

# Corpora and Language Learning with the Sketch Engine and SKELL

Adam Kilgarriff, Fredrik Marcowitz, Simon Smith, James Thomas

## Abstract

We introduce the idea of using corpora – the linguist’s name for ‘big data’ – in language research, and sketch its history, first in linguistics in general, then in language learning and teaching. We then take a careful look at the hazards of using corpora in language learning, and arrive at some maxims for when and how they have a place: firstly, *don’t scare the students*; then, use the corpus *when the dictionary does not tell you enough*, and moreover, *disguise the corpus as a dictionary*. We then introduce Sketch Engine, and show how it implements these ideas through SKELL, its language-learner interface. We show how corpora can be used, both in the classroom, and in the background, for syllabus design, where we have used corpora of learner output to identify patterns of overuse and underuse, with implications for what needs teaching.

## 1 Introduction and motivation

### History

Big data is big news. It is the big new thing in science: newspapers run supplements on what it means and how it will change the world, and politicians announce initiatives and inaugurate new research centres so that their country’s scientists will be leading the charge. So – how might that relate to language teaching?

Within linguistics, the ‘big data’ movement antedates the name ‘big data’ and is called, rather, *corpus* linguistics. A ‘corpus’ is a ‘body’ of data, and linguists call their big datasets ‘corpora’ (the latinized plural of ‘corpus’). A corpus is a collection of pieces of language, so-called when used for language research, and it may be anything from newspaper articles, to transcripts of everyday conversations or chat shows or lectures, to novels or letters or advertising brochures or shopping lists.

Back in the 1980s, lexicographers and dictionary publishers were making the case for big data long before it was a familiar one. They knew, as James Murray, editor of the Oxford English Dictionary from 1879 to his death in 1915, had before them, that languages are big: for dictionaries to get good coverage of a language, and not to make embarrassing omissions, they need very, very large amounts of data, and tools to support finding all the words and phrases in them. James Murray resorted to an army of volunteers to gather twenty million ‘slips’ – examples of words in use, with a sentence written out, with details of where it had been found, on an index card - which were then filed under the word being exemplified, so when he started work on the word, he would go to the filing system, find all the slips for it, and use them as the basis for the entry.

By the 1980s, it was evident that computers could assist and streamline this process enormously. In the UK, this technological development coincided with a

commercial one: English Language Teaching was blossoming worldwide, and there was money to be made in dictionaries for the ELT market. The Oxford Advanced Learners Dictionary had, until then, dominated the field, but others were keen to challenge Oxford for a share of the action. In particular, Collins, who found an academic partner in John Sinclair, professor at Birmingham University and already arguing vigorously for corpus methods. Between these two the COBUILD (Collins Birmingham University International Language Database) project was established. It went on to revolutionize dictionary-making,<sup>1</sup> showing that dictionaries could give a more accurate and fuller account of a language if they were based on a corpus.

In 1987, when the COBUILD dictionary was published, Collins were leading the field. By this stage Longman and Chambers also wanted to compete in the EFL dictionaries market, and realized, as did Oxford, that they needed a corpus if they were to keep up. Several Universities were also interested, in particular Lancaster, where Geoff Leech had been an early advocate of corpora in linguistics, for studying topics ranging from grammatical change to stylistics. Out of this alliance the British National Corpus project was born. The corpus, an unimaginably vast (by the standards of the time) 100 million words, of which 10% were transcribed everyday speech, published in 1994, was the big data of its day. It remains a landmark and model for how a corpus ought to be.

By the 1990s, others were starting to see the potential of corpora, in particular the computational linguists and the technology companies. The computational linguists (also called NLPers<sup>2</sup>) asked questions like “if we have a good grammar, shouldn’t it account for most of the sentences that we find in a corpus? Shouldn’t we be able to treat the grammar as a scientific hypothesis, and, if the hypothesis is good, shouldn’t we be able to write a program to find the grammatical structures of the sentences in a corpus?” The technologists were interested both because corpora provided samples of the language that they wanted to be able to handle for purposes of search, or spell-checking, or typesetting, or automatic speech transcription, and because the large corpora and high technological demands that they imposed were a challenge for the technical development of the computers. One innovative project from the early 1990s, HECTOR, was a joint venture between Oxford University Press and DEC, a leading Silicon Valley technology company of the time. DEC wanted to see if their hardware had the potential to index and display results from a very large corpus fast enough, and across enough monitors to meet the lexicographers’ needs.

### **The TALC movement**

Naturally, language teachers who came across the excitement generated by corpora in linguistics and dictionary-publishing asked “can I use this in my

---

<sup>1</sup> *At least for English: at this point let us apologize for the anglocentric flavor of this version of the history. There is also a long history of corpus use in lexicography for Dutch, German, French and Swedish, amongst others. The particular commercial situation of English has meant that, over time, English has often been the best-funded area for corpus work, with many advances happening first for English. Also the home of EFL has tended to be the UK rather than the USA, with British publishers dominating the market, so there is also some justification for the British-centric flavour.*

<sup>2</sup> *NLP is Natural Language Processing (where ‘natural’ is in contrast to ‘computer’).*

teaching". A leading early proponent of 'corpora in the classroom' was Tim Johns, an inspiring teacher also working at Birmingham University though in the language-teaching rather than the English department. He made the case for 'Data Driven Learning': for the student discovering the rules and patterns of the language, and their exceptions and limitations, through examining examples of the phenomenon they were looking at. He would bring printouts of concordances into the classroom – this was twenty years before there were computers in the classroom for the students to work on. Others were inspired, people wrote and distributed software to handle corpora and do the concordancing, and a community developed. In 1994, at Lancaster University, it had its first conference, 'Teaching and Language Corpora' (TALC). TALC has taken place every two years since, recently revisiting Lancaster for its twentieth anniversary.

### A delicate question

English language teaching (ELT) is a huge international business, employing 11 million people worldwide (British Council 2010) and aiming to meet the hopes and aspirations for English-language fluency of a large share of the world's young people. In that context, one might expect corpora, as a dynamic and innovative approach to getting students to grapple with the language, to grow and grow, and to become, itself, a big sub-area. But it has not. It has remained a tiny, specialist niche. Why is that?

First, we can distinguish two kinds of use of corpora: direct – in the classroom, with students looking at concordances – and indirect, with corpus use by people preparing dictionaries, syllabi, coursebooks and other teaching materials (Boulton 2009). The success of corpora in indirect use, starting with dictionaries, is clear to see and largely now beyond question. To give an accurate picture of a language, we need evidence of how the language is patterned. To know which phenomena are the common ones, we need language data. For this we need a corpus.<sup>3</sup>

Let us turn to direct use: 'corpora in the classroom'. It is here that we might have expected to find explosive growth, in the millions of ELT classrooms worldwide – but we have not.

