Before you attempt to use corpus linguistics for your own investigation, it's a good idea to practise so you're familiar with its use. Here we will practise with two corpus tools: Antconc and Sketch Engine. You'll be using these tools to analyse aspects of gender representation in two *Star Wars* films. By the end of this you'll be able to put these skills to use for your own investigation.

Setting up Antconc

Download Antconc from here: www.laurenceanthony.net/software.html

1. Download Antconc from the website. Save it. Then start it up.

2. Set up Antconc to ignore 'tags', which are special codes used to annotate the texts in the corpus. Click on 'Global settings' on the menu bar, click on 'tags', and then select 'hide tags'. Then click 'apply'.

3. You also need to tell Antconc to include numbers. To do this, click on 'Global settings', click on 'token (word) definition', and select 'number' in the 'number token classes' box. Then click on 'apply'.

4. Now you need to tell Antconc which file(s) you want to work with. Select 'file' and then 'open file(s) …' A standard 'file-open' box will appear. Antconc can only deal with plain text files – so any files in different formats need to be saved as plain text files first.

Collecting texts to be used with Antconc: practice with *Star Wars* scripts

You can of course collect any texts that you wish to investigate, e.g. online news articles about Trump, Brexit etc., tweets from particular celebrities, tweets relating to a particular hashtag such as #metoo.

For this exercise, you'll use the scripts from the very first *Star Wars* film, *Star Wars: A New Hope* and the first film in the latest trilogy, *Star Wars: The Force Awakens*. You can access film scripts at www.imsdb.com. Copy the entire script and paste it into a Word document.

Save the document as **PLAIN TEXT**, which is the format Antconc can read.

- Start with the *Star Wars: A New Hope* script. You will see that the file is now listed in the main Antconc window. Open Antconc again and do the same with the *Star Wars: The Force Awakens* script. You cannot have the two scripts in the same programme, as Antconc will treat them together and you are going to compare the two scripts, so they need to be kept separate.

- Follow the steps on the following sheet. You will be using Antconc to analyse *Star Wars*. Answer the questions in as much detail as you can.

Step one - word frequency lists

To create a word frequency list, click on the 'Word List' tab. Make sure you tick the box near the bottom of the screen that says 'treat all data as lowercase'. (If you don't click it, Antconc will treat 'the' and 'THE' as if they were two different words.) Now click 'start'. The most frequent words tend to be function words such as 'the'. Make a note of any **proper nouns** that appear in the top 10 frequent words for each script.

1. What do you notice about the proper nouns listed in the frequency lists for each script? **Hint:** if you are unfamiliar with the films, research the names of the main human characters for both films.

2. What is the most striking difference with regard to frequencies for the main characters' names in the older and more recent films?

3. The Antconc Wordlist gives the most frequent words in rank order and their % frequencies. Note the % frequency is calculated as follows: %frequency = raw frequency (= the total number the particular word appears in the corpus) / total number of word tokens (= total number of all words in the corpus) x 100.

4. Why do you think it is useful to work out the frequencies as % and not simply use the raw frequency (the total number of appearances of one word in the whole corpus)?

Step two – concordances

You may have noticed that your word list contains letters such as 's' and 'm'. In order to understand why a particular item appears in a wordlist, it's often useful to investigate, using concordances. A concordance is a list of all the occurrences of a certain word, in the context of the sentence(s) that word appears in. You can look at a concordance list simply by clicking on any word in the word frequency list (your cursor should change to a 'pointing finger' icon when you do so).

Alternatively, you can click on the 'concordance' tab in the toolbar at the top of the Antconc screen to take you to the concordance screen. Then type the word you want in the box labelled 'search term' and click 'start'.

Even though you may be able to guess the meaning of individual letters such as 'm' and 's', use the concordance function to find the meaning of any individual letters.

Question: What issues could arise with regard to individual letter 's' in word lists ('s' can mean many things in English!) so that it may be important to always check the 's' in the concordance?

Step three – collocates

Return to the word list for *Star Wars: A New Hope* and click on the 'Collocates' tab. At the bottom, set the search for 'luke' in the box under 'search term', and to the right set the 'window span' to 5L to 5R (this means Antconc will look for collocates with 'luke' in five places to the left and five places to the right of our search term – it is a commonly used setting for collocates).

A collocate is a very strong kind of collocation. The corpus software uses statistical analysis to calculate how strong the co-occurrence of collocates are.

1. Write down the top 10 collocates (words that appear statistically more frequently with 'luke') with their statistical measure rounded to two figures after the decimal point.

2. Now type in the collocation search box 'leia' and keep the same settings. Again, write down the top 10 collocates for 'leia'.

