 5, 7 and 8 arose through a different word formation process than compounding, while Type 6 already existed in a classical language in antiquity, with approximately the same meaning. In this study, the textual density of prototypical NCCs, and of the extended set NCCs, are discussed and analysed separately.

4. METHODOLOGY

4.1. Preliminary word-length analysis

The BAWE corpus has quite a rich set of annotations. For all texts, the native language of the author, the grade (distinction or merit) obtained by the author, text genre (essay, report etc.), the subject and broader discipline area, as well as the year level of the student, are all indicated. At the word level, POS mark-up is also available; however, information about NCC status of vocabulary is not (and we are not aware of any large corpora where it is; those conducting corpus-informed research on NCC distribution, such as Lasserre (2013) and Warren (1990), tend to create their own corpora).

The corpus was too big for us to carry out a full annotation for the purposes of this paper, so our initial analysis was based on word length. We assumed that most NCCs in the corpus could be identified if the search was confined to a set of longer words, and after some experimentation established that a word length of 13 allowed us to capture most NCCs. At length 12 and below, we found only a very small proportion of NCCs, and many of these were general language words such as *telephone* which one would not expect to be representative of academic vocabulary.

In order to demonstrate that an analysis of shorter words was not worthwhile, we examined the list of all words in the corpus of a length between 8 and 11, inclusive. The first bona fide NCC was *biological*, occurring at frequency position 462, with *parameters* a few entries below. More frequent than this on the list were items that qualify only marginally as NCCs, or not at all: *introduced*, *introduction* and *hypothesis*, which existed in Latin or Ancient Greek with substantially the same meaning as English (Type 6 in Table 1 above); *economic*, *technology* and *diagnosis*, which existed in the classical language with a somewhat different but cognate meaning (Type 4); and *agricultural*, which entered English and other modern languages through Late Latin, although it was not attested as a combined form in classical times (Type 3).

We then analysed sets of comparable subcorpora in terms of the proportions of words of length 13 and over. We compared subjects (English,

Business and Engineering); broader discipline (humanities, physical science and so on); author gender; grade; university year level of the author; and a selection of author native languages (Chinese, French, Finnish and English, the choice of the first three being corresponding to the L1s of students whose knowledge of Graeco-Latin vocabulary was investigated in Smith & Keng (2014)).

Having established which of these sets of subcorpora promised the most revealing analyses (that is, the greatest variation in number of longer words), we worked through the wordlists highlighting the NCCs. We also categorized the NCCs found into the 8 different categories, as shown in the typology of Table 1.

As noted above, we included only items with a length of 13 or more letters. A further constraint was that we only recorded NCCs that occurred two or more times; this was because hapax legomena (items occurring once) are unlikely to bring any added value statistically, given the extra overhead in analysing what would have become extremely long lists of words.

4.2. Data analysis procedures

After the NCCs had been identified, tokens and token counts were compiled in a spreadsheet, grouped according to author's L1 and discipline (these were the subcorpus categories chosen for more detailed analysis), as well as NCC Type, per Table 1. As noted above, we treated Types 1 and 2 as prototypical NCCs, and first investigated the different patterns of prototypical NCC use by subcorpus (L1 and academic discipline). We also investigated the patterning of the extended set of NCCs, that is all of Types 1-8.

It was also thought appropriate to include a measure of difficulty or degree of specialization in the analysis (since it is clear that NCCs are more likely to denote difficult or technical concepts than certain other types of morphological construct). One possible approach would have been to use frequency statistics from a reference corpus, but given the variety of lexis from different domains in the BAWE corpus under analysis, a simpler procedure was decided on. The author is a native English speaker with a notional vocabulary of 34100 words, according to http://testyourvocab.com/. This is probably an average sized vocabulary for a UK academic. With the exception of vocabulary in the author's own discipline – linguistics – it could be assumed that NCCs not known to him were above some arbitrary, yet constant, level of difficulty. Therefore, all NCCs in our spreadsheet were annotated with a code K or U, denoting whether they were known/familiar or unknown/unfamiliar to the author.

5. RESULTS

We first discuss the selection of subcorpora for detailed NCC analysis based on word length, before turning to the outcomes of the NCC analysis itself, based in turn on the comparison of discipline subcorpora and native language subcorpora.

