An introduction to corpus linguistics

Colm Smyth*

Abstract

In any language, it can often be difficult to ascertain as to which word should be used in a given context. In the past, people have mostly made these vocabulary choices by using their intuition. Nowadays, we can use a corpus to help in this decision making process. This paper will give an introduction to the area of corpus linguistics and its methodologies. A brief look will also be taken at the implications for language teachers.

Keywords : corpus, collocation, word-form, mutual information, T-score.

Introduction

In any language it can often be confusing as to which word or phrase should be used in a given situation or, indeed, what the exact meaning of a given word is. Moreover, oftentimes, words which at first appear to have similar meanings and usages may actually be used in slightly different ways. They may have a pattern of usage that is unique to them. In the past these kinds of situations were generally judged by using intuition. However, given the advent of technology and corpus linguistics, it is now possible to study and analyse these patterns of usage. In the past, it simply was not feasible to manually do a meaningful study of this kind.

In this paper, a look will be taken at the area of corpus linguistics. Firstly, a brief outline of what corpus linguistics is will be given. There will then be a description of some of the methodologies behind corpus research, with an emphasis placed on the word- based approach. An outline of collocation and the measurements used to strengthen assumptions will be made from the collocations. Next, there will be a discussion of patterning, usage and phraseology in text. Finally, there will be a brief discussion of implications for the language teacher.

Corpus Linguistics

Corpus linguistics is a relatively new field of linguistic research. It involves the collection of data; spoken, written, or both, and collating it into one or more text files. These text files are then searchable and the resulting data can be further studied for the purpose of linguistic research. Kennedy (1998:1) describes a corpus as 'a body of written text or transcribed speech which can serve as a basis for linguistic research.' An important point to remember, as pointed out by Hunston and Laviosa (2000), is that any information found from research done on a corpus is only applicable for the data studied. It cannot necessarily be applied to the language as a whole. They also point out that any results

＊未来科学部英語系列講師　Lecturer, Department of English Language, School of Science and Technology for Future Life

corpus itself and that when it comes to a corpus, the bigger it is the better. When measuring the size of a corpus, we are interested in the total word count. Aijmer and Altenberg (2001) describe corpus linguistics as '…the study of language on the basis of text corpora.'

A corpus, while having the potential to be limitless in size, is created for the explicit purpose of research and can be tailored to the study of one particular area, for example tabloid or broadsheet journalism, novels, radio broadcasts etc. This applicability to the area of study is, according to Leech (2001:9), what makes a corpus different from a large archive of random data.

Methodologies

There are two main methodologies used for the study of corpora. These are, according to Hunston and Laviosa (2001), **category based** and **word-form based**. A look will now be taken at both of these methodologies.

Category based

This approach to corpus data analysis, according to Hunston and Laviosa (2001), necessitates the putting of all words in the corpora into a particular category, such as verb, adjective, noun, conjunction etc. before any work can be carried out on the corpus. This can be carried out automatically by software known as a **tagger**. Hunston and Laviosa (2001) also point out that this process is not 100% fool proof and there may be some slight errors in the tagging of some words. This necessitates the manual intervention by the researcher to correctly tag any words that were erroneously tagged by the tagging software. This work can be extremely time consuming depending on the size of the corpus being used and, as Hunston and Laviosa (2001:93) point out, it will inevitably be a significant factor in the size of corpus used.

This tagged corpus can now be easily searched for instances of any type of grammatical word. A key point to remember is that once the corpus has been annotated with the word tags for grammatical class it is no longer in its raw, unprocessed, form. The result of this being that, according to Leech (2001:19), words are no longer searched for, instead it is '…grammatical abstractions…' that are examined. This represents a slight shift in the assumed idea of how corpus research might normally be carried out. It allows for the comparison of categories, such as the usage of past and present tense in a selected corpus. This method is best represented by the pioneering work of Biber (1986 and 1988). It is worth noting that once a corpus has been tagged, it cannot be untagged. Therefore, it may be advisable for the researcher to make a back-up copy of the corpus before taking the step of tagging it.

Word-form based

According to Hunston and Laviosa (2001), this approach differs from **category based** in that there is a very minimal tagging of the corpora and any tagging done is fully automated, there is no manual intervention by the researcher to do, or amend, any tagging. The overall result of this difference in approach is that the subject of the study is moved away from the grammatical abstractions of the **category based** approach and instead the focus is placed on the individual words, or phrases, and the ways in which they act within the text.

The **word-form based** approach can help a researcher determine the different meanings which a word has and furthermore the patterns in which this differing meaning tends to occur. To help with this research **collocation** is used.

Collocation

Hunston and Laviosa (2001), state that **collocation** is the propensity for words to occur near each other in a text. In other words, they co-occur, or they are co-located. However, they also point out that just because two words frequently occur near each other, this does not necessarily mean that there is a high significance to this co-occurrence. For instance, for any given word of which the collocates are searched for, there is a high probability that it will collocate with the some of the most frequently occurring words in the English language e.g. the, a, etc. Therefore, the collocate list should not be taken at face value. Hunston (2002:68) states that **collocation** is: '…the tendency of words to be biased in the way they co-occur.' To gain a true idea of the important collocates which a word has, two measurements are applied; these are **mutual information** and **T-score**. These will be discussed in a little more detail later. When calculating the collocates of a word, the search is usually performed within the four words to the left and four words to the right of the search, or **node**, word. This space within which the search is performed is known as the **span** and its idea was put forward by Sinclair *et al* (1970). As noted by Baker (2006:103), the size of the span will have a bearing on the collocates found. In other words, venturing into a bigger span increases the chances of finding words which are not true collocates being included in the results.