Consider Figure 1: a basic concordance, for the English word, *belt*, from the British National Corpus, in a standard 'key word in context' form.

|  |              |   |
|--|--------------|---|
| tried to stand up quickly despite his safety | <b>belt</b>  | . We thought the bumpy flight must have         |
| form a world-wide network of mountain        | <b>belts</b> | far higher ( above ocean floor level ) and more |

---

<sup>3</sup> A study by Biber and Reppen (2002) showed that the tendency, in language courses, to teach the present continuous (the appropriate form of *be* + the *ing* form of the verb: "I am going...") very early, was based on a false assumption that it is a very common verb tense (in the kinds of language that students are likely to encounter). It turns out they will find the base form of the verb altogether more useful. Course materials are now being widely revised to follow this finding.

|   |               |   |
|---|---------------|---|
| should have been awarded a Lonsdale         | <b>belt</b>   | for his first-half haymaker on Francis . The    |
| the courage to go out and play and not just | <b>belt</b>   | the ball up field . ' Saunders , whose penalty  |
| were a relief Like heather flowers His      | <b>belt</b>   | could not endure the siege - it burst And       |
| Chile-Peru Trench . These steeply-inclined  | <b>belts</b>  | or zones of earthquake sites are known now      |
| development of ` brown ' areas of the Green | <b>Belt</b>   | . UNDERUSED AND SURPLUS PROPERTY IN THE         |
| to reconcile the irreconcilable but green   | <b>belt</b>   | is one of the most successful all-purpose tools |
| and put it into his pocket . He found the   | <b>belt</b>   | at the bottom of the bag . It was a brown       |
| which management regimes are best . Tree    | <b>belts</b>  | offer many benefits over conventional           |
| Nielsen , the augmented embonpoint , `      | <b>belted</b> | ' a ` fan ' who tried to give her a ` kiss ' ;  |
| of the width of the contemporary climatic   | <b>belts</b>  | . But it can hardly be argued that either       |

Figure 1: Concordance of *belt* (lemmatized) from the British National Corpus. Random sample of twelve lines.

There are several things to note.

- Each line is a fragment of text. It is not a full sentence.
- Each line is from a different text, which is not an authentic experience of language at work.
- Each line has been stripped from its context. If the student wants to look further for more context, to gather more clues to interpretation, they have more work to do.

Thus each line is not in any sense a self-contained piece of language. Let us now look at the lines one by one.

- the first shows a standard compound for *belt*, *safety belt*
- the second line shows a rarer, more specialized compound, possibly familiar only to geologists and similar: *mountain belt*
- the third line alludes to belts as indications of levels of achievement in some sports, particularly oriental martial arts such as judo and karate, where the practitioner progresses from a beginner's white belt, via a range of colours, to the champion's black belt. This is in fact a report on a rugby match where the journalist is adding colour with a joke from a different sport.
- the fourth is a verbal, informal use: to belt something can mean 'to hit it hard', with the etymology based on schoolchildren being hit with a belt as punishment for bad behaviour.
- the fifth is from a poem.
- the sixth, like the second, is geological.
- in the seventh and eighth, the word is part of a technical term, probably only in use in the UK, related to town planning policies, where it is often thought desirable to leave a 'green belt' of agricultural or otherwise 'green' land around towns and cities, and there is UK legislation to make this happen. One of these instances is capitalized, the other is not. In addition a further half of the first line is fully capitalized.
- the ninth is (finally) a prototypical and straightforward belt:

- a narrow piece of leather, cloth etc that you wear around your waist, for example to keep your clothes in place or for decoration<sup>4</sup>
- the tenth is a technical term from forestry.
- the eleventh is in quotation marks and is another use of the verbal, informal sense where 'belt' means 'hit'.
- The twelfth is a technical term (which may also be seen as metaphorical) from climatology.

From a lexicographer's point of view this is wonderful. In just twelve corpus lines, we have already found nine meanings, most of which need covering in any good dictionary (depending on the size and scope of the dictionary). But from a language learning perspective, it is alarming. Students could not recognize these different meanings based on one meeting with each, and teachers would not encourage them to jump to any conclusion without more evidence. Looking at these lines is not an efficient route to understanding the straightforward English word *belt*. If we are to succeed in bringing concordances into the classroom, a central challenge is how we do so without scaring the students.

A review of the TALC literature supports this central finding. Time and again, we find it is only the advanced and motivated students who engage and benefit. Gao (2011) finds that corpus texts "are difficult to understand and use for language learners". Boulton (2009) finds that "the main difficulty lay not in the DDL approach or in the KWIC presentation, but rather in the difficulty of using authentic language." "Using" is the thrust of what follows. It is worth noting that not every native speaker understands every concordance line.

### **When to turn to the corpus**

To find out about a word, like English *belt*, the place to look is the dictionary. The dictionary is designed to give the user the information they need. For learners of English in particular, a number of very high quality monolingual learner dictionaries have evolved to give the student a word's meaning(s), pronunciation, grammar, patterns of use, details of inflection, distributions according to genre such as *literary*, *informal*, and regions such as *US*, *UK*, *Aus*. In addition to the traditional book format, they are nowadays available through internet and mobile phone, with additional searching functionality as well as pronunciation. On the phone, they are always at hand.

So when should a learner turn to the corpus? The short answer is: when the dictionary does not tell us enough. Dictionaries, even online ones, are limited in how much they can say. If entries are too long, information will be hard to find, and users will struggle to find the information they seek. There are various strategies being explored by dictionary-makers to address the issue (see, for example Serge Verlinde's 'Base lexicale du français' (Verlinde 2010)). Nonetheless, there will always be occasions where the quandary the user wants help with is not covered in the dictionary. There are simply too many cases that different users want help with. This is easier to see when the user is producing language (speaking or writing) rather than receiving it (listening or reading). If

---

<sup>4</sup> From Macmillan Dictionary, <http://www.macmillandictionary.com>, January 17, 2015.

the user is writing a chemistry essay and is aware that the appropriate verb for the chemical process is *coalesce*, but is not sure how the verb should be used, none of the main online dictionaries gives us a chemistry example: it is not such a common word, and gets brief treatment, with examples only in a couple of cases. As we see shortly, if the user looks to corpora they promptly see relevant cases which can be used as models for the sentence they are writing.

### **“But I want to learn English, not find out about corpora”**

We suspect, the most common reason why students have used corpora in the classroom is because their teacher is a corpus enthusiast.

This is not such a good reason, and the student might well say, “but I want to learn English (or French, or Chinese...), not to find out about corpora”. How might we overcome students’ natural resistance to finding out about something which is, on the face of it, a distraction from the language-learning task? Here is a possible response. Corpora and dictionaries are both language resources on a spectrum: a spectrum from the raw to the refined. The lexicographer takes the corpus evidence, and then analyses, filters, sorts and selects - and the end product is the dictionary entry.

Students don’t know what a corpus is, and maybe do not want to put effort into finding out. But they know what a dictionary is: they are taught how to use them, they regularly do use them, and they quite often even express affection for them.

Why not sidestep the question entirely, **not** introduce corpora, and present corpus evidence as if it was dictionary evidence. Let’s disguise the corpus as a dictionary.

To show how that might work, we introduce Sketch Engine and SKELL.

