Can I have a word please? – Supporting learning at Foundation level through use of a corpus of student-generated texts.

Megan Bruce and Simon Rees
Durham University Foundation Centre

Abstract
International students typically expect to encounter academic language problems and to have to acquire a new vocabulary for their studies. However, the need to expand their (native speaker) vocabulary to include general academic as well as discipline-specific language comes as a shock to many home students. The FOCUS project is a concordancing programme which we have created at Durham University to allow Foundation students to explore a corpus of student-generated texts in order to improve their own writing. The project also provides a suite of activities to guide students through corpus searches in order to enhance their understanding of both subject-specific and general academic vocabulary.

Keywords: ***academic vocabulary development, higher education, student-generated text***

Background evidence
All teachers have the task of inducting students into their Community of Practice. In these disciplinary communities, members help one another to establish knowledge and norms (November & Day, 2012; Wenger 1998). It is widely acknowledged that helping students to acquire the vocabulary they will need both to study their subject and to write academic English in a more general sense is one of the first challenges that staff and new students have to face (Berkenkotter, Huckin, & Ackerman, 1991; Drury & Webb, 1991; Freedman, 1987; November & Day, 2012; Woodward-Kron, 2004).

This paper outlines the FOCUS project which has recently been established at Durham University Foundation Centre. In this paper we will firstly outline the rationale for the project, before moving on to explain how it has been created. We will then discuss the use currently being made of the corpus tool and conclude by discussing the possible future direction of the project.

Project rationale
How important is academic vocabulary?
Freedman (1987) acknowledges that lack of familiarity with discipline-specific terminology can be a significant barrier to students who are trying to enter into a new discipline (November & Day, 2012). We have found this to be the case in our department where students are often unable even to frame their question about the content of a module because they do not have the vocabulary to do so.

Our initial response to helping students with vocabulary issues was to develop a suite of activities (www.dur.ac.uk/foundation.science) including an online glossary to which students could contribute, and some online activities covering areas such as meanings of affixes to enable students to determine the meaning of unfamiliar words (Rees & Bruce, 2012). These continue to be used by students studying science

subjects with some success. However, using these activities and other similar activities from the EAP Toolkit (http://www.elanguages.ac.uk/eap_toolkit.php) highlighted a concern amongst our home students that they do not have the meta-language to talk about the problems they encounter with language and thus it can be very difficult to address the problem.

Joan Didion famously said "Grammar is a piano I play by ear" and this sums up the feeling of our native speaker students towards any overt teaching of language. They are proficient speakers of English (though their writing is often markedly less proficient) but, to follow the analogy, cannot read music. They cannot comfortably discuss how method sections of reports need to be written in the passive voice, or how it is important to use Standard English past participles in academic writing rather than the dialect variants. Our home students do not expect to encounter linguistic difficulty when they begin their programme of study and often have significant confidence issues where grammar and lexis are concerned. They are surprised, for example, that words that they already know such as "heat", "process" and "energy" also have specific academic meanings and they have difficulty learning new language information.

In this situation, our international students have two distinct advantages over the home students. Firstly, because English is not their first language they expect to encounter linguistic difficulties and to have to work to solve them. Secondly, they have meta-language: because they are already language learners, they have the vocabulary to talk about their difficulties in order to receive support.

What vocabulary do students need to learn?

Given a need for students to learn new vocabulary in order to become members of their Community of Practice, the next obvious question to ask is what lexical items students need to learn. Nation (2001) divides vocabulary into three groups: high frequency words (covering about 80% of most texts); academic vocabulary (words which are most often found in academic writing and which comprise 8-10% of academic texts); and technical vocabulary which is dictated by subject area and typically covers around 5% of academic texts (Hyland & Tse, 2007, p.236).

Hyland and Tse outline that there is significant evidence to suggest that first year students tend to find academic vocabulary harder to learn than technical vocabulary because it is less likely to be taught explicitly (Flowerdew, 1993) and the large number of "academic" words mean that each lexical item occurs relatively infrequently (Worthington & Nation, 1996). Hyland and Tse also dispute Nation's categorisation of vocabulary into the above three sections. Their research shows that Nation's "academic vocabulary" is still discipline specific. For example, they show that "the word *process* is far more likely to be encountered as a noun by science and engineering students than by social scientists" (Hyland & Tse, 2007, p. 244). Drawing on Trimble's (1985) claim that "in different disciplinary environments words may have quite different meanings (ibid p. 247), Hyland and Tse conclude that they cannot "support the division between academic and technical vocabulary" (ibid p. 249).