5.1. Word length analysis

Table 2 shows the counts of words of length 13 or more. We decided that there was enough variation in the proportion of long words in the different student L1s to merit further investigation. Interestingly, the distribution of long words in the L1 subcorpora did not entirely tally with the distribution of NCCs. We return to this question later in this section.

L1	**Total words**	**Total long words**	**Percentage of long words**
French	204,063	2018	0.99%
Chinese	741,048	7,124	0.96%
Finnish	38,669	317	0.82%
English	5,765,739	49,762	0.86%
Gender			
Female	4,989,426	45,996	0.92%
Male	3,346,836	28,210	0.84%
Discipline groups	**Total words**	**Total long words**	**Percentage of long words**
AH	2,243,330	16,519	0.74%
LS	1,754,545	17,850	1.02%
SS	2,727,126	27,970	1.03%
PS	1,611,261	11,867	0.74%
Grade			
D	3,759,740	33,727	0.90%
M	4,152,814	36,453	0.88%
unknown	423,708	4,026	0.95%
Disciplines			
Engineering	698,927	4,565	0.65%
Business	408,254	4,074	1.00%
English	329,853	1,910	0.58%

Table 2. – Words above 13 letters long, by subcorpus

NCCs were not analysed by gender. Although there was a difference in the proportion of long words used by men and women, it is almost certain that this could be explained by the imbalance of genders in the particular academic subjects, and a direct analysis of this was deemed of greater interest. Another difficulty was the scale of such an investigation, which would have meant analysing the entire BAWE corpus. For this reason, an analysis of the four discipline groups into which the corpus is divided was also foregone, even though the differences in distribution of long word usage are somewhat counterintuitive: one might have expected to see length features in common in natural sciences (LS and PS in Table 2) on the one hand, and on Humanities and Social Sciences (AH and SS) on the other.

There seemed to be no clear relationship between use of long words and grade awarded, so that subcorpus was not further analysed either. The specific disciplines Engineering, Business and English seemed to offer great promise in terms of the explanatory power of long word distribution, so this subcorpus was selected for further exploitation.

5.2. NCC analysis based on discipline subcorpora

In this section, first the prototypical NCC distribution is discussed, followed by the extended set.

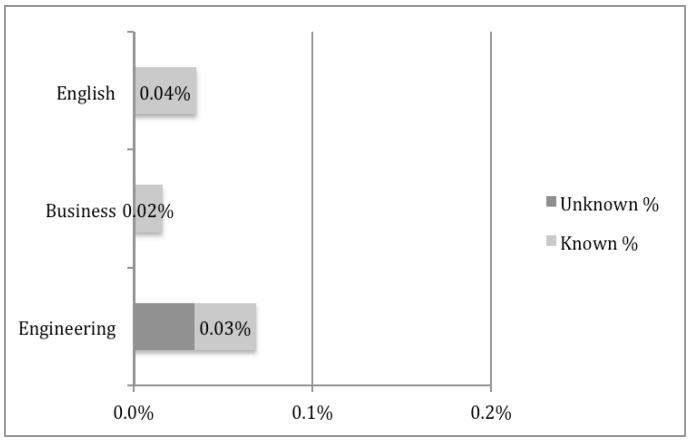

Figure 1[2]. – Percentage of prototypical NCCs in discipline subcorpora, known and unknown to author

Figure 1 shows the proportion of prototypical NCCs against total words in the writing of English, Business and Engineering major students. The class of prototypical NCCs includes combinations of 2 bound forms, or of a

2 In all figures in this paper, the *x*-axis (abscissa) shows the total percentages for known and unknown; the percentage printed on the bar shows the known words only.

Bound form + free form. The NCCs must first have arisen in a modern language, which is in many cases English, although some of the NCCs arrived through other European languages, for example *infrastructure* (via French) and *chemotherapeutic* (German). The component {infra}, incidentally, is an example of a morpheme that could have been described as a prefix under some accounts (being a preposition of Latin). However, we felt, when designing the NCC typology, that it makes a sufficiently important lexical, rather than grammatical, contribution to the resulting combined form to be classed as an NCC component, and lacks the productivity of a prefixed form. This was alluded to in Section 3.1.