## **2 The Sketch Engine and SKELL**

The Sketch Engine<sup>5</sup> (Kilgarriff et al 2004) is a leading corpus tool which has been in use for lexicography and language research since 2004. It has two parts: one for exploring corpora, the other for building and managing corpora. For a partner paper to this one, introducing the Sketch Engine in full, see Kilgarriff et al. (2014).

Since its inception, it has also been used in various ways for language learning. However the constant feedback has been ‘it is too complicated; it puts the students off’. With this in mind, SKELL, ‘Sketch Engine for Language Learners’ has been designed as a stripped-down, non-scary version of Sketch Engine for use by learners, in keeping with the argument of this paper.

SKELL<sup>6</sup> is a language learning website in which all the reports are corpus-based, using fully automated methods, and are designed to avoid scaring the students. It is currently available only for English.

---

<sup>5</sup> <http://www.sketchengine.co.uk>

<sup>6</sup> <http://skell.sketchengine.co.uk>

It offers three reports: word sketch, examples and 'similar words'.

## 2.1 Word sketch

| object of catch           | subject of catch          | adjectives with catch      | modifiers of catch           | words and/or catch       |
|---------------------------|---------------------------|----------------------------|------------------------------|--------------------------|
| <a href="#">glimpse</a>   | <a href="#">angler</a>    | <a href="#">red-handed</a> | <a href="#">unawares</a>     | <a href="#">bowl</a>     |
| <a href="#">sight</a>     | <a href="#">surprise</a>  | <a href="#">unprepared</a> | <a href="#">freshly</a>      | <a href="#">punish</a>   |
| <a href="#">pass</a>      | <a href="#">fisherman</a> | <a href="#">alight</a>     | <a href="#">finally</a>      | <a href="#">imprison</a> |
| <a href="#">fish</a>      | <a href="#">ear</a>       | <a href="#">cold</a>       | <a href="#">behind</a>       | <a href="#">convict</a>  |
| <a href="#">breath</a>    | <a href="#">eye</a>       | <a href="#">most</a>       | <a href="#">up</a>           | <a href="#">chase</a>    |
| <a href="#">fire</a>      | <a href="#">camera</a>    | <a href="#">more</a>       | <a href="#">locally</a>      | <a href="#">throw</a>    |
| <a href="#">attention</a> | <a href="#">dryer</a>     | <a href="#">many</a>       | <a href="#">dead</a>         | <a href="#">trap</a>     |
| <a href="#">eye</a>       | <a href="#">peloton</a>   | <a href="#">early</a>      | <a href="#">accidentally</a> | <a href="#">execute</a>  |
| <a href="#">ball</a>      | <a href="#">Anyone</a>    | <a href="#">several</a>    | <a href="#">eventually</a>   | <a href="#">eat</a>      |
| <a href="#">hold</a>      | <a href="#">fielder</a>   |                            | <a href="#">suddenly</a>     | <a href="#">kill</a>     |
| <a href="#">cold</a>      | <a href="#">fish</a>      |                            | <a href="#">quickly</a>      | <a href="#">release</a>  |
| <a href="#">train</a>     | <a href="#">police</a>    |                            | <a href="#">soon</a>         | <a href="#">arrest</a>   |
| <a href="#">off-guard</a> | <a href="#">touchdown</a> |                            | <a href="#">unaware</a>      | <a href="#">try</a>      |
| <a href="#">prey</a>      | <a href="#">wind</a>      |                            | <a href="#">wild</a>         | <a href="#">hang</a>     |
| <a href="#">trout</a>     | <a href="#">anyone</a>    |                            | <a href="#">short</a>        | <a href="#">shoot</a>    |

Figure 2: SKELL word sketch for English *catch* (verb)

The function that gives the Sketch Engine its name is the word sketch: a one-page summary of a word's grammatical and collocational behaviour (Figure 2).

This is a feast of information on the word. For *catch (verb)* just looking at the first column (objects of the verb) we immediately see a number of meanings, idioms and set phrases. *We catch a glimpse of or catch sight of* something. Sportsmen and women, in a range of sports, *catch passes* and *balls*. Anglers and fisherman (column 2) *catch fish, prey* and *trout*. When surprised or shocked, one *catches one's breath*. Things *catch fire*. You often want to *catch someone's attention*. Things sometimes *catch your eye*. People *catch hold of* other people when they don't want to let them go. We all sometimes *catch a cold* and *catch trains*. In the single parsing error in this column, we are often *caught off-guard* by unexpected events. (*Off-guard* has been mis-classified as a noun, hence an object, rather than an adverb, hence a modifier.)

This is all of great value to learners. The COBUILD dictionary offers thirty meanings of *catch*, most of which are the verb. It lists the meanings in order of frequency, and, the first two are, in brief,

1. catch a person/animal (capture),
2. catch an object that is moving through the air.

These frequent and prototypical meanings have an equivalent in other languages e.g., for COBUILD sense 1, French *attraper*, German *fangen*, Chinese 抓 *zhuā*. Beyond this, most of the thirty meanings do not use the "catch" verb e.g. catch a bus: French - *prendre le bus*, German - *den Bus erwischen*, Chinese - 乘公共汽车 *chéng gōnggòngqìchē*. The word sketch for *catch* presents learners with clues to the multiple uses of the word, most of which manifest quite differently in their first language. This is valuable for language learning.

Moving on to the subjects column, *surprise* relates to the expression *caught by surprise*. *Ears* and *eyes* catch things when we hear or see; likewise *cameras*. *Dryers*, it turns out, frequently *catch fire*, and this is often reported, as the cause of a fire, in short local-news reports: we find this is the explanation for *dryer* being in the list by clicking on the word *dryer*: we are then shown the 'examples' report at Figure 3. In this way the underlying evidence is always available at a mouse-click. In all 40 examples of *dryer+catch*, it is always *caught*.