A well-known study in the area of academic vocabulary is Coxhead's compilation of an Academic Word List (AWL) (Coxhead, 2000). This research has been the basis of

several learner textbooks, such as Schmitt & Schmitt (2005). The AWL identified 570 word families which are frequently used in academic texts. The list was based on a corpus compiled by Coxhead which consisted of 3.5 million words of academic writing, from 414 academic texts written by more than 400 authors across 28 subject areas (Schmitt & Schmitt, 2005 p. vi). Despite their disagreement regarding the classification of vocabulary, both Coxhead (2000) and Hyland and Tse (2007, p. 251) agree that lexical items need to be learned in context and this idea formed the basis of our FOCUS project.

How to teach vocabulary in the absence of meta-language

Having decided to find a way to help Foundation students learn new vocabulary more successfully, our initial challenge was to identify what to "know" a word means. Knowing a word typically requires an understanding of the following to enable someone to have both receptive and productive knowledge: what the word means; how it is spelled; how it is pronounced; which other words it typically collocates with; what part of speech it is; what register it is; its grammatical characteristics (e.g. countable/uncountable, transitive/intransitive); how frequently it is used; what its derivative forms are; whether it has any connotations (Schmitt 2000 in Schmitt & Schmitt 2005, p. vii). Trying to teach some of these elements explicitly requires grammatical knowledge and a meta-language that our home students typically do not possess, as discussed above. Therefore, we decided that our vocabulary project needed to circumvent explicit teaching and focus instead on allowing students to notice the relevant characteristics of the words in question by seeing them in context.

Schmitt and Schmitt (2005, p. vi) outline Nation's (1990) conclusion that "it takes from five to sixteen or more [encounters] for a word to be learned". Nation claims that a learner needs to notice a word in context a number of times in order to understand and learn all its different characteristics. It can take a very long time to notice a word in context on up to 16 different occasions, so using a concordancing programme to search corpus data has the advantage of reducing this time and speeding up learning. An obvious question to ask, however, is whether concordancing really provides the same experience for the learner. Cobb (1997, p. 314) states that concordancing is able "to mimic the effects of natural contextual learning" and other studies suggest that corpus searches can be effective as "noticing" opportunities for students.

Data-Driven Learning (DDL)

Johns (1991) coined the term "data driven learning" to describe a learning situation where "*the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data – hence the term "data-driven learning" (DDL) to describe the approach*" (Johns, 1991, p. 2). In DDL the learner uses data to uncover the rules behind the language while the teacher "*provides a context in which the learner can develop strategies for discovery*" (ibid).

DDL using corpora and concordancing programmes has been commonly used in language classrooms for the last twenty years. However, although its use is wide-spread, it has been focused mainly in the learning of second languages. As has been mentioned, some of our Foundation students are learning in a second language but the majority are native speakers who are trying to learn a subject-specific vocabulary in their own language. The use of concordancing in a native speaker context is much

less well established, yet fits well with the needs of our native speaker students who do not have the meta-language to be taught about their linguistic deficits explicitly.

Building the corpus – the FOCUS project

Departmental profile

Durham University Foundation Centre has an annual intake of approximately 200 non-traditional students. The department aims to widen participation in and access to Higher Education and therefore accepts students who lack some of the formal qualifications. The majority of students are either mature learners returning to education or international students who are unable to study to a sufficient level in their own country for direct entry to UK degree programmes (https://www.dur.ac.uk/foundation.centre/). The Foundation Centre offers progression to all the departments in Durham University, and consequently we have students studying an extremely broad range of subject areas.

We decided that the best way to help our students learn the vocabulary they needed for their studies was to build a corpus of student texts which our students could then search using a concordancer. This decision was informed by the following:

- Our students did not have the knowledge to benefit from being explicitly taught about language and yet their written work contained significant errors;
- Words need to be "noticed" between 5 and 16 times in order to be learned (Nation 1990);
- "Noticing" needs to take place in authentic contexts;
- DDL can allow the individual learner to make their own discoveries about language.

Our project is entitled "FOCUS" which is short for "**FO**undation **C**orp**US**". It is a corpus of academic writings produced by Durham University students (undergraduate and postgraduate) in various (initially STEM) subjects. Tribble (1997) cautions against using apprentice performances as corpus data. However, we would argue that since the function of our corpus is to teach Foundation students to write like conventional university students, in this case student writings are expert rather than apprentice performances.