Engineers, it will be seen, use a higher proportion of NCCs in their writing than students in the other two majors. It might be supposed that this is due to the relatively large number of technical or scientific terms in their academic vocabulary; the fact that the author is unfamiliar with around half the words certainly seems to bear this out. An inspection of the data reveals that there are 44 NCCs of Type 2, a bound form followed by a free form, and almost all of them are technical in nature. The most frequent, *microcontroller*, occurs 69 times, followed by *microstructure* with 40. In third and fourth place came *piezoelectric* and *interferometer*, neither of which were familiar to the author. In total, there were 304 uses of Type 2 NCCs.

Surprisingly, there were no uses of Type 1 NCCs, those that consist of two bound forms, in the Engineering domain. Business students used two, *phenomenologists* and *demographical*, on two occasions each, while English majors used ten, totalling 55 uses. Some of these terms would have been from linguistics essays (*morphological* and *heteroglossia*, for example), and that is why the author was familiar with them. Even though there were no terms unfamiliar to the author in the Business and English major subcorpora, a cursory inspection shows that all the NCCs in the Business subcorpus were actually words in general rather than specialist usage, with the exception of the two Type 1 NCCs mentioned above, and economics terms *macroeconomic* and *supranational.* The morpheme {supra} had been treated as an NCC component, like {infra}, even though the {inter} of *international* is a prefix, because of the former's specialist meaning and lack of productivity.

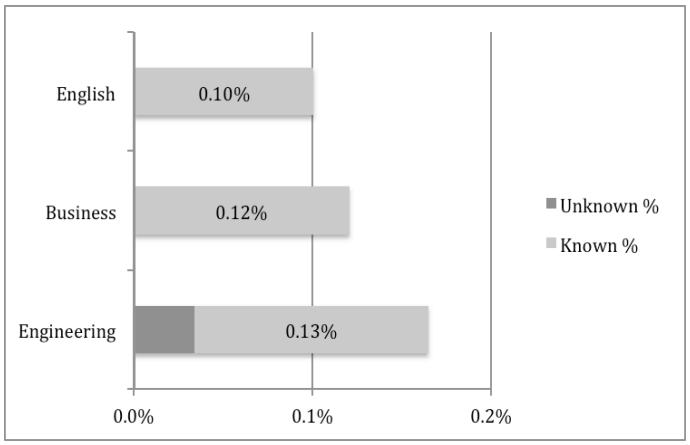

Figure 2. – Percentage of extended set NCCs in discipline subcorpora

The most striking point about Figure 1 was perhaps the oblique contrast with the results presented in Table 2: that is, that Business students used a much larger proportion of longer words than either engineers or students of English, but very few prototypical NCCs. To try to account for this, we present Figure 2, which shows the usage of the extended set of NCCs, including all eight types from Table 1, in the three disciplines. We can see that the number of NCCs does increase substantially, under the new analysis, in the Business subcorpus, although still not to the extent shown in Table 2, which includes longer words which are not part of the present analysis. An inspection of the corpus reveals that many of the longer words used by Business students are indeed G-L words, but without a transparent morphological structure (or incorporating only POS-changing affixes). The four most frequent long words in the subcorpus, for example, are *organisations*, *implementation*, *organisational* and *responsibility*. This finding appears to reinforce the earlier conclusion that Business student writing includes more general vocabulary, and fewer technical terms than other domains.

In English and Engineering, too, Figure 2 indicates a substantial increase in the use of NCCs when the extended set taken into account, but the difference is not as marked as for Business students. In Engineering, the effect is largely due to the incorporation of 319 occurrences of *manufacturing* and *manufacturers*, a loan to English from late Latin (Type 3 NCC, per Table 1).