3. Now do the same with the script for *Star Wars: The Force Awakens* and the two characters Rey and Finn.

Question: Which of the two films, based on your findings with step one above and step three's tasks 1, 2 and 3, suggests a more stereotypical representation of male and female characters? What are your reasons for this?

Step four

Although software such as Antconc allows you to find patterns in large texts that are impossible to detect by reading the scripts yourself, you should still use your language knowledge and investigate findings by hand.

1. Return to the collocations for 'luke'. Look at 'sweetheart' – how is this used in the actual script? Is it spoken by Luke or used as a pre or post modifier to modify the proper noun 'Luke' or is there something else going on?

2. How could we have avoided 'sweetheart' appearing in our collocates list?

3. In both the Luke and Leia collocations, the token 'vantage' appears as a strong collocation. Look at where this word is used and explain why this particular noun is likely to occur with characters' names in a film script.

Step five - concordance plot

This function of Antconc gives a picture of the file as a box with black lines marking every instance of a search term occurring in the file.

1. Click on the 'concordance' tab and create a concordance plot for 'Luke' – this gives you a graphic image of how frequently this character is mentioned in the script. You also get the total number of hits. Write this number down.

2. Now make concordance plots for another two characters in *A New Hope*, Leia and Han. Make a note of the total number of hits for each.

3. Now do the same for Rey and Finn from *The Force Awakens*. Also make another concordance plot for Han, who also appears in *The Force Awakens*. Again, make a note of the total number of hits for all characters.

Question: What do the findings from the concordance plot suggest about representation of gender in fantasy/science fiction films?

Setting up Sketch Engine:

You can access Sketch Engine at: www.thesketchengine.eu. You can sign up for a free trial, which gives you access to corpora including English Web 2013, a more contemporary corpus than the BNC. You can also use its corpus making tools, including 'WebBootCaT' which allows you to build your own corpus by searching for texts featuring the search terms you have asked for.

In order to explore some functions of Sketch Engine, you will continue to use the two *Star Wars* scripts that you have already made when working with Antconc as this will allow for comparing the two tools.

- Create a free trial account with Sketch Engine and login. On the homepage there's a menu on the left in turquoise. Click on 'new corpus' in the top right corner and upload the *Star Wars: A New Hope* script by following the onscreen instructions. You will need to click on 'compile' to fully upload the script and make it a corpus on Sketch Engine. Do the same with the *Star Wars: The Force Awakens*. You should now have a dashboard which lists the two scripts as your corpora.

- Sketch Engine makes all the corpora made by its users available everyone else. If you and all your classmates are doing the same activity, there will be several *Star* Wars script corpora appearing on Sketch Engine.

- However, if you are interested in finding out how Sketch Engine works, it is a useful exercise to try to upload your own texts/corpus. Just make sure that you are all members of a class uploading the same texts to give your version a distinctive name. When you have finished with the activities on the *Star Wars* scripts, you can remove the corpora by clicking 'manage corpus' and then 'delete'.

Word Sketch: Click on the corpus of *A New Hope* and you will get a search box to type in a word you'd like to investigate (this is the concordance function). Ignore this and instead click in the left hand menu on 'Word Sketch'. Complete the following steps.

1. In the blank 'lemma box', fill in the name 'luke' and select 'noun' in the drop down menu (from the 'advanced' tab) for 'PoS' (part of speech). Click on 'go'. Sketch Engine will now show you how the lemma (basic word form, including all possible versions such as 'Luke', 'luke', 'luke's' etc.) 'Luke' is used in the script. You can save it as a PDF by clicking on the download symbol in the top right hand corner.

2. Do the same for the names of 'Leia' and 'Han' for *A New Hope*.

3. Return to the home page and now select the *Force Awakens* script corpus. Make a Word Sketch for 'Rey' and 'Finn' respectively.

Now that you have Word Sketches for all the main human characters in the two films answer the following questions:

a. When you look at the category 'verb with [name] as object' you are given a score in the top box (above the list of actual examples). Write down the scores for Luke, Leia, Rey and Finn. This score is a statistical measure, expressing the number of times per million instances of the name, it is used as an object of a verb. Do these findings confirm or contradict any ideas you may have about gender representation in the two *Star Wars* films following your investigations with Antconc?

b. Now look at the table of modifiers for each name. You'll notice that some of these words that modify the names are not modifiers in the grammatical sense – they are not all attributive adjectives or attributive nouns. For example, the name 'Finn' in its Word Sketch is modified by 'Han' and 'Chewie' – the names of other characters. Can you explain why the name of one character might appear in a film script in such a way as to suggest to Sketch Engine that it is pre modifying the name you are investigating?

c. List for each of the four characters (Luke, Leia, Rey and Finn) the attributive adjectives and attributive nouns. What do these descriptions tell us about each character in question with regard to gender representation?