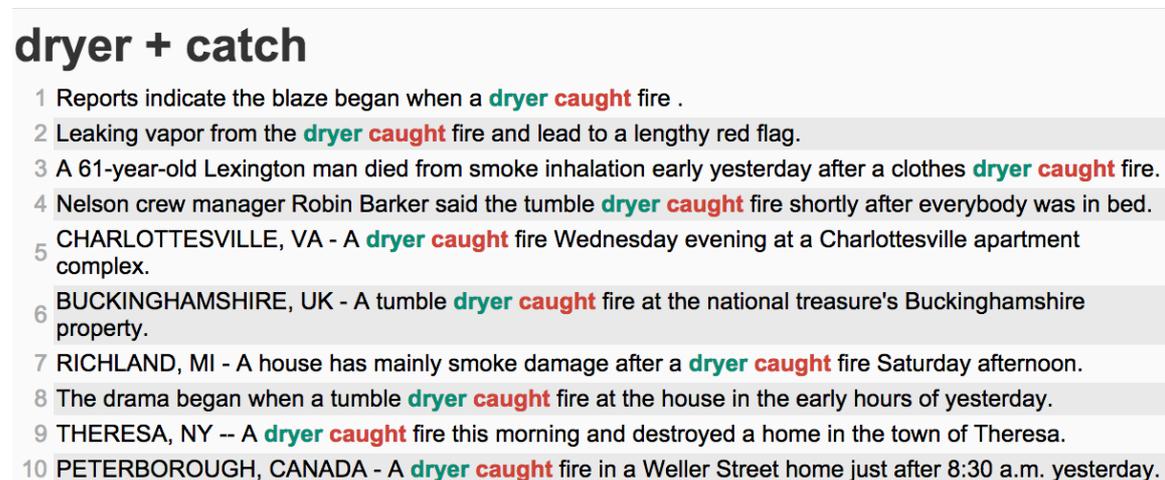


Figure 3: SKELL Examples report for *dryer* as object of *catch*

Turning back to the subjects list: *peloton* is a specialist term from the *Tour de France* cycling race, referring to the racer who is in the lead and wears the yellow jersey. They often *catch* the man in front. *Fielder* and *touchdown* are also sport terms, both from American football (with another minor mis-analysis: *touchdown catch* is a compound nominal rather than a subject-verb pair). *Anyone* and *Police* introduce a new meaning of the verb: police catch criminals. *Anyone* brings to our attention the related pattern *Anyone caught [doing X] will be [punished]*.

The remaining columns tell us more about the police meaning (*red-handed, unawares, punish, imprison, convict, chase, trap, execute, kill, release, arrest, try, hang, shoot*), the sports/hunting/fishing meaning (*locally, wild, bowl, throw, eat*) and several other idioms.

For professional and budding lexicographers, the word sketch is a draft dictionary entry. The system has worked its way through the corpus to find all the recurring patterns for the word and has organized them – not by meaning, as

a dictionary does: that is too ambitious to do automatically – but at least by grammar, which is some help. As noted above, this is not a report for a beginner learner: all the information that the beginner needs will be provided in a good dictionary. It is for intermediate and advanced learners looking for information they could not find in the dictionary. We find it meets that need well.

## 2.2 Examples

On the spectrum from corpus to dictionary, the Examples report is closer to the corpus end. It presents example sentences for the search term. However, in contrast to the concordance for *belt* in Figure 1:

- each instance is a full sentence, not a series of fragments
- the sentences are chosen as ‘good’ examples (insofar as this can be done automatically).

The algorithm for choosing good examples is called GDEX (Kilgarriff et al 2008) and works as follows:

- if there is a word sketch for the search term, use that as a starting point and show the best example sentence for each collocation in the word sketch. This will mean that all examples will exemplify a common collocation for the word.
- ‘score’ each sentence containing the search term, and show the user the highest-scoring sentences. Factors in the scoring algorithm include:
  - sentence length: not too long (so there is too much for the user to struggle through), nor too short (so there is not enough context to be helpful)
  - it is a sentence, starting with a capital letter and ending with a full stop, ! or ?
  - no or not too many rare or unrecognized words. This constraint tends to rule out sentences with spelling errors
  - mostly comprising common words
  - not too many non-letter characters: a large number of numbers, punctuation marks and other non-letter characters tend to indicate a non-standard sentence
  - not too many capital letters: a large number suggests names and acronyms, which often indicate that the user will need domain-specific knowledge to understand the sentence
  - context-free-ness: ideally we would like to promote sentences which stand alone, rather than needing context to be comprehensible. A sentence like “So they decided it was not worth it” is problematic because the reader has no context to help them understand what *they*, *it* or *it* refers to. It was hard to implement this constraint, and we are unsure if we have been at all successful.

One of the first questions that a corpus linguist will ask about SKELL is: “what corpus do you use?” This is not a question which SKELL immediately presents an answer to. Its users are language learners, not corpus linguists, and we do not expect them to know what a corpus is, let alone to have detailed ideas about what corpus might be good for what task. The learner takes it on trust that a



converting receptive (passive) vocabulary into vocabulary available for productive (active) use.

### 3.1 *germane*

The word *germane* does not appear in lists for foreign learners. We often meet words in our reading and conversations that pique our interest, but since "authenticity does not automatically entail typicality" (Gries 2008), it is necessary to look beyond that single first encounter.

So let's say a learner wants to know more about this rare bird. The COBUILD dictionary tells us

Something that is **germane** to a situation or idea is connected with it in an important way; a formal word"

and gives us two examples and a grammar code

ADJ QUALIT: PRED +*to*.

This is probably not enough information to make it available for confident, active use. The curious student, having direct access to the data, finds concordances of *germane* in SKELL, as shown in Figure 5.

- 1 That is a much more **germane** analogy.
- 2 Du Bois addressed several problems **germane** to black existential philosophy.
- 3 These things are true but not **germane** .
- 4 This makes transpersonal art criticism **germane** to mystical approaches to creativity.
- 5 It selects what is **germane** , pertinent, and related.
- 6 But to revert to matters more **germane** to the lakes.
- 7 The entire body took a vote on whether the amendment was **germane** .
- 8 Regarding CEO pay ... This is a **germane** argument.
- 9 He is in the business, so his comment is **germane** .
- 10 All of section 407.640 seems **germane** to the original reference.
- 11 Not all disclosures are **germane** to the ICWPA.
- 12 The two inquiries were so **germane** that they helped him reciprocally.
- 13 Such a principle of deep ecology is therefore **germane** to indigenous Celtic spirituality.
- 14 The second aspect is **germane** to all plans: the effect on utilization.
- 15 Stories are stories, and their relative proximity to reality is not **germane** .
- 16 Such an act he deemed entirely **germane** to Zoraida's dark methods.
- 17 We have identified three types of instructional resources **germane** to implementing Modeling Instruction.
- 18 Its a popular meme , but not all that **germane** .
- 19 As such, my observation was only partially **germane** to your present Article.
- 20 His religious background was complicated but **germane** to his marginal status in German academia.

Figure 5. SKELL concordance for *germane*.

It is at the next stage that the role of the teacher is critical. The learner cannot be expected to know what linguistic features to look for, nor which ones are *germane* to the question, at least, not until they have been trained in the requisite metacognitive strategies. Johns recommends the teacher taking on the role of research organizer. (Johns 1991b: 31). This sees the teacher proffering a series of leading questions that the learner can answer from the data:

- Is it typically followed by *to*?

- When it is, what precedes it?
- Does it follow the nouns it qualifies?
- When does it occur at the end of information units?
- Are there any typical nouns or semantic types whose company it keeps?
- Is it used mostly in formal contexts?

The teacher can also create a learning activity using this data: select some sentences and ask if *germane* could be replaced by any similar words. The learners arrive not only at semi pre-constructed phrases for *germane*, but a host of other features. Many mental processes are happening during such work. The student is learning language, about language and about using data.