5.3. NCC analysis based on L1 subcorpora

Figure 3 shows the distribution of Type 1 and 2 NCCs in the L1 subcorpora. The very low proportion of NCC use among Finnish L1 students could be an effect of the inadequate sample size, which at 38,669 words is

many times less than any of the other L1 groups, as shown in Table 2. The English native speaking group make the most use of NCCs generally and of difficult NCCs (defined earlier as those unfamiliar to the author). At first glance, this seemed to be because of the apparently large number of medical terms found, with *lymphadenopathy* appearing 52 times, and *cardiovascular* 48 times. Closer inspection, however, revealed that the proportion of native / non-native texts in the Medicine subcorpus is scarcely different from that in any other discipline. It is, therefore, possible that the extensive use of NCCs, particularly technical vocabulary, represents a greater degree of sophistication or confidence amongst students writing in their own native language. Another interpretation is simply that (intuitively enough), native speaking students command a larger vocabulary, including a larger range of less frequent and more difficult words, including NCCs.

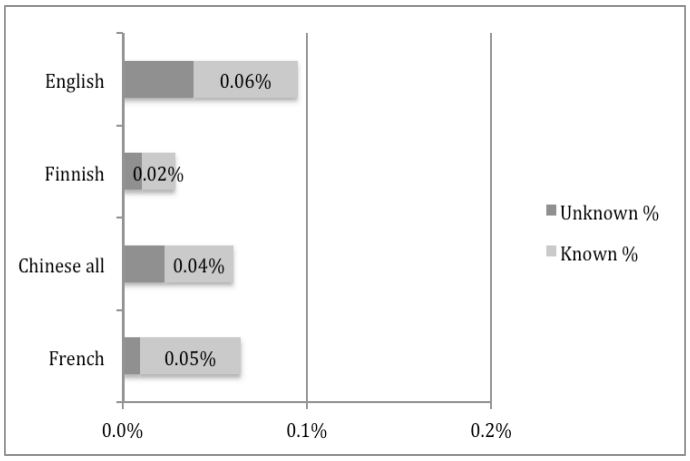

Figure 3. – Percentage of prototypical NCCs in L1 subcorpora, known and unknown to author

Since French is a Romance language, attesting many words of classical origin, it is surprising that we do not find a larger proportion of NCCs in the French students' work, in Figure 3. When we look at the extended set of NCCs, a class which includes many French cognates, we see that the usage by French students shoots up to 0.29%, as shown in Figure 4. The figure also indicates that the proportion of difficult words is relatively low, indicating that the words selected here are probably non-specialist language cognates.

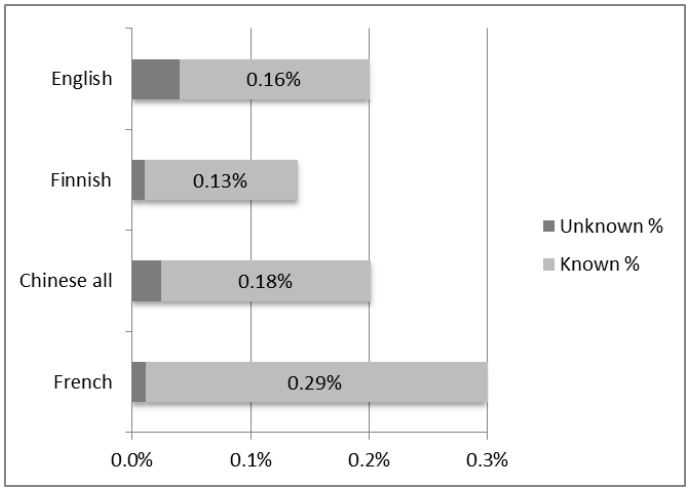

Figure 4. – Percentage of extended set NCCs (Types 1-8) in L1 subcorpora

6. DISCUSSION

The research questions are repeated here, for the convenience of the reader.

RQ1. Is there a significant difference between native and non-native student writers' work, in terms of NCC density?

RQ2. Do students of different non-English L1s use NCCs differently?

RQ3. Are any NCCs limited to specific disciplines or discipline groups in the BAWE corpus? If so, do they share any distinctive formal characteristics?

It was demonstrated in Figure 3 that English native speakers make considerably greater use of NCCs than non-native speakers. Native speaking students can be expected to have a larger vocabulary size than colleagues of other linguistic backgrounds; higher frequency vocabulary will be shared with non-native speakers, and it is at the level of infrequent, difficult vocabulary that the difference is expected to emerge.