Teacher notes

Step one: word frequency lists

Questions 1,2 and 3

'Luke' is the only proper noun in the Top 10 for *A New Hope*. Interestingly, the character 'Han' appears in sixteenth place, 'Vader' in twenty-eighth and Leia in thirty-eighth place. This is especially interesting as Leia features more frequently in the film (a concordance plot of the script shows this very easily).

'Rey' (fifth), 'Finn'(sixth) and 'Han' (eighth) all occur in the Top 10. This suggests a more balanced (in terms of the two characters Rey and Finn) approach to representing female and male characters.

'Luke' appears about third of the total number of occurrences of the most frequent word, 'the', which fits with Zipf's Law for word frequencies. It is useful to use a statistical measure such as percentages when looking at frequencies rather than absolute numbers, simply because when you compare different texts or corpora, you can make direct comparisons. After all, the absolute frequency of a word depends on how big the text is that is investigated.

It is extremely unusual for a proper noun to feature in a word frequency list – when looking at word frequencies on the BNC, the only words appearing in the top 10 were function words. This finding that proper nouns are so frequent is probably typical for the genre of the text we are investigating – film scripts.

Step two: concordances

A concordance is an essential tool for corpus linguistics, as it shows the word under investigation in its direct context (usually four or five words to left and to the right of it), which can give a researcher a good insight into how a word might be used in a particular corpus. However, this activity focuses on perhaps lesser known corpus linguistics tools. There is a wealth of instructional material such as YouTube video tutorials available on concordances for students who wish to pursue this in more depth.

The activity here reminds students to be careful when looking at word frequency lists or similar and not to jump to conclusions.

Question: it is not surprising that the letter 's' features so commonly in an English corpus – considering it can be used to mark possession, as a contractible auxiliary and a contractible copula.

Step three: collocations

It is important to bear in mind that the collocations provided by a concordancing tool such as Antconc (but also when using CQPWeb or Sketch Engine) are based on a statistical calculation – the number gives a quantifiable measure as to how strong the collocation pattern is for the corpus.

Star Wars: A New Hope – Results from Antconc

Collocates for 'luke'		
rank	word token that collocates with 'luke'	statistical score for strength of collocation
1.	yanked	7.14
2.	wink	6.55
3.	vantage	6.55
4.	unseen	6.55
5.	uniform	6.55
6.	sweetheart	6.55
7.	surfaces	6.55
8.	intrigued	6.55
9.	greeted	6.55
10.	grab	6.55

Collocates for 'leia'		
rank	word token that collocates with 'leia'	statistical score for strength of collocation
1.	softly	8.92
2.	frustrating	8.92
3.	contracting	8.92
4.	announcing	8.92
5.	wrapped	7.92
6.	victim	7.92
7.	vantage	7.92
8.	translate	7.92
9.	traitor	7.92
10.	tracking	7.92

Star Wars: The Force Awakens – Results from Antconc

Collocates for 'rey'

rank	word token that collocates with 'rey'	statistical score for strength of collocation
1.	dazed	7.44
2.	yours	6.85
3.	tussle	6.85
4.	tennis	6.85
5.	swiftly	6.85
6.	pilex	6.85
7.	notices	6.85
8.	none	6.85
9.	mother	6.85
10.	match	6.85

Collocates for 'finn'

rank	word token that collocates with 'finn'	statistical score for strength of collocation
1.	suppress	7.55
2.	yours	6.96
3.	targeting	6.96
4.	sceptical	6.96
5.	shocks	6.96
6.	selection	6.96
7.	searches	6.96
8.	rushed	6.96
9.	possibly	6.96
10.	plops	6.96

Looking at the frequency lists and the collocations for the two films, there appears to be some suggestion of gender stereotyping in *A New Hope* – Leia's character is frequently featured in the film's narrative, but the character's name appears much less frequently than the male characters. Of course, this could be the result of the fact the film has three central hero characters: Luke, Han and Leia. However, in *The Force Awakens*, there are two central heroic characters, Rey and Finn, but a cast of supporting heroic characters, including Han, Pau (both male) and Leia (female). So, it appears that Rey in the 2015 follow-up is more of a central character than Leia was in the first film.

The collocates for Rey and Finn do not feature any example of words with particular gendered connotations. When looking at the collocates for Luke and Leia, however, there are some examples of collocates for Leia that suggest femininity e.g. 'victim' and 'softly'.

Step four: collocations – continued

1, 2 and 3

The collocate 'sweetheart' for Luke could be used to refer to Luke or could be said by him to another character. However, when looking at the concordance for 'sweetheart', it is clear that the word is actually spoken by another character (Han) to address Leia.