### 3.2 "the problem lies"

Producing language requires a productive knowledge of vocabulary. Just as we are about to use the word *problem*, we find that we do not know the collocating verb that expresses the meaning we have in our first language. Here we consider the case where the Czech verb would be *spočívát*, for which a bilingual dictionary offers ten or so English equivalents. The word sketch provides these fifteen verb collocates with *problem* as subject:

solve arise face stem occur plaque lie persist confront exist affect beset result concern relate

Only one, *lie*, means something similar to the Czech verb. The learner clicks on *lie* in the word sketch and finds example sentences as in Figure 3, which confirm that they have found the right verb.

- 1 But here is where the **problem lies** .
- 2 This is where the real **problem lies** .
- 3 The second half is primarily where the **problem lies** .
- 4 This is where the real **problems lie** .
- 5 The actual **problem lies** elsewhere and are complex .
- 6 **Problems lie** somewhere between puzzles and policy issues.
- 7 The **problem lies** in their very nature.
- 8 She said the **problem lies** with previous owners.
- 9 The backs were mainly pedestrian but the fundamental **problem lay** elsewhere.
- 10 That is not where the **problem lies** .

Figure 6. SKELL concordance for *problem lie*.

The productive use of the collocation may require a little more guidance – it is not enough to know *problem lie*. The teacher may further guide the student with the questions

- what tense is the verb in?
- what article is used?
- where in the sentence is the expression?

The curious student might even wonder what other abstractions *lie*? An answer to this question could lead to seeing this use of *lie* as a pattern, not just an

idiosyncratic collocation. Such an observation triggers a new appreciation of how words are combined to create text that sounds idiomatic and native-speaker like. The nouns given in the word sketch for *lie*, in the 'subject' column, are

answer fault snow strength body future island blame difficulty land danger village loyalty  
interest problem

By asking the students to divide this list into concrete and abstract nouns, the curious students have answered their question. To develop their understanding, they click on the abstract nouns to see what follows *lie* in these sentences. One that they see is

The solution to this **problem lies not in** legislation but in effective screening of potential surrogate mothers.

To observe the patterned nature of this, we type *lie not in* into the Examples search field and get Figure 7.

- 1 The real danger may **lie not in** Ukraine but farther west.
- 2 The Lachesis is its strength **lies not in** public.
- 3 According Agoncillo and Palma, his interest **lies not in** politics.
- 4 The fault **lies not in** our ties, but in our selves.
- 5 The difficulty **lay not in** identifying the issues but in tackling them resolutely.
- 6 Salvation **lies not in** a returning Jew or damnation in its opposition.
- 7 A text's unity **lies not in** its origin but in its destination.
- 8 The solution to the problem **lies not in** axing international or European fixtures.
- 9 Its economic base **lay not in** manufacturing but in commerce, contracting and land.
- 10 True happiness **lies not in** wanting great things or even in achieving your dreams.

Figure 7. SKELL concordance for *lie not in*.

Many abstract nouns are the subject of *lie*, and *not* predicts a following *but*. These concordances reveal a pattern  
the X<sup>abstr</sup> lies not in Y, but in ...  
which the learner may use in a sentence they write.

It may be objected that no-one has time for these shenanigans in class, as there is a strict syllabus to be followed. This makes it all the more important for teachers to appreciate how many learning experiences are happening at the same time. All being well, the teachers have inducted their students into the basics, and the students can practice discovery tasks for homework.

Teachers set tasks for students so that students become familiar with the many kinds of question that SKELL data can answer. Without this guidance, students will find little of interest. But once they have a framework, they become able to ask interesting and pertinent questions, and answer them.

#### 4 Vocabulary development for Business English

We now present a study of university students learning English in Taiwan and the UK, and show how learners may engage with a specialist corpus consultation task when they have built the corpus themselves.

Tyne (2009) and Charles (2012) argue that the process of creating a corpus gives the learner ownership of the corpus and thereby motivates them to explore it. This is especially true when the topic of the corpus is of personal interest to the learner, or coincides with their major field of study. Learners may pursue language study for only a short period of their university career, but once the corpus is constructed, students may be sufficiently motivated to consult it and add to it when needed.

Similar work includes Tyne (2009) who reports on a course in which British students were asked to create French corpora based on spoken data; Seidlhofer (2002) who described the use of a collaborative learner corpus in her class of trainee English teachers, making the students' own work the "primary objects of analysis" (p. 217); and Castagnoli (2006).

Castagnoli had translation trainees use the BootCaT toolkit (Baroni & Bernardini, 2004) to generate web corpora on specific topics, and extract lists of terms which could be used to compile glossaries and term databases. BootCat works as follows:

- the user decides on a domain
- they identify a set of around six 'seed' words or phrases which are specific to the domain: 'seeds', because they will be used to 'grow' the corpus
- the seed words are used, typically three at a time, to construct queries
- the queries are sent to a search engine (such as Google)
- the top pages (typically ten) that the search engine finds are downloaded the pages are cleaned and filtered to remove noise (such as non-text, foreign text, advertisements and duplicates)

Castagnoli's students found that a larger number of relevant terms could be extracted when the domain was highly specialized. By way of assessment, the students were given a technical translation task, and were asked to prepare for it by building a web corpus in the relevant domain, and extracting from it a glossary of terms.

#### **4.1 The first study: National Chengchi University, Taiwan**

In our first study, we applied Castagnoli's (2006) approach with non-major English learners in a Taiwanese university, using the Sketch Engine implementation of BootCaT, WebBootCat (WBC; Baroni, Kilgarriff, Pomikálek, and Rychlý 2006). Because it is included in the Sketch Engine web platform, the full range of corpus query tools described in Section 2 is available to analyse user-constructed corpora, as well as the pre-loaded corpora provided on the platform.

In this study, 19 Taiwanese learners of English used WBC to create a web corpus based on an area of interest by selecting appropriate seed words. They were asked to complete a number of tasks, including (1) comparison of patterns found

in their corpus with those in a general corpus (using word sketches and concordancing), (2) extraction and inspection of terms from their corpus, , and (3) evaluating the terms as potential seed words to be used to build a larger corpus in the same domain, in the process that gives BootCaT its name: 'Bootstrapping' (i.e. defying physics by picking itself up by its own bootstraps) Corpora and Terms'.

WebBootCat and Sketch Engine instruction was delivered by means of mini-lectures (10 to 15 minutes at a time, of a two-hour General English class). Students were given an introduction to corpora and concordancing, with some examples taken from Chinese corpora in an effort to make the material more accessible. The value of corpora as a source of authentic English was emphasized, as was the importance of learning from context and collocation, as opposed to memorizing English vocabulary and Chinese translations. To this end, the students were asked to explore the meaning and usage of unfamiliar words in their regular reading assignments, by studying common collocations in Sketch Engine word sketches. Large corpora such as the British National Corpus and the web corpus ukWaC (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008) were used for this purpose.