It was expected that NCC use would be more widespread among speakers of a Romance language, such as French, which itself attests NCCs. Thus, the finding that Finns use fewer NCCs than French speakers was expected (although as noted in the Results section, the Finnish sample size was probably inadequate). With Chinese speakers, though, the level of NCC usage was approximately the same as among the French, whereas one would expect a lower usage because of the lack of NCC cognates in Chinese. One possible explanation for this is that the Chinese language does include a class of compound-like words incorporating bound forms, which is in many

ways analogous to NCCs: this is noted and discussed by Arcodia (2007). Briefly, a large proportion of multi-syllabic words of Chinese are composed of morphemes which cannot stand alone as a word, at least not in the spoken language. While the concept of NCC formation might in principle be familiar to Chinese learners, it is very unlikely that the similarity would ever have been pointed out to them by teachers. We will return briefly to this point when discussing implications for teaching.

Another, perhaps more plausible, explanation for the unexpectedly similar levels of NCC usage among Chinese and French students was suggested by Figure 4: Many prototypical NCCs belong to a discipline-specific sublanguage which is, by and large, newly acquired by all non-native speakers in the course of their study. When the extended set of NCCs is taken into account, we see increased usage among the French students. This set includes more non-specialist words, and many of them are French cognates.

This last point also serves to address the third research question: In the Business subcorpus, it was noted earlier, many of the NCC terms seemed to belong to general language; however, the Engineering subcorpus was dominated by technical or specialist terms. In the English major subcorpus, too, a reasonable proportion of linguistics and other specialist vocabulary was found amongst the NCCs. In Engineering, the set of NCCs used was overwhelmingly dominated by terms incorporating one bound and one free form (Type 2), while the English major subcorpus was somewhat better represented by what Bauer (1988) and others would view as the more canonical Type 1 NCCs, that is those consisting of two bound morphemes. Further research will perhaps answer why that should be.

7. CONCLUSION

In concluding, we address the limitations of the research and some implications for language teaching and learning, before making closing remarks.

The BAWE corpus includes English academic texts by native and non-native student writers. The non-native speaker component consists of fairly high quality writing, which has been rewarded with an above average grade in UK university assessment. Originally, we had planned to make use of a second corpus, composed of writing by students of a lower level of English proficiency, for comparative purposes. Unfortunately, this corpus contained only essays on one topic, and it soon became apparent that it did not offer the breadth of vocabulary needed to make a realistic comparison with the BAWE work. Where NCCs were used at all (and in the work of some students, they were not) the same words cropped up repeatedly, with little variation. In future work, it will be beneficial to locate a multi-discipline, multi-topic corpus of work by lower proficiency students.

Also, given greater time and resources, a thorough annotation of the BAWE and learner corpora, by NCC type, should be attempted. Our analysis

of NCC use by major subject and L1 has been somewhat piecemeal, and it was difficult to inspect categories within subcorpora. For example, it would have been useful to conveniently extract all uses of Type 1 NCCs by French L1 users in the Business domain. This would also have permitted an analysis by broad discipline area (Humanities, Physical Sciences and so on) as was noted earlier. In future, attention could also be paid to academic writing genre (essay, case study, experimental report, and so on) to determine whether interesting patterns of NCC usage emerged in different types of student text. Any differences in NCC usage due to gender of the student could have been investigated, too.

Finally, the study was limited to NCCs of more than 12 letters in length, which occurred more than once in the corpus. It is possible that without these constraints, the findings would have been more representative.

This study has implications for the teaching of English, especially English for Academic Purposes, to lower proficiency students who are learning academic subjects through the medium of English. Exposure to examples of high quality academic writing is in the opinion of this author of benefit to such students; his professional practice suggests that they are generally happy to learn from the work of peer role models. Using a corpus approach such as that we have adopted, it would be possible to generate lists of discipline-relevant NCCs (and other specialist vocabulary). These lists can then be used to search for example texts within the corpus. They can also be used to seed a tool such as WebBootCat (Baroni *et al.*, 2006), which can find relevant texts on the web, to be used as study materials. Such lists can also be used for learning and memorization of specialized vocabulary (students from China, for example, tend to favour this type of learning approach, unconventional as that may seem in a Western pedagogical context).