By setting our collocates search originally at 5 Left and 5 Right, the speech by Han, which included the address 'sweetheart' was included. In a film script, the name of the character is frequently repeated (every time they are meant to speak or do something), so Han's speech was linked to the name 'Luke' in the margin of the script when the character of Luke was intended to follow Han's speech. By setting our search window a little smaller, e.g. at 2 Left and 2 Right, this might be avoided. Although it could still happen as the nature of film scripts makes the data rather unusual in terms of structure.

Step five: concordance plots

Luke:	appears 726 times in the whole script for *A New Hope*
Leia:	appears 141 times in the whole script for *A New Hope* (R2 – a robot character - appears 150 times in the whole script)
Han:	appears 269 times in the whole script for *A New Hope*
Rey:	appears 446 times in the whole script for *The Force Awakens*
Finn:	appears 414 times in the whole script for *The Force Awakens*
Han:	appears 321 times in the whole script for *The Force Awakens*

Working with word sketch

a. Word Sketch results - characters as object

Luke:	4.15 per million as object of a verb in *A New Hope*
Leia:	5.67 per million as object of a verb in *A New Hope*
Rey:	5.91 per million as object of a verb in *The Force Awakens*
Finn:	6.00 per million as object of a verb in *The Force Awakens*

The older film, *A New Hope*, appears to be more stereotypical in its representation as the central female character as passive – Leia is more frequently the object of a verb in the script. However, in the more recent *The Force Awakens* the gap between Rey and Finn is much smaller and it is the male character of Finn, who appears to be more frequently featured as an object of a verb (in the context of the story, this makes sense – Finn is a soldier at the start of the film who is told what to do and he is also taken prisoner when he refuses to execute his duties).

Stereotypically, Hollywood action and adventure films tend to feature far fewer female characters than male characters and research by corpus linguists at Huddersfield University has demonstrated some stereotypical representations of male and female characters in this particular type of film, of which the *Star Wars* franchise is a sub genre. The original trilogy produced in the late 1970s and early 1980s featured far fewer female characters as well as very few ethnic minority characters (although in some respects the original trilogy of *Star Wars* was quite ground-breaking in that Leia is not just a passive damsel in distress and the second and third film feature a significant part for African-American actor Billy Dee Williams). The contrast with the 2015 'reboot' of *The Force Awakens* is striking - the main characters feature actors from a range of different ethnicities, for example and the choice of Rey as the central heroic character marks a real shift away from the gender stereotypes of the genre. The initial findings from applying corpus tools to the scripts support this.

b and c: Word Sketch results - modifiers with character names

The finding that one character's name is modified (according to Sketch Engine's 'Word Sketch' function) with the name of another character can be explained by the fact that the text investigated is a film script. Characters appear together in scenes and the script will list their names at the start of each scene – this could explain why Sketch Engine 'misinterprets' names as modifying other names.

Here are the attributive nouns and adjectives that modify each of the four characters:

A New Hope	
Character name	**Modifiers (att. adjectives and att. nouns)**
Luke	master, sir
Leia	princess, senator, concerned, beautiful, young

The Force Awakens	
Character name	**Modifiers (att. adjectives and att. nouns)**
Rey	no examples – no nouns/adjectives to describe Rey in script.
Finn	wide-eyed, grateful, embarrassed, unconscious, exhausted, tight

These findings suggest a change with regard to representation of gender in the 38 years that separate the release of the original *Star Wars* film (now titled *A New Hope*) and the 2015 instalment of *The Force Awakens*. A quick comparison of Leia with Rey – 'Leia' is modified twice with adjectives that tend to be more associated with females: 'beautiful' and 'young'. According to the story, Luke and Leia are twins, so in theory, Luke could also be modified with the adjective 'young', but this does not happen.

However, while it appears that Luke is pre-modified by honorific terms indicating respect such as 'master' and 'sir', it needs to be borne in mind that it is only the character of Luke's robot, C3PO, who addresses him as such in the film. The character of Leia is a princess and a senator and in some of the scenes featuring her, she is addressed by her titles.

A concordance search of the terms 'master' and 'sir' would show this and while this activity has not touched on concordances, it may be worthwhile pointing out to students that researchers will use these findings as a starting point for further investigations rather than as firm conclusions! Similarly, the use of Leia's formal titles is mostly in scenes where she is the prisoner of the evil empire and is being interrogated by the film's villain of Darth Vader. The formal register of these scenes could be regarded as reflecting the uniformity and oppressive nature of the empire – as the rebels, including the film's protagonists are noticeably more informal in their interactions. Again, further investigation could focus on the lexical choices to create heroic and villainous characters.