The sequence of teaching and learning procedures is set out in Table 1.

| Step | What   | When                              | Notes   |
|------|--|-----------------------------------|---|
| 1.   | Five 10-15 minute mini-lectures on corpora and concordancing | Weeks 2-6 of the course.          | One conducted during each 2 hour English class  |
| 2.   | General exposure to corpora                                  | Throughout the course             | Students encouraged to use concordancing during class preparation and other reading.  |
| 3.   | Preliminary WebBootCat workshop                              | Week 8                            | Students practise using WBC, and create a small, on-topic web corpus.   |
| 4.   | Corpus comparison task                                       | Week 10                           | Students create a new on-topic corpus. Compare output (word sketches) from it with output from a large general corpus, such as BNC.                                   |
| 5.   | Final project  | Completed by Week 18 (final week) | Students compile a new corpus based on their own subject specialism. They answer questions about the corpus, and demonstrate understanding of corpora and WBC itself. |

Table 1. Teaching and learning steps

Until Task 5, it was left up to students to choose their own corpus topic, and this led to some novel and sometimes bizarre selections: one student named her corpus "Haha" and used seed words *funny, giggle, laugh, smile*. Other corpora built included "cake", "chocolate", "ghosts", and the somewhat unsettling

“killing”. Not all students were comfortable choosing the keywords which represent a domain, or even selecting a plausible domain for investigation. Although students had been told that more specialized domains would yield better key term lists and richer corpora, we did not predict just how *un-specialized* some students’ domains would turn out to be.

In the final project, students were required to construct and consult a corpus relating to their own academic discipline. This task yielded more fruitful results in terms of student analysis and commentary, with one student, for example, commenting “Creating a specialized corpus could be useful when it comes to researching a particular subject or learning a subject in English. It is useful because of the different results which are much more relevant than searching on a much more general English corpus.”

#### **4.2 The second study: Coventry University**

The first study had established that learner corpus construction based on academic major was more successful than construction of corpora for less well-defined domains. In the second study, the students were all enrolled on the same academic course, Accounting and Finance for International Business (AFIB). A group of six AFIB students, all from China, undertook corpus construction as part of an English for Academic Purposes (EAP) class at Coventry University. There were two women and four men.

The work was conducted over a period of four teaching weeks (2 hours per week). In the first two lessons, an introduction to the use of corpora and the reading of concordance lines was given, with students looking at academic writing samples from the BAWE corpus.<sup>7</sup> In weeks 3 and 4, students constructed and consulted their own corpora.

The corpora were seeded from lecture PowerPoints, seminar discussion notes and other materials provided by teachers in the AFIB department, and not from user-selected (and, as noted above, sometimes arbitrary) keywords. Figure 8 shows a typical lecture PowerPoint.

---

<sup>7</sup> **British Academic Written English Corpus** : <http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/>

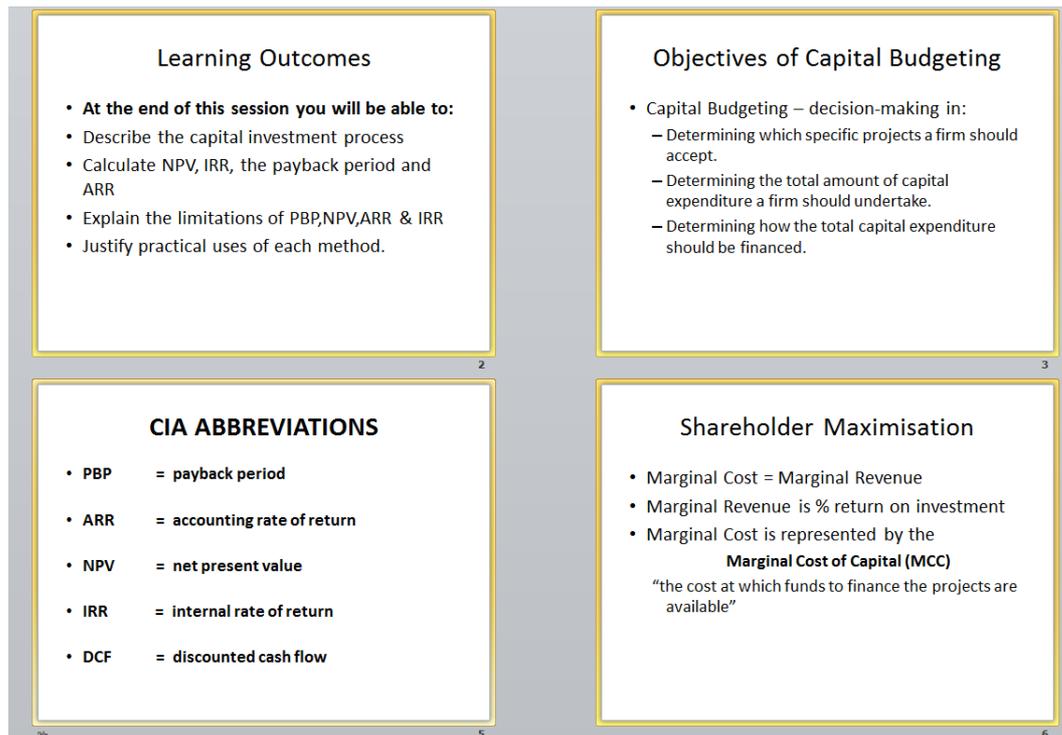


Figure 8: Management Accounting lecture slides

Sketch Engine can be used to upload text from a variety of document types to create and add to a corpus. In most UK universities, students are given online access to teaching and learning materials via a Virtual Learning Environment, and Coventry University, like many institutions, uses Moodle for this purpose.

First, the user uploads the text content of teaching materials to form a mini-corpus. Because of the nature of lecture slides, the resulting corpus does not contain many full sentences, but it will include the key vocabulary for the particular topic. Students could opt to create a more specialized corpus, consisting of perhaps just one PowerPoint, for example on “Capital Investment Appraisal”, to which two lectures were devoted. Alternatively they might decide to create a whole-module corpus, such as “Management Accounting for Business Decisions”.

The Sketch Engine is then used to generate a list of the most salient words in the corpus: the words found most frequently in this corpus, compared with their frequency in a reference corpus: the word *the* is not salient, because it is found with equal relative frequency in both specialist and reference corpora. These words are then used to 'bootcat' a much larger corpus, consisting of texts from the web. Figure 9 illustrates this process.