If it is the case that Chinese students are particularly amenable to the study of NCCs because of the existence of a broadly similar category of morphological structures in their own language, it may be worth teachers pointing this out as part of vocabulary instruction. The value of explicit teaching of productive word components, including elements of G-L words, has been described elsewhere, for example by Zheng & Nation (2013), and it is likely that study of productive NCC components would be of benefit too.

This paper has presented a corpus-based analysis of NCC use in academic English writing. A typology of NCCs, based on BAWE corpus data, was presented, and some findings regarding the distribution of NCCs by discipline and linguistic origin of writer were described and discussed. Some limitations of the work and directions for future research were given, along with some pedagogical implications of the study.

REFERENCES

AMIOT D., DAL G. (2007). Integrating neoclassical combining forms into a lexeme based morphology. In: G. Booij, L. Ducceschi, B. Fradin, E. Guevara, A. Ralli, S. Scalise (eds), *On-Line Proceedings of the 5th Mediterranean Morphology Meeting.* Fréjus, 15-18 September 2005.

ARCODIA G. (2007). Chinese: A Language of Compound Words? In: F. Montermini, G. Boyé, and N. Hathout (eds), *Selected Proceedings of the 5th Décembrettes: Morphology in Toulouse*, 79-90. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1617.

BEELEN H. (2004). Van leenwoord tot inheemse nieuwvorming: De herkomst van neoklassieke composita op *–cratie. Neerlandistiek.nl* 04(02). Available at: http://www.neerlandistiek.nl/?000078.

BARONI M., KILGARRIFF A., POMIKÁLEK J., RYCHLÝ P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. Oslo: *Proceedings of EAMT 2006*. 247-252.

BAUER L. (1983). *English word-formation.* Cambridge: Cambridge University Press.

BAUER L. (1988). *Introducing linguistic morphology*. Edinburgh: Edinburgh University Press.

BAUER L. (1998). Is there a class of neoclassical compounds in English and is it productive? *Linguistics* 36: 403-422.

BAESKOW H. (2006). Lexical properties of selected non-native morphemes of English. *Tübinger Beiträge zur Linguistik 482.*

BOOIJ G. (2005). *The grammar of words: An introduction to linguistic morphology*. Oxford: Oxford University Press.

CORSON D. (1982). The Graeco-Latin (G-L) instrument: a new measure of semantic complexity in oral and written English. *Language and Speech* 25: 1, 1-10.

CORSON D. (1985). *The lexical bar*. Oxford: Pergamon Press.

HARPER D. (2013). *Online etymology dictionary*. Available at: http://www.-etymonline.com/.

IACOBINI C. (2004). Composizione con elementi neoclassici. In: M. Grossmann, F. Rainer (eds), *La formazione delle parole in italiano*. Tübingen: Max Niemeyer Verlag, 69-95.

JACKSON H. (2002). *Lexicography: An introduction*. London: Routledge.

LASSERRE M. (2013). Neoclassical compounds and language registers. In: *Proceedings of JéTou 2013*, Toulouse, May 16th – 17th 2013, 98-108.

LÜDELING A. (2006). Neoclassical word-formation. In: K. Brown (ed.) *Encyclopedia of language and linguistics, 2nd Edition*. Oxford: Elsevier.

McCRAY A.T., BROWNE A.C., MOORE D.L. (1988). The semantic structure of neo-classical compounds. *Symposium on computer applications in medical care*, November 9: 165-168.

NAMER F., ZWEIGENBAUM P. (2004). Acquiring meaning for French medical

terminology: contribution of morphosemantics. In: Annual Symposium of the American Medical Informatics Association (AMIA), San Francisco.

SMITH S., KENG N. (2014). Acquisition of Classical Origin Words by Chinese, French and Finnish Learners. *Language Education in Asia*, 4 (2), 122-134. Available at: http://dx.doi.org/10.5746/LEiA/13/V4/I2/A03/Smith_Keng

WARREN B. (1990). The importance of combining forms. In: W. Dressler, H. Luschützky, O. Pfeiffer, J. Rennison (eds) *Contemporary morphology*. Berlin: Walter de Gruyter.

QUIRK R., GREENBAUM S., LEECH G., SVARTVIK J. (1985). *A comprehensive grammar of the English language*. London: Longman.

ZHENG W., NATION P. (2013). The word part technique. *Modern English teacher*, 22(1), 12-16.