Criteria for text inclusion

Acquiring texts to include in a project of this nature is always difficult as it requires the co-operation of a range of different people (Alsop & Nesi, 2009, p. 76-81). Some of our texts are PhD theses which are freely accessible within the university. Since our corpus can currently only be viewed by registered members of the university, we were allowed to include theses. However, the majority of our texts have been sent to us by students for inclusion in the corpus. We have targeted one university department at a time, explained our project to them and asked for permission to contact their students and ask them to send us copies of strong examples of academic writing. We define "strong example" as a piece of work which was assessed at 60% or above by the department. With the help of each department, we identified particular assignments which exemplified (a genre of) academic writing in that subject area. We obtained a list of students who had scored 60% or more in that particular assignment and contacted them to ask them to send us a copy of their assignment. The contact email outlined the aims of and ethical procedures in of the project. To incentivise students to send us their writing, we entered all names of

contributors into a draw for a £100 Amazon voucher each term. So far we have successfully obtained texts from Chemistry, Earth Science, Sociology, Criminology and Sport and are in the process of liaising with Medicine and Pharmacy.

During our initial approach, some departments expressed a concern that our corpus could turn into an essay bank which students could plagiarise. In fact, this would not be a possible use of the corpus data. As can be seen from the screenshot below, a search for a particular word reveals the keyword plus around 40 characters of text to the left and to the right of the keyword (totalling 100 characters per line).

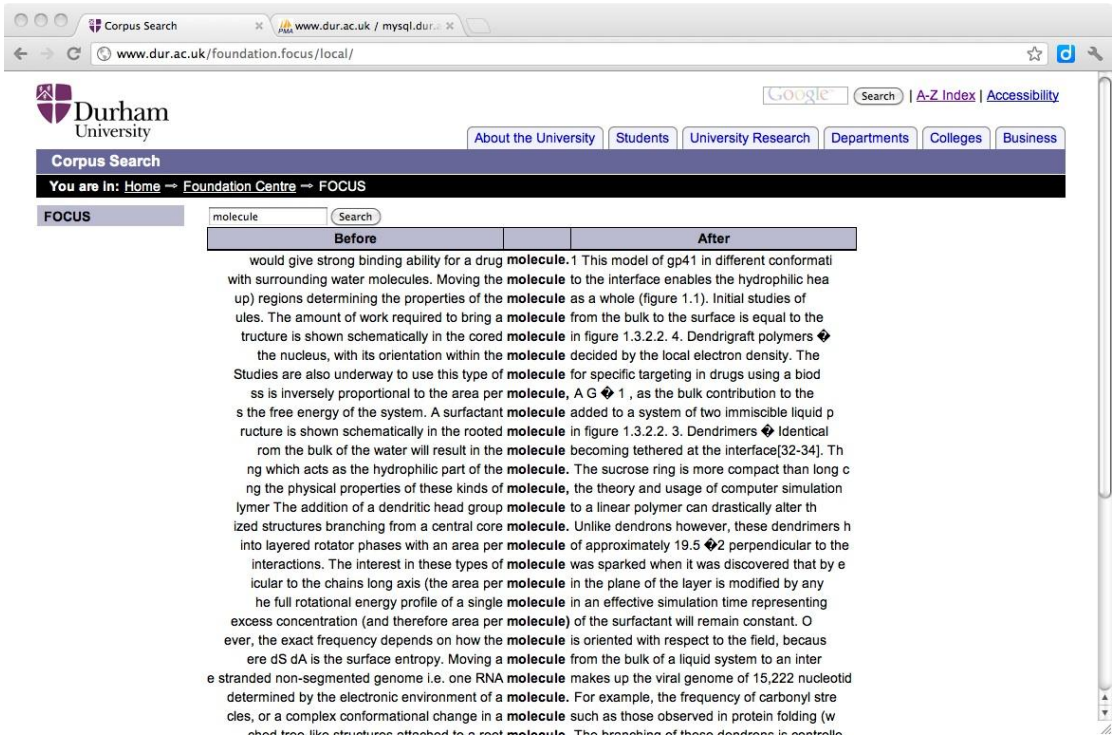

Figure 1: Screenshot of "molecule"

Clicking on a single instance of the keyword gives a slightly larger text fragment, but still only around 200 characters. No coding is made available to the user which would allow them to identify and piece together fragments to form a complete assignment.