Mutual Information

Mutual Information, henceforth referred to as **MI score**, is used to calculate the number or actual occurrences of a word against the number of times that word was predicted to occur. Hunston (2002:71) says that '…MI score measures the amount of non-randomness present when two words occur.' Hunston and Laviosa (2000:16-17), state that this gives a more accurate idea of the relationship between two words. They go on to say that MI score assesses the importance of a collocation and that it shows a clearer picture of the relationship between words than that given by a simple collocation list alone. It is a measurement of two-way attraction. Walter (2010:435) states that because a word that occurs infrequently collocates with another word, it is unlikely that this collocation happens by chance. However, according to Baker (2006:102), one drawback of MI score is that it tends to attach a high significance to words that occur rarely in a text, therefore giving somewhat misleading results. It is therefore not immediately clear how accurate, or usable, the results are. According to Hunston (2002), only MI scores of 3 or higher should be considered to be important. To help verify the importance of any given collocation, as well as calculating MI score, another measurement called T-score is used.

T−Score

This measurement takes into account evidence for the collocation throughout the corpus. Hunston (2002:72) points out that T-score is used to analyse and validate a collocation when we: '…need to know how much evidence there is for it…how certain we can be that the collocation is the result of more than vagaries of a particular corpus.' This differs from MI score in that it gives a clearer insight to which words have a strong attraction to the node word and words which do not occur frequently in the corpus are not given a high significance. Therefore, it is more explicit about the importance of a collocation. But as Hunston and Laviosa (2000) point out, T-score only shows the words which are important to the

node word, not which words the node word is important to. It is a measurement of one-way attraction. According to Hunston (2002), a T-score of 2 or higher should be considered important.

Patterns

When talking about patterns in text, Hunston and Laviosa (2000) state that it is referring to the grammatical patterns in which a word occurs. Regardless of whether a word is a noun, adjective, adverb, pronoun, preposition etc., they all occur in some form of grammatical pattern. These patterns can be analysed and coded into a standardised form. The coding used by Hunston and Laviosa (2000) is that which is also employed by Collins.

The analysing and coding of grammatical patterns helps to show how a word is used and ultimately shows the meaning, or meanings, which a word has in a given pattern, or context. According to Hunston and Laviosa (2000:29), Hunston (2002:138-139) and Sinclair (1991), these different meanings are generally highlighted by being part of differing grammatical patterns. Furthermore, as Hunston (2002:139) points out, a pattern is not necessarily exclusive to one meaning of a word. Differing meanings may share the same pattern, however Hunston (2002:139) reassures that the relationship between the pattern and the meaning still holds true.

Hunston and Laviosa (2000:28) also say that the study of patterns affords us the opportunity to verify whether or not our native speaker intuition is correct and allows for the recognising of a possible change in language behaviour earlier than may otherwise be possible.

Implications for the language teacher

Corpus linguistics has the potential to be a powerful tool in the arsenal of a teacher, whether or not the course in question in specifically linguistics related or not. In particular, a writing class is ideally suited to such study as the teacher could set out rules for the type of files that students submit and dictate the format that file names should take. These files would be immediately ready for inclusion in a specialised corpus for both individual classes and a group of classes. This would allow the teacher to tailor future lessons to the needs of the students as the corpus would help highlight any common or frequent errors and, hopefully, aid in discovering in why this type of error was made. The corpus could also be student specific, which would greatly enhance feedback that a teacher gives.

Creating a corpus for a communication course would, naturally, be more time consuming, but would also offer the same potential benefits. However, it would be quite difficult to create the type of student specific corpus mentioned above.

Bibliography

Aijmer, K. & Altenberg, B. (2001), 'English Corpus Linguistics.' Longman, London.

Baker, P. (2006), 'Using Corpora in Discourse Analysis.' Bloomsbury Academic, London.

Hunston, S., (2002), 'Corpora in Applied Linguistics.' Cambridge University Press, Cambridge.

Hunston, S. & Laviosa, S. (2000), 'Corpus Linguistics.' Birmingham: School of English, CELS.

Kennedy, G. (1998), 'An Introduction to Corpus Linguistics.' Longman, London.

Leech, G. (2001), 'The state of the art in corpus linguistics.' In Aijmer & Altenberg (2001), 'English Corpus Linguistics.' Longman, London.

Sinclair, J., (1991), 'Corpus Concordance Collocation.' Oxford University Press, Oxford.

Sinclair, J., Daley, R. and Jones, S., (1970), 'English lexical studies.' Report No. 5060, Office of Scientific and Technical Information, London.

Walter, E.(2010), 'Using a corpus to write dictionaries.' In O'Keefe & McCarthy (eds.) 'The Routledge Handbook of Corpus Linguistics.' (2010).