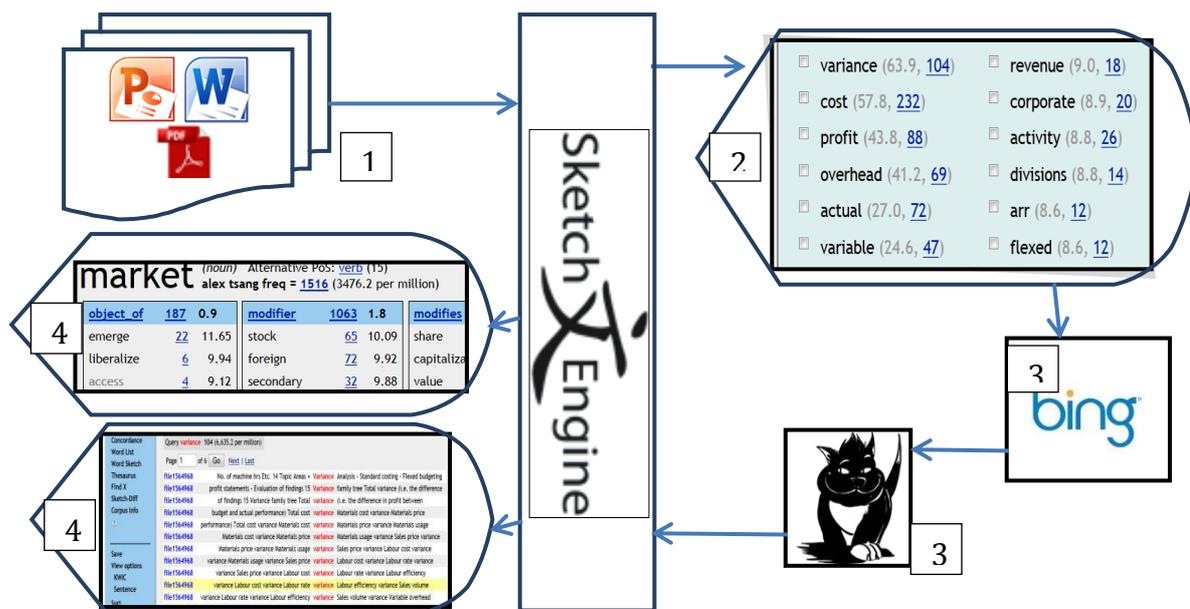


Figure 9: Corpus construction and consultation. Key: 1. Text input. 2. Wordlist from mini-corpus. 3. Bing API interacts with BootCat. 4. Word sketch and concordance displays from web corpus.

The large corpus can then be used in the following ways

1. To produce lists of subject area words and terms for study
2. To view word sketches, which give a one-page view of the collocations and grammatical structures in which a word or term participates.
3. To view the words and terms in context, using concordancing
4. To link back to the original texts on the web.

Although the small number of participants was a clear limitation of the study, feedback was positive. Representative comments included:

*The work is useful for my AFIB study. Because the software lists the word which I do not understand very clearly. I can learn the speech and the meaning of this word.*

In the current academic year, a third study is being conducted. This study includes four EAP classes, totaling 90 students. Pre-tests have been conducted, with post-tests scheduled; two of the classes are being taught in computer labs, and students are actively engaging in in-class corpus construction and analysis, while two control group classes are being given regular vocabulary lists for private study, extracted from corpora constructed in the same way by the researcher. The results of this experiment will be published elsewhere in due course.

## 5 Error analysis on Business English essays by Chinese adult learners

The data for this study comprised 40 Human Resource Management coursework essays, written in English by Chinese learners studying International Business at Guangdong University of Foreign Studies. The aim was to identify which aspects of English grammar students had the most difficulty with when writing for business.

### Methodology

The essays were prepared as a corpus in Sketch Engine. Sketch Engine was used as it had a number of features that made compiling and analysing a corpus efficient and effective, including an automatic part of speech (POS) tagger, and search functions that enabled the user to search for different grammatical constructions. It also allowed the user to add Microsoft Word documents to the corpus, which was beneficial since all coursework assignments had originally been submitted as Word files.

Prior to being compiled into the corpus, all essays were checked for evidence of plagiarism and the use of translation software. The page numbers, headers, reference lists, and front covers were removed from every essay to prevent these from affecting the results of the study. References and quotations within the essays were also deleted and replaced with the symbols [R] and [\*] respectively, an approach undertaken in the creation of the International Corpus of Learner English (Nesselhauf 2005: 45), to make the corpus as authentic as possible; any quotations copied were not examples of learner output and may have distorted the study's results. Any references that were part of a paraphrased sentence were left within the text, because it would have otherwise involved the deletion of some of the learner output as well.

In total eight grammatical categories were investigated. They were: articles, subject-verb agreement, copula verbs, quantifier-noun combinations, tense, verb forms, prepositions, and adjectives of emotion and state. These categories were chosen following past studies including Chuang and Nesi (2006), Gessling (2010), Shuang and Shang (2010), and Zheng and Park (2013).

A set of searches was developed in order to identify the number of times the different grammatical components had been used both correctly, and incorrectly. It was important to identify the number of times the learners had correctly used a component, as certain features such as articles are more commonly used compared to others than others like quantifiers. Examples of the searches are as below.

#### *Articles*

| Search Parameters  | Reason  |
|--|---|
| Search for the word - "a"<br>Search for the word - "an"<br>Search for the word - "the" | identifies every instance when "a"/"an"/"the" was used, which helps to identify every instance in which "a"/"an"/"the" was used correctly, and incorrectly. |

|                                      |  |
|--------------------------------------|--|
| <code>[tag!="DT"][tag="NN.*"]</code> | identifies every instance when a noun was used without an article preceding it, which helped to identify every instance in which an article was incorrectly omitted, or where the zero article was used correctly. |
| <code>[tag!="DT"][tag="NP"]</code>   | all noun phrases that do not have an article in front of them.   |
| <code>[tag="DT"][tag="NP"]</code>    | all noun phrases that have an article in front of them.  |

### Verb Forms

In order to identify every correct and incorrect verb form, each type of verb was searched for individually, and the correct and incorrect usages calculated.

| Search Parameters   | Reason  |
|---|---|
| <code>[tag="VVG"]</code>  | Shows all instances of the continuous verb form.      |
| <code>[tag="VVP"]</code><br><code>[tag="VVZ"]</code><br><code>[tag="VV"]</code> | Shows all instances of the present verb form.         |
| <code>[tag="VN"]</code>   | Shows all instances of the past participle verb form. |
| <code>[tag="VVD"]</code>  | Shows all instances of the past verb form.            |

Tense, subject-verb agreement, and prepositions were identified in similar ways. Copula verbs, quantifiers, and adjectives of state and emotion were found through the word searches.

### Findings

As hypothesized at the outset, learners had the greatest difficulties in the categories of articles, preposition, and subject-verb agreement: see Table 2.

| Grammatical Category                   | ✓                | X            | Total Instances |
|--|------------------|--------------|-----------------|
| <b>Articles</b>                        | 5680<br>(88.20%) | 760 (11.80%) | 6440 (100%)     |
| <b>Subject-Verb Agreement</b>          | 482 (78.22%)     | 135 (21.88%) | 617 (100%)      |
| <b>Copula Verbs</b>                    | 828 (96.73%)     | 28 (3.27%)   | 856 (100%)      |
| <b>Tense</b>                           | 1767<br>(96.40%) | 66 (3.60%)   | 1833 (100%)     |
| <b>Verb Form</b>                       | 3782<br>(95.61%) | 174 (4.39%)  | 3956 (100%)     |
| <b>Quantifiers<br/>(Much/Many)</b>     | 86 (81.38%)      | 21 (19.62%)  | 107 (100%)      |
| <b>Adjectives of<br/>State/Emotion</b> | 10 (50%)         | 10 (50%)     | 20 (100%)       |
| <b>Prepositions</b>                    | 3033<br>(91.49%) | 286 (8.61%)  | 3319 (100%)     |