Concordancer design

Alongside the building of the corpus of texts we needed to decide which concordancer we would use to run queries on the corpus. Johns (1991) defines a concordancer as:

> "*able to recover from the text all the contexts for a particular item (morpheme, word or phrase) and to print them out in a way which facilitates rapid scanning and comparison. The most usual format is the keyword-in-context (KWIC) concordance in which the keywords are arranged one below the other down the centre of the page, with a fixed number of characters of context to the left and to the right. A useful refinement, particularly where one is concerned with regularities and patternings in large numbers of citations, is the ability to sort alphabetically the contexts to the left or right of the keyword so that similar contexts are grouped together*." (Johns 1991:2)

A more concise definition from Tribble (1990, p. 11) states that: "*What the concordance does is make the invisible visible*".

In our initial phase of the FOCUS project we were awarded a Higher Education Academy and UK Council for International Student Affairs (HEA/UKCISA) grant to explore existing online concordancing programmes. This evaluation led us to conclude that the freely available programmes did not contain all the functionality that we wanted for our project, but the programmes that we would need to purchase were going to be too expensive given that we did not have a guaranteed income-stream for our project.

Our next step was to apply for Durham University Enhancing the Student Learning Experience Funding which we were awarded and which we were able to use to pay a colleague in the Computer Science department to design a bespoke concordancing programme for our project.

Functionality of FOCUS

Our concordancer can be accessed from this webpage (http://www.dur.ac.uk/foundation.focus/) but requires a Durham University login. Any visitors to the site can, however, see a YouTube demonstration of the tool and access our blog and twitter feed.

A user can perform a simple search by just entering a keyword. This can be a word, morpheme (using our wildcard symbol %) or phrase. A more advanced search can also be performed which limits the findings by level (e.g. UG/PG), type (e.g. essay/lab report) or subject (e.g. Chemistry/Earth Sciences).

The search results are displayed in a screen like this:

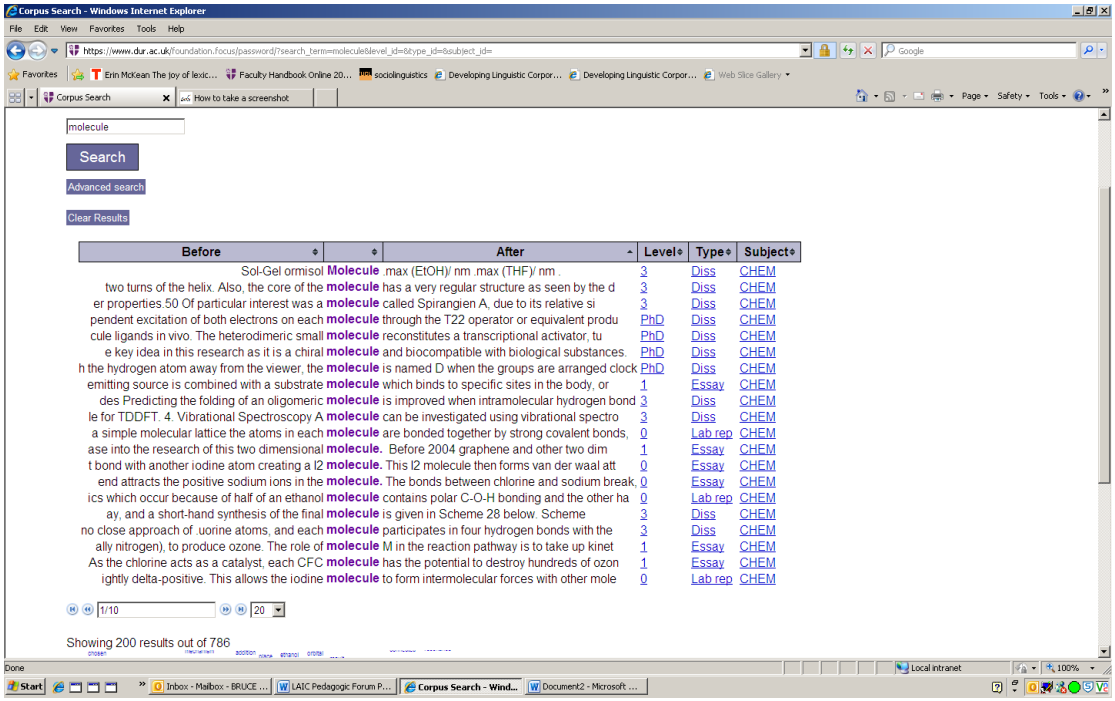

Figure 2: Screenshot of molecule showing level, type and subject

The screen can be set up to show 20, 40 or 100 concordance lines and it chooses a random 200 lines from the total held in the corpus.

Users can sort the data alphabetically using the "Before" or "After" tabs to identify common collocations. So, for instance, we can see that the most common collocate of "molecule" is "water":

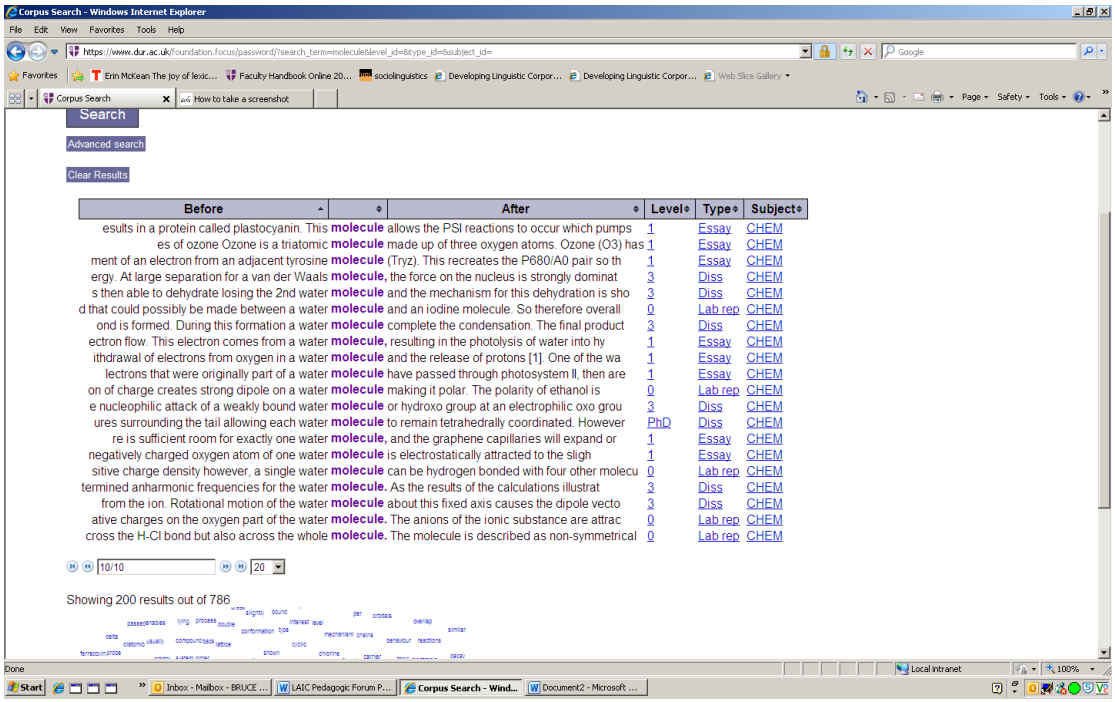

Figure 3: Screenshot of molecule sorted left to show prevalence of collocate "water"

The tool also includes a word cloud feature to identify common collocations and guide users into useful explorations about their chosen key word:

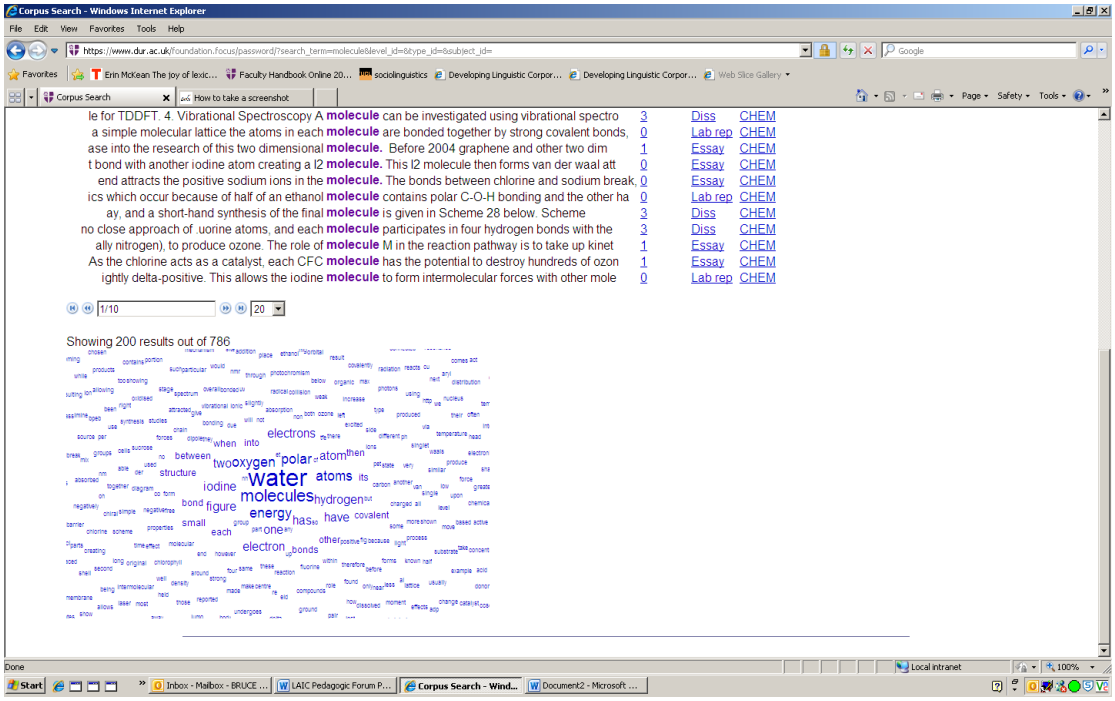

Figure 4: Screenshot of word cloud for molecule

Another feature of the tool is its wild card function using the % symbol. This allows users to search for all forms of a word family, so searching for combin% would give: combination, combinations, combinatorial, combine, combines, combined, and combining.

The wild card symbol also allows users to explore particular affixes, such as the search below for %icity:

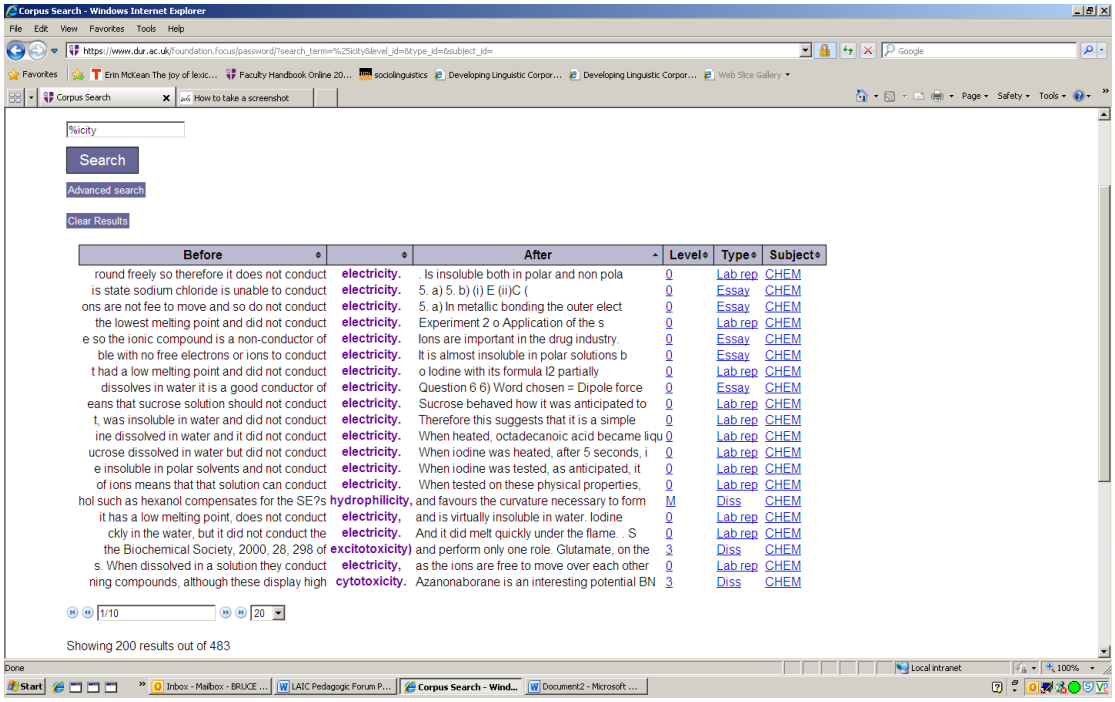

Figure 5: Unsorted hits from the wildcard %icity query

Clicking on the central column would alphabetise the keywords with the –icity suffix and allow users to see which ones occur most frequently. Students can also use this type of search to deduce the meaning of various affixes, which in turn improves their ability to guess at the meaning of unknown words in the future. This activity is available as a YouTube screencast (http://www.youtube.com/watch?v=s3Ep06t3e-M).