Table 2: Frequencies for correct and incorrect uses of grammatical patterns

Article errors were predominant in raw frequency, with 760 in total. When compared to the number of times the learners used the article system correctly however, it could be seen that the article system had been correctly applied to the work 88.20% of the time. Despite only 11.80% of the articles having been incorrectly used, such a frequently used grammatical feature deserved more attention. Whilst the articles "the", "a", and "an" had been used correctly 96.15%, 95.85%, and 90% of the time correctly, the zero article had been used correctly only 78.25%, making up most of the article errors. Errors in which the zero article was misused can be seen in:-

[a] \* *Restaurant and company operate is the same*

[b] \* *Second problem is the health & safety*

It seems likely that learners insert an article only when they are certain they should be using one. Other areas of grammar that the learners had difficulty with were subject-verb agreement, and quantifiers, as can be seen in examples [c] and [d]:-

[c] *"Hire people is more important because it affect\*(s) their bonuses"*

[d] *"It conclude so many\* (much) knowledge, for example, what..."*

In [c], the difference between adding an -s, and not adding one does not affect coherence, and likewise the difference between *many* and *much* in [d]. Learners and teachers probably concentrate more on areas of grammar that affect coherence, so one hypothesis about the prevalence of these errors is that they are kinds of errors that have often gone uncorrected. A second reason why subject-verb agreement errors occur in the first place is that the learners' first language, Chinese, does not have subject-verb agreement (or indeed any inflectional morphemes) so the whole system is challenging for Chinese learners.

The proportion of errors within the other grammatical categories when compared to correct usages was low, with the exception of adjectives of state/emotion. Prior to the investigation it had been hypothesized that there would be a large number of verb form and tense errors within the corpus since Chinese has no morphological system, and from the results of similar studies by Shuang and Shang (2010: 88), and Ning (2011). Upon closer investigation it was found that the learners had made use of the present simple tense 66.66% of the time, and the simple form of verbs 73.68% of the time. From this one could hypothesize that the students relied upon the simple tense and simple form of verbs as an avoidance strategy as they knew how to use these, but may not have been confident in using other tenses and verb forms and were not willing to take the risk of making an error.

#### 4 Conclusion

The messages to take away from this paper are:

- indirect use of corpora in language teaching is clearly a good thing

- regarding direct use of corpora in the classroom:
  - o **don't scare the students**
  - o corpora are for where *the dictionary does not tell you enough*
  - o give some thought to **disguising** the corpus as a dictionary.

We have shown how one tool, the Sketch Engine (also in its SKELL variant) supports the use of corpora in language learning and teaching, and have given detailed examples of

- teaching lexis and collocation (as well as a set of learning strategies, for students to use outside the classroom) using SKELL
- building a corpus in the students' area of interest, and using it to identify the key vocabulary in that domain, for the students to focus on
- using student essays as a corpus, and identifying the grammatical patterns that they overuse and underuse.

## References

- Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of 4<sup>th</sup> International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 1313–1316. available from [http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat\\_lrec\\_2004.pdf](http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf) > [20th February 2015]
- Baisa, V. & Suchomel, V. (2014) SkELL: Web Interface for English Language Learning. In Eighth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2014. pp. 63–70. ISSN 2336-4289.
- Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006) WebBootCaT: instant domain-specific corpora to support human translators. *Proceedings, 11th Annual Conference of the European Association for Machine Translation Conference*. Oslo, Norway, 247-252. available from <http://www.mt-archive.info/EAMT-2006-Baroni.pdf> > [20th February 2015]
- Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching?. *Studies in Second Language Acquisition*, 24(02), 199-208.
- British Council (2010). Annual Report, 2009-2010.
- Castagnoli, S. (2006). Using the Web as a source of LSP corpora in the terminology classroom". In M. Baroni & S. Bernardini (Eds.), *Wacky! Working papers on the Web as corpus* (pp. 159-172). Bologna: Gedit. available from <http://wackybook.sslmit.unibo.it/pdfs/castagnoli.pdf> > [20th February 2015]
- Charles, M. (2012). Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31, 93-102.
- Chuang, F. Y., and Nesi, H. (2006) 'An Analysis of Formal Errors in a corpus of L2 English produced by Chinese students'. *Corpora* 1 (2), 251-271.
- Ferraresi, A., Zanchetta, E., Baroni, M. & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008*, Marrakech, Morocco. available from

- <<http://clic.cimec.unitn.it/marco/publications/lrec2008/lrec08-ukwac.pdf>> [20th February 2015]
- COBUILD (1987). Sinclair, J. (ed.). Collins Cobuild English Language Dictionary. London: HarperCollins.
- Gressang, J. E. (2010) *A frequency and error analysis of the use of determiners, the relationships between noun phrases, and the structure of discourse in English essays by native English writers and native Chinese, Taiwanese, and Korean learners of English as a Second language*. Unpublished PhD thesis. Iowa: University of Iowa, available from <<http://ir.uiowa.edu/etd/507>> [20th February 2015]
- Gries, S. (2008) Corpus-based methods in analysis of SLA data. In *Handbook of cognitive linguistics and second language acquisition*. Edited by P. Robinson and N. Ellis. Routledge.
- Hanks, P. (2013) *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, Mass. and London.
- Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press
- Nesselhauf, N. (2005) *Collocations in a Learner Corpus*. Amsterdam: John Benjamins
- Ning, M. (2012) 'Implications of Interlanguage Error Analysis and Research on English Language Testing and Teaching'. *Higher Education of Social Science* [online] 2 (2), 4-7. available from <<http://www.cscanada.net/index.php/hess/article/view/j.hess.1927024020120202.1344/2416>> [20th February 2015]
- Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learner-driven data. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213-234). Amsterdam: John Benjamins.
- Sun, J., and Shang, L. (2010) 'A Corpus-based Study of Errors in Chinese English Majors' English Writing'. *Asian Social Science* [online] 6 (1), 86-94. available from <<http://www.ccsenet.org/journal/index.php/ass/article/view/4781/4026>> [20th February 2015]
- Tyne, H. (2009). Corpus oraux par et pour l'apprenant [Spoken corpora by and for the learner]. In A. Boulton (Ed.), *Des documents authentiques oraux aux corpus: Questions d'apprentissage en didactique des langues* (pp. 91-111). Nancy, France: Mélanges CRAPEL. available from <[http://www.atilf.fr/IMG/pdf/melanges/05\\_Tyne.pdf](http://www.atilf.fr/IMG/pdf/melanges/05_Tyne.pdf)> [20th February 2015]
- Verlinde, S. (2010). The Base lexicale du français: A Multi-Purpose Lexicographic Tool. *eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve: Cahiers du CENTAL, 335-342.
- Zheng, C., and Park, T. J. (2013) 'An Analysis of Errors in English Writing Made by Chinese and Korean University Students' in *Theory and Practice in Language Studies* Vol 3 No 8 pp 1342-1351 [online] available from <<http://ojs.academypublisher.com/index.php/tpls/article/view/tpls030813421351/7470>> [20th February 2015]