Teaching activities with FOCUS

We are still in the relatively early stages of building this corpus and beginning to use it in our classrooms. Feedback from initial trials with students has been positive and we are in the process of trialling the activities within course programmes.

Over the course of the past twelve months we have worked with a group of Foundation Year students to develop a range of concordance based self-study activities. These can be viewed at www.dur.ac.uk/foundation.science (click on link to language skills for learning) and are summarised in the table below. The design cycle of activity identification, development and review was effective over the course of the year to develop meaningful and engaging activities. It has enabled us to demonstrate how FOCUS can be used effectively in teaching in different contexts and how it can be used to develop study and language skills.

Activity	**Sub-Activities**
Writing Laboratory Reports	The Method Control Variables The Conclusion
Affixes in Science	
Words with Multiple Meanings	
Developing Reading	Scientific Vocabulary
Types of Scientific Language	Scientific Language Questionnaire Personal Glossary
Using Connectives	General vs Subject-Specific Vocabulary Typical Problems using connectives Using a corpus to explore connectives Writing practice using connectives

Table 1: Outline of activities designed to accompany the corpus

The set of activities address a range of subject-specific language skills such as improving understanding of common affixes in science to different writing genres such as writing effective laboratory reports. Therefore, the activities represent a blend of those that are strategic in their value and those that aim to develop more holistic skills in scientific literacy.

Central to the success of the project has been the development of activities that are interactive and a variety of software applications were utilised to do this. The basis of the activity design uses WIMBA create (an add-in to Microsoft word) which provides a straightforward way to write interactive content using a variety of question styles with feedback. Google docs was used to write a subject specific language questionnaire which enables the author to record and analyse the student responses. Camtasia was used to produce tutorial screencasts to guide the student through the

activity. The content was further enriched through the embedding of material such as links to the FOCUS concordancing tool, relevant articles, and extension material (see Fig.6.).

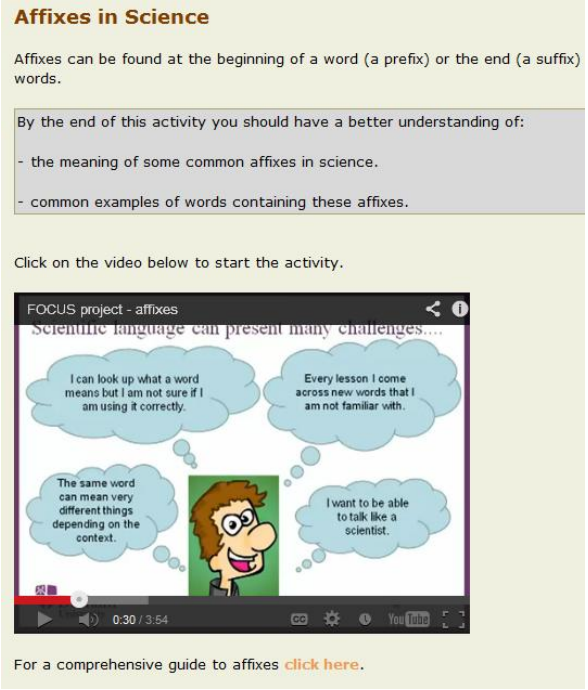

Figure 6: Screen capture of one of the activities with embedded video tutorial

In class questionnaires were completed and produced very positive feedback from the students (see fig. 7).

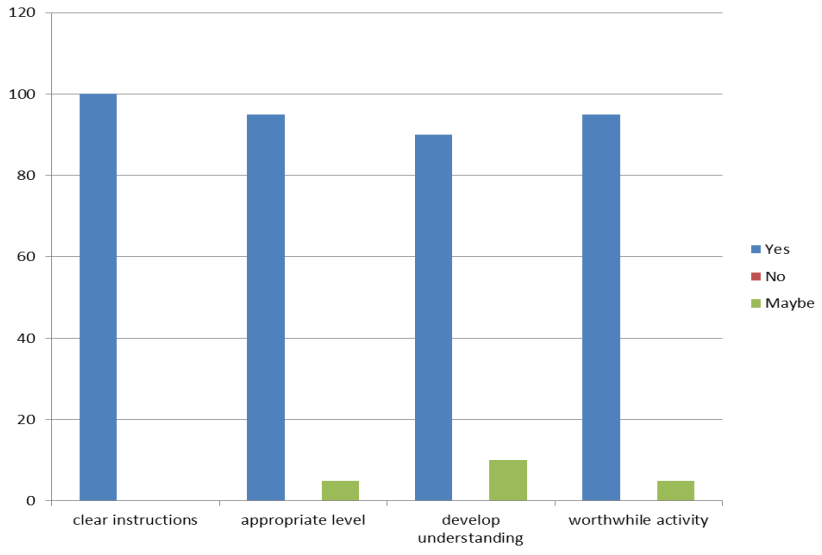

Figure 7: Results of an in class questionnaire about the activities.

This was supported by the focus group who felt that the design of the activities was clear and well organised with an appealing blend of different forms of media. They felt that the activities would be useful for the next student cohort.

The development of the FOCUS concordancing tool has provided the opportunity to access a large body of student work to enhance student understanding of subject specific language. However, its presentation as a stand-alone tool is not sufficient to engage with students. It can be difficult for a non-specialist to appreciate the value of the tool and how it can impact on learning. The development of these activities however, addresses this issue because it demonstrates how structured learning activities can be designed which are enriched by the corpus content. These activities can be used to provide meaningful subject specific language support within departments.

The next stage of the FOCUS project

There are three important stages to address next in the development of this project. The first stage is to continue corpus-building: to obtain more texts from more departments so that we continue working towards our aim of having texts from all the departments to which Foundation students progress.

The second stage is to continue to develop teaching activities to allow students to learn from the corpus data both before and during their Foundation studies. Related to this is the third stage of the project: to develop a diagnostic language assessment based on FOCUS data which will allow us to provide more tailored language support to our students on arrival in the department.

REFERENCES

Alsop, S. & Nesi, H. (2009). 'Issues in the development of the British Academic Written English (BAWE) Corpus'. *Corpora 4*(1), 71-83.

Berkenkotter, C., Hukin, T. & Ackerman, J. (1991). Social context and socially constructed texts: The initiation of a graduate student into a writing research community. In C. Bazerman & J.Paradis (Eds.) *Textual dynamics of the professions* (pp. 191-215). Madison, WI: The University of Wisconsin Press.

Cobb, T. (1997). 'Is there any measurable learning from hands-on concordancing?' In *System 25*(3), 301-315.

Coxhead, A . (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Drury, H. & Webb, C. (1991). *Literacy at tertiary level: Making explicit the writing requirements of a new culture.* Paper presented at the Inaugural Systematic Linguistics Conference, Deakin University.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System, 21*, 231-244.

Freedman, A. (1987). 'Learning to write again: Discipline specific writing at university'. *Carleton Papers in Applied Language Studies*, *4*, 45-65.

Hyland, K. & Tse, P. (2007). 'Is there an Academic Vocabulary?' *TESOL Quarterly 41*(2), 235-253.

Johns, T. F. (1991). 'Should you be persuaded: Two examples of data-driven learning'. In Johns, T. F. & P King (Eds) *Classroom Concordancing*. (Pp1-13). Birmingham: ELR.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York, NY: Heinle and Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York, NY: CUP.

November, N. & Day, K. (2012). 'Using Undergraduates' Digital Literacy Skills to Improve their Discipline-Specific Writing: A Dialogue.' *International Journal for the Scholarship of Learning and Teaching*, *6*(2), 1-21.

Schmitt, D. & Schmitt, N. (2005). *Focus on vocabulary: Mastering the Academic Word List.* London: Longman.

Rees, S & Bruce, M. (2012). "The development of online resources to enhance understanding of subject specific language in non-traditional students". Talk at Durham University Blackboard Users' Conference: January 2012.

Tribble C. (1997). 'Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching.' in Melia J. & B. Lewandowska -Tomaszczyk (ed.) PALC '97 Proceedings, Lodz: Lodz University Press. Retrieved from http://www.ctribble.co.uk/text/Palc.htm (Accessed 5th February 2013).

Trimble, L. (1985). *English for science and technology: a discourse approach*. Cambridge: CUP.

Wenger, E.C. (1998). *Communities of Practice: Learning, Meaning and Identity*. New York, NY; Cambridge: CUP.

Woodward-Kron, R. (2004). 'Discourse communities and writing apprenticeship: An investigation of these concepts in undergraduate Education students' writing. *Journal of English for Academic Purposes*, *3*, 139-161.

Worthington, D. & I. S. P. Nation (1996). 'Using texts to sequence the introduction of new vocabulary in an EAP course'. *RELC Journal*, *27*, 